Original Article | Thoracic Imaging

eISSN 2005-8330
https://doi.org/10.3348/kjr.2020.0603
Korean J Radiol 2021;22(2):281-290

Korean Journal of Radiology
KJR

Content-Based Image Retrieval of Chest CT with Convolutional Neural Network for Diffuse Interstitial Lung Disease: Performance Assessment in Three Major Idiopathic Interstitial Pneumonias

Hye Jeon Hwang, MD, PhD, Joon Beom Seo, MD, PhD, Sang Min Lee, MD, PhD, Eun Young Kim, MD, PhD, Beomhee Park, MS, Hyun-Jin Bae, PhD, Namkug Kim, PhD

All authors: Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Objective: To assess the performance of content-based image retrieval (CBIR) of chest CT for diffuse interstitial lung disease (DILD).

Materials and Methods: The database was comprised by 246 pairs of chest CTs (initial and follow-up CTs within two years) from 246 patients with usual interstitial pneumonia (UIP, n = 100), nonspecific interstitial pneumonia (NSIP, n = 101), and cryptogenic organic pneumonia (COP, n = 45). Sixty cases (30-UIP, 20-NSIP, and 10-COP) were selected as the queries. The CBIR retrieved five similar CTs as a query from the database by comparing six image patterns (honeycombing, reticular opacity, emphysema, ground-glass opacity, consolidation and normal lung) of DILD, which were automatically quantified and classified by a convolutional neural network. We assessed the rates of retrieving the same pairs of query CTs, and the number of CTs with the same disease class as query CTs in top 1–5 retrievals. Chest radiologists evaluated the similarity between retrieved CTs and queries using a 5-scale grading system (5-almost identical; 4-same disease; 3-likelihood of same disease is half; 2-likely different; and 1-different disease).

Results: The rate of retrieving the same pairs of query CTs in top 1 retrieval was 61.7% (37/60) and in top 1–5 retrievals was 81.7% (49/60). The CBIR retrieved the same pairs of query CTs more in UIP compared to NSIP and COP (p = 0.008 and 0.002). On average, it retrieved 4.17 of five similar CTs from the same disease class. Radiologists rated 71.3% to 73.0% of the retrieved CTs with a similarity score of 4 or 5.

Conclusion: The proposed CBIR system showed good performance for retrieving chest CTs showing similar patterns for DILD.

Keywords: *Content-based image retrieval; Multidetector computed tomography; Convolutional neural network; Interstitial lung disease*

INTRODUCTION

Diffuse interstitial lung disease (DILD) is a heterogeneous group of fibrotic lung diseases with a complex variety of imaging and clinical features. Radiologic exams, especially high resolution computed tomography (HRCT), have a key role in the diagnosis and follow-up of DILD. However, the interpretation of HRCT of DILD requires much clinical experience; DILD HRCTs are interpreted qualitatively with an inherent degree of subjectivity and the various DILDs are differentiated only by subtle changes in the lung parenchyma, with complexity of the various imaging

Received: March 5, 2020 **Revised:** May 8, 2020 **Accepted:** June 3, 2020
Corresponding author: Joon Beom Seo, MD, PhD, Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea.
• E-mail: seojb@amc.seoul.kr

features and characteristic distribution within the lung. It is well known that the inter-observer agreements for the diagnosis of DILD are moderate even among experienced chest radiologists (1, 2). To overcome these limitations, the objective and reproducible assessment of DILD HRCTs using automated quantification systems have been proposed and their usefulness regarding the diagnosis, follow-up of disease or prediction of patients' survival has been reported (3-6). Recently, the deep learning based classifier was applied in the classification of the regional pattern of DILD, and it showed good performance (7).

The content-based image retrieval (CBIR) system is an image search engine with tools for classifying, indexing, and retrieving images from the database that depict similar imaging appearances. In the application of CBIR in DILD, the automated quantification and classification of regional disease patterns of DILD in HRCT can be integrated into the measurement of the similarity index. As the interpretation of HRCT of DILD requires a lot of experience, the searching for similar CT images from the CT database with confirmed diagnosis based on the objectively measured similarity index would be helpful for radiologists in accurately diagnosing DILD and reducing reading time. By presenting similar CT images, rather than presenting the scores of the quantified results, CBIR is a more intuitive way to identify similar CT images. CBIR system for HRCT of DILD is also in accordance with the actual clinical diagnostic workflow of radiologists searching for similar CT images in textbooks, journals, and personal databases. However, until now the CBIR system has rarely been applied to chest CT imaging, including HRCT of DILD (8-10).

We have developed a fully-automated CBIR by incorporating the deep convolutional neural network (CNN)-based pattern classification for CT images of DILD. In this study, we aimed to assess the performance of our CBIR system for retrieving similar CTs as query CTs in DILD, and visually assess the similarities between the retrieved CTs and the query CTs.

MATERIALS AND METHODS

This retrospective study was approved by the Institutional Review Board of Asan Medical Center which waived the requirement for patient informed consent.

CT Database for CBIR

Altogether, 492 HRCTs from 246 patients with DILD were collected for the CBIR database, which included patients with three different disease classes, including 100 patients with usual interstitial pneumonia (UIP), 101 patients with nonspecific interstitial pneumonia (NSIP), and 45 patients with cryptogenic organic pneumonia (COP). The 492 HRCTs consisted of 246 pairs of HRCT which contained the initial and follow-up (within two years) HRCTs of the same patient. The 101 patients were diagnosed based on histologic results of surgical biopsy and 145 patients were diagnosed through a multidisciplinary approach. Patients with collagen vascular disease, or occupational lung disease were not included in this study. HRCT scans were performed in both supine and prone positions at full inspiration without contrast materials, and CT images obtained in the supine position were only used for analysis (Supplementary Materials). A radiologist (6-years of experience) reviewed all of the pairs of HRCTs in order to evaluate the stability of the parenchymal abnormalities of DILD and rated them using four types of score (0, exactly same; 1, same; 2, similar; and 3, progress or improvement of disease observed).

