Document Image Retrieval from Incomplete Queries Using Texture Features

Ph.D.(Assist.Prof.) Matheel Emaduldeen *, Mohammad Talib Hashim*

Abstract

Document image retrieval (DIR) is an important part of many document image processing systems such as paperless office systems, digital libraries and so on. It helps the users to find out the most similar document images from a document image database. Most of the researches have been carried out with complete queries which were present in the database, but in many cases distorted or incomplete images can be encountered. This distortion or incompetence is due to some missing information, some undesirable objects, blurring, noise due to document printing, scanning etc. This paper describes an approach for retrieval of incomplete and distorted document images based on visual features using texture information for retrieval from large document image database. A Gray Level Co-occurrence Matrix (GLCM) features for texture analysis were proposed and provide a comprehensive experimental evaluation.

Keywords:

Document image retrieval, Texture Features, GLCM, Incomplete queries.

*Computer Science Department, University of Technology

1. Introduction

Search and retrieval of relevant documents from a large collection of document images has been a problem of interest for many years. Document images are documents that typically start out on paper and are then electronically scanned. These documents have rich internal structure and might only be available in image form. Additionally, they may have been produced by a combination of printing technologies (or by handwriting); and include diagrams, graphics, tables and other non-textual elements. The indexing and analysis of large document collections is currently limited to textual features based on Optical Character Recognition (OCR) data and ignore the structural context of the document as well as important non-textual elements such as logos, tables, diagrams, and images. Recently, many researchers have been focused on document images retrieval based on different kinds of low level image features. These methods are OCR free. The great challenge for these methods is how to extract some discriminative features from document images and use them to develop a robust retrieval algorithm. However, this is not an easy work partly due to the fact that we cannot find such a feature which can well represents the original images and thus providing a good similarity measurement between two document images.

2. Related Works

Several researchers have considered the use of such features for imageretrieval.John F. Cullen et al. presented a system that uses texture to retrieve and browse images stored in a large document image database [3]. A method of graphically generating a candidate search image is used that shows the visual layout and content of a target document. H. Liu et al. proposed a retrieval method based on density distribution feature and key block feature of document images [1]. The features are very simple and robust to document images with different resolutions, formats and multiple languages. M. W. Lin et al. present a hybrid approach to segment and classify contents of document images [4]. The image of a document is subdivided into blocks and for each block five GLCM features are extracted. Meng et al. proposed document imagesretrieval method based on multiple features combination [2]. In

their method, two new kinds of document image features are proposed based on projection histogram and crossings number histogram respectively. Abu Sayeed et al. present an effective solution for content-based retrieval and classification of ultrasound medical images [6]. They extract low level ultrasound image features combining histogram moments with GLCM based statistical texture descriptors and use of multiclass support vector machine for classifying image features into their corresponding high level categories. B. K. Singh et al. investigated utility of content based image retrieval techniques for retrievean incomplete and distorted queries that uses Hue-Saturation-Value color space model and shape features to represent the image [5].

In this paper, GLCM features were proposed to construct a new document image retrieval system. Experiments show that the system is very efficient for retrieving incomplete document images.

The organization of the paper is as follows: Sections 3explains the document image retrieval, Section 4 explains the method descriptor, Section 5deals with experimental evaluation and discussion, and finally Section 6presents the conclusions.

3. Document Image Retrieval

Document Image Retrieval (DIR) aims at finding relevant documents relying on image features only. DIR is performed in two steps: indexing and searching. In indexing step contents (features) of the image are extracted and are stored in the form of a feature vector in the feature database. In the searching step, user query image feature vector is constructed and compared with all feature vectors in the database for similarity to retrieve the most similar images to the query image from the database [7].

3.1Texture Feature Extraction

The first issue in DIR is to extract the features of the image efficiently and then represent them in a particular form to be used effectively in the matching of images.

Texture is a natural property of surfaces and it provides visual patterns of the image. It has repeated pixel of information and it contains vital information regarding the structural arrangement of the surface (example clouds, leaves bricks). It also gives the relationship between the surface and external environment.

Texture features can be extracted in several methods, using statistical, structural, model-based and transform information. The statistical texture features are considered useful for the classification and retrieval of similar images. These texture features provide the information about the properties of the intensity level distribution in the image like uniformity, smoothness, flatness, contrast and brightness [7]. The identification of specific textures in an image is achieved primarily by modeling texture as a two-dimensional gray level variation. This two dimensional array is called as Gray Level Co-occurance Matrix (GLCM) [21].

