Recognition of Printed Text Based on Hidden Markov Model

Ghayda A.A. Al-Talib **Armanesa Nuaman Hasson**

ghaydabdulaziz@uomosul.edu.iq

College of Computer Sciences and Mathematics

University of Mosul *University of Tikrit*

Received on: 26/5/2010 **Accepted on:25/10/2010**

ABSTRACT

Automatic recognition of printed text is of high importance in modern IT applications. Recognition of text for lateen scripted language is readily in use for a long time. For cursive script languages (such as Arabic language) recognition of text is not available as a robust one with a reliable performance. More improvements still exist to reduce average of incorrect words, rather then no constraints on the limit of words of a specific language.

Numerous approaches were tried in recognition of text but recognition of Arabic text based on Hidden Markov model seems to be the most promising one because of its ability to discriminate cursive scripts.

This paper provides an off-line system to recognize printed Arabic text by using hidden Markov model with the aid of the algorithm that segment the text lines into connected parts then into characters.

By looking on the results given by the designed recognition system it is found that a recognition rate (94.9 %) can be achieved. Such rate is in the same order of rates of recognition researches viewed in previous studies. This rate can still be improved. The language used in building the system is Matlab V7.6 (R2008a).

Keywords: Character Recognition, HMM

التعرف على النص العربي المطبوع باستخدام نموذج ماركوف الخفي

غيداء عبد العزيز الطالب **ارمانيسة نعمان حسون**

قسم علوم الحاسوب *قسم علوم الحاسوب*

جامعة الموصل *جامعة تكريت*

تاريخ استلام البحث: 2010/5/26 **تاريخ قبول البحث: 2010/10/25**

الملخص

التعرف الآلي على النص المطبوع له أهمية كبيرة في تطبيقات تكنولوجيا المعلومات الحديثة. فالتعرف على النص المكتوب باللغة اللاتينية تم استخدامه منذُ فترة طويلة. أما بالنسبة للغات المكتوبة بأحرف متصلة (كاللغة العربية) فان نظام التعرف على النص غير متوفر كنظام قوي موثوق في أدائه. فما يزال هناك متسع للتحسينات فيما يتعلق بتخفيض معدل الكلمات الخاطئة، فضلاً عن عدم التقيد بحصيلة لغوية معينة.

لقد جُرِبَت عدة مناهج في مجال التعرف على النص، ويبدو أن التعرف على النص العربي القائم على نموذج ماركوف الخفي هو الأكثر وعداً وذلك بسبب قدرته على تمييز الكتابة المتصلة.

نُقدِم في هذا البحث نظام يعمل بأسلوب off-line للتعرف على النص العربي المطبوع باستخدام نموذج ماركوف الخفي مع الاستعانة بخوارزمية تقطيع السطر النصي إلى مقاطع ثم حروف.

حقق النظام المقترح نسبة انجاز قدرها (94.9%) وهي نسبة تقع ضمن بحوث التعرف المنجزة، وتبقى هذه النسبة قابلة للتحسين.

استخدمت Matlab V7.6 (R2008a) كلغة برمجية في بناء النظام المقترح.

الكلمات المفتاحية: تمييز الانماط، نموذج ماركوف الخفي

1- المقدمة

يُعد الذكاء الاصطناعي من أهم المجالات العلمية التطبيقية في علوم الحاسبات فقد تعددت وتنوعت التطبيقات البرمجية في هذا المجال فشملت معالجة اللغات الطبيعية والترجمة الآلية وتمييز الأنماط... الخ. ويعد تمييز الأنماط(patterns) احد التطبيقات البرمجية للذكاء الاصطناعي.[1]

كما أن تمييز الأنماط هو دراسة كيف يُمكن للمكائن أن تُلاحظ البيئة، فتتعلم إظهار أنماط ترغب بتمييزها وتتخذ قرارها المعقول حول أصناف تلك الأنماط.[2]

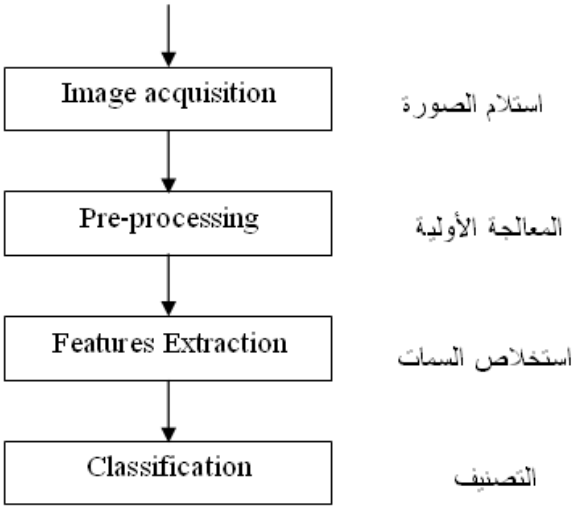
وقد عرف Watanbe النمط كنظير للفوضى، بأنه كيان مُعرف بشكل غير واضح مُمكن أن يُعطى اسماً معيناً. [3]

وبالرغم من التحسينات التدريجية في تطبيقات علم تمييز الأنماط في العقد الأخير من القرن العشرين وأوائل هذا القرن، يبقى تمييز الحروف واحد من أهم مسائل تمييز الأنماط. [4]

ومن تلك التطبيقات تمييز الحروف بصرياً (OCR Optical Character Recognition)- ويتم فيها قراءة العنوان البريدي على المظروف، أرشفة واسترجاع النص، ترقيم المكتبات ... الخ. يمر OCR بعدة مراحل و آخر مرحلة فيه هي التمييز حيثُ توجد عدة طرق لإجرائها، وسوف نستخدم في هذا البحث نموذج ماركوف الخفي (Hidden Markov Model-HMM) في تمييز النص العربي المطبوع. فنموذج ماركوف الخفي HMM واحد من النماذج المستخدمة في معالجة الكلام واللغات. [5] ويُعرف HMM كعملية تصادفيه مزدوجة فيها حالات مخفية يُمكن مشاهدتها فقط من خلال مشاهدات معينة. [6]

