Journal of Biomedical Informatics 84 (2018) 123–135

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Relevance feedback for enhancing content based image retrieval and automatic prediction of semantic image features: Application to bone tumor radiographs

Imon Banerjee[a,*], Camille Kurtz[c], Alon Edward Devorah[a], Bao Do[b], Daniel L. Rubin[a,b,1], Christopher F. Beaulieu[b,1]

[a] Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
[b] Department of Radiology, Stanford University, Stanford, CA, USA
[c] Department of Computer Sciences, LIPADE (EA 2517), Université Paris Descartes, Paris, France

ARTICLE INFO

Keywords:
Content based image retrieval
Pixel-level features
Radiomics
Semantic features
Relevance feedback
Bone tumors
Radiography

ABSTRACT

Background: The majority of current medical CBIR systems perform retrieval based only on “imaging signatures” generated by extracting pixel-level quantitative features, and only rarely has a feedback mechanism been incorporated to improve retrieval performance. In addition, current medical CBIR approaches do not routinely incorporate semantic terms that model the user’s high-level expectations, and this can limit CBIR performance.
Method: We propose a retrieval framework that exploits a hybrid feature space (HFS) that is built by integrating low-level image features and high-level semantic terms, through rounds of relevance feedback (RF) and performs similarity-based retrieval to support semi-automatic image interpretation. The novelty of the proposed system is that it can impute the semantic features of the query image by reformulating the query vector representation in the HFS via user feedback. We implemented our framework as a prototype that performs the retrieval over a database of 811 radiographic images that contains 69 unique types of bone tumors.
Results: We evaluated the system performance by conducting independent reading sessions with two subspecialist musculoskeletal radiologists. For the test set, the proposed retrieval system at fourth RF iteration of the sessions conducted with both the radiologists achieved mean average precision (MAP) value ~0.90 where the initial MAP with baseline CBIR was 0.20. In addition, we also achieved high prediction accuracy (> 0.8) for the majority of the semantic features automatically predicted by the system.
Conclusion: Our proposed framework addresses some limitations of existing CBIR systems by incorporating user feedback and simultaneously predicting the semantic features of the query image. This obviates the need for the user to provide those terms and makes CBIR search more efficient for inexperience users/trainees. Encouraging results achieved in the current study highlight possible new directions in radiological image interpretation employing semantic CBIR combined with relevance feedback of visual similarity.

1. Introduction

Medical information retrieval is important for research and potentially for clinical care, but finding similar cases is largely an unassisted and time-consuming process, and precision is established through many years of training and experience. Even despite this training, substantial inter-reader variation in determining case similarity is challenging. Moreover, the volume of medical information is growing faster than the ability of professionals to do this task themselves without the support of computerized search mechanisms [28]. In radiology, image retrieval has particular importance because the radiologist commonly confronts rare abnormalities for which diagnosis is difficult. Finding similar images from large imaging archives, such as picture archiving and communication system (PACS), can potentially assist in suggesting diagnoses of many similar cases, and the evidence supplied by the similar cases can assist the radiologist to improve interpretation of rare abnormalities and may help in determining diagnosis [18].

To serve this purpose, Medical content-based image retrieval (CBIR) systems have been developed that typically operate by comparing the query image to other images present in the imaging archives, based on

[*] Corresponding author.
E-mail address: imonb@stanford.edu (I. Banerjee).
[1] Equal contributions.

https://doi.org/10.1016/j.jbi.2018.07.002
Received 5 April 2018; Received in revised form 6 June 2018; Accepted 3 July 2018
Available online 05 July 2018

comparing "*imaging signatures*" generated based only on the quantitative features (*radiomics features*), such as shape and textures in the image or within regions of the image [26]. While these quantitative features describe the low-level pixel-based information in an automated fashion, they are often not specific enough to capture high-level radiological concepts (*semantic features*). Therefore, the performance of medical CBIR systems is often constrained by the low-level properties of medical images and cannot effectively model the user's high-level expectations. Since this challenge remains unsolved, traditional CBIR systems are not equipped to support the current advancement of cross-sectional clinical studies with thousands of cases, and current CBIR systems still require exhaustive manual filtering of the retrieval results.

Recently, as alternatives to traditional CBIR, some methods [18,19,22], use a combination of quantitative features and qualitative descriptive terms used by radiologists ("*semantic image features*") to serve as the imaging signatures for CBIR. The combination of high-level and low-level image descriptions may improve performance of CBIR; however, such hybrid CBIR systems are limited by two core constraints. *First*, in order to use the semantic features, the end-user must annotate query images with semantic terms, which is not only a tedious process but also requires considerable domain expertise for inferring the appropriate semantic characteristics of an abnormality [10]. This may restrain the similarity-based diagnosis workflow only to the expert radiologists and therefore diminishes its core purpose: *evidence-based diagnosis of rare/unseen abnormality* (PDQ [24]). Some computerized methods Banerjee et al. [2,6] have proposed to apply machine learning techniques to *predict* semantic terms by utilizing the low-level pixel data. However most of these studies were pursued on a narrow imaging domain with limited expert-knowledge, and were validated on a relatively small number of cases which limit their generalizability for other domains. A *second* limitation is that often the retrieved images where quantitative features are only measures for similarity, are of insufficient resemblance to the query image to be clinically relevant. The radiologist must therefore spend a large amount of time sifting through irrelevant retrieved images to identify those that are semantically similar to the query image according to the clinical task. Perhaps the most important limitation is that the performance of current clinical CBIR systems cannot be improved based on user feedback. This limits the ability to customize retrievals to match individual reader's expectations and for a given image database, puts an upper bound on the accuracy of image retrieval.

A "Relevance feedback" mechanism has been proposed as a strategy for clinical CBIR systems to improve with use and to add more flexibility for personalizing the retrieval results [20]. The key idea is to incorporate user feedback about the relevance of retrieved results produced by an initial CBIR search (based on only semantic and/or quantitative features) to refine subsequent search results. User feedback can be gathered across multiple iterations of search, with the user evaluating the quality each retrieved image to the query image in each iteration [35]. Several approaches to relevance feedback in CBIR have been reported [4,7,22,32] focusing on using various combinations of quantitative image features, but to our knowledge, no prior systems have leveraged integration between semantic and quantitative features. Nonetheless, several retrospective studies [2,3,21] have advocated that bridging the "semantic gap" between complex image features and the human-perceived semantic features will enable construction of a single, unified, and searchable data structure for automated reasoning on both image content and their semantic descriptors. We hypothesized that an efficient integration may also play a critical role in maximizing semantic accuracy in a CBIR system for radiological images, yet no prior study exists that can fully support our claim.

Our goal is to extend the traditional relevance feedback mechanism by incorporating semantic information in a hybrid feature space (HFS) along with the quantitative features to improve the retrieval outcome. In addition, we seek to predict the semantic features of query images with the implicit knowledge collected via the user feedback in the HFS, which would reduce the need for radiologist annotation of images for CBIR. We make two key research contributions:

First, we create a system that efficiently aggregates three levels of information - *quantitative image features*, *semantic features*, and *user feedback*, bridging the current "semantic gap" in medical image retrieval and simultaneously producing personalized search results.

Second, we propose an approach to predict automatically the semantic features of the query image by exploiting the relevance feedback and the quantitative features that have been computed from the raw pixel data of the region-of-interest (ROI) draw by the user.

The remaining article is organized as follows: Section 2 describes the database employed and the proposed methodology; Section 3 describes experimental results; and Section 4 presents a summary of the work, highlights limitations, and provides some concluding remarks.

2. Material and methods

2.1. Data

The study was approved by our Institutional review board (IRB). The requirement for informed consent was waived as this was a retrospective review of historical images and patient data. The data set is a collection of 1664 radiographic cases of bone tumors at a tertiary-care teaching hospital (Stanford medical center) that were collected by one Professor approximately between the year 1955 and 2005. The original images were hard copy (conventional X-ray film) radiographs, and a transparency film scanner (Pacsgear – Lexmark, Pleasanton, CA) was used to digitize all images at 600 dpi. A total of 22,864 images were captured from the 1664 cases. Upon review by an experienced musculoskeletal radiologist, cases were subjectively categorized into 124 low, 675 medium, and 865 high quality cases. High quality cases included excellent representation of the bone lesion in terms of radiographic exposure and resolution, as well as lack of extraneous markings such as wax pencil or film labels. Low quality cases included under- or over-exposed images that may have exhibited motion artifact or interfering overlying markings. Taking the high quality and a selection of the medium quality cases, a "top 1000" collection was constructed which included the relevant radiographic projections that best shows each lesion (see Fig. 1).

During an initial semantic annotation phase ("offline processing") prior to the current work and described further in Section 2.2.1a below, 189 cases were not annotated because of limited visibility of the lesions or subjectively lower overall image quality. This curation process resulted in 811 cases with 69 unique bone tumor diagnoses that were either confirmed by histology or by pathognomonic features. In Table 1, we present the distribution of the cases according to the bone tumor diagnosis to demonstrate the heterogeneous nature of the dataset. For creating the test image pool, we randomly selected 20 cases from the first and second columns of Table 1 to make sure that the database contains a significant number of images with same bone tumor diagnosis for the retrieval. The limited number of cases in the test image pool is mainly influenced by the complexity and the size of the database.

