Content-based image retrieval of multiphase CT images for focal liver lesion characterization

Yanling Chi
Singapore Bioimaging Consortium, Agency for Science, Technology and Research, 30 Biopolis Street, #07-01, Matrix, Singapore 138671

Jiayin Zhou
Institute for Infocomm Research, Agency for Science, Technology and Research, 1 Fusionopolis Way, #21-01, Connexis, Singapore 138632

Sudhakar K. Venkatesh
Department of Diagnostic Imaging, National University Hospital, 5 Lower Kent Ridge Road, Level 2, Main Building 1, Singapore 119074 and Mayo Clinic, Rochester, Minnesota 55905

Qi Tian and Jimin Liu
Singapore Bioimaging Consortium, Agency for Science, Technology and Research, 30 Biopolis Street, #07-01, Matrix, Singapore 138671

(Received 8 April 2013; revised 21 July 2013; accepted for publication 23 August 2013; published 11 September 2013)

Purpose: Characterization of focal liver lesions with various imaging modalities can be very challenging in the clinical practice and is experience-dependent. The authors’ aim is to develop an automatic method to facilitate the characterization of focal liver lesions (FLLs) using multiphase computed tomography (CT) images by radiologists.

Methods: A multiphase-image retrieval system is proposed to retrieve a preconstructed database of FLLs with confirmed diagnoses, which can assist radiologists’ decision-making in FLL characterization. It first localizes the FLL on multiphase CT scans using a hybrid generative-discriminative FLL detection method and a nonrigid B-spline registration method. Then, it extracts the multiphase density and texture features to numerically represent the FLL. Next, it compares the query FLL with the model FLLs in the database in terms of the feature and measures their similarities using the L1-norm based similarity scores. The model FLLs are ranked by similarities and the top results are finally provided to the users for their evidence studies.

Results: The system was tested on a database of 69 four-phase contrast-enhanced CT scans, consisting of six classes of liver lesions, and evaluated in terms of the precision-recall curve and the Bull’s Eye Percentage Score (BEP). It obtained a BEP score of 78%. Compared with any single-phase based representation, the multiphase-based representation increased the BEP scores of the system, from 63%–65% to 78%. In a pilot study, two radiologists performed characterization of FLLs without and with the knowledge of the top five retrieved results. The results were evaluated in terms of the diagnostic accuracy, the receiver operating characteristic (ROC) curve and the mean diagnostic confidence. One radiologist’s accuracy improved from 75% to 92%, the area under ROC curves (AUC) from 0.85 to 0.95 (p = 0.081), and the mean diagnostic confidence from 4.6 to 7.3 (p = 0.039). The second radiologist’s accuracy did not change, at 75%, with AUC increasing from 0.72 to 0.75 (p = 0.709), and the mean confidence from 4.5 to 4.9 (p = 0.607).

Conclusions: Multiphase CT images can be used in content-based image retrieval for FLL’s categorization and result in good performance in comparison with single-phase CT images. The proposed method has the potential to improve the radiologists’ diagnostic accuracy and confidence by providing visually similar lesions with confirmed diagnoses for their interpretation of clinical studies.

© 2013 American Association of Physicists in Medicine. [http://dx.doi.org/10.1118/1.4820539]

Key words: focal liver lesion characterization, multiphase image retrieval, similarity query, multiphase representation, clinical decision support system

1. INTRODUCTION

Multiphase contrast-enhanced computed tomography (CT) has been widely used for the characterization of focal liver lesions (FLLs), such as hepatocellular carcinoma (HCC), metastases and hemangioma (HEM), etc., due to its various advantages.1–5 A multiphase CT study can be acquired in four phases: noncontrast (NC) enhanced phase, arterial (ART) phase, portal venous (PV) phase, and delayed (DL) phase. The visual characteristics of FLLs and their evolutions over the four phases carry important diagnostic information. Generally, FLLs similar in image appearance correspond to the same disease category.6,7 hence radiologists can, to a certain extent, characterize the lesions based on their radiological
2. METHODOLOGY

The multiphase-feature-based CBIR is proposed to assist FLL characterization. Its flowchart is shown in Fig. 1.

The system first detects FLLs using a hybrid generative-discriminative method. Then, a FLL of interest is selected in FLL characterization. Its flowchart is shown in Fig. 1.

In this paper, we present a CBIR prototype system based on multiphase CT images to help radiologists in characterizing focal liver lesions. A preliminary version of this work has been reported. The rest of the paper is organized as follows: Sec. 2 presents the methodology, Sec. 3 describes experiments, Sec. 4 reports the experimental results, Sec. 5 discusses the work, and Sec. 6 draws conclusions.

2.A. FLL detection

Finding FLLs in the CT images and labeling them is the preprocessing of the CBIR system. In this study, a hybrid generative-discriminative method is employed to detect FLLs directly in 3D. Initially, it uses a generative model to describe the nonlesion components, such as the normal liver parenchyma and the enhanced vessels, and then identify all candidate FLLs within a 3D liver volume by eliminating the nonlesion components. It subsequently employs a discriminative approach to suppress the false positives with the advantage of a lesion-likelihood measure combining three shape features: spherical symmetry, compactness, and size. The detection algorithm detects FLLs on images in four phases and highlights them for the user’s interpretation. Among the detected lesions, the user can select one FLL for subsequent processing.

