Document Image Retrieval Based on Texture Features: A Recognition-Free Approach

Fahimeh Alaei  
School of ICT, Griffith University, Australia  
fahimeh.alaei@griffithuni.edu.au

Alireza Alaei  
School of ICT, Griffith University, Australia  
ali20alaei@yahoo.com

Umapada Pal  
CVPR Unit, Indian Statistical Institute, India  
umapada@isical.ac.in

Michael Blumenstein  
University of Technology Sydney, Australia  
michael.blumenstein@uts.edu.au

Abstract—The tendency of current technology is towards a paperless world. Due to the rapid increase of digitized documents, providing a fast and easy method for retrieval is in high demand. The aim of this paper is to examine the effectiveness of texture features for document image retrieval. Thus, segmentation-free document image retrieval using a binary texture method is proposed. In the proposed approach, local features are extracted, local grey-level structures are summarised, and their distribution is characterised using global features. The assumption is that texture properties in the text regions and non-text regions of the document images are different. This assumption is used to rank the available document images and retrieve only those, which have greatest visual similarity to a given query. The under-sampled image and sub-images of the original image are further considered to improve the retrieval results, which are up to 76.0% in the first ranking and 96.2% in the Top-10 ranking. The Media Team Oulu Document Database, which is a heterogeneous database that offers a great variety of page layouts and contents, is used for experimentation.

Keywords: Document image retrieval; Texture features; Local binary pattern.

I. INTRODUCTION

Fast and efficient access to a massive and daily increasing number of digitized documents is one of the necessities in the current digital world. Document image retrieval (DIR) is a research area between classic information retrieval (IR) and content-based image retrieval (CBIR) [1]. DIR has commonly been performed in two different ways in the literature: recognition-based and recognition-free based approaches. Optical character recognition (OCR), as a recognition-based approach, has widely been used as the basis and heart of many DIR systems. However, this approach has some weaknesses, such as high computational cost, language dependency, and sensitivity to image resolution, when employed with complex images [2]. Thus, recognition-free based methods, which rely on document image features, have been proposed for better efficiency.

Several methods developed in the literature for document image retrieval can be relatedly grouped into different categories, such as page layout similarity using graph and modified XY (MXY) tree representation [2, 3]; layout structural similarity using visual features and distance measures [4, 5]; signature-based feature methods [6]; horizontal and vertical traverse density of the character objects [7]; proportion of the black pixel area in character bounding box areas [8]; and density distributions using bag-of-visual words features [9]. The drawbacks of these approaches are remarkable, as the documents should have structured layouts, and similar fonts and size; furthermore, they need to be from the same language source.

Volume of data is the main issue in DIR problems. Therefore, the DIR methods that are conceptually simple and easy to implement and, most importantly, fast and computationally inexpensive, are preferred in the literature [10]. Texture features are generally fast to compute and they are computationally suitable for large volume data. Texture feature analysis has played a significant role in other research areas such as medical imaging [11], industrial inspection, and remote sensing [12]. Many researchers from the document image analysis (DIA) community have used various texture features in different stages, such as segmentation, layout analysis, and recognition of different document analysis systems [13-15].

The important information about the structural arrangement of each document and their relationships to the surrounding area is represented using texture features [16]. In other words, the spatial arrangement of grey-levels of pixels in a region of an image can be described as a texture feature. In addition, texture features have been used to detect and classify images based on local spatial variations of intensity and colour [17]. Different texture information extracted from document images alone or in a combination can be used to characterize document images for their further retrieval [18]. Texture approaches have also been applied to retrieve pictures in historical documents in [13]. Five features related to frequencies and orientations from different parts of a page at different resolutions have been considered for retrieval purposes. In [14], characterization of the content of a document image has been carried out without considering the structure and characteristics of treated pictures. Measurement of similarity of documents to a query image based on query-by-example has been computed. Texture features such as visibility and complexity have been provided to determine visual similarity of document images. In [19, 20], texture approaches have been used for segmentation and classification of document images. Text and graphic content have been characterized in two different textures in the documents. Supervised and un-supervised segmentation schemes have been considered, and a better classification performance has been shown using supervised segmentation [19].
There are various approaches for extraction and analysis of texture features in the related literature. These approaches are traditionally divided into four categories: a) statistical approaches, which consider the distributions of the pixel and relationships between the grey-levels of an image; b) structural approaches, which provide a good symbolic description of the image and these approaches are more useful for synthesis than analysis tasks; c) model-based approaches, which use a weighted average of pixel intensities in their neighbourhood, and can generate an empirical model of each pixel in the image; and d) transform-based approaches which use the spatial frequency properties of pixel intensity variations, converting the image into a new form [17]. Among these categories, statistical methods have provided better accuracy at a reasonable cost, and as a result they have become more popular in comparison to other approaches [21].

