

[Egyptian Informatics Journal (2011) 12, 37](http://dx.doi.org/10.1016/j.eij.2011.02.001)–[43](http://dx.doi.org/10.1016/j.eij.2011.02.001)

Cairo University

Egyptian Informatics Journal

[www.elsevier.com/locate/eij](http://www.elsevier.com/locate/eij) [www.sciencedirect.com](http://www.sciencedirect.com/science/journal/11108665)

ORIGINAL ARTICLE

Recognition for old Arabic manuscripts using spatial gray level dependence (SGLD)

Ahmad M. Abd Al-Aziz [a](#_bookmark0),[\*](#_bookmark3), Mervat Gheith [b](#_bookmark1), Ayman F. Sayed [c](#_bookmark2)

a *Business Information System Department, Canadian International College, Egypt*

b *Computer Science and Information Department, Institute of Studies and Statistical Researches, Cairo University, Egypt*

c *Oriental Archaeology Department, French Institute of Oriental Archaeology, Egypt*

Received 14 July 2010; accepted 11 November 2010

Available online 22 March 2011

Abstract Texture analysis forms the basis of object recognition and classification in several domains, one of these domains is historical document manuscripts because the manuscripts hold our culture heri- tage and also large numbers of undated manuscripts exist. This paper presents results for historical doc- ument classification of old Arabic manuscripts using texture analysis and a segmentation free approach. The main objective is to discriminate between historical documents of different writing styles to three dif- ferent ages: *Contemporary* (Modern) Age, *Ottoman* Age and *Mamluk* Age. This classification depends on a Spatial Gray-level Dependence(SGLD) techniquewhichprovideseightdistincttexturefeaturesforeach sample document. We applied Stepwise Discriminant Analysis and Multiple discriminant analysis meth- ods to decrease the dimensionality of features and extract training vector features from samples. To clas- sify historical documents into three main historical age classes the decision tree classification is applied. The system has been tested on 48 Arabic historical manuscripts documents from the *Dar Al-Kotob Al- Masri*a Library. Ourresults sofar yield 95.83% correctclassification for the historical Arabic documents.

KEYWORDS

Spatial gray-level depen- dence (SGLD);

Old Arabic manuscripts recognition;

Document analysis; Segmentation free approach

© 2011 Faculty of Computers and Information, Cairo University. Production and hosting by Elsevier B.V. All rights reserved.

\* Corresponding author.

E-mail addresses: [ahmed\_abdelaziz@cic-cairo.com](mailto:ahmed_abdelaziz@cic-cairo.com) (A.M. Abd Al-Aziz), [mervat\_gheith@yahoo.com](mailto:mervat_gheith@yahoo.com) (M. Gheith), aymansayyid@ hotmail.com (A.F. Sayed).

1110-8665 © 2011 Faculty of Computers and Information, Cairo University. Production and hosting by Elsevier B.V. All rights reserved.

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

doi:[10.1016/j.eij.2011.02.001](http://dx.doi.org/10.1016/j.eij.2011.02.001)

Production and hosting by Elsevier

1. Introduction

In the last decade scientists are interested in studying and ana- lyzing the historical manuscripts from various points of views. They also study the manuscripts written in various languages (Latin, Arabic, Chinese, Hebrew, etc.), in various geographical areas and in different ages (early, medieval and modern age), to help researchers, historians and librarians to extract and re- trieve information which help them in editing, indexing and archiving processes [[1,2]](#_bookmark22). This work is important for paleogra- phers because it may confirm or invalidate their work and bring more objective conclusions when they face an undated historical document manuscript recognition problem.

The recognition of the handwriting style which represents the historical period of manuscripts raises different questions:

¯How to define a reliable ‘‘style similarity’’ between complex writings?

¯Which features to use in order to characterize a writing style and only the writing style independently from the writer, the text content and the image quality? [[3]](#_bookmark16)

We are interested more in ancient Arabic manuscripts of three different ages (Contemporary, Ottoman and Mamluk age). We agree with Eglin in [[3]](#_bookmark16), that for paleographers, the change from writing to another was not made in a radical way but by a slow and progressive evolution, which explains that it is difficult to identify categorically a given writing as shown in [Fig. 1](#_bookmark4).

