Interactive Content-Based Document Retrieval Using Fuzzy Attributed Relational Graph Matching

Ramzi Chaieb, Karim Kalti, Najoua Essoukri Ben Amara

Interactive Content-Based Document Retrieval Using Fuzzy Attributed Relational Graph Matching

Ramzi CHAIEB
Engineering School of Tunis
SAGE Research Unit, ENISo
University of Sousse
Sousse, Tunisia
ramzi.chaieb@hotmail.com

Karim KALTI
Faculty of Science of Monastir
SAGE Research Unit, ENISo
University of Sousse
Sousse, Tunisia
karim.kalti@gmail.com

Najoua ESSOUKRI BEN AMARA
Engineering School of Sousse (ENISo)
SAGE Research Unit, ENISo
University of Sousse
Sousse, Tunisia
najoua.benamara@eniso.rnu.tn

***Abstract*—The evolution of digital technology and the desire to maintain an easy access to scanned documents implies the interest of Document Image Retrieval (DIR). In this paper, we propose a fuzzy graph-based document retrieval approach for determining similarity between document images for content-based document retrieval. To model the structure of document images, we have opted to use Attributed Relational Graphs (ARGs). For each document region (text, graphic, etc) we associate a node which is characterized by a set of fuzzy membership degrees reflecting low-level properties (texture, shape, color, etc). We use this fuzzy description in order to guarantee more robustness against the eventual segmentation errors which may be occurred after the segmentation of document regions. ARGs edges represent spatial relationships between regions which have common boundaries. Finally, we have developed a tree-search based optimal matching algorithm, which allows the search for document according to its structure. The database used for experiments is composed of segmented Coran document images from the National Library of Tunisia. Different weights are assigned for regions and edges according to their relative importance. The results obtained demonstrate the effectiveness of fuzzy graph-based document retrieval. Such an approach is very useful for several applications in many fields.**

Keywords—Fuzzy Attributed Relational Graph indexing; Graph matching distance; Document image retrieval; User interaction

I. INTRODUCTION

In recent years, with the rapid development of digital devices for image creation, storage and transmission, huge amounts of images are produced every day (for example in administrations, museums, etc). The huge amount of scanned documents stored on the internet is growing rapidly. Manual organization of documents is too costly and sometimes even impossible. Automated content-based document management methods, generally known as DIR techniques, are needed to deal with these amounts of data.

Motivated by the enormous demand of information accessing in document image databases, more and more research efforts have been devoted to DIR in recent years [1, 2, 3]. In DIR systems, the user specifies a query either in the form of image, text or sketch and the system is expected to return the items that are visually and semantically most similar to the query. Traditional approaches for DIR are text-based [2, 6, 7, 8]. A key task of text-based approaches is to annotate and index the images with textual descriptions or keywords for retrieving images [2, 6, 7, 8]. However, many limitations of the text-based approaches make them still far from working in real-world applications. Describing the rich content of a document image by only a few keywords is almost difficult. In addition, text descriptions reflect the subjectivity of the annotator and the annotation process is prone to be incomplete, ambiguous and very difficult to be automated. Hence, fully text-based approach is not practical enough for DIR applications. This has created a demand for effective and flexible techniques for automatic DIR to overcome the shortcomings of the text-based retrieval mechanism. Content-Based Document Retrieval (CBDR) has been suggested as an alternative approach for information retrieval to find document images which are visually most similar to a given query from a document image database. At present, CBDR has become an extremely active area of research in the visual information retrieval [4, 5]. In CBDR, the document images in the database are indexed by extracted visual features (such as color, shape, texture, etc.). Based on the indexing scheme with visual descriptions of low-level features, visual queries can be formulated and similarity of document images can be measured by employing a distance function defined in the CBDR systems [1, 2, 3].