Co-occurrence features are popular and effective texture descriptor using statistical approach. Given an image of n gray levels, characteristics of images are estimated from the second-order statistical features by considering the spatial relationship of pixels in the image. A GLCM element $P_{d,\theta}(i, j)$ is the joint probability of the gray level pairs i and j in a given direction θ separated by distance of d units. Multi-distance and multi-direction can be used to extract a large number of features. The probability measure can be defined as equation 1: [8, 9, 10].

$$\Pr(x) = \{Cij|\ (d, \theta)\} \quad (1)$$

Where, Cij (the co-occurrence probability between gray levels i and j) is defined as equation 2:

$$Cij = \frac{Pij}{\sum_{i,j=1}^{G} Pij} \quad (2)$$

Where:

Pij: Represents the number of occurrences of gray levels
i and j : Within the given image window, given a certain (d, θ) Pair
G : The quantized number of gray levels.The mathematical formulas used to calculate the Co-occurrence features are shown inTable 1.First 13

features suggested by Haralick et al. [8], features 14 -18 suggested by Leen-Kiat et al. [9].

Table 1: The mathematical definitions of GLCM features.

Feature	Equation	No.
Angular Second Moment (Energy)	$f_1 = \sum_{i,j=0}^{G-1} (\mathrm{P}(i,j))^2$	(3)
Contrast	$f_2 = \sum_{i,j=0}^{G-1} (i-j)^2 \mathrm{P}(i,j)$	(4)
Correlation	$f_3 = \frac{\sum_{i,j=0}^{G-1} (ij)\mathrm{P}(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$ where $\mu_x = \sum_{i,j} i.p(i,j)$ (5a) $\mu_y = \sum_{i,j} j.p(i,j)$ (5b) $\sigma_x = \sum_{i,j}(i-\mu_x)^2.p(i,j)$ (5c) $\sigma_y = \sum_{i,j}(j-\mu_y)^2.p(i,j)$ (5d)	(5)
Sum of Squares (Variance)	$f_4 = \sum_{i,j=0}^{G-1} (i-\mu)^2 \mathrm{P}(i,j)$	(6)
Inverse Difference Moment (Homogeneity)	$f_5 = \sum_{i,j=0}^{G-1} \frac{\mathrm{P}(i,j)}{1+(i-j)^2}$	(7)
Sum average	$f_6 = \sum_{i=0}^{2G-2} i\mathrm{P}_{x+y}(i)$	(8)

Sum variance	$f_7 = \sum_{i=0}^{2G-2} (\mathrm{i} - \mathrm{SENT})^2 \mathrm{P}_{x+y}(\mathrm{i})$	(9)
Sum entropy	$f_8 = -\sum_{i=0}^{2G-2} \mathrm{P}_{x+y}(\mathrm{i}) \log(\mathrm{P}_{x+y}(\mathrm{i}))$	(10)
Entropy	$f_9 = -\sum_{i,j=0}^{G-1} \mathrm{P}(\mathrm{i},\mathrm{j}) \log(\mathrm{P}(\mathrm{i},\mathrm{j}))$	(11)
Difference variance	$f_{10} = \text{Variance of } p_{x-y}$	(12)
Difference entropy	$f_{11} = -\sum_{i=0}^{G-1} \mathrm{P}_{x+y}(\mathrm{i}) \log(\mathrm{P}_{x+y}(\mathrm{i}))$	(13)
Information measures of correlation 1	$f_{12} = \frac{HXY - HXY1}{\max(HX, HY)}$ Where HX, HY: entropies of p_x and p_y $HXY = -\sum_{i,j=0}^{G-1} \mathrm{P}(\mathrm{i},\mathrm{j}) \log(\mathrm{P}(\mathrm{i},\mathrm{j}))$ (14a) $HXY1 = -\sum_{i,j=0}^{G-1} \mathrm{P}(\mathrm{i},\mathrm{j}) \log(\mathrm{P}_\mathrm{x}(\mathrm{i})\mathrm{P}_\mathrm{y}(\mathrm{j}))$ (14b)	(14)
Information measures of correlation 2	$f_{13} = (1 - \exp[-2HXY2 - HXY])^{1/2}$ where $HXY2 = -\sum_{i,j=0}^{G-1} \mathrm{P}_\mathrm{x}(\mathrm{i})\mathrm{P}_\mathrm{y}(\mathrm{j}) \log(\mathrm{P}_\mathrm{x}(\mathrm{i})\mathrm{P}_\mathrm{y}(\mathrm{j}))$ (15a)	(15)
Autocorrelation	$f_{14} = \sum_{i,j=0}^{G-1} (ij) P(i,j)$	(16)

