Information Processing and Management 47 (2011) 176–185

Contents lists available at ScienceDirect

Information Processing and Management

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

Brain CT image database building for computer-aided diagnosis using content-based image retrieval

Kehong Yuan [a,*], Zhen Tian [a], Jiying Zou [a], Yanling Bai [b], Qingshan You [b]

[a] Graduate School at Shenzhen, Tsinghua University, China
[b] Tumor Hospital, Harbin Medical University, Harbin, 150080, China

ARTICLE INFO

Article history:
Received 3 August 2008
Received in revised form 5 June 2010
Accepted 23 June 2010
Available online 15 July 2010

Keywords:
Content-based image retrieval
Brain CT image
Database management
Computer-aided diagnosis
Non-negative tensor factorization

ABSTRACT

Content-based image retrieval for medical images is a primary technique for computer-aided diagnosis. While it is a premise for computer-aided diagnosis system to build an efficient medical image database which is paid less attention than that it deserves. In this paper, we provide an efficient approach to develop the archives of large brain CT medical data. Medical images are securely acquired along with relevant diagnosis reports and then cleansed, validated and enhanced. Then some sophisticated image processing algorithms including image normalization and registration are applied to make sure that only corresponding anatomy regions could be compared in image matching. A vector of features is extracted by non-negative tensor factorization and associated with each image, which is essential for the content-based image retrieval. Our experiments prove the efficiency and promising prospect of this database building method for computer-aided diagnosis system. The brain CT image database we built could provide radiologists with a convenient access to retrieve pre-diagnosed, validated and highly relevant examples based on image content and obtain computer-aided diagnosis.

1. Introduction

Content-based image retrieval (CBIR), also known as query by image content, is to search for images with similar content in a large collection of images with computer vision technologies (Gudivada & Raghavan, 1995). Thus, it is paramount for CBIR system to store and manage collected images efficiently. Meanwhile, the use of computer-aided diagnosis (CAD) system by radiologists in practice has led to an increasing demand for efficient medical imaging data storage and retrieval techniques (Müller, Michoux, Bandon, & Geissbuhler, 2004; Siadat, Soltanian-Zadeh, Fotouhi, & Elisevich, 2005). CBIR of medical images based on its specific image features is an important alternative and complement to text-based retrieval using keywords (Shyu et al., 1999; Kim, Cai, Feng, & Wu, 2006; Zheng, Wetzel, Gilbertson, & Becich, 2003). Fig. 1 shows the architecture of CBIR system.

In recent years, various CBIR systems (Liu, Lazar, & Rothfus, 2004; Müller et al., 2004; Smeulders, Worring, Santini, Gupta, & Jain, 2000; Veltkamp & Tanase, 2002) that aim at efficient retrieval of relevant images based on automatically extracted image features have been applied to medical images, which significantly improves the development of CAD system. Müller et al. (2004) have reviewed the increasing research on content-based retrieval approaches in medical applications and found that the majority of the research mainly focuses on particular image content, modality, body region, or pathology with some special algorithms but ignores the database construction, which is the first and key step in CBIR. It is pointed out that feature extraction and database building should and would be paid more attention to. Although Stefanescu and Leventon (2003) introduced the database organization for CBIR system, they did not discuss it in detail.

* Corresponding author.
E-mail address: yuankh@sz.tsinghua.edu.cn (K. Yuan).