2- النموذج العام لنظام تمييز الحروف

يتكون نظام التمييز بصورة عامة من أربعة مراحل أساسية يوضحها المخطط في الشكل- 1، حيث تبدأ بإدخال الوثيقة التي تحتوي على النص المراد تمييزه وتنتهي بتصنيف حروف الوثيقة المُدخلة. [7]



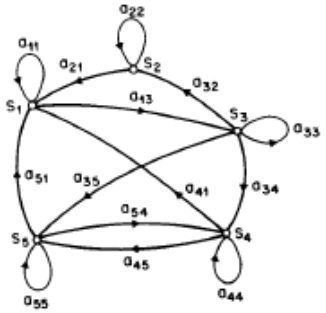
(الشكل–1) المخطط العام لنظام التمييز

وقد لا يحتوي نظام التمييز على جميع هذه المراحل إذ يتم اختزال بعض المراحل بدون أن يؤثر ذلك على عملية التمييز. فمثلاً يقوم النظام بالتمييز دون أن يحتاج إلى مرحلة استخلاص السمات، و يستخدم بدلاً عنها مطابقة القوالب (Templates Matching).

3– سلاسل ماركوف Markov Chains

النماذج الرياضية ممكن أن تكون محددة (Deterministic) أو تصادفيه (Stochastic). ومع ذلك فأن في عدة حالات اجتماعية وحياتية هُناك ظواهر تصادفيه (وهي ظواهر ذات سلوك غير قطعي لا يمكن السيطرة عليها بشكل تام أو التنبؤ بسلوكها المستقبلي بشكل مؤكد ويُطلق عليها مصطلح العمليات التصادفية. [8] فيصبح النموذج التصادفي هو الأكثر ملائمة لتمثيلها.

المنظومة الموضحة في الشكل–2 يُمكن أن توصف خلال أي فترة زمنية، كأن تكون موصوفة في واحدة من مجموعة الحالات المتقطعة(N) (Discrete states) ($S_1,S_2,\ldots,S_N$).



(الشكل–2) سلسلة ماركوف لـ(5) حالات مع انتقالاتها

وخلال تلك الأزمنة المتقطعة، تخضع المنظومة إلى تغيرات في الحالة (من الممكن الرجوع إلى الحالة نفسها) وفقاً لمجموعة من الاحتمالات المرتبطة بالحالة. ويُرمز إلى الزمن المرتبط بتغير الحالة بـ(t=1,2,…)، ويُرمز للحالة الحقيقية خلال الزمن(t) بـ (Q_t). إن وصف الاحتمالية بصورة كاملة للمنظومة أعلاه يتطلب وصف الحالة الحالية عند الزمن(t)، فضلاً عن كل الحالات السابقة لها. [6] فيُنظر إلى سلسلة ماركوف كنوع من مخطط

الاحتمالات (Probabilistic Graphical Model) أو طريق لتمثيل الفرضيات الاحتمالية. وسلسلة ماركوف محددة بالمكونات التالية: [5]

1- مجموعة N من الحالات وتمثل بـ $Q=\{q_1,q_2,\ldots,q_N\}$

2- المصفوفة الاحتمالية الانتقالية A (transition probability matrix) وتمثل بـ

$$A=\begin{pmatrix} a_{11} & \ldots & a_{1N} \\ . & . & . \\ . & . & . \\ . & . & . \\ a_{N1} & \ldots & a_{NN} \end{pmatrix}$$

حيث أن كل a_{ij} تُمثل احتمالية الانتقال من الحالة i الى الحالة j بحيث تحقق الشرط التالي:

$$\forall i \qquad \sum_{j=1}^{N} a_{ij} = 1$$

3- حالات خاصة هي حالة البداية q_0 وحالة النهاية q_F التي لا ترتبط مع أية مشاهدات (Observations).

4- التوزيع الاحتمالي الابتدائي على الحالات (Initial probability distribution)

$$\pi = \pi_1, \pi_2, ..., \pi_N ;$$

وكذلك $\sum_{i=1}^{N} \pi_i = 1$

وتكون الاحتمالية (probability) التي تبدأ بها سلسلة ماركوف عند الحالة i في بعض الحالات $\pi_i = 0$ يعني لا يُمكن أن تكون الحالة ابتدائية (initial state).

وتعرّف فرضية ماركوف بالعلاقة التالية:

$$P(q_i \backslash q_1 \ldots q_{i-1}) = P(q_i \backslash q_{i-1})$$

حيثُ أن

$P(q_i \backslash q_1 \ldots q_{i-1})$ تمثل احتمالية حدوث الحالة q_i عند توفر الحالات $q_1 \ldots q_{i-1}$

و$P(q_i \backslash q_{i-1})$ تمثل احتمالية حدوث الحالة q_i عند توفر الحالة q_{i-1} التي تسبقها فقط

4- نموذج ماركوف الخفي Hidden Markov Model

نموذج ماركوف الخفي (HMM) عبارة عن نظام محطات الآلة المحدودة (finite state machine) القادر على توليد مشاهدات باحتمالية انتقال الحالة عند الزمن t التي تعتمد فقط على الحالة السابقة لها عند الزمن t-1. علماً أن تسلسل الحالة التي تنتج المشاهدة المُعطاة مجهول. [9] لذا، ففي نموذج ماركوف الخفي تكون الحالة ليست مرئية، لذلك سُميَّ بنموذج ماركوف الخفي والانتقالات بين الحالات تحكمها مجموعة من الاحتمالات يُطلق عليها احتمالات الانتقال من حالة معينة والتي يُمكن أن تنتج نتيجة أو مشاهدة وحسب توزيع الاحتمالية المرتبط بتلك الحالة. [10] والاختلاف بين نموذج ماركوف الخفي ونموذج ماركوف هو وجود الاحتمالات الإضافية. ويُمثل هذا الجزء الخفي للنموذج ويرتبط بالمشاهدة الناتجة من كل حالة [11]. فنموذج ماركوف الخفي هو نموذج تصادفي قادر على التصنيف الإحصائي. ولذلك فقد طُبِقَ في تمييز الصوت وتمييز الكتابة اليدوية بسبب قدرته على التكيف وتعددية الاستخدام في معالجة الإشارات المتسلسلة [12]. كما طُبِقَ نموذج ماركوف الخفي في مجتمع المعلوماتية الحياتية (Bioinformatics) لإيجاد سلاسل DNA. [13] وكذلك طُبِقَّ في تصميم

