Accurate Determination of Imaging Modality using an Ensemble of Text- and Image-Based Classifiers

Charles E. Kahn Jr. · Jayashree Kalpathy-Cramer · Cesar A. Lam · Christina E. Eldredge

Published online: 6 July 2011
© Society for Imaging Informatics in Medicine 2011

Abstract Imaging modality can aid retrieval of medical images for clinical practice, research, and education. We evaluated whether an ensemble classifier could outperform its constituent individual classifiers in determining the modality of figures from radiology journals. Seventeen automated classifiers analyzed 77,495 images from two radiology journals. Each classifier assigned one of eight imaging modalities—computed tomography, graphic, magnetic resonance imaging, nuclear medicine, positron emission tomography, photograph, ultrasound, or radiograph—to each image based on visual and/or textual information. Three physicians determined the modality of 5,000 randomly selected images as a reference standard. A “Simple Vote” ensemble classifier assigned each image to the modality that received the greatest number of individual classifiers’ votes. A “Weighted Vote” classifier weighted each individual classifier’s vote based on performance over a training set. For each image, this classifier’s output was the imaging modality that received the greatest weighted vote score. We measured precision, recall, and F score (the harmonic mean of precision and recall) for each classifier.

Individual classifiers’ F scores ranged from 0.184 to 0.892. The simple vote and weighted vote classifiers correctly assigned 4,565 images (*F* score, 0.913; 95% confidence interval, 0.905–0.921) and 4,672 images (*F* score, 0.934; 95% confidence interval, 0.927–0.941), respectively. The weighted vote classifier performed significantly better than all individual classifiers. An ensemble classifier correctly determined the imaging modality of 93% of figures in our sample. The imaging modality of figures published in radiology journals can be determined with high accuracy, which will improve systems for image retrieval.

Keywords Computer vision · Content-based image retrieval · Digital libraries · Image analysis · Image retrieval · Classification · Data mining

Introduction

Image retrieval is a growing area of research in medical informatics [1]. As imaging becomes increasingly prevalent in all aspects of health care and medical research, there has been a substantial growth in the number of biomedical images being created every day. It is important to manage the storage and retrieval of these images, whether they are stored in Picture Archive and Communication Systems, in patient health records, or on the web. Effective image annotation and retrieval can be useful in the clinical care of patients, education, and research [2]. Image retrieval can be used by clinicians to generate differential diagnoses, monitor patient response to therapy, and for quality control. Medical practitioners and trainees have indicated that effective image retrieval can be useful for self-education and for patient education [2].

C. E. Kahn Jr. (✉) · C. A. Lam
Department of Radiology, Medical College of Wisconsin,
9200 W. Wisconsin Ave.,
Milwaukee, WI 53226, USA
e-mail: kahn@mcw.edu

J. Kalpathy-Cramer
Department of Medical Informatics and Clinical Epidemiology,
Oregon Health and Sciences University,
Portland, OR, USA

C. E. Eldredge
Department of Family and Community Medicine,
Medical College of Wisconsin,
Milwaukee, WI, USA

Many areas of medicine, such as radiology, dermatology, and pathology, are visually oriented, yet surprisingly little research has been conducted on how clinicians find and use images. The lack of standardized test collections has limited medical image retrieval research [3]. The annual ImageCLEF conference, begun in 2003 as part of the Cross-Language Evaluation Forum (CLEF), has addressed the need for standardized test collections and multi-institutional evaluation forums, and has grown to become the pre-eminent venue for image retrieval evaluation [4, 5].

A major goal of ImageCLEF has been to foster development and growth of multimodal retrieval techniques, that is, retrieval techniques that combine visual, textual, and other methods to improve retrieval performance. Traditionally, image retrieval systems have been text-based, and have relied on the textual annotations or captions associated with images [3]. Several general purpose commercial systems, such as Google Images (images.google.com) and Yahoo! images (images.yahoo.com), employ this approach, as do radiology-specific image search engines such as ARRS GoldMiner [6] and Yottalook (yottalook.com).