A document image is generally composed of several components such as logo, text, figure, table, signature, etc. Usually, document images belonging to the same type have similar layout structures. For example, a letter is generally composed of a letter head, text area and a signature. An article contains columns of text blocks, figures, tables, etc. Traditional CBDR approaches are not powerful enough to capture these high-level concepts because of the semantic gap between high-level concepts and low-level features. A segmentation phase is usually necessary to separate adjacent regions. However, segmentation errors may occur frequently. From the extracted regions, a structural signature is needed to define document image regions and their topological relationships. Graph based approaches usually provide a rich and holistic description of the layout and content of the analyzed document images. In the literature, a lot of research has been done on graph-based representation and graph matching algorithms [6, 8, 9, 10].

In this study, we propose a fuzzy graph-based document image representation in order to reduce the problem of possible

978-1-4799-1805-8/15/$31.00 ©2015 IEEE

segmentation errors. Each segmented region is characterized by a set of fuzzy membership degrees derived from low-level features (texture, shape, color, etc). For example, a segmented region may be a 80% text and 20% graphic. In addition, a similar set of fuzzy membership degrees is used to specify the spatial relationship (above, below, right, left, surrounded-by, etc) between two adjacent segmented regions. Another contribution of this work is the human's interaction in order to surmount the limited capacity of traditional CBDR systems. User's weights are involved in the retrieval process.

The remaining of this paper is organized as follows. The next section gives an overview of the proposed approach. Section 3 presents the document FARG representation. Sections 4 describes the process of FARG matching algorithm. Experimental results are evaluated and discussed in section 5. Finally, concluding remarks are given in the last section.

II. Methodology

The proposed CBDR approach is composed of two main phases: indexing and retrieval phases (Fig. 1).

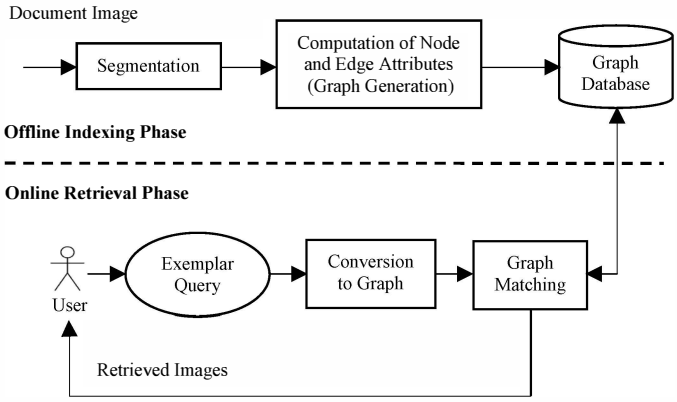

Fig. 1. Block diagram of the proposed CBDR approach

A. Offline Indexing Phase

The used document images are indexed into a database. Each document image is represented by a FARG that describes the document image as a set of regions with relationships between them. Both the nodes and the edges are represented with their attributes.

B. Online Retrieval Phase

Once a user specifies an example document image, the proposed query is then represented by a FARG. The retrieval task is to find the closest matches of the query FARG representing the structure of a document image components to a database of FARGs already generated. A similarity measure is used in the comparison of two document images. The obtained costs of graph matching are used to sort the database document images in a decreasing order of similarity. Finally, the obtained results are ranked as the final retrieval result and returned to the user.

III. Fuzzy Attributed Relational Graph Representation

We have used FARGs as a structural representation of document layout. One motivation for the FARG representation is to reduce the effect of inaccurate segmentation on retrieval. FARGs have an advantage over vector-based features as they can model the spatial relationship between different regions [9, 10].

A. Attributed Relational Graph

An ARG is a graph whose nodes and edges are both represented with attributes. It can be represented as a 3-tuple $(N; E; A)$ where N is the set of nodes (vertices), $E \subseteq N \times N$ is the set of edges (relations) and A is the set of attributes. Each node (n_i) attributes are defined as a vector $x_i=[x_i^{(u)}]$, $(u=1,2,3,...,U)$, where U is the total number of node attributes associated with node n_i. Similarly, for each edge (e_j) a set of attributes are represented as a vector $y_j=[y_j^{(v)}]$, $(v=1,2,3,...,V)$, where V is the total number of edge attributes associated with edge e_j. An example of an ARG is shown in Fig. 2.