To the best of our knowledge, little attention has been paid to texture features for document image retrieval. Considering the amount of research on DIR using texture features, and their applicability regarding CBIR, a texture-based approach for DIR is proposed in this research work. We investigate the suitability of texture features for DIR, providing an extensive experimental analysis on a benchmark database. The considered notion in these experiments is the difference between texture properties in the text regions and non-text regions in the document images. The first objective of this research work is to rank the documents in the database, and find the document images which have the greatest visual similarity to a query image. A document that has the greatest similarity to a query image can be placed in the first ranking, which is the best condition, or in one of the Top-10 ranking. The second objective is to organize a retrieval procedure with a simple, fast and, low computational processing. For these reasons, a group of local binary methods, which are the most successful texture features due to their high performance and discriminative abilities, is considered [22]. The steps in the proposed method are as follows. Firstly, texture features are extracted from the whole a document image. Secondly, by dividing the image into four sub-images based on the centre of gravity of the image, the spatial distribution of the grey-scale values are analysed and presented in terms of histograms. To improve the retrieval results, the same features are extracted from the under-sampled images of the original image and included in the knowledge-based information. Finally, similarity between a query and the knowledge-based information are measured to rank the retrieval results.

The remainder of the paper is organized as follows. In Section II, the proposed approach for document image retrieval is explained. Experimental analysis based on a variety of LBP is discussed in Section III and finally conclusions and future work are presented in Section IV.

II. PROPOSED APPROACH

The stages involved for document image retrieval in this research work are demonstrated in Figure 1. The proposed approach is composed of two phases, training and testing. In the training phase initially texture features are extracted from the whole a document image by applying a variety of LBP-based techniques without considering any pre-processing or segmentation methods. The features extracted from the training document images are then kept as knowledge-based information to be used in testing phase. In testing phase, for a given query image, the same feature extraction technique used in the training phase is employed to characterize the query image. Similarity between the features extracted from the query image and the knowledge-based information obtained during the training phase are computed. Finally the relevant image(s), which have maximum visual similarity to the query image, are retrieved and sorted accordingly.

Figure 1. Block diagram of the retrieval process

A. Feature Extraction

As mentioned, texture features have widely been used in various computer vision and pattern recognition problems in the literature including object recognition, face recognition and background subtraction [25-27]. However, these features have not thoroughly been considered for document image retrieval. In this research work a variety of texture features are employed for document image retrieval to study the effectiveness of texture feature particularly Local Binary Patterns (LBP) features and its family. Textural features are mostly extracted from 3×3 and 5×5 block sizes of document images. Since sufficient information about the spatial structure of local texture cannot be provided using a single texture measure, joint occurrences of two or more features have been used for better discrimination of textures [23]. Difference between pairs of grey-levels in a variety of orders is considered in each block of document images. To measure the frequency of occurrence of these differences over the image, a histogram is provided. The histogram is computed based on movement of the pixel and the neighborhood through the image [10]. The LBP-based feature extraction techniques are statistical approaches in nature and the texture features are also calculated in the spatial domain.

1) Local Binary Pattern (LBP)

The local binary pattern texture feature [23, 24] has been effectively used for various computer vision problems and shown its efficiency in different applications [25-27]. It, however, has not been used for document image retrieval so far. The LBP method has powerful discrimination, low computational complexity, and less sensitivity to changes in illumination [28].
The LBP method has been introduced as a means of summarizing local grey-level structure in [24]. For image I, the difference between the grey values of a centre pixel C and the grey values of its N neighbourhood, placed by a patch on a circle with a radius of R, have been considered.

