An Effective WSSENet-Based Similarity Retrieval Method of Large Lung CT Image Databases

Yi Zhuang[1*], Shuai Chen[1], Nan Jiang[2], Hua Hu[3]
[1] School of Computer & Information Engineering, Zhejiang Gongshang University, Hangzhou, P.R.China
[e-mail: zhuang@zjgsu.edu.cn]
[2] Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, P.R.China
[e-mail: zy158cn@163.com]
[3] Hangzhou Normal University, Hangzhou, P.R.China
[e-mail: huhua@hdu.edu.cn]
[*]Corresponding author: Yi Zhuang

Received February 9, 2022; revised March 21, 2022; accepted May 23, 2022; published July 31, 2022

Abstract

With the exponential growth of medical image big data represented by high-resolution CT images(*CTI*), the high-resolution *CTI* data is of great importance for clinical research and diagnosis. The paper takes lung *CTI* as an example to study. Retrieving answer *CTI*s similar to the input one from the large-scale lung *CTI* database can effectively assist physicians to diagnose. Compared with the conventional content-based image retrieval(CBIR) methods, the CBIR for lung *CTI*s demands higher retrieval accuracy in both the contour shape and the internal details of the organ. In traditional supervised deep learning networks, the learning of the network relies on the labeling of *CTI*s which is a very time-consuming task. To address this issue, the paper proposes a Weakly Supervised Similarity Evaluation Network (*WSSENet*) for efficiently support similarity analysis of lung *CTI*s. We conducted extensive experiments to verify the effectiveness of the *WSSENet* based on which the CBIR is performed.

Keywords: CT images, spatial transformer layer, deep learning, content-based medical image retrieval

http://doi.org/10.3837/tiis.2022.07.013 **ISSN : 1976-7277**

1. Introduction

With the rapid development of medical image technology and the advent of the era of big data, the number of high-resolution lung CT image(*CTI*)s increased exponentially. Finding similar lung *CTI*s from the large lung *CTI* repository enables physicians to efficiently and fully learn previous similar cases, thus enabling assisted diagnosis and treatment.

Generally, a set of lung *CTI*s is derived from routine cross-sectional scans that acquire images of lung and mediastinal windows at various cross-sectional levels of the chest. So a set of lung *CTI*s contains hundreds of lung *CTI*s covering various parts of a patient's lungs. These lung *CTI*s saved in the database contain more than just information about the images themselves, but also the patient-specific therapy measures and results. Through the similarity comparison of the *CTI*s, physicians can find similar *CTI*s from previous patients who are likely to have the same disease since they have the same pathological symptoms. The treatment can be continuously improved by referring to previous cases based on which finding similar lung sections from the lung *CTI* database will greatly help physicians make correct diagnosis.

Compared with the conventional content-based image retrieval(CBIR) methods, the content-based medical image (e.g., *CTI*) retrieval(CBMIR) requires higher retrieval accuracy. In most cases, all of the lung *CTI*s are generally similar. The main differences between lung lesions in patients lie in the shape of the lung lobes and the detail information (e.g., the bronchi, blood vessels, etc) inside the lungs. Therefore, it is urgent to develop a novel CBMIR method for lung *CTI*s with higher retrieval accuracy. In addition, due to the complex characteristics of the objects in the lung lobes (i.e., location and shape, etc), it is not easy to accurately describe and quantify them. Furthermore, deep learning-based similarity measure is based on a large amount of data and labels which are provided by medical professionals manually. The labeling process of the medical images, however, is a very time-consuming and expensive task. For the CBMIR processing, firstly, a deep learning network capable of extracting medical image feature descriptors is first trained, then a series of distance formulas is applied to calculate the similarity between feature descriptors, finally finding the most similar medical images. However, this retrieval method has certain shortcomings. The first one is that a large number of labels are necessary to train the network, and the second one is that the extracted image descriptors need to be saved in the disk, which entails the larger storage overhead.

To tackle the above challenges, the paper presents a *Weakly Supervised Similarity Evaluation Network* called the *WSSENet* which is a two-layer-based hierarchical network structure: the first layer is responsible for the shape similarity measure of lung lobes, called the *shape similarity calculator*(SSC); while the second layer is to compute the similarity metric of the details in the lung lobe, called the *detail similarity calculator*(DSC). The training process of the network is also carried out layer by layer. First, the training set required for the SSC is created based on the Spatial Transformation Layer(STL). The labeling accuracy in this training set, however, does not reach the required similarity assessment accuracy for the lung *CTI*s, and therefore it belongs to an inexact supervised learning [1]. To further improve the effectiveness of similarity measure by focusing the accuracy on the soft tissues inside the lung lobes, the DSC needs to be trained. Meanwhile, to automatically construct a high accuracy training set for DSC, we need to resort to the trained SSC. After the SSC and the DSC are trained by the above process, they are combined together to build the hierarchical network structure (i.e., *WSSENet*), which is used in evaluating the similarity between two lung *CTI*s. Our proposed *WSSENet* proved to be highly effective for similarity measurement of lung *CTI*s.

The following is a summary of our contributions:

1. We introduce *WSSENet* which is a new weakly supervised deep learning network for lung *CTI* similarity assessment.
2. We present a novel automatic labeling approach for similarity labeling for the lung *CTI*s.
3. We conduct extensive experimental evaluation based on real datasets to verify the effectiveness and efficiency of our techniques.

The rest of the paper is laid out below: Section 2 reviews the previous work directly related to ours. Section 3 presents a systematic study of the *WSSENet*-based method for lung *CTI* similarity analysis. Section 4 introduces the training process of the *WSSENet*. Section 5 evaluates our algorithms with extensive experiments. Finally, Section 6 concludes the paper with a summary of findings and the directions of future work.

2. Related Work

Content-based medical image retrieval essentially computes the similarity between two medical images by some similarity metrics (e.g., Euclidean distance). Due to the high-dimensional characteristics of medical images, however, the efficiency of calculation using the aforementioned method is quite low. Researchers have begun to consider approaches for extracting the internal features of medical images. The similarity retrieval of medical images may be done by extracting the feature vectors that can represent the features of medical images and then computing the similarity between the feature vectors. The feature extraction research has gone through two stages.

The feature extraction in the first stage was primarily focused on the extraction of low-level visual features that are the fundamental image features that the human eye can see, such as color, texture, and shape, etc. Such basic image features, however, are vulnerable to contamination by noise. Representative features of medical images, on the other side, are localized, therefore researchers began to concentrate on extracting local visual features. Mizotin et al. [2] combined SIFT features with the bag of visual words (BoVWs) algorithm to obtain an excellent retrieval method of brain MRI images, especially for Alzheimer's disease images [3].The Idiap research team [4] combined local binary patterns (LBP) with modSIFT [5], which is used to characterize image textures. This fused feature achieves a satisfactory retrieval on IRMA dataset. Pan et al. [6] first modeled brain CT images and proposed an uncertain location graph (ULG) structure that can be used to better express multiple textures of the brain. Using an index structure, a method for computing the ULG similarity was proposed to speed up the retrieval performance with 80% search precision rate. Karthik et al. [7] tried to combine the image features with different modalities to build a hybrid feature model based on which the CBIR processing can be effectively performed. Sampathila et al. [8] presented a CBIR method using image features (e.g., color, shape, and texture) to represents and retrieves images in a large database that are similar to a given query image. These features are determined based on grayscale co-occurrence-based Haralik features and histogram-based cumulative distribution function (CDF) with excellent results on radiological image retrieval.