2.B. Multiphase image registration

Multiphase CT scans encode the properties of lesion’s blood supply, carrying essential diagnosis information. Hence, FLL representation based on multiphase images can be more discriminative than the representation based on images in single phase. To exploit the evolution of radiological appearance in the four-phase CT scans as a discrimination factor, density and texture features need to be extracted from images of lesions in the four acquisition phases. However, a FLL is usually easy to be observed in one or two phases while difficult to be observed in other phase(s). Localizing lesions precisely is especially difficult in the phases where FLLs are in very low contrast to the liver parenchyma. In a clinical CT study, the four-phase images might not be well aligned due to the possible displacements in patient’s positions and respiratory phases. The image alignment is desirable. With image registration, images in different phases can be aligned globally. Thus, the position of a lesion in one phase can correspond to the same position in other phases. In this study, the images, in the phase that the query FLL being labeled, are fixed and images in the other three phases are aligned with the fixed images using the existing algorithm: nonrigid B-spline registration or affine registration. An example is shown in Fig. 2. The boxes are suspicious regions detected in the fixed phase and drawn on images in the other three phases using the same coordinates. The lesion positions, before and after image registration, can be compared in the four-phase images. It can be observed that global registration, using nonrigid B-spline registration, is able to well align the lesion regions across the four phases, while affine registration fails to align the arterial phase images with the portal venous phase images. Possible reasons might be that the lesion is at the bottom of the right lobe, where the liver is prone to large deformation due to the patient’s respiration. Nonrigid B-spline registration is employed in this work.

2.C. FLL representation

FLLs’ imaging appearance depends on their biological properties. A cyst is a thin-walled structure that contains fluid, thus, it has internal attenuation near water on CT and typically does not enhance following contrast administration. A hemangioma consists of large, thin-walled, blood-filled vascular spaces separated by fibrous septa, in which blood flow is extremely slow. On CT, the lesion...
usually enhances from the periphery, gradually fills in centrally, and becomes isodense or hypodense compared to the liver parenchyma. Thrombosis within the vascular channels may result in central fibrosis and calcification. Most focal nodular hyperplasia (FNH) is hypervascular, in which hemorrhage, necrosis, and infarction are extremely rare. Thus, a FNH demonstrates a brief (approximately 1 min) and uniform tumor enhancement during arterial phase and becomes isodense in portal venous and delayed phases. A liver abscess (ABS) is a pus-filled cavity, which demonstrates either a single or multiple low-density areas separated by enhanced septa on CT. A HCC demonstrates three major growth patterns: diffuse infiltrative, solitary massive, and multinodular. It commonly has intratumoral hemorrhage and necrosis, and sometimes has calcifications (punctuate, stippled, or rimlike). Most of HCCs are hypervascular and demonstrate contrast enhancement in arterial phase, with diminishing enhancement in delayed phase. Metastases demonstrate a wide spectrum of appearance on CT images. They may be uniformly solid, necrotic, cystic, or calcified. They may be avascular, hypovascular, or hypervascular. They are commonly irregular and poorly marginated but may be sharp and well defined. Typical characteristic is bandlike peripheral enhancement in arterial phase, with rapid washout in delayed phase.

Generally, the imaging appearance of a FLL is determined by several factors: (a) tissue, such as water-filled, pus-filled, or blood-filled lesions, (b) heterogeneous additional content, such as fat, calcification, fibrosis, hemorrhage, infarction, or necrosis, and (c) lesion vascularity, etc. These factors make the FLL demonstrate variable densities and textures because of their differences in enhancement following intravenous contrast administration. Therefore, in total, four multiphase features, i.e., density feature, density derivative feature, texture features, and texture derivative feature are extracted from one FLL to represent the lesion’s heterogeneity and enhancement pattern in four phases.

2.C.1. Density feature

The density feature measures the lesion density enhancement pattern with respect to the liver parenchyma. It is defined as the ratio of the average lesion density to the average density of the liver parenchyma, and is calculated as a four-dimensional feature vector

\[ f_1 = \{d_{NC}, \ d_{ART}, \ d_{PV}, \ d_{DL}\}, \]

where \( d_{NC} = \text{density}_{\text{lesion}} / \text{density}_{\text{liver}} \), \( d_{ART} \), \( d_{PV} \), and \( d_{DL} \) are defined. The density feature is used to describe a hyperdense or hypodense focal lesion. If a lesion is with the density feature \( d_{ART} \) larger than 1 and \( d_{PV} \) less than 1, it is hyperdense in arterial phase and hypodense in portal venous phase. The lesion may be supplied by the hepatic artery and have contrast agent washout in portal venous phase.

2.C.2. Texture feature

The texture feature measures the lesion’s heterogeneity, including six textural coefficients: Energy, Entropy, Inverse Difference Moment (IDM), Inertia, Cluster Shade, and Correlation.\(^{12, 13}\) The texture coefficients are calculated based on Gray Level Co-occurrence Matrix (GLCM), which is an estimation of a joint probability density function (PDF) of gray level pairs in an image. GLCM can be expressed by \( G_{\theta, d}(i, j)(i, j = 0, 1, 2, \ldots, N - 1) \), where \( i, j \) indicate the gray levels of two pixel, \( N \) is the number of gray levels, and the value of \((\theta, d)\) decides the direction and distance of two pixels. The computation of the texture coefficients is described in Table I. The texture coefficients of a FLL, extracted in arterial phase, portal venous phase, and delayed phase, are integrated into an 18-dimensional texture feature vector

\[ f_2 = \{T_1, \ T_2, \ T_3, \ T_4, \ T_5, \ T_6\}, \]

where \( T_k = \{t_{k, ART}, \ t_{k, PV}, \ t_{k, DL}\}, k = 1, 2, 3, 4, 5, 6 \). The proposed method extracts texture coefficients on axial plane, in four directions (0°, 45°, 90°, and 135°) and with four offsets (3 pixels, 5 pixels, 7 pixels, and 9 pixels). Given an offset, i.e., the distance of two pixels, four-directional texture coefficients are averaged to obtain the rotation invariance to a certain degree. The optimal value for the offset is determined experimentally, which is described in Sec. 4.