Dissimilarity	$f_{15} = \sum_{i,j=0}^{G-1} \lvert i - j \rvert P(i,j)$	(17)
Cluster shade	$f_{16} = \sum_{i,j=0}^{G-1} (i + j - \mu_i - \mu_j)^3 P(i,j)$	(18)
Cluster prominence	$f_{17} = \sum_{i,j=0}^{G-1} (i + j - \mu_i - \mu_j)^4 P(i,j)$	(19)
Maximum probability	f_{18} = MAX ($p(i,j)$) for all (i,j)	(20)

3.2 Feature Normalization

The scales of individual features can differ drastically. This disparity can be due to the fact that each feature is computed using a formula that can produce various ranges of values. Another problem is that, features may have the same approximate scale, but the distribution of their values has different means and standard deviation. The normalization is performed to enable all the features to have the same range of values that will result in an equal contribution of weight for the similarity measure [4].

Statistical normalization has been adopted that independently transforms each feature in such a way that each transformed feature distribution has means equal to 0 and variance equal to 1 as in equation (21). [4]

$$Xnorm_{ij} = \frac{(x_{ij} - \bar{x_j})}{\sigma_j} \quad (21)$$

where

$Xnorm_{ij}$: normalized feature
x_{ij}: feature value
x_j: mean value of jth coordinate, estimated on the training set.
σ_j: standard deviation of jth coordinate, estimated on the training set.

Note: the normalization is applied both to the training and the test sets, using x_j and σ_j computed on the training set.

3.3 Feature Selection

The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection methods can be decomposed into three broad classes. One is Filter methods and another one is Wrapper method and the third one is Embedded method [11]. One of the most promising feature selection methods for wrappers is the Sequential Floating Forward Selection algorithm (SFFS) [10]. The SFFS finds an optimum subset of features by insertions (i.e. by appending a new feature to the subset of previously selected features) and deletions (i.e. by discarding a feature from the subset of already selected features) that partially avoid the local optima of the correct classification rate (CCR) [11].

3.4 Similarity Measurements

Similarity measurement is the second issue in DIR in which the query image is compared with other database images. To measure the similarity between the query image and the database images, the difference is calculated between the query feature vector and the database feature vectors by using distance metrics. The small difference between two feature vectors indicates the large similarity and the small distance. The vectors of the images with small distances are most similar to the query image [7]. The distance metrics which are included in this work are the Euclidean distance and it can be calculated as in equation 22: [7]

$$\Delta \boldsymbol{d} = \sqrt{\sum_{i=1}^{n} (|\boldsymbol{Qi} - \boldsymbol{Di}|)^2} \qquad (22)$$

Where:

Qi: the i^{th} query image feature

Di: the corresponding feature in the feature vector database

n: refers to the number of images in the database

3.5 K-Nearest Neighbor

Given a query image, DIR system tries to find out the most similar images in a database. Essentially, document images retrieval can be regarded as a process of searching the nearest neighbors (NNs) in a given feature space [13].

In the K-Nearest Neighbor (KNN) algorithm, in order to classify a new input pattern, its K nearest neighbors from the training set are identified. The new pattern is classified to the most frequent class among its neighbors based on a similarity measure that is usually the Euclidean distance [13].

3.6 Evaluation Measures

Precision and recall measures are widely used for evaluation the retrieval performance of the image retrieval system. They are defined as follows [7]:

$$\text{precision} = \frac{\#(\text{relevant retrieved records})}{\text{total number of retrieved record}} \quad (23)$$

$$\text{recall} = \frac{\#(\text{relevant retrieved records})}{\text{total number of relevant record}} \quad (24)$$

4. The Proposed Method

The proposed method consists of two processing stages: (i) images indexing, and (ii) images retrieval. The block diagram of the system is shown in Figure 1.

4.1 Images Indexing

An image is indexed by a feature vector representing the texture descriptor in this image. The steps of such process are summarized in the following:

Input: dataset of images

Output: an indexed dataset of images.

