Inf Retrieval (2014) 17:229–264
DOI 10.1007/s10791-013-9235-2

Multimodal biomedical image indexing and retrieval using descriptive text and global feature mapping

Matthew S. Simpson · Dina Demner-Fushman · Sameer K. Antani · George R. Thoma

Received: 22 October 2012 / Accepted: 16 October 2013 / Published online: 13 November 2013

Abstract The images found within biomedical articles are sources of essential information useful for a variety of tasks. Due to the rapid growth of biomedical knowledge, image retrieval systems are increasingly becoming necessary tools for quickly accessing the most relevant images from the literature for a given information need. Unfortunately, article text can be a poor substitute for image content, limiting the effectiveness of existing text-based retrieval methods. Additionally, the use of visual similarity by content-based retrieval methods as the sole indicator of image relevance is problematic since the importance of an image can depend on its context rather than its appearance. For biomedical image retrieval, multimodal approaches are often desirable. We describe in this work a practical multimodal solution for indexing and retrieving the images contained in biomedical articles. Recognizing the importance of text in determining image relevance, our method combines a predominately text-based image representation with a limited amount of visual information, in the form of quantized content-based visual features, through a process called global feature mapping. The resulting multimodal image surrogates are easily indexed and searched using existing text-based retrieval systems. Our experimental results demonstrate that our multimodal strategy significantly improves upon the retrieval accuracy of existing approaches. In addition, unlike many retrieval methods that utilize content-based visual features, the response time of our approach is negligible, making it suitable for use with large collections.

M. S. Simpson (✉) · D. Demner-Fushman · S. K. Antani · G. R. Thoma
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
e-mail: simpsonmatt@mail.nih.gov

D. Demner-Fushman
e-mail: ddemner@mail.nih.gov

S. K. Antani
e-mail: santani@mail.nih.gov

G. R. Thoma
e-mail: gthoma@mail.nih.gov

Keywords Multimodal image retrieval · Image indexing · Clustering

1 Introduction

Images and illustrations are sources of essential information within the biomedical domain. For example, images can be found in the articles appearing in biomedical publications and in the case reports contained in electronic health records. Within these resources, images are informative for a variety of tasks, and they often convey information not otherwise mentioned in surrounding text. Following the rapid progress in science and medicine, the volume of biomedical knowledge that is represented visually is constantly growing, and it is increasingly important that we provide a means for quickly accessing the most relevant images for a given information need. Not surprisingly, biomedical image retrieval systems have been developed to address this challenge.

Generally, image retrieval systems enable users to access images using one of several strategies. First, text-based image retrieval methods represent images with their associated descriptions or annotations. Using traditional text-based retrieval techniques, users search for images by providing a system with a description of the image content they desire to retrieve. Second, content-based image retrieval (CBIR) methods represent images with numeric feature vectors describing their appearance. Following a "query-by-example" paradigm, users query a CBIR system with some example image, and the system ranks retrieved images according to their visual similarity with the example. Finally, in an attempt to combine the strengths of these two approaches, multimodal image retrieval systems represent images with both descriptive text and content-based features.[1] This approach allows users to construct multimodal information requests consisting of a textual description of the image content they desire to retrieve that is augmented by visual features extracted from one or more representative images.

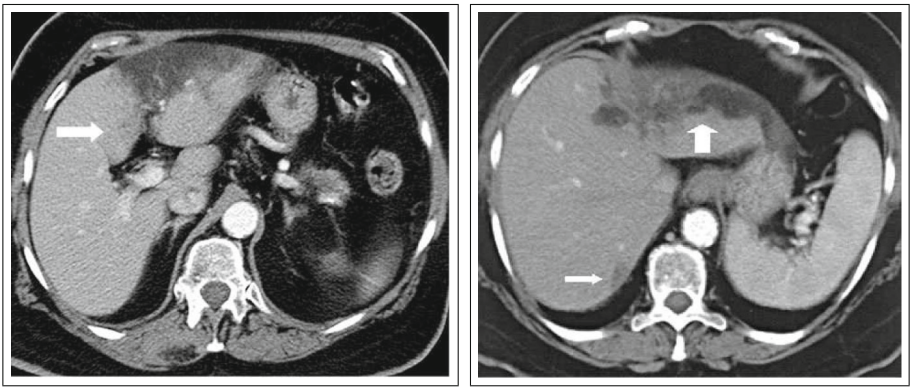
Figure 1 shows an example multimodal information request taken from the 2010 ImageCLEF[2] medical retrieval track data set (Müller et al. 2010). The textual description of this retrieval topic asks for "CT images containing a fatty liver," and the example images are visual depictions of this request: the large gray mass in each of the abdominal CT scans is a liver and the white arrows indicate areas of fat accumulation (Hamer et al. 2006). A multimodal image retrieval system must process the textual description of the topic and extract content-based features from the example images in order to generate a multimodal query.

As evidence of the significant contribution the combination of text-based and content-based features can provide, a variety of multimodal image retrieval strategies have been proposed. Unfortunately, developing a multimodal retrieval system becomes challenging when the usability of the system and the quality of the results are primary and equal concerns. The limitations of existing methods can be attributed to deficiencies in the following areas:

[1] In this work, we use the term "multimodal" to specifically refer to retrieval techniques that combine both textual and visual information. We use the term "content-based features" to denote only the visual content of images, as it is also possible to extract content-based textual features from images that have overlain text.

[2] ImageCLEF is a community-wide forum for evaluating image retrieval methods, and we discuss the 2010 and 2012 collections in more detail in Sect. 5.

Example Images



Textual Description

Topic 6: CT images containing a fatty liver

Fig. 1 Example multimodal topic taken from the 2010 ImageCLEF medical retrieval track data set

– *Practicality* Comparing content-based features across images generally requires significantly more computation time than comparing text-based features across documents. Because of this expense, some multimodal systems are unable to retrieve relevant images from large collections in an amount of time consistent with traditional text-based retrieval, the efficiency of which many users have come to expect.
– *Precision* Effectively combining content-based and text-based features in a way that actually improves retrieval precision has proven to be challenging. In particular, for the task of retrieving images from biomedical articles, many multimodal systems are unable to significantly improve upon the precision of simple text-based methods.

The above drawbacks were recognized by Datta et al. (2008) in their survey of image retrieval trends when they remarked that "the future lies in harnessing as many channels of information as possible, and fusing them in smart, practical ways to solve real problems."

We describe in this work global feature mapping (GFM), a practical solution for performing multimodal biomedical image retrieval. GFM advances the state of the art by enabling efficient access to the images in biomedical articles while simultaneously improving upon the average retrieval precision of existing methods. Recognizing the importance of text in determining image relevance, GFM combines a predominantly text-based image representation with a limited amount of visual information through the following process:

1. Our system extracts a set of global content-based features from a collection of images and groups them into clusters.
2. Our system maps each cluster to a unique alphanumeric code word that it then assigns to all images whose features are members of the cluster.
3. Our system combines the code words assigned to an image with other text related to the image in a multimodal surrogate document that is indexable with a traditional text-based information retrieval system.
4. Our system searches the index using a textual query generated from a multimodal topic by first assigning code words to the topic's example images and then by combining these words with the topic's textual description.

We experimentally validated the success of GFM using the 2010 and 2012 (Müller et al. 2012) ImageCLEF medical retrieval track data sets. Our results show that on both collections, GFM achieves statistically significant improvements in mean average precision (averaging 6.31 %, $p < 0.05$) over a competitive text-based retrieval approach. When configured for performing content-based retrieval, GFM also demonstrates a significant improvement in precision compared with standard methods. As evidence of its practicality, our results show that GFM requires a response time comparable to that of text-based retrieval, suggesting that it is an appropriate technique for indexing large image collections.