2.2. System architecture

In Fig. 2, we present the workflow of the system which is divided into two core operating phases: (1) offline processing, and (2) online operation. In the offline processing phase, with the help of radiologists, we built our annotated database by identifying the regions-of-interest (ROI) from the sample images, and recording the semantic (radiological observations) and radiomics features. In the online processing, our proposed system inputs a query image, and, based on refinement with user feedback, retrieves '*n*' similar images, where '*n*' is specified by the users. In addition, the system also predicts the pre-defined set of radiological observations for the query images. In the following subsections,

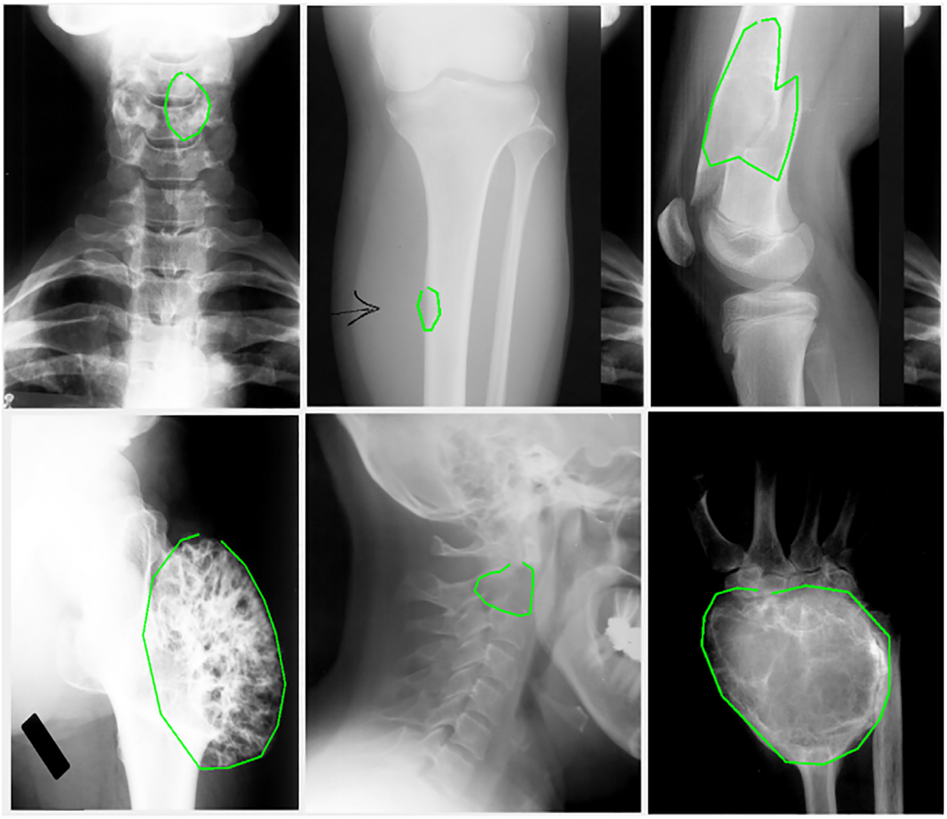


Fig. 1. Sample images of "top 1000" collection.

we illustrate each processing block of Fig. 2 in detail.

2.2.1. *Offline processing*

a. Manual annotation of ROI with semantic features

For each of the 801 known cases (excluding the 20 images in the test pool), 1 image (single image with best visualization of the tumor) was selected for annotation with the 18 clinical and qualitative features by either of two experienced radiologists ([initials withheld], 22 years' experience and [initials withheld], 6 years' experience). An ROI was also drawn to circumscribe the lesion. For consistent encoding, annotations were performed using ePad [27], a freely available quantitative imaging informatics platform, and stored in Annotation and Image Markup (AIM) standard [26]. Attribute values for each field were derived from RadLex, if possible [21], and supplemented by attributes derived from clinical experience or literature descriptions of bone tumor observations [8].

A total of 18 semantic features (2 clinical and 16 qualitative radiographic features) were recorded for each case (see Table 2). The semantic features used in this work were chosen by an expert musculoskeletal radiologist having knowledge of radiologic observations. The selections were made to highlight various visual image features contributing to bone tumor diagnosis, such as lesion boundary, internal texture, density, bony expansion or cortical erosion, bone location (transverse and longitudinal), and patient age (in decades).

b. Radiomics feature extraction

Following the manual annotation process of drawing free-form ROIs on images and recording semantic features of each lesion, an image feature extraction module cropped the ROI from the whole image. In order to capture boundary features of the lesion, the module padded the ROI margin with 50 neighboring pixels external to the ROI. We chose this sized margin empirically to capture enough boundary features of the lesion. Radiomics features were then computed from raw pixel data of the cropped images.

We considered two types of radiomics features: (i) *photometric features* that capture statistics of intensity values and texture information of the ROI and its adjacent region as well as quantifies the intensity gradient along the contour, and (ii) *geometric features* that make use of shape information of the ROI. We created a MATLAB module to extract 18 quantitative feature classes (see Table 3) from the images.

In addition to these standard radiomics features (Table 3, we also implemented two sets of case-specific edge-based features that appear to correlate well with the major complementary characteristics of bone tumors – increased bone density and bone loss relative to surrounding normal bone.

The first set of edge-based feature is calculated from the Sobel gradient image, where the value at each point in the image is the result of the corresponding gradient vector. The Sobel edge operator [14] consists of a pair of 3 × 3 convolution kernels ($Sobel\_G_x$ *and* $Sobel\_G_y$) where one kernel ($Sobel\_G_y$) is simply the other($Sobel\_G_x$) rotated by 90°. These kernels are designed in a way to respond maximally to edges running vertically and horizontally relative to the pixel grid. The kernels ($Sobel\_G_x$ *and* $Sobel\_G_y$) are applied separately to the input crop image (A), to produce separate measurements of the gradient component in each orientation (x, y) as: $G_x = Sobel\_G_x * A$ and $G_y = Sobel\_G_y * A$. Finally, the measurements of gradient component are combined to find the absolute magnitude of the gradient at each pixel i of Sobel gradient image ($Sobel\_G$) as: $|G_i| = \sqrt{G_{xi}^2 + G_{yi}^2}$.

In addition to Sobel operator, we applied another edge detection operator - *Laplacian of Gaussian (LoG)*, for simultaneously reducing the sensitivity to noise and highlighting regions of intensity change in vertical and horizontal directions. The 2-D *LoG* function centered on zero and with Gaussian standard deviation σ has the form:

Table 1
Bone tumor diagnosis/classes (69 total) represented in the final 811 cases according to the number of samples for each diagnosis.

Bone Tumor Diagnosis (# No. of Samples)		
No. of Samples ≥20	No. of Samples ≥5	No. of Samples < 5
Osteosarcoma (83) Enchondroma (65) Metastasis (55) Osteochondroma (46) Aneurysmal bone cyst (41) Chondrosarcoma (41) Giant cell tumor (41) Nonossifying fibroma (38) Ewing sarcoma (38) Fibrous dysplasia (30) Lymphoma (22) Chondroblastoma (21)	Simple bone cyst (19) Eosinophilic granuloma (19) Osteoid osteoma (16) Non-Hodgkin lymphoma (16) Malignant fibrous histiocytoma (14) Chondromyxoid fibroma (11) Osteomyelitis (11) Periosteal chondroma (10) Multiple myeloma (10) Osteoblastoma (10) Unknown (12) Ganglion cyst (9) Paget disease (9) Giant cell reparative granuloma (8) Hemangioma (7) Intraosseous lipoma (7) Adamantinoma (7) Plasmacytoma (6) Pigmented villonodular synovitis (6) Ossifying fibroma (6) Hodgkin lymphoma (5) Sarcoma (5) Synovial sarcoma (5)	Cystic angiomatosis (4) Mastocytosis (4) Epithelioid hemangioendothelioma (4) Brodie abscess (3) Angiosarcoma (3) Fibrosarcoma (3) Benign fibrous histiocytoma (3) Ameloblastoma (2) Giant cell tumor of tendon sheath (2) Tuberculosis (2) Caffey disease (2) Leukemia (2) Central osteosarcoma (2) Rosai-Dorfman disease (2) Telangiectatic osteosarcoma (2) Osteopetrosis (1) Avascular necrosis (1) Osteonecrosis (1) Erdheim-Chester disease (1) Hereditary multiple exostoses (1) Osteoma (1) Desmoplastic fibroma (1) Glomus tumor (1) Granuloma (1) Sclerosing osteomyelitis (1) Rheumatoid nodule (1) Enchondromatosis (1) PIndborg Tumor (1) Chronic sclerosing osteitis (1) Sarcoidosis (1) Schwannoma (1) Rhabdomyosarcoma (1) Maffucci syndrome (1)

$LoG(x, y) = \frac{1}{\pi\sigma^4}\left(1-\frac{x^2+y^2}{2\sigma^2}\right)e^{\frac{x^2+y^2}{2\sigma^2}}$. We designed a pair of discrete 10 × 10 kernels, where the first one approximates the LoG function (LoG_G_x) and second one(LoG_G_y) is a transpose version of the original. Because these kernels are approximating a second derivative measurement on the image, they are very sensitive to noise. To counter this, the image (A)is smoothed using Gaussian smoothing before applying the Laplacian filter. This pre-processing step reduces the high frequency noise components prior to the differentiation step. We convoluted the Laplacian kernel into $x$and $y$direction of the image(A) separately as: $L_x = LoG\_G_x{*}A$ and $L_y = LoG\_G_y{*}A$. Finally, we combined them to find the absolute magnitude of the gradient at each pixel i of Laplacian gradient image ($Laplacian\_G$) as: $|Laplacian\_G_i| = \sqrt{LoG\_G_{xi}^2 + LoG\_G_{yi}^2}$.

Both Sobel and LoG operators take only the masked ROI region of a single gray level image as input and produce another gray level gradient magnitude map image as output. In Fig. 3, we present the Sobel ($Sobel\_G$) and Laplacian of Gaussian ($Laplacian\_G$) magnitude map image for various types of bone tumors which shows that the edge operators were able to extract internal spatial arrangements of intensities within the tumor, and may also be helpful in differentiating the tumor characteristics.

We condensed the information of the magnitude map images $Sobel_G$and $Laplacian_G$in the form of two statistical measures – (1) *total variational energy* (TV) which is L_1 norm of the gradient of the edge image with 4-connected neighborhood that has been computed as: $TV\,(Sobel\_G) = \sum_x ||\nabla Sobel\_G(x)||_1$ and $TV\,(Laplacian\_G) = \sum_x ||\nabla Laplacian\_G(x)||_1$, (2) entropy which measure the randomness of the gradient edge image that has been computed as: $Entropy = -\sum_{k=1}^{K} p_k \log_2(p_k)$,where K is the number of gray level in the magnitude map images $Sobel_G$and $Laplacian_G$and $p_k$is the probability associated with the gray level k. Finally, we normalized the statistical measures according to the area of the ROI region.