Development of CBIR for DILD

Lung segmentation on HRCT was performed using our method with deep CNN (11). To classify the regional disease patterns of DILD in HRCT, we utilized our two-dimensional fully automated DILD classifier using CNN (7) (Supplementary Materials). HRCT datasets not included in the database used in this study were used for the development of the CNN algorithm for lung segmentation of HRCT and classification of regional disease patterns of DILD on HRCT in previous studies (7, 11). After applying lung masks, we performed a pixel-level disease pattern classification using the DILD classifier and each slice of HRCT scan was classified into six classes of DILD disease patterns (honeycombing, reticular opacity, emphysema, ground-glass opacity, consolidation and normal lung) (Fig. 1). Therefore, we obtained the regional fraction of DILD disease patterns in each slice of HRCT scans.

To quantify the spatial distribution of the disease patterns, we divided each x-, y-, and z-axis of the lung volume in HRCT into a quarter at the regular intervals, and thus divided the whole lung volume into 64 cuboids (4x x 4y x 4x). We calculated the fraction of each of the six DILD patterns in each cuboid. As a result, each CT scan had 384 features (64 cuboids x 6 image patterns) thus comprising a unique feature for each CT scan (Fig. 2). Similarity was calculated from the Euclidian distance among the feature

https://doi.org/10.3348/kjr.2020.0603

Korean Journal of Radiology
KJR

vectors of the CT images to be compared. For example, the most similar images would have the least distance between their feature vectors, and vice versa.

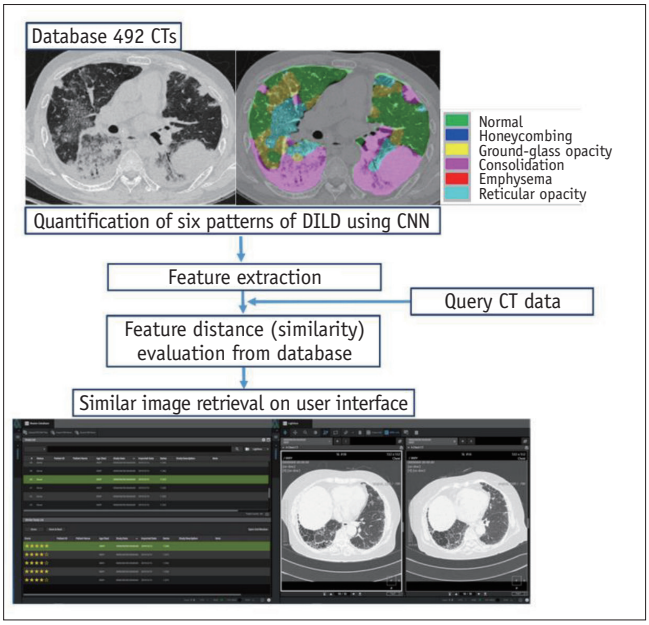

Fig. 1. Schematic flow diagram of our CBIR system. Lung segmentation on HRCT was automatically performed and six image patterns (honeycombing, reticular opacity, emphysema, ground-glass opacity, consolidation and normal lung) of DILD were quantified by the CNN. Feature extraction was performed by calculating the fraction of each of the six DILD patterns (Fig. 2) in all the CTs from the database. After calculation of the feature distance (similarity) from the Euclidian distance among the feature vectors of query CT and CTs of the database, the CBIR system retrieved five similar CTs based on the calculated similarity in a descending order. It displayed query CT image and retrieved CT images one-by one and side-by-side on the user interface. CBIR = content-based image retrieval, CNN = convolutional neural network, DILD = diffuse interstitial lung disease, HRCT = high resolution computed tomography

The feature extraction was performed in all CTs of the database through the above process, and the results were indexed in CBIR system. Then, given the query CT given to the CBIR system, the top five similar CTs were retrieved from the database in descending order by calculating similarities with the query CT, and displayed the user interface (Fig. 1).

Assessment of Retrieving Accuracy

For assessment of the accuracy of CBIR in retrieving similar CTs, we defined similar CTs as the pairs of HRCTs with stable parenchymal abnormalities. The query cases were selected from the 132 pairs of HRCTs with stability scores of 0 to 1, and sixty cases (30 UIP cases, 20 NSIP cases, and 10 COP cases; 6 cases with score 0 and 54 cases with score 1) were used for the queries. For a given query HRCT, the top 1–5 similar CTs as a given query in descending order were retrieved from the database of 491 HRCTs (except a given query HRCT from the database of 492 HRCT) by comparing the calculated similarity by CBIR. For assessing the retrieval accuracy of CBIR, we assessed the rates of retrieving the same pairs of a query CT, and the number of CTs which were classified with the same disease class as a query CT in recalled top 1–5 CTs. The performance of CBIR was also evaluated according to the disease patterns.

Visual Similarity Assessment

The visual similarity assessment of the retrieved five CTs and the query CT was independently performed by