Step 1: Loading images.

Step 2: Convert the Red, Green and Blue (RGB) color image into gray level image.

Step 3:Feature extraction: for each image do the following:

a. Construct a co-occurrence matrix $P_{d,\theta}(i, j)$ in which the (i,j)th element describes the frequency of occurrence of two pixels (i and j) that are separated by distance d=1 in the direction θ. Four directions are performed in this work (θ=0°,45°,90° and 135°). This produces 4 matrices of ($Q_1 \times Q_1$) integer elements per matrix.

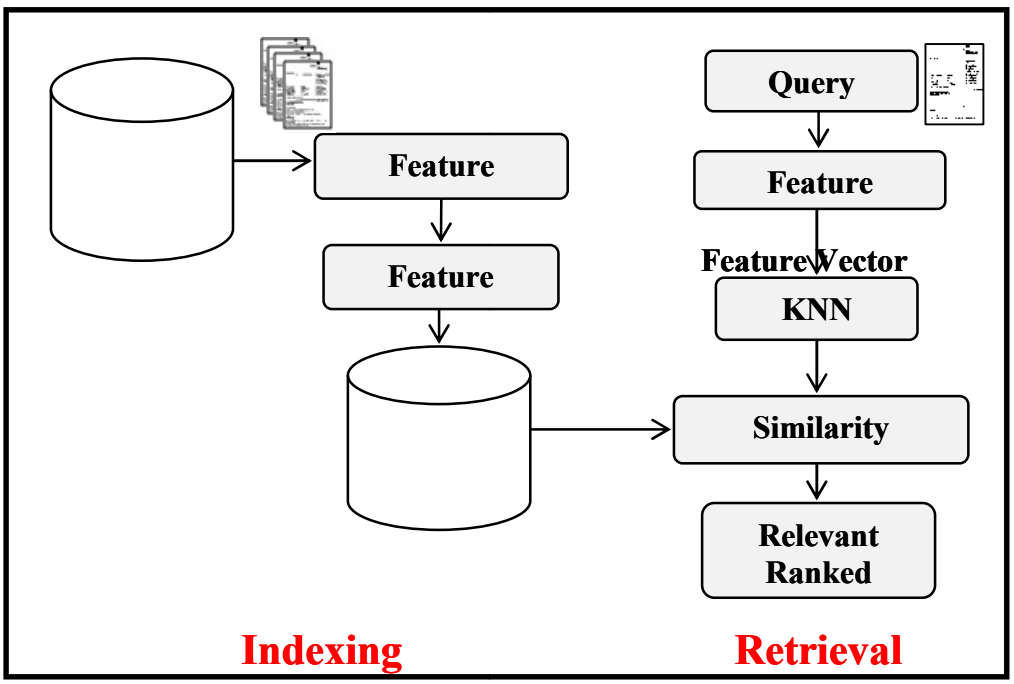

Figure1:The block diagram of the proposed system.

b. Normalize the 4 co-occurrence matrix $P_{d,\theta}(i, j)$ by dividing each entry by the summation of all entries in the matrix of the same direction. Hence, treating the matrix as a probability density function.

c. Feature extraction for each normalized co-occurrence matrix:

Eighteen features were chosen from the normalized co-occurrence matrix, the mathematical definitions of these features are illustrated in table 1.

d. Compute the mean of each feature from each direction ($\theta = 0^{o}, 45^{o}, 90^{o}$ and 135^{o}), this value of the feature is transformed into a suitable vector form of features.

e. After extracting features, then the feature vector is constructed as a collection of 18 features.

Step 4: Repeat steps (1 to 3) for all images in the database.

Step 5:Features normalization: all the features vectors are normalized (i.e.mapped to the range [0,1]) in order to make all features having equal effect on the similarity comparison process as in equation (21)

Step 6: Feature selection algorithm is applied to select best feature subset from the feature vector. In this work, we use SFFS algorithm to select best feature subset with KNN algorithm as a base classifier in the experiment.

After applying feature selection algorithm, six features are selected as a best subset features as follows:

1. Contrast
2. Entropy
3. Cluster prominence
4. Variance
5. Difference variance
6. Information measure of correlation 2

Then the feature vectors shrink into a collection of these 6 features and stored into features database.

4.2 Images Retrieval

Given a query image, DIR system tries to find out the most similar images in a database. The steps of such process are summarized in the following:

Input: query image.