أنظمة كشف التطفل على الشبكات (حيثُ تُعنى أنظمة كشف التطفل بحماية الشبكات من الهجمات و/أو سرقة البيانات المهمة من قبل المستخدمين المخولين أو الغرباء). [14]

تُعرف عناصر نموذج ماركوف الخفي (HMM) كالتالي: [6]

N: عدد الحالات في النموذج، فبالرغم من أن الحالات مخفية إلا أن للعديد من التطبيقات الطبيعية هناك في اغلب الأحيان بعض الأهمية المتعلقة بالحالات أو بمجموعة الحالات من النموذج ويُمكن تمثيل فضاء الحالة (S) كما يلي:

$$S=\{S_1,S_2,\ldots,S_N\}$$

حيثُ يُرمز للحالة عند الزمن (t) بـ (q_t).

- M: عدد رموز مشاهدات الحالة الواحدة. ويُمكن تمثيل رموز المشاهدة الواحدة كما يلي:

$$V=\{v_1,v_2,\ldots,v_M\}$$

- التوزيع الاحتمالي للحالة الانتقالية (A):

$$A=\{a_{ij}\}$$

حيثُ

$$a_{ij}=p[q_{t+1}=S_j \backslash q_t=S_i] \quad , \quad 1 \le i, j \le N$$

- التوزيع الاحتمالي لرمز المشاهدة عند الحالة j

$$B=\{b_j(k)\}$$

حيثُ

$$B_j(k)=p[v_k \text{ at } t \backslash q_t=S_j] \quad , \quad \begin{matrix} 1 \le j \le N \\ 1 \le k \le M \end{matrix}$$

- توزيع الحالة الابتدائية:

$$\pi = \{\pi_i\}$$

حيثُ

$$\pi_i = p[q_1 = S_i] \quad , \quad 1 \le i \le N$$

وبإعطاء القيم المناسبة لكل من (N,M,A,B,π) يكون بالإمكان استخدام نموذج ماركوف الخفي كمولد لمتسلسلة المشاهدات O.

$$O=O_1\ O_2\ \ldots O_T$$

ويُمكن أن يُمثل نموذج ماركوف الخفي بالمعلمة $\lambda = (\pi, A, B)$ [15]

حيث أن:

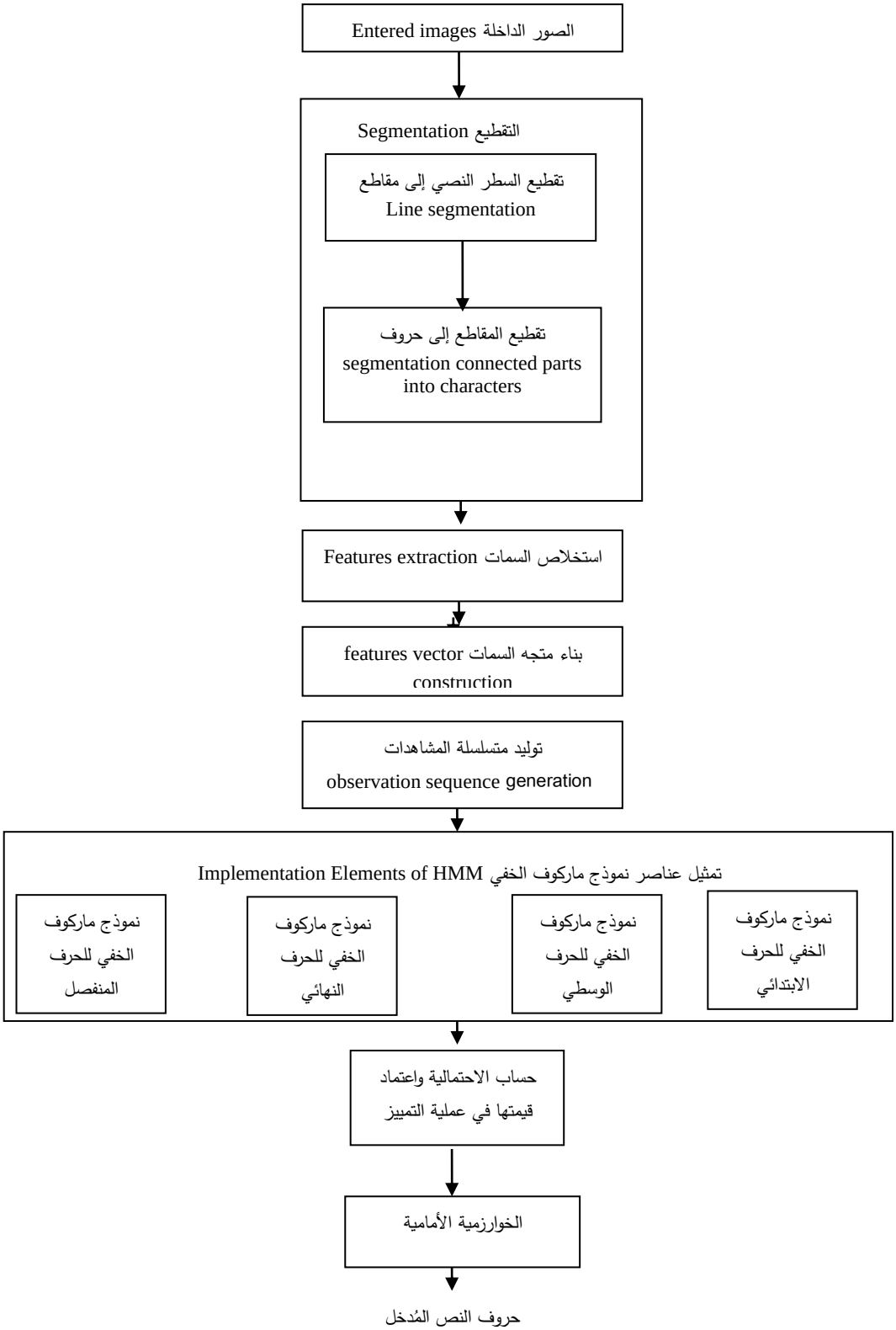
π تُمثل احتمالية الحالة الابتدائية.