Although text-based information retrieval methods are mature and well researched, they are limited by the quality of the annotations applied to the images. There are other important limitations facing traditional text retrieval techniques when applied to image annotations: (1) image annotations are subjective and context-sensitive, and can be quite limited in scope or even completely absent; (2) manually annotating images is labor- and time-intensive, and can be very error prone; (3) image annotations are very “noisy” if they are automatically extracted from the surrounding text; and (4) there is far more information in an image than can be abstracted using a limited number of words.

In addition to text-based indexing and retrieval of images, new techniques in computer vision have led to a second family of methods for image retrieval: content-based image retrieval (CBIR). In a CBIR system, an image’s visual features—such as color, shape, or texture—are mathematically abstracted and compared to similar abstractions of other images in the database [1, 7]. Typically, such systems present the user with a list of images that are visually most similar to the given image. However, purely content-based retrieval methods have had limited success in the clinical domain as visual similarity between images does not always translate to clinical similarity.

Many studies of web search engines have noted that most users typically view only the first page of search results [8], indicating that high “early precision” is the goal for many of search engines. Precision is the fraction of retrieved documents that are relevant to the search, and

“early precision” is typically defined as precision of the first five to 30 documents retrieved. Combining visual and textual methods to aid in image retrieval has shown to improve search performance, especially for early precision [4, 9]. Often, the information contained in the image itself and the associated annotations or captions can be complementary. Thus, combining these sources can improve search performance.

A commonly stated limitation of web-based image search engines is the lack of precision of the search results. The ability to restrict search results by certain attribute filters would thus be beneficial. Some such attributes associated with the image include the imaging modality, anatomical location, view and the observation or pathological finding contained within the image. Although some of this information may have been associated with the image in the form of DICOM headers or other meta-data at the time of acquisition, this information is often lost when the image is compressed or prepared for presentation in on-line journals or websites. Although the textual annotation associated with the image may no longer contain information about the acquisition modality or view, these attributes can be inferred based on the visual appearance and can be useful in improving the precision of the search results.

Qualitative studies have identified the imaging modality as an important attribute by which clinicians would like to limit the search results [4, 9]. Popular clinical image search engines such as ARRS GoldMiner and Yottalook currently allow search results to be filtered by imaging modality. Previous research has demonstrated that modality-based filtering is a useful technique in improving the precision of search results for clinical queries using a collection of images from the radiology literature [10]. Thus, a primary goal of the present work is to improve the classification of modality of images for this collection of primarily radiographic images, thereby improving the performance of image-retrieval systems.

Materials and Methods

Our investigation was conducted under the auspices of the ImageCLEF 2010 medical image retrieval challenge [5]. The medical task within ImageCLEF was initiated in 2004 with the goal of providing an evaluation forum for researchers in both text- and content-based image retrieval techniques to compare their techniques using a common dataset and a well understood set of evaluation metrics. A variety of data sources, including teaching files and user-supplied images from websites, have been used for this challenge over the years. In 2009, the Radiological Society of North America granted ImageCLEF participants permission to use 77,495 images previously published in its

journals *Radiology* and *RadioGraphics*. Images were identified using the database of the ARRS GoldMiner image search engine [6], and included all figures published in the two journals from January 1999 to June 2008. Information presented to ImageCLEF participants included the image's uniform resource locator, caption text, article title, and PubMed ID. This rich resource consisted of high-quality images as well as highly relevant captions.

In 2010, the medical task within ImageCLEF comprised of three subtasks: a modality classification task, an ad hoc retrieval task and a case-based retrieval task. Based on the contest organizers' experience from previous years as well as user surveys of a set of clinicians [4], imaging modality was identified as an important filter and a precursor to effective retrieval.

ARRS GoldMiner uses heuristic techniques to determine the imaging modality of each figure based on its figure caption and other associated text. For this challenge, we sought to classify images into the eight imaging modalities used by GoldMiner: computed tomography (CT), graphic (e.g., chart and drawing), magnetic resonance imaging, nuclear medicine, positron emission tomography (PET), photograph, radiograph, and ultrasound. GoldMiner's modality assignment for the images in this collection was stored for comparison.