The features extracted from slices of the lesion are averaged to represent the lesion. As a FLL region is a 3D volume, it is interesting to know the system performance with the texture coefficients extracted in 13 directions, since a voxel has 26 neighbors in the three-dimensional space. This issue will be discussed in Sec. 4.
Correlation illustrates the spatial arrangement of gray levels. It is an indication of the linearity of the relationship of gray levels of pixel pairs:

\[ t_1 = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} g(i, j)^2. \]

Entropy measures the disorder or randomness of the image:

\[ t_2 = -\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} g(i, j) \log_2(g(i, j)). \]

IDM measures the local homogeneity of the image:

\[ t_3 = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{1}{1 + (i - j)^2} g(i, j). \]

Inertia measures the local variations presented in the image:

\[ t_4 = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i - j)^2 g(i, j). \]

Cluster shade gauges the perceptual concepts of uniformity and proximity:

\[ t_5 = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i + j - \mu_i - \mu_j)^2 g(i, j). \]

Correlation illustrates the spatial arrangement of gray levels. It is an indication of the linearity of the relationship of gray levels of pixel pairs:

\[ t_6 = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \sigma_i \sigma_j, \]

where \( \mu_i = \sum_i \sum_j g(i, j), \sigma_i = \sum_i \sum_j (i - \mu_i)^2 g(i, j), \mu_j = \sum_i \sum_j g(i, j), \) and

\[ \sigma_j = \sum_i \sum_j (j - \mu_j)^2 \sum_i g(i, j). \]

### 2.C.3. Texture derivative feature

The six texture properties, constituting the texture feature, may be measured in different value scopes. A difference in quantity may represent a small change of texture properties measured in large values, while it may represent a big change of texture properties measured in small values. If the six texture coefficients are compared and equally integrated into a similarity/distance measurement, in certain cases, some coefficients with large values may dominate the similarity measurement. Normalization is desirable for simple feature comparison. Lesions of same type have similar characteristics in enhancement during four phases, while the enhancement, in terms of quantity, may be different due to the individual variances of patients or scanning protocols. We choose to normalize the texture changes in three phases by the largest change so as to weaken the influence of individual variations of patients. Thus, the texture derivative feature, derived from the texture feature vector \( f_2 \), is designed to capture the evolution trend of lesions’ texture appearance relatively. It is calculated by

\[ f_3 = \{ F_1, F_2, F_3, F_4, F_5, F_6 \}, \]

where \( F_k = \{ f_k^{ART}, f_k^{PV}, f_k^{DL} \}, k = 1, 2, 3, 4, 5, 6, \)

\[ f_k^{ART} = \frac{t_k^{ART} - \text{median}(t_k^{ART}, t_k^{PV}, t_k^{DL})}{\max(t_k^{ART}, t_k^{PV}, t_k^{DL}) - \min(t_k^{ART}, t_k^{PV}, t_k^{DL})}, \]

\[ f_k^{PV} = \frac{t_k^{PV} - \text{median}(t_k^{ART}, t_k^{PV}, t_k^{DL})}{\max(t_k^{ART}, t_k^{PV}, t_k^{DL}) - \min(t_k^{ART}, t_k^{PV}, t_k^{DL})}, \]

and

\[ f_k^{DL} = \frac{t_k^{DL} - \text{median}(t_k^{ART}, t_k^{PV}, t_k^{DL})}{\max(t_k^{ART}, t_k^{PV}, t_k^{DL}) - \min(t_k^{ART}, t_k^{PV}, t_k^{DL})}. \]

### 2.C.4. Density derivative feature

The density derivative feature, derived from density feature \( f_1 \), measures the density changes of a lesion in three enhanced phases relative to noncontrast enhanced phase. It is calculated by

\[ f_4 = \{ \text{diff}_{ART}^{NC}, \text{diff}_{PV}^{NC}, \text{diff}_{DL}^{NC} \}, \]

where \( \text{diff}_{ART}^{NC} = \frac{\text{density}_{ART} - \text{density}_{ART}^{NC}}{\text{density}_{ART}^{NC}}, \)

\( \text{diff}_{PV}^{NC} = \frac{\text{density}_{PV} - \text{density}_{PV}^{NC}}{\text{density}_{PV}^{NC}}, \)

and

\( \text{diff}_{DL}^{NC} = \frac{\text{density}_{DL} - \text{density}_{DL}^{NC}}{\text{density}_{DL}^{NC}}. \)

The overall representation of a FLL is an integration of the four features: \( \{ f_1, f_2, f_3, f_4 \} \).