As visual features often do not accurately capture the high-level semantic features of medical images well, so in the second stage, with the advancement of deep learning techniques, researchers have focused on employing the deep learning techniques to explore the high-level semantic features in medical images. Shin et al. [9] have fine-tuned the CNN model and pre-trained it on the ImageNet dataset, and then used this pre-trained model to extract features, which is a typical transfer learning idea. Sundararajan et al. [10] used a variant of the CNN model to extract features from avascular necrosis(AN) images to implement retrieval. They adopted the median filter (MF) in image preprocessing to remove the noise in the image, and then obtained the features of medical images, the similarity between features was measured by Hamming distance. Khatami et al. [11] came up with a hierarchical structure for medical image retrieval. In the first layer, they use CNN to get the categories to which the images are most likely to belong, then in the second layer, the images of the same category are formed into a search space by Radon transform to further implement the retrieval. Since then, Khatami et al. [12] has proposed a two-step hierarchical shrinking technique using CNN transfer learning and a Radon projection pool for medical image retrieval. Ma et al. [13] fused the semantic and visual similarities between the query image and each image in the database as the similarity between images. For the semantic features, the images were classified into multiple cases and classified with support vector machine. The visual features (i.e., HOG-based BoVW, wavelet features, LBP and CT-valued histogram) were extracted, from which the optimal sub-features were selected.

To quickly retrieve medical images from large datasets, hash-based methods were proposed which project high-dimensional features into a low-dimensional space and then generate compact binary codes. Lai et al. [14] presented a deep neural network hashing (DNNH) method that describes more complex semantic information by using triplet-based constraints. Liu et al. [15] proposed a deep supervised hashing method (DSH) that takes pairs of images (similar/dissimilar) as training input to a CNN and encourages the output of each image to be close to the discrete value, and their extracted features obtain excellent retrieval results. Cai et al. [16] designed a new loss function based on CNN with hash coding to learn models to make images belonging to the same class with similar features, and the proposed method achieved satisfactory results.

3. The WSSENet

In this section, we first provide the preliminaries and a system framework of the *WSSENet* for lung *CTI* similarity analysis. Next, the dataset generator and similarity calculator are presented in Sections 3.2 and 3.3, respectively.

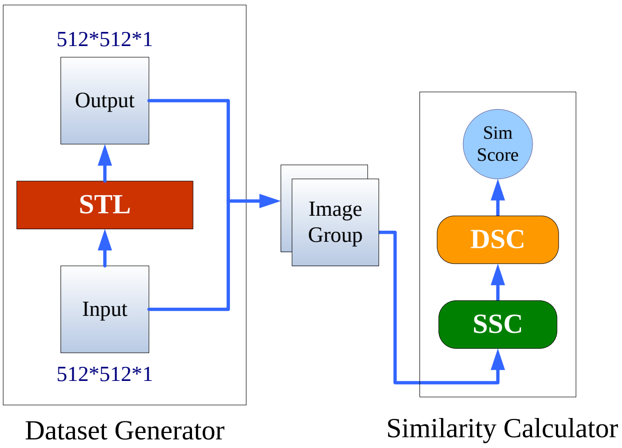


Fig. 1. The whole framework of the *WSSENet*

3.1 Preliminaries and Overall Framework

Firstly, **Table 1** summarizes the major symbol notations.

Table 1. Primary notation used throughout the paper

Notation	Meaning
SSC	shape similarity calculator
DSC	shape similarity calculator
PA	pathological area
IP_{nm}	the *m*-th image patch in the *n*-th *CTI*
STL	spatial transformer layer
STN	spatial transformer network
VT	vision transformer
S_i	the *i*-th white streak
CTI_i	the *i*-th CT image

As depicted in **Fig. 1**, the *WSSENet* consists of two major modules: *the dataset generator* and *the similarity calculator*. For the dataset generator, a STL is employed to generate an initial training set for network training. In such a way, the follow-up network training tends to be inexact supervision. As the key part of the *WSSENet* used for network training, a SSC and a DSC are introduced, which are trained and tested over the LUNA16 dataset [22].

3.2 Dataset Generator

As stated in Section 1, the purpose of *Spatial Transformation Network* (STN) [17] is to enable the model to be unaffected by changes in object pose or position in computer vision. In this subsection, we introduce a new structure called the *spatial transformer layer*(STL) which is the simplified version of the STN by removing the localization network module. As the weak supervised characteristic exhibited by the STL, it can generate the *CTI*s similar to the input one in the original dataset for dataset enhancement. The STL can provide for each input spatial transformation based on a *thin plate spline* (TPS) [18].

Fig. 2 shows the overall architecture of the STL in which *UM* and *VM* represent input and output *CTI* matrices (512*512*1), respectively. The *Grid Generator* and *Sampler* together constitute the STL. θ is a 25*2 tensor, and the tensor elements are randomly generated, which are normalized and input to the *Grid Generator*. The *Grid Generator* internally produces the corresponding values between every pixel point among the input and output *CTI*s. Once entered into the sampler, the corresponding points are inserted into the new matrix *UM* using the thin slice sampling interpolation transformation. Finally, after spatial transformation, the output matrix *VM* is obtained.

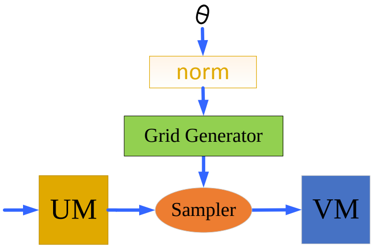


Fig. 2. The overall architecture of the STL

3.3 Similarity Calculator

In this subsection, we focus on the study of the similarity calculator including two calculators: 1) the SSC, and 2) the DSC.

