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,