### 2.D. Similarity query

After FLLs are represented using multiphase density and texture features, the similarity between a query lesion and a model lesion in the database can be numerically measured by a L1-norm-based similarity score between their feature vectors. It is calculated as

\[ \text{Similarity} = 1 - D_{L1} = 1 - \sum_{i=1}^{4} \left( w_i \frac{1}{M_i} \sum_{n=1}^{M_i} | f_i^{\text{input}}(n) - f_i^{\text{model}}(n) | \right), \]

where \( D_{L1} \) represents the L1-norm distance between feature vectors of a query lesion and a model lesion. \( f_i^{\text{input}}(n) \) represents the \( n \)th value in the \( i \)th feature vector of a query FLL. \( f_i^{\text{model}}(n) \) represents the \( n \)th value in the \( i \)th feature vector of a model FLL. \( M_i \) is the dimension of the \( i \)th feature vector, and \( w_i \) is the weight of the \( i \)th feature vector. The weight selection is discussed in Sec. 4.

The model FLLs are ranked according to their similarity scores with the query FLL, and the top FLLs are returned as the retrieval results. Based on retrieval results, the system
can also predict the pathological type of a query lesion. In this study, the “Bull’s Eye Percentage Score” (BEP) (Ref. 14) is employed to measure the retrieval accuracy, which is defined, for each query, as the percentage of correct matches with respect to the query’s class that appear within the top 2C retrieved results. C denotes the size of the considered query class. The query is predicted as the class with the highest retrieval accuracy. The classification is conducted using

\[ \text{query} \subseteq T_i \quad \text{if} \quad \text{BEP}(T_i) = \max_{k=1,2,3,4,5,6} \left( \text{BEP}(T_k) \right), \]  

(1)

where \( T_i \) is the class of the \( i \)th pathologies, \( k \) is the number of pathologies of the model lesions. BEP \( (T_i) \) is the BEP score when the query is considered belonging to \( T_i \). The similarity of a query FLL with the class of the pathology \( T_i \) is measured using the average similarity score of model FLLs of \( T_i \), ranked within the top 2C retrieval results. It can also be viewed as the algorithm’s confidence to predict a lesion of \( T_i \). It is calculated as

\[ \text{Similarity}(T_i) = \frac{1}{N(T_i)} \sum_{j=1}^{N(T_i)} \text{similarity}(j), \]  

(2)

where \( N(T_i) \) is the number of the recalled class \( T_i \)s within the top 2C retrieval results.

An interpretation of the query lesion, with respect to its position, size, and pathological type, is finally given based on the lesion detection and the similarity query results. Even though the system can predict the pathological type of a lesion, it is important for a radiologist not only to refer to the predicted type, but also to look into the pathologies which are easily confused. Most queries return more than one type of focal lesions within the top retrieval results. Each pathological type has the accuracy calculated using Eq. (1). The type with the highest accuracy will be the system’s prediction, and the other types with a bit lower accuracies are frequently confused with the query type. For example, when querying a liver abscess, both liver abscesses and liver metastases can be returned within the top results. Liver abscess may have a high accuracy, while liver metastasis (METS) has a relatively low accuracy. It indicates that liver abscess and liver metastasis need a differential diagnosis. This is possible in a clinical scenario since these two types of lesions can have similar radiological features in certain cases.

3. EXPERIMENTS

3.A. Experimental data

Approved by the Institutional Review Board, 69 clinical four-phase CT scans from 69 patients were used in this study. CT scans of the liver were acquired on one 64-detector scanner (SOMATOM Sensation, Siemens Medical Solutions, Forchheim, Germany), using a standard four-phase contrast-enhanced imaging protocol with the slice collimation of 0.6 mm, matrix of 512 × 512 pixels, and inplane resolution of 0.59–0.78 mm. The raw data were then reconstructed into 3 mm-thick slices without interslice gap. All the data, used in this study, were anonymous and were identified by a preallocated ID. The final confirmation of pathology was based on clinical features, imaging characteristics including those on CT and other imaging modalities, and a histological confirmation with a percutaneous biopsy or surgical excision wherever indicated. The database consisted of 69 lesions with confirmed diagnoses and the distribution of the pathologies was shown in Table II. It was divided into two data sets: the training data and the testing data. Twelve training datasets, including two each of liver metastases, HCC, FNH, hemangioma, cyst and liver abscess, were randomly selected for parameter selection. Each class had the same number of scans, thus, the parameters determined were fair to all classes. The training data were excluded from the database and the rest was the testing data. The raw data of the 12 training data were reconstructed into 0.6 × 0.6 × 0.6 and 3 × 3 × 3 mm³ isotropic volumes for 3D texture testing.

3.B. Experimental design

The retrieval system was evaluated by a common performance measure of the precision-recall plot. Precision is defined as the number of matched lesions divided by the total number of lesions retrieved. Recall is defined as the number of matched lesions divided by the total number of lesions in the same class in the database. The ideal precision-recall curve is a curve when precision equals one at all values of recall. The leave-one-out strategy was used to generate the precision-recall curves and the BEP scores.

3.B.1. Experiments on task subjectivity

FLL characterization is experience dependent. It is interesting to study the subjectivity of the task. Two radiologists (radiologists 1 and 2) with three-year and five-year experience were invited to review all data in the database. Both radiologists reviewed the CT scans and clinical information and performed characterization of lesions. Their diagnoses were compared with each other in terms of Dice coefficient15 to evaluate the task subjectivity.