و A تُمثل مصفوفة احتمالية انتقال الحالة.

و B تُمثل احتمالية مشاهدة الرمز عند الحالة i.

5- النموذج المقترح لنظام تمييز النص العربي

يتكون نظام التمييز المقترح من مراحل أساسية تبدأ بإدخال النص المراد تمييزه وتنتهي بمرحلة التمييز باستخدام نموذج ماركوف الخفي. والمخطط في الشكل-3 يوضح مراحل تنفيذ النظام المقترح للتمييز.



(الشكل-3) مخطط نظام التمييز المقترح

6- خطوات تنفيذ النظام المقترح

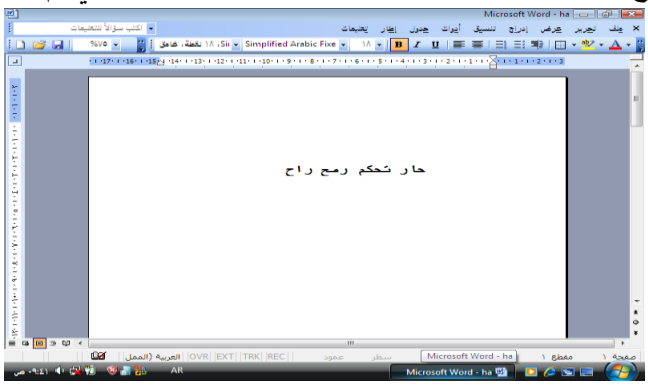
تم في هذا البحث اقتراح خوارزمية لتمييز النص العربي المطبوع باستخدام نموذج ماركوف الخفي Hidden Markov Model. حيثُ تمت البرمجة باستخدام لغة Matlab V7.6 (R2008a) لعدة خطوات نُفِذت بالتعاقب لأجراء عملية التمييز (Recognition).

1-6 مرحلة التدريب

يبدأ عمل النظام بمرحلة تدريب نماذج ماركوف المصممة والتي تضمنت الخطوات التالية:

1-1-6 إدخال الصور

تم في هذه الخطوة طباعة 28 سطراً نصياً كل سطر تمت طباعته على حدا باستخدام خط من نوع (simplified Arabic fixed) بحجم (18) في برنامج معالج النصوص (Word 2003) بحيث يتضمن كل سطر نصي حرف معين بجميع أشكاله. يبين الشكل-4 سطر واحد من 28 سطر نصي تم إدخالها.



(الشكل -4) حرف الحاء بأشكاله الأربعة

وبعد ذلك يُخزن كل سطر نصي على شكل صورة ثنائية في ملف نوعه BMP. تُخزن البيانات في الصورة الثنائية بصيغة (1,0) إذ تُمثل النقطة السوداء التي تكون جزءاً من النمط بالقيمة 0 والنقطة البيضاء بالقيمة 1 والتي لا تكون جزءاً من أي نمط. بعدها تُقرأ الصورة وتُخزن في مصفوفة ثنائية، ليتم بعدها إجراء العمليات اللاحقة عليها لغرض الحصول على النص المقابل للصورة.

لم نُجر عملية تقليل الضوضاء بسبب عدم إدخال الصورة عن طريق أجهزة المسح البصري مثل الماسح الضوئي (scanner) أو أجهزة الكتابة مثل القلم الضوئي (penlight) التي تُسبب وجود الضوضاء.

2-1-6 مرحلة التقطيع

وتعد مرحلة التقطيع مرحلة مهمة ضمن مراحل نظام تمييز النص العربي بسبب طبيعة الكتابة العربية المتصلة التي تتطلب فصل تراكيب أنماط الحروف الواحدة عن الأخرى. ويتم التقطيع الآلي بخطوتين: في الخطوة الأولى يتم تقطيع السطر النصي إلى كلمات و/أو مقاطع وذلك باستخدام المدرج التكراري العمودي، بعدها يتم تقطيع كل كلمة و/أو مقطع إلى الحروف المكونة لها، وفيما يلي شرح تقطيع المقاطع إلى حروف:

إن عملية استقطاع الحرف تتم بعد تحديد بدايته ونهايته فضلاً عن إيجاد:

1- خط الأساس Base line: يكون عند الخط line الذي يملك اكبر عدد من النقاط الضوئية السوداء.

2- الخط العلوي Top line لكل عمود في المقطع.

3- الخط السفلي Bottom line لكل عمود في المقطع.

4- حد العتبة Threshold: يُقابل اكبر قيمة مكررة في المدرج التكراري لكل عمود الذي تم إيجاده في خطوة التقطيع السابقة.

5- عدد الانتقالات العمودية من (0-1) أو (1-0).

عمود البداية للحرف يكون المدرج التكراري له اكبر من حد العتبة، بينما عمود النهاية يجب أن يُحقق شروط هي:

أ- الخط العلوي لهذا العمود يكون اقل أو يساوي خط الأساس.

ب - الخط السفلي لهذا العمود يكون اكبر من أو يساوي خط الأساس.

ج- الفرق بين الخط السفلي والخط العلوي يكون اقل أو يساوي حد العتبة.

ء- المدرج التكراري له اقل أو يساوي حد العتبة .

هـ- عدد الانتقالات العمودية تساوي اثنين.