To further demonstrate its practicality, we implemented GFM in two information retrieval systems: a general purpose system based on the vector space retrieval model and a biomedical system that utilizes a probabilistic retrieval model. We obtained statistically significant improvements using both systems. Finally, we have incorporated GFM into the OpenI system (Demner-Fushman et al. 2012). OpenI is a multimodal biomedical image retrieval platform that currently indexes over one million images taken from the articles included in the open access subset of PubMed Central.® OpenI is a publicly accessible service[3] developed by the U.S. National Library of Medicine.

The remainder of this article is organized as follows. We review in Sect. 2 existing image retrieval strategies. We present the details of GFM in Sect. 3, and we discuss the reuse of text-based systems for performing multimodal retrieval in Sect. 4. We describe our evaluation of GFM on the ImageCLEF data sets in Sect. 5, our two prototype implementations of GFM in Sect. 6, and our experimental results in Sect. 7. Finally, we discuss the significance of our results in Sect. 8.

2 Background and related work

Image retrieval is a broad and well-researched topic whose scope is far greater than our immediate task of efficiently retrieving biomedical images. In this section, we first briefly review the general strengths and weaknesses of well-known text-based and content-based image retrieval strategies. We then discuss current work related to multimodal retrieval.

2.1 Text-based image retrieval

Text-based image retrieval systems represent images using descriptive text. For example, the images contained in biomedical articles can be represented by their associated captions. Using this text, a collection of surrogate documents is created to represent a given set of images. These documents are indexed with a traditional text-based information retrieval system, and they are searched using text-based queries.

There are several advantages to indexing and retrieving images using text. First, text-based retrieval is a well-understood topic, and the knowledge gained in this area is easily applied to the retrieval of images when they are represented by related text. Second, text-based retrieval is efficient. Because words are discrete data, image surrogates can be indexed in data structures that allow for low latency retrieval, such as inverted file indices. Additionally, because text-based image queries are typically sparse, only a fraction of the surrogates in an index must be scored and ranked for a given query. Finally, a text-based representation allows for semantic image retrieval, enabling us to search for images by providing a system with a description of the content we desire. By "semantic image

[3] http://openi.nlm.nih.gov/.

retrieval" we are referring to the ability of a system to reason beyond the surface form of an information request. For example, it is common for text-based biomedical retrieval systems to perform query expansion using ontological resources such as the Unified Medical Language System®(UMLS®) (Lindberg et al. 1993).[4] A query for the term "heart attack" could then be used to retrieve documents mentioning the term "myocardial infarction" since these terms, although having different surface representations, refer the same concept.

Unfortunately, text can often be a poor substitute for image content. For example, authors sometimes do not write meaningful captions for the images they include in their articles. For a text-based image retrieval system to be effective, the surrogate documents with which it represents images must adequately reflect the content that is requested of it.

2.2 Content-based image retrieval

CBIR systems represent images as numeric vectors. These multidimensional visual descriptors characterize features of the images' content, such as their color or texture patterns. A CBIR system queries a collection of images using an example image, and it ranks the images according to their visual similarity with the example.

The advantage of CBIR systems compared to text-based image retrieval systems is their ability to perform searches based on visual similarity. Such an ability is useful, for example, for finding within a collection of images all images that are nearly identical to one another, regardless of the context in which they appear. Müller et al. (2004) survey the use of CBIR systems in medical applications.

However, there are several disadvantages to retrieving images using their content. First, the visual similarity of a retrieved image with some example image is not always indicative of its relevance to a query. Whereas text-based image retrieval systems provide a means to access relevant images using descriptive text, semantic retrieval is difficult to achieve using visual similarity alone.[5] Second, CBIR systems are usually not as efficient as text-based image retrieval systems. Because visual descriptors can be highly dimensional, dense, and continuous-valued, computing the similarity between any two images can often be a computationally intensive task.[6]

CBIR systems judge the similarity of images using a distance measure computed between their extracted visual descriptors. Although many distance measures have been proposed (e.g., Rubner et al. 2000), below we illustrate visual similarity using Euclidean distance. Assume the vectors $\mathbf{f}_q^x$ and $\mathbf{f}_j^x$ represent the visual descriptors of some feature x extracted for images I_q and I_j. The similarity of these two images for feature x is defined as:

$$\text{sim}(I_q, I_j) = 1 - \frac{\|\mathbf{f}_q^x - \mathbf{f}_j^x\|}{\max_{m,n} \|\mathbf{f}_m^x - \mathbf{f}_n^x\|} \quad (1)$$

Thus, their similarity is equal to one minus the normalized Euclidean distance between their visual descriptors. The denominator of the above function computes the maximum

[4] The UMLS is a collection of controlled vocabularies in the biomedical domain, and its Metathesaurus® represents synonymy relationships among the terms in the various vocabularies.

[5] Although it is distinct from image retrieval, content-based image annotation can provide an efficient means of accessing images by high-level concepts, and we discuss work related to image annotation in Sect. 3.5.

[6] This computation can be lessened through the use of spatial data structures, approximate similarity models, or "visual words," and we discuss the advantages of these methods in Sects. 2.2.1–2.2.3.

distance between the descriptors extracted for all images I_m and I_n within some collection. This min–max normalization ensures the computed value is always defined on the interval [0,1].

A naïve content-based retrieval approach involves using the above equation to compute the visual similarity between an example image and every image within some collection. The images are then ranked by sorting them in decreasing order of their similarity with the example. We refer to this approach as the brute-force retrieval strategy. Although the brute-force strategy is adequate for small image collections, it does not scale to large collections, and its use is impractical for many retrieval tasks. A variety of techniques exist for reducing the cost associated with the brute-force retrieval approach. Below, we briefly review some of these general approaches before discussing in Sect. 2.3 work specifically related to GFM's multimodal retrieval strategy.

2.2.1 Exact representations

Spatial data structures provide an efficient means of storing and retrieving visual descriptors, and they are well-understood within the metric space approach to similarity search (Zezula et al. 2006). These data structures are commonly organized as search trees created by first recursively partitioning a space into regions and then assigning objects to these regions. While many spatial data structures share similar organizations, they often differ in the way in which they partition a space. Examples of spatial data structures include vantage-point trees (Yianilos 1993), generalized hyperplane trees (Uhlmann 1991), geometric nearest-neighbor access trees (Brin 1995), and M-trees (Ciaccia et al. 1997). Within Euclidean spaces, *k*-d trees (Bentley 1975) and R-trees (Guttman 1984) are common.

Unfortunately, the use of such data structures for indexing image content is not always appropriate. We frequently represent the content of an image collection using more than one feature. Because the descriptors of these features are often highly dimensional, the improvement in response time realized through the use of spatial data structures does not always justify their use. Spatial data structures generally perform no better than brute-force algorithms for finding nearest neighbors in highly dimensional spaces (Indyk 2004).

2.2.2 Inexact representations

Dimensionality reduction techniques, especially when they are used in combination with spatial data structures, can effectively reduce the cost associated with maintaining multidimensional data. By representing visual descriptors with fewer attributes, the response time incurred by spatial data structures can be significantly improved. Examples of dimensionality reduction techniques include principal component analysis (Ng and Sedighian 1996), singular value decomposition (Pham et al. 2007), self-organizing maps (Kohonen 2001), multidimensional scaling (Beatty and Manjunath, 1997), and locality-sensitive hashing (Indyk and Motwani 1998).