We merged all the radiomics feature vectors for our 801 cases into one large feature matrix, normalized column-wise, to obtain zero means and standard deviation one, resulting in a 496-dimensional feature matrix. Finally, all the semantic and radiomics features together with the images are stored in a database.

c. Hybrid Feature Space (HFS) formulation

During the initialization, the system (see Fig. 2) processes the semantic and radiomics image features associated with all images in the database and creates a single hybrid feature space. Among the original 18 semantic features available, we only consider 9 semantic features that can be associated with the radiomics features computed from the images (see Table 4). We discard the demographic (e.g. age, gender) and any of the location dependent (e.g. longitudinal or transverse location) information that cannot be characterized by the quantitative modeling of pixel data because there is no intrinsic information in the images that can represent these factors. We convert the 9 semantic features into categorical variables with possible values 0/1, and create a semantic feature matrix: $S\_Matrix_{801\times9}$. We create radiomics feature matrix: $R_Matrix_{801\times496}$by parsing the normalized radiomics features resulting from the previous step. Finally, we augment $S\_Matrix_{801\times9}$ and $R\_Matrix_{801\times496}$ to create a larger feature matrix***: $\gamma \leftarrow [R_{Matrix801\times496}|S_{Matrix801\times9}]$, which is used to build an initial hybrid feature space (HFS) model of the database, in which each image (j) is represented as a vector of 505 dimensions: $\gamma_j = (\gamma_{j,1}, \gamma_{j,2}, \cdots\cdots..,\gamma_{j,505})$.

2.2.2. Online operations

a. Query image and ROI, Radiomics feature extraction, Inserting query in HFS

We built a graphical UI that allows the user to load a query image as a DICOM file and create a free-form ROI circumscribing the lesion with mouse clicks. There is no need for the user to provide the semantic features of the query image, since our system infers them via relevance feedback in HFS (see Section 2.2.2c). Elimination of the need to annotate the query image with semantic features makes the initial query formulation process reasonably simple and fast for the end-user.

The system calls the '*Radiomics feature computation*' module (Section 2.2.1) for computing the radiomics features matrix: $R_{Matrix_query_{1\times496}}$of the ROI. The initial query vector is formulated as: $\gamma_{query\_round1} \leftarrow [R_{Matrix\_query_{1\times496}}|S_{Matrix\_query_{1\times9}}]$, where $S_{Matrix\_query_{1\times9}}$ is the semantic feature matrix but it is only padded with zeros since no semantic features are available yet for the query image. Thus, the initial query vector is built only based on the quantitative information of the ROI region. If multiple abnormalities are present, the system only supports analysis of a single ROI that has been marked by the user. Finally, the system embeds the initial query vector (γ_{query_round1}) in the HFS (see Fig. 3a).

b. Cosine distance computation, Ranked retrieval of top 'n' similar images

In this module, the system measures the cosine similarity of the database images with the query vector in the high dimensional HFS by computing the Cosine Angle Distance between the feature vector of database entities $\gamma$and the query vector γ_{query}. The cosine distance of two vectors γ_{query} and γ_i is defined as:

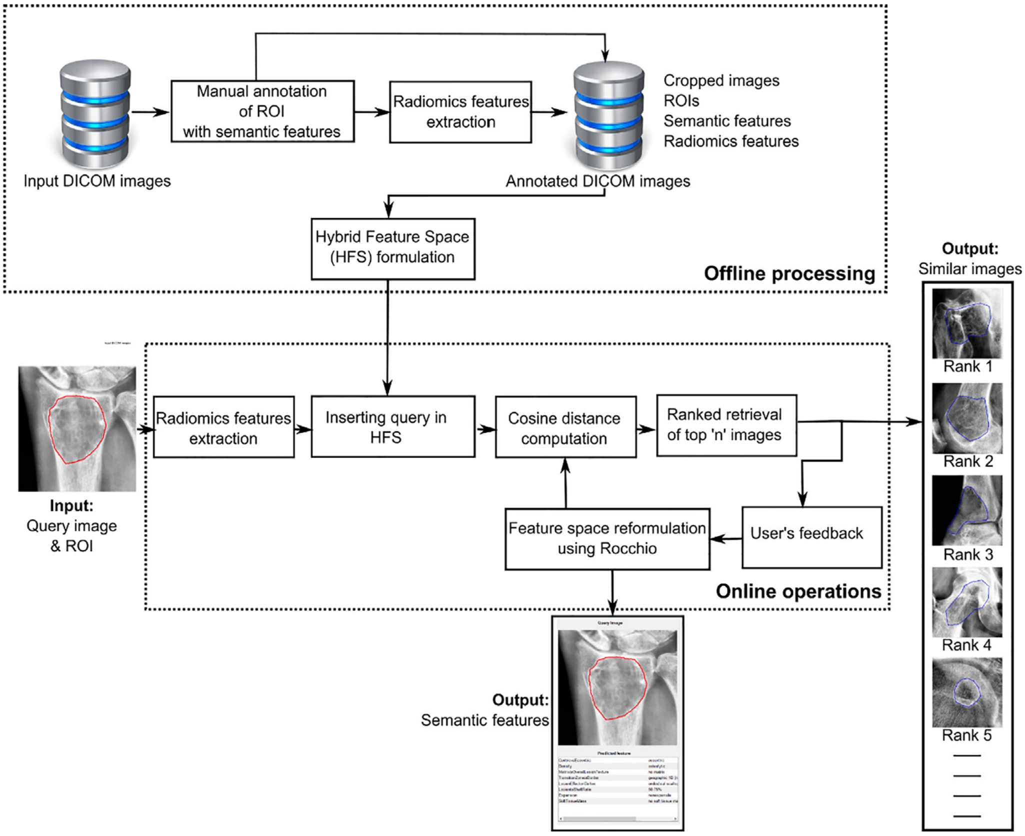


Fig. 2. Workflow of the proposed system.

$$\sigma(\gamma_{query}, \gamma_i) = \frac{\gamma_{query}\cdot\gamma_i}{||\gamma_{query}||||\gamma_i||} \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm and γ_i denotes the feature vector of the ith image in the dataset. We choose to use Cosine distance to measure the similarity since one important property of vector cosine angle is that it gives a metric of similarity between two vectors unlike Euclidean distance, which give metrics of *dissimilarities* instead. Finally, the similarity values between query and the stored images are used to rank the results, and the '*n*' top ranked images are retrieved and presented to the user in the subsequent round to collect the feedback.

c. User feedback, Feature space reformulation using Rocchio, Semantic features prediction

The goal of this step is to incorporate feedback from the user about the relevancy of initial retrieved images so as to improve the final query results according to user preferences and, ultimately, to improve the

Table 2
List of 18 semantic features and all possible values.

Type	Semantic Feature	Values
Clinical	Age	Decade bins: 0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–100+
	Gender	Male, female
Radiographic	Number of lesions	Solitary, multiple
	Bone location	Carpals, clavicle, femur, fibula, foot, hand, humerus, iliac bone, ischium, mandible, patella, pubis, radius, rib, sacrum, scapula, skull, sternum, tarsals, tibia, ulna, vertebrae
	Longitudinal location	Apophysis, diaphysis, epiphysis, metadiaphysis, metaphysis, n/a
	Proximal vs. distal	Proximal, middle, distal, not applicable
	Transverse location	Medullary cavity, endosteum, cortex, periosteum, sessile, pedunculated, juxtacortical, soft tissue
	Distribution	Central, eccentric, n/a
	Density	Normal, ground glass, lytic, sclerotic, mixed lytic and sclerotic
	Matrix/texture	Normal, bone forming or osteoid, chondroid, septated, coarse trabeculae, central calcification
	Transition zone/border	Geographic 1A (narrow sclerotic), geographic 1B (narrow nonsclerotic), geographic 1C (wide non sclerotic), permeative/destructive/punched out, unable to determine border or n/a (e.g.: osteochondroma does not have a border)
	Cortex	Endosteal scalloping grade: 0 = none, 1 = 0–25%, 2 = 25–50%, 3 = 50–75%, 4 = 75%+, where % = approximate depth of scalloping; cortical thickening, periosteal scalloping (any degree), n/a
	Periosteum	No periosteal reaction, solid periosteal reaction, lamellated periosteal reaction, interrupted periosteal reaction, codman triangle, sunburst
	Lesion to shaft ratio	0–25%, 25–50%, 50–75%, 75–100%, > 100%, n/a
	Physis	Closed, open
	Expansion	Non-expansile, expansile
	Soft tissue mass	Yes, no
	Pathologic fracture	Yes, no

Table 3
Description of the quantitative features extracted by the developed MATLAB module.

Type	Radiomics Feature (Name and citation)	Dimension	Represents
Photometric features	Intensity median inside lesion	1	Quantifies 1st order intensity distribution within the lesion
	Entropy inside lesion	1	
	Proportion of pixels with intensity larger than pre-defined threshold	1	
	Intensity different between lesion and its neighbouring tissue (3 scale analysis)	3	
	Haralick features [11]	12	Captures occurrence of gray level pattern within the lesion.
	Gabor features [34]	32	
	Daubechies features (Wang et al. [31])	324	
	Haar wavelets	1	
	Run Length Matrix [29]	7	
	Local binary pattern [23]	12	Computes marginal distribution of gray values with in lesion
	No of pixels in different Hist. bins	20	
	Edge sharpness	60	Quantifies edge sharpness along the lesion contour
	Histogram on edge	1	
Geometric features	Compactness [9]	1	Describes the morphology of the lesion
	Eccentricity	1	
	Roughness [16]	1	
	Local area integral invariant [13]	15	
	Radial distance signatures [25]	2	

accuracy of image retrieval. The main novelty of the current approach is that an initial automatic prediction of semantic terms is generated from the HFS without any user input other than an ROI about the lesion. Subsequently, the set of predicted semantic terms is improved with iterations of relevance feedback. This can be done by moving the query point in the HFS towards the contour of the user's preference which is defined as the cluster formed in HFS by the relevant images identified by the user feedback (see Fig. 4). After a few iterations with changes of location and contour, the reformulated query point should be close to a convex region of the user's preference in HFS, and, as a consequence, the reformulated query vector should also better capture the semantic axis of the query image.