Circularly symmetric neighbour sets for different (N, R), can produce different patch sizes. The values of N = {4, 8, 12, 16, 24} and R = {1, 1, 1.5, 2, 3}, respectively, have been taken into account for defining the patches. By considering the number of neighbours, LBP\textsubscript{N,R} produces 2\textsuperscript{N} output values which correspond to the different binary patterns. It produces a binary pattern based on the pixels around. When the number of transactions between 0 and 1 of the sequence is less than or equal to two, a binary pattern is called uniform, which is an important concept in the LBP. However, there is a similarity between LBP operators and LBP\textsubscript{N,1} since they form a circular chain [24].

The central pixel value of each patch has been considered as a threshold for that specific patch. Moreover, the central pixel gets the resulting binary valued image patch as a local image descriptor. Thus, the result of the LBP operator has been made a rotation invariant, performing n - 1 bitwise shift operations on the binary pattern.

\[
LBP_{N,R} = \sum_{n=0}^{N-1} 2^n s(i_n - i_c) \quad (1)
\]

\[
s(i_n - i_c) = \begin{cases} 
1 & \text{if } (i_n - i_c) \geq 0, \\
0, & \text{otherwise}
\end{cases} \quad (2)
\]

where n includes the number of neighbours of the central pixel c, i\textsubscript{c} and i\textsubscript{n} are the grey-level values at c and n. Let S represent a matrix of 3x3 whereby neighbour pixels are indexed.

\[
S = \begin{bmatrix} 
i_7 & i_6 & i_5 \\
i_4 & i_c & i_4 \\
i_1 & i_2 & i_3
\end{bmatrix} \quad (3)
\]

2) Median Binary Pattern (MBP)

The term MBP is another form of the LBP method [22], which generally refers to a localized binary pattern by comparing the pixels to their median as a threshold.

\[
MBP = 2^n f(i_c - M) \sum_{n=0}^{N-1} 2^n f(i_n - M) - 1 \quad (4)
\]

\[
f(i_c, n) - M) = \begin{cases} 
1 & \text{if } (i_c, n - M) \geq 0, \\
0, & \text{otherwise}
\end{cases} \quad (5)
\]

where M is the median value over a 3x3 neighbourhood. Thus, MBP refers to the operations of 3x3 patches and produces 2\textsuperscript{N} binary patterns. Based on the median value, the MBP operator divides the image into two groups and captures contrast between two intensity ranges that impact the local structure.

3) Improved Local Binary Pattern (ILBP)

The idea of ILBP [29] is also similar to LBP; the only difference is that grey-level neighbourhood pixel values are being thresholded by the local average grey-level value (instead of centre pixel value). This notion has been applied on LBP to reduce the effect of noise.

\[
ILBP = 2^n f(i_c - V) \sum_{n=0}^{N-1} 2^n f(i_n - V) \quad (6)
\]

\[
f(i_{c,n} - V) = \begin{cases} 
1 & \text{if } (i_{c,n} - V) \geq 0, \\
0, & \text{otherwise}
\end{cases} \quad (7)
\]

where V is defined as the average value of a specific patch. In this operation, 2\textsuperscript{N} - 1 binary patterns are produced.

\[
V = \frac{1}{N} \sum_{n=0}^{N-1} i_c \quad (8)
\]

4) Binary Gradient Contours (BGC)

The BGC method is a simple texture descriptor whereby intensity invariance plays an important role in that. The gradients between the closed paths around the central pixel in a 3x3 patch have been considered. To gain the gradient, a pairwise comparison has been carried out in three different paths. The paths have been created based on pattern mapping which is called a single-loop, double-loop, and triple-loop [21].