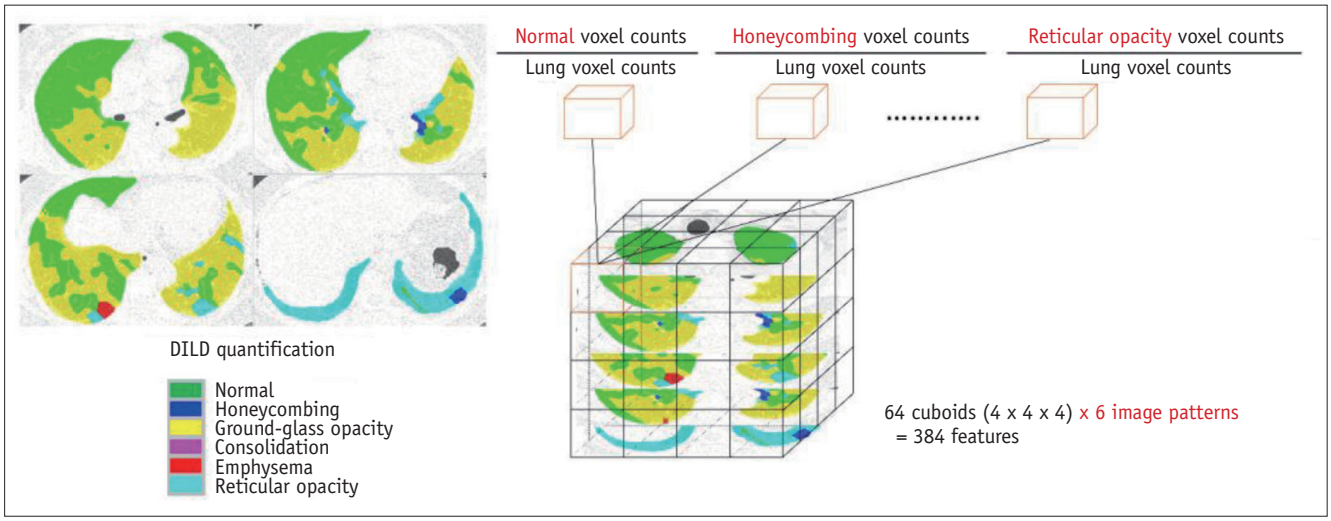

Fig. 2. Feature extraction of the HRCT patterns with DILD for similarity calculation. After dividing each three-dimensional lung volume of a CT scan into 64 cuboids (4x x 4y x 4x), the fraction of each of the six DILD patterns in each cuboid was calculated. As a result, each CT scan had 384 features (64 cuboids x 6 image patterns), which comprise a unique feature for each CT scan.

https://doi.org/10.3348/kjr.2020.0603

three chest radiologists (with 25, 15, and 12 years of experience, respectively). The query CT and retrieved CTs were displayed and compared one-by-one and side-by-side using the dedicated in-house interface (Coreline Soft) (Fig. 1). The five retrieved CTs were displayed in random order next to the query CT, not in the order of the calculated similarity. Radiologists were blinded to the CT's diagnosis, and patients' information. The similarity was subjectively graded using the 5-scale scoring system. The definition of the similarity score is described in Table 1.

Statistical Analysis

The rates of retrieving the same pairs of query CTs and the number of CTs with the same disease classes of DILD as the query CT in the retrieved top 1–5 CTs were compared among the three disease classes using Pearson's chi-square test. Similarity scores were compared among the top 1–5 CTs and among three disease classes using the Kruskal-Wallis test. Post hoc analysis with Mann-Whitney test was performed. Bonferroni's correction was performed for multiple comparisons. The statistical analyses were performed using SPSS version 21.0 (IBM Corp.). Inter-reader agreement of three radiologists for similarity scores of the retrieved CTs were assessed using the weighted k statistics in MedCalc version 19.2.0 (MedCalc Software). The k values were interpreted as slight (0–0.20), fair (0.21–0.40), moderate (0.41–0.60), good (0.61–0.80), or excellent (0.81–1.0) (12). A p value < 0.05 was considered to indicate a statistical significance.

RESULTS

Retrieving Accuracy of CBIR System

The rate of retrieving the same pairs of query CTs was 61.7% (37/60) in the top 1 retrieval, 76.7% (46/60) in the top 1–3 retrievals, and 81.7% (49/60) in the top 1–5 retrievals (Table 2, Fig. 3). Our CBIR system retrieved 93.3% of the top 1 retrievals from the same disease class as the query CT, and it retrieved an average of 4.17 of five retrieved similar CTs from the same disease class (83.2%). The CBIR system performed better in CTs with the UIP pattern than in CTs with other disease patterns. Of the 30 queries with the UIP pattern, CBIR retrieved the same pairs of the 24 query CTs in the top 1 retrieval (80.0%) and the same pairs of the 29 query CTs in the top 1–5 retrievals (96.7%). The retrieving rate of the same pairs of the query CTs in the top 1–5 was significantly higher in UIP cases (96.7%) compared to NSIP (70.0%) and COP (60.0%) cases (p = 0.008 and 0.002). The retrieving rate of the same pairs of the query CTs in the top 1–5 of the NSIP and COP cases were not significantly different (p = 0.584). The rate of retrieving CTs with the same disease class as the query CT in the top 1–5 were also significantly higher in UIP (88.7%) and NSIP (90.0%) cases than in COP (54.0%) cases (UIP vs. COP, p < 0.001 and NSIP vs. COP, p < 0.001). The CBIR system always retrieved the CTs with the same disease class as the query CT for the top 1 retrieval in cases with UIP or NSIP patterns (30/30 for the UIP pattern and 20/20 for the NSIP pattern). However, in the COP pattern, it retrieved the CTs with the same disease group as a top 1 retrieval in six of 10 queries. The CBIR system retrieved three CTs with the NSIP pattern and one CT with the UIP pattern as top 1 retrievals in the 10 queries with the COP pattern.

Similarity Scores by Three Thoracic Radiologists

Three radiologists rated 71.3% to 73.0% of the retrieved CTs with a similarity score of 4 or 5 (Figs. 4, 5). They rated 86.7% to 90.0% of top 1 retrieved CTs with a similarity score of 4 or 5 (Table 3). The mean similarity scores of the top 1–5 retrieved CTs were 3.88 ± 0.98 overall, which is close to a score of 4, thus indicating that "it is classified as the same disease, subjectively." The similarity score was

Table 1. Definition of Similarity Score

Score 5	The presence of the regional disease patterns (honeycombing, reticulation, ground-glass opacity, consolidation, emphysema, normal looking lung), as well as the regional distribution and the regional extent are almost identical.
Score 4	The presence of the regional disease patterns is almost identical, although there are differences in the regional distribution and the regional extent. However, it is subjectively classified to be the same disease.
Score 3	Some of the regional disease patterns are identical. The regional distribution or the regional extent differ. Radiologically, the likelihood of the same disease is approximately half.
Score 2	There are two inconsistencies in the regional disease patterns, the regional distribution, and the extent of disease patterns. It is more likely to be a different disease than the same disease.
Score 1	The regional disease patterns, the regional distribution, and the regional extent are different and appear to be a different disease.