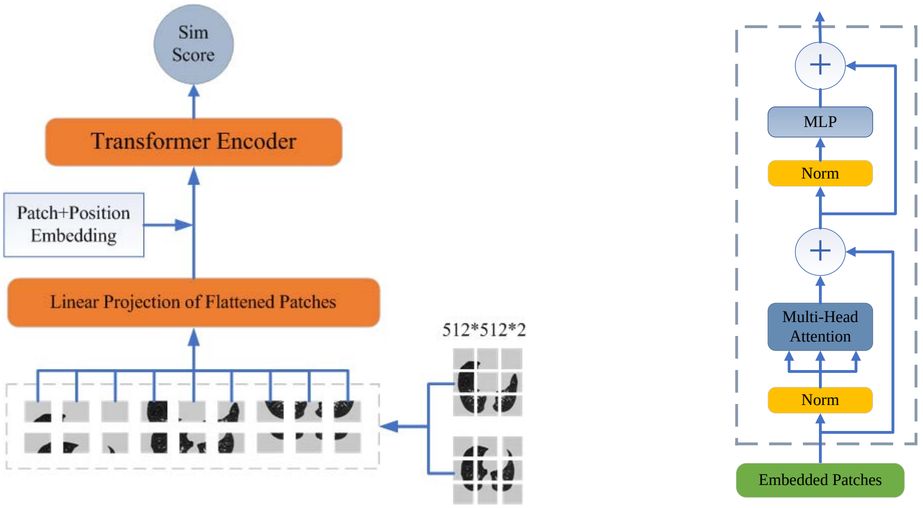
3.3.1 The SSC

The aim of the SSC is to calculate the shape similarity of the two input lung *CTI*s without considering the interior structure of the lungs.

Definition 1. *Given two CTIs: CTI_i and CTI_j, their corresponding shape similarity(Sim_S) can be defined in Eq.(1):*

$$Sim_S = SSM(CTI_i, CTI_j) \tag{1}$$

where $SSM(CTI_i,CTI_j)$ is a function for the shape similarity measurement of CTI_i and CTI_j.



(a). Overview framework of the VT

(b). Transformer encoder

Fig. 3. The network architecture for the VT

Note that, the function $SSM(CTI_1,CTI_2)$ in Definition 1 is defined based on a deep learning network to be introduced below. The scale sizes of the above two *CTI*s are 512*512*1.

First of all, as shown in **Fig. 4**, the lung *CTI*s are divided into several image patch(*IP*)s. Based on the continuity between different *IP*s of the lung *CTI*s and the diversity of lung contour shape, there is a contextual association between different *IP*s. Formally, given an input CTI_i, and its corresponding retrieved CTI_j, then we have $CTI_i = \{IP_{i1}, IP_{i2}, ..., IP_{ik}\}$, $CTI_j = \{IP_{j1}, IP_{j2}, ..., IP_{jk}\}$, where IP_{nm} means the *m*-th *IP* in the *n*-th *CTI*, and *k* is the number of the *IP*s in the *CTI*, and $m \in [1,k]$. The IP_{im} is flattened and expanded to form a vector x_{im} that is synthesized into an *IP* pair vector $x_m = (x_{im}, x_{jm})$ with the vector x_{jm}, which is patterned by IP_{jm}.

Definition 2. *Given k IP pair vectors(i.e., $(x_1, x_2, ..., x_k)$), their corresponding contextual association relationship(CAR) is derived in Eq.(2,*

$$CAR = \alpha_1^T x_1 + \alpha_2^T x_2 + ... + \alpha_k^T x_k \tag{2}$$

where α_i is the weight coefficient of the i-th vector , and x_i is a vector of the i-th IP pair.

As stated in Definition 2, some *IP*s are crucial while others are secondary according to the contextual relationship of distinct *IP*s of the lung *CTI*s. Some vectors' weight coefficients may be significantly higher than those of other vectors. Based on the above analysis, it's critically important to introduce a self-attentive mechanism [19] to filter out a small quantity of critical information from a vast quantity of information in the calculation of shape similarity. In **Fig. 3**, the *Vision Transformer* (VT) is applied as a network architecture for the SSC based on the self-attentive mechanism [20]. Note that, the transformer encoder in **Fig. 3(a)** is illustrated in **Fig. 3(b)** in detail. In the self-attentive mechanism, the more changeable lung lobe shape part was assigned more attention.

3.3.2 The DSC

A set of shape similar image set *C* is obtained after the shape similarity calculation, where *C*={CTI_1, CTI_2, …, CTI_n}. The *CTI*s in *C* that included similar information on *bronchi*, *vessels*, *nodules*, and some other soft tissues inside the lung lobes of the input *CTI* had to be identified further. Since the parenchymal section of the lung is defined as a *pathological area* (PA) in CTI_i, as illustrated in **Fig. 4**, the detail similarity computation concentrates only on that part of the lung. The following is how the similarity of details is defined.

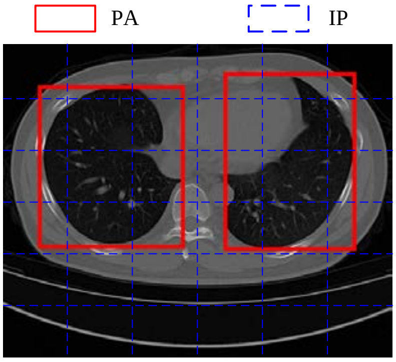


Fig. 4. An example of a *PA* in a lung *CTI*

Definition 3. *Given a lung CTI, its corresponding pathological area(PA) is modeled by a vector:*

$$PA = \{S_1, S_2, ..., S_m\} \tag{3}$$

where S_i means the i-th white streak in the lung and m is the number of white streaks in PA.

The white streaks in above definition correspond to the soft tissues inside the lung lobes of a lung *CTI*. A graph model is applied to describe the while streak in a *CTI* because the white streaks possess different shape features.

Definition 4. *Given a white streak S_i, it is modeled by a graph: $S_i = \{V, E\}$, where*

- *V represents a set of vertices in S_i, and $V = \{v_1, v_2, ..., v_{|V|}\}$, where v_i is the i-th pixel value;*
- *E refers to a collection of edges, and $E = \{e_1, e_2, ..., e_{|E|}\}$, in which $e_k =< v_i, v_j >$ means that v_i and v_j are connected.*

Based on Definition 4, two pixel points are considered to be adjacent if their corresponding Euclidean distance does not exceed $\sqrt{2}$. Let a vertex v_i be a search center, its corresponding search processing is performed around it to connect all neighboring ones and generate the edge collection *E*. As a result, a connected graph *S* is created, which is the streak formation process. Therefore, given two *PA*s (e.g., PA_1 and PA_2), their corresponding detail similarity of the two lung *CTI*s can be derived in Definition 5.

Definition 5. *Given two PAs(i.e., PA_i and PA_j), their corresponding similarity is measured as follows:*

$$Sim(PA_i, PA_j) = \begin{cases} 1, & if \sum_{k=1}^{N(PA_i)} \sum_{t=1}^{N(PA_j)} S_{ik} \sim S_{jt} \geq \theta * N(PA_i) \\ 0, & otherwise \end{cases} \tag{4}$$

where S_{ik} refers to the k-th streak in the PA_i, S_{jt} denotes the t-th streak in the PA_j. N(●) denotes the number of streaks in ●, and θ refers to the similarity threshold. $S_{ik} \sim S_{jt}$ indicates that two streaks are similar.