In the case of a single-loop (BGC1), the pixel order for a path has been defined as \{i_0, i_7, ..., i_1, i_0\}. Hence, the binary patterns are extracted from the corresponding pairs \{(i_7 - i_0), (i_6 - i_1), ..., (i_0 - i_7)\}. The order of pixels for the triple-loop (BGC3) has been defined as \{i_0, i_2, i_5, i_3, i_1, i_4, i_0, i_2, i_5\}. It has been noted that based on the inequality in the definition, at the same time eight components cannot be 0’s [21]. Thus, both of the operators have generated 2\textsuperscript{8} - 1 gradient binary contour features. However, in double-loop (BGC2) two closed paths have been provided, which include \{i_1, i_0, i_5, i_4, i_1\} and \{i_1, i_0, i_5, i_4, i_2, i_2, i_0\}. In this case, the operator has generated 2\textsuperscript{8} - 1 gradient binary contour features in each path. Thus, the combination of two loops produce (2\textsuperscript{4} - 1)\textsuperscript{2} = 225 bins.

5) Under-sampled Image Creation

One of the issues in image retrieval system is “how to handling image retrieval when images with different resolutions are present in a database”. In this research work, we proposed to use the concept of under-sampling used in early stage of OCR [30]. The under-sampling method is a simple process that has been used to reduce the dimension and complexity of images. To do so, an image is divided into a number of non-overlapping blocks of similar size (b\times b). In each block, the average grey intensity of all pixels is then computed and the block with the average grey intensity is considered as a pixel of the new image. Each element of the new image has a value in the range 0 to 255. The generated image obtained employing the under-sampling process is called under-sampled image, which is also a lower dimension image compared to the original image. The generated image provides invariance to small distortion as well. Using the under-sampled image, the same feature extraction technique considered to extracted features from the original images is employed on the under-sampled image to characterize this lower dimension image.

B. Creation of Knowledge-based Feature

All the above mentioned features provide global features and a global representation of a document image by the means of a feature histogram. The feature histogram contains different numbers of bins according to each feature extraction method employed for characterizing the document image. To consider spatial relationship between different
parts of the document image and also preserve local information of the document, the document image is divided into four sub-images based on the centre of gravity of the image. The same feature extraction method employed to extract the global features, is then applied on each sub-image. The feature sets extracted from the original image and all sub-images are concatenated to obtain the final feature set representing the document image.

To take care of the variation in resolutions for document image retrieval, two lower dimension document images are obtained from each training image considering two different block sizes for under-sampling the original image. The very similar feature extraction process employed on the original image is also applied on each lower dimension image to characterize this image. The feature sets obtained from the lower dimension (under-sampled) images are further included in the knowledge-based features. By doing so, more information from the image can be preserved in the training phase for creation of the knowledge-based to be used for the retrieval purposes.

C. Similarity Measure and Retrieval Process

To measure the similarities between a query image and the knowledge-based features obtained from the trained samples, the Euclidean distance and the Tanimoto coefficient [34] are considered. In a preliminary experiment, we noted that the Tanimoto distance method provides better results compared to the Euclidean distance for finding the most similar documents to a query image. Thus, all the ensuing results are based on the Tanimoto distance method [34].

By giving a query to the system, the distance of a query image and all the trained documents are computed by the help of the Tanimoto distance. Then, the documents are listed based on a high degree of similarity to a query image. The best results can be obtained, when the document with the greatest visual similarity is in the same class as the query document image.

III. EXPERIMENTAL ANALYSIS

A. Database and Evaluation Metric

To technique evaluate the proposed, we considered the Media Team Document Database (MTDB) composed of heterogeneous document images [31]. Various types of scanned documents including address list, advertisement, article, business cards, music notes, and so on are available in the database. A great diversity of page layout and contents is provided in the database, in addition to the ground truth for physical segmentation. For any document image in high-resolution, there is a low-resolution and a thumbnail size document image available in the database that are also tested. However, some of the classes have less than a dozen images. In our experiments, we focused on eleven of nineteen predefined classes of the MTDB that contain mainly text documents. The other 8 classes are of small in number (2 images in the class) or of the music notes and maps.

We performed different experiments considering different training and testing sets. In each experiment, the training and testing sets did not have any overlapping. For training the proposed retrieval method, we initially used only one sample from each group of document images for training. As a result, only 154 images were considered for training and the rest (1168 images) were taken into account for testing. We then performed the training by considering 2 and 3 document images from each group. The results are demonstrated in different tables in the following.