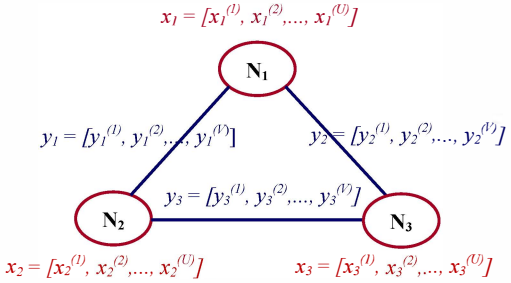

Fig. 2. An example of an ARG with node attribute vectors $\{x_i, i = 1, 2, 3\}$ and edge attribute vectors $\{y_j, j = 1, 2, 3\}$

B. Regional Adjacency Graph

Regional Adjacency Graph (RAG) is an ARG whose nodes represent document image regions and edges represent spatial relationships between regions which have common boundaries. Each node has a unique identifier number. Node attributes represent informations about the region such as label, type of the region (which may be text, graphic, etc), area, perimeter, shape, color, etc. Edge attributes specify relationships between two regions like the distance between their centroids, common boundary length, orientation which may be above, below, right, left, surrounded-by, etc. An example of a RAG is shown in Fig.3.

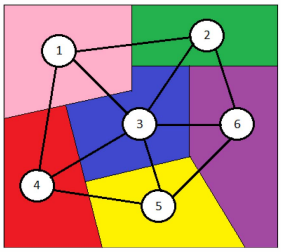

Fig. 3. An example of a RAG generation, regions are represented by nodes

C. Fuzzy Attributed Relational Graph

One advantage of FARG is the representation of document images with a fuzzy approach, which is similar to human perception [10, 11]. Fuzzy logic approaches are proposed to map from the low-level features to high-level concepts in order to bridge the semantic gap between them. Document image regions often possess properties that are fuzzy in nature and it has been extended to include fuzzy information into the attributes.

In a FARG, each vertex may take attributes from the set $Z=\{z_i \mid i = 1, 2, ..., I\}$. For each attribute z_i, it will take values from $S_i = \{s_{ij} \mid j = 1, 2, ..., J_i\}$. The set of all possible fuzzy attribute-value pairs is $L_v = \{(z_i , A_{Si}) \mid i = 1 , ..., I\}$, where A_{Si} is a fuzzy set on the attribute-value set S_i. A valid pattern primitive is just a subset of L_v in which each attribute appears only once. The set of all those valid pattern primitives is denoted as Π. Each vertex will be represented by an element of Π.

Similarly, each arc may take attributes from the set $F=\{f_i|i=1, 2, ..., I\}$ in which each f_i may take values from $T_i=\{t_{ij}| j = 1, 2, ..., J_i\}$ and $L_a = \{(f_i , B_{Ti}) \mid i = 1 , ..., I\}$ denotes the set of all possible relational attribute value pairs, where B_{Ti} is a fuzzy set on the relational attribute value set T_i. A valid relation is just a subset of L_a in which each attribute appears only once. The set of all those valid relations is denoted as Φ.

A Fuzzy Attributed Graph (FAG) over $L=(L_v,L_a)$, with an underlying graph structure $H=(N,E)$ is defined to be an ordered pair (V,A), where $V=(N,\sigma)$ is called a fuzzy vertex set and $A=(E,\delta)$ is called a fuzzy arc set. The mapping $\sigma: N \rightarrow \Pi$ and $\delta:E \rightarrow \Phi$ are called fuzzy vertex interpreter and fuzzy arc interpreter, respectively. This definition also applies when there are non-fuzzy attributes, since a crisp (non-fuzzy) set can always be represented as a special case of a fuzzy set.

IV. Fuzzy Attributed Relational Graph Matching Algorithm

In this section, a fuzzy graph matching algorithm is used to calculate the graph matching distance between a query document image and a document image in the database. Given a query document image, its FARG is matched to each FARG in the database. A similarity measure between matched FARGs is assigned to each pair of FARGs.