* 1. *Ancient Arabic manuscripts*

The ancient Arabic manuscripts are considered to be more complex than the other manuscripts written in other languages [[4]](#_bookmark16). This complexity comes from two things; first complexity applies to all manuscripts which include (character degrada- tion, stains, low quality images, etc.) and the second complex- ity applies only to the Arabic language which includes (cursiveness of the text, character overlapping, diacritics and variety of calligraphic Arabic handwriting, ornaments, frames, decorations, two columns layout, etc.)

* 1. *Segmentation-free approach*

Due to the complexities in ancient Arabic manuscripts the seg- mentation approach which is the traditional and regular ap- proach will be inefficient to characterize the historical Arabic documents. In addition the accuracy of segmentation results are linked to quantity and the quality of the threshold manu- ally set [[5]](#_bookmark16). To avoid these difficulties; we have applied a

segmentation-free approach which considers the texture fea- tures analysis of document images independent from the text content by analyzing statistically the whole image. During the last years the segmentation free approach based on statis- tical feature analysis has been used in different domains and achieved promising results.

Said et al. [[6]](#_bookmark16) propose a text-independent approach and de- rive writer-specific texture features using multichannel Gabor filtering and (SGLD). The method requires uniform blocks of text that are generated by word de-skewing, setting a prede- fined distance between text lines/words and text padding. Two sets of 20 writers, 25 samples per writer are used in the evalu- ation. Nearest-centroid classification using weighted Euclidean distance and Gabor features achieved 96% writer identifica- tion accuracy.

Eglin et al. [[3]](#_bookmark16) applied the SGLD to extract several features to characterize the writing style of ancient Latin and Arabic manuscripts of the middle-ages. They propose that The SGLD is identical on different text areas of the same document and is robust to noise and does not require any image segmentation or layout analysis. They reduced the features size by using Haralick descriptors. From these combinations of features, they defined a ‘‘style similarity’’ measure and formed a large database of images samples of writings with a paleographic description to develop a reliable image retrieval system for medieval writings styles.

Srihari et al. [[7]](#_bookmark16) propose a large number of features divided into two categories. Macro-features operating at document / paragraph / word level and Micro-features operating at word

/ character level. Text dependent statistical evaluations are per- formed on a dataset containing 1000 writers who copied 3 times a fixed text of 156 words. The results have shown that micro-features are better than macro-features in identification tests with a performance exceeding 80%.

Marius Lucian Bulacu [[8]](#_bookmark16) proposed an automation process of writer identification using scanned images of handwriting and thereby to provide a computer analysis of handwriting