The system is evaluated using the accuracy value measured based on the results obtained from the proposed method ranked among the first, Top-3, Top-5, and Top-10 similar documents to a query document image.

B. Results and Discussion

Since the MTDB is composed of document images with different resolutions, we initially trained the proposed system using only one document image from each group with the same resolution. The system then, was trained by different resolution images to see the effect of image size and image resolution on texture features for retrieval purposes. Here, we have not included the training features obtained from the under-sampled (lower dimension) images into the knowledge-based features. Considering different feature extraction techniques proposed in this research work, the results obtained from the proposed system are shown in Tables I.

From Table I, it can be noted that by considering only one high-resolution document image from each group for the

<table>
<thead>
<tr>
<th>Methods</th>
<th>High-resolution</th>
<th>Low-resolution</th>
<th>Thumbnail size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1 (%)</td>
<td>Top-3 (%)</td>
<td>Top-5 (%)</td>
</tr>
<tr>
<td>LBP</td>
<td>55.1</td>
<td>66.4</td>
<td>68.3</td>
</tr>
<tr>
<td>LBP_{w1}</td>
<td>55.7</td>
<td>65.8</td>
<td>67.6</td>
</tr>
<tr>
<td>LBP_{h1}</td>
<td>55.8</td>
<td>66.4</td>
<td>68.3</td>
</tr>
<tr>
<td>LBP_{1215}</td>
<td>55.7</td>
<td>64.6</td>
<td>67.3</td>
</tr>
<tr>
<td>LBP_{162}</td>
<td>54.6</td>
<td>64.0</td>
<td>66.9</td>
</tr>
<tr>
<td>MBP</td>
<td>45.5</td>
<td>58.5</td>
<td>61.5</td>
</tr>
<tr>
<td>ILBP</td>
<td>40.7</td>
<td>52.2</td>
<td>55.2</td>
</tr>
<tr>
<td>BGC1</td>
<td>39.6</td>
<td>52.5</td>
<td>55.7</td>
</tr>
<tr>
<td>BGC2</td>
<td>42.3</td>
<td>57.8</td>
<td>61.7</td>
</tr>
<tr>
<td>BGC3</td>
<td>49.7</td>
<td>63.0</td>
<td>66.1</td>
</tr>
</tbody>
</table>
training, the LBP$_{8,1}$ method provided a correct document retrieval of 55.8%, whilst only the first similar document images to the queries were considered. The correct retrieval accuracy increases to 69.5% in the Top-10 similar document images to the query image when LBP used as feature extraction. We further noted that original LBPs comparably provided better results in high resolution images compared to the MBP, ILBP, and BGCs feature extraction techniques.

Considering the results obtained from the proposed system while the system is trained by low-resolution images, it is evident that the highest accuracy in the first rank belongs to the BGC3 method. The MBP method provided the best retrieval accuracy of 72.9% when the first Top-10 samples were considered for evaluating the accuracy.

Out of 1168 samples, which are given as a query to the system, the system using the LBP$_{16,2}$ technique could correctly retrieve 52.0% of document at the first rank and 67.7% in the Top-10 ranks of samples, while training samples are thumbnail size.

Comparing the results shown in Table I, it can further be noted that the results in Top-1 were decreased when the low resolution images and thumbnail size were considered for training. Moreover, as the LBP-based methods considered the relationship between a pixel and its surrounding pixels, in a high-resolution image, these relationships can provide more precise and more accurate information compared to those using a small image size.

![Figure 2. Samples and their class are listed based on the most similar to a query document image.](image)

One example of a query document image and of the Top-10 similar documents to the query, as well as their classes, is listed from top left to bottom right in Figure 2. Since the exact version of a query is not present between trained documents, another resolution of the query image was retrieved in the fifth rank. It can also be noted that among ten retrieved documents, eight documents belong to the same class as the query and the results are encouraging.

Since better results were obtained with the LBP$_{8,1}$ feature extraction method using the high-resolution document images, the rest of the experiments were explored with this method. Only 12% of the data were initially considered for training the proposed system, we have increased the training samples and the results obtained from the proposed system, are demonstrated in Table II. It is obvious that the increasing the training size improves the retrieval accuracy.

