Content-Based Retrieval of Arabic Historical Manuscripts Using Latent Semantic Indexing

Mohammad H. Yahia * and Wasfi G. Al-Khatib **

* King Faisal University, Mathematics and Statistics Department, Hofuf, Saudi Arabia. Email: mnajeeb@kfue.edu.sa
** King Fahd University of Petroleum and Minerals, Information and Computer Science Department. Dhahran, Saudi Arabia. Email: wasfi@kfupm.edu.sa

Abstract—Large archives of historical Arabic manuscripts cannot be manually searched because of the difficulty of manual indexing construction. Optical character recognition (OCR) techniques are available, but due to the characteristics of the historical Arabic manuscripts and some contents such as figures and drawings of manuscripts, OCR techniques may not yield satisfactory results. Due to these reasons, we suggest in this research work, employing content-based image retrieval (CBIR) techniques and latent semantic indexing (LSI) approaches to query these manuscripts and to make them better accessible to the public. The results showed that our system was able to retrieve relevant words with a 72.5% accuracy.

Keywords—Historical Manuscripts Indexing; CBIR; LSI; SVD

I. INTRODUCTION

Arabic historical manuscripts represent an important source of knowledge that needs to be preserved. Most of these manuscripts are archived and stored in scattered libraries and universities worldwide. Even the ones that were digitized and stored electronically cannot be automatically searched since they were stored in their original image forms. Due to the nature of these manuscripts, applying optical character recognition techniques to automatically recognize text is currently infeasible. With the availability of current technologies such as digitization and image processing techniques, it is possible to digitize them and provide better and easier ways of access.

A. Content-Based Image Retrieval (CBIR)

Generally, CBIR systems retrieve images from a database by comparing the features that were automatically extracted from the images themselves. Content-based image retrieval depends on the concepts of query by image example (QBE). It involves a matching process that matches the features of a query image with the features of images stored in the database. The system then determines the stored images whose feature values match those of the query most closely. Furthermore, the structure of the content-based image retrieval system as shown in [1] involves the following four parts in system realization: (i) Data collection, (ii) Buildup of the feature database, (iii) Search in the database and (iv) Order arrangement and deal with the results of the retrieval.

II. LITERATURE REVIEW

In the literature, there is a number of several proposed systems for indexing historical manuscripts. The following paragraphs concentrate on indexing Arabic historical manuscripts.

The work of [2] has shown the difficulty of historical handwritten manuscript recognition. Authors of [3] proposed a computer aided retrieval and indexing system for Arabic historical manuscripts. The proposed system used different combinations of five feature categories: projection profiles, concentric circle features, angular line features, Hu’s moment and geometric features. Several similarity measures such as Euclidean distance, Manhattan distance, and angular separation have been tested. The developed system has shown encouraging results with recognition rate equal to 76%.

In [4], a framework for content-based retrieval of historical documents in the Ottoman Empire archives was developed. The system used different combinations of five feature categories: projection profiles, concentric circle features, angular line features, Hu’s moment and geometric features. Several similarity measures such as Euclidean distance, Manhattan distance, and angular separation have been tested. The developed system has shown encouraging results with recognition rate equal to 76%.

III. LATENT SEMANTIC INDEXING AND IMAGE RETRIEVAL

The main reason for employing Latent Semantic Indexing (LSI) is its ability to carry out similarity matches using thresholds that are largely independent from the underlying features used.

A. Mathematical Background

Latent Semantic Indexing (LSI) model is used to exploit geometric relationships between document and term vectors in order to explain both similarities and differences. LSI attempts to project very high dimensional document and query vectors into a low dimensional space. This uses a Singular Value Decomposition (SVD) of the term-by-document matrix. Let us define SVD
Definition A.1: The singular value decomposition (SVD) of any matrix $A^{m \times n}$ of rank $\leq q = \min(m, n)$, denoted by $SVD(A)$, is defined as:

$$A = U \Sigma V^T$$ (1)

Where $U^{m \times q}$ and $V^{q \times n}$ are orthogonal matrices. The first $r$ columns of $U$ and $V$ are called the left and right singular vectors respectively. The elements $\sigma_i$ are the nonnegative square roots of the $n$ eigenvalues of $AA^T$ or $A^TA$.

B. Constructing of the LSI Index

In text retrieval (TR), a document collection comprised of $n$ documents which are indexed by $m$ terms, can be represented as an $m \times n$ term-by-document matrix $A$. Thus the matrix element could be the weighted frequency at which term $i$ occurs in document $j$ [6].