Output: most similar images in the database.

1. Loading query image.
2. Convert the RGB color image into gray level image.

3. Feature extraction: repeat steps (a to e) as in indexing process, but only for the 6 selected features.

4. Features normalization: the normalization is applied to the query image features, using x_j and σ_j computed on the training set.

5. Matching: this step is based on finding the minimum distance between the input query image features and the preserved features in the database, the distance metrics which are included in this work are the Euclidean distance and it can be calculated as in equation (22).

The matching process using KNN algorithm includes two main steps. The first step is to classify the input query image into its category. The second step is to retrieve the most similar images from the image database with the same category as the query image.

6. Output Stage: this stage is responsible for displaying the relevant ranked documents.

5. Experimental Evaluation and Discussion

Experiments were carriedoutonaset of600images, digitized with 200 dpi resolution. The images came from different academic publications, brochure and business letter, with mixed texts, graphics and pictures.Figure 2 showed an example of document images in the database. Besides, to make the evaluation more convenient, a special testing set is also added into the database. Each document had different versions by adding different amount of handwritten notes, different kinds of noises,different skew angles (-3° ~ +3°),and different parts deleted from the images.

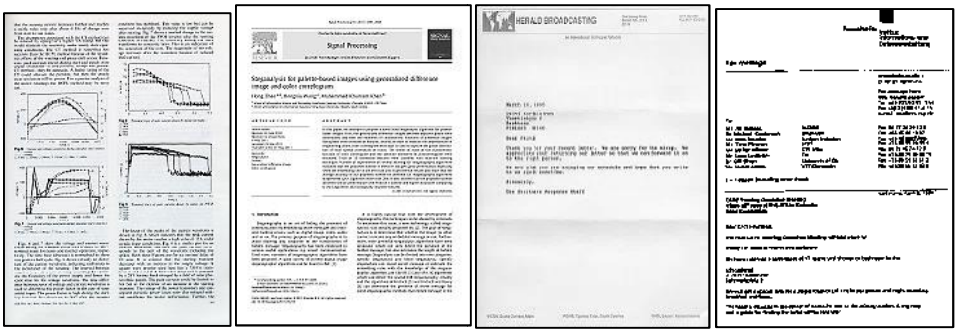

Figure 2: Examples of document images used in the experiments.

Discussions

Experiments are designed to test the performance of the proposed DIR system based on the special testing set. To evaluate the retrieval performance, we have randomly selected 50 images as the query images. We adopted the “Query by Example” method for submitting the query to the retrieval system. A retrieved image was considered a match if it belongs to the same category as that of the query image.Figure 3illustrated the retrieval results for the top 5 similar images of the query image.