As part of this task, the organizers had provided participants with a set of 2,390 training images that had been manually classified into one of the above classes by the organizers. Another set of 2,630 unclassified "test" images had been provided to the participants, who then submitted "runs" consisting of the purported image class for these images. These image classes were then compared against the manually verified image classes. Participants were allowed to use the caption (text) or the image itself (visual) or both (mixed) for the classification. Participants were encouraged to submit the results of their classifiers on not just the test set of 2,630 images but also for all 77,495 images in the database. Seven research groups submitted a total of 46 runs (21 visual, 15 textual, and 10 mixed)

Of the 46 image classifiers explored by ImageCLEF participants, we limited our analysis to the 17 classifiers that were run on the entire collection of 77,495 images. Of these classifiers, nine used textual information, such as the figure caption, article title, and PubMed metadata. Six classifiers used visual information derived from the image contents. Two classifiers used both text- and image-based information.

The individual classifiers employed a variety of commonly used image processing techniques. Imaging features included local binary patterns, Tamura texture features, Gabor features, GNU Image Finding Tool software, MPEG-7 Color Layout Descriptor and Edge Histogram Descriptor

techniques, Color and Edge Directivity Descriptor, and Fuzzy Color and Texture Histogram using the Lucene Image Retrieval library, Scale Invariant Feature Transform, and various combinations thereof. The classifiers used machine learning techniques such as simple k -nearest neighbors, Ada-Boost, multilayer perceptrons, support vector machines, and a variety of techniques to combine the output from multiple classifiers including those derived from Bayes theory such as product, sum, maximum, and mean rules.

The goal of this work was to explore if the combination of these different techniques in a post hoc manner would provide improved accuracy over any of the individual classifiers. We did not have access to the individual classifiers or algorithms used by the participants, but rather the output of these individual classifiers on the entire set of images.

To create a reference standard to evaluate the performance of the classifiers, a sample of 5,000 reference images was selected randomly from the image collection. Two radiologists and one family medicine physician (all with more than 10 years of experience) reviewed each image and its associated textual information to assign an imaging modality as the reference standard. The reviewers were blinded to the results of the individual and ensemble classifiers; the text-based assignment from the ARRS GoldMiner system was used as a default value. Reviewers were able to view the original published images with their full captions. Each image was reviewed by at least two physicians. If there was a question or discrepancy, the imaging modality was determined by consensus of all three reviewers.

We created two ensemble classifiers to combine the results of the 17 individual classifiers. The simple vote classifier assigned each image to the modality that received the greatest number of votes by the individual classifiers. The weighted vote classifier assigned each image to the modality that received the highest score; each individual classifier's vote was weighted by its accuracy on a set of training cases. The weighted vote classifier was evaluated using tenfold cross-validation; for each fold, the classifier was trained using the remaining nine folds.

For individual and ensemble classifiers, we computed the conventional information-retrieval measures of precision and recall. Precision is defined as the number of relevant images retrieved by a search divided by the total number of images retrieved by that search. It quantifies the quality of the search results, and is analogous to positive predictive value. Recall is defined as the fraction of all relevant images in the database that have been retrieved; it is analogous to sensitivity [11]. The F score is defined as the harmonic mean of precision and recall; thus, for precision (P) and recall (R), we define $F=2 \cdot P \cdot R / (P+R)$.

Table 1 Imaging modalities

Modality	Description	Reference Images
CT	Computed tomography images, including reformatted and 3-D CT images	1,438 (29)
Graphic	Charts, diagrams, illustrations	740 (15)
MRI	MR imaging, including spectroscopy, and functional imaging	1,223 (24)
Nuclear Medicine	Planar and tomographic radionuclide images, excluding PET	52 (1)
PET	Positron emission tomography images, including PET-CT fusion images	90 (2)
Photo	Photographs, including microscopic and endoscopic images	412 (8)
Ultrasound	Ultrasound images, including color Doppler and 3D ultrasound images	371 (7)
Radiography	Radiographic images, including fluoroscopic and angiographic images	674 (13)

The eight modalities considered in this investigation are listed, along with their descriptions and the number (and percentage) of images in the reference sample

The 95% confidence interval (CI_{95}) was computed for all values.

Results

The physician reviewers assigned an imaging modality to each of the 5,000 reference images. The numbers of images in each class are shown in Table 1, along with sample images of each class. The performance of the individual

text-based and visual classifiers is shown in Table 2. The weighted vote classifier is compared to the reference standard as a “confusion matrix” (Table 3).