We used Rocchio's algorithm [5], which is a query modification method that aims to find an optimized query vector, denoted as $\widehat{\gamma_{query}}$, by maximizing the similarity with relevant images while minimizing similarity with irrelevant images as: $\widehat{\gamma_{query}} = \text{argmax}[\text{sim}(\gamma_{\text{query}}, \gamma_{\text{rel}}) - \text{sim}(\gamma_{\text{query}}, \gamma_{\text{nonrel}})]$, where γ_{rel} and $\gamma_{nonrel}$are the set of relevant and irrelevant feature vectors, respectively. In our system, similarity between the vectors ($sim(u, v)$) is measured as cosine similarity (Eq. (1)) as described in the previous section. Under the cosine similarity, Rocchio's algorithm computes the optimized query vector as:

$$\widehat{\gamma_{query}} = \alpha\gamma_{query0} + \beta\frac{1}{|\gamma_{rel}|}\sum_{\gamma_i\in\gamma_{rel}}\gamma_i - \delta\frac{1}{|\gamma_{nonrel}|}\sum_{\gamma_i\in\gamma_{nonrel}}\gamma_i \qquad (2)$$

where γ_{query0} is the original query vector, $\gamma_{rel}$and $\gamma_{nonrel}$are the set of relevant and irrelevant feature vectors, respectively, and α, β, $\delta$are the weights associated with each term. Starting from γ_{query0}, the new query $\widehat{\gamma_{query}}$ moves some distance toward the centroid of the relevant feature vector and some distance away from the centroid of the irrelevant vectors, as presented in Fig. 4. This new refined query vector should better represent user's current information needs and simultaneously should expand the query vector to capture the semantic information integrated in the vector space. This modified query vector can be used for retrieval in the vector space model as well as it can provide a reasonable judgment about the query image semantics.

We constitute the user feedback procedure as:

(1) The query vector (γ_{query_round1}) for round 1 contains only the radiomics features, and the semantic features are padded with zeros;
(2) The 'n' top similar images present in the current dataset are

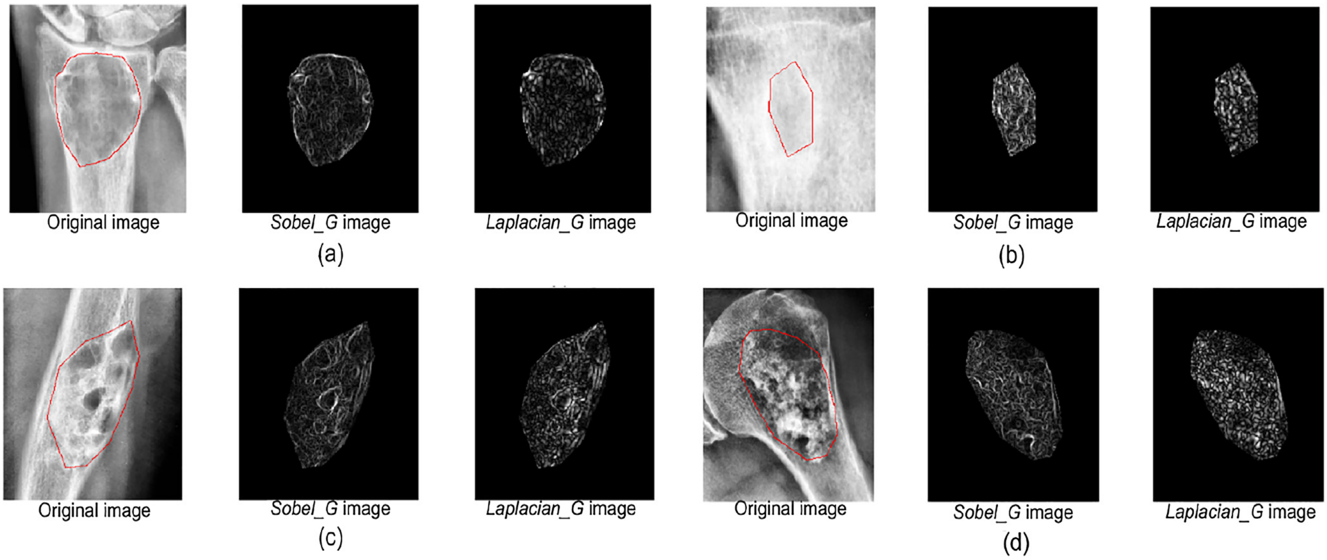


Fig. 3. Sobel_G and Laplacian_G images of different types of bone tumors accentuating different image features of the lesions – (a) osteolytic; (b) osteosclerotic; (c) mixed lytic and sclerotic; (d) mixed lytic and sclerotic.

Table 4
Selected semantic features incorporated in the hybrid feature space.

Targeted Semantic Feature	Values
Distribution	Central, eccentric, n/a
Density	Normal, ground glass, lytic, sclerotic, mixed lytic and sclerotic
Matrix/texture	Normal, osteoid, chondroid, septated, coarse trabeculae, central calcification
Transition zone/border	Geographic 1A (narrow sclerotic), geographic 1B (narrow nonsclerotic), geographic 1C (wide non sclerotic), permeative/ destructive/punched out, unable to determine border or n/a
Cortex	Endosteal scalloping grade: 0 = none, 1 = 0–25%, 2 = 25–50%, 3 = 50–75%, 4 = 75%+, where % = approximate depth of scalloping; cortical thickening, periosteal scalloping (any degree), n/a
Lesion to shaft ratio	0–25%, 25–50%, 50–75%, 75–100%, > 100%, n/a
Expansion	Non-expansile, expansile
Soft tissue mass	Yes, no

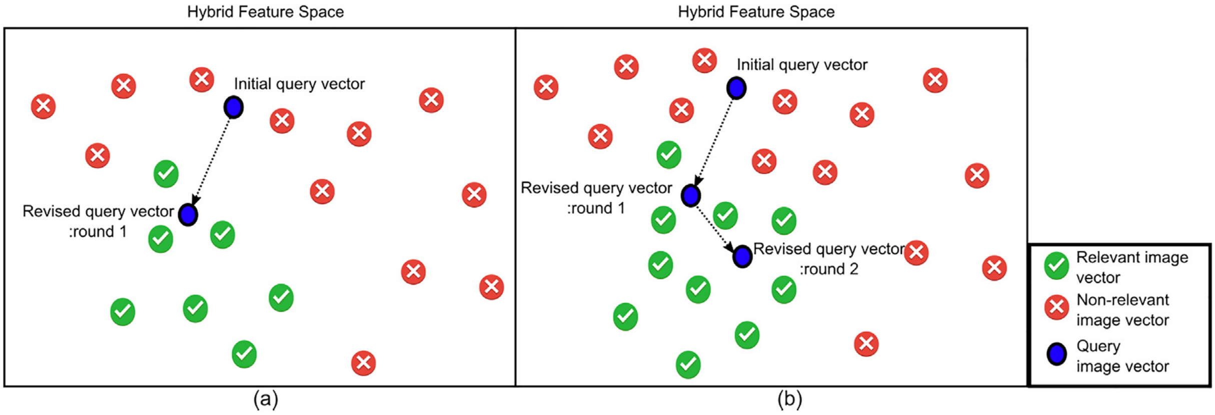


Fig. 4. Conceptual representation of Rocchio's query reformulation. User marks some images as relevant and non-relevant and the initial query vector is moved in response to this feedback – (a) relevance feedback round 1 employing only radiomics features, and (b) relevance feedback round 2, employing both radiomics and semantic features.

retrieved based on cosine similarity with the initial query vector where '*n*' can be specified by the user and is typically ~ 10;

(3) The user gives feedback on the retrieved result set by marking relevant or non-relevant images;
(4) Based on the user feedback, the initial query vector position in the HFS will be altered according to Rocchio's query reformulation algorithm (Eq. (2)),
(5) The new updated query vector ($\widehat{\gamma_{query}}$) incorporates both radiomics and semantic axis of the HFS;
(6) In the subsequent round, the '*n*' top similar images will be retrieved based on cosine distance (σ) between images present in the dataset and the new query vector $\widehat{\gamma_{query}}$.

The relevance feedback procedure (from 2 to 4) can go through one or more iterations, if requested by the user. The process exploits the idea that it may be difficult to formulate a comprehensive representation of the query, particularly if we do not collect semantic features from the user with the query image, but it is natural and efficient for a user to visually judge the similarity of retrieved images. Thus, it should be practical to engage the user in iterative query refinement of this sort. In such a scenario, relevance feedback can also be effective in tracking a user's evolving information need: seeing some images may lead users to refine their understanding of the information they are seeking.