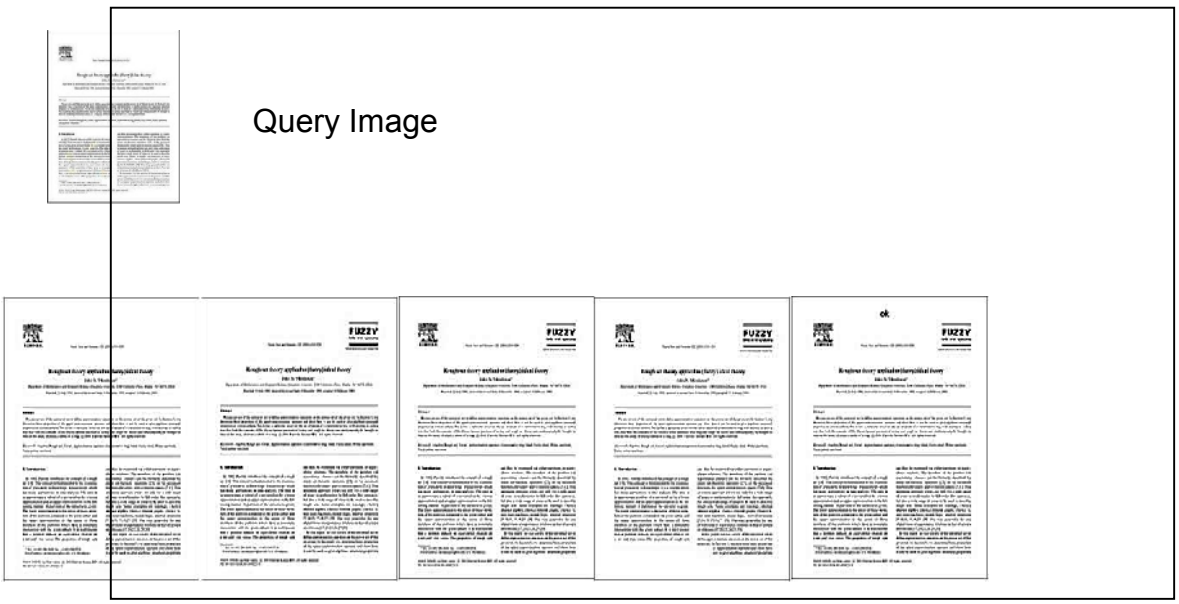

Figure 3: Retrieval result for the top 5 similar images of the query image.

Figure 4 demonstrates the precision and recall curves drawn by calculating the average precision values from the retrieved images. As can be observed from this graph, the average precision value lies above 70% for the first 20 retrieved images, which indicates very satisfactory retrieval performance.

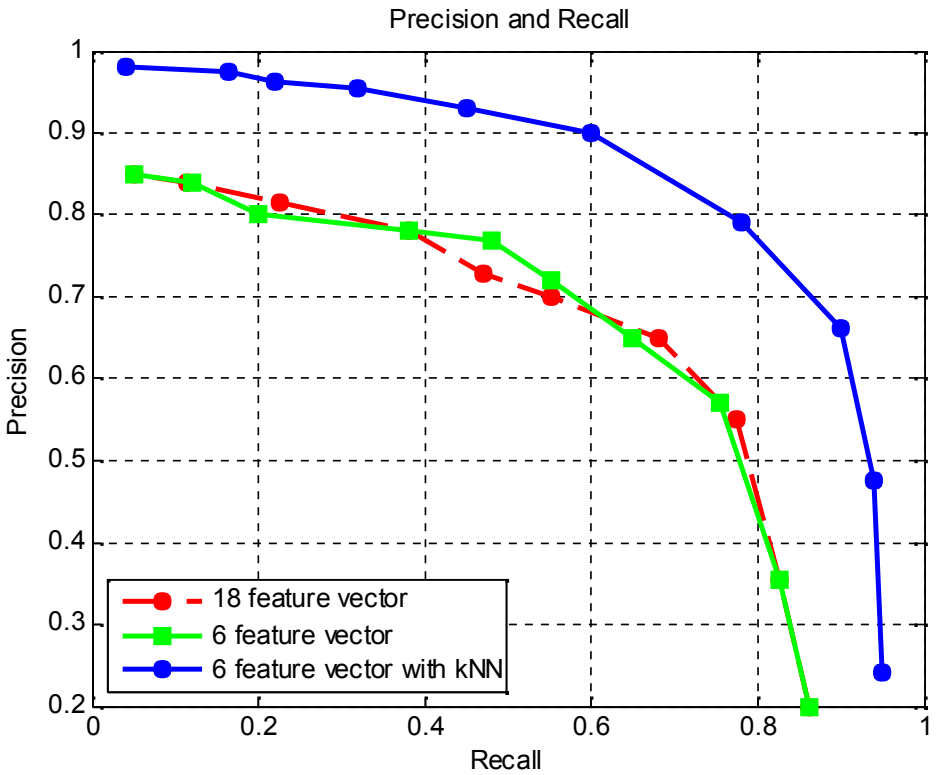

Figure 4:Performance of the proposed method in retrieving document images.

It can be found out that the overall performance of system after features selection using only six features is the same as the performance of system using all eighteen features, but with significant decrease in computation time. These results strongly indicate that the problem of feature selection is of primary importance for machine learning tasks. It seems especially important in such practical problems where one searches for classification models in large data sets with many feature without filtering them in preprocessing phase. A good choice of features may not only improve predictive accuracy but also results in smaller models which are easier to understand and interpret. On the other hand, it should notice that the

better of forward or backward results should not be taken as the best performance possible from the wrapper model. More comprehensive search strategies could examine a larger part of the search space and might even yield improved performance.

Also it can be observed from figure 4 that when using KNN algorithm the overall performance of system is outstandingly improved over the best performance giving an average precision value above 88.87%.