https://doi.org/10.3348/kjr.2020.0603

Korean Journal of Radiology
KJR

Table 2. Retrieval Accuracy of the Content-Based Image Retrieval System

Measurement	Total (n = 60)	UIP (n = 30)	NSIP (n = 20)	COP (n = 10)
Presence of the same pair (%)				
Top 1	37 (61.7)	24 (80.0)	10 (50.0)	3 (30.0)
Top 1–3	46 (76.7)	27 (90.0)	14 (70.0)	5 (50.0)
Top 1–5*	49 (81.7)	29 (97.0)	14 (70.0)	6 (60.0)
The number of CTs with the same disease class in recalled top, n (%)				
Top 1	56 (93.3)	30 (100.0)	20 (100.0)	6 (60.0)
Top 1–3	154 (85.6)‡	82 (91.1)	54 (90.0)	18 (60.0)
Top 1–5†	250 (83.3)§	133 (88.7)	90 (90.0)	27 (54.0)
Presence of the same class in the recalled top, n (%)				
Top 1	56 (93.3)	30 (100)	20 (100)	6 (60.0)
Top 1–3	60 (100)	30 (100)	20 (100)	10 (100)
Top 1–5	60 (100)	30 (100)	20 (100)	10 (100)

*Pearson's chi-square test for the presence of the same pair in top 1–5 among the three disease classes: UIP vs. NSIP vs. COP, $p = 0.009$; UIP vs. NSIP, $p = 0.008$; UIP vs. COP, $p = 0.002$; and NSIP vs. COP, $p = 0.584$, †Pearson's chi-square test for the number of CTs with the same disease class in top 1–5 among the three disease classes: UIP vs. NSIP vs. COP, $p < 0.001$; UIP vs. NSIP, $p = 0.455$; UIP vs. COP, $p < 0.001$; and NSIP vs. COP, $p < 0.001$, Significance level of 0.0167 takes into account the Bonferroni's correction for Pearson's chi-square test, ‡The number of denominator is changing to 180 (= 60 x 3) for top 1–3 due to the category of measurement, §The number of denominator is changing to 300 (= 60 x 5) for top 1–5 due to the category of measurement. COP = cryptogenic organizing pneumonia, NSIP = nonspecific interstitial pneumonia, UIP = usual interstitial pneumonia

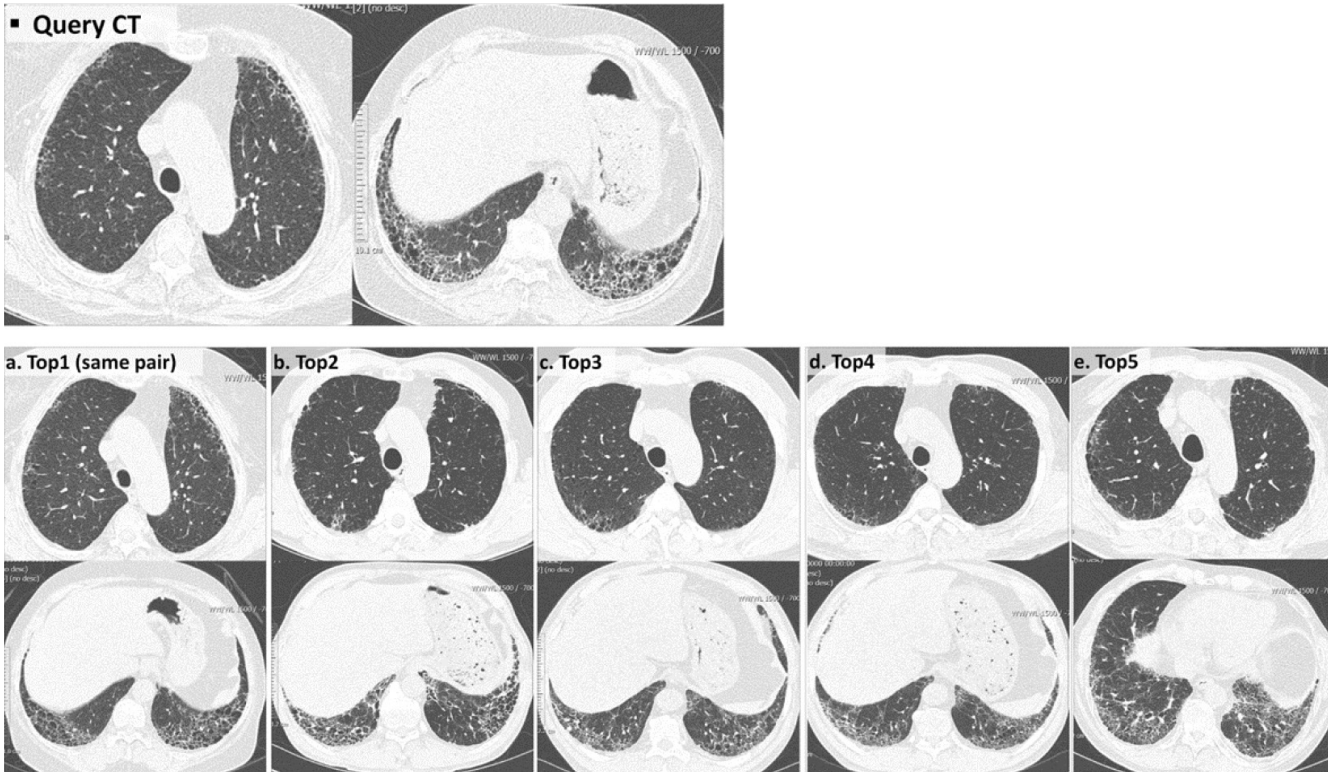

Fig. 3. Example of retrieval of HRCTs similar to query CT using our CBIR system. The query CT shows lower lobe predominant reticulation, minimal honeycombing and mild ground-glass opacities with traction bronchiectasis in the subpleural areas of both lower lobes and both upper lobes. (a-e) Images represent the retrieved five similar CTs based on the measured similarity in descending order using the CBIR system. (a) Our CBIR system retrieved the pair (follow-up CT) of query CTs as the top 1 retrieval. (b-e) CT images of top 2 to top 5 retrievals also show quite similar CT patterns and distribution of the pulmonary parenchymal abnormalities of DILD. The query case and retrieved cases were diagnosed as DILD with UIP pattern. In the visual similarity assessment, two chest radiologists rated the similarity scores of the top 1 to 5 as 5, 5, 4, 4, and 4, respectively, and the other chest radiologist as 5, 5, 5, 4, and 5. UIP = usual interstitial pneumonia