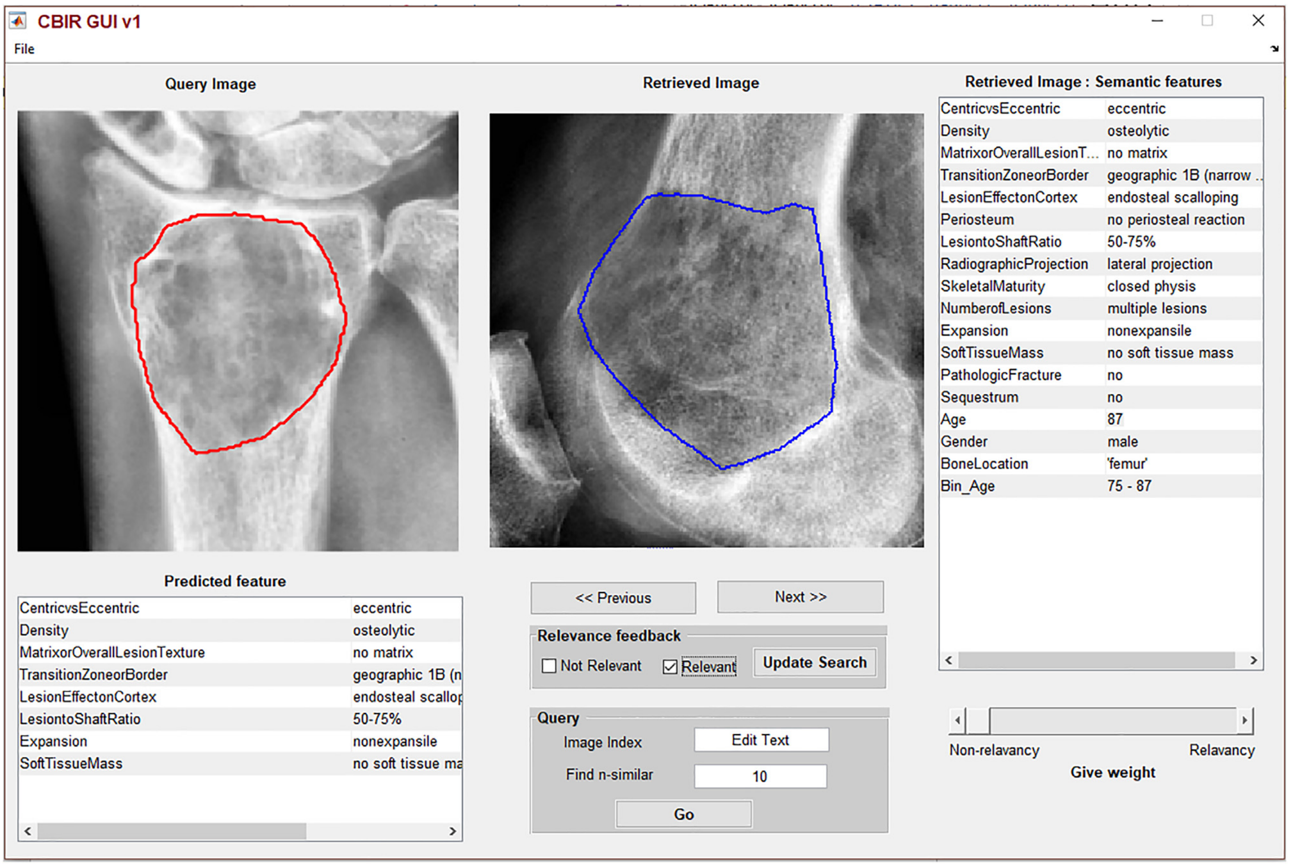
We also exploit the idea of Rocchio's query vector reformulation to predict the semantic features of query image from the altered query vector $\widehat{\gamma_{query}}$. The altered query vector $\widehat{\gamma_{query}}$ is basically composed of modified radiomics and semantic query vector components: $\widehat{\gamma_{query}} \leftarrow [\widehat{R_{Matrix\_query_{1\times496}}}|\widehat{S_{Matrix\_query_{1\times9}}}]$. We extract the $\widehat{S_{Matrix\_query_{1\times9}}}$ component from the optimized query vector, and, owing to the categorical representation of the semantic features, we can project the numeric values back to the original semantic feature space for predicting the semantic features of the query images. Conceptually, the new semantic feature vector ($\widehat{S_{Matrix\_query_{1\times9}}}$) is formulated in a way that it should be close to the mean semantic features of cluster of relevant image set in the semantic feature space. Therefore, if the relevance feedback iteration manages to saturate, at that point the predicted semantic features can capture an implicit correlation between radiomics and semantic axis in the HFS by incorporating the high-level expert preference.

2.2.3. UI prototype implementation

We implemented the components of our system (Fig. 2) as a graphical MATLAB application that allows a user to load query DICOM images, draw ROIs, and collect user feedback on the relevance of retrieved images to their query image. The prototype retrieves the top '*n*' similar images. Fig. 5 shows a screenshot of the application, where a user can initiate a retrieval operation by drawing an ROI on the query image and specify the number '*n*' of retrieved images to be displayed. By default, we set the value of '*n*' to 10, and the weight of the non-relevant images (δ) in Eq. (2) is set to 0. The user can modify the pre-set values at any round during the retrieval process. The markup and the semantic features for each retrieved image are shown within the application.

2.3. Evaluation

As mentioned earlier in Section 2.1, we created a held-out test set by randomly selecting 20 images from the first and second columns of Table 1. To evaluate the performance of our system, two experienced radiologists independently validated the held-out 20 test cases by uploading the images (see Section 2.1). For the sake of uniformity during the experimentation, both radiologists used identical pre-drawn ROIs on the query images. The relevance feedback was collected on the top-10 retrieved images returned by the system. This number was selected



Retrieved Image : Semantic features

CentricvsEccentric	eccentric
Density	osteolytic
MatrixorOverallLesionT...	no matrix
TransitionZoneorBorder	geographic 1B (narrow ..
LesionEffectonCortex	endosteal scalloping
Periosteum	no periosteal reaction
LesiontoShaftRatio	50-75%
RadiographicProjection	lateral projection
SkeletalMaturity	closed physis
NumberofLesions	multiple lesions
Expansion	nonexpansile
SoftTissueMass	no soft tissue mass
PathologicFracture	no
Sequestrum	no
Age	87
Gender	male
BoneLocation	'femur'
Bin_Age	75 - 87

Predicted feature

CentricvsEccentric	eccentric
Density	osteolytic
MatrixorOverallLesionTexture	no matrix
TransitionZoneorBorder	geographic 1B (n
LesionEffectonCortex	endosteal scallop
LesiontoShaftRatio	50-75%
Expansion	nonexpansile
SoftTissueMass	no soft tissue ma

Image Index: Edit Text
Find n-similar: 10

Fig. 5. MATLAB UI snapshot showing semantic feature prediction result and the top ranked similar image of a sample data.

as a realistic number of images for a reader to review quickly as part of an interpretation workflow. We also assume that if most relevant images are not returned within the first 10 hits, the retrieval system is not performing well. For each of the 10 test cases selected as query images, the users determined the number of iterations of relevance feedback that he/she wanted to pursue. The first iteration was without relevance feedback, with ranking of image similarity to the query image based on quantitative features alone. The relevance feedback was provided beginning with the 2nd round of retrieval, indicating for each of the top-10 hits whether it was relevant or irrelevant, resulting in a new ranked list of image query results. The users repeated the process as many times as they desired until they felt they had a visually similar set of query results. According to our experiment with the test set, the system usually saturated at 4th round of iteration. The user can determine the weights of the relevant and non-relevant feedback (β, δ) by interacting with the graphical UI. However, for sake of consistency during the evaluation, we consider only positive feedback, which is equivalent to setting $\delta = 0$ in Eq. (2), and α is set to 1.

To evaluate the performance of the proposed system, we adopted several standard evaluation measures in information retrieval – *Mean Average Precision (MAP)*, which provides a single-figure measure of quality across precision levels for the ranked information retrieval. According to the literature [30], MAP has been shown to have especially good discrimination and stability. MAP is defined for a set of queries as the mean of the average precision scores for each query image:

$$MAP = \frac{\sum_{q=1}^{Q} AvgPrecision(q)}{Q} \tag{3}$$

where Q is the number of queries and $AvgPrecision(q)$ is the average precision score for qth query. For a single query image, Average Precision score is defined as the average of the precision values obtained for the set of top n retrieved images:

$$AvgPrecision = \frac{1}{n}\sum_{i=1}^{n} P@i \tag{4}$$

where $P@i$ is the precision at the ith position when precision is defined by the relevance feedback provided by the rater. For example, if the rater marks the 1st ranked retrieved image 'relevant' and 2nd ranked image 'nonrelevant', $P@i$ values will be $P@1 = 1/1$ and $P@2 = 1/2$. We measure the *AvgPrecision* value of each round for each query, and compute the MAP for the whole test set to report the performance summary of the retrieval.

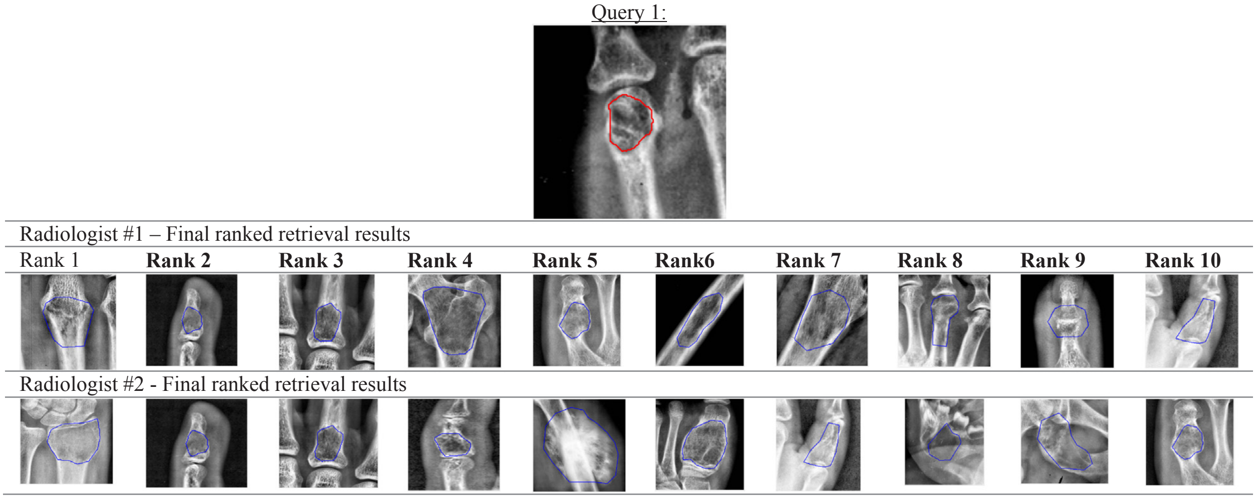
To evaluate the accuracy of semantic feature prediction, we collected the ground truth of eight semantic features (see Table 4) for the test images in the same way as described in Section 2.2.1. We measure prediction accuracy by comparing the predicted value with the ground truth features, and overall prediction score for each semantic feature category is computed as average prediction accuracy over all the test images.

3. Results

3.1. Retrieval of similar images

Table 5 shows the ranked retrieval results of final round (4th) for the first query image (Query 1) for both the radiologists (named as Radiologist #1 and Radiologist #2). As seen from the table, for the same query images, the retrieved similar images may vary between Radiologist #1 and #2, as the feedback differs based on the individual judgment of similarity. Table 6 shows the corresponding results of Rocchio's query vector reformulation at each iteration of the user feedback (we used Principal Component Analysis to reduce the

Table 5
Ranked retrieval results for the first query image (Query 1).


