**Writer based Handwritten Document Image Retrieval**

**Vijayalaxmi.M.B and B.V.Dhandra**

Department of P.G.Studies and Research in Computer Science
Gulbarga University, Gulbarga
Gulbarga, India

**ABSTRACT**

In this paper a method is proposed for retrieval of handwritten document images based on the writer’s handwriting using texture features of input handwritten document image block. Typically it can be observed that the patterns of any handwritten text blocks encompass spatial texture primitives. The conventional two-dimensional (2-D) discrete wavelet transforms (DWTs) and Correlation of GLCM is used to extract spatial features. Handwritten documents are collected from 100 writers each in English, Kannada and Hindi scripts. These handwritten documents are segmented into image blocks and 2000 image blocks of each script writers are used separately for validation of the proposed method. The similarity measures viz., Euclidean and City block distances are used and achieved Top-1 retrieval rates as 100% for each of the Kannada, English and Hindi writers’ document image blocks.

**General Terms**

Pattern Recognition, Retrieval, Handwritten Documents.

**Keywords**

Texture, Discrete wavelet transform, Correlation of GLCM, document similarity measurement, handwritten documents, document retrieval, writer identification.

1. INTRODUCTION

With the aid of modern technology it is possible to efficiently produce, process, store, and transmit document images. As the world is moving towards the paperless office, large quantities of printed or handwritten documents are digitized and stored as images in databases. Many organizations currently use and depend on document image database and hence the document images as information source have retained their importance even today. Searching for relevant document from large complex document image repositories is a crucial problem in document image analysis and retrieval. Handwritten based document image retrieval is a process of retrieval of the handwritten documents that are similar to query image and has many applications in indexing and document image retrieval from archival of digital library of handwritten documents.

Document image retrieval is an important field of research with the continuous progress and increasing security requirements for the development of the modern society. In this paper, we have proposed a method for writer based automatic handwritten document image retrieval from the features extracted over an entire textual handwriting image. Hence, the proposed method is a global approach of writer based handwritten document image retrieval using the features based on correlation of GLCM of input image and unique directional multi-resolutionality property of DWT.

2. REVIEW OF LITERATURE

A number of approaches to retrieval and writer identification have been proposed in the literature.

Pirlo et al. [1] designed a system for layout based document image retrieval for retrieval of commercial forms as invoices, waybills and receipts, to optimize document management and sustainability. It used a morphologic filtering technique and Radon Transform for layout description and Dynamic Time Warping for document image matching. Shirdhonkar et al. [2] used Countertop Transform based features for Handwritten document image retrieval and achieved precision accuracy of 100% using Canberra distance. Daramola et al. [3] proposed an item content image retrieval system. Two image content attributes in the form of texture and shape were extracted. Extraction of low level features was achieved by decomposition of images using Haar wavelet transform. Images at detailed bands were partitioned into non-overlapping blocks before phase congruency feature was extracted. Retrieval of images based on query images was done using Support Vector Machine (SVM) and fused feature from image shape and texture. Atanasiu et al. [4] used a set of 10 features extracted from the probability density function of the orientations of the writing contours for retrieving the documents belonging to the same writer. The experiment was conducted on IAM database. Ralph Niels et al. [5] proposed a method for writer identification, by recasting the traditional information retrieval (IR) problem of finding documents based on the frequency of occurrence of particular character shapes: allographs. Top-1 performances of almost 60% were achieved for small query documents containing only 10 characters and achieved perfect top-1 writer identification rates for larger databases. Stefan Fiel et al. [6] calculated local features of the image and with the help of a predefined codebook an occurrence histogram was created. This histogram is compared to determine the identity of the writer or the similarity of other handwritten documents and the method has been evaluated on IAM and TrigraphSlant datasets. They also proposed a writer identification and writer retrieval method [7] by calculating a vocabulary by clustering features using a Gaussian Mixture Model and applying the Fisher kernel. For each document image the features were calculated and the Fisher Vector was generated using the vocabulary. Chawki et al. [8] used texture based edge-hinge and run-length features to characterize the writing style of an individual and shown that by reducing the search space using a writer retrieval mechanism prior to identification improves the identification rates. Ajinkya et al. [9] proposed image retrieval methods based on shape features extracted using gradient operators like Robert, Sobel, Prewitt and Canny. A database of 1000 variable sized images spread across different categories has been used. Robert gradient operator based retrieval method has shown higher rate of precision and recall. Rajiv Jain et al. [10] proposed a method for document image retrieval based on Local feature extraction using Speeded Up Robust Features (SURF), feature indexing and geometric verification of documents and used Complex Document Information Processing (CDIP) Tobacco dataset.
The properties of GLCM of input handwritten document images are used by Dhandra et al. [11,12] for writer identification. In [11] they extracted a set of texture features based on correlation-homogeneity properties of gray level co-occurrence matrices of the input handwritten document images of Roman, Kannada and Devanagari writers and obtained accuracies above 80% for writer identification in documents of single, two and three scripts written by the same 100 writers and also they have shown [12] that among four properties of GLCM namely Correlation, Contrast, Energy and Homogeneity, when single property is used Correlation has shown higher writer identification rate and hence may be used as a potential property in writer identification problems.