However, inexactly representing the visual descriptors of a collection of images forces a trade-off between retrieval precision and efficiency. Because descriptors must be transformed to enable their efficient storage and retrieval, we can no longer rank images according to their exact similarity with a query. Instead, we must rely on an approximate similarity, which may result in a significant reduction in retrieval precision. Moreover, as we increase the number of descriptors with which we represent images, or the dimensionality of these descriptors, we can expect the precision of an approximate similarity

search to worsen. The lower-dimensional approximation becomes increasingly inexact as we increase the descriptiveness of the original representation.

2.2.3 Bag of visual words representations

Of the existing techniques for reducing the response time of brute-force CBIR, the use of "visual words" (e.g., Yang et al. 2007) is most similar to GFM's processing of content-based image features. Using the bag of visual words (BVW) approach, an image is first segmented into a set of regions, commonly by overlaying a regular grid onto the image. Alternatively, an interest point detector can be used to detect salient local patches within an image (Nowak et al. 2006). Widely used interest point detectors include the Harris affine region detector (Harris and Stephens 1988), Lowe's difference of Gaussians detector (Lowe 2004), and the Kadir–Brady saliency detector (Kadir and Brady 2001). Once an image has been segmented into regions, local features, especially SIFT (Lowe 1999) features, are extracted from each region, and these features are mapped to visual words. An image is represented as a collection of the words assigned to its constituent regions, which is a description of the image that can be efficiently maintained in an inverted file index.

Having been inspired by the success of text-based retrieval, numerous content-based retrieval strategies have been proposed that implement the BVW approach. A well-known example is the Video Google system (Sivic and Zisserman 2003), which utilizes the BVW strategy and inverted file indices to efficiently retrieve all occurrences of user-outlined objects in videos. Another example demonstrating the use of text retrieval models for performing content-based retrieval is the Mirror DBMS (de Vries 1999), which generates visual words for independent local feature spaces and then applies a retrieval model based on the INQUERY system (Callan et al. 1992). de Vries and Westerveld (2004) describe a similar content-based retrieval system based on the language modeling approach of information retrieval. Finally, the Viper project (Squire et al. 2000) demonstrated that inverted file indices permit the use of extremely high-dimensional feature spaces for performing content-based image retrieval. The MedGIFT (Müler et al. 2003) system is a more recent incarnation of this work that has been adapted to the biomedical domain.

The difference between BVW representations and the content-based feature processing of GFM can be summarized as follows. BVW representations decompose images spatially into patches, representing the *local* features extracted from each patch with a word. Conversely, GFM decomposes images conceptually into complimentary "views" of the images' content according to various *global* features. Each view is then represented as a set of words. The two models are orthogonal, with the former mapping local patches within an image to words and the latter mapping global views of an image to words.

BVW approaches provide a convenient representation for region-based computer vision tasks where the spatial orientation of an image's local features is not an essential consideration. For example, BVW models are commonly used for categorizing the objects within images. However, for retrieval tasks in which the overall appearance of images is important, BVW models often do not perform well in isolation, and the use of global features can improve performance. Within the biomedical domain, images of a particular medical imaging modality commonly exhibit a similar global appearance. For example, physicians usually perform chest X-ray using standard medical imaging equipment on patients oriented in the same direction. A result of this uniform examination procedure is that chest X-ray images can be distinguished from other medical imaging modalities using global features, such as color and texture. For multimodal retrieval tasks within the biomedical domain, it is often not necessary to consider the local features of images within a

particular modality. Detecting nodules within a chest X-ray query, for example, is not needed if the query text already mentions the concept "tuberculoma," a pulmonary nodule found in patients having tuberculosis. For such retrieval tasks, GFM is an appropriate technique for efficiently combining the exact representation of an image's global content-based features with descriptive text so as to improve average retrieval precision.

2.3 Multimodal image retrieval

Multimodal image retrieval systems, the last of the three image retrieval methods we will discuss, represent images as a "fusion" of descriptive text and numeric feature vectors. Fusion can either be performed *early* in the analysis process by creating a unified data representation or *late* in the process, after each data type has been analyzed independently. GFM is an instance of early fusion because it combines an image's text-based and content-based features into a single indexable representation. Retrieval strategies that filter or re-rank images retrieved using a text-based query based on their visual similarity with some example image are instances of late fusion. These methods perform retrieval separately for each modality and later merge the results into a single ranked list of images. Atrey et al. (2010) survey fusion methods that have been proposed for a variety of different data types.

Because multimodal image retrieval systems combine the aforementioned text-based and content-based image retrieval approaches, these systems inherit the strengths and weaknesses of each method. Advantages of multimodal retrieval include the ability to search for images both semantically and by visual similarity. A disadvantage is the inherent difficulty in determining an effective fusion strategy that simultaneously improves retrieval precision while remaining practical for use in real systems.

One of the most active areas of multimodal image retrieval research has been the biomedical domain. Although an exhaustive account of the multimodal biomedical image retrieval strategies is not feasible, a popular topic has been the retrieval of images from biomedical articles. An appropriate starting point for surveying this work is Müller et al.'s (2010a) retrospective of the ImageCLEF evaluations. This volume describes the various ImageCLEF tracks and the evolution of strategies used by the ImageCLEF participants. A recurring theme—not only of the medical retrieval track, but of the other tracks as well—is the difficulty encountered by the participants in meaningfully combining text-based and content-based image features. The prototype implementation of GFM is based on our own past experiences, documented by Simpson et al. (2009, 2010, 2011, 2012a), at developing multimodal retrieval strategies for these evaluations.

While many multimodal fusion-based retrieval strategies have been proposed within the biomedical domain, we review the following as being representative methods that have also been evaluated on the ImageCLEF data sets. Kalpathy-Cramer and Hersh (2010) demonstrate an effective late fusion approach for improving the early precision of a medical image retrieval system. The method first assigns image modality labels (e.g., X-ray) to a collection of images based on their content-based features, and it then uses these labels to re-rank images retrieved using text-based queries. Clinchant et al. (2010) and Alpkocak et al. (2012) also describe techniques that use image modality to re-rank results obtained by a text-based image search. Whereas the above approaches perform late fusion using image modality, Demner-Fushman et al. (2009) describe a medical image retrieval system that first performs a text-based query to retrieve an initial set of images and then re-ranks the retrieved images according to their visual similarity with an example query image. Similarly, Gkoufas et al. (2011) perform brute-force CBIR to re-rank the one thousand highest ranked images retrieved using a text-based retrieval approach. Caicedo

et al. (2010) utilize latent semantic kernels to construct combined text-based and content-based feature vectors, which they then use for performing a brute-force retrieval strategy. Finally, Rahman et al. (2010) describe a fusion-based query expansion method, and Zhou et al. (2010) evaluate the effectiveness of classical information fusion techniques for biomedical image retrieval.

GFM is distinct from the above multimodal approaches in several ways. First, whereas the above methods primarily rely on late fusion techniques to filter or re-rank the results of text-based retrieval, GFM is an early fusion approach. GFM's combination of image code words with image-related text enables the creation of multimodal surrogate documents that are indexable by traditional text-based retrieval systems. The reuse of text-based systems for performing multimodal retrieval contributes to GFM's low search latency and suggests that it is an appropriate technique for indexing large image collections. Second, our results demonstrate that GFM consistently achieves statistically significant improvements in retrieval precision over text-based approaches whereas existing methods show mixed results.

3 Global feature mapping

GFM is a practical solution for enabling the retrieval of biomedical images using both descriptive text and visual similarity. By mapping the exact representation of an image's global content to code words, and then by combing these words with other image-related text, GFM creates a multimodal image representation that is efficiently indexed and retrieved using a traditional text-based information retrieval system. Reusing a text-based system for performing multimodal image retrieval ensures the efficiency of GFM and improves upon the average retrieval precision of existing methods.