https://doi.org/10.3348/kjr.2020.0603

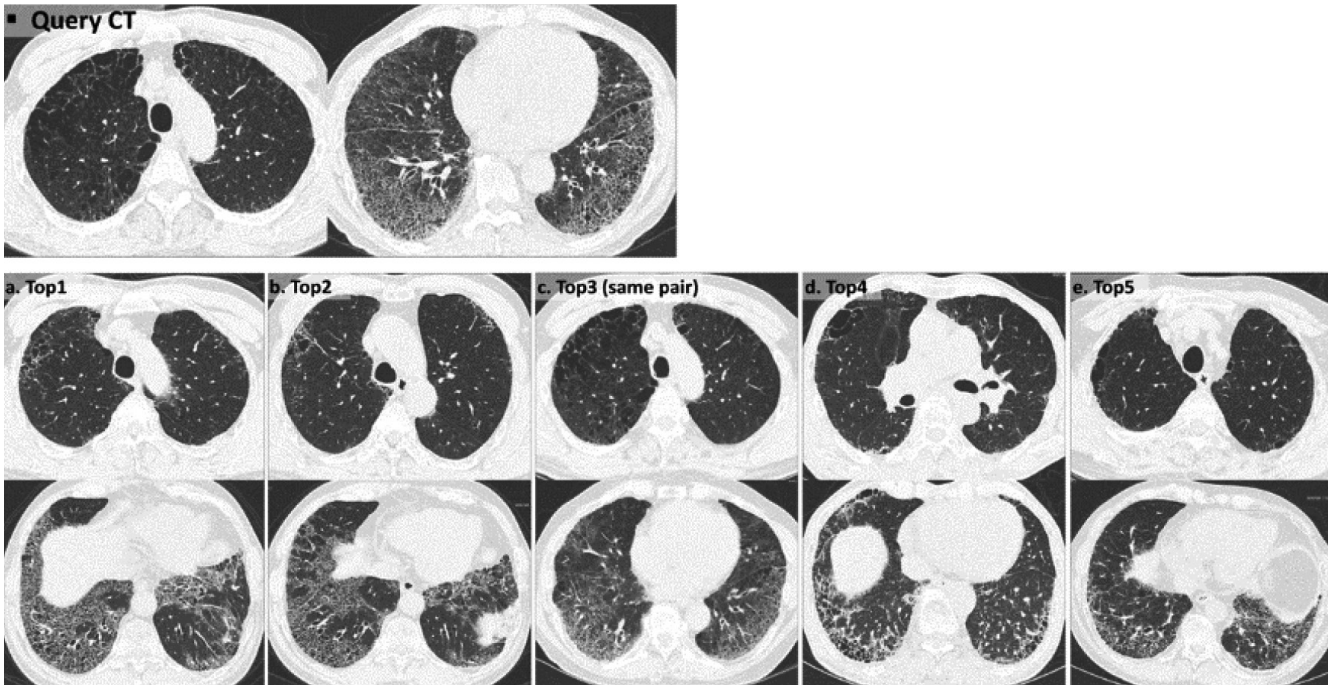

Fig. 4. Example of retrieval of HRCTs similar to query CT using our CBIR system. The query CT shows upper lobe dominant centrilobular and paraseptal emphysema and lower lobe predominant diffuse ground-glass opacity, reticulation, and mild traction bronchiectasis in both lungs which suggest DILD with a desquamative interstitial pneumonia pattern or atypical UIP pattern or NSIP pattern. The query case was confirmed as UIP pattern with clinical data, HRCT features, and histologic results. (a-e) Among five retrieved similar CTs, our CBIR system retrieved the pair (follow-up CT) with query CT as (c) the top 3 retrieval. CT images of the retrieved cases show quite similar distribution of upper lobe predominant emphysema and lower lobe predominant ground-glass opacity and reticulation. All retrieved cases were diagnosed as having a UIP pattern. For visual similarity scores, two chest radiologists rated the top 1 to 5 searches as 4, 4, 5, 4, and 4, respectively, and the other chest radiologist as 5, 4, 5, 4, and 4. NSIP = nonspecific interstitial pneumonia

different among the top 1–5 retrievals ($p < 0.001$); it was highest in the top 1 retrieval, followed by the top 2, top 3, top 5, and top 4 retrievals. The similarity score differed significantly among three disease classes ($p < 0.001$), and it was highest in the UIP cases followed by the NSIP pattern and the COP pattern (Table 4). In post-hoc analysis, the similarity score was significantly higher in the UIP cases compared to the COP cases in two readers and compared to the NSIP cases in two readers ($p \leq 0.001$, respectively). Observer agreement for similarity score is shown in Table 5. The weighted k values for similarity scores were moderate.

DISCUSSION

Our CBIR system enables retrieval of similar CT images by objective inter-case similarity assessment of DILD CTs. The CBIR system retrieved 4.17 of five similar CTs from the same disease class (83.2%). It retrieved the same pairs of query CTs as top 1 retrieval in 37 of 60 queries (61.7%) and as top 1–5 retrievals in 49 of 60 queries (81.7%). Thoracic radiologists rated 71.3% to 73.0% of the retrieved CTs with a similarity score of 4 or 5 (scores indicating 'classified to be the same disease' or 'almost identical,' respectively).

A large amount of medical imaging data is stored in the picture archiving and communication system (PACS); however, searching for similar images from PACS is challenging. Thus, there have been attempts to apply the concept of CBIR to medical imaging (13-18). However, CBIR has been rarely applied to chest CT (8, 13). In the study by Depeursinge et al. (8), the three-dimensional-based automated categorization of the lung in HRCT was applied in their CBIR, and retrieving of similar cases was enabled based on the proportions of categorized lung as well as clinical information. In this study by Aisen et al. (13), they investigated whether "automated search and selection engine with retrieval tools" can aid radiologists in the interpretation of chest CT images. However, the number of cases in the database of previous studies (128 cases and 173 cases, respectively) were too small to assess the performance of CBIR. Besides, the similarity between retrieved CTs and query CTs was not precisely evaluated.