As stated in Definition 5, detail similarity is decided by the count of similar streaks in the two *PAs*. Similar streaks are determined by the position and number of all vertices that generate streaks in the graph. The *DSC* structure is given in **Fig. 5**. Since the fully connected layer at the resnet18 [21] has some translation invariance, it can be substituted for a resnet18 variant using which the DSC is built.

In this figure, a round rectangle represents a convolution operation with certain convolution kernel parameters such as convolution kernel size and number of convolution kernels. The number of convolution kernel move steps is denoted by '*NS*'. '*BS*' stands for the image's blank fill size. '*LBN*' refers to a layer for batch normalization. We adopt linear rectification(*RL*) function as activation function, and '*MPL*' represents the maximum pooling layer. The DSC is broken into five blocks: *conv*1 and *four basicblocks* in which the gradient disappearance problem in network training can be solved using the residual network. After inputting PA_1 and PA_2 to the DSC via the above five blocks, the '*Sim Score*', which represents the similarity between PA_1 and PA_2, is finally the output.

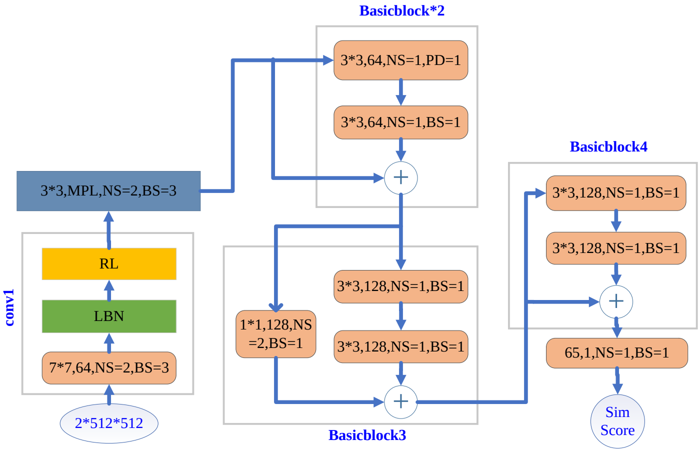


Fig. 5. The overall architecture of the DSC

4.Training

4.1 Pre-processing Step

In this subsection, as a part of the lung nodule detection dataset launched in 2016, we adopt the LUNA16 public dataset [22] with a total of 1018 cases as a training dataset, which is generated from a bigger dataset LIDC-IDRI.

For a whole lung *CTI*, there are more than just two lung lobes in it, along with other useless information. Here, we define human muscles or other soft tissues as useless information in the *CTI*. In this subsection, we study how to extract the *PA*s in the lung *CTI* and make the *WSSENet* focus on learning the characteristics of the *PA*s.

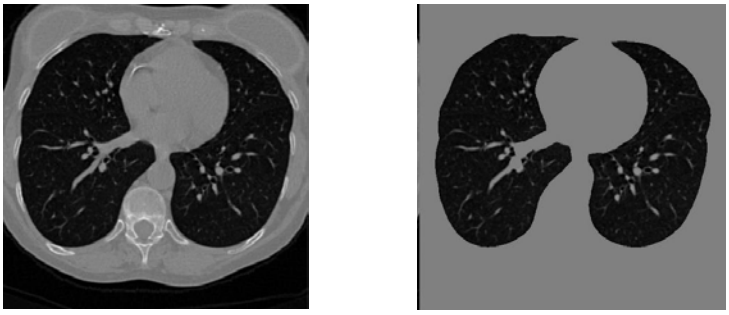
The degree of X-ray absorption by organs or tissues in the *CTI* is quantified as a CT value in HU(Hounsfield unit). To extract the *PA*, the *PA* mask in *CTI* is required to be accessed first. The *PA* mask extraction steps are as follows:

(1) In a *CTI*, different CT values correspond to different grayscale values. The *CTI* is binarized with the CT value of *PA* as the demarcation.

(2) Erase the boundary information of the binarized *CTI*.

(3) The soft tissue information within the *PA* is removed utilizing the closure operation in morphology to get a *PA* mask.

Fig. 6 illustrates the image effect before and after preprocessing.



(a). Before (b). After

Fig. 6. Comparison between before and after image pre-processing

4.2 Training Set and Test Set

A. The SSC dataset

First, the *STL* is used to construct the training set required for training the *SSC*. The *STL* transforms the input CTI_1 (512*512*1) slightly to generate CTI_2 (512*512*1), and combines CTI_1 and CTI_2 into a 2*512*512 tensor with the label 1 (similar). After that, we also have to find the image pairs that are not similar in shape. A CT scan case is composed of hundreds of layers of sections, if the level of the section changes significantly, then the shape of the lung in the section will also change obviously. Based on this, CTI_3, which has a significant change in the number of layers from CTI_1, can be found in the same case. CTI_1 and CTI_3 are a pair of dissimilar images, and they are synthesized as a 2*512*512 tensor with a label of 0 (dissimilar). The *CTI* pairs with similar and dissimilar shapes identified by the proposed method are given in **Fig. 7** and **Fig. 8**, respectively.

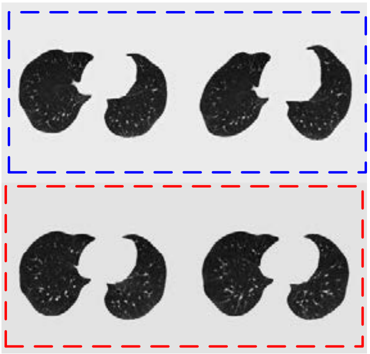


Fig. 7. Shape similar *CTI* pairs

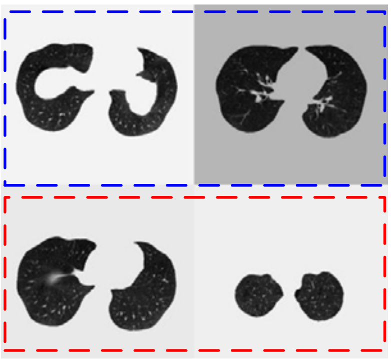


Fig. 8. Shape dissimilarity *CTI* pairs

B. The DSC dataset

After the *SSC* is trained, it can be used to create a training set for the *DSC*. Given CTI_1, the *SSC* is utilized to randomly find CTI_2, an image in the dataset with a similar lung lobe shape to CTI_1, synthesize CTI_1 and CTI_2 into a 2*512*512 tensor, and label it as 0 (not similar). Then a slice *CTI* (i.e., CTI_3) adjacent to CTI_1 was found in the case of CTI_1, synthesized CTI_1 and CTI_3 as a 2*512*512 tensor and labeled as 1 (similar).

The aforementioned method has an error in finding image pairs with dissimilar details: the *SSC* can probably obtain a *CTI* with similar details to the original *CTI* and compose an image pair of these two, incorrectly labeling them as 0(not similar). In order to lessen the erroneous training labels brought by SSC, while allowing DSC to learn more details of similarities within the lung lobes, it is required to make the quantity of similar image pairs in the training set larger than the quantity of dissimilar image pairs, and we tune the ratio of these two quantities to 3:1.