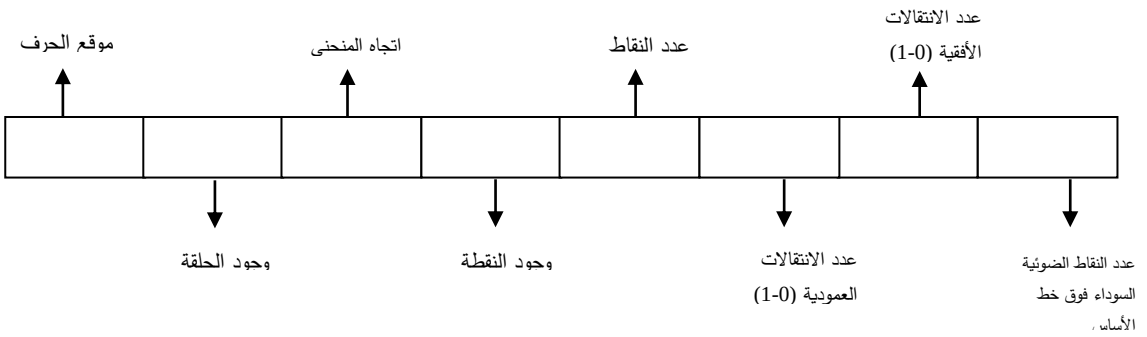
و- الخط العلوي لهذا العمود يكون اكبر من الخط العلوي لعمود البداية.

3-1-6 استخلاص السمات

تم في المرحلة السابقة الوصول إلى كل حرف ومعرفة بدايته ونهايته والمساحة التي يشغلها الحرف، وفي هذه المرحلة تجرى عملية استخلاص السمات لغرض توليد متسلسلة المشاهدات، ثم استدعاء نموذج ماركوف الخفي المُصمم حسب موقع الحرف، لكي يتم بعدها حساب الاحتمالية لمتسلسلة مشاهدات الحرف وإخراج الحرف المميز. وتُكرر هذه الخطوات على بقية الحروف بالتتابع.

4-1-6 بناء متجه السمات

يتكون متجه السمات من ثمانِ متغيرات كل متغير يمثل سمة من السمات التي وجدناها سابقاً ويكون ترتيب عناصر متجه السمات كما في الشكل-5.



(الشكل-5) عناصر متجه السمات

وبعد إيجاد متجه السمات لكل حرف وبجميع أشكاله تتكون أربعة جداول للسمات حسب موقع الحرف في الكلمة.

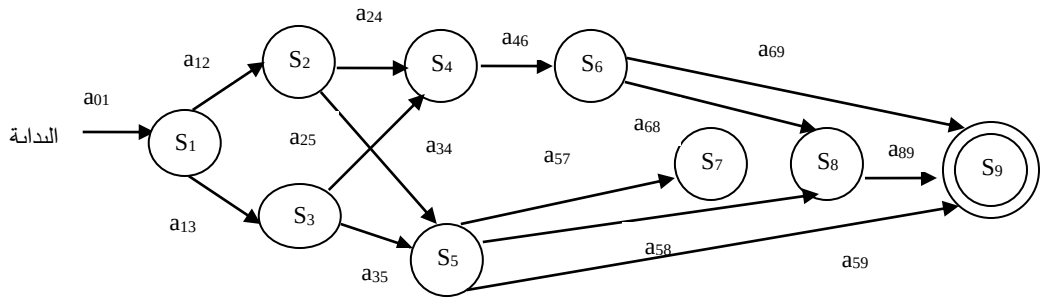
5-1-6 تمثيل عناصر نموذج ماركوف الخفي

تم تصميم أربعة نماذج لتمييز النص العربي المطبوع حسب موقع الحرف في الكلمة (ابتدائي، وسطي، نهائي، أو منفصل).

ونوع نموذج ماركوف المستخدم هو نموذج اليسار–إلى–اليمين المتوازي (parallel left-to-right) الذي ينسجم مع هيكل النماذج المصممة للتمييز. وفيما يلي شرح عناصر كل نموذج.

6-1-6 عناصر نموذج ماركوف الخفي المصمم للحرف الابتدائي

لقد تم تصميم النموذج بحيث يضم تسع حالات يوضحها الشكل–6.



(الشكل–6) نموذج ماركوف الخفي للحرف الابتدائي– تسع حالات

وكانت عناصر نموذج ماركوف الخفي للحرف الابتدائي كما يلي:

1– احتمالية توزيع الحالة الابتدائية: وهي احتمالية حدوث الحالة S_i عندما i=1,2,…9 عند الزمن t=1. وتوضع في متجه π تكون أبعاده 1*N، حيثُ N تُمثل عدد الحالات (N=9).

$$\pi = [1.0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0]$$

2– مصفوفة احتمالية الانتقال بين الحالات A حجمها N*N وحسب النموذج المصمم كان حجمها 9*9، حيث ان N تمثل عدد حالات نموذج ماركوف المصمم. والجدول–1 يوضح قيم المصفوفة A.

(الجدول–1) احتمالية الانتقال بين الحالات للحرف الابتدائي

State	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
S_1	**0**	**0.3636**	**0.6364**	**0**	**0**	**0**	**0**	**0**	**0**
S_2	**0**	**0**	**0**	**0.5**	**0.5**	**0**	**0**	**0**	**0**
S_3	**0**	**0**	**0**	**0.6429**	**0.3571**	**0**	**0**	**0**	**0**
S_4	**0**	**0**	**0**	**0**	**0**	**1.0**	**0**	**0**	**0**
S_5	**0**	**0**	**0**	**0**	**0**	**0**	**0.2222**	**0.2222**	**0.5556**
S_6	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0.3077**	**0.6923**
S_7	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1.0**
S_8	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1.0**
S_9	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1.0**

يليها حساب التوزيع الاحتمالي لمسار رموز مشاهدات الحروف في مواقعها الأخرى. حيث كانت مصفوفة التوزيع الاحتمالي لرموز المشاهدات B (حجمها N*M) وحسب بيانات التدريب 17*9 حيث ان N تمثل عدد حالات نموذج ماركوف المصمم وان M تمثل عدد رموز المشاهدات المتوقعة عند كل حالة، كما في الجدول–2 حيث يوضح قيم المصفوفة B.