Let $G(N,E,\alpha,\beta)$ and $G'(N',E',\alpha',\beta')$ be respectively a query FARG and a target FARG. For each node n_i $(i \in 1, 2, ..., N)$ in G, the set of nodes membership degrees is given as α_i:

$$\alpha_i = \alpha_i^{(1)}, \alpha_i^{(2)}, \dots, \alpha_i^{(a)}, a \in \{1, 2, \dots, A\} \quad (1)$$

where N is the number of nodes in G and A is the total number of nodes membership degrees associated with each node n_i. We denote by α the set of all vectors α_i.

$$\alpha = \alpha_1, \alpha_2, \alpha_3, \dots, \alpha_N \quad (2)$$

Nodes membership degrees indicate how much a region belongs to a graphic zone, text zone, etc. Different weights are assigned for each node features according to their relative importance to describe regions. If we denote by w_{n_i} the weight of node n_i, the query image can be represented as a weight vector W_N in the weight space by:

$$W_N = \left[w_{n_1}, w_{n_2}, \dots, w_{n_N}\right], i \in \{1,2,\dots,N\} \; and \sum_{i=1}^{N} w_{n_i} = 1 \quad (3)$$

Node n_i with the highest value of W_N is considered the most important node during the matching of nodes.

Similarly, for each edge e_j $(j \in 1, 2, ..., E)$ in G, the set of edges membership degrees is given as β_j:

$$\beta_j = \beta_j^{(1)}, \beta_j^{(2)}, \dots, \beta_j^{(b)}, b \in \{1, 2, \dots, B\} \quad (4)$$

where E is the number of edges in G and B is the total number of edges membership degrees associated with each edge e_j. We denote by β the set of all vectors β_j.

$$\beta = \beta_1, \beta_2, \beta_3, \dots, \beta_E \quad (5)$$

Edges membership degrees specifies the spatial relationship between two nodes (above, below, right, left, surrounded-by, etc). If we denote by w_{e_i} the weight of edge e_i, the query image can be represented as a weight vector W_E in the weight space by:

$$W_E = \left[w_{e_1}, w_{e_2}, \dots, w_{e_E}\right], i \in \{1,2,\dots,E\} \; and \sum_{i=1}^{E} w_{e_i} = 1 \quad (6)$$

Edge e_i with the highest value of W_E is considered the most important edge during the matching of edges.

The graph matching is implemented through a tree-based search. Assume that G an G' with order m and n $(m \leq n)$, respectively. In the special case where $m=n$, the problem is to find the optimal isomorphism between G and G'. First, the decision tree is constructed. It has height m, and n-p sons for each node at level $p = 0, 1,..., m$. Therefore, at any level $p>0$, there is a path that consists of ordinal nodes in the decision tree from the root to node N. The first level of the tree contains all possible combinations of matched nodes. A matching between a query node and a target node is permissible only if they have the same nature (text, graphic, etc). Then, the cost of each node pair is computed using weighted Euclidean distance (Equation 7).

$$d_n(q_n, s_n) = \sqrt{\sum_{i}^{T_n} w_{n_i} \times (q_{n_i} - s_{n_i})^2} \quad (7)$$

where q_n is a node of G, s_n is a node of G', T_n is the number of node membership degrees, q_{n_i} is a membership degree of the query node, s_{n_i} is a membership degree of the mapped node. In order to increase the accuracy of our approach, we introduced weighting w_{n_i} to our similarity measurement (w_{n_i} is the importance weight for a node n_i). Node pair (q_i,s_i) with the lowest cost is expanded.

$$(q_i, s_i) = arg\,min\, d_n(q_i , s_i) \quad (8)$$

The cost of each edge pair is computed using weighted Euclidean distance (Equation 9).

$$d_e(q_e, s_e) = \sqrt{\sum_{i}^{T_e} w_{e_i} \times (q_{e_i} - s_{e_i})^2} \quad (9)$$

where q_e is an edge of G, s_e is an edge of G', T_e is the number of edge membership degrees, q_{e_i} is a membership degree of the query edge, s_{e_i} is a membership degree of the mapped edge. Similarly to node weighting, we introduced weighting w_{e_i} to our similarity measurement (w_{e_i} is the importance weight for an edge e_i).