It is difficult to evaluate the performance of CBIR for

https://doi.org/10.3348/kjr.2020.0603

Korean Journal of Radiology
KJR

retrieving similar CTs because assessing the similarity of CT images is conceptual and prone to subjectivity. To overcome these difficulties, we designed a multi-level performance assessment of CBIR. First, we prepared HRCT pairs with DILD comprised of initial and follow-up HRCTs. Our CBIR showed good performance for retrieving the same pairs of query CTs from the database. Secondly, we tested if the CBIR retrieved the same disease cases, which were confirmed by multidisciplinary diagnosis. The CBIR retrieved 4.17 of the five retrievals (83.3%) from the same disease class as queries on average. Therefore, this system can be considered to be able to accurately search CTs similar to the query CT.

Although our CBIR system showed good performance in retrieving CTs with the same disease class as the query CTs by comparing the objectively quantified similarity, it is also important whether the radiologists accept the retrieved CTs as similar images with queries. In our study, thoracic radiologists assessed the CTs retrieved by CBIR as quite similar to query CTs and rated them as being classifiable to the same disease class as the query CTs. The similarity scores given by the radiologists were also in agreement with

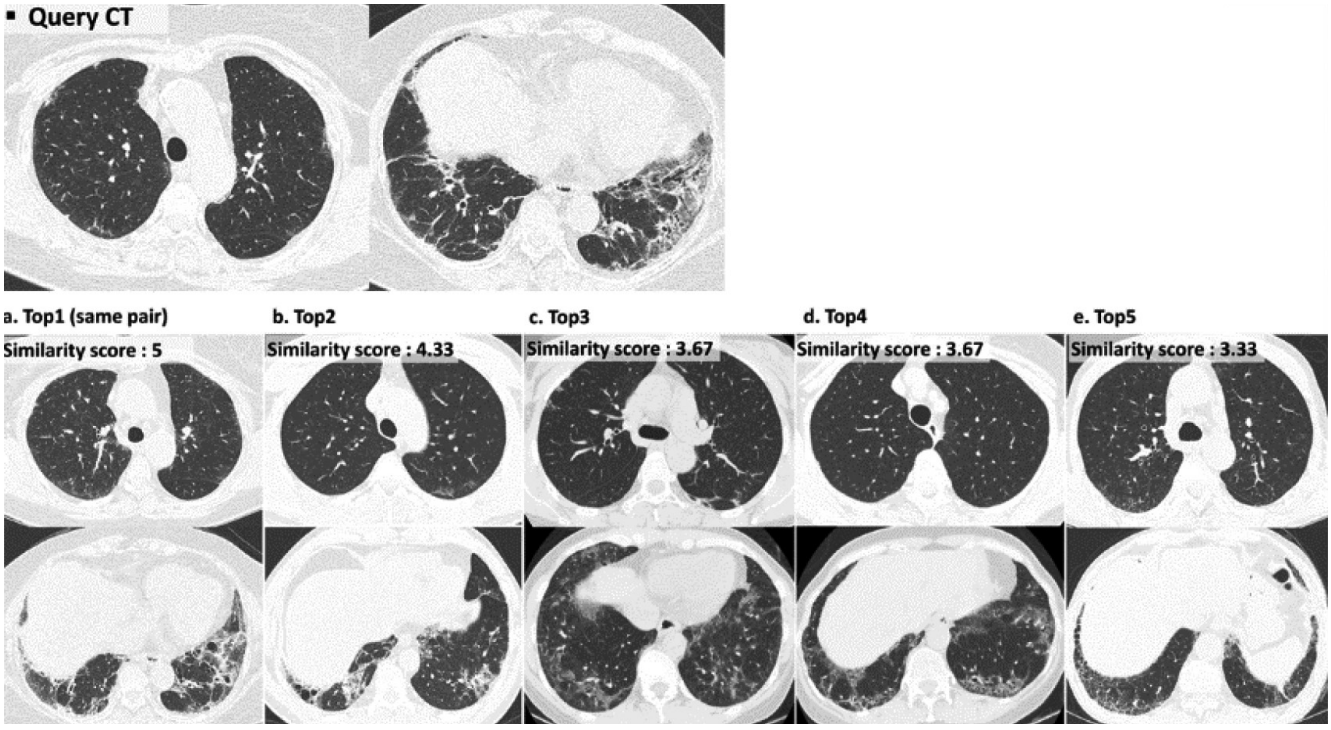

Fig. 5. Mean similarity scores of the retrieved five similar CTs assessed by the three thoracic radiologists. The query CT shows ground-glass opacity and mild traction bronchiectasis in both lower lobes which are indicative of a NSIP pattern. The query case was confirmed as NSIP based on HRCT features and histologic result. Among the five retrieved similar CTs, the mean similarity score was highest in (a) the top 1 retrieval which was the pair with query CT. The mean similarity score was lowest in (e) the top 5 retrieval with score 3.33. All of the five similar CTs retrieved by CBIR were HRCTs with NSIP pattern.

Table 3. Similarity Scores of Top 1–5 Retrieved CTs

Top N	CTs with Similarity Score ≥ 4				Similarity Score*			
	Reader 1	Reader 2	Reader 3	Overall	Reader 1	Reader 2	Reader 3	Overall
Top 1 (60)	54 (90.0)	54 (90.0)	52 (86.7)	160 (88.9)	4.60 ± 0.72	4.52 ± 0.77	4.45 ± 1.00	4.52 ± 0.84
Top 2 (60)	46 (76.7)	43 (71.7)	47 (78.3)	136 (75.6)	4.08 ± 1.00	4.02 ± 0.95	3.78 ± 1.12	3.96 ± 1.03
Top 3 (60)	41 (68.3)	39 (65.0)	44 (73.3)	124 (68.9)	3.85 ± 0.86	3.80 ± 0.82	3.67 ± 1.04	3.77 ± 0.91
Top 4 (60)	38 (63.3)	39 (65.0)	36 (60.0)	113 (62.8)	3.68 ± 0.81	3.65 ± 0.76	3.35 ± 1.12	3.56 ± 0.92
Top 5 (60)	40 (66.7)	39 (65.0)	39 (65.0)	118 (65.6)	3.72 ± 0.80	3.57 ± 0.83	3.53 ± 1.00	3.61 ± 0.89
Total (300)	219 (73.0)	214 (71.3)	218 (72.7)	651 (72.3)	3.99 ± 0.90	3.91 ± 0.89	3.76 ± 1.10	3.88 ± 0.98

Data are number of CTs with the similarity score more than or equal to 4 except similarity scores which are the means ± standard deviations. Data in parentheses are percentages. *Kruskal-Wallis test showed that the similarity scores of three radiologists were significantly different among the Top 1 to 5 retrievals, respectively ($p < 0.001$).

https://doi.org/10.3348/kjr.2020.0603

the order of similarity given by the CBIR system. Therefore, the similarity specified by the CBIR seems to be similar to the visual similarity determined by thoracic radiologists.