(الجدول–2) التوزيع الاحتمالي لرموز المشاهدات– الحرف الابتدائي

S \ 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
S_1	1.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S_2	0	0.375	0.375	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0
S_3	0	0	0.6429	0.1429	0.1429	0.0713	0	0	0	0	0	0	0	0	0	0	0
S_4	0	0	0	0	0	0	0.7692	0.2308	0	0	0	0	0	0	0	0	0
S_5	0	0	0	0	0	0	0	0	1.0	0	0	0	0	0	0	0	0
S_6	0	0	0	0	0	0	0	0	0	0.6154	0.2308	0.1538	0	0	0	0	0
S_7	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0	0	0
S_8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0
S_9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0

بالطريقة نفسها تم تصميم نموذج ماركوف الخفي للحرف الوسطي والنهائي والمنفصل وحساب عناصر النموذج وعناصر النموذج هي:

1– احتمالية توزيع الحالة الابتدائية

2– مصفوفة احتمالية الانتقال بين الحالات A

3– ومصفوفة التوزيع الاحتمالي لرموز المشاهدات B

7-1-6 حساب الاحتمالية

بعد استخلاص المشاهدات وتوليد متسلسلة المشاهدات لكل حرف حسب موقعه في الكلمة، تم تطبيق الخوارزمية الأمامية (forward algorithm) لحساب احتمالية المشاهدات ، حيث تعمل الخوارزمية الأمامية على حساب احتمالية المشاهدات وذلك بجمع احتمالات مسارات جميع الحالات الخفية التي بإمكانها أن تنتج متسلسلة المشاهدات' وطُبِقت على بيانات التدريب لحساب احتمالية متسلسلة المشاهدات O لكل الحروف عند وجود النماذج المصممة حسب موقع الحرف ضمن الكلمة فتكونت أربعة جداول خاصة بالاحتمالية المحسوبة لمتسلسلة المشاهدات لكل حرف حسب موقعه في الكلمة، ندرج منها الجدول–3 الخاص بالحرف الابتدائي.

(الجدول–3) نموذج من نتائج تطبيق الخوارزمية الأمامية للحرف الابتدائي

الحرف	الاحتمالية المحسوبة للحرف الابتدائي
بـ	0.007235288619995
تـ	0.05878829956054
ثـ	0.03919219970703
جـ	0.007235288619995
حـ	0.0230809432983398
خـ	0.024117469787598
سـ	0.07792207792208

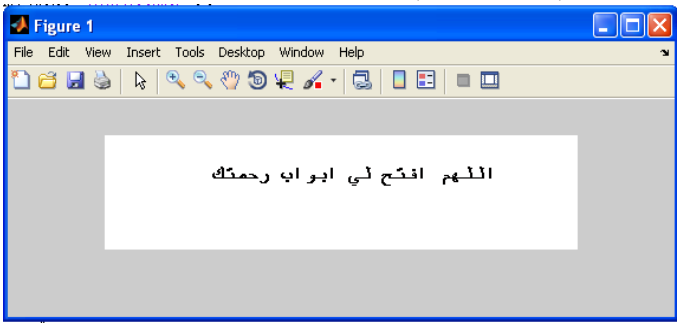
كما تم حساب الجدول أعلاه للحرف بمواقعه الأخرى (الوسطي، النهائي، والمنفصل) كمتطلب للخوارزمية الأمامية.

2-6 مرحلة الاختبار

لغرض اختبار كفاءة النموذج المصمم في التمييز تطبق الخطوات الآتية:

6-2-1 إدخال صورة السطر النصي

تم اختيار صورة السطر النصي الموضحة في الشكل–7 كمثال لتطبيق نظام التمييز المُقترح عليها.

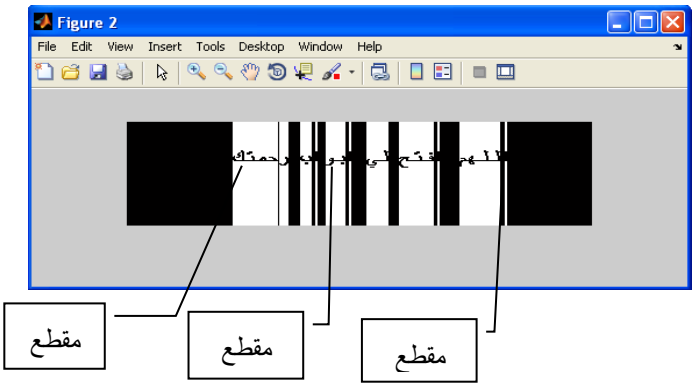


(الشكل–7) صورة السطر النصي المدخل

عملية الإدخال تبدأ بطباعة سطر نصي في برنامج معالج النصوص (Microsoft word 2003) استخدم خط من نوع Simplified Arabic fixed بحجم 18، ومن ثم خزنه على شكل صورة ثنائية في برنامج paint بملف نوعه BMP. بعدها يُقرأ السطر النصي في البرنامج المكتوب بلغة Matlab V 7.6 R2008a ويخزن في مصفوفة ثنائية.