The tree is extended until path with least total cost giving optimal mapping is found. An example of the tree-based search approach is given in Fig. 4.

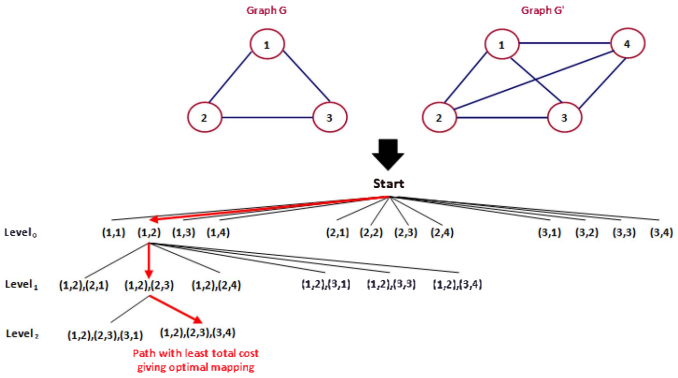

Fig. 4. An example of tree-based matching process

The total node and edge distances are computed using equations 10 and 11, respectively.

$$\mathrm{TND} = \sum_{i=1}^{N} d_{n_i}(q_n, s_n) \quad (10) \qquad \mathrm{TED} = \sum_{i=1}^{W} d_{e_i}(q_e, s_e) \quad (11)$$

where N and W are the number of matched node pairs and the number of matched edge pairs, respectively.

The total similarity measure between a FARG pair can be defined by the formula 12:

$$SIM(G, G') = w_G \times TND + w_{G'} \times TED \quad (12)$$

where w_G and $w_{G'}$ are properly selected weights of G and G'.

V. Experimental Results

Experiments were conducted on a database composed of 200 initially segmented Coran document images from the National Library of Tunisia [12]. This database is created in collaboration with the National Library of Tunisia [12]. The segmentation method is based on the exploitation of connected components and morphological operators. It allows the extraction of three types of regions (graphic, text and frame). The database is divided into 5 classes. Each class contains 40 Coran document images with similar structures and contents. Variability between classes is caused by differences between the regions' specific characteristics and their positions in the Coran document images. Five examples from the five different classes are shown in Fig. 5. Two sample images from the used database and their segmented regions are shown in Fig. 6.

(a) (b) (c) (d) (e)

Fig. 5. Five examples from the five used classes [12]: (a) containing only text regions, (b) containing text regions surrounded by a frame, (c) containing graphic regions followed by text regions, (d) containing text, graphic and frame regions, (e) containing only graphic regions

In the retrieval phase, firstly user assigns an importance degree to each region by giving an appropriate value between 0 and 1 for each region. Region with the highest value of importance degree is considered the most important region during the matching of regions. Similarly, user assigns an importance degree to each relationship. So, the search will rather focuses the most important relationships selected by the user. Assigning importance degrees provides a more precise formulation of user's queries. Figure 7 shows an example of the used importance degrees related to the nodes and edges of the query graph G. In this example, region 1 is considered the most important region in the query document image. Each image is represented by a FARG. Then, images whose FARGs are closest in distance to a given query FARG are selected and the images belonging to each selected FARG are retrieved and displayed to user.

In these experiments, each region i is described by two membership degrees $\alpha_i^{(1)}$ and $\alpha_i^{(2)}$ which corresponds respectively to the percentage of each region to be a graphic or a text. Similarly, each relationship j between adjacent regions is determined using two metrics $\beta_j^{(1)}$ and $\beta_j^{(2)}$ which corresponds respectively to the distance and the horizontal angle between the two centroids of adjacent regions.