**LSI in Image Retrieval**

In image retrieval, we use all subword images and their feature values as a document collection. Thus, $n$ columns represent the images and $m$ rows represent the feature values. Figure 1 demonstrates how an $8 \times 3$ term-by-document matrix is constructed from a small collection of subwords of the image of the word prophet (رسول).

$$A = \begin{pmatrix} 0.22 & 0.34 & 0.08 \\ 0.17 & 0.05 & 0.34 \\ 0 & 0.14 & 0 \\ 0 & 0.10 & 0 \\ 0.12 & 0.11 & 0.23 \\ 0.27 & 0.20 & 0.12 \\ 0.07 & 0 & 0.12 \\ 0.12 & 0.02 & 0.09 \end{pmatrix}$$

**Figure 1. The construction of a term-by-document matrix $A$**

**Feature Normalization**

Feature normalization techniques aim at normalized extracted features in such a way that the normalized features are better suited for classification. We use a simple technique to carry out the features normalization. Firstly, we count the number of black pixels within a given area. Secondly, this counter is divided by the count of all black pixels within the word image.

**Feature Occurrences**

The major problem of applying LSI to image retrieval is that terms must be replaced by image features. In image feature values, there is no existence of occurrence, thus, we have to convert these normalized feature values into feature occurrences. In order to do that, we apply the algorithm of [7].

C. Query Matching

Query matching in the vector space model can be viewed as a search in the column space of the matrix $A$. The cosines used as a distance measure to retrieve relevant documents. The Cos $\theta$ between the query vector $q$ and the $n$ document vectors are defined by:

$$\cos \theta_j = \frac{a_j^T q}{\|a_j\|_2 \|q\|_2} = \frac{\sum_{i=1}^{m} a_{ij}q_i}{\sqrt{\sum_{i=1}^{m} a_{ij}^2} \sqrt{\sum_{i=1}^{m} q_i^2}}$$ (2)

for $i = 1, 2, ..., n$. In contrast, this query matching can be formulated using the component matrices of the SVD. Suppose we have a query vector $q$ and $n$ columns of the reduced-rank matrix $A_k$, $\cos \theta$ can be defined as follows:

$$\cos \theta_j = \frac{(U_k \Sigma_k V_k^T e_j)^T q}{\|U_k \Sigma_k V_k^T e_j\|_2 \|q\|_2}$$ (3)

where $e_j$ is the $j$th canonical vector of dimension $n$ and $(j = 1, 2, ..., n)$.

IV. SYSTEM FOR INDEXING HISTORICAL ARABIC MANUSCRIPTS

In this research work, we develop a system that provides retrieval and indexing facilities of Arabic historical manuscripts. In this system, a user can specify a query word and the system will search and retrieve occurrences of the query word found in the database. The prototype system consists of five main modules: i) image pre-processing module, ii) image segmentation module, iii) feature extraction module, iv) LSI index construction module, and v) the query matching module.

A. Preprocessing

In this stage, two necessary operations are carried out on manuscript images in order to represent them in a way which can be measured for enhancing the efficiency of the developed system. These operations include: i) Binarization and ii) noise removal and smoothing. Furthermore, binarization process binarization, is done by two steps: i) converting RGB color space to a gray-scale image, and ii) converting a gray-scale image to a binary image.

B. Segmentation

The objective of this process is to segment all of the manuscript pages, i.e., images, into connected components (i.e. words/subwords). In order to achieve this objective, two modules are implemented. The former one is to segment the whole text within a page into lines by detecting the base line using the concept of horizontal projection. The later one is to segment these resulted lines into words/subwords by applying the connected component algorithm [8]. In Figure 2, the Arabic word (الرسول), means prophet, has four connected parts c1, c2, c3, and c4.

**Figure 2. The four connected components in the word prophet**
C. Features Extraction

The objective of this process is to extract the essential attributes which describe each word/subword resulted from the previous process i.e. the segmentation process. Here, we use three sets of low-level features: i) Concentric circle features, ii) angular line features, and iii) rectangular region features.

Concentric Circle Features

These features, which are four, are computed by counting the number of black pixels contained between concentric circles centered at centroid with radius \( r, 2r, ..., (n-1)r, nr \), where \( n \) is the number of concentric circles. Then, the values are normalized by dividing the count of black pixels within each region by the count of black pixels within the subword image. Figure 3 depicts these features while Table I includes these normalized features

\[
\begin{array}{cccc}
\text{f1} & \text{f2} & \text{f3} & \text{f4} \\
0.0949 & 0.3609 & 0.3179 & 0.2234 \\
\end{array}
\]

Angular Line Features

We compute these features by counting the number of black pixels in 45° degree slices centered at the centroid of the subword image with respect to the horizontal axis. As a result of this slicing, there are eight features. Then, these values are normalized by the total number of black pixels of the image. Figure 4 depicts these angular line features of a given word while Table II includes these normalized features

\[
\begin{array}{cccccccc}
\text{f1} & \text{f2} & \text{f3} & \text{f4} & \text{f5} & \text{f6} & \text{f7} & \text{f8} \\
0.3668 & 0.0018 & 0.0055 & 0.1122 & 0.3445 & 0.0064 & 0.0044 & 0.1600 \\
\end{array}
\]

Rectangular Region Features

These are nine feature values resulted from dividing the image area into nine rectangular regions. Then, the count of black pixels is computed for each region. These values are normalized by the count of all black pixels of the image. Figure 5 depicts these features while Table III includes these normalized features

\[
\begin{array}{cccccccc}
\text{f1} & \text{f2} & \text{f3} & \text{f4} & \text{f5} & \text{f6} & \text{f7} & \text{f8} \\
0 & 0 & 0.20 & 0.05 & 0.26 & 0.21 & 0.23 & 0.02 \\
\end{array}
\]

V. EXPERIMENTAL EVALUATION

This chapter presents the testing results of the experimental work of this research work.