6-2-2 تقطيع السطر النصي إلى مقاطع

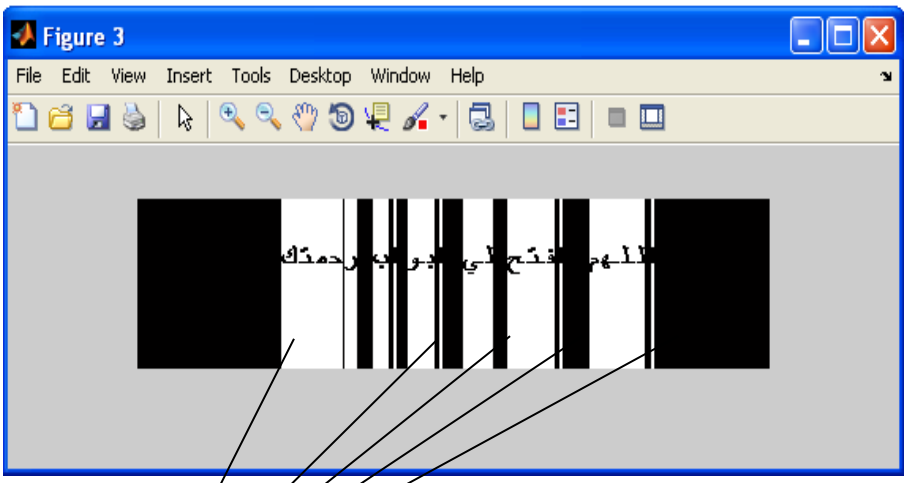
يقطع السطر النصي في هذه الخطوة إلى المقاطع المكونة له، وبعد تحديد قيمة عمود البداية والنهاية لكل مقطع يتم خزنها للاستفادة منها في خطوة التقطيع إلى حروف. وناتج هذه الخطوة هو الحصول على المقاطع المكونة للسطر النصي المُدخل كما في الشكل–8 الذي يوضح صورة السطر النصي بعد تقطيعه إلى مقاطع.



(الشكل–8) صورة السطر النصي بعد تقطيعه إلى مقاطع

6-2-3 تقطيع المقاطع إلى حروف

يتم تقطيع كل مقطع تم الحصول عليه من الخطوة السابقة إلى الحروف المكونة له بعد تحديد عمود بداية ونهاية كل حرف. وناتج هذه الخطوة يوضح في الشكل–9 حيثُ يتم تحويل نقطة سوداء واحدة إلى بيضاء لتكوين فراغات تفصل بين الحروف المقطعة.



(الشكل–9) ناتج تقطيع السطر النصي إلى حروف

4-2-6 استخلاص السمات

وبعد إكمال عملية التقطيع (segmentation) والوصول إلى الحروف المكونة للسطر النصي تم تحديد عمود البداية والنهاية لكل حرف، وبذلك تحددت مساحة العمل على الحرف. يتم إجراء هذه الخطوة والخطوات اللاحقة على كل حرف في المقطع ومن اليمين إلى اليسار ولكل مقطع في السطر النصي، وهكذا يتم استخلاص سمات الحرف.

5-2-6 بناء متجه السمات

وعند تطبيق خوارزميات استخلاص السمات حصلنا على سمات كل حرف في السطر النصي المُدخل، حيث يتم خزن هذه السمات في متجه السمات الخاص بكل حرف.

وكانت قيم متجه السمات للسطر النصي المُدخل موضحة في الجدول–4.

(الجدول–4) متجه السمات لجزء من السطر النصي المُدخل

الحرف	متجه السمات							
	موقع الحرف	وجود الحلقة	اتجاه المنحني	وجود النقطة	عدد النقاط	عدد الانتقالات العمودية	عدد الانتقالات الأفقية	عدد النقاط الضوئية السوداء فوق خط الأساس
ـلـ	2	0	1	0	0	2	1	20
ـهـ	2	1	0	0	0	1	2	10
ـم	3	0	1	0	0	1	1	6
ا	4	4	0	0	0	1	1	16
ف	1	2	0	1	1	3	1	17
ـتـ	2	0	1	1	2	2	1	20
ـح	3	0	3	0	0	2	1	19
ـل	1	0	1	0	0	2	1	20
ي	3	0	2	2	2	2	2	2

6-2-6 توليد متسلسلة المشاهدات

يتم في هذه الخطوة توليد متسلسلة المشاهدات لكل حرف. وذلك بتحويل السمات المستخلصة إلى متسلسلة من الرموز، حيث يتم استدعاء الدالة الخاصة بتوليد متسلسلة المشاهدات لكل حرف في السطر النصي المُدخل. ويوضح الجدول–5 نموذج من متسلسلة المشاهدات لحروف السطر النصي المُدخل

(الجدول–5) نموذج من متسلسلة المشاهدات لحروف السطر النصي المُدخل

الحرف	متسلسلة المشاهدات				
لـ	1	3	9	13	-
هـ	1	2	9	14	15
م	1	3	9	14	-
ا	1	5	9	-	-
ف	1	3	7	10	-
ت	1	3	7	11	-
ج	1	5	9	14	-
لـ	1	3	9	15	-

7-2-6 التمييز باستخدام نموذج ماركوف الخفي

بعد الحصول على متسلسلة المشاهدات للحرف يتم إدخالها إلى نموذج ماركوف الخفي المُقابل لموقع الحرف المُدخل، وتُستدعى الدالة الخاصة بحساب الاحتمالية بتطبيق الخوارزمية الأمامية (forward algorithm), وسيتم فيما يلي توضيح نتائج التمييز لصورة السطر النصي المُدخل باستخدام الخوارزمية الأمامية.

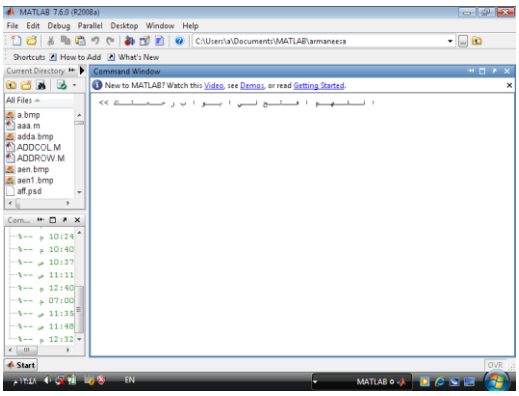
التمييز باستخدام الخوارزمية الأمامية

يتم استدعاء الدالة الخاصة بالخوارزمية الأمامية لتحسب احتمالية متسلسلة مشاهدات الحرف المُدخل وتُقارن مع الاحتمالية المحسوبة مُسبقاً في مرحلة التدريب فإذا حصل تطابق نطبع الحرف المُقابل له أما إذا لم يحصل التطابق فيطبع عبارة حرف غير معروف. نُطبق هذه الخطوة على جميع الحروف المكونة للنص المُدخل. بينما نتائج تطبيق الخوارزمية الأمامية لبقية حروف السطر النصي المُدخل توضح في الجدول–6.