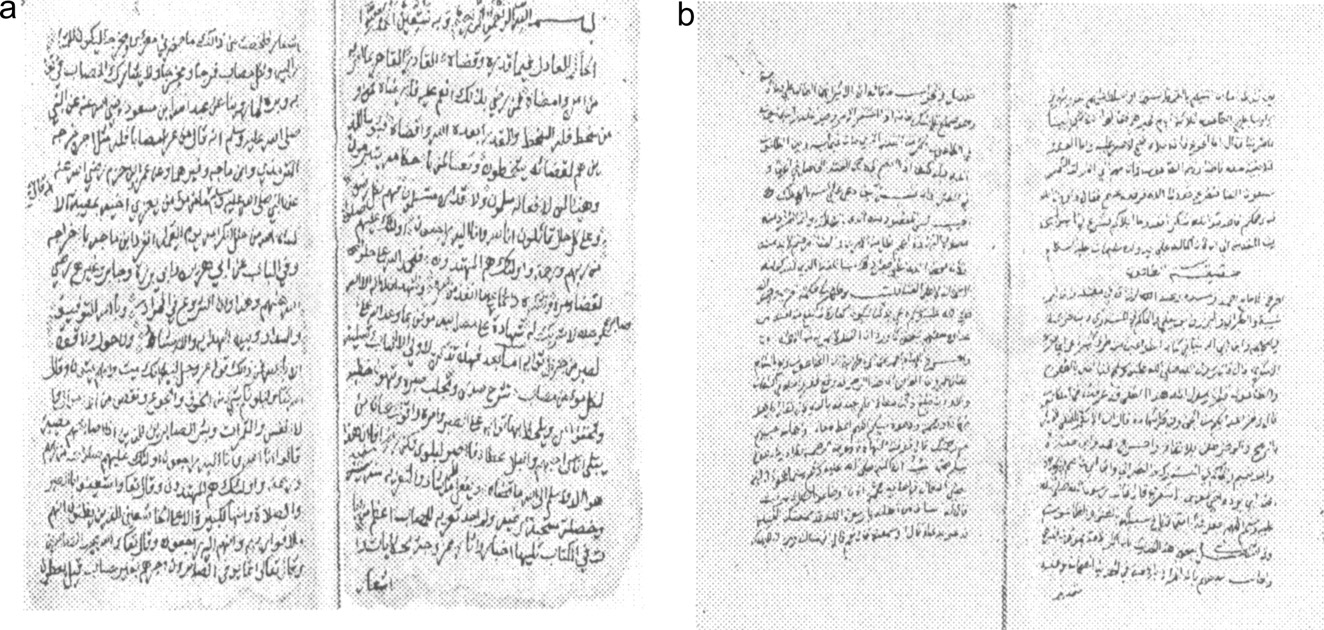


Figure 1 Different Arabic writing styles (a) 1339 hijri, (b) 1251 hijri.

individuality. The similarity in handwriting style between any two samples is computed by using appropriate distance mea- sures between their corresponding feature vectors. The features and writer classification operate in the general framework of statistical pattern recognition. Combining texture-level and allograph-level features yields very high writer identification and verification performance, with usable rates for datasets containing 103 writers.

Another system for classifying ancient manuscripts without using segmentation approach is proposed in [[9]](#_bookmark17). The aim of this scientific work is to propose a suitable assistance tool for paleographers and historians to help them in their intuitive and empirical work of identification of writing styles (for medi- eval handwritings) and authentication of writers (for humanis- tic manuscripts). In that context, they have chosen to propose a dual methodology based on two complementary approaches: a *texture based approach* that considers the handwriting in its global and homogeneous environment and a *local shape based approach* that considers the handwriting as a series of loops and right segments. They propose a global approach of writer’s classification based on Curvelets based features in rela- tion with two discriminative shapes properties, the *curvature* and the *orientation*. The main experimental results are very promising and show 78% of good retrieval (as precision) on the Middle-Ages database and 89% on the humanistic database. Another system proposed to characterize handwritten his- torical document and indexation without any a priori graph- emes segmentation is the system proposed by Bres et al. [[10]](#_bookmark18).This work is a response to scientific problems of historical handwritten corpus digitalization. It deals with the handwrit- ing denoising and indexation (finding documents coming from the same author) and is applied to a multi-language and multi- alphabet corpus. The developed perception based model lies

on the Hermite frequencial decomposition for image denoising and indexing. In that way, they have chosen a segmentation free approach that is global and generic. The current results of handwriting denoising and classification with orientation Hermite based features are very promising (more than 83%). The reviewing of previous different researches shows that the segmentation-free approach and statistical texture analysis method is being efficient in dealing with the historical docu- ment analysis problems. This approach is used in different problems related to the ancient document such as writer iden- tification of ancient manuscript documents [[6–8]](#_bookmark16). The research- ers also used the approach for indexation [[10]](#_bookmark18) and some papers focus on historical documents classification and recognition

[[6,9,5]](#_bookmark16).

1. Proposed system

The literature review shows the importance and advantages of the segmentation-free and statistical feature analysis ap- proaches in developing historical document analysis systems. Spatial Gray level Dependence (SGLD) is one method which considers the features analysis of a historical document image independent from the text content and letters frequencies. So we propose an automatic system for recognizing the old Arabic document manuscripts in three different ages: Contemporary (Modern), Ottoman and Mamluk age using SGLD technique.

The system accepts the undated ancient Arabic document image, after the preprocessing step the system applies SGLD to this image, then extracts a set of features from it, and then classifies it into three main different classes: Contemporary (Modern) age, Ottoman Age or Mamluk Age. A schematic of this system is shown in [Fig. 2](#_bookmark5).