The similar and dissimilar lung *CTI* pairs that were found by the aforementioned method are illustrated in **Fig. 9** and **Fig. 10**, respectively.

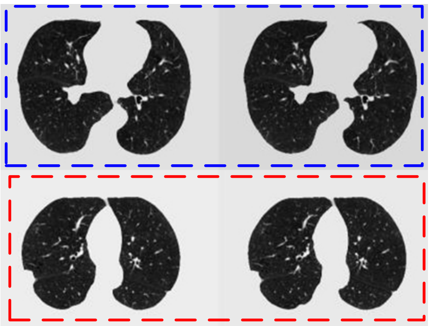


Fig. 9. Detail similar *CTI* pairs

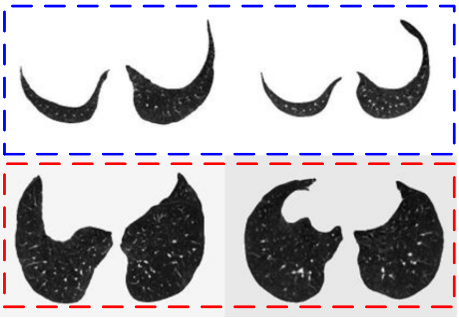


Fig. 10. Detail dissimilarity *CTI* pairs

4.3 Loss Function

In the task of lung *CTI*s similarity calculation, the goal of the *WSSENet*-based similarity measure of lung *CTI*s is essentially a classification task into one category. In this subsection, we design a cross-entropy-based loss function represented in Eq.(5) to assess the difference between the predicted and the ground truth. Note that, the *sigmoid* function ($g(x)$) is used as an activation function for training the similarity calculators (i.e., the SSC and the DSC)..

$$Loss(S(x),\sigma) = \frac{1}{n}\sum_{x\in|B|} -\sigma\log[g[S(x)]] - (1-\sigma)log[1-g[S(x)]] \tag{5}$$

where

- $|B|$ means the batch size in the network training;
- $S(x)$ denotes the similarity score of the output in the network, and $S(x)\in[0,1]$;
- σ refers to the ground truth, and $\sigma\in\{0,1\}$;
- $g(x) = \dfrac{1}{1+e^{-x}}$.

5. Experiments

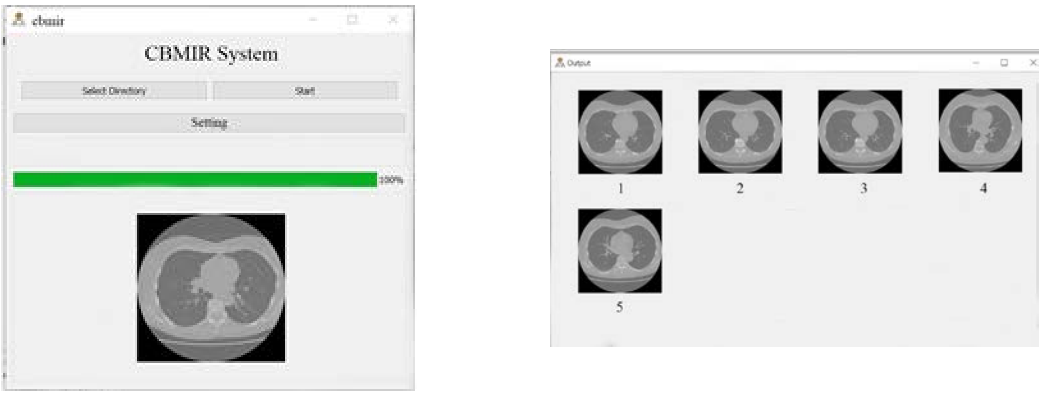
To verify the retrieval performance of our proposed *WSSENet*, in this section, we conduct comprehensive empirical study in practical scenarios.

5.1 Experiment Setup

Our prototype system was implemented using the PyTorch library with an NVIDIA 1080Ti GPU. The platform was equipped with an Intel i7-11400F CPU, 16 Gigabyte RAM, and a 4 Terabyte hard disk.

Datasets. The dataset comes from the LUNA16 dataset [22] containing 44522. In this dataset, there were 179 pulmonary CT cases. The average number of lung *CTIs* in each case was 249.

Algorithms Evaluated. For comparative evaluation, the *WSSENet*-based retrieval method is compared with four competitors, including two CNN-based hashing methods: the CNNSH [16] and the DSH [15]. Two unsupervised methods: the Locality Sensitive Hashing (LSH) and the SIFT-BoVWs [3].



(a). Submission interface (b). Result *CTIs*

Fig. 11. The prototype system demo

5.2 A Prototype System

Fig. 11 illustrates our prototype system. **Fig. 11(a)** is the submission interface in which the ‘Setting’ button changes the value of *k* in the Top-*k* retrieval. The right side window contains the output *k* similar *CTI*s. A simple *Top-k* retrieval algorithm is designed in the system. The algorithm uses the *WSSENet* as the similarity evaluation function. After inputting a *CTI* into the system, the system computes the similarity with the all *CTIs* in the database and finally outputs the *k CTIs* with the highest similarity which is shown in **Fig. 11(b)**.