Through exhaustive experimentation the features based on discrete wavelet transform and correlation property of GLCM of input handwritten document image are found to give encouraging results in the document image retrieval process and hence these textural features are extracted in the proposed writer document image retrieval method.

The paper is organized as follows: In Section III data collection and Feature Extraction. The training and retrieval phases are explained in Section IV. In Section V the results of writer document retrieval are discussed. Finally, Section VI contains conclusion.

3. DATA COLLECTION AND FEATURE EXTRACTION

3.1 Data Collection

The standard database for writer's handwritten documents of Kannada, Hindi scripts are not available. This enforced us to create a database of the Kannada, Hindi writers. Hence handwritten documents are collected from 100 writers of Kannada, Hindi script belonging to different age groups. The collected documents are scanned through HP scanner to obtain digitized images. The scanning is performed at 300 dpi resolution. For English writers, IAM database [13] of handwritten document images of 100 writers are considered.

The 20 blocks of size 512X512 are segmented from the digitized document images of each writer amounting to 2000 blocks in each script.

We employed preprocessing steps such as removal of non-text regions and noise. In the proposed model, text portion of the document image was separated from the non-text region manually. Noise removal by morphological operations. A global thresholding approach is used to binarize the scanned gray scale images where black pixels having the value 0’s correspond to object and white pixels having value 1’s correspond to background.

Sample blocks of three different Kannada writers and architecture of proposed document retrieval system are shown in Figure 1 and Figure 2 respectively.

![Figure 1. Kannada Writer’s Sample Blocks of size 512X512 pixels of three different writers (a), (b) and (c)](image_url)

![Figure 2. System Architecture for Proposed Document Retrieval of Writers](image_url)

3.2 Feature Extraction

For feature extraction, we computed correlation of GLCM of input image along 4 directions and 5 distances, DWT with Wavelet family (Coiflet-5) basis function to get the four sub band images namely Approximation (A) and three detail coefficients - Horizontal (H), Vertical (V) and Diagonal (D). The details of feature extraction process are described below.

3.2.1 Correlation of Gray Level Co-occurrence Matrix

A statistical method that considers the spatial relationships of pixels is the gray-level co-occurrence matrices (GLCM) of the image, also known as the gray-level spatial dependence matrix. We use five distances d = 1, 2, 3, 4, 5 and four directions θ = 0°, 45°, 90°, 135° to construct twenty GLCMs. For each GLCM matrix common statistical property correlation can be extracted. where h(x,y) is the (i,j)th entry in the GLCM and is probability that a pixel with value x, will be found adjacent to a pixel with value y [11].

\[ Correlation = \sum \sum \frac{(x - \mu_x)(y - \mu_y)h(x,y)}{\sigma_x \sigma_y} \] (1.1)

where, \( \mu_x, \mu_y, \sigma_x \) and \( \sigma_y \) are the means and standard deviations of \( h_x \) and \( h_y \).

\[ h(x) = \sum h(x,y) \] \( h(y) = \sum h(x,y) \)

\[ \mu_x = \sum x \cdot h(x,y) \] \( \mu_y = \sum y \cdot h(x,y) \)

\[ \sigma_x = \sqrt{\sum (x - \mu_x)^2 \cdot h(x,y)} \] and

\[ \sigma_y = \sqrt{\sum (y - \mu_y)^2 \cdot h(x,y)} \]

It is a measure that a pixel is correlated to its neighbor over the whole image. The correlation feature is a measure of gray tone linear dependencies in the image.