Below, we detail GFM's image indexing and retrieval process. The primary components of this process include (1) a method for generating a "codebook" of words with which to represent the global features extracted from a collection of images, (2) a method for assigning these code words to images, (3) a method for indexing the images' assigned code words with their related descriptive text, and (4) a method for querying the resulting multimodal index. We follow this section with a discussion of the treatment of image code words in a traditional text-based information retrieval system.

3.1 Codebook generation

GFM's codebook generation process defines a mapping of the global content of a collection of images to a set of indexable code words. Assume a collection of images $\{I_1, I_2, \ldots, I_m\}$ and a set F of global content-based features, such as color and texture.[7] We represent each image in the collection as a set of numeric vectors extracted for each of the features:

$$I_j = \left\{ \mathbf{f}_j^x : x \in F \right\} \tag{2}$$

[7] GFM is generally applicable and does not require the use of specific content-based features. However, the features used with GFM must be representable as numeric vectors. Refer to Sect. 6.1 for a description of the features we use with our prototype implementation.

The vector $\mathbf{f}_j^x$ is a visual descriptor of feature x of image I_j, $1 \leq j \leq m$. For each of the global features, the codebook generation process clusters the corresponding vectors extracted from the images in the collection and then maps the resulting cluster centroids to unique code words. The complete mapping of cluster centroids to code words defines GFM's codebook.

The codebook generation process proceeds as follows. First, GFM optionally partitions each vector $\mathbf{f}_j^x$ into p lower-dimensional vectors of equal dimensionality. $\mathbf{f}_j^x$ is written in terms of its constituent partitions as:

$$\mathbf{f}_j^x = \begin{bmatrix} \mathbf{f}_{1,j}^x & \mathbf{f}_{2,j}^x & \cdots & \mathbf{f}_{p,j}^x \end{bmatrix} \tag{3}$$

where the row vector $\mathbf{f}_{l,j}^x$ is partition l of $\mathbf{f}_j^x$, $1 \leq l \leq p$. The dimensionality of each of these lower-dimensional vectors is equal to $1/p$ times the dimensionality of the original descriptor. Thus, the lower-dimensional vector $\mathbf{f}_{l,j}^x$ contains dimensions of the original vector that are within the range $[lp - p + 1, lp]$. GFM partitions descriptors when their dimensionality and the number of images in the collection combine to make clustering the vectors prohibitively expensive with available resources. Also, because partitioning increases the number of vectors representing each image, it increases the number of ways in which these images can differ, thereby improving the intracluster ranking of the retrieved images (Sect. 7.2.4).

GFM's use of lower-dimensional feature vectors is related to the notion of product quantization (Jégou et al. 2011) and the dimensionality reduction technique proposed by Ferhatosmanoglu et al. (2001) that partitions vectors after having transformed them using the Karhunen–Loeve transformation (KLT). However, unlike these methods, GFM does not partition visual descriptors so as to improve the performance of approximate nearest neighbor search. Instead, GFM partitions the vectors and maps them to code words as a practical means of combining visual and textual information in a form indexable by a traditional text-based retrieval system. We retrieve images not by approximating the Euclidean distance between their descriptors, but by relying upon the underlying text-based retrieval model. For example, the well-known vector space model computes the cosine distance between *tf* and *idf* term vectors. Using GFM, each of these vectors is constructed from the combined term statistics of an image's code words as well as its related text.

After GFM partitions the extracted visual descriptors, the clustering process begins. GFM requires a centroid-based algorithm, such as k-means (Lloyd 1982), to cluster each set of lower-dimensional vectors corresponding to a given partition and feature. Assume $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m\}_l^x$ is the set of lower-dimensional vectors representing partition l of the visual descriptors extracted from the collection of images for feature x. We denote the clustering of these vectors as $\{C_1, C_2, \ldots, C_k\}_l^x$, where $C_{i,l}^x$ is the set of vectors belonging to cluster i, $1 \leq i \leq k$. We denote the centroid of cluster $C_{i,l}^x$ as the vector $\mathbf{c}_{i,l}^x$.

Once the clustering process is complete, GFM generates the codebook. GFM stores in the codebook a mapping from each cluster centroid to a unique code word. Because each centroid $\mathbf{c}_{i,l}^x$ is uniquely identified by the feature x and the tuple (i, l), GFM combines these values to construct textual code words of the form "x:kipl." For example, having extracted the color layout descriptor (CLD) (Chang et al., 2001) for all images in the collection, GFM maps the centroid $\mathbf{c}_{1,2}^{\text{cld}}$ to the text string "cld:k1p2." In this way, each code word in GFM's codebook is uniquely associated with a given feature, partition, and cluster. Although various other techniques can be envisioned for generating unique code words, GFM follows the aforementioned strategy to map images to sets of words.

The diagram shown in Fig. 2 summarizes GFM's codebook generation process. Assume $\{a, b, c\} \subseteq F, 1 \leq l \leq p$, and $1 \leq i \leq k$. A collection of images $\{I_1, I_2, \ldots, I_m\}$ first undergoes feature extraction (FE) to produce a set of visual descriptors $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m\}^x$ for a feature x. This set of descriptors then undergoes feature partitioning (FP) to produce p sets of lower-dimensional vectors. For a partition l, the set of lower-dimensional vectors $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m\}_l^x$ is grouped into k clusters, resulting in a set of centroids $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}_l^x$. GFM stores in the codebook a mapping from each centroid $\mathbf{c}_{i,l}^x$ to a unique code word of the form "x:kipl. " In Fig. 2, the clustering process is represented with the k-means algorithm (KM), but any centroid-based clustering algorithm is sufficient for generating the codebook.

3.2 Code word assignment

After generating and storing in the codebook unique words representative of the cluster centroids, GFM then assigns the words to each image in the collection. GFM assigns the code word "x:kipl" to all images whose partition l of the descriptor for feature x lies within the cluster whose centroid is $\mathbf{c}_{i,l}^x$. Specifically, the set of images to which GFM assigns this word is given by $\{I_j : \mathbf{f}_{l,j}^x \in C_{i,l}^x\}$.

While it is useful to know the set of images to which GFM assigns a given code word, we often must consider the set of all code words assigned to a given image. Recall that the code word representing centroid $\mathbf{c}_{i,\,l}^x$ is defined by the feature x and the tuple (i, l). The set of defining tuples for all code words assigned to an image I_j for feature x is given by:

$$W_j^x = \left\{ \left(\underset{i}{\operatorname{argmin}} \|\mathbf{f}_{l,j}^x - \mathbf{c}_{i,l}^x\|, l \right) : 1 \leq l \leq p \right\} \tag{4}$$