Our CBIR system showed different performances depending on the disease class of DILD; it showed better performance in the UIP or NSIP cases than in the COP pattern. The similarity scores by radiologists were also higher in the UIP and NSIP patterns than in the COP pattern. This difference may be due to the slight altered extent of the lung lesion after treatment on follow-up CT in some COP cases, although the radiologist rated the stability score as 'same.' Moreover, fewer number of COP cases were included in the database compared to other disease patterns. Using the database with a relatively small number of COP pattern, CBIR might select the other disease patterns for some cases of top 1–5 similar CTs, and that might result in the lower similarity score given by the thoracic radiologists. Therefore, due to the limited number of COP cases, the retrieving performance of CBIR with COP cases could be poor compared to cases with other patterns. The performance of our CBIR system according to the DILD patterns should be evaluated with a larger database.

Table 4. Similarity Scores in the Three Disease Classes

Reader	UIP	NSIP	COP	*P**
Overall	4.08 ± 0.84	3.78 ± 1.00	3.50 ± 1.16	< 0.001
Reader 1	4.18 ± 0.83	3.87 ± 0.91	3.64 ± 0.99	0.001
Reader 2	4.09 ± 0.79	3.75 ± 0.91	3.70 ± 1.02	0.005
Reader 3	3.98 ± 0.89	3.72 ± 1.17	3.16 ± 1.40	< 0.05

Data are presented as the mean ± standard deviation. *Similarity scores of each radiologist were compared among the three disease groups using the Kruskal-Wallis test, Post-hoc analysis with the Mann Whitney test was performed to compare the similarity scores between the two groups, Reader 1: UIP vs. NSIP, $p = 0.006$; UIP vs. COP, $p = 0.001$; NSIP vs. COP, $p = 0.189$. Reader 2: UIP vs. NSIP, $p = 0.004$; UIP vs. COP, $p = 0.019$; NSIP vs. COP, $p = 0.884$. Reader 3: UIP vs. NSIP, $p = 1.70$; UIP vs. COP, $p < 0.001$; NSIP vs. COP, $p = 0.013$, Significance level of 0.0167 takes into account the Bonferroni's correction for post-hoc analysis.

By presenting similar cases of DILD for reviewing at a single time the CBIR system has potential in many areas such as self-learning and educating students, radiologists, and clinicians. Furthermore, if the CBIR system is linked to the medical records system, it can contribute to better treatment decisions by providing comprehensive information regarding treatment options and effects, as well as the prognosis of cases similar to those of the patient. It has been reported that combining imaging features with text data improves the accuracy of the case search (19). Also, new studies can be conducted by enabling statistical research based on identifying similar patterns of DILD CT images. In addition to the typical CT findings of each DILD, various CT findings are presented, even in the same DILD disease class. Therefore, CBIR can be used to classify image-based phenotypes showing different image features even in the same disease class, and then the research according to these new image-based phenotypes may also be conducted.

Our study has several limitations. First, our research was performed with a limited dataset consisting of three disease patterns that are frequently encountered in clinical practice. However, in addition to the above three patterns, DILD includes various other disease patterns and it is often difficult to make a differential diagnosis. Thus, it is necessary to evaluate the performance of our CBIR using a dataset containing a larger number of cases with various DILD patterns. Second, our research was performed with our own private database obtained using our CT protocol. Therefore, the CBIR could be difficult to apply to different study protocols. Third, the effect of CBIR on the clinical workflow was not evaluated. Computer-aided image interpretation has been reported to substantially enhance diagnostic sensitivity in various diseases (20-23). Thus, the impact of CBIR on diagnostic performance should be evaluated in future studies. Fourth, clinical parameters, such as patient age, or sex, were not provided in our CBIR. Clinical parameters as well as the similarity

Table 5. Inter-Reader Agreement of the Three Chest Radiologists for Similarity Scores

Top N	Reader 1 vs. Reader 2	Reader 2 vs. Reader 3	Reader 3 vs. Reader 1
Overall	0.542 (0.472–0.612)	0.496 (0.429–0.564)	0.541 (0.477–0.605)
Top 1	0.724 (0.594–0.854)	0.701 (0.565–0.837)	0.618 (0.478–0.758)
Top 2	0.477 (0.320–0.634)	0.439 (0.285–0.594)	0.513 (0.387–0.639)
Top 3	0.311 (0.141–0.481)	0.297 (0.155–0.439)	0.364 (0.194–0.535)
Top 4	0.461 (0.282–0.641)	0.404 (0.246–0.561)	0.522 (0.388–0.656)
Top 5	0.492 (0.334–0.649)	0.401 (0.200–0.603)	0.406 (0.210–0.603)

Data are k values, with weighted 95% confidence intervals in parentheses.

https://doi.org/10.3348/kjr.2020.0603

Korean Journal of Radiology
KJR

of CT images may have an important role in the diagnosis of DILD. Fifth, although radiologists were blinded to the diagnosis obtained from CTs, they may have been able to recognize the diagnosis of some CTs, which in turn may have influenced the visual similarity scores. Thus, kappa scores for inter-reader agreement may have also been overestimated.