5.3 Evaluation of Precision

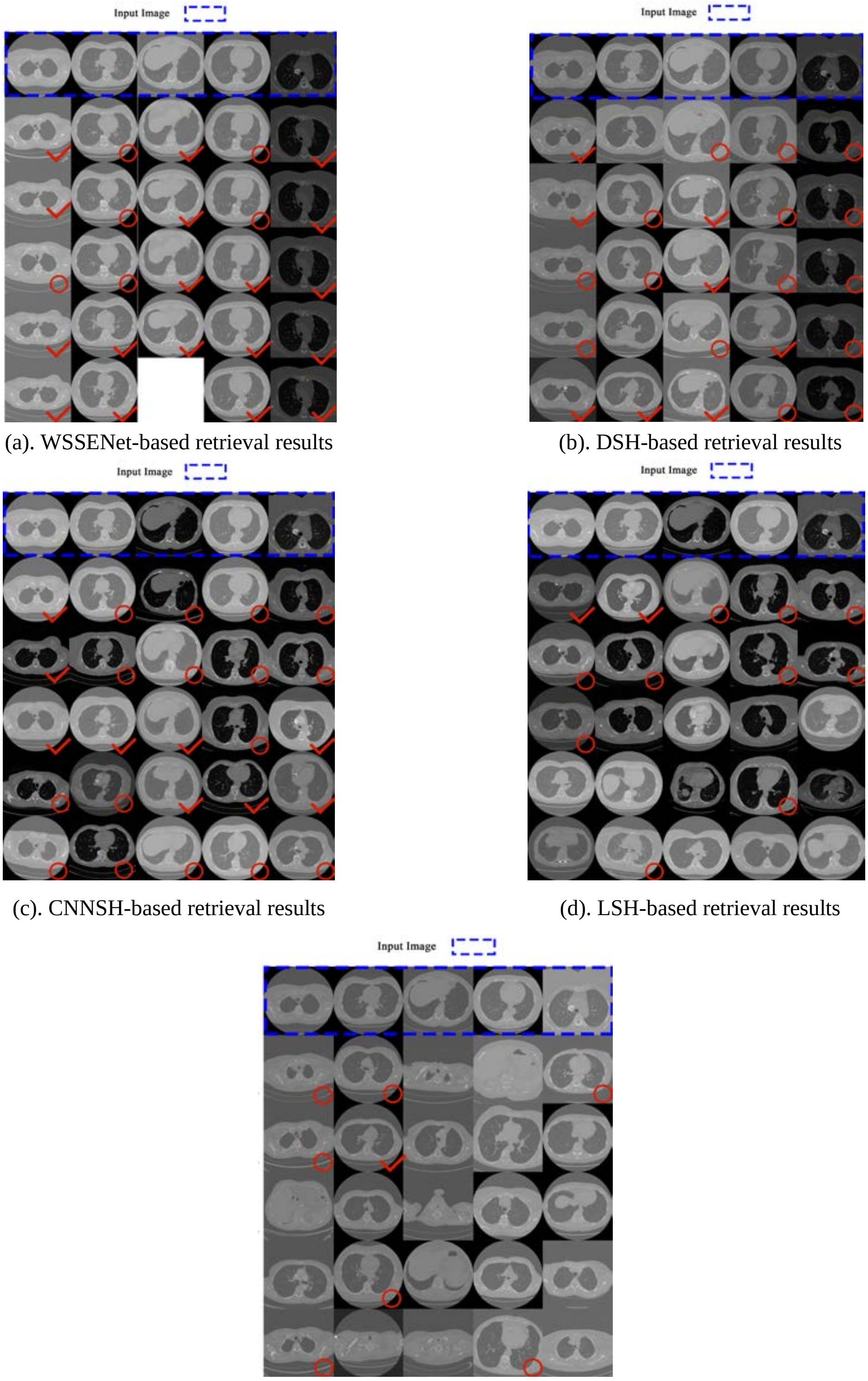
In this subsection, we empirically validate the accuracy of the proposed network on similarity evaluation by the *precision* of the Top-*k* retrieval.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

where *TP* and *FP* denote the number of correct and incorrect *CTI*s as the outputs, respectively.

To begin with, for the Top-*k* retrieval, **Fig. 12** shows retrieval examples for different retrieval methods when *k* is 5 in which every column denotes one retrieval. The input lung *CTI* is illustrated in the 1st row in this figure, and the following 5 rows are the 5 retrieved *CTI*s. In **Fig. 12(a)**, there are only 4 outputs in the 3rd column, indicating that the system retrieved only 4 similar *CTI*s in the database. **Figs. 12(b-e)** represent the result *CTI*s of the DSH, the CNNSH, the LSH and the SIFT-BoVWs, respectively. A red circle in the retrieved result means that the result is similar to the shape of the input image, and a red tick indicates that the result is similar to the input image in shape and detail. If only the shape similarity is concerned, the accuracy of the *WSSENet*-based method can reach 95%, while the DSH-based method is 92%, the CNNSH-based method is 100%, the LSH-based method is 48%, and the SIFT-BoVWs method is only 32%. If both of the shape and detail similarities are considered, the accuracy of the *WSSENet*-based method is 72%, while the DSH-based method is 32%, the CNNSH-based method is 36%, the LSH-based method is 8%, the SIFT-BoVWs method is only 4%. Therefore, for the retrieval by similar shapes, the proposed *WSSENet*, the DSH and the CNNSH have quite high retrieval precisions, while the retrieval precisions of the LSH and the SIFT-BoVWs are much lower than the former three. The *WSSENet* achieves the highest precision compared to other retrieval methods based on the shape and detail similarity.

Next, when *k* in the Top-*k* retrieval equals to 10, the retrieval accuracies of the above five methods are empirically compared in 50 retrievals. The specific retrieval accuracies and their corresponding statistics of the five methods are shown in **Fig. 13** and **Fig. 14**, respectively. In **Fig. 13(a-b)** and **Figs. 14(a-b)**, the *WSSENet* only considers the retrieval precision of similar shape, while the *WSSENet+* also considers the similar precision of details at the same time, and the rest of the methods are the same. Three metrics (i.e., *max*, *min* and *avg*) of the retrieval accuracy are provided. It is observed that if only the shape similarity is considered, the retrieval precisions of the deep learning-based retrieval methods (i.e., the *WSSENet*, the DSH, and the CNNSH) are significantly higher than that of the LSH and the SIFT-BoVWs. The average and minimum precisions of the *WSSENet* are slightly higher than the DSH and the CNNSH. If both of the shape and detail similarities are considered, the retrieval precision of the *WSSENet* is significantly better than the other four methods in all three metrics.