(الجدول–6) الاحتمالية عند تطبيق الخوارزمية الأمامية لحروف السطر المُدخل

الحرف	الاحتمالية المحسوبة
لـ	0.0638542175229297
هـ	0.005050420761108
م	0.037048339843750
ا	0.015872955322266
ف	0.156768798828125
ت	0.054752349853516
ج	0.031698226928711

وناتج تمييز صورة النص المُدخل بتطبيق الخوارزمية الأمامية موضحاً في الشكل–10.



(الشكل–10) ناتج تمييز صورة النص المدخل بتطبيق الخوارزمية الأمامية

ومن خلال متابعة نتائج النظام في تمييز حروف 20 سطرا نصيا تضمنت اغلب حروف اللغة العربية وقد بلغت دقة التمييز (94.9%) للحروف ذات الحجم والنمط الخطي الواحد، مما يجعل منه أساس عمل أو لبنة أولى لبناء نظام تمييز حروف مختلفة الأنماط والأحجام.

10- الاستنتاجات

يُمكننا من خلال العمل الحالي استنتاج ما يلي:

1- قدرة نموذج ماركوف الخفي المصمم على التعرّف بسرعة محسوسة وأداء عاليين كما مُبين من خلال متابعة أداء النظام في التعرّف على السطر النصي المُدخل الذي تم توضيحه في المقطع السابق.

2- قدرة الخوارزمية الأمامية في التعرّف على صورة الحرف المُدخل بعد تحويله إلى متسلسلة مشاهدات من خلال حساب احتماليته ومقارنتها مع احتمالات محسوبة مسبقاً (من بيانات التدريب) وإخراج الحرف المقابلة احتماليته لاحتمالية الحرف المدخل.

11- الأعمال المستقبلية

هنالك عدة اقتراحات لتحسين أداء النظام وهي:

1- تطوير النظام ليقوم بالتعرّف على النص العربي المكتوب بخط اليد.

2- توسيع النظام ليشمل التعرّف على علامات التشكيل بالإضافة إلى الحروف المتداخلة مثل(آ، أ، لا، لآ، لاٌ، لاٍ).

3- تطوير النظام ليُطبق على أنماط مختلفة من الخطوط والأحجام.

4- تطوير النظام ليتعامل مع صفحات تحتوي على أنواع الرسوم أو الأشكال أو الصور ومن ثم فصل الصور والأشكال عن النص والتعامل معه بصورة مستقلة.

5- التعرّف بدون التقطيع إلى حروف وذلك لتجاوز الأخطاء التي تُسببها مرحلة التقطيع.

المصادر

[1] عجرش، آمال سفيح 2004، "**استخدام المنطق المضبب آلية لتمييز الحروف العربية**"، رسالة ماجستير غير منشورة، قسم علوم الحاسبات، كلية العلوم، جامعة البصرة، العراق.

[2] Sharma Amit Kumar and kishor Mr.R Rama, 2007, "**pattern recognition: Different available approaches**", proceeding of National conference on challenges & opportunities in information technology (COIT-2007) RIMT-IET, Mandi Gobindrh. www.rimtengg.com/coit 2007/.../coitindex.html

[3] Jain Anil K., Duin Robert P.W. and Mao Jain chang, 2000, "**statistical pattern recognition: A review**", IEEE Transaction pattern analysis and Machine intelligence, vol.22, No.1.

[4] Jannoud, Ismael Ahmed, 2007, "**Automatic Arabic Handwritten Text Recognition System**", American Journal of Applied sciences 4(11): 857-864, ISSN 1546-9239.

[5] Jurafsky Daniel and Martin James H., 2006, "**speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition**", 2nd Ed., prentice-Hall 2000, ISBN: 0-13-095069-6.

[6] Rabiner Lawrence R., 1989, "**A Tutorial on Hidden Markov Models and selected Applications in speech recognition**", proceedings of the IEEE, vol.77, NO.2.

[7] الكيم، سلوان تحسين فالح 2005، "**تصميم نظام لتمييز الحروف العربية باستخدام الخوارزميات الجينية**"، رسالة ماجستير غير منشورة، قسم علوم الحاسبات، كلية العلوم، جامعة البصرة، العراق.

[8] الكسو، ابتهاج عبد الحميد محمد 2005، "**استخدام الشبكات العصبية في تقدير رتب سلاسل ماركوف مع التطبيق على سلسلة جبل بطمة في محافظة نينوى**"، أطروحة دكتوراه غير منشورة، قسم الإحصاء، كلية علوم الحاسبات والرياضيات، جامعة الموصل، العراق.

[9] Sofia, Fatin Basher Abdul Ahad 2003, "**An Implementation of Arabic speech recognition**", Unpublished Ph.d. Thesis, Department of mathematical science, college of computer and mathematical science, university of Mosul, IRAQ.

[10] Aazami, Farshideh Einsele 2008, "**Recognition of ultra low resolution, Anti-aliased text with small font sizes**", Unpublished Ph.d. thesis, Scientarium informaticarum, Faculty of science, University of Fribourg, Switzerland.

[11] Dunham, Margaret H., 2002, "**Data Mining introductory and advanced Topics**", prentice Hall.

[12] Li xiaolin, Parizeau Marc and plamondon Rejean, 2000, "**Training Hidden Markov Models with multiple observations-A combinational Method**", IEEE Transactions on PAMI, vol.PAMI-22, NO.4, pp.371-377.

[13] Attaluri, srilatha, 2007, "**Detecting Meta Morphic Viruses using profile Hidden Markov Models**", Unpublished M.Sc. thesis, computers science, the faculty of the department of computer science, university of San Jose State.

[14] Jecheva Vaselina, 2006, "**A bout some Application of Hidden markov Model in intrusion detection system**", International conference and computer systems and Technologies-compsys tech'06.

[15] Khorsheed M.S., 2003, "**Recognizing handwritten Arabic Manuscripts using a single Hidden markov Model**", Pattern Recognition letters 24.