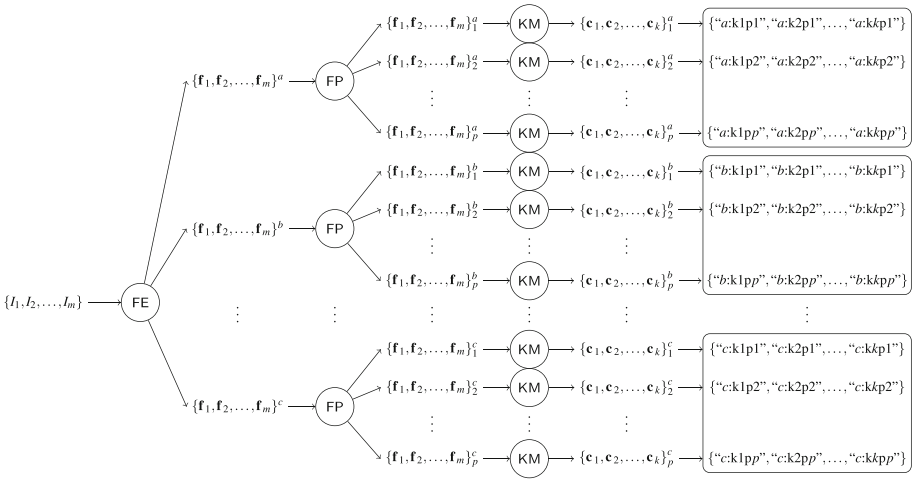


Fig. 2 Codebook generation. Visual descriptors representative of a set of global features are first extracted from a set of images via some feature extractor (FE). Then, the descriptors of each feature are partitioned via some feature partitioner (FP) to form several sets of lower-dimensional vectors. Finally, the sets of lower-dimensional vectors are clustered via the k-means algorithm (KM), and the resulting cluster centroids are mapped to unique textual code words in the codebook

W_q^x identifies the centroids that are nearest to each of the p lower-dimensional vectors representing feature x of image I_j. The set of all code words representing I_j is then given by:

$$D_j^c = \left\{ \text{“}x\text{: kip}l\text{”} : x \in F \wedge (i,l) \in W_j^x \right\} \tag{5}$$

Intuitively, D_j^c can be thought of as a text document containing the code words for the content-based features extracted for image I_j.

The diagram shown in Fig. 3 summarizes the code word assignment process of GFM for an image I_j and a single global feature x. Assume $x \in F, 1 \leq l \leq p$, and $1 \leq i \leq k$. The image first undergoes feature extraction to produce a visual descriptor for feature x. The extracted descriptor then undergoes feature partitioning (FP) to produce p lower-dimensional vectors $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_p\}_j^x$. Feature extraction and partitioning are performed during the codebook generation process. The codebook entries for partition l of feature x are given by $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}_l^x$. For each lower-dimensional vector $\mathbf{f}_{l,j}^x$, the code word assignment process performs a nearest-neighbor search (NN) to select among these entries the cluster centroid to which the vector is nearest. Each selected centroid $\mathbf{c}_{i,l}^x$ is represented in the set W_j^x as a tuple (i, l). The textual representation of the tuples for all the content-based features extracted for an image I_j is then given by D_j^c.

3.3 Multimodal image representation

Because images are seldom self-evident, they are frequently accompanied by text. In general, this text can provide meaning to the visual characteristics of the images and can place the images within a broader context. For example, the images found in biomedical

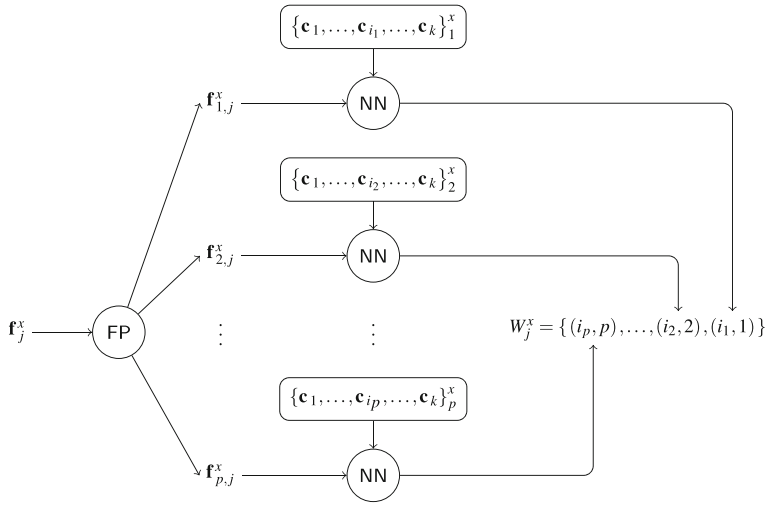


Fig. 3 Code word assignment. The visual descriptor representative of a global feature is extracted from a query image and partitioned via some feature partitioner (FP), forming several lower dimensional vectors. For each lower-dimensional vector, a nearest-neighbor search (NN) is performed to select among the codebook entries the cluster centroid to which the vector is nearest. The query image is then represented by a set of tuples that uniquely define the selected centroids. Note that this diagram only depicts the code word assignment process for one of the features used for codebook generation

articles are usually accompanied by descriptive captions, and their relevance is commonly discussed in passages within the articles' full text that mention the images.

GFM provides efficient access to both the meaning of images as well as their visual characteristics by representing images as a combination of their text-based and content-based features. Assume GFM has completed the codebook generation process for a collection of images and a set of global content-based features. Furthermore, assume GFM has assigned each image in the collection one or more code words corresponding to each of the global features. To prepare the collection of images for indexing, GFM combines the code words assigned to the images with natural words taken from their related text.[8] Let D_j^t represent a document containing descriptive text related to an image I_j. We define a multimodal surrogate document D_j for image I_j as the following:

$$D_j = D_j^t \cup D_j^c \tag{6}$$

After constructing multimodal surrogates for each image in the collection, we index the resulting documents with a traditional text-based information retrieval system.

The image representation used by GFM is best understood with an example. Figure 4 shows a multimodal surrogate document created for an image in the 2010 ImageCLEF medical retrieval track data set. The image, taken from an article by Helbich et al. (1999), is a CT scan depicting morphologic abnormalities in a 9-year-old boy with cystic fibrosis. The image is represented by both its text-based and content-based features in a document that is indexable with a traditional text-based information retrieval system. The image's text-based features include its caption, passages from the full text of the article that mention the image (i.e., Fig. 2 in this example), the title of the article, the article's abstract, and the article's assigned medical subject headings (MeSH® terms).[9] The image's global content-based features are represented as code words derived from five visual descriptors, each of which are described in Sect. 6.1. Note that in this particular example, the descriptors have not been partitioned into lower-dimensional vectors (i.e., $p = 1$). Thus, the image is only assigned one code word for each of the five features.

3.4 Multimodal image retrieval

Having utilized a traditional text-based information retrieval system to index the collection of multimodal surrogate documents produced by GFM, we can efficiently retrieve images from the collection. Assume we would like to retrieve the most relevant images for a multimodal topic, such as the one shown in Fig. 1. In order to formulate a query for this retrieval task, we must first assign code words to the topic's example images and process the textual description of the topic. Then, we can combine the images' code words with natural words taken from the topic description and use this text to search the collection of multimodal image surrogates. We describe GFM's query formulation process below.

Like it does for images in the collection, GFM represents queries as multimodal surrogate documents containing both text-based and content-based features. For a given topic or information request, GFM processes the topic's example images in the same way as it does images in the indexed collection, extracting visual descriptors for an identical set of global content-based features and partitioning these descriptors into the same number of

[8] GFM does not require the use of specific text-based features, and we define image-related text broadly. Refer to Sect. 6.1 for a description of the features we use with our prototype implementation.

[9] MeSH is a controlled vocabulary used by the U.S. National Library of Medicine for indexing articles in the MEDLINE® database.

Image Caption

Figure 2: CT scan at the level of the upper lobes in a 9-year-old boy (group 2 [6-16 years]) demonstrates mild to severe signs of bronchiectasis (curved arrows) and mild to moderate signs of bronchial wall thickening. In addition, CT scan shows mucous plugging (straight arrows) and mosaic perfusion (*).

