# A content-based digital mammography retrieval using inexact graph matching

Fradj Ben Lamine
MARS Research Group
University of Monastir
Tunisia
benlamine.fradj@gmail.com

Karim Kalti
Computer Sciences department
Faculty of sciences of Monastir
University of Monastir
Tunisia
Karim.kalti@gmail.com

Lotfi Ben Romdhane
MARS Research Group
ISITcom of Hammam Sousse
University of Sousse
Tunisia
lotfi.ben.romdhane@gmail.com

***Abstract*—Content-Based Image Retrieval (CBIR) is becoming one of the most vivid research area in computer vision. It is widely used in medical applications especially in computer aided diagnostic systems (CAD). CBIR systems in digital mammography take an important part of these works. The work presented in this paper aims to propose a CBIR approach based on inexact graph matching algorithm for mammographic images. To achieve this task, we represent a mammogram as an Attributed Relational Graph (ARG) based on ImageMap approach where each node of the graph represents a semantic object. Objects that are considered in mammogram are: Background, Breast, Pectoral Muscle, Masses and Calcifications. Then, for each node, we compute a signature that describes the selected object. In order to retrieve the most similar images to a query one, a graph matching technique is applied based on the Hungarian algorithm. To Evaluate our approach 100 mammographic images from the MIAS database were used and six metrics were computed. Experiments demonstrate that the proposed method using Hamming distance gives the most promising results.**

## I. Introduction

Computer Aided Diagnosis systems (CADx) are nowadays used in hospital to give a second opinion to radiologists. Content based image retrieval systems (CBIR) are included in CADx to retrieve the most similar cases to a query one. This vivid area takes a lot of importance in computer vision researches. In fact, big quantities of images are produced each day in hospital. Archiving and retrieval in this big volume of data require an appropriate indexing method. Content based image retrieval is composed of two main steps: indexing and retrieval. Text-based approaches are the first used in literature for indexing image databases. But, these approaches became unrealizable in large database. The necessity of an automated approach to describe the content of images has emerged. In such approaches, several features are computed to describe the image content, like shape, intensity and texture. These features cannot correctly describe the semantic content of the image. This problem is known in literature as the semantic gap. The semantic gap is the distance between the interpretation of the user and features computed to describe the image content. To deal with the semantic gap, researchers use pattern recognition modeling techniques. In this work we adopt one of these techniques which is attributed relational graphs (ARGs) where semantic content is described by nodes and edges labels. Images are then modeled as ARGs and the problem of content retrieval becomes a graph matching problem.

This paper is organized as follows: the second section presents the most related works that deals with CBIR in mammography. It focuses particularly on systems using inexact graph matching. The third section introduces our approach with its detailed steps. The fourth section presents experimental results. The paper ends by a conclusion and by highlighting future works.

## II. Related works

Image indexing with bag of words became a hard task due to the big quantities of image produced each day. This task takes a lot of time if we need to describe correctly the diversity of image content. In the last decade, content based image retrieval approaches became automated. These approaches use low level features to describe visual image content. In this case image is indexed by a vector of features. Each component in this vector corresponds to a texture feature, a shape feature or an intensity feature. CBIR approaches can be classified into two categories depending on the type of the query. The query can be a region of interest (ROI) or the entire image.
When the query is a ROI, a vector of low level features is used to describe its content. In digital mammography, most of related works are of this type. De Oliveira and al. developed a content based image retrieval called MammoSVD [1]. This approach is based on breast density. El naqa and al. developed many CBIR approach for mammogram retrieval [2], [3], [4]. Konishita and al. developed a content based retrieval of mammograms using visual features related to breast density patterns [5]. Mendes and al. developed a content based mammography images retrieval using ripleys k function[6]Honda and al developed a content-based image retrieval in mammography: using texture features for correlation with BI-RADS categories. Bin Zheng gives a detailed Computer-Aided Diagnosis in Mammography Using Content-Based Image Retrieval Approaches. He discusses the current status and gives future perspectives [24]. Severin and al. developed a Content based mammogram retrieval using Gray Level Aura Matrix [25]. Other works are developed in literature [8], [9], [10], [11], [18]. The majority of these approaches uses a ROI in retrieval process. For the second category, the totality of the content of the image is indexed. The disadvantage of the use of global features, in this case, is their inability to describe in details the content of image in terms of objects that are present in. Thus, this kind of approaches are rarely used for databases containing images representing the same scene. The adequate approach to

deal with this kind of databases proceeds by the segmentation of the image into regions. The result of segmentation is then represented by an attributed relational graph. In ARG, nodes represent the region and edges represent the spatial relationship that existing between them.
Using this representation model, the retrieval step is transformed into a graph matching problem. Graph matching consists in comparing two graphs in order to find the best correspondence between their nodes and edges. The matching can be exact or inexact. Exact graph matching algorithms give a strict correspondence between nodes and edges. In inexact graph matching algorithms, a tolerance is allowed. In the case of medical images, the exact graph matching is not appropriate because these images are not entirely similar. Thus, the majority of CBIR approaches involves the use of inexact graph matching.
In literature, several CBIR based on graph matching are developed. FReBIR [14] is an image retrieval system based on fuzzy region matching. Each image is represented by ARG and an inexact graph matching algorithm is applied in the retrieval step. In [15], a CBIR by matching hierarchical attributed region adjacency graphs is proposed. Authors apply their approach in medical context. Similarity flooding approach and Hopfield neural network are used in the graph matching step. The relationship between adjacency regions is taken into account in this work. The disadvantage of this approach is the presence of multiple scale of graphs for one image. The performance of this approach depends on the choice of the graph's scale. In [22], a graph-based approach to the retrieval of dual-modality biomedical images using spatial relationships is developed by Ashnil Kumar and al. A novel method of querying by using both the functional and anatomical information and their structural relationships is used. An inexact graph matching algorithm is applied to ARGs. Petrakis and al. [13] developed a series of works in CBIR using inexact graph matching. All of them take into account the relationship between regions of interest. The ARG representation is used in these works. [23] proposes an image retrieval via probabilistic hypergraph ranking. Images are taken as vertices in a weighted hypergraph. The disadvantage of this approach is the absence of relationship between regions of interest in each image. Broumandnia and al. [19] developed a CBIR with graph theoretic approach. An image is presented by a set of regions and then each image is represented by a graph. Matrix matching is proposed to determine the best matching by maximizing the likelihood probability. In [17], a CBIR with relevance feedback based on graph theoretic region correspondence estimation is proposed. An inexact graph matching is applied to find the most similar cases. In breast cancer domain, graph matching is applied in [16]. This work is the first approach in breast cancer using biopsies images.The main goal of the authors of this approach is to determine the histological similarity between images using graph based representation for each image. Then, a graph matching method based on A* algorithm is applied to get the most similar images.

## III. The proposed approach

Content based image retrieval is a lively domain and an extremely active area in digital mammography. Its novelty resides in the integration of pattern recognition in the retrieval process. The graph based representation takes the first key of these works. In digital mammography, the majority of the approaches developed, till now, are based on distance between vectors that describe images.
This work is in his preliminary phases. Our approach belongs to the category where the query is an entire image. In this case, the image can be described by global features or local features. When the database is specific, generally all images represent the same scene. In this case, describing the image using global features is not appropriate because the global characteristics of the database images are almost the same. Thus, image characterization must be then done locally inside the image. For that, we adopt to the representation of the image using a graph. Each node of the graph describes an object or a region that may be present in the image. Edges represent the spatial relationship existing between them. Two types of spatial relationship are considered: "contains" or "outside". To obtain the graph, we proceed by using the annotated database describing image objects. Objects that we consider in our approach are: Background, Breast, Pectoral Muscle, Masses and Calcifications.
The annotation step is done using prior information that

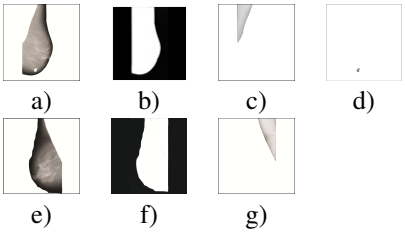


Fig. 1. Annotated database mdb080.pgm and mdb047.pgm:a) Breast of mdb080, b) Background of mdb080, c) Pectoral muscle of mdb080, d) Masse of mdb080, e) Breast of mdb047, f) Background of mdb047, g) Pectoral muscle of mdb047

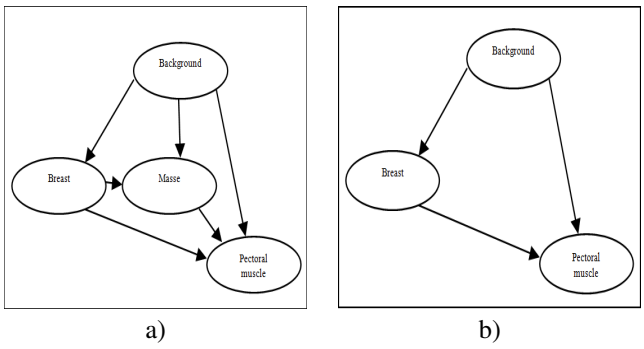


Fig. 2. Graph representation of the MIAS database images : a) graph of mdb080.pgm, graph of mdb047.pgm

describes the MIAS database. In this work, we don't use BIRADS description to annotate the mammogram. So, the density of mammogram is not taken into account. The construction of the graph is performed using the ImageMap method [13] where each object represents a node of the graph. The figure 1 shows the detailed steps of our approach. Each node of the graph is attributed. A set of 10 features is used to describe the characteristics of each object. These features are: Area ($S_1$), Perimeter ($S_2$), Euler Number ($S_3$), Orientation ($S_4$), Convex Area ($S_5$), Filled Area ($S_6$), Eccentricity ($S_7$), Solidity ($S_8$), a label giving the nature of the object($S_9$) and the degree of

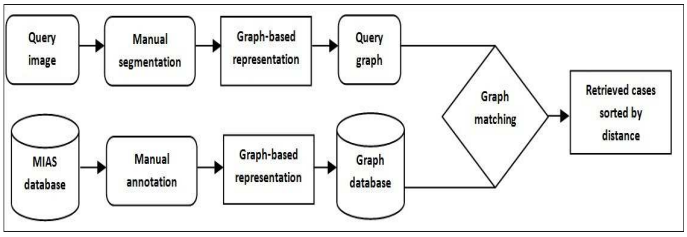


Fig. 3. The approach model

TABLE I. COST MATRIX

| | $V1_1$ | $V1_2$ | ... | $V1_n$ |
|---|---|---|---|---|
| $V2_1$ | M(1,1) | M(2,1) | ... | M(n,1) |
| $V2_2$ | M(1,2) | M(2,2) | ... | M(n,2) |
| | | | | |
| $V2_m$ | M(1,m) | M(2,m) | ... | M(n,m) |

each node ($S_{10}$). Thus, each node is described by a vector where each component is the value $S_i$. This vector is called the signature of the node which is formally described by:

$$Signature = \{S_i\} \quad i \in \{1..10\} \tag{1}$$

In order to compare two attributed graph G1 (V1, E1) and G2 (V2, E2) where $V1 = \{V1_1, V1_2, \ldots, V1_n\}$ and $V2 = \{V2_1, V2_2, \ldots, V2_m\}$ , we construct the cost matrix M where each element M(i, j) is the distance D between the signature of $V1_i$ and the signature of $V2_j$[12].

$$D(V1_i, V2_j) = D(Signature_{V1_i}, Signature_{V2_j}) \tag{2}$$

In this step, the use of an appropriate algorithm to find the cost minimizing assignment between nodes of the two graphs is necessary. The Hungarian algorithm solves the assignment problem in polynomial time. This algorithm applies modifications in raw and column values in order to optimize the assignment cost. Thus, the cost matrix M is the input for the Hungarian algorithm [20]. The output of this algorithm is the permutation matrix P that describes the matching between two sets of nodes V1 and V2. Each element of P takes the value of 1 if there is a matching between two nodes and 0 otherwise. The distance between the two graphs G1 and G2 is then computed as the sum of values of M(i, j) when P(i, j) is equal to 1.

$$distance(G1, G2) = \sum_{i=1}^{n} \sum_{j=1}^{m} M(i,j) * P(i,j) \tag{3}$$

In our approach , different distances are used between nodes signatures for experimental test. These distances are: the Hamming distance, the Jaccard distance, the Euclidean distance, the Cityblock distance, the Correlation distance, the Cosine distance and the Chebychev distance.

## IV. RESULTS

To evaluate our work, we have used 100 images from an annotated database. Each image in the database contains a description of the list of objects that it contains. 20 images are used as a queries image and 80 images for the constitution of the database. In our approach, six metrics are used to evaluate the performance of the proposed content based image retrieval approach: Accuracy, Sensitivity, Specificity, Precision, Recall and F−measure [21].

The preliminary results show that the Hamming distance has the best performance comparing to other distances. In fact this distance gives a precision of 0.6 and a recall of 1. Besides, the use of Jaccard distance has a nearly results to Hamming distance. The experimental results show also that when we use the correlation distance or the cosine distance, the most similar images are inverted i.e when the query image contain tumor, the most similar images are without tumor. These results are medium comparing to classical approaches that use global or/and local features to describe the images. But, our results are consistent in term of getting images containing tumor or not. Besides the classical approaches cannot give this result because the features used are symbolic and describe the image with statistical values. The weakness of the results of our approach is noticed when the query image contains tumor. In this case, precision takes the value of 20% when we use the euclidean distance. Besides, when the query image is without tumor, the average precision is near 60%. Comparing to other CBIR approaches using graph based representation in breast cancer, our approach takes into account the entire image however the work developed in [16] takes the portion of image. The disadvantage of [16] is that in the retrieval step, original images are divided into small size blocks. These blocks have different sizes in experimental step. The next table show the difference between our approach and the work developed by Sharma and al. Our approach can be optimized in future works using an adaptive inexact graph matching algorithm in the retrieval step. Therefore, other graph based representation may be selected to model digital mammography images.

## V. CONCLUSION

The goal of this work is the development of a content based digital mammography retrieval approach using inexact graph matching. Our approach uses annotated database. The considered annotations concern the following objects: breast region, pectoral muscle, background, masses and calcifications. Each mammogram is modeled by an annotated graph. The Hungarian algorithm is then used to get the most similar images to a query one. Different distances have been tested in this context. Experiments have shown that the most promising result is achieved by the use of the Hamming distance. Our work is still in progress. Many perspectives are being explored in the retrieval process including the consideration of the spatial distribution of the annotated objects and the integration of the medical records information.

## REFERENCES

[1] De Oliveira, J. E. E., A. P. B. Lopes, G. C. Chvez, A. de Albuquerque Arajo, T. M. Deserno (2009). Mammosvd: A content-based image retrieval system using a reference database of mammographies. In CBMS, pp. 14. IEEE.

[2] El-Naqa, I., Y. Yang, N. P. Galatsanos, R. M. Nishikawa,M. N. Wernick (2004). A similarity learning approach to content based image retrieval: application to digital mammography.

[3] El-Naqa, I., Y. Yang, N. P. Galatsanos, M. N.Wernick (2003). Relevance feedback based on incremental learning for mammogram retrieval. In ICIP (1), pp. 729732.

[4] El-Naqa, I., Y. Yang, M. N.Wernick, N. P. Galatsanos (2002). Content-based image retrieval for digital mammography. In ICIP (3), pp. 141144.

TABLE II. EXPERIMENTAL RESULTS

| | Distances | | | | | | |
|---|---|---|---|---|---|---|---|
| | Hamming | Jaccard | Euclidean | Cityblock | Correlation | Cosine | Chebychev |
| Accuracy | 0.6 | 0.5 | 0.2 | 0.25 | 0.1 | 0.1 | 0.3 |
| Sensitivity | 1 | 0.84 | 0.25 | 0.33 | 0.17 | 0.17 | 0.4167 |
| Specificity | 0 | 0 | 0.125 | 0.125 | 0 | 0 | 0.125 |
| Precision | 0.6 | 0.56 | 0.3 | 0.36 | 0.2 | 0.2 | 0.4167 |
| Recall | 1 | 0.84 | 0.25 | 0.34 | 0.17 | 0.17 | 0.4167 |
| F-mesure | 0.75 | 0.66 | 0.27 | 0.35 | 0.18 | 0.18 | 0.4167 |

TABLE III. COMPARISON BETWEEN OUR APPROACH AND [16]

| | Our approach | [16] |
|---|---|---|
| Segmentation | annotated database | 64*64-128*128-256*256-512*512 |
| Database | MIAS | Biopsies images |
| Node attributes | Area, Perimeter, Euler Number, Orientation, Convex Area,Filled Area, Eccentricity, Solidity,label | Area, Perimeter, label |
| Edge attributes | Distance between centroid, angle between objects | Distance between centroid, Common boundary length |
| Matching | Hungarian algorithm | A* algorithm |


[5] Kinoshita, S. K., P. M. de Azevedo Marques, R. R. Pereira, J. A. H. Rodrigues, R. M. Rangayyan (2007). Contentbased retrieval of mammograms using visual features related to breast density patterns. J. Digital Imaging 20(2), 172190.

[6] Mendes, L. S. K. A., A. C. de Paiva, C. de Souza Baptista, A. C. Silva (2009). Content based mammography images retrieval using ripleys k function.

[7] Marcelo Ossamu Honda, Paulo Mazzoncini de Azevedo Marques, Jos Antonio Hiesinger Rodrigues, Content-based image retrieval in mammography: using texture features for correlation with BI-RADS categories, IWDM 2002 6th International Workshop on Digital Mammography.

[8] De Oliveira, J. E. E. D., A. D. A. Arajo, T. M. Deserno. Iwssip 2010- 17th international conference on systems, signals and image processing Mammosyslesion: a content-based image retrieval system for mammographies.

[9] Qi, H., S. W. E. (1999). Content-based image retrieval in picture archiving and communications systems. J. Digital Imaging 12, 8183.

[10] Swett, H. A., P. G. Mutalik, V. P. Neklesa, L. Horvath, C. Lee, J. Richter, I. Tocino, P. R. Fisher (1998). Voice-activated retrieval of mammography reference images. J. Digital Imaging 11(2), 6573.

[11] Tourassi, G.D., V.-V. R. D. J. C. M. C. J. E. Floyd. Computerassisted detection of mammographic masses: a template matching scheme based on mutual information. Med Phys, 21232130.

[12] Jouili, S.; Mili, I. Tabbone, S. (2009), Attributed Graph Matching Using Local Descriptions., in Jacques Blanc-Talon; Wilfried Philips; Dan C. Popescu Paul Scheunders, ed., 'ACIVS' , Springer, , pp. 89-99.

[13] Petrakis, E. G. M.; Faloutsos, C. Lin, K.-I. (2002), ImageMap: An Image Indexing Method Based on Spatial Similarity, IEEE Trans. Knowl. Data Eng. 14 (5) , 979-987.

[14] Philipp-Foliguet, S.; Gony, J. and Gosselin, P. H. (2009), 'FReBIR: An image retrieval system based on fuzzy region matching.', Computer Vision and Image Understanding 113 (6) , 693-707 .

[15] Benedikt Fischer; Christian J. Thies Mark O. Guld and Thomas M. Lehmann,Content−based image retrieval by matching hierarchical attributed region adjacency graphs. SPIE,Medical Imaging 2004: Image Processing,pp.598-606.

[16] Harshita Sharma, Alexander Alekseychuk, Peter Leskovsky, and al. Determining similarity in histological images using graph-theoretic description and matching methods for content-based image retrieval in medical diagnostics Diagnostic Pathology, Vol. 04 October 2012, 134.

[17] Li, C.Y. Hsu, C.T. (2008), Image Retrieval With Relevance Feedback Based on Graph-Theoretic Region Correspondence Estimation, IEEE Transactions on Multimedia 10 (3) , 447-456 .

[18] C.-H. Wei, C.-T. Li, and Y. Li, Content-based retrieval for mammograms, Artificial Intelligence for Maximizing Content-Based Image Retrieval, 313-339,2008.

[19] Ali Broumandnia, Mostafa Cheraghi, Mohsen Azararjmand,5) Content-Based Image Retrieval with Graph Theoretic Approach ,International Journal of Recent Technology and Engineering (IJRTE),Volume-1, Issue-3, January 2013.

[20] Munkres J.,Algorithms for the assignment and transportation problems.J.Soc.Indust.Appl.Math.5(1975),32-38.

[21] David M. W. Powers, Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation No. SIE-07-001. (2007).

[22] A. Kumar, J. Kim, M. Fulham, D.D. Feng, Graph-Based Retrieval of Multi-Modality Medical Images: A Comparison of Representations Using Simulated Images, IEEE Computer based Medical Systems (CBMS) 2012.

[23] Huang, Y.; Liu, Q.; Zhang, S. and Metaxas, D. N. (2010), Image retrieval via probabilistic hypergraph ranking., in CVPR , IEEE, , pp. 3376-3383 .

[24] Bin Zheng, Computer-Aided Diagnosis in Mammography Using Content-Based Image Retrieval Approaches: Current Status and Future Perspectives, Algorithms 2009, 2(2), 828-849; doi:10.3390/a2020828.

[25] Severin Wiesmller and D. Abraham Chandy, Content based mammogram retrieval using Gray Level Aura Matrix, International Journal of Computer Communication and Information System ( IJCCIS) Vol2. No1. ISSN: 09761349 July Dec 2010.