6. Conclusions

In this paper, retrieval of incomplete/distorted document image queries using texture analysis is addressed. Experiments were conducted on 600 document images database. It is found that the proposed features perform well for retrieving document images with different types and different document variations giving precision of 70.42%. The result shows that retrieval accuracy is highly increased by using KNN classifier giving precision of 88.87%. Experiments show that the proposed DIR system is efficient for retrieving document images with different types, different document variations and noises caused by scanning and printing.

References

[1] H. Liu, S. Q. Feng, H. B. Zha, and X. P. Liu, “Document image retrieval based on density distribution feature and key block feature”, in Proc. 8th ICDAR, pp. 1040-1044, 2005.

[2] GaofengMeng, Nanning Zheng, Yonghong Song, Yuanlin Zhang “Document images retrieval based on multiple features combination”, in Proc. 9th ICDAR, pp. 1040-1044, 2007.

[3] John F. Cullen, Jonathan J. Hull and Peter E. Hart, “Document image database retrieval and browsing using texture analysis”, IEEE 1997.

[4] M-W Lin, J-R Tapamo, B Ndovie, “A texture-based method for document segmentation and classification”, in ARIMA/SACJ, No. 36., 2006.

[5] Bikesh Kumar Singh, A. S. Thoke, KeshriVerma and AnkitaChandrakar, “Image information retrieval from incomplete queries using color and shape features”, Signal & Image Processing : An International Journal (SIPIJ) Vol.2, No.4, December 2011.

[6] Abu Sayeed Md. Sohail, Md. MahmudurRahman, Prabir Bhattacharya, Srinivasan Krishnamurthy, Sudhir P. Mudur, “Retrieval and classification of ultrasound images of ovarian cysts combining texture features and histogram moments”, IEEE, 2010.

[7] Fazal Malik and BaharumBaharudin, “Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain”, Production and hosting by Elsevier B.V. on behalf of King Saud University, 2013.

[8] Haralick, R.M., Shanmugan, K., and Dinstein, I. “Textural features for image classification”. IEEE Transactions on Systems, Man, and Cybernetics, Vol. 3, No. 6, pp. 610–621, 1973.

[9] Leen-KiatSoh, and Costas Tsatsoulis, “Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices”, IEEE transactions on geoscience and remote sensing, Vol. 37, No. 2, March 1999.

[10] David A. Clausi, “An analysis of co-occurrence texture statistics as a function of grey level quantization”, in Can. J. Remote Sensing, Vol. 28, No. 1, pp. 45–62, 2002.

[11] P. Pudil, J. Novovicova, J. Kittler, “Floating search methods in feature selection”, Pattern Rec. Letters 15 (1994) 1119–1125.

[12] L.Ladha, and T.Deepa, “Feature selection methods and algorithms”, in International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 5 May 2011.

[13] Dasarathy, B., Notions, N., Sheela, B., and Bangalore, I., “Visiting nearest neighbors,”Nearestneighbor(NN) norms: nn pattern classification techniques1, 434 (1991).

إسترجاع صور الوثائق من الاستفسارات الغير كاملة بأستعمال خصائص القوام النسيجي

أ.م.د.مثيل عماد الدين* (طالب دكتوراة) محمد طالب هاشم*

المستخلص

يعتبر إسترجاع صور الوثائق جزء مهم في الكثير من أنظمة معالجة صور الوثائق مثل أنظمة المكاتب الخالية من الورق، المكتبات الرقمية وغيرها. يساعد المستخدمين في ايجاد صور الوثائق الأكثر تشابه من قاعدة بيانات صور الوثائق. أغلب البحوث نُفذت بتقديم إستفسارات كاملة موجودة في قاعدة البيانات، لكن في العديد من الحالات يمكن مواجهة صور مشوهة أو ناقصة. هذا التشويه أو النقص يكون بسبب بعض المعلومات المفقودة، بعض الأجسام الغير مرغوبة، التشويه، الضوضاء نتيجة لعملية الطباعة، المسح الضوئي للوثيقة، ...الخ. يصف هذا البحث طريقة لإسترجاع الوثيقة الناقصة والمشوهة مستندة على الخصائص البصرية بأستعمال معلومات القوام النسيجي للإسترجاع من قاعدة بيانات كبيرة لصور الوثائق. تم اقترح إستعمال خصائص مصفوفة الـ Co-occurrence لتحليل القوام النسيجي وأوردنا تقييم اختباري شامل.

*قسم علوم الحاسوب / الجامعة التكنولوجية