The ARRS GoldMiner search engine assigned a text-based imaging modality to 4,363 of the 5,000 sample images, of which 3,747 were classified correctly. Its precision was 0.859 and its recall was 0.749, with an F score of 0.800. F score of the other text-based classifiers ranged from 0.402 to 0.810. Image-based classifiers had F scores of 0.691–0.802. The two “mixed” classifiers had F scores of 0.696 and 0.892.

Table 2 Performance of individual and ensemble image-modality classifiers

Type	ID	Correct	Classified	Precision	Recall	F score
Text	GoldMiner	3,747	4,363	0.859	0.749	0.800
	4	3,774	4,999	0.755	0.755	0.755
	26	3,560	5,000	0.712	0.712	0.712
	27	3,465	5,000	0.693	0.693	0.693
	28	2,697	5,000	0.539	0.539	0.539
	29	2,583	5,000	0.517	0.517	0.517
	30	2,549	5,000	0.510	0.510	0.510
	31	2,008	5,000	0.402	0.402	0.402
	42	4,052	4,999	0.811	0.810	0.810
	44	3,652	5,000	0.730	0.730	0.730
Visual	35	3,634	4,996	0.727	0.727	0.727
	36	3,454	4,996	0.691	0.691	0.691
	37	3,634	4,996	0.727	0.727	0.727
	38	917	4,996	0.184	0.183	0.183
	39	3,643	4,996	0.729	0.729	0.729
	41	4,009	4,999	0.802	0.802	0.802
Mixed	3	3,480	4,999	0.696	0.696	0.696
	43	4,460	4,999	0.892	0.892	0.892
Ensemble	Simple vote	4,565	5,000	0.913	0.913	0.913
	Weighted vote	4,672	5,000	0.934	0.934	0.934

The “Type” indicates whether the classifier uses text-based, visual, or mixed input. The table indicates the number of sample images correctly classified and the total number of classified images in the sample. Precision, recall, and F score are defined in the text

Table 3 “Confusion” matrix compares the results of the weighted vote ensemble classifier with the reference standard

		Reference Standard							
		CT	Graphic	MR	Nuc Med	PET	Photo	US	Xray
Weighted vote	CT	1,354	28	47	5	1	46	3	17
	Graphic	3	692	4	–	–	5	–	–
	MR	31	4	1,142	1	1	4	4	1
	Nuc Med	1	–	1	45	1	1	1	1
	PET	6	–	1	–	86	2	–	–
	Photo	11	13	8	–	–	340	–	2
	US	9	–	12	1	1	2	363	3
	Xray	23	3	8	–	–	12	–	650
	Total	1,438	740	1,223	52	90	412	371	674
Total		1,438	740	1,223	52	90	412	371	674
		1,501	704	1,188	51	95	374	391	696
		5,000							

For example, the weighted vote classifier incorrectly identified 28 “Graphic” images as “CT”

The Simple Vote classifier correctly determined imaging modality in 4,565 of 5,000 images to yield an F score of 0.913 (CI_{95} , 0.905–0.921). The weighted vote classifier assigned the correct modality to 4,672 of 5,000 images (F score 0.934; CI_{95} , 0.927–0.941). The F score values are summarized graphically in Fig. 1.

Discussion

The acquisition modality is an important characteristic by which to classify, index, and retrieve images. This study shows that the modality of images published in radiology journals can be identified with high accuracy based on the text of the figure captions and the visual features of the images themselves. Images available on the web and in image libraries may not include header information such as imaging modality, and hence the modality must

be inferred from a figure caption, image annotations, or other metadata.

Ensemble classifiers, which combine the results of individual classification programs, can yield superior results. In this study, two ensemble classifiers both performed significantly better than any individual classifier. The weighted vote ensemble classifier achieved significantly better performance than the simple vote ensemble classifier by weighting the votes of its constituent classifiers based on their performance on the training set.