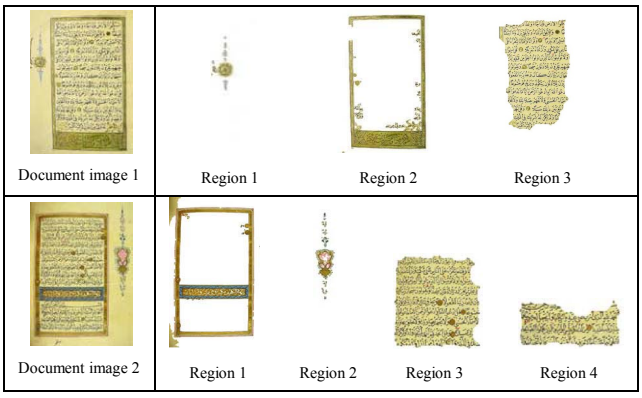

Fig. 6. Two sample images from the database used in experiments [12]

As shown in Fig. 7, Document image 1 and Document image 2 are represented by two graphs G and G', respectively. Their tree-based matching process is illustrated in the same figure. The total similarity measure between the two graphs G and G' is computed as follows:

$$TND = d_{n5}\,(of\ Level\ 0) + d_{n5}\,(of\ Level\ 1) + d_{n1}\,(of\ Level\ 2)$$
$$= 0.0077 + 0.0219 + 0.0379 = 0.0675$$
$$TED = d_{e1} + d_{e2} = 1.8828 + 1.0678 = 2.9506$$
$$SIM(G,G') = w_G \times TND + w_{G'} \times TED$$
$$= 0.7 \times 0.0675 + 0.3 \times 2.9506 = 0.9324$$

Precision and recall are the most popular measures used for evaluation of Content-Based Image Retrieval system performance [8, 9]. Precision and Recall are calculated only for one query. To evaluate a system for several queries, we use the mean precision and the mean recall. Examples of recorded Precision and Recall values are given in table 1. The obtained results are satisfactory. They are related to the number and the nature of regions (text, graphic or frame) and their relative positions in each Coran document image.

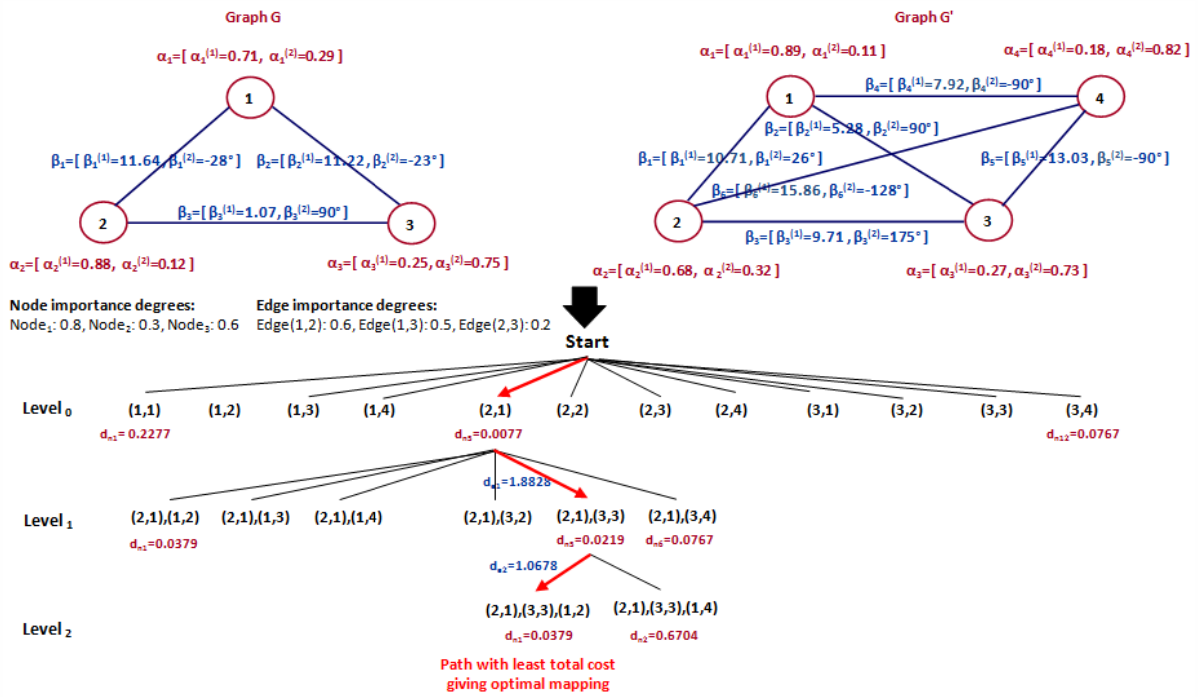

Fig. 7. An example of two FARGs (G and G') matching which corresponds to "Document image 1" and "Document image 2", respectively