![Table II. The results obtained by increasing the number of training samples using the LBP$_{8,1}$ method as the best features in high-resolution.](image)

<table>
<thead>
<tr>
<th>No. of samples for training</th>
<th>No. of samples for test</th>
<th>Top-1 (%)</th>
<th>Top-3 (%)</th>
<th>Top-5 (%)</th>
<th>Top-10 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>154</td>
<td>1168</td>
<td>55.8</td>
<td>65.1</td>
<td>68.2</td>
<td>69.3</td>
</tr>
<tr>
<td>308</td>
<td>1014</td>
<td>57.6</td>
<td>70.4</td>
<td>73.2</td>
<td>77.9</td>
</tr>
<tr>
<td>462</td>
<td>860</td>
<td>58.3</td>
<td>76.5</td>
<td>81.1</td>
<td>88.1</td>
</tr>
</tbody>
</table>

When training samples were increased to 34% of samples, and considering three different resolutions available in the database for training, the accuracy of the system using the LBP$_{8,1}$ features was improved to 58.3% for the first similar documents. Meanwhile, the Top-10 similar documents provide accuracy up to 88.1%.

The under-sampled version of document images by choosing the block sizes of 3×3 and 5×5 were further included in the training set to create the knowledge-based features. The results obtained from the new knowledge-based features considering the under-sampled images are shown in Table III. Overall, these results indicate that an accuracy of 76.0% in the first rank and an accuracy of 96.2% in the Top-10 ranking were achieved from our proposed system using a complex and unstructured database.

![Table III. The results obtained by including the under-sampled version of document images in the training set.](image)

<table>
<thead>
<tr>
<th>Samples Methods</th>
<th>Top-1 (%)</th>
<th>Top-3 (%)</th>
<th>Top-5 (%)</th>
<th>Top-10 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP$_{8,1}$</td>
<td>76.0</td>
<td>87.6</td>
<td>90.9</td>
<td>96.2</td>
</tr>
</tbody>
</table>

C. Analysis of Errors

The confusion matrix of the system, trained by 462 samples using the LBP$_{8,1}$ method (with the accuracy of correct retrieval in each class), is shown in Table IV. By looking at Table IV, it can be observed that more confusion occurs in Class 3, as many of the documents in other classes have some similarities with documents in that class. Therefore, the system recognizes many of the documents as Class 3 in the first rank. This poor performance can also be noted due to the paucity of available data for training purposes in the classes. However, by providing under-sampled versions of the documents and getting more spatial information from the document images, this problem could be alleviated. The other main point for the error in the proposed system might be the distribution of data samples in each class of the MTDB. The MTDB is an imbalance database as can be seen from Table IV.
D. Comparative Analysis

In previous research on document image retrieval [4, 5, 7, 32, 33], more attention has been devoted to structured documents for the retrieval process. One of the advantages of the proposed system in this research work is obtaining positive results when it is applied to heterogeneous and complex databases such as the MTDB. In [15], Transform-based approaches have been used for classifying the MTDB. The wavelet transform method, along with an SVM classifier and the ratio of 1:1 for training and testing have provided 25.4% in the case of text-only images. For hybrid images and accuracy of 45.3% has also been recorded. The accuracy of our proposed method in the first trial is, however, 55.8% considering the LBP_{8,1} features and eleven most frequent classes of document images of the MTDB. It may not be an altogether equivalent comparison of the results, since the whole database is divided into three types for evaluation. However, as the same database is used, along with texture features for classification, these are the reasons for the analogy.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, the effects of binary texture methods to retrieving the document image(s) that have greatest visual similarity to a query image are investigated. The spatial arrangements of the grey-levels of the pixels in a region of an image are described as a texture feature to measure this similarity. In the proposed method the structure of the document image is taken into account without performing any segmentation method. Subsequently, considering undersampled versions of the images in the knowledge-based features produced better document image retrieval results. It can also be noted that texture features demonstrate their strength for retrieval of digitized documents.

In future work, the plans are to perform more experimentation and investigation about other categories of texture descriptors to achieve favourable results. Moreover, the use of prior knowledge about the document classes may also be integrated into the feature set to improve the retrieval performance.

REFERENCES