Image Mentions

Bronchiectasis (80%), peribronchial wall thickening (76%), mosaic perfusion (64%), and mucous plugging (51%) were the most frequently observed morphologic CT abnormalities in the 117 patients (Table 2; Figs 1–5)

Evaluation of the three age groups demonstrated significant trends for progression of disease (P < . 05) as revealed in the overall CT score, in frequency of specific CT signs, and in the severity of the specific abnormalities (Table 3; Figs 1–5).

Global Image Features

cedd:k167p1 cld:k24p1 ehd:k451p1 fcth:k1308p1 sconcept:k60p1

Article Title

Cystic fibrosis: CT assessment of lung involvement in children and adults

Article Abstract

Purpose: To compare a computed tomographic (CT)-based scoring system with nonimaging indexes of pulmonary status in patients with cystic fibrosis.

Materials and Methods: Pulmonary CT findings were assessed in 117 patients with cystic fibrosis, with cases classified according to three groups by age; 0-5 years, 6-16 years, and 17 years and older. Images were examined for specific abnormalities, and the severity and anatomic extent of each sign were used to generate a score. Scores in each category and the global score for each patient were correlated with pulmonary function test results, clinical status, serum immunoglobulin levels, and genotype, all obtained within 2 weeks of CT.

Results: The most frequent individual CT abnormalities were bronchiectasis in 94 (80.3%), peribronchial wall thickening in 89 (76.1%), mosaic perfusion in 71 (63.9%), and mucous plugging in 56 (51.3%) patients. The percentage of patients with specific CT findings and the overall CT scores increased significantly (P < . 05) with progressively increasing age groups. All CT findings and the overall CT scores correlated significantly (P < . 05) with the pulmonary function test results, serum immunoglobulin levels, and clinical scores. No relationship was observed between genotype and CT scores.

Conclusion: Scoring of CT studies in patients with cystic fibrosis seems to offer a reliable way to monitor disease status and progression and may provide a reasonable tool to assess treatment interventions.

Article MeSH Terms

Adolescent; Adult; Age Factors; Child; Child, Preschool; Cystic Fibrosis/radiography*; Female; Humans; Infant; Male; Prospective Studies; Tomography, X-Ray Computed*

Fig. 4 Multimodal image representation. The image from Fig. 2 of the article "Cystic fibrosis: CT assessment of lung involvement in children and adults" by Helbich et al. (1999) is shown represented by a combination of text-based and content-based features

lower-dimensional vectors. GFM assigns words to a query image I_q following the code word assignment process, which produces a set of code words D_q^c. GFM combines D_q^c with a set of natural words D_q^t taken from the topic's textual description[10] in order to form a multimodal query D_q. Finally, we submit D_q as a query to the text-based information retrieval system we used to index the collection of multimodal surrogates and retrieve a set of images.

3.5 Relation with semantic image annotation

Much recent work in the CBIR community has dealt with bridging the so-called "semantic gap" between an image's content and its meaning. The idea is that by automatically labeling an image or the interesting regions of an image with semantically meaningful

[10] The exact method by which we process the textual descriptions of topics is not important for understanding GFM. Refer to Sect. 6.2 for a discussion of the text processing we perform for our prototype implementation.

concepts, such concepts could then be leveraged in order to retrieve conceptually similar images. Our assumption in this work has been that semantic descriptions of the images in our collections are already accessible: the images found in biomedical articles are surrounded by meaningful text (e.g., their captions), and we are using meaningful text (e.g., topic descriptions) as the primary means of retrieving them. Thus, the semantic gap associated with our collection is narrow if it exists at all, and bridging it is not a problem that GFM attempts to solve. Although the code words GFM assigns to images can be thought of as annotations, they convey no obvious meaning beyond cluster membership. Even so, our experiments have shown that, for our data sets, we can improve the retrieval of relevant images by incorporating these image code words into a text-based retrieval process. However, it is beneficial for us to briefly survey some representative work related to semantic annotation.

Many approaches to image annotation attempt to create joint probabilistic models of text-based and content-based features. Typical of these approaches, image-related text is represented as a bag of words, and image content is represented as "blobs," which are quantized content-based feature vectors extracted from important image regions. Conceptually, blobs are similar to GFM's code words, but whereas GFM may represent the global content of a single image with several code words, a single blob represents the local content of one region. The goal of an annotation model is then to learn joint word-blob probabilities from a collection of images and their associated text. Perhaps inspired by techniques from natural language processing, Duygulu et al. (2006) formulate the modeling problem as an instance of machine translation, and Lavrenko et al. (2003) apply the language modeling framework of information retrieval to learn the semantics of images. Barnard et al. (2003) investigate various correspondence models as well as a multimodal extension of Latent Dirichlet allocation (LDA). Blei and Jordan (2003) also propose the use of LDA for modeling associations between words and images. Finally, though not directly related to annotation, Rasiwasia et al. (2010) model the correlations between images' content-based features and their related text in support of cross-modal retrieval. The authors demonstrate that their cross-modal model can outperform systems when evaluated on unimodal retrieval tasks.

Instead of modeling the associations between text-based and content-based image features, semantic annotation can also be achieved using supervised machine learning techniques. Datta et al. (2007) describe a structure-composition model for categorizing image regions. The authors annotate images with tags corresponding to recognized regions and use the annotations for retrieving semantically similar images. They use a bag of words distance measure based on WordNet (Miller 1995) for computing semantic similarity. Li and Wang (2008) present ALIPR (automatic linguistic indexing of pictures—real time), a real time image annotator that uses hidden Markov models to capture the spatial dependencies of content-based features associated with a given set of semantic categories. A related approach is described by Chang et al. (2003), who use Bayes point machines (Herbrich et al. 2001) to assign "soft" annotations to images based on category confidence measures estimated from a training set of labeled images.

Within the biomedical domain, region classification has been a popular approach for improving image retrieval. Lacoste et al. (2007) index images using a combination of UMLS concepts extracted from image-related text and VisMed (Lim and Chevallet 2005) terms derived from image content. VisMed terms are semantic labels generated by classifying the appearance of image regions. The authors demonstrate that a multimodal fusion approach that utilizes VisMed terms is capable of outperforming systems evaluated on the 2005 ImageCLEF medical retrieval track data set. However, unlike GFM's unsupervised

method of generating code words, semantic annotation using VisMed terms is an instance of supervised learning and requires a sufficient set of training from which to derive the terms. Additionally, Simpson et al. (2012b) discuss the creation of a "visual ontology" of biomedical imaging entities using supervised learning. The authors utilize natural language and image processing techniques to automatically create a training set of annotated image regions by pairing the visible arrows in images with the caption text describing their pointed-to regions. They then use this data set to train a classifier to label regions in images having no associated text. This approach has yet to be evaluated for its use in improving medical image retrieval.

Finally, Wang et al. (2008) discuss a search-based approach to image annotation. To annotate an image, the method first performs a content-based search to retrieve visually similar images, and it then uses text related to the retrieved images to form a list of candidate annotations for the original.

4 Images as words

When represented as code words, images become subject to the underlying models used by traditional text-based information retrieval systems. While this may not seem immediately desirable, the well-understood concepts of text-based retrieval are easily adapted for use with image code words, and they prove to be beneficial for improving upon the retrieval performance and efficiency of existing content-based and multimodal image retrieval systems. Below, we discuss how common text-based retrieval techniques—namely, query expansion and relevance ranking—operate when we represent images as words.

4.1 Code word expansion

Text-based retrieval systems often perform query expansion in an effort to improve retrieval performance. A commonly used technique involves expanding a query to include the synonyms and morphological variants of existing terms. Text-based query expansion methods, however, are not directly applicable to image code words because, as they are not natural words, they do not have conventional synonyms or variants. Instead, we define the relatedness of two code words as the distance between their representative cluster centroids, and we expand a query of code words to include those corresponding to nearby centroids.