Query 1:
Radiologist #1 – Final ranked retrieval results
Rank 1
Rank 2
Rank 3
Rank 4
Rank 5
Rank6
Rank 7
Rank 8
Rank 9
Rank 10
Radiologist #2 - Final ranked retrieval results

dimensionality of the HFS for visualization purpose). Table 6 also shows the corresponding *P@i* graph for the ranked retrieval results where round 1 does not yet incorporate any user feedback. Therefore, the performance of round 1 represents a baseline CBIR performance with no relevance feedback from the user. As seen from Table 6, with the increasing number of relevance feedback rounds, the query vector is moving towards the cluster of relevant images in the HFS, and there is a consistent improvement in the ranked retrieval performance for both raters – *P@i* is increasing for each ranked retrieval position.

In Table 7, we summarize the retrieval performance in the test set in terms of Average Precision score as defined in Eq. (4). Starting in round 2, given the user relevance feedback and the predicted semantic features, the system's performance improved substantially. Ultimately, retrieval leads to a convergence for most of the query images. To prove this fact statistically, we performed two-sided T-test by considering the null hypothesis (H_0) that baseline (round 1) have identical expected AvgPrecision value as with the subsequent rounds with user feedback, and report the p-values in Table 8. This test assumes that the populations have identical variances by default. As seen from Table 8, the null hypothesis has always been rejected when baseline is compared with the subsequent round of feedback that means there is a significant gain in Average Precision score with iterative user relevance feedback. More importantly, null hypothesis has been rejected with very high statistical significance ($p \ll 0.001$) when baseline is compared with the final round of feedback. The significant improvement in retrieval performance compare to the quantitative baseline CBIR also proves the fact the predicted semantic features could capture the radiological observations of the query image content.

Finally, we compute the MAP value as defined in Eq. (3) and present the MAP value for each round as a bar plot (Fig. 6) which clearly shows the iterations having user relevance feedback produce results much higher in MAP value than the baseline round 1 (~ 0.2). For the final round, the computed MAP value is ~ 0.90 for both the raters which show a great improvement over the initial ~ 0.2 MAP suggesting that relevance feedback on the hybrid feature space CBIR is very helpful to incorporate user preferences.

3.2. *Prediction of the semantic features*

We consider the semantic features computed at the final round of relevance feedback (round 4 in the current study) as the prediction outcome. For each rater, we present the prediction results of the test query images for 8 different semantic features as heat maps in Figs. 7 and 8, where 1 means predicted feature matches with the actual annotation and 0 means no match. The heatmaps indicate that prediction of most of the semantic features matches nicely with the ground truth annotations.

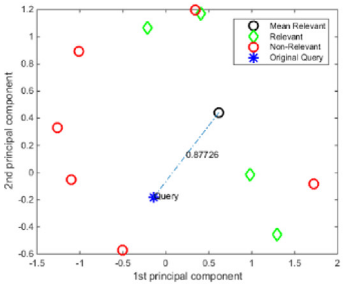
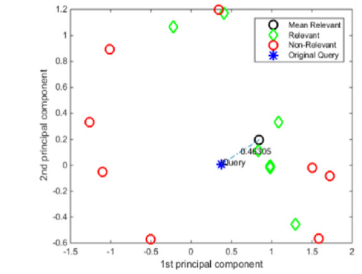
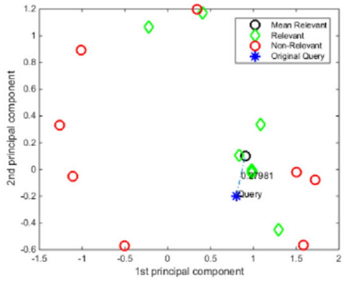
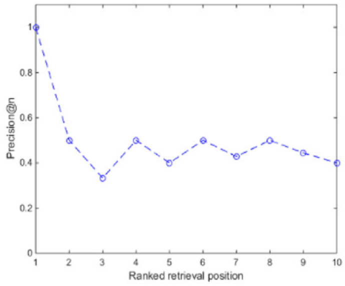
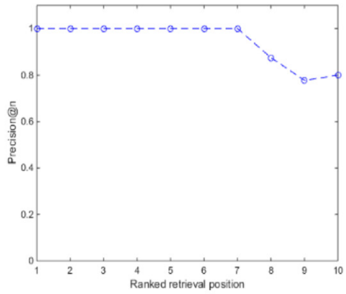
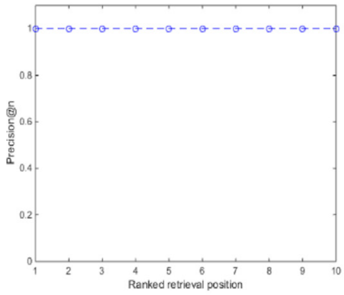
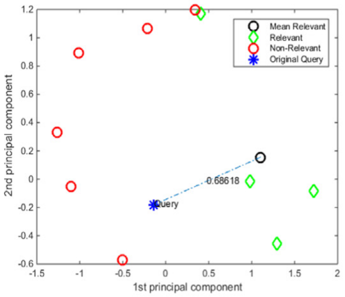
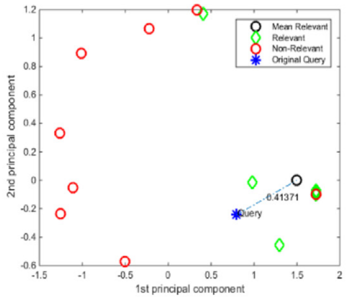
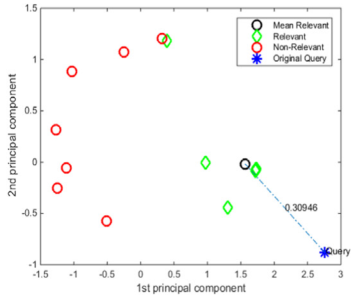
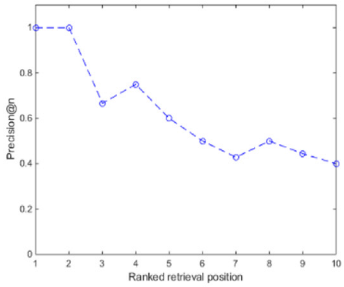
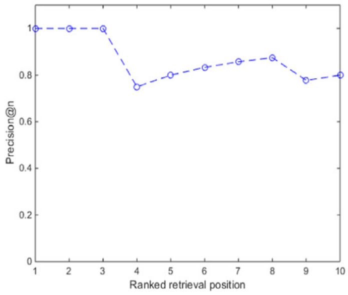
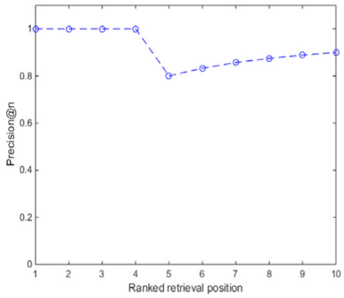
We also computed the average accuracy for the semantic feature prediction in Table 9. Other than '*Distribution*' and '*Transition zone/border*', all the semantic features achieved > 0.8 prediction accuracy. This is a promising result since not only the semantic features are predicted in a semi-supervised iterative way, they were also incorporated on the feature space to improve the baseline CBIR performance. As seen from the retrieval results, the predicted features successfully captured the user preferences as well as increase the precision of the system. The lower accuracy of '*Distribution*' and 'Transition zone/border' may be caused by the fact that they are more challenging to estimate from the cropped images than other features. For instance, information on the bone center is not captured on the cropped images which makes computation of '*Distribution*' semantic feature (sub-category – eccentric/concentric) less conceivable. Thus, the two radiologists did not rely as heavily on these features in refining the results.

4. Discussion and conclusion

In this work, relevance feedback was employed as a mechanism to achieve good accuracy in CBIR by incorporating semantic and quantitative image features, but not requiring the user to provide the semantic features in the query image. The implications of this approach are important in that by removing the requirement of collecting semantic features, it may be practical to introduce our approach to CBIR into the clinical workflow, since busy radiologists rarely have time to input such data into the system. Furthermore, it is relatively easy to judge visual similarity seeing a different set of images, and this process, combined with existing knowledge in previously annotated images, enables our system to perform well in the CBIR task with relevance feedback.

While the ultimate goal of machine learning algorithms and artificial intelligence may be to automatically learn from the data with limited or no human interaction, it needs to be recognized that achieving accurate results for complex image interpretation tasks such as medical images may require higher levels of cognitive processing [12]. Our system shows promising results in this regard by retrieving visually similar images in the challenging area of bone tumor radiography by incorporating a human expert as part of the iterative process. This type of "*human-in-the-loop*" integration or so-called "*interactive machine learning*" has promise for other complex interpretation tasks in radiology. In addition, if humans recognize their integral role in the iteration towards improved retrieval, this may boost their confidence in

Table 6
Round-wise retrieval evaluation measures for the first query image (Query 1) – first row shows the Rocchio query vector reformulation and second row presents *P@i* graph.

Query 1- Radiologist #1		
Round 1 (Baseline CBIR)	**Round 2**	**Round 3**
		
		
Query 1- Radiologist #2		
Round 1 (Baseline CBIR)	**Round 2**	**Round 3**
		
		

applying the results clinically. The latter remains to be investigated and was not part of the current project. In the current work, performance was measured in terms of retrieval of similar images, regardless of the final histological diagnosis. We believe this approach has the potential to aid in differential diagnosis of bone tumors by presenting a number of possible lesion types that match the query image. We integrated quantitative and semantic features extracted from the images to build the hybrid feature space. By incorporating relevance feedback into the hybrid feature vector space model, we not only get a very substantial gain over the baseline CBIR performance but we also predict the semantic features of the query image with good accuracy.

Much prior work in developing relevance feedback methods used non-medical images, but little research has been conducted on the application of relevance feedback in medical images using a fusion of semantic and quantitative features. This is mainly due to the heterogeneity and complexity of medical images and their contents [1,32]. We are aware of an encouraging prior work [19] that incorporates a “soft” prediction of ontological terms that describe the image contents from image features and retrieves similar images by evaluating the similarity that takes into account both image-based and ontological term relations. However, that framework depends on deterministic learning of visual signatures, and no dynamic relevance feedback mechanism is

Table 7
Round-wise retrieval performance measured in-terms of AvgPrecision value.