3.B.2. Experiments on parameter selection

Experiments were conducted to determine the optimal value of the offset in texture coefficients extraction, and the

<table>
<thead>
<tr>
<th>Pathologies</th>
<th>No. of FLLs</th>
<th>No. of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>METS*</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>HCC</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>HEM</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Cyst</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>ABS</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>FNH</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>69</td>
<td>69</td>
</tr>
</tbody>
</table>

*The primary for liver metastases includes pancreatic carcinoma (2), sigmoid carcinoma (5), rectal carcinoma (1), colorectal carcinoma (1), and gallbladder carcinoma (1).
optimal weighting in the similarity measurement. The system, employing different offset values, was evaluated on the training data and the corresponding precision-recall curves were generated. The offset value that can result in the best performance was used in the rest experiments. In the experiments to select the offset, an equal weighting was set to calculate the similarity score. A Brute-force search was then conducted to select the optimal weighting. Its objective was to maximize the BEP score under the conditions of the 0.1 increment/decrement of the weights and the sum of the weights being 1. The system between using optimal weights and an equal weighting was also compared to evaluate its performance on different weightings.

3.B.3. Experiments on 3D textures

To evaluate 3D texture features, the GLCM-based texture coefficients were extracted in 13 directions from FLLs of the isotropic training data. They were averaged to obtain the rotation invariance to a certain degree. The systems, with lesions represented using 3D texture coefficients and using the averaged 2D texture coefficients, were compared on performance in terms of the precision-recall plot. The representation, resulting in a better performance, was more discriminative and will be employed in this study.

3.B.4. Experiments on performance evaluation

After parameter selection and texture feature evaluation on the training data, the system performance on FLL characterization, with single-phase representation and multiphase representation, was evaluated on the testing data. The system when employing multiphase features, and when employing the single-phase features extracted from arterial phase, portal venous phase, and delayed phase, respectively, were compared in terms of the BEP score and the precision-recall curve. Experiments were also conducted to analyze the saliency of the proposed features. The system, employing various feature combinations, was evaluated in terms of the precision-recall curve and the BEP score.

3.B.5. Pilot trial

Two radiologists (radiologists 3 and 4) with about five-year experience were invited to evaluate whether the retrieval results can affect their decisions and to what extent they were influenced. To avoid learning effects, it is important to make no query data appear in the retrieved results during the trial. The leave-one-out evaluation method is not suitable here, and the query data and the query database need to be well separated. To be fair to each of six classes, we used the 12 training data (two lesions for each class) as the query data and the 57 testing data as the database in this pilot trial. The procedure was as follows. First, random numbers were allocated to the 12 training lesions, and ranked the lesions in a list. Each entry in the list, corresponding to one lesion, consists of the random number, the CT scans’ IDs, the lesion’s coordination in CT volumes. Then, the list was passed to the radiologists to do first-round diagnoses in sequence from top to bottom. They, continuously, accessed CT scans stored on the PC and given their diagnoses and confidence level to the lesions based on the images. Their diagnoses were recorded. The confidence level was from 1 to 10, where 1 represented that the radiologist was not confident at all about his/her diagnoses and 10 represented that he/she was absolutely certain of the diagnoses. Two weeks later, those lesions were ranked randomly again in a list. The list was first used to retrieve the testing database by an engineer. Top five model lesions, returned for each query, were recorded, as well as their corresponding confirmed diagnosis and similarity scores. The retrieval results were attached to the corresponding entry in the list. The list was then provided to the radiologists to do the second-round diagnoses. During this round, the radiologists referred to the CT’s ID and lesion locations.

The diagnoses, without and with the knowledge of the top five retrieved results, were compared with the confirmed diagnoses in terms of the diagnostic accuracy: the ratio of the number of true positives to the total number of query lesions. The higher the ratio, the better the overall diagnoses. No false positive rate was calculated since all the queries were lesions. Alternatively, for each type of lesions, the ROC curves were generated from the two-round diagnoses. The average ROC curves of the six types of lesions and the areas under the curves (AUC) were compared between two-round results using the method of Delong et al. The radiologists’ diagnostic confidence levels were given positive signs if the diagnoses were true positives. The values were given negative signs if the diagnoses were false negatives. An increase in the second-round diagnostic confidence indicated that the radiologists corrected the previous diagnoses, gained more confidence with the previous right diagnoses, or lose some confidence with the previous wrong diagnoses. The radiologists’ confidence levels of two-round diagnoses were compared and statistically analyzed using the Wilcoxon signed-rank test. A $p$-value $\leq 0.05$ was considered to represent statistical significant. The inter-radiologist agreements of the two-round diagnoses were measured using the Dice coefficient. All statistical analysis in this pilot study was performed using MedCalc® (http://www.medcalc.org/).

4. EXPERIMENTAL RESULTS

4.A. Results on task subjectivity

The Dice coefficient between radiologist 1 and 2’s diagnoses was 71%. FLL characterization using CT images was, to a certain extent, subjective.

4.B. Results on parameter selection

The precision-recall plots over different offset values were displayed in Fig. 3. It can be observed that the best performance was resulted from using $d = 7$ pixels, i.e., the curve using $d = 7$ pixels was closest to the ideal curve. Therefore, the offset was set to 7 pixels in the texture feature extraction.
The optimal weighting was determined as \([0.1, 0.1, 0.4, 0.4]\), which was set in the rest experiments. The equal weighting was our initial selection, i.e., the weight of each feature was set to 0.25. It was interesting to compare the precision-recall plots, between the optimal weighting and the identical weighting, displayed in Fig. 4. It can be observed that the proposed method was, to a certain extent, sensitive to the weights selections.