(a). WSSENet-based retrieval results

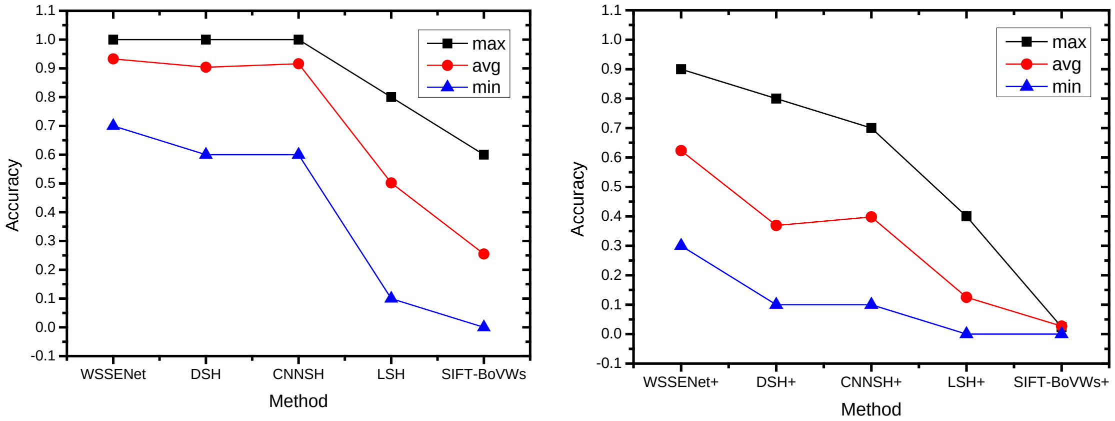
(b). DSH-based retrieval results

(c). CNNSH-based retrieval results

(d). LSH-based retrieval results

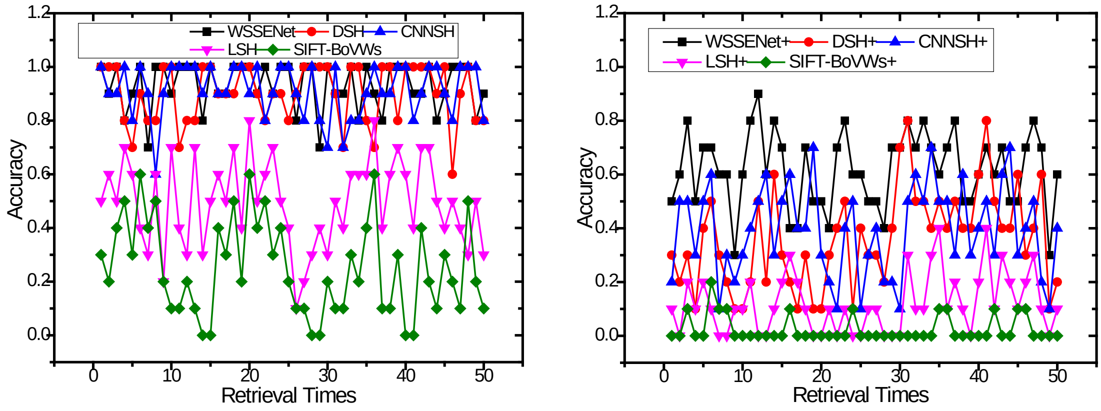
(e). SIFT-BoVWs-based retrieval results

Fig. 12. Five examples of the Top-5 retrievals



(a). the shape (b). the shape & the detail

Fig. 13. Effect of accuracy rate



(a). the shape (b). the shape & the detail

Fig. 14. Comparison of the accuracy rates of the 50 retrievals

5.4 Evaluation of *mAP*

To further evaluate the effectiveness of the retrieval system, the mean average precision (*mAP*) is provided that is derived as follows:

$$mAP = \frac{\sum_{i=0}^{n} AP_i}{n} \tag{7}$$

where *n* is the number of retrieval samples. The definition of *AP* is derived below:

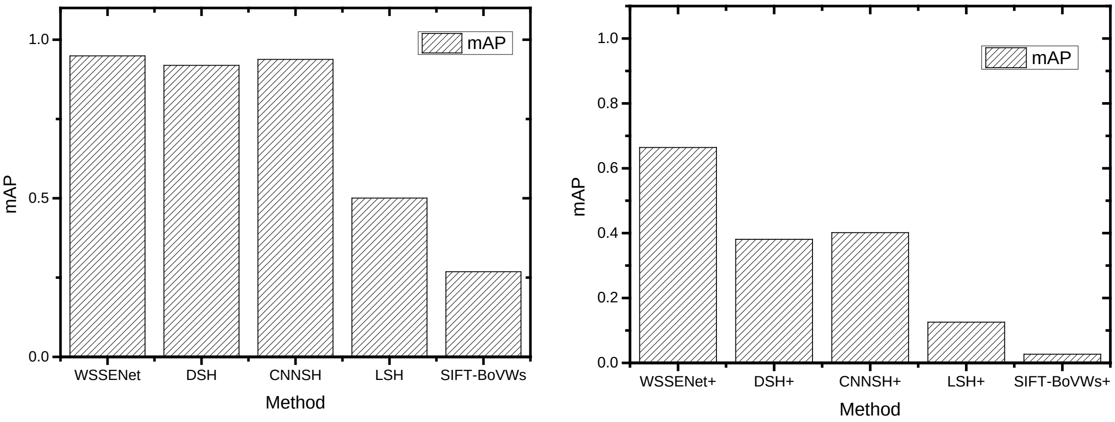
$$AP = \frac{\sum Precision_{idx}}{N} \tag{8}$$

where *idx* denotes the index of the *CTI* among all the *CTI*s retrieved, $Precision_{idx}$ is the precision until the output *CTI* with the index of *idx*, and *N* is the total amount of the outputs.

As in **Figs. 15-16**, if only shape similarity is considered, the mAP@10 of our method can reach 94.91%, while the mAP@10 of the DSH is 91.92%, the mAP@10 of the CNNSH is 93.78%, the mAP@10 of the LSH is 50.08%, and the mAP@10 of the SIFT-BoVWs is only

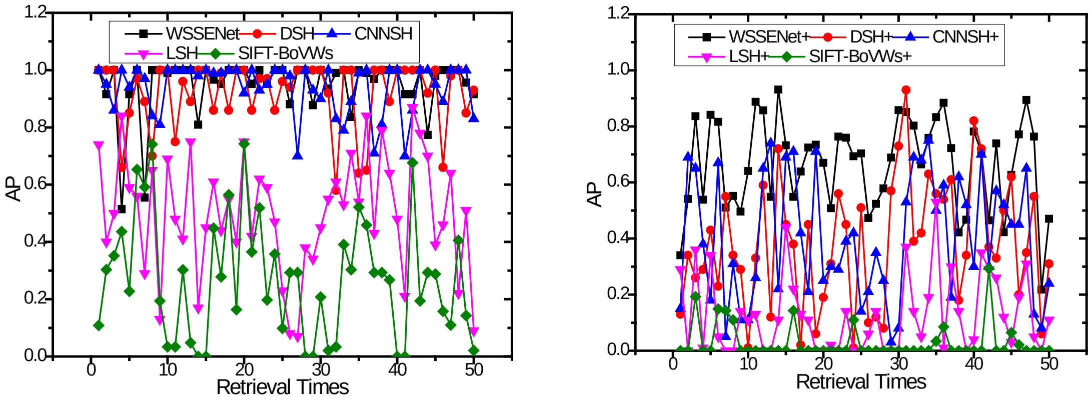
26.84%. If detail similarity is further considered, the mAP@10 of our method can reach 66.40%, while the mAP@10 of the DSH is 38.1%, the mAP@10 of the CNNSH is 40.16%, the mAP@10 of the LSH is 12.56%, and the mAP@10 of the SIFT-BoVWs is only 2.68%. Considering both shape and detail similarity as a very high precision similarity requirement, it is likely that results similar to the input *CTI* details do not exist in the database, and then it is not possible for the retrieval system to retrieve these results correctly, resulting in a lower *AP*. Therefore, further considering the detail similarity may lead to the decrease of the *mAP.*

Based on the comparison of the five methods on *AP* and *mAP*, it can be seen that the mAP of the *WSSENet*-based method is slightly better than the DSH and the CNNSH in shape similarity, and significantly better than the LSH and the SIFT-BoVWs. If further considering the similarities in detail, the mAP of the *WSSENet* is significantly better than the remaining four methods.