TABLE I. OBTAINED RESULTS (PRECISION, RECALL AND MATCHING TIME)

Query Number	Relevant Retrieved Images	Retrieved Images	Relevant Images	Precision (%)	Recall (%)	Matching Time (s)
1	26	34	40	78.47	65	0.0108
2	25	32	40	78.12	62.5	0.0134
3	29	38	40	76.31	72.5	0.0174
4	26	35	40	74.28	65	0.0117
5	27	37	40	72.97	67.5	0.0142
6	27	38	40	71.05	67.5	0.0157
7	29	41	40	70.73	72.5	0.0153
8	31	44	40	70.45	77.5	0.0143
9	33	48	40	68.75	82.5	0.0162
10	34	51	40	66.66	85	0.0118
Average				72.77	71.75	0.0141

VI. CONCLUSION

In this paper, we have proposed a CBDR approach based on FARG representation to address the problem of matching between document images. FARG representation has helped in reducing the effect of inaccurate segmentation on retrieval. The proposed approach is composed of two main modules that work in pipeline mode. The first module consists of query construction and FARGs generation. The second module consists of FARG matching approach. A tree-based approach has been developed to calculate the matching cost between two FARGs. Finally, best matching results are retrieved and displayed to the user. The experimental results are promising. They can be improved, as future work, using relevance feedback techniques in order to learn from the user's evaluation.

Besides, the proposed approach can be generalized to many types of document images such as administrative documents, journals, books, etc.

REFERENCES

[1] L. Ying, Z. Dengsheng, L. Guojun and M. Wei-Ying, "A survey of content-based image retrieval with high-level semantics", Pattern Recognition 40, pp. 262 – 282, 2009.

[2] S. Pattanaik and D.G. Bhalke, "Beginners to Content Based Image Retrieval", International Journal of Scientific Research Engineering & Technology (IJSRET), pp. 040-044, 2012.

[3] S. Deb and Y. Zhang, "An Overview of Content-based Image Retrieval Techniques", IEEE, 18th International Conference on Advanced Information Networking and Application (AINA'04), 2004.

[4] L. Zhao and J. Tang, "Content-Based Image Retrieval Using Optimal Feature Combination and Relevance Feedback", IEEE, International Conference on Computer Application and System Modeling (ICCASM'2010), 2010.

[5] N. Singhai and S. K. Shandilya, "A Survey On: Content Based Image Retrieval Systems". International Journal of Computer Applications, pp. 0975 – 8887, 2010.

[6] A. Sameriya and B. Sharma, "A Survey on Content Based Image Retrieval (CBIR) Schemes". International Journal of Engineering Sciences & Management (IJESM), pp. 75-82, 2014.

[7] G. F. Ahmed and R. Barskar, "A Study on Different Image Retrieval Techniques in Image Processing". International Journal of Soft Computing and Engineering (IJSCE), pp. 2231-2307, 2011.

[8] R. Datta, D. Joshi, J. Li and J. Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age". ACM Transactions on Computing Surveys, 2008.

[9] Sharma *et al.*, "Determining similarity in histological images using graph-theoretic description and matching methods for content-based image retrieval in medical diagnostics". Diagnostic Pathology, 2012.

[10] S. Philipp-Foliguet, J. Gony and PH. Gosselin, "FReBIR: An image retrieval system based on fuzzy region matching". Computer Vision and Image Understanding, pp. 693-707, 2009.

[11] V. Aiswarya, T. Senthil Kumar, "Survey on Content Based Image Retrieval Techniques". IJRET: International Journal of Research in Engineering and Technology, 2014.

[12] National Library of Tunisia (web site: http://www.bibliotheque.nat.tn).