Query no.	Radiologist #1: round-wise *AvgPrecision* value				Radiologist #2: round-wise *AvgPrecision* value			
	Round 1)	Round 2	Round 3	Round 4	Round 1	Round 2	Round 3	Round 4
	(Baseline CBIR	CBIR + Predicted semantic features + Relevance feedback			(Baseline CBIR	CBIR + Predicted semantic features + Relevance feedback		
1	0.45	0.94	1	1	0.63	0.87	0.96	0.96
2	0.04	0.51	0.56	0.88	0.01	0.85	1	1
3	0.15	0.25	0.37	0.64	0.06	0.13	0.40	0.58
4	0.05	0.23	0.34	0.73	0.3	0.34	0.67	0.70
5	0.63	0.34	0.91	0.98	0.53	0.64	0.76	0.89
6	0.26	0.97	1	1	0.23	0.28	0.85	0.85
7	0.02	0.55	0.51	0.85	0.02	0.16	0.62	0.9
8	0.06	0.53	0.64	0.81	0.1	0.71	0.71	0.87
9	0.14	0.86	0.99	1	0.01	0.21	0.78	1
10	0.22	0.30	0.45	0.98	0.30	0.45	0.62	0.89
11	0.12	0.81	0.71	0.98	0.02	0.55	0.75	0.79
12	0.33	0.85	0.98	1	0.16	0.92	0.94	0.94
13	0.01	0.86	1	1	0.31	0.79	0.89	0.88
14	0.2	0.67	0.9	1	0.15	0.7	0.89	0.96
15	0.1	0.47	0.86	0.99	0.24	0.65	0.86	1
16	0.01	0.69	0.75	0.86	0.1	0.69	0.83	0.86
17	0.24	0.82	0.9	0.96	0.25	0.83	0.92	1
18	0.53	0.60	1	1	0.25	0.61	0.91	0.91
19	0.28	0.85	0.91	0.91	0.23	0.56	0.92	0.96
20	0.21	0.63	0.76	0.92	0.14	0.42	0.63	0.94

involved. This limits the effectiveness and similarity search personalization capability of the system. Relevance feedback applied in the semantic feature space also showed promise for performing retrieval of medical images [17,33]. However, these systems require users to manually interpret the targeted image and report semantic observations, which is time consuming and potentially disruptive of the current radiology workflow.

Our proposed framework addresses some limitations of existing clinical CBIR systems by enabling search for similar images with iterative relevance feedback, and at the same time, predicting the semantic features of the targeted image. The system requires only a query image and an ROI marked by the user, and by iteration with user feedback, it simultaneously predicts the semantic features by query vector reformulation and uses the predicted features to improve the retrieval results. Thus, the primary contribution of this paper is that it proposes a framework where a seamless integration between low-level features and high-level semantic terms exploited through relevance feedback to perform radiological image interpretation and semi-automatic image annotation. In addition, we calibrated state-of-the-art edge-based features and design a case-specific quantitative feature computation pipeline that automatically analyze the gradient of edges present inside ROI and extract statistical measures which appear to correlate well with the major complementary characteristics of tumors - bone density and bone loss.

The study has several limitations. First, our dataset is unique and heterogeneous, with a total of 69 unique bone tumor diagnoses, and some diagnoses have few examples (< 5), which makes accurate retrieval very challenging. Therefore, we randomly selected 20 test cases from the tumor types that have more than five samples and evaluated the CBIR performance with the same tests for both the radiologists.

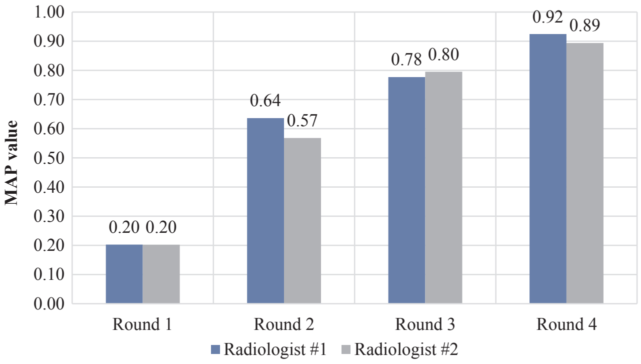
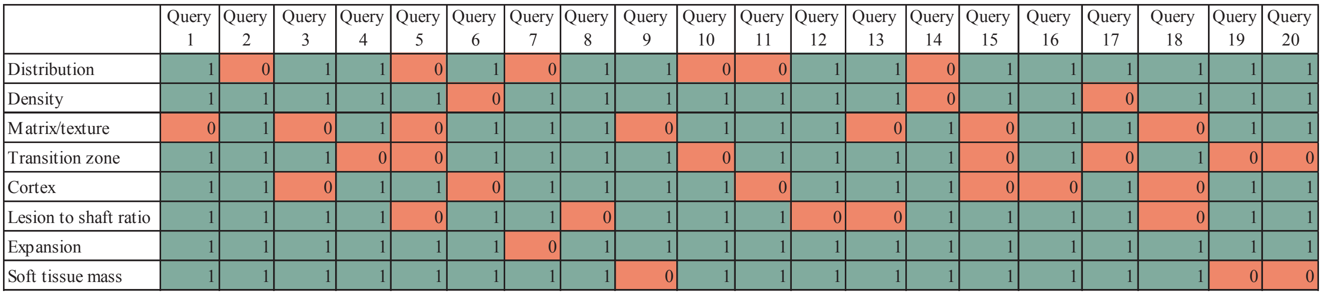


Fig. 6. MAP measures of the retrieval based on the relevance feedback iterations.

A second limitation is that we only collected the feedback on the top 10 similar images returned by the baseline CBIR since our system heavily depends on active user engagement, and it did not seem practical to ask the radiologists to evaluate more than 10 cases at each round. We plan to build a web-based user interface in the future to facilitate collecting relevance feedback from users on a large scale heterogeneous image database. This may yield means to couple the proposed short-term learning RF strategy to a long-term learning strategy that collects and stores the feedback of the radiologists for

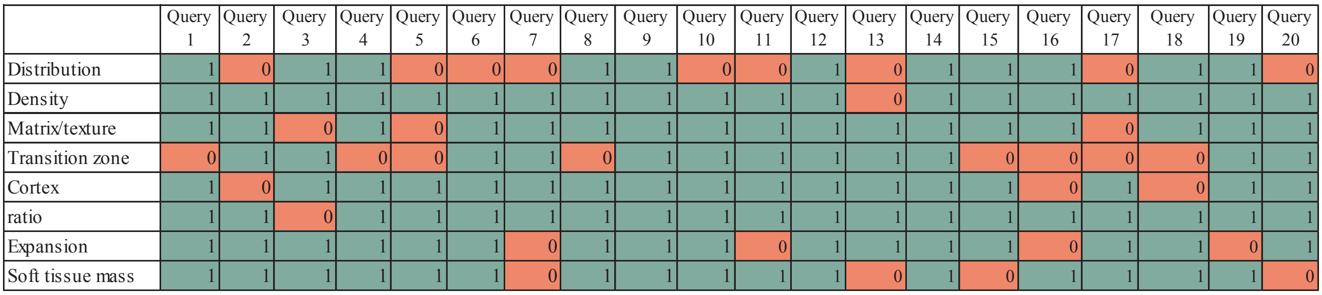
Table 8
Pairwise *t*-test of the AvgPrecision values (Table 9) between the baseline CBIR performance with increasing round of relevance feedback – Null hypothesis (H_0) assumes identical expectation of AvgPrecision value between baseline and subsequent round of relevance feedback.

Raters	Measure	p-Value	Derived hypothesis
Radiologist #1	AvePrecision [Baseline (Round 1) – 1st round of relevance (Round 2)]	0.0066	H_0 rejected with > 99% confidence
	AvePrecision [Baseline (Round 1) – 2nd round of relevance (Round 3)]	0.0003	H_0 rejected with > 99% confidence
	AvePrecision [Baseline (Round 1) - Final round of relevance (Round 4)]	6.14e-08	H_0 rejected with > 99% confidence
Radiologist #2	AvePrecision [Baseline (Round 1) – 1st round of relevance (Round 2)]	0.0439	H_0 rejected with 96% confidence
	AvePrecision [Baseline (Round 1) – 2nd round of relevance (Round 3)]	1.62e-05	H_0 rejected with > 99% confidence
	AvePrecision [Baseline (Round 1) - Final round of relevance (Round 4)]	2.80e-07	H_0 rejected with > 99% confidence



	Query 1	Query 2	Query 3	Query 4	Query 5	Query 6	Query 7	Query 8	Query 9	Query 10	Query 11	Query 12	Query 13	Query 14	Query 15	Query 16	Query 17	Query 18	Query 19	Query 20
Distribution	1	0	1	1	0	1	0	1	1	0	0	1	1	0	1	1	1	1	1	1
Density	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	0	1	1	1
Matrix/texture	0	1	0	1	0	1	1	1	0	1	1	1	0	1	0	1	1	0	1	1
Transition zone	1	1	1	0	0	1	1	1	1	0	1	1	1	1	0	1	0	1	0	0
Cortex	1	1	0	1	1	0	1	1	1	1	0	1	1	1	0	0	1	0	1	1
Lesion to shaft ratio	1	1	1	1	0	1	1	0	1	1	1	0	0	1	1	1	1	0	1	1
Expansion	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Soft tissue mass	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0

Fig. 7. Heatmap showing accuracy of semantic feature prediction for the queries performed by Radiologist #1 – green cells represent correct prediction and white cells represents incorrect prediction.



	Query 1	Query 2	Query 3	Query 4	Query 5	Query 6	Query 7	Query 8	Query 9	Query 10	Query 11	Query 12	Query 13	Query 14	Query 15	Query 16	Query 17	Query 18	Query 19	Query 20
Distribution	1	0	1	1	0	0	0	1	1	0	0	1	0	1	1	1	0	1	1	0
Density	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
Matrix/texture	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
Transition zone	0	1	1	0	0	1	1	0	1	1	1	1	1	1	0	0	0	0	1	1
Cortex	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1
ratio	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Expansion	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	1	1	0	1
Soft tissue mass	1	1	1	1	1	1	0	1	1	1	1	1	0	1	0	1	1	1	1	0

Fig. 8. Heatmap showing accuracy of semantic feature prediction for the queries performed by Radiologist #2 – green cells represent correct prediction and white cells represents incorrect prediction.

Table 9
Overall accuracy of semantic feature prediction for each rater.