A problem with traditional centroid-based clustering algorithms is the requirement that each element belongs to exactly one cluster. This restriction is unfortunate for GFM because it implies that images that may be similar in appearance can be assigned different code words. Consider a Voronoi diagram representing the clustering of the visual descriptors extracted from a collection of images for a particular global feature. Descriptors lying close to and on either side of the boundary between two adjacent cells are more similar to each other than either one is to its respective cell center. Thus, because GFM assigns different code words to images whose descriptors lie within different cells, it may not retrieve the most visually similar set of images to a given query image if the query image's descriptor lies far away from a cell's center.

The goal of code word expansion is to minimize the negative impact rigid cluster membership has on retrieval performance. Similar to fuzzy cluster analysis (Bezdek et al. 1999), code word expansion allows GFM to assign more than one code word to a query image for a given feature and partition. Recall that the set W_q^x contains all tuples (i, l) that

define the code words GFM assigns to a query image I_q for feature x. For each partition l of feature x, W_q^x identifies the single centroid $\mathbf{c}_{i,l}^x$ to which $\mathbf{f}_{l,q}^x$ is closest. In order to expand a code word query, we parameterize W_q^x with a code word expansion factor ε. $W_q(\varepsilon)^x$ identifies the ε nearest cluster centroids to each $\mathbf{f}_{l,q}^x$ and is defined by:

$$W_j^x(\varepsilon) = \left\{(e_i, l): 1 \leq i \leq \varepsilon \wedge 1 \leq l \leq p \wedge e_i \in E_{l,j}^x\right\} \tag{7}$$

$E_{l,j}^x$ is a set of identifiers corresponding to the cluster centroids after they have be sorted in order of increasing distance from $\mathbf{f}_{l,q}^x$:

$$E_{l,j}^x = \{e_1, e_2, \ldots, e_k\} \quad \text{ordered by} \quad \mathbf{f}_{l,j}^x - \mathbf{c}_{e_n,l}^x \leq \mathbf{f}_{l,j}^x - \mathbf{c}_{e_{n+1}^x,l} \tag{8}$$

Thus, the set $W_q^x(\varepsilon)$ contains all tuples (i, l) representative of the ε nearest centroids to each $\mathbf{f}_{l,j}^x$ for $1 \leq l \leq p$. Similarly, $D_q^c(\varepsilon)$ contains the actual expanded set of code words for a query image I_q. The code words assigned to an image are subject to the term weighting strategy of the underlying text-based retrieval model. However, because many retrieval systems allow terms to be weighted manually, a weighting strategy that allocates less weight to the expanded code words could potentially be realized that simulates the probabilistic cluster membership obtainable by fuzzy clustering techniques.

4.2 Image similarity

The most apparent consequence of using a traditional text-based retrieval system to index images is that retrieved images are ranked according to some text-based similarity measure. Whereas existing content-based and multimodal image retrieval systems commonly rank images by the Euclidean distance between their extracted visual descriptors, this ranking is not directly possible when we instead represent images as words. Modern text-based retrieval systems implement a variety of set-theoretic, algebraic, and probabilistic retrieval models. Though not always the best-performing approaches, many text-based systems, such as Apache Lucene,[11] implement a combination of the Boolean and vector space models, especially variants of these models that utilize *tf–idf* term weighting. Below, we briefly discuss the treatment of image code words within these well-known models.

4.2.1 Boolean model

The Boolean model (Lancaster and Fayen 1973) was one of the first and most widely adopted information retrieval strategies, and many modern retrieval systems provide a mechanism for constructing queries that utilize standard Boolean operators. If we assume query images to be the disjunction of their code words, then the set of images retrieved by the model for a query image I_q is given by $\{I_j: D_j^c \cap D_q^c(\varepsilon) \neq \emptyset\}$. Thus, the model retrieves all images from the collection that are represented by a code word contained in the set of expanded code words GFM assigns to the query image. Alternatively, if we assume query images to be the conjunction of their code words, the set of images retrieved by the model is given by $\{I_j: D_j^c \subseteq D_q^c(\varepsilon)\}$.

The use of Boolean operators is especially useful for creating queries for topics containing more than one example image, such as the one shown in Fig. 1. For such topics, we might like to retrieve all images that are visually similar to at least one of the example

[11] http://lucene.apache.org/.

images, or we might, instead, prefer to retrieve images that are similar to all of the examples. We can construct result sets for complex image queries by first retrieving a set of images for each example image according to either the disjunctive or conjunctive query formulation strategy and then applying the Boolean operators to the retrieved sets of images.

4.2.2 Vector space model

The vector space model (Salton et al. 1975) is a well-known algebraic model of information retrieval where documents and queries are represented as term vectors. We can construct code word vectors for images following the classical formulation. D_j^c, the code words GFM assigns to an image I_j, is represented as a set of term vectors $\{\mathbf{v}_j^x : x \in F\}$, where each $\mathbf{v}_j^x$ corresponds to the codebook entries for feature x. Furthermore, each $\mathbf{v}_j^x$ is partitioned into p lower-dimensional term vectors:

$$\mathbf{v}_j^x = \begin{bmatrix} \mathbf{v}_{1,j}^x & \mathbf{v}_{2,j}^x & \cdots & \mathbf{v}_{p,j}^x \end{bmatrix} \tag{9}$$

Each $\mathbf{v}_{l,j}^x$ corresponds only to those codebook entries of feature x that are defined for partition l. The attributes of these lower-dimensional vectors are weights corresponding to the codebook entries they represent:

$$\mathbf{v}_{l,j}^x = \begin{bmatrix} w_{1,l,j}^x & w_{2,l,j}^x & \cdots & w_{k,l,j}^x \end{bmatrix} \tag{10}$$

Vector space retrieval systems commonly implement the *tf–idf* term weighting strategy. Because GFM only assigns images one word per feature and partition combination, the term frequency of each codebook entry contained in document D_j^c is equal to one. Thus, code words are weighted by their inverse document frequency, which is defined by:

$$w_{i,l,j}^x = \begin{cases} \log \frac{m}{C_{i,l}^x} & \text{if}(i,l) \in W_j^x(\varepsilon) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

The inverse document frequency of a given code word is related to the number of images whose visual descriptors are members of the cluster the code word represents. Because code words are uniquely defined by feature x and tuple (i, l), this number is equal to $|C_{i,l}^x|$. The *tf–idf* weighting strategy implies that code words representing clusters having few members are weighted more heavily than those representing clusters with many members. Thus, the retrieval system favors images that are more unique within the collection.

Code word expansion can either be performed when indexing images or when mapping query images to their associated code words; it is not necessary to perform code word expansion during both the indexing and retrieval steps. By convention, we perform code word expansion during the retrieval process. Thus, we let $\varepsilon = 1$ for the m images in the collection and $\varepsilon \geq 1$ for query images.

Having defined the term vectors used by the vector space model and the weights assigned to each code word, we can compute the similarity between two images. The similarity between a query image I_q and an image I_j from the collection is given by the average cosine similarity between the code word vectors representing I_q and I_j for all features. For a given query image, the retrieval system computes its similarity with each image in the collection and then ranks the collection of images accordingly.

5 Data and evaluation

The medical retrieval track of ImageCLEF has been an important catalyst for advancing the science of image retrieval within the biomedical domain (Hersh et al. 2009). For the *ad hoc* retrieval task, participants are provided with a set of topics, and they are challenged with retrieving for each topic the most relevant images from a collection of biomedical articles. As was shown in Fig. 1, each topic is multimodal, consisting of a textual description of some information need as well as one or more example images. Although the best-performing systems at ImageCLEF evaluations have historically relied upon text-based retrieval methods, recent systems have shown encouraging progress towards combing these methods with content-based approaches, especially since the introduction of an image modality classification task (Müller et al. 2010b). For the classification task, the goal is to classify images according to medical imaging modalities such as "Computerized Tomography" or "X-ray."