In conclusion, by comparing the extent and distribution of regional lung parenchymal patterns which is automatically quantified and classified by CNN in DILD, the proposed CBIR system is able to retrieve CTs which are similar to the query CTs from the database. The CBIR showed good performance in retrieving CTs with similar patterns to the query DILD CTs. Thus, by searching for similar CTs of confirmed cases, the CBIR may be helpful for the diagnosis of DILD.

Supplementary Materials

The Data Supplement is available with this article at https://doi.org/10.3348/kjr.2020.0603.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

ORCID iDs

Hye Jeon Hwang
https://orcid.org/0000-0003-3508-2870
Joon Beom Seo
https://orcid.org/0000-0003-0271-7884
Sang Min Lee
https://orcid.org/0000-0002-2173-2193
Eun Young Kim
https://orcid.org/0000-0002-7280-8856
Beomhee Park
https://orcid.org/0000-0002-6548-2392
Hyun-Jin Bae
https://orcid.org/0000-0001-5134-5517
Namkug Kim
https://orcid.org/0000-0002-3438-2217

REFERENCES

1. Walsh SL, Calandriello L, Sverzellati N, Wells AU, Hansell DM, Consort UIPO. Interobserver agreement for the ATS/ERS/JRS/ALAT criteria for a UIP pattern on CT. *Thorax* 2016;71:45-51
2. Watadani T, Sakai F, Johkoh T, Noma S, Akira M, Fujimoto K, et al. Interobserver variability in the CT assessment of honeycombing in the lungs. *Radiology* 2013;266:936-944
3. Jacob J, Bartholmai BJ, Rajagopalan S, Kokosi M, Nair A, Karwoski R, et al. Mortality prediction in idiopathic pulmonary fibrosis: evaluation of computer-based CT analysis with conventional severity measures. *Eur Respir J* 2017;49:1601011
4. Lee SM, Seo JB, Oh SY, Kim TH, Song JW, Lee SM, et al. Prediction of survival by texture-based automated quantitative assessment of regional disease patterns on CT in idiopathic pulmonary fibrosis. *Eur Radiol* 2018;28:1293-1300
5. Salisbury ML, Lynch DA, van Beek EJ, Kazerooni EA, Guo J, Xia M, et al. Idiopathic pulmonary fibrosis: the association between the adaptive multiple features method and fibrosis outcomes. *Am J Respir Crit Care Med* 2017;195:921-929
6. Yoon RG, Seo JB, Kim N, Lee HJ, Lee SM, Lee YK, et al. Quantitative assessment of change in regional disease patterns on serial HRCT of fibrotic interstitial pneumonia with texture-based automated quantification system. *Eur Radiol* 2013;23:692-701
7. Kim GB, Jung KH, Lee Y, Kim HJ, Kim N, Jun S, et al. Comparison of shallow and deep learning methods on classifying the regional pattern of diffuse lung disease. *J Digit Imaging* 2018;31:415-424
8. Depeursinge A, Vargas A, Gaillard F, Platon A, Geissbuhler A, Poletti PA, et al. Case-based lung image categorization and retrieval for interstitial lung diseases: clinical workflows. *Int J Comput Assist Radiol Surg* 2012;7:97-110
9. Dhara AK, Mukhopadhyay S, Dutta A, Garg M, Khandelwal N. Content-based image retrieval system for pulmonary nodules: assisting radiologists in self-learning and diagnosis of lung cancer. *J Digit Imaging* 2017;30:63-77
10. Oosawa A, Kurosaki A, Kanada S, Takahashi Y, Ogawa K, Hanada S, et al. Development of a CT image case database and content-based image retrieval system for non-cancerous respiratory diseases: method and preliminary assessment. *Respir Investig* 2019;57:490-498
11. Park B, Park H, Lee SM, Seo JB, Kim N. Lung segmentation on HRCT and volumetric CT for diffuse interstitial lung disease using deep convolutional neural networks. *J Digit Imaging* 2019;32:1019-1026
12. Kundel HL, Polansky M. Measurement of observer agreement. *Radiology* 2003;228:303-308
13. Aisen AM, Broderick LS, Winer-Muram H, Brodley CE, Kak AC, Pavlopoulou C, et al. Automated storage and retrieval of thin-section CT images to assist diagnosis: system description and preliminary assessment. *Radiology* 2003;228:265-270
14. Müller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. *Int J Med Inform* 2004;73:1-23
15. Müller H, Rosset A, Garcia A, Vallée JP, Geissbuhler A. Informatics in radiology (infoRAD): benefits of content-based visual data access in radiology. *Radiographics* 2005;25:849-858

16. Sasso G, Marsiglia HR, Pigatto F, Basilicata A, Gargiulo M, Abate AF, et al. A visual query-by-example image database for chest CT images: potential role as a decision and educational support tool for radiologists. *J Digit Imaging* 2005;18:78-84
17. Shyu CR, Brodley CE, Kak AC, Kosaka A, Aisen AM, Broderick LS. ASSERT: a physician-in-the-loop content-based retrieval system for HRCT image databases. *Comput Vis Image Underst* 1999;75:111-132
18. Brodley C, Kak A, Shyu C, Dy J, Broderick L, Aisen AM. *Content-based retrieval from medical image databases: a synergy of human interaction, machine learning and computer vision*. Menlo Park: AAAI Press, 1999:760-767
19. Névéol A, Deserno TM, Darmoni SJ, Güld MO, Aronson AR. Natural language processing versus content-based image analysis for medical document retrieval. *J Am Soc Inf Sci Technol* 2008;60:123-134
20. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology* 2001;220:781-786
21. Lo SB, Freedman MT, Gillis LB, White CS, Mun SK. JOURNAL CLUB: computer-aided detection of lung nodules on CT with a computerized pulmonary vessel suppressed function. *AJR Am J Roentgenol* 2018;210:480-488
22. Roos JE, Paik D, Olsen D, Liu EG, Chow LC, Leung AN, et al. Computer-aided detection (CAD) of lung nodules in CT scans: radiologist performance and reading time with incremental CAD assistance. *Eur Radiol* 2010;20:549-557
23. Warren Burhenne LJ, Wood SA, D'Orsi CJ, Feig SA, Kopans DB, O'Shaughnessy KF, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* 2000;215:554-562

https://doi.org/10.3348/kjr.2020.0603