On an earlier sample of 1,000 images, GoldMiner exhibited precision of 0.972 and recall of 0.864 [12]. That sample included other journals in which the imaging modality was specified more frequently in the figure captions. Limitations of the current study included the use of eight modality categories. PET images and other tomographic radionuclide images, such as single photon emission computed tomography, may not be distinguishable using visual features, and thus must be classified using words that appear in their figure captions. Three-dimensional images generated from CT or MR studies were classified according to their acquisition modality, even though the images were not visually similar to conventional cross-sectional images. Image classification techniques, such as those shown here, may not be useful where the imaging modality is known or can be discerned from the examination’s DICOM header information.

Fig. 1 Classification accuracy (F score) of image modality classifiers. ARRS GoldMiner® (“GoldMiner”) uses text-based information. Data for the other text-based (“Text”), image-based (“Image”), and mixed text- and image-based (“Mixed”) classifiers are shown. Performance of the two ensemble classifiers (“Simple Vote” and “Weighted Vote”) are shown in the right-most columns

Conclusions

Imaging modality is an important part of an image’s metadata and is a useful feature for image retrieval. As the number of clinical images in online journals, image

libraries, and other repositories continues to grow rapidly, determination of imaging modality has become an important area of research. Improvements in image classification can enhance image search engines, and allow search engines to explore a wider array of images. The current work shows that an ensemble classifier that combines both text- and image-based features can achieve high accuracy in classifying images by imaging modality.

Acknowledgments This work was supported in part by a supplement to National Science Foundation (NSF) grant ITR-0325160, a National Library of Medicine Training grant (2 T15 LM07088), a National Library of Medicine grant (5 K99 LM009889), the Swiss National Funds (grant 200020-118638/1) and the BeMeVIS project of the University of Applied Sciences Western Switzerland (HES-SO). Henning Müller of HES-SO helped organize and conduct the ImageCLEF medical image challenge. The authors express their gratitude to the Radiological Society of North America (RSNA) for the use of images published in the journals *Radiology* and *RadioGraphics*, and to the American Roentgen Ray Society (ARRS) for access to the ARRS GoldMiner® database.

References

1. Müller H, Michoux N, Bandon D, Geissbühler A: A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *Int J Med Inform* 73:1–23, 2004
2. Müller H, Despont-Gros C, Hersh W, Jensen J, Lovis C, Geissbühler A: Health care professionals' image use and search behaviour. *Proceedings of Medical Informatics Europe (MIE)*, Maastricht, Netherlands, 2006, pp 24–32
3. Hersh WR, Müller H, Jensen JR, Yang J, Gorman PN, Ruch P: Advancing biomedical image retrieval: development and analysis of a test collection. *J Am Med Inform Assoc* 13:488–496, 2006
4. Müller H, Kalpathy-Cramer J, Kahn CE Jr, Hatt W, Bedrick S, Hersh W: Overview of the ImageCLEFmed 2008 medical image retrieval task. *Lecture Notes in Computer Science* 5706:500–510, 2009
5. Müller H, Kalpathy-Cramer J, Eggel I, Bedrick S, Kahn CE Jr, Hersh W: Overview of the CLEF 2010 medical image retrieval track. http://clef2010.org/resources/proceedings/ImageCLEF2010_medOverview.pdf. Accessed 20 Sept 2010
6. Kahn CE Jr, Thao C: GoldMiner: a radiology image search engine. *AJR Am J Roentgenol* 188:1475–1478, 2007
7. Akgül CB, Rubin DL, Napel S, Beaulieu CF, Greenspan H, Acar B: Content-based image retrieval in radiology: current status and future directions. *J Digit Imaging* 24:208–222, 2011
8. Jansen BJ, Spink A, Saracevic T: Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf Process Manag* 36:207–227, 2000
9. Sedghi S, Sanderson M, Clough P: A study on the relevance criteria for medical images. *Pattern Recognit Lett* 29:2046–2057, 2008
10. Kalpathy-Cramer J, Hersh WR: Automatic image modality based classification and annotation to improve medical image retrieval. *Stud Health Technol Inform* 129(Pt 2):1334–1338, 2007
11. Hersh W: Evaluation of biomedical text-mining systems: lessons learned from information retrieval. *Brief Bioinform* 6:344–356, 2005
12. Kahn CE Jr: Effective metadata discovery for dynamic filtering of queries to a radiology image search engine. *J Digit Imaging* 21:269–273, 2008