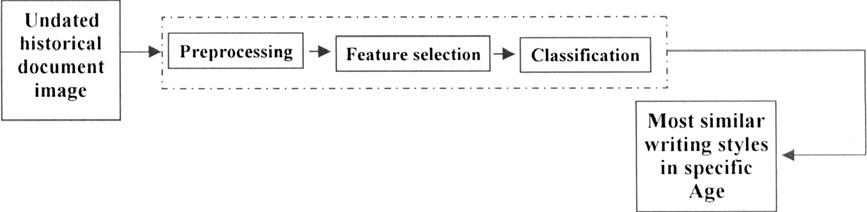


Figure 2 Block diagram of ancient document classification system.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 1 The sample of historical books and different writing styles in contemporary, Ottoman and Mamluk ages. | | | | |
| Age | Book’s title | Year | No. of pages | No. of books |
| Contemporary | *Thab Al-Amyr* | 1379 h | 20 | 2 |
|  | *Brd Al-Akabr* | 1339 h | 10 |  |
| Ottoman | *Sharh Kafyt ze Al-lob wa Al-fahm fy Asol ‘lm Al-tb* | 1250 h | 10 | 4 |
|  | *Ma Rawah Al-wa‘wn fy Akhbar Al-Ta‘wn* | 1251 h | 15 |  |
|  | *Mqn‘Al-mhtaj fy Adab Al-zwaj* | 995 h | 10 |  |
|  | *Al-drr fy ‘irab Aw,l Al-swr* | 1196 h | 5 |  |
| Mamluk | *Moftah Al-‘lwm* | 734 h | 5 | 4 |
|  | *Fwa*‘*id Al-‘lwm* | 894 h | 5 |  |
|  | *Moftah Al-‘lwm* | 877 h | 5 |  |
|  | *Al-mfsl fy Al-nahw* | 649 h | 5 |  |
| Total |  |  | 90 | 10 |
|  |  |  |  |  |

* 1. *Preprocessing*

To prepare the dataset of the experiment, we set all document images as gray level, and in terms of scale all document images are of the same size (400 · 320), and all the document images are of the same orientation because the SGLD technique is sensitive to the orientation of image. Finally all the pages are of the same ratio in order to guarantee that we compare the same information. All samples are collected from *Dar Al-Kotob Al-Masria*. We select from the page the text area only without the margins.

The system consists of 90 document images collected from 10 books or manuscripts related to three main different Ages: *Contemporary* (Modern) from 1220 Hijri till present, *Ottoman* from 923 Hijri to 1220 Hijri, and finally *Mamluk* Age from 648

3 SGLD is sensitive to rotation due to (Dx, Dy) parameters, so for the same document the SGLD matrix differs when direction is changed.

In our experiment we choose one offset vector [0 1], which means 1 neighbor pixel with 0 in all document images.

* 1. *Features selection*

Haralick [[12]](#_bookmark20) describes 14 statistics that can be calculated from SGLD matrix, we used the following eight Haralick’s texture features to characterize the document, and the other six statis- tics were not calculated due to computational instability for each image [[13]](#_bookmark21):

Hijri to 923 Hijri. The previous classification is compared by a paleographer expert on ancient Arabic manuscripts as shown in [Table 1](#_bookmark6).

*2.2. Applying spatial gray level dependence (SGLD)*

Correlation : (*i* — l*i*)(*j* — l*j*)*p*(*i*; *j*)

*i*;*j* r*i*r*j*

X

Inverse difference moment(homogeneity) : X *p*(*i*; *j*)

X

*i*;*j*

1 + |*i* — *j*|

(2)

(3)

The co-occurrence can be evaluated from the SGLD which is a joint probability to observe the same intensity value between two different pixels according to their spatial relation [[11]](#_bookmark19). By default, the spatial relationship is defined as the pixel of interest and the pixel to its immediate right (horizontally adja- cent) [[6]](#_bookmark16), SGLD is nothing but a matrix or distribution that is defined over an image to be the distribution of co-occurring

values at a given offset (distance between pixels). Mathemati-

Contrast : |*i* — *j*|2*p*(*i*; *j*) (4)

*i*;*j*

Angular second moment(energy) : *p*(*i*; *j*)2 (5)

X

*i*.*j*

Entropy : — *p*(*i*; *j*) log(*p*(*i*; *j*)) (6)

X

*i*.*j*

Mean : X X *i*.*p*(*i*; *j*) (7)

X

*i*

*j*

cally, a co-occurrence matrix C is defined over an N · M image

I, parameterized by an offset (Dx, Dy) as

Sum of squares(variance) : (*i* — l)2 (8)