We chose the 2010 and 2012 ImageCLEF medical retrieval track data sets for the evaluation of GFM. The 2010 collection contains 77,479 images taken from a subset of the articles appearing in the *Radiology* and *Radiographics* journals, and the 2012 collection contains 306,539 images taken from a portion of the articles in the open access subset of PubMed Central. Each image is associated with its caption and the title, identifier, and URL of the article in which it appears. The 2010 collection identifies each article by its PubMed identifier (PMID) whereas the 2012 collection uses its PubMed Central identifier (PMCID). There are sixteen multimodal *ad hoc* topics in the 2010 collection and twenty-two topics in the 2012 data set. The organizers of the ImageCLEF evaluations categorize the topics in roughly equal proportions as being "Visual," "Mixed," or "Semantic" according to their expected benefit from content-based or text-based retrieval techniques.

Following the TREC evaluation methodology (Voorhees and Harman 2005), the highest ranked images retrieved by each ImageCLEF participant for a given topic were pooled and manually judged as either being relevant to the topic or not relevant. Using these judgements, we report a system's performance for a topic as binary preference (bpref), judged mean average precision (MAP′), and judged precision-at-ten (P′@10) over the one thousand highest-ranked images. Although they are highly correlated, because bpref and MAP′ do not always agree, we assume these metrics to be complimentary and use them both as an indicator of average system performance. However, Sakai (2007) has determined that average precision, when computed on only the images having relevance judgements, is at least as robust to incomplete judgements as bpref but more discriminative. To measure the statistical significance between the average performance of two or more systems, we applied Fisher's two-sided, paired randomization test (Smucker et al. 2007), which is a recommended statistical test for evaluating information retrieval systems.

We evaluate the efficiency of each retrieval system having measured the time in milliseconds needed to produce a ranked list of results for each topic. To conduct the experiments, we organized the retrieval systems in a client/server architecture networked via a Gigabit Ethernet connection. The GNU/Linux server had 2 Intel Xeon 5160 processors (2 cores, 3 GHz, 4 MB L2 cache) and 10 GB of memory. The Microsoft Windows XP client had a single Intel Xeon W3520 processor (4 cores, 2.66 GHz, 8 MB L3 cache) and 3 GB of memory.

6 Implementation

Owing to its practicality, we have implemented GFM within two text-based information retrieval frameworks. The first, Apache Lucene, is a general purpose vector space system widely recognized for its ease of use and reasonable performance. The second, Essie (Ide et al. 2007), is a biomedical retrieval system developed by the U.S. National Library of Medicine. Essie scores documents using a probabilistic retrieval model, and automatically expands query terms along the synonymy relationships in the UMLS. The retrieval models of both Lucene and Essie support queries that utilize standard Boolean operators. We evaluate our Lucene implementation of GFM on the 2010 ImageCLEF collection and our Essie implementation on the 2012 data set. Because the medical retrieval track of ImageCLEF is a domain-specific retrieval task, we have also implemented UMLS synonymy expansion for Lucene. Both of the above retrieval systems allow documents to be composed of multiple fields and provide low latency access to documents using inverted file indices. Below we describe how we represent images as multi-field documents indexable by these two systems.

6.1 Image representation

We represent the images in the ImageCLEF collections using a combination of text-based and content-based features. Our text-based features include an image's caption and mentions as well as the title, abstract, and MeSH terms of the article in which it is contained. The ImageCLEF data sets provide image captions and article titles. To obtain each article's abstract and MeSH terms, we utilize its associated PMID/PMCID with the Entrez programming utilities (NCBI 2010) to retrieve MEDLINE citations containing the required elements. To obtain image mentions, we extract passages that refer to the images from the full text articles retrieved using the provided URLs. We identify image mentions using regular expression patterns that match image labels. For example, if an image's caption identifies it as "Fig. 2a," we extract sentences that contain variants of this label.

Our content-based features primarily describe color and texture information, and they include the descriptors listed in Table 1. We used the "Core" features with our Lucene implementation and both the "Core" and "Additional" features with our Essie implementation. Although we recognize that no single combination of features is adequate for describing the content of all images, a detailed analysis of the strengths and weaknesses of these particular sets is beyond the scope of our current evaluation. However, note that the dimensionality of many of the features is prohibitively large for maintaining them in spatial data structures. To efficiently extract these features, we utilized the MapReduce framework on an eight-node Apache Hadoop[12] cluster. For convenience, we extracted the features for both collection and topic images offline, prior to performing our indexing and retrieval experiments. However, given the extracted features for a topic image, our GFM implementations compute the associated code words online, and this computation time is accounted for in our results.

Once the content-based features have been extracted, our GFM implementations cluster them using the k-means++ algorithm (Arthur and Vassilvitskii 2007), which we chose for its simplicity, efficiency, and accuracy. Additionally, because k-means++ uses Euclidean distance as its clustering metric, GFM's retrieval results can be compared with those obtained by other retrieval systems using Euclidean distance without the need for

[12] http://hadoop.apache.org/.

Table 1 Content-based features used for our global feature mapping implementations

Type	Name	Dims.
Core	Color layout descriptor* (Chang et al. 2001)	16
	Semantic concept (Rahman et al. 2009)	30
	Edge histogram descriptor* (Chang et al. 2001)	80
	Color and edge directivity descriptor* (Chatzichristofis and Boutalis 2008a)	144
	Fuzzy color and texture histogram* (Chatzichristofis and Boutalis 2008b)	192
Additional	Color moment	3
	Primitive length	5
	Shape moment	5
	Tamura moment* (Tamura et al. 1978)	18
	Gray-level co-occurrence matrix moment (Srinivasan and Shobha 2008)	20
	Autocorrelation	25
	Edge frequency	25
	Gabor moment*	60
	Scale-invariant feature transformation* (Lowe 1999)	256
	Local binary pattern (Mäenpää 2003)	512
	Local color histogram	1,024

* Feature computed using the Lucene Image Retrieval Library (Lux and Chatzichristofis 2008)

considering potential differences in similarity metrics. However, GFM's indexing and retrieval method is compatible with other centroid-based clustering techniques, and we have experimented with some of these algorithms, such as hierarchical k-means. In addition to k-means and its variants, many other clustering techniques have been proposed for image retrieval tasks. Datta et al. (2008) survey the strengths and weaknesses of several popular algorithms.

We experimentally determined reasonable values for the number of partitions and clusters for each feature based on preliminary observations. Due to the computational complexity associated with clustering the higher-dimensional features vectors we used for the 2012 ImageCLEF collection, we let the maximum number of feature partitions equal six ($p = 6$) for this data set, whereas we let the number of partitions equal two ($p = 2$) for the 2010 ImageCLEF collection. To ensure the scalability of our method, we let the number of clusters for each partition be logarithmic in the total number of images. Thus, the number of clusters for a partition p is given by:

$$k = \left\lceil \frac{d}{p} \times \log m \right\rceil \tag{12}$$

where d is the dimensionality of a content-based feature shown in Table 1, and m is the total number of images in the collection.

We include the images' text-based features and the code words corresponding to their content-based features as unique fields in multi-field text documents. In this way, each image in the ImageCLEF collections is represented as a surrogate document indexable by a typical text-based retrieval system. Figure 4 shