Analyzing Tree-Like Structures in Biomedical Images Based on Texture and Branching: An Application to Breast Imaging

Michael Barnathan[1], Jingjing Zhang[1], Despina Kontos[2], Predrag Bakic[2],
Andrew Maidment[2], and Vasileios Megalooikonomou[1]

[1] Data Engineering Laboratory, Department of Computer and Information Sciences,
Temple University, 1805 N. Broad St., Philadelphia, PA 19122, USA
{michael.barnathan,jjzhang,vasilis}@temple.edu
[2] Department of Radiology, University of Pennsylvania,
3400 Spruce St., Philadelphia, PA 19104, USA
{despina.kontos,predrag.bakic,andrew.maidment}@uphs.upenn.edu

Abstract. We propose an approach for analyzing tree-like structures in biomedical images. Our analysis is based on Vector Quantization (VQ), an image compression technique. Here, we approach VQ from a different perspective: we use the histogram of the codeword usage as a feature vector representing the initial image. As ductal tree topology has predictive value for a variety of diseases, such as papilloma, ductal ectasia, and ductal carcinoma, we chose to apply this technique to compare texture of the breast ductal tree in x-ray galactograms against the same tissue in corresponding unenhanced mammograms, which do not visualize the ductal tree. We also investigate the relationship between texture and the underlying ductal branching topology using descriptors adapted from the data mining literature. We believe that our method has the potential to assist the interpretation of clinical images and deepen our understanding of relationships among structure, texture, function, and pathology.

Keywords: mammography, galactography, image texture, tree-like structures.

1 Introduction

Analysis of natural and biomedical tree-like structures presents special challenges, as surroundings may obscure branching patterns. Examples of such tree-like structures include the bronchial tree, the blood vessel network, the nervous system, and the breast ductal network. Properties such as topology, spatial distribution of branching, and tortuosity have been analyzed in the literature and associated with altered function and/or pathology. For example, regional changes in vessel tortuosity have been used to identify early tumor development in the human brain [1]. Similarly, studies have shown that the morphology of the ductal network can provide valuable insight to the development of breast cancer and assist in diagnosing pathological breast tissue [2].

E.A. Krupinski (Ed.): IWDM 2008, LNCS 5116, pp. 25–32, 2008.

However, imaging techniques that clearly visualize tree-like structures may be impractical in terms of cost, safety, and comfort. Galactography, for example, can be performed to visualize the breast ductal network by injecting a contrast agent into the lactiferous ducts of the breast (see Figure 1).

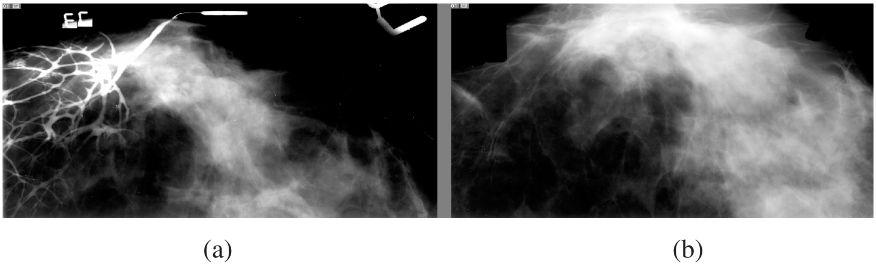


Fig. 1. (a) A galactogram, an x-ray image of the contrast-enhanced breast ductal network, (b) a mammogram of the same breast acquired without contrast-enhancement of the ducts

Galactography can be useful for visualizing early symptoms of papilloma or ductal ectasia, which cause spontaneous nipple discharge in the absence of identifiable mammographic lesions. Nevertheless, such a procedure is not frequently performed and is considered painful and complicated.

In order to overcome such obstacles in the analysis of tree-like structures, we propose an approach that attempts to correlate branching topology with corresponding image texture. Our ultimate hypothesis is that if such a relationship can be quantitatively established, analysis of texture can be used to infer properties of the underlying tree structure, leading to more effective analysis of tree-like structures. In this paper, we focus our analysis on breast imaging due to the particular challenging task of visualizing and characterizing the breast ductal network. Our approach has the potential benefit of advancing our understanding of breast anatomy and physiology, can greatly assist early cancer detection and cancer risk estimation, and may even improve computer simulations of breast tissue for the purpose of evaluating novel breast imaging modalities. Moreover, the proposed representation and methodology could be extended to study other tree-like structures, such as the blood vessel network and airways in the lungs. To summarize, we believe that our approach can help expand upon the current understanding of the relationship between morphology, structure, and function of tree-like structures in medical images.

2 Background

Several approaches have been proposed in the literature for characterizing tree-like structures in images for a wide range of scientific disciplines. *Ramification (R)* matrices model the probability of branching at various tree-levels. R-matrices are computed using the *Strahler number* of a tree t, denoted s_t, which is defined recursively as follows:

- If t is a leaf, $s_t = 1$.

- If there exist two children of s with unequal Strahler numbers, $s_t = \max(s_{\text{children}(t)})$.
- Otherwise, $s_t = s_{\text{children}(t)} + 1$.

The R matrix of a tree with Strahler number s is a lower triangular matrix, defined as:

$$R_{s-1,s} = \lfloor r_{k,j} = b_{k,j} / a_k \,, k \in (2,s), j \in (1,k) \rfloor \tag{1}$$

where a_k is equal to the number of branches with Strahler number k. For $j<k$, $b_{k,j}$ is the number of pairs of branches with labels k and j, while for $j=k$, $b_{k,j}$ is the number of pairs of branches both labeled k-1, descending from a node. *R-matrices* were initially used for studying botanical trees [3]. More recently, the R-matrix approach has been also used for simulating breast ductal trees [4] and classifying radiological findings in clinical breast images [5].

String-encoding descriptors have also been used in the current literature to characterize and classify tree-like structures in breast images [6][7]. By using an encoding scheme, the problem of tree classification is reduced to string classification where node labels comprise the string terms. These characterization strings capture properties of the branching patterns and the topological structure of the corresponding tree.

The *depth-first string encoding* (DFSE) is a straightforward encoding scheme which assigns each node in the tree an ascending integer label based on its position in a preorder traversal. A more sophisticated tree encoding scheme that reflects branching frequencies of the tree nodes is the *Prüfer* encoding. To construct this encoding, each node is visited in preorder and, for all nodes but the root, the label of the parent node is used to represent it.

By treating the string encoding as a document vector, the features of the vector (string elements) are considered to be a collection of terms. *tf-idf* weighting assigns each term in the string a weight determined by its relative frequency within the document and the inverse of its frequency among all documents. The new representation becomes a vector with the corresponding weight at each term's feature position $d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$, where t is the size of the document's vocabulary. More specifically, the main idea of *tf-idf* weighting is that:

(i) more frequent terms in a document are more important, i.e. more indicative of the topic,

(ii) we may want to normalize term frequency (*tf*) across the entire corpus and

(iii) terms that appear in many different documents are less indicative of overall topic.

The weights derived by this approach are given by the following formula:

$$w_{ij} = tf_{ij}\, idf_i = tf_{ij} \log_2 (N / df_i), \tag{2}$$

where:

f_{ij} is the frequency of term i in string j,
$tf_{ij} = f_{ij} \;/ max\{f_{ij}\}$,
df_i = document frequency of term i = number of strings containing term i,

idf_i = inverse document frequency of term $i = \log_2 (N/ df_i)$, and
N is the total number of strings.

3 Method

Our approach for analyzing tree-like structures is based on investigating the relationship between image texture and branching topology of the tree. We consider a framework of methodological steps that aim to quantify this relationship using texture and branching descriptors.

3.1 Segmentation of Regions of Interest (ROIs) and Delineation of the Corresponding Tree Structures

Several preprocessing steps are necessary before the tree-like structures are available for analysis. First, the boundary of these structures needs to be traced to distinguish these structures from the rest of the tissue. This process of segmentation can be performed manually, automatically or semi-automatically using methods such as fuzzy segmentation and the concentric circle analysis proposed by Sholl [8]. Since our methodology focuses on finding descriptive features for classification rather than segmentation, we performed manual segmentation. Once the tree is extracted, a canonicalization step is performed to avoid the tree isomorphism problem, as described in the literature [7].

3.2 Computation of ROI Texture Features Using VQ

We first decompose all ROIs into blocks of equal size and use Vector Quantization (VQ) to represent each block with the closest codeword from a codebook generated by the Generalized Lloyd Algorithm, which produces a “locally optimal” codebook by iterative refinement based on two conditions: the Nearest Neighbor Condition (NNC) and the Centroid Condition (CC). More specifically, this algorithm operates as follows:

Given a codebook $C_m = \{y_i\}$, an improved codebook C_{m+1} is generated by partitioning a training sequence T into cells R_i according to the Nearest Neighbor Condition:

$$R_i=\{x : d(x,y_i) \leq d(x,y_j); \quad \forall j \neq i\} \tag{3}$$

where $d(x,y)$ is the *distortion* between x and y and is generally computed via the Mean Squared Error. In other words, no two neighbors x and y may quantize to the same codeword if there exists a nearer neighbor of x that does not quantize to that codeword. C_{m+1} is then set to the centroids of the new cells:

$$C_{m+1}=\{cent(R_i)\} \tag{4}$$

The algorithm then calculates the average distortion of C_{m+1}, denoted D_{m+1}, and stops if the fractional drop:

$$(D_m - D_{m+1}) / D_m \tag{5}$$

is below a user-defined threshold. Otherwise, the algorithm runs again.

Once the optimal codebook is computed, each image is encoded using the codebook and is represented as a vector of codeword usage frequencies, as illustrated in Figure 2.

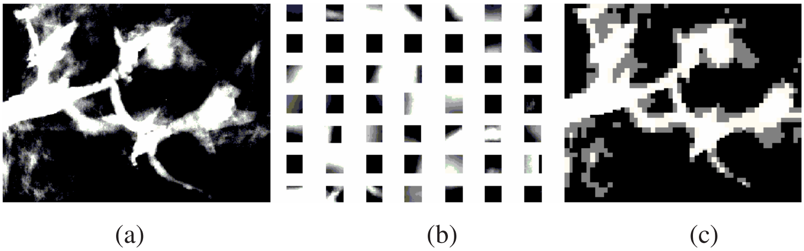
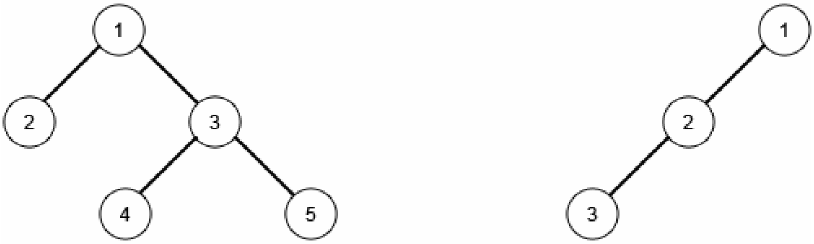


Fig. 2. Vector quantization on a galactogram, depicting (a) a region of interest, (b) part of the codebook generated by GLA algorithm, (c) the quantized representation of the ROI

3.3 Computation of Tree Branching Descriptors Using String Encoding

We then encode the tree using the *Depth-First String Encoding* (DFSE) and the *Prüfer encoding*, representing the string representation as a document vector. We use the *tf-idf* text mining technique to assign a weight of significance to each string term in the encoded tree, indicating terms that form discriminative branching patterns. This step is shown in Figure 3:



Prüfer code: { 1 1 3 3 }

Prüfer code: { 1 2 }

$$df_3 = 1$$

$$idf_3 = \log_2(2 / 1) = 1$$

$tf_{31} = 2 / 2 = 1$

$w_{31} = tf_{31} * idf_3 = 1$

$tf_{32} = 0 / 2 = 0$

$w_{32} = tf_{32} * idf_3 = 0$

Fig. 3. Tree descriptor computation on a sample forest, showing two trees, their Prüfer encodings, and calculation of *tf-idf* weights

3.4 Estimate the Similarity between Texture and Branching Topology Measures

Having obtained texture descriptors for each pair of images, we then analyzed texture-to-texture similarity between all mammograms and galactograms using the summed Euclidean distance metric. Additionally, following the calculation of the *tf-idf* vectors and texture descriptors, we analyzed the similarity between texture and branching descriptors using the cosine similarity measure.

4 Results

We analyzed 8 image pairs, each consisting of one mammogram and one x-ray galactogram of the same tissue, acquired from 5 women (mean age 49.5 years, range 40-95), at the Hospital of the University of Pennsylvania in the period from January 1994 to May 2000. Pairs 1 and 2 corresponded to one woman, as did pairs 3 and 4 and pairs 6 and 7. The mammograms and galactograms in each image pair were labeled $M_{1,}...,M_8$ and $G_{1,}...,G_8$, respectively. Two women (image pairs 1, 2, 6, and 7) had no radiological findings; one woman (image pair 5) was diagnosed with a benign mass, and a benign cyst was suspected (but not pathologically confirmed) in one woman (image pair 8). A diagnosis was not readily available for one woman (image pairs 3 and 4). We computed the VQ descriptors and obtained *tf-idf* weight vectors from the branching *Prüfer encoding* of 7 of the galactographic trees.

Having obtained texture descriptors for each pair of mammograms and galactograms, we assessed the similarity between texture in mammograms and galactograms using summed Euclidean distance. We then derived a similarity measure by normalizing the distance and subtracting from 1. These results, shown in Table 1, were encouraging, with all galactograms except G_5 and G_6 showing statistically significant associations with their corresponding mammograms at $\alpha = .05$. The associated p-values are shown in Table 2. These results support our hypothesis that the underlying texture of the breast ductal network may be inferred from an unenhanced mammogram. These are the mammogram-to-galactogram results; the galactogram-to-mammogram results were also computed but are not shown, as they are identical due to the metric nature of Euclidean distance.

Table 1. Normalized similarity (1 - distance)

	Normalized Texture Similarity							
	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8
M_1	**.55**	.18	.13	.13	.09	.11	.12	.12
M_2	.18	**.49**	.12	.13	.10	.11	.12	.12
M_3	.12	.12	**1.0**	.17	.09	.10	.10	.10
M_4	.13	.13	.17	**.76**	.09	.10	.10	.10
M_5	.10	.10	.09	.09	.20	.09	.12	.11
M_6	.10	.10	.10	.10	.12	.11	.11	.10
M_7	.11	.11	.10	.10	.12	.11	**.43**	.14
M_8	.13	.13	.10	.10	.11	.12	.14	**.46**

We have also computed the correlation between galactographic texture descriptors and the branching descriptors (*tf-idf* weights) of the Prüfer-encoded trees. G_5 was not included in the branching analysis as the position of the nipple was unclear. Unfortunately, a significant correlation has not yet emerged. We are in the process of analyzing these correlations in a larger dataset. We are also developing alternative topological and texture descriptors to improve our ability to investigate this problem.

Table 2. Associated texture distance p-values. Bold values significant at α = .05.

	p-values							
	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8
M_1	**.01**	.44	.59	.56	.68	.63	.61	.61
M_2	.44	**.02**	.61	.59	.65	.63	.61	.61
M_3	.61	.61	**<.01**	.47	.68	.65	.65	.65
M_4	.59	.56	.47	**<.01**	.68	.65	.65	.65
M_5	.65	.65	.68	.68	.40	.68	.61	.63
M_6	.65	.65	.65	.65	.61	.63	.63	.65
M_7	.63	.63	.65	.65	.61	.63	**.05**	.56
M_8	.59	.59	.65	.65	.63	.61	.54	**.03**

5 Discussion

We presented a new methodology for capturing the topology of tree-like structures using texture analysis techniques, demonstrating its efficacy through analysis of the breast ductal network. We utilized encoding schemes, such as DFSE and Prüfer encoding, to represent the tree structure as a string, then performed *tf-idf* weighting to yield vector descriptors of the branching structure. We then performed texture analysis using vector quantization. Our results suggest that it is possible to deduce the texture of the lactiferous ductal network from ordinary mammograms. Considering the small size of our dataset, additional study needs to be performed on larger collections of data to further evaluate our approach.

Acknowledgments. This work was supported in part by the National Science Foundation under grant IIS-0237921. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

1. Bullitt, E., Zeng, D., Gerig, G., Aylward, S., Joshi, S., Smith, J.K., Lin, W., Ewend, M.G.: Vessel Tortuosity and Brain Tumor Malignancy: A Blinded Study. Academic Radiology 12, 1232–1240 (2005)
2. Pereira, B., Mokbel, K.: Mammary ductoscopy: past, present, and future. International Journal on Clinical Oncology 10, 112–116 (2005)
3. Vannimenus, J., Viennot, X.G.: Combinatorial tools for the analysis of ramified patterns. Journal of Statistical Physics 54, 1529–1538 (1989)
4. Bakic, P.R., Albert, M., Brzakovic, D., Maidment, A.D.: Mammogram synthesis using a three-dimensional simulation. III. Modeling and evaluation of the breast ductal network. Medical Physics 30, 1914–1925 (2003)
5. Bakic, P.R., Albert, M., Maidment, A.D.: Classification of galactograms with ramification matrices: preliminary results. Academic Radiology 10, 198–204 (2003)

6. Kontos, D., Megalooikonomou, V., Javadi, A., Bakic, P., Maidment, A.: Classification of Galactograms Using Fractal Properties of the Breast Ductal Network. In: Proc. of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), Arlington, Virginia (2006)
7. Megalooikonomou, V., Kontos, D., Danglemaier, J., Javadi, A., Bakic, P., Maidment, A.D.A.: A representation and classification scheme for tree-like structures in medical images: An application on branching pattern analysis of ductal trees in x-ray galactograms. In: Proceedings of the SPIE Conference on Medical Imaging, San Diego, California, February 2006, vol. 6144, 61441H (2006)
8. Sholl, D.A.: Dendritic Organization in the Neurons of the Visual and Motor Cortices of the Cat. Journal of Anatomy 87, 387–406 (1953)