4.C. Results on 3D textures

The system performance with lesions represented by only texture features, i.e., 3D texture coefficients and averaged 2D texture coefficients, was evaluated and the corresponding precision-recall plots were illustrated in Fig. 5. It can be observed that the system performed best with the averaged 2D feature. The 3D texture, extracted from a volume \(3.0 \times 3.0 \times 3.0\) mm\(^3\), can result in a better performance than the texture, extracted from the volume of \(0.6 \times 0.6 \times 0.6\) mm\(^3\). Possible reasons for the relatively low performance of 3D texture features might be that the \(0.6 \times 0.6 \times 0.6\) mm\(^3\) data were relatively noisy, and the \(3.0 \times 3.0 \times 3.0\) mm\(^3\) data might overlook too much spatial information.

4.D. Results on performance evaluation

The system performance on FLL characterization was evaluated and the corresponding results were presented in

4.E. Results in pilot trial

The two-round results were analyzed. The diagnostic accuracy of the radiologist 3 increased from 75% to 92%, while the diagnostic accuracy of the radiologist 4 remained unchanged, at 75%. The average ROC curves were plotted in Fig. 9. The AUC of the radiologist 3’s diagnoses increased from 0.85 to 0.95 with a \(p\)-value of 0.081 (\(\geq 0.05\)), and that of the radiologist 4’s diagnoses increased from 0.72 to 0.75 with a \(p\)-value of 0.709 (\(\geq 0.05\)). The mean diagnostic confidence of the radiologist 3 increased from 4.6 to 7.3 with a \(p\)-value of 0.039 (\(\geq 0.05\)). The increase was statistically significant. The mean confidence of the radiologist 4 increased from 4.5 to 4.9 with a \(p\)-value of 0.607 (\(\geq 0.05\)). The increase was not statistically significant. More datasets may be desirable to detect the significant differences of the radiologist 4’s two-round diagnoses. In first-round diagnoses, the inter-radiologist agreement was 75%, and it dropped to 58% in the second-round diagnoses. The retrieved images resulted in a decrease of the
FIG. 6. System performance comparisons of various phases regarding (a) the precision-recall curves; (b) BEP scores; the multiphase features obtain the highest accuracy.

inter-radiologist agreement probably due to the increased diagnostic accuracy of the radiologist.

We did not conduct significance tests on the diagnostic accuracy as the data set in the pilot trial was small. In the two-round diagnoses, the radiologists changed their decisions in only 2–3 cases out of 12 cases. The effective sample size is too small to statistically detect a significant difference between the two-round diagnoses. This is one of the limitations of this pilot study.

5. DISCUSSION

5.A. Discussion on FLL classification vs CBIR

There is work being reported on the automatic classification of FLLs. Duda et al.\(^\) proposed to represent liver/tumor tissue using texture features of three acquisition phases on contrast-enhanced CT scans. Support vector machine classifiers and dipolar decision trees were used to classify lesions. A fuzzy support vector machine (FSVM) was proposed by Xian\(^\) to classify malignant and benign liver tumors on ultrasonography. The FLL was manually marked by the physician using the region of interest (ROI) and represented by the gray level co-occurrence matrix based texture features. The system developed by Shiraishi et al.\(^\) extracted the morphologic and density features from ultrasonography to represent FLLs and trained six independent artificial neural networks to classify three types of FLLs. In their study, the decisions were made in the cascade which restricted their method’s generalities to other pathological types. In the work of Gletsos et al., three feed forward neural networks were used to classify hepatic lesions on noncontrast enhanced CT images. Texture features were employed to represent liver lesions. In another work from the same group, also tested on noncontrast enhanced phase images, the performance of various computer-aided diagnosis (CAD) architectures were assessed for the discrimination of normal liver tissue and three types of lesions. The highest accuracy was 85% achieved by an ensemble of five different primary classifiers, a fused feature set, and the weighted voting scheme. However, detecting lesions on noncontrast enhanced images, sometimes, is challenging due to their low contrast to the liver parenchyma. The diagnostic information observed in noncontrast enhanced phase is very limited. In general, these above summarized work developed methods to classify unknown liver tissue into a certain pathological type and to provide qualitative diagnostic information as a “second” reader. These systems can only provide users with classification results in decision support. The important evidences for differential diagnosis, e.g., the similar cases that were of different types from the query, were lacking in the

![Graphs showing system performance comparisons](image-url)
classification results, thus, their potential in clinical practice was limited.

Content-based image retrieval (CBIR) systems have been increasingly used in the medical domain. CBIR retrieves images, visually similar to the query submitted to the system, from a database. Different from the FLL classification that classify a FLL to a certain type, a CBIR-based system for FLL characterization presents users retrieval results ranked by the visual similarity. The retrieved FLLs can be of the same type as the query or of the different pathological types from the query, which can support radiologists’ decisions or alert them heuristically with other options. The retrieval results are important evidences for differential diagnosis. Recently, there are studies of FLL characterization using CBIR systems. Yu et al. developed a CBIR to differentiate HCC from cyst and hemangioma on CT images, using local co-occurrence-matrix-based texture and density features as well as wavelet-transform-based global features. This method was tested on around 1700 images from more than 200 patients. It is a single-phase based method. The global features were extracted from an abdominal image, which consists of the lesion and other anatomical structures. The anatomical structures varied on lesion’s positions and may dominate the image content, thus, the effectiveness of global features on lesion characterization can be very limited. The study lacks detailed descriptions on the calculation of the local features and the data they tested on, such as, which phase images they used, how the abdominal images were selected, etc. A recent work from this group was reported by Yang et al. It retrieved an image database of three types of FLLs: hepatoma, cyst, and hemangioma using the bag-of-visual-words (BoW) based method. The image database, in total, consisted of 1375 CT slices from 189 patients. One to five representative slices were selected from one lesion. The average precision of 90% was obtained. Napel et al. retrieved a database of CT images including three types of liver lesions, i.e., hemangioma, liver metastasis, and cyst. In their work, lesions were queried on the single PV phase CT images. Thirty images were tested on and the mean precision was greater than 90% at all values of the recall. A novel CBIR method was proposed by Costa et al. to help to characterize the indeterminate liver metastasis and cyst, based on the intrinsic random forest similarity. In the...
above work, the study of features from multiphase CT images was relatively neglected. To the best of our knowledge, our system is the first in managing to use multiphase CT images in CBIR for FLL characterization.