3.2.2 Discrete Wavelet Transform

Discrete wavelet transform (DWT) performs sub-band coding on an image in terms of spatial and frequency components and analysis of image from coarse to finer level. The literature on wavelet-based methods continue to be powerful mathematical tools in texture classification problems.

The different wavelet transform functions filter out different range of frequencies (i.e., sub-bands). Thus, wavelet is a
powerful tool, which decomposes the image into low frequency and high frequency sub band images. We have considered only three sub band images namely Approximation (A) and Horizontal (H), Vertical (V) of DWT with Coiflet-5 family.

query image. The ratio of number of relevant documents retrieved to total number of documents retrieved gives the document retrieval rate based on writer.

End

5. EXPERIMENTAL RESULTS

An exhaustive experimentation is conducted by using Energy, Contrast, Correlation and Homogeneity properties of GLCMs of input handwritten document images and found that among those only the correlation property based feature which is giving higher identification accuracy and discrete wavelet transform of the input image are used as features in the proposed method.

As each input image is of 512X512 pixels size and consists of 2 to 5 lines of text, each line may contain one to three number of words depending on the length of the word. The feature vector of input query block is compared against feature vectors of all the train blocks stored in the feature library, the nearest neighbor is used to retrieve the document. The top-1 writer document retrieval rate is 100% using both Euclidean distance and City block distance based similarity measures because in top-1, query image is matched only with one nearest train block image. The blocks having lesser information content and writers having similar handwriting causes lesser writer document retrieval rates from same writer in case of Top 2, 3, etc.

Writer document image retrieval algorithm is applied to single script documents at a time. In each script, the Top-k documents belonging to particular writer are retrieved using Euclidean and City block distance measures and are tabulated in tables and illustrated in figures below. From Table 1, 2 and Figure 4, Figure 5 it is observed that City Block distance based writer retrieval rates are little high for English and Kannada. From Table 3 and Fig 6 it is observed that Euclidean distance based writer retrieval rates are little high for Hindi.  

<table>
<thead>
<tr>
<th>Top k</th>
<th>Writer Document Retrieval Rate % (Euclidean)</th>
<th>Writer Document Retrieval Rate % (City Block)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>82.625</td>
<td>84.075</td>
</tr>
<tr>
<td>3</td>
<td>74.8167</td>
<td>76.9667</td>
</tr>
<tr>
<td>4</td>
<td>69.8875</td>
<td>72.55</td>
</tr>
<tr>
<td>5</td>
<td>66.54</td>
<td>69.35</td>
</tr>
<tr>
<td>6</td>
<td>63.6083</td>
<td>66.95</td>
</tr>
<tr>
<td>7</td>
<td>61.2143</td>
<td>64.8071</td>
</tr>
<tr>
<td>8</td>
<td>58.9688</td>
<td>62.8563</td>
</tr>
<tr>
<td>9</td>
<td>57.0111</td>
<td>61.95</td>
</tr>
<tr>
<td>10</td>
<td>55.205</td>
<td>59.55</td>
</tr>
</tbody>
</table>
6. CONCLUSION
A new approach is proposed for texture based handwritten document image retrieval based on writer using directional multiresolution property of DWT and features of correlation of GLCM along four directions and five distances of image. The document matching using Euclidean and City block distance measures is performed and obtained higher retrieval results using City Block measure for English and Kannada documents, using Euclidean measure for Hindi documents. Although the text content in each image block is very less the retrieval results are high for writer document retrieval in each English, Kannada and Hindi script document writers.

7. REFERENCES