(a) the shape (b) the shape & the detail

Fig. 15. Mean Average Precision



(a) the shape (b) the shape & the detail

Fig. 16. Comparison of the average precision of 50 retrievals

6.Conclusion and Future Work

In this paper, we proposed the *WSSENet* model which is a new weekly supervised deep learning network for the effective similarity matching of the lung *CTI*s. The major advantage of our system over the traditional machine learning-based CBMIR systems is its weak supervision for the network's training. Furthermore, the *WSSENet*-based retrieval system achieves excellent performance through weakly supervised training. We conducted empirical experiments on the LUNA16 dataset to verify that our proposed *WSSENet* scheme can achieve the shape similarity *mAP@10* with more than 94% and both shape as well as detail similarity *mAP@10* with 66.4%, which essentially meets the demands of CBMIR. Comparing with the existing methods, the *WSSENet*-based retrieval method has obvious advantages in high precision (similar shape and details) and mAP between lung *CTI*s.

Since the *WSSENet* is a general network that can be extended to other medical images (ie.g., X-ray, MRI, etc.) for retrieval tasks. Future work may focus on reducing the matching time of the *WSSENet*-based retrieval system to improve efficiency.

References

[1] ZH. Zhou, “A brief introduction to weakly supervised learning,” *National Science Review*, 5(1), 2018, 44-53, 2018. Article (CrossRef Link)

[2] DG Lowe, "Distinctive image features from scale-invariant keypoints,” *Int’l J. of Computer Vision*, 60(2), 91-110, 2004. Article (CrossRef Link)

[3] M Maxim, et al, “Feature-based brain MRI retrieval for Alzheimer disease diagnosis,” in *Proc. of IEEE ICIP*, 2013. Article (CrossRef Link)

[4] H. Müller, J. Kalpathy-Cramer, et al, “Overview of the CLEF 2009 medical image retrieval track,” in *Proc. of Workshop of the Cross-Language Evaluation Forum for European*, pp. 72-84, 2009. Article (CrossRef Link)

[5] A. Krizhevsky, I Sutskever, and GE. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017. Article (CrossRef Link)

[6] H. Pan, P. Li, Q. Li, Q. Han, X. Feng, L. Gao, “Brain CT image similarity retrieval method based on uncertain location graph,” *IEEE J. of Biomedical and Health Informatics*, 18(2), 574-584, 2014. Article (CrossRef Link)

[7] K. Karthik, and S.S. Kamath, “A hybrid feature modeling approach for content-based medical image retrieval,” in *Proc. of 2018 IEEE 13th Int’l Conf. on Industrial and Information Systems(ICIIS)*, 2018. Article (CrossRef Link)

[8] N. Sampathila, and R J Martis, “Computational approach for content-based image retrieval of K-similar images from brain MR image database,” *Expert Systems*, vol. 39, no. 7, e12652, 2022. Article (CrossRef Link)

[9] HC. Shin, HR. Roth, M. Gao, et al, “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,” *IEEE Trans. on Medical Imaging*, 35(5), 1285-1298, 2016. Article (CrossRef Link)

[10] SK. Sundararaja, B Sankaragomathi, DS Priya, “Deep belief CNN feature representation based content based image retrieval for medical images,” *J. of Medical Systems*, 43(6), 1-9, 2019. Article (CrossRef Link)

[11] A. Khatami, et al, “A deep-structural medical image classification for a radon-based image retrieval,” in *Proc. of 2017 IEEE 30th Canadian Conf. on Electrical and Computer Engineering* (CCECE), 2017. Article (CrossRef Link)

[12] A. Khatami, M. Babaie, HR. Tizhoosh, A. Khosravi, T. Nguyen, S. Nahavandi, "A sequential search-space shrinking using CNN transfer learning and a Radon projection pool for medical image retrieval," *Expert Systems with Applications*, vol. 100, pp. 224-233, 2018. Article (CrossRef Link)

[13] L. Ma, et al, "A new method of content based medical image retrieval and its applications to CT imaging sign retrieval," *J. of biomedical informatics*, vol. 66, pp. 148-158, 2017. Article (CrossRef Link)

[14] H. Lai, et al, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. of the IEEE conf. on computer vision and pattern recognition*, 2015. Article (CrossRef Link)

[15] H. Liu, et al, "Deep supervised hashing for fast image retrieval," in *Proc. of the IEEE conf. on computer vision and pattern recognition*, 2016. Article (CrossRef Link)

[16] YH. Cai, et al, "Medical image retrieval based on convolutional neural network and supervised hashing," *IEEE access*, vol. 7, pp. 51877-51885, 2019. Article (CrossRef Link)

[17] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, "Spatial transformer networks," in *Proc. of NIPS*, pp. 2017-2025, 2015. Article (CrossRef Link)

[18] FL. Bookstein, "Principal warps: thin-plate splines and the decomposition of transformations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(6), 567-585, 1989. Article (CrossRef Link)

[19] A. Vaswani, et al, "Attention is all you need," in *Proc. of NIPS*, pp. 6000-6010, 2017. Article (CrossRef Link)

[20] A. Dosovitskiy, et al, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. Article (CrossRef Link)

[21] K. He, et al, "Deep residual learning for image recognition," in *Proc. of the IEEE CVPR*, 2016. Article (CrossRef Link)

[22] A.A.A. Setio, A. Traverso, T. de Belo, et al, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge," *Medical Image Analysis*, vol. 42, pp. 1-13, 2017. Article (CrossRef Link)

Yi Zhuang received Ph.D degree in Computer Science from Zhejiang University in May 2008. He is currently a full professor at the School of Computer & Information Engineering in Zhejiang Gongshang University where he joined as faculty member since May 2008. Dr. Zhuang is a recipient of the CCF Doctoral Dissertation Award conferred by Chinese Computer Federation in 2008 and IBM Ph.D. Fellowship 2007–2008. From January 2008 to March 2008, supported by IBM Ph.D. Fellowship, Dr. Zhuang has spent 3 months to participate in the study of an optimal hybrid storage model based on DB2 as a research intern in IBM China Research Lab. His research interests mainly focus on multimedia database and cloud computing. He has published 40+ papers in the leading journals and conferences, i.e., ACM TOIT, TALIP, Information Sciences, KAIS, and ESWA, etc.

Shuai Chen is now a master student in Computer Science at Zhejiang Gongshang University. Before that, he received the bachelor degree of Computer Science from Hangzhou Dianzi University. His research interests mainly focus on medical image processing, machine learning, etc.

Nan Jiang received the bachelor degree of medical science and the master degree of medical science both from the Zhejiang University in 2004 and 2007, respectively. Ms. Jiang is with Affiliated Hangzhou First People’s Hospital, Zhejiang University School of Medicine. Her research interests mainly focus on medical image processing. She has published 10+ papers in international conferences and journal.

Hua Hu received Ph.D degree in Computer Science from Zhejiang University in 1998. He is currently a full professor at the College of Information Engineering in Hangzhou Normal University where he joined as faculty member since 2019. Before that, Dr. Hu was a faculty member in Zhejiang Gongshang University from 1998-2008, and faculty member in Hangzhou Dianzi University from 2009-2018, respectively. His research interests mainly focus on medical image processing, machine learning, workflow, etc.