5.8. Discussion on study comparison

These three studies, are closely related to our work, and it is interesting to compare them with ours in terms of the framework. The comparison was listed in Table III. In the study of Yang et al., FLLs were described using the BoW histograms, with the visual vocabulary being image patches of the training lesions in terms of the raw intensity without normalization. The distance metric learning algorithms were designed to measure the similarity of the BoW histograms. Good system performance was obtained. The method represented the multiphase lesion using a merged BoW histogram of single phase images, which may overlook the time information of the images acquired in sequence. The visibility of a lesion in four phases can be different. To avoid lesions’ mismatching in multiphase representation, finding good correspondence of a lesion in multiple phases is essential. Basically, Yang et al. developed a single-phase-based method. In the study of Napel et al., employing both semantic and low level features made the retrieval of visually similar lesions perform well. The visual similarity between any two lesions was defined by two senior radiologists based on similarities in textures, boundary shape, and sharpness. Three types of similarities were defined as 3/2/1 for very similar, somewhat similar, and not similar, respectively. The system was evaluated in terms of the precision-recall curve on how well the CBIR system can retrieve visually similar lesions in comparison with expert radiologists. In their work, only one slice in portal venous phase was manually selected to represent a 3D lesion. In addition, the system was not optimized in terms of the retrieval performance for FLL categorization. The work of Costa et al. also made use of both high level and low level features, on a much larger FLL database. It employed an elaborated machine learning method, i.e., random forest, to learn a discriminant distance in classifying FLL attributes such as hyperdense vs hypodense in terms of FLL density, benign vs malignant in terms of FLL benignancy, cyst vs liver metastasis in terms of type. Their method was evaluated, in terms of ROC curve, over the classification performance on lesion attributes. Three orthogonal 2D cuts of the lesion ROI were used to represent a 3D liver lesion. In their work, the performance on small subcentimeter FLLs was considered, while clinically very small lesions are usually not diagnosed. Our work employed multiphase CT images in CBIR for FLL categorization. Global registration was employed to align a lesion in four phases. Six types of FLLs have been tested and good performance has been achieved. The retrieval performance in our work demonstrated that CBIR using multiphase features performed much better than using single-phase features in retrieving same category lesions for the given queries.

In summary, Yang’s work proposed a BoW-based method and illustrated that the BoW representation was superior to intensity histogram and Gabor filters based representation. Napel’s work defined a set of high level semantic features for the liver lesion description, and studied how these features can be combined with low level features to retrieve visually similar liver lesions. The work showed the CBIR system performed well in comparison with expert radiologists in assessing liver lesions’ visual similarity. Costa’s work made use of discriminant distance learning to cluster liver lesions and applied CBIR to lesion attribute classification. Our work demonstrated that multiphase CT images can be used in CBIR for FLL’s categorization and resulted in good performance in comparison with single-phase CT images.

5.9. Discussion on feature selection

In this work, we employed gray level co-occurrence matrix based texture coefficients to describe FLLs since they have been widely used to characterize FLLs and obtained good results. Other texture features were not tested on our multiphase image retrieval system, but they will be investigated in our future work.

Shape features are important in describing liver lesions. Slowly growing benign lesions are often well circumscribed or encapsulated, while rapidly growing malignant lesions tend to have an indistinct irregular shape and are not encapsulated. Shape features are good at discriminating benign lesions from malignant lesions, while this work is to characterize six types of FLLs, including four types of benign lesions and two types of malignant lesions. Shape features have insufficient power to differentiate among benign lesions, or among malignant lesions. In addition, FLLs are generally well observed on one-phase images, thus, accurate shape features can only be extracted from single-phase images. They are basically single-phase based features, while multiphase image retrieval is the focus of this work. Multiphase-based features are desirable. To avoid distracting our focus, shape features are not included in this work.

Besides density, texture, and shape features, clinical history and high level semantic information also contain useful information on the clinical diagnosis. High level semantic features have been well addressed by Napel et al. In their work, FLLs were annotated by radiologists using the semantic features from a list of 12 categories comprising 161 possible descriptors, which were selected from the RadLex (Ref. 31) controlled terminology. The semantic features were not tackled in this paper, since our focus was on CBIR based on multiphase features. Considering that the semantic features selected by individual radiologists may not be able to cover all the information carried by the images, it is necessary to explore objective image analysis to bridge the semantic gap of CBIR.