*i*;*j*

X

*n m*

X

X

*C*D*x*;D*y*(*i*; *j*)=

1; if *I*(*p*; *q*)= *i* and *I*(*p* + D*x*; *q* + D*y*)= *j*

0; Otherwise

Covariance : (*i* — l*i*)(*j* — l*j* )*p*(*i*; *j*) (9)

*i*;*j*

*p*=1 *q*=1

(1)

In order to choose a subset of features to decrease the

dimensionality from Haralick’ texture features, two feature

From the output we can recognize a set of characteristics of SGLD method:

1. The SGLD is identical on different text areas of the same document and is robust to noise and does not require any image segmentation or layout analysis [[3]](#_bookmark16).
2. The SGLD is not only identical on different text areas of the same document but also similar to all documents in the entire manuscript of the same writer. Consequently, we do not have to study all document images from the entire manuscript.

selection methods were applied to the sample training data.

The first method is a modified version of multiple discriminant analysis and the second method is a stepwise discriminant analysis.

* + 1. *A modified version of multiple discriminant analysis*

The aim of this method in our research is to select those fea- tures that have the largest ratio of the variance of that feature calculated using all samples in the training set to the sum of the variances of that feature calculated for each class in the train- ing set, according to the following formula.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 2 Result of multiple discriminant analysis of features. | | | |
| Feature | Variance of feature in all classes | Sum of variances of feature in three classes | Ratio |
| Correlation | 0.0092 | 0.01416314 | 0.649573[\*](#_bookmark9) |
| Homogeneity | 0.0044 | 0.009510077 | 0.462667[\*](#_bookmark9) |
| Entropy | 0.0013 | 0.004703039 | 0.276417 |
| Contrast | 1113500.00 | 1352804.494 | 0.823105[\*](#_bookmark9) |
| Energy | 1.80E—06 | 4.26678E—06 | 0.421372[\*](#_bookmark9) |
| Mean | 0.899889365 | 4.95965163 | 0.181442 |
| Variance | 80328.15858 | 330560.1531 | 0.243006 |
| Covariance | 88328.1606 | 300560.1431 | 0.293878 |
| Mean value of all ratios: 0.4242.  \* Selected feature. | | | |

K = det(*W*(*x*)) (11)

|  |  |  |  |
| --- | --- | --- | --- |
| Table 3 The best significant features combination. | | | |
| Step | Feature (s) combination | Sig. of *F*-to-enter | Wilks’s Lambda |
| 1 | Contrast | 0.000 |  |
| 2 | Contrast | 0.000 | 0.635 |
|  | Correlation | 0.000 | 0.519 |
| 3 | Contrast | 0.000 | 0.512 |
|  | Correlation | 0.000 | 0.363 |
|  | Homogeneity | 0.000 | 0.285 |
| 4 | Contrast | 0.000 | 0.373 |
|  | Correlation | 0.000 | 0.245 |
|  | Homogeneity | 0.000 | 0.226 |
|  | Energy | 0.000 | 0.149 |
| Maximum significance of *F* to enter is 0.05. | | | |

|  |  |  |
| --- | --- | --- |
| Table 4 The total number and percentage of tested documents. | | |
| Age | Number of documents | Percentage (%) |
| Contemporary | 7 | 14.58 |
| Ottoman | 22 | 45.83 |
| Mamluk | 15 | 31.25 |
| Unknown age | 4 | 8.33 |
| Total | 48 |  |
|  |  |  |

det(*T*(*x*))

where *x* = [*x*1; *x*2; .. . ; *xp*] is a vector of the features that are currently included in the system.

*q ng*

var(*f*)

*W*(*i*; *j*)= X X(*xigt* — *xig*)(*xjgt* — *xjg*) (12)

P var(*fc*)

(10)

*g*=1 *t*=1

∀*c*

Where *fc* contains only feature values from class *c* and *f* con-

tains features from all document image classes. The goal of this criterion is to identify features that widely separate the classes

from one another (total variance) while keeping the classes

is within-groups sum of squares (variance) and cross products

for the features under consideration, and

*q ng*

*T*(*i*; *j*)= (*xigt* — *x*¯*ig*)(*xjgt* — *x*¯*j*) (13)

*g*=1

*t*=1

X X

themselves as tightly clustered as possible (sum of within class

variances) [[12]](#_bookmark20), the results of multiple discriminant analysis of previous 8 features as calculated from SGLD matrix are shown in [Table 2](#_bookmark7).