A. System Implementation

Our system is implemented in the MATLAB 2009b environment. It is a GUI based easy to use application. Users can perform several tasks such as preprocessing steps, segmentation and feature extraction steps, and searching steps.

B. Historical Arabic Manuscripts

Within the experimental work of this research work, a pre-scanned historical Arabic manuscript was used. The title of this historical manuscript is "Sahih Al-Bukhari" (صحيح البخاري). We used 34 pages from the section of "Mawaqeet Al-Haj wa Al-Umra" (مواقيت الحج والعمرة).

C. Query Words

Thoughtfully, we choose 20 words as query words. Some of these words such as (الحج), meaning pilgrimage, may be considered as keywords in order to search particular topics. Names of places and names of people are also selected to be query words. Moreover, phrases such as "peace be upon him" (صلى الله عليه) and "be pleased with him" (رضي الله عنه) are selected to issue queries in order to search prophetic traditions.

D. Performance Evaluation

The standard way to evaluate the performance of a system is to compute recall and precision. Recall is measured as the ratio of the number of relevant documents retrieved to the total number of relevant documents which exist in the collection. Computations of these measures are as follows:

\[
\text{Recall} = \frac{\text{Relevant Images Retrieved}}{\text{Total Relevant Images Present}} \quad (4)
\]
E. Feature Sets

As mentioned before, three different features are used in the experimental work within this research work. Each one of these features has two different calculated values. The first value is the normalized features, and the second one is feature occurrence values. Also, we group our query words depending on the number of connected components. Thus we have 6 feature sets that should be tested on each query word. The goal of the tests is to indicate which feature sets and feature value types will be most useful in indexing and retrieval of manuscripts. The main results from these tests are shown in Table IV, while Figure 6 shows one of these tests using rectangular region features. In this figure, the cells marked with true symbol correspond to relevant words, while the cells marked with false symbol correspond to non-relevant words. It is clearly that rectangular region feature sets are the best features over the other feature sets.

<table>
<thead>
<tr>
<th>Feature Sets</th>
<th>Normalized Features</th>
<th>Feature Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concentric Circles</td>
<td>47.5%</td>
<td>56.5%</td>
</tr>
<tr>
<td>Angular Lines</td>
<td>53.7%</td>
<td>60%</td>
</tr>
<tr>
<td>Rectangular Regions</td>
<td>53.7%</td>
<td>60%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Sets</th>
<th>Normalized Features</th>
<th>Feature Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concentric Circles</td>
<td>55%</td>
<td>57.5%</td>
</tr>
<tr>
<td>Angular Lines</td>
<td>58.7%</td>
<td>62.5%</td>
</tr>
<tr>
<td>Rectangular Regions</td>
<td>60%</td>
<td>62.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Sets</th>
<th>Normalized Features</th>
<th>Feature Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concentric Circles</td>
<td>61.2%</td>
<td>60%</td>
</tr>
<tr>
<td>Angular Lines</td>
<td>62.5%</td>
<td>70%</td>
</tr>
<tr>
<td>Rectangular Regions</td>
<td>65%</td>
<td>67.5%</td>
</tr>
</tbody>
</table>

The main evident observations from Table IV are:

- Feature occurrences of rectangular region feature set, with recall equal to 72.5%, performs better than other feature sets. While normalized concentric circle feature set is the worst with recall equal to 47.5%.
- There is a steadily increasing relationship between the number of connected components and the recall.

VI. CONCLUSIONS

An indexing and searching system for Arabic historical manuscript is developed and implemented. There are several modules such as preprocessing module, segmentation module, and features extraction module, applied before constructing the term-by-document matrix. Features such as concentric circular, angular lines, and rectangular regions are extracted from the word images. Latent Semantic Indexing (LSI) is employed to represent subword images and their corresponding features as a term-by-document matrix. Query matching mechanisms are applied in order to retrieve candidate words. Our system shows encouraging results and is able to retrieve relevant words.

VII. FUTURE WORK

For the future work, we are planning to:

- Use database management systems in order to increase the efficiency of the retrieval operations.
- Investigate other features that can improve the precision-recall of the system.
- Use the concepts of the relevance feedback in order to enhance the indexing process.

REFERENCES