Costa et al. claimed that searching results based solely on the similarity of low level features (computer generated) would be rarely clinically meaningful. They employed the histogram and various moments in the low level liver lesion descriptor and their experimental results showed that the descriptor had weak discrimination power on METS and cyst. The moments are well-known quantitative measures of the global shape of a set of points, but FLL shapes are rarely used
TABLE III. Comparison of studies.

<table>
<thead>
<tr>
<th></th>
<th>Yang’s work (Ref. 27)</th>
<th>Napel’s work (Ref. 28)</th>
<th>Costa’s work (Ref. 29)</th>
<th>Our work</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goal</strong></td>
<td>A bag-of-visual-words-based CBIR for FLL characterization</td>
<td>A similarity retrieval of liver lesions</td>
<td>A learning-based CBIR for FLL characterization</td>
<td>A multiphase feature-based CBIR for FLL characterization</td>
</tr>
<tr>
<td><strong>Materials</strong></td>
<td>• 1375 FLL images</td>
<td>• 30 FLL images</td>
<td>• 1201 FLL volumes</td>
<td>• 69 FLL volumes</td>
</tr>
<tr>
<td></td>
<td>• Single phase</td>
<td>• PV phase</td>
<td>• Single phase</td>
<td>• Four phases</td>
</tr>
<tr>
<td></td>
<td>• Three types (87 HCCs, 62 cysts, and 60 HEMs)</td>
<td>• Three types (13 cysts, seven HEMs, and ten METS)</td>
<td>• Two types (98 cysts and 1103 METS)</td>
<td>• Six types (cyst, METS, HCC, HEM, FNH, and ABS)</td>
</tr>
<tr>
<td><strong>FLL localization</strong></td>
<td>Manual</td>
<td>Manual</td>
<td>Manual</td>
<td>Automated</td>
</tr>
<tr>
<td><strong>FLL description</strong></td>
<td>• Low level features</td>
<td>• High and low level features</td>
<td>• High and low level features</td>
<td>• Low level features</td>
</tr>
<tr>
<td></td>
<td>• Single phase</td>
<td>• Single phase</td>
<td>• Single phase</td>
<td>• Multiphase</td>
</tr>
<tr>
<td><strong>System evaluation</strong></td>
<td>• Leave-one-out on a database classified by pathological types</td>
<td>• Leave-one-out on a separate reference standard</td>
<td>• Leave-one-out on a database classified by pathological types</td>
<td>• Leave-one-out on a database classified by pathological types</td>
</tr>
<tr>
<td><strong>Pilot trial</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

5.D. Discussion on system performance

The system can rank most true positives in high positions, as presented in Sec. 4. It may rank some false negatives in high positions, as illustrated in Fig. 10. In Fig. 10(a), given a query of liver abscess, the system ranked liver metastasis first. It may be due to liver abscess and liver metastasis being similar in the appearance of hypodense, and heterogeneity in the four phases. In Fig. 10(b), hemangioma was ranked second when the system queried a HCC. It was observed that the query HCC demonstrated a density enhancement from lesion margin to inside. In Fig. 10(c), liver metastasis and cyst presented similar enhancement pattern through the four phases, resulting in cyst being rank first. In Figs. 10(d) and 10(e), lesions in three phases have similar appearance, thus, hemangioma and liver abscess were ranked in relatively high positions.

Even though most lesions of the same pathological types are similar in imaging appearance, there exist variations. The overlapping features of different FLLs are well known and it is in fact our motivation to perform this study. The false negatives motivate us for further work to fine-tune the effort toward an automatic method which is promising in helping radiologists.
6. CONCLUSION

We have presented a CBIR prototype system based on multiphase contrast-enhanced CT images. We have demonstrated the system’s potential to help the radiologists to improve their diagnostic accuracy and diagnostic confidence from a test perspective. To be a practical CAD, several technical challenges still need to be solved. Currently, it takes more than 10 min for one query and 90% of the time taken is for the FLL detection and representation. Efficient image processing algorithms and the implementation are necessary for a real-time application. Also, the clinical validation has to be conducted and the system’s robustness on unknown cases should be tested. Although our work has been tested on six common lesion types for this initial study, it has not been tested on the other less common FLLs, such as cholangiocarcinoma. Our database does not include the entire spectrum of FLLs and this is one of the limitations of the study. We will include these less common FLL types in our future study and expanding database is desirable. With the expanded database, calculating similarity between the query and each model in the database will be time consuming, thus the indexing techniques will be urgently needed. Moreover, with more pathological types added in, increasing the discriminative power of the FLL representation is on demand. Nonetheless, we have proposed a general framework to automatically identify and localize FLLs on the four-phase CT scans and represent lesions using the multiphase information. The system has been validated on real clinical CT scans classified according to pathological types, and good performance in FLL characterization has been achieved. A pilot trial has been conducted and encouraging results have been obtained.

ACKNOWLEDGMENTS

This research work was supported by a grant (Grant No. JC0A03_FG05_2009) from the Joint Council Office, Agency for Science, Technology and Research, Singapore. The authors would like to thank Dr. Tiffany Hennedige, Department of Diagnostic Imaging, National University Hospital, Singapore, for her great help on data collection.

41 Author to whom correspondence should be addressed. Electronic addresses: yanling.chi@gmail.com and CHI_yanling@sbic.a-star.edu.sg; Telephone: (0065) 64788351.


51 B. S. Manjunath, Introduction to MPEG-7: Multimedia Content Description Interface (Wiley, Chichester, 2002).