From the previous results we select those features that have the largest ratio (based on the mean value of all ratios), these features are in order: Contrast, Correlation, Homogeneity and Energy.

* + 1. *Stepwise discriminant analysis*

In order to identify combined features that maximize a crite- rion which describe their ability to separate classes from one another while at the same time keeping the individual classes as tightly clustered as possible, we applied the stepwise dis-

is a total sums of squares (total variance) and cross products. Where *q* is the number of classes, *ng* is the number of sam- ples in class *g*, *xigt* is the value of feature *i* for sample *t* of class *g*, *xig* is the mean of feature *i* over class *g*, and *xi* is the mean of feature *i* over all classes. Low values of K indicates features

that better discriminant the classes.

We notice that criterion K is nothing but the value calcu- lated from a modified version of multiple discriminant analysis (see Eq. [(10)](#_bookmark8)), but here we search for a set of features that min- imize the criterion value of K. So to accommodate the stepwise nature of the process, the partial K statistic is used. The statis- tic describes the increase in the discrimination ability of a sys- tem after adding a new feature, *xp+1*

K(*x*; *x* )

criminant analysis, the criterion used is Wilks’s K [[12]](#_bookmark20) which is defined as

K(*x*

*p*+1

· *x*)=

*p*+1

K(*x*)

(14)

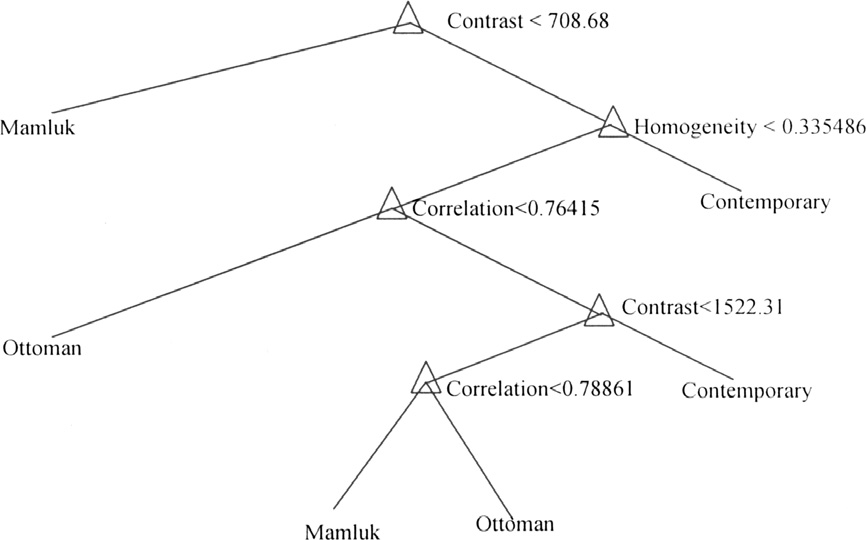


Figure 3 Decision tree classification.

To facilitate the ability to decide whether adding a new feature to the system will increase the discrimination significantly, Wilk’s partial K is converted to an *F*-statistic for which it is possible to assign a level of statistical significance. We applied sequential forward selection algorithm [[12]](#_bookmark20) to calculate *F*-to- enter statistic for each feature *j* not already included:

*F* (*j*)= *n* — *p* — *q* 1 — K(*xp*+1 · *x*) (15)

enter

*q* — 1

K(*xp*+1

· *x*)

where *n* is the number of data samples in all classes, *p* is the num- ber of features currently in the analysis, *q* is the number of clas- ses, [Table 3](#_bookmark10) shows the best significant features used in the system.

* 1. *Classification*

We apply a decision tree classification method in order to clas- sify the 90 training samples based on selected features previ- ously. Decision trees are multistage decision systems in which classes are sequentially rejected until we reach a finally accepted class [[14]](#_bookmark23).Before classification, the document image features data were separated into three main distinct training classes: Contemporary (Modern), Ottoman and Mamluk age as shown in [Fig. 3](#_bookmark12):

Table 5 The results of tested ancient Arabic documents.