0306-4573/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2010.06.003

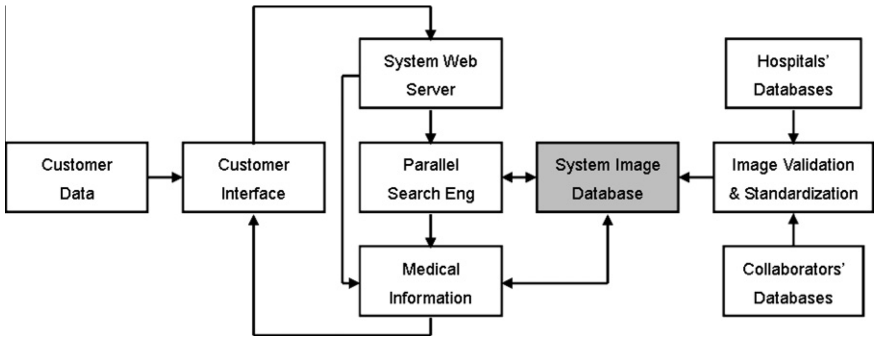

Fig. 1. The construction of CBIR system.

In this study, we summarized our successful experiences of building a database for computer-aided brain CT images diagnosis system based on CBIR technology. We introduced the process of database building including data collection and validation, image normalization, registration and feature extraction in detail. In image normalization, we found an ellipse to best match the whole brain region and use its long radius and short radius to correct the rotation angle and rescale the ellipse region. Besides, we used non-negative tensor factorization (NTF) algorithm to extract image features and also compare this approach with other three feature extraction methods: principal components analysis (PCA), Gabor filters and non-negative matrix factorization (NMF). Three datasets were used to build database respectively and the accuracy and speed of image retrieval with the built database were evaluated.

This paper is organized as follows: In Section 2, the data acquisition and verification process is introduced. Image preprocessing including normalization and registration is described and illustrated in Section 3. Feature extraction and management is summarized in Section 4. Image retrieval experimental results and evaluation are presented in Section 5. Finally, in Section 6, discussion and concluding remarks are provided.

2. Image collection and management

The function of our CBIR system crucially depends on the image database that we built. The database consists of a large collection of normalized, high quality, anonymous medical imaging cases. It is continuously and monotonically updated through a distributed process in which data from luminary sites are securely acquired, cleansed, validated, enhanced and tested. Here, a case is defined as one or more series of scans acquired from the same patient in the same medical institution within a relatively short time. A series is a collection of two-dimensional digital slice images. Images within the same series are acquired during a single scan of a patient. In addition to digital image data, our database also contains a higher-level feature representation or saliency map for each image, clinical reports and diagnosis information for each case.

2.1. Data acquisition and validation

The images of our database are provided by collaborators and partnering hospitals. Currently, the pathologies presented in the database are the diseases often seen in conventional radiological imaging centers. However, it would become increasingly important to preferentially populate the database with specific diseases. To populate the database, several types of data are acquired in our system at the same time, including image data in DICOM format, the radiological report, histopathology report and any other relevant reports if available. All data are collected anonymously by removing all patient confidential information before being sent to our system via DICOM push, CDs, Zip disks, or FTP. All internet-based transfers are done on secure lines. Upon arrival at the system, data is processed through a number of preparatory steps into a standard format. The radiological clinical reports and integer values labeling pathology classifications are amalgamated to the corresponding image data. Next, the aggregate dataset representing a case is processed through a validation and enhancement stage, performed by Registered/Licensed Radiologic Technologists who are also registered in specialty imaging. The technologists review the images for quality (sub-optimal images are discarded), pulse sequence accuracy, and contrast indicators. Finally, clinical image data is reviewed for accuracy and consistency. During the review process, each pulse sequence (series) is reviewed individually and compared to the clinical report with the identified abnormalities in the images. The abnormalities or pathological areas are of particular interest to the image retrieval algorithm, therefore, the precise location of the pathologies are delineated as region of interest (ROIs).

Throughout this process, difficult cases are reviewed by neuroradiologists. When appropriate, ROIs for these difficult cases are delineated under the direction of the neuroradiologist. For further validation, random samples of the recently entered cases are reviewed on a regular basis. Furthermore, regression analysis is performed on every base before entry into the database. Finally, random samples of the entire database are reviewed regularly to ensure the accuracy and integrity of both the data and the processing. The final stage of data loading is to process and store the resulting data in the database.