Semantic feature	Radiologist #1	Radiologist #2
Distribution	0.7	0.55
Density	0.85	0.95
Matrix/texture	0.65	0.85
Transition zone/border	0.65	0.6
Cortex	0.7	0.85
Lesion to shaft ratio	0.75	0.95
Expansion	0.95	0.8
Soft tissue mass	0.85	0.8

future use of the system by another user [15].

A third limitation is that the benefits of relevance feedback are not guaranteed to be realized for all iterations due to the randomness of human interaction, and the user may need to continue providing relevance feedback until the system reaches a saturation point. It is also possible that the reformulated query vector may actually move farther from the relevant vectors in the feature space after multiple iterations and the optimal saturation point will never be reached, since we only considered the positive relevance for this study. However, in the user interface, we have incorporated a sliding control bar that in future testing will allow the users to tune the weights of positive and negative relevance.

Finally, the relevance feedback approach may not work as well if the relevant vectors create several disconnected clusters within the hybrid feature vector space. Thus, the reported performance may not generalize to other types of datasets or with different set of users. Further studies on larger image collections will be needed. Despite these limitations, we believe our results show the potential value of our approach to enhancing CBIR with relevance feedback of visual similarity as a new direction for helping radiological image interpretation. Looking toward immediate future work, one of our goals is to test the ability of our system to improve diagnostic accuracy across readers with varying levels of clinical experience, with the aim of improving the performance of less experienced users through effective similar image retrieval. We are also interested in studying whether this type of interactive machine learning increases the confidence of readers in applying the results in clinical applications beyond that obtained with entirely machine-based methods.

Acknowledgements

This work was supported in part by grants from the National Cancer Institute, National Institutes of Health (1U01CA190214 and 1U01CA187947) and by the French program Investissement d’Avenir run by the Agence Nationale pour la Recherche (ANR-11-INBS-0006). We acknowledge the use of the bone tumor carefully collected over his career by the late Henry H. Jones, M.D. (aka “*Bones Jones*”).

Conflicts of interest

None.

References

[1] C.B. Akgül, D.L. Rubin, S. Napel, C.F. Beaulieu, H. Greenspan, B. Acar, Content-based image retrieval in radiology: current status and future directions, J. Digit. Imaging 24 (2011) 208–222, https://doi.org/10.1007/s10278-010-9290-9.

[2] I. Banerjee, C.F. Beaulieu, D.L. Rubin, Computerized prediction of radiological observations based on quantitative feature analysis: initial experience in liver lesions, J. Digit. Imaging 30 (2017) 506–518, https://doi.org/10.1007/s10278-017-9987-0.

[3] I. Banerjee, G. Patané, M. Spagnuolo, Combination of visual and symbolic knowledge: a survey in anatomy, Comput. Biol. Med. 80 (2017) 148–157, https://doi.org/10.1016/j.compbiomed.2016.11.018.

[4] P.H. Bugatti, D.S. Kaster, M. Ponciano-Silva, C. Traina, P.M. Azevedo-Marques, A.J.M. Traina, PRoSPer: Perceptual similarity queries in medical CBIR systems through user profiles, Comput. Biol. Med. 45 (2014) 8–19, https://doi.org/10.1016/j.compbiomed.2013.11.015.

[5] Christopher D. Manning, Raghavan Prabhakar, Schütze Hinrich, Introduction to Information Retrieval, Cambridge University Press, Cambridge, 2008.

[6] A. Depeursinge, C. Kurtz, C. Beaulieu, S. Napel, D. Rubin, Predicting visual semantic descriptive terms from radiological image data: preliminary results with liver lesions in CT, IEEE Trans. Med. Imaging 33 (2014) 1669–1676, https://doi.org/10.1109/TMI.2014.2321347.

[7] C. Despont-Gros, H. Mueller, C. Lovis, Evaluating user interactions with clinical information systems: a model based on human–computer interaction models, J. Biomed. Inform. 38 (2005) 244–255, https://doi.org/10.1016/j.jbi.2004.12.004.

[8] B.H. Do, C. Langlotz, C.F. Beaulieu, Bone tumor diagnosis using a naïve bayesian

model of demographic and radiographic features, J. Digit. Imaging 30 (2017) 640–647, https://doi.org/10.1007/s10278-017-0001-7.

[9] R. Duda, P. Hart, Pattern classification and scene analysis, Wiley, 1976.

[10] R. Fitzgerald, Error in radiology, Clin. Radiol. 56 (2001) 938–946, https://doi.org/10.1053/crad.2001.0858.

[11] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybernet. SMC-3 (1973) 610–621, https://doi.org/10.1109/TSMC.1973.4309314.

[12] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inf. 3 (2016) 119–131, https://doi.org/10.1007/s40708-016-0042-6.

[13] B.-W. Hong, E. Prados, S. Soatto, L. Vese, Shape representation based on integral kernels: application to image matching and segmentation, in: Presented at the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006, pp. 833–840. 10.1109/CVPR.2006.277.

[14] Irwin Sobel, History and definition of the sobel operator. Retrieved from the World Wide Web (2014).

[15] W. Jiang, G. Er, Q. Dai, J. Gu, Hidden annotation for image retrieval with long-term relevance feedback learning, Pattern Recogn. 38 (2005) 2007–2021, https://doi.org/10.1016/j.patcog.2005.03.007.

[16] J. Kilday, F. Palmieri, M.D. Fox, Classifying mammographic lesions using computerized image analysis, IEEE Trans. Med. Imaging 12 (1993) 664–669, https://doi.org/10.1109/42.251116.

[17] B.C. Ko, J. Lee, J.-Y. Nam, Automatic medical image annotation and keyword-based image retrieval using relevance feedback, J. Digit. Imaging 25 (2012) 454–465, https://doi.org/10.1007/s10278-011-9443-5.

[18] C. Kurtz, C.F. Beaulieu, S. Napel, D.L. Rubin, A hierarchical knowledge-based approach for retrieving similar medical images described with semantic annotations, J. Biomed. Inform. 49 (2014) 227–244, https://doi.org/10.1016/j.jbi.2014.02.018.

[19] C. Kurtz, A. Depeursinge, S. Napel, C.F. Beaulieu, D.L. Rubin, On combining image-based and ontological semantic dissimilarities for medical image retrieval applications, Med. Image Anal. 18 (2014) 1082–1100, https://doi.org/10.1016/j.media.2014.06.009.

[20] C. Kurtz, P.-A. Idoux, A. Thangali, F. Cloppet, C.F. Beaulieu, D.L. Rubin, Semantic retrieval of radiological images with relevance feedback, in: H. Müller, O.A. Jimenez del Toro, A. Hanbury, G. Langs, A. Foncubierta Rodriguez (Eds.), Multimodal Retrieval in the Medical Domain, Springer International Publishing, Cham, 2015, pp. 11–25, , https://doi.org/10.1007/978-3-319-24471-6_2.

[21] C.P. Langlotz, RadLex: a new method for indexing online educational materials, RadioGraphics 26 (2006) 1595–1597, https://doi.org/10.1148/rg.266065168.

[22] D. Markonis, R. Schaer, H. Müller, Evaluating multimodal relevance feedback techniques for medical image retrieval, Inf. Retrieval J. 19 (2016) 100–112, https://doi.org/10.1007/s10791-015-9260-4.

[23] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 971–987, https://doi.org/10.1109/TPAMI.2002.1017623.

[24] PDQ Pediatric Treatment Editorial Board, Osteosarcoma and Malignant Fibrous Histiocytoma of Bone Treatment (PDQ®): Health Professional Version, PDQ Cancer Information Summaries. National Cancer Institute (US), Bethesda (MD), 2002.

[25] R.M. Rangayyan, Biomedical Image Analysis, CRC Press, 2004.

[26] D.L. Rubin, P. Mongkolwat, V. Kleper, K. Supekar, D.S. Channin, Annotation and image markup: accessing and interoperating with the semantic content in medical imaging, IEEE Intell. Syst. 24 (2009) 57–65, https://doi.org/10.1109/MIS.2009.3.

[27] D.L. Rubin, D. Willrett, M.J. O'Connor, C. Hage, C. Kurtz, D.A. Moreira, Automated tracking of quantitative assessments of tumor burden in clinical trials, Transl. Oncol. 7 (2014) 23–35, https://doi.org/10.1593/tlo.13796.

[28] G.D. Rubin, Data explosion: the challenge of multidetector-row CT, Eur. J. Radiol. 36 (2000) 74–80, https://doi.org/10.1016/S0720-048X(00)00270-9.

[29] X. Tang, Texture information in run-length matrices, IEEE Trans. Image Process. 7 (1998) 1602–1609, https://doi.org/10.1109/83.725367.

[30] E.M. Voorhees, Variations in relevance judgments and the measurement of retrieval effectiveness, Inf. Process. Manage. 36 (2000) 697–716, https://doi.org/10.1016/S0306-4573(00)00010-8.

[31] J.Z. Wang, G. Wiederhold, O. Firschein, S.X. Wei, Content-based image indexing and searching using Daubechies' wavelets, Int J Digit Libr 1 (1998) 311–328, https://doi.org/10.1007/s007990050026.

[32] R. Wang, H. Pan, Q. Han, J. Gu, P. Li, Medical Image Retrieval Method Based on Relevance Feedback, in: S. Zhou, S. Zhang, G. Karypis (Eds.), Advanced Data Mining and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 650–662, , https://doi.org/10.1007/978-3-642-35527-1_54.

[33] Lu. Ye, Hongjiang Zhang, Liu Wenyin, Hu. Chunhui, Joint semantics and feature based image retrieval using relevance feedback, IEEE Trans. Multimedia 5 (2003) 339–347, https://doi.org/10.1109/TMM.2003.813280.

[34] C.G. Zhao, H.Y. Cheng, Y.L. Huo, T.G. Zhuang, Liver CT-image retrieval based on Gabor texture, in: Presented at the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2004, IEMBS '04, 2004, pp. 1491–1494. 10.1109/IEMBS.2004.1403458.

[35] X.S. Zhou, T.S. Huang, Relevance feedback in image retrieval: a comprehensive review, Multimedia Syst. 8 (2003) 536–544, https://doi.org/10.1007/s00530-002-0070-3.