Tested cases

System’s annotation

Total number of tested documents: 48.

Percentage of correct classification: 95.83%. Percentage of wrong classification: 4.17%.

\* Misclassified documents.

\*\* Consistent result between expert and system.

1. Results

To test our system, we applied the testing phase on 48 ancient Arabic documents in different three main ages as shown in [Table 4](#_bookmark11):

To guarantee the reliability of our system we took into our consideration two main parameters: the first one is the level of

recognition and the second one is the type of manuscript.

* 1. *Levels of recognition*

We applied 3 levels of testing. Level 1: in this level we had been testing dated document images not included in the sample training data but their dated manuscripts included in the sys- tem. In this level the system is responsible for classifying this document image. Level 2: in this level we tested dated docu- ment images their manuscripts not included in the system. In this level the system is also responsible for classifying this doc- ument images. Finally level 3: in this level we tested undated document images their undated manuscripts not included in the system. This level is the most difficult level of recognition; we relied on both system’s result and the expert’s knowledge.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Manuscript name | Year (Hijri) | Age | Page number | Total number of page(s) | result |
| Resala Fe Al- Hanadsa | 748 | Mamluk | 4, 57, 58 | 3 | Mamluk |
| Bord Al Akabr | 1339 | Contemporary | 6,7 | 2 | Contemporary |
| Moftah Al-Olom | 877 | Mamluk | 308 | 1 | Mamluk |
| Al-Masadr | 1202 | Ottoman | 1, 2 | 2 | Ottoman |
| Sharh Lob Al-Albab | 855 | Mamluk | Part from final page | 1 | Mamluk |
| Al-Montakhab | 1228 | Contemporary | 10, 11 | 2 | Contemporary |
| Al-Sayae Al-mahmoud | 1330 | Contemporary | 1, 2 | 2 | Ottoman[\*](#_bookmark14) |
| Al-Moknea | – | – | 1, 2 | 2 | Ottoman[\*\*](#_bookmark15) |
| Untitled | 1379 | Contemporary | Part from page1 | 1 | Contemporary |
| Olom Al-Hadeth | 661 | Mamluk | First page and Final page | 2 | Mamluk |
| Al-Bayan Al-Mofed | 845 | Mamluk | 1 | 1 | Mamluk |
| Nathl Al-Hemian | 836 | Mamluk | Before last page and last page | 2 | Mamluk |
| Asaned Al-Kotob Al-Seta | 1190 | Ottoman | 1, 2 | 2 | Ottoman |
| Quran | 866 | Mamluk | Soret Al-Zaryat 1 | 1 | Mamluk |
| Quran | 866 | Mamluk | Soret Al-Zaryat 2 | 1 | Mamluk |
| Quran | 9th century | Mamluk | Soret Al-Moamenon-1 | 1 | Mamluk |
| Quran | 9th century | Mamluk | Soret Al-Moamenon-2 | 1 | Mamluk |
| Sharh Kafyt ze Al-lob | 1250 | Ottoman | 5, 7, 8, 9, 14, 15, 16, 17, 18, 24, 25 | 11 | Ottoman |
| wa Al-fahm fy Asol ‘lm Al-tb |  |  |  |  |  |
| Ma Rawah Al-wa‘wn | 1251 | Ottoman | 12, 17, 25 | 3 | Ottoman |
| fy Akhbar Al-Ta‘wn |  |  |  |  |  |
| Mqn‘Al-mhtaj fy | 995 | Ottoman | Page 29-30 and page 31-32 | 2 | Ottoman |
| Adab Al-zwaj |  |  |  |  |  |
| Al-mfsl fy Al-nahw | 649 | Mamluk | 188 | 1 | Mamluk |
| Nozat Al Albab | 843 | Ottoman | Before last page and last page | 2 | Ottoman |
| Fe Al Alkab  Al-Amsela Al-Mokhtalefa | – | – | 1, 2 | 2 | Ottoman[\*\*](#_bookmark15) |
| Fe Al-Sarf |  |  |  |  |  |
| Total |  |  |  | 48 |  |

* 1. *Types of ancient document images*

In tested ancient documents we selected a variety of complex types of documents to show the efficiency of the segmentation free approach such as: ancient document with decoration, small part from document image, document image with degra- dation, and text-image document.

The following table shows that 95.83% of correct recogni- tion from different 48 tested ancient documents ([Table 5](#_bookmark13)).

1. Conclusions

This work has been done with the objective to help experts in paleography to recognize the undated Arabic manuscripts by approximating them to the similar writing styles in a certain age. Through the present research, we proposed a recognition system for ancient Arabic documents in three main historical ages: Contemporary, Ottoman and Mamluk ages based on segmentation free approach which considers the texture fea- tures analysis of document image independent from the text

content by analyzing statistically the whole image.

The SGLD is used to calculate a set of features that helped us to classify and recognize the ancient Arabic documents. The introduced system shows a high ability to recognize the ancient Arabic documents in different complex conditions such as: deg- radation, decoration or text-figure document in different ages.

The limitation of SGLD is that it should be applied on the same scale, ratio, and offset direction to recognize the undated ancient document efficiently. As future work SGLD will be tested to recognize the undated ancient documents in more an- cient ages such as: Abbasid and Fatimid ages.

Finally, we can extend the idea of free segmentation ap- proach and use more methods such as curvelets with SGLD that we can join more than one type features into a single coherent feature set, which can be used for writer identification and document indexation.

References

1. Boussellaa W, El-Abed H, Zahour A. A concept for the separation of foreground/background in arabic historical manu-

scripts using hybrid methods. In: 7th International symposium on virtual reality, Archaeology and Cultural Heritage, 2006. p. 132.

1. Perantonis S, Gatos B, Ntzios K, Pratikakis I, Vrettaros I, Drigas A, Emmanoulidis C, Kesidis A, Kalimmirakis D. A system for processing and recognition of old greek manuscripts D-SCRIBE project. In: WSEAS international conference on applied infor- matics and communications, 2004.
2. Eglin V, Lebourgeois F, Bres S, Emptoz H, Leydier Y, Moalla I, Drira F. Computer assistance for digital libraries: contributions to middle-ages and authors’ manuscripts exploitation and enrich- ment. In: 2nd International conference on document image analysis for libraries, 2006, pp. 265–80.
3. Adnan A. Recognition of printed and handwritten Arabic charac- ters, Advances in Document Image Analysis, Springer, Berlin, 1997.

p. 40–59.

1. Journet N, Eglin V, Ramel JY, Mullot R. Dedicated texture based tools for characterization of old books. In: 2nd International conference on document image analysis for libraries, 2006.
2. Said H, Tan TN, Baker KD. Personal identification based on handwriting. J Pattern Recog Soc 2000;33:149–60.
3. Srihari S, Beal M, Bandi K, Shah V, Krishnamurthy P. A statistical model for writer verification. In: 8th International conference on document analysis and recognition, vol. II, 2005. p. 1105–9.
4. Bulacu M. Statistical pattern recognition for automatic writer identification and verification, <[http://dissertations.ub.rug.nl/](http://dissertations.ub.rug.nl/faculties/ppsw/2007/m.l.bulacu/) [faculties/ppsw/2007/m.l.bulacu/](http://dissertations.ub.rug.nl/faculties/ppsw/2007/m.l.bulacu/)>, last visited date 12/4/2010.
5. Joutel G, Eglin V, Bres S, Empotz H. Curvelets based feature extraction of handwritten shapes for ancient manuscripts classi- fication. Doc Recog Retr 2007;XIV:6500.
6. Bres S, Eglin V, Volpilhsv-Auger C. Evaluation of handwriting similarities using hermite transform. In: 10th International work- shop on frontiers in handwriting recognition, 2006.
7. Vasantha M, Subbiah Bharathi V, Dhamodharan R. Medical image feature, extraction, selection and classification. Int J Eng Sci Technol 2010;2(6):2071–6.
8. Boland M. Quantitative description and automated classification of cellular protein localization patterns in fluorescence microscope images of mammalian cells, PhD dissertation, Pittsburgh, USA, 1999.
9. Breuer P, Dislaire G, Filatoff J, Steemans P. A classification of spores by support vectors based on an analysis of their ornament spatial distribution. In: Recent advances in palynology carents de geologie / Notebooks on Geology, Brest, 2007.
10. Theodoridis S, Kourtroumbas k. Pattern Recog. 4th ed. 146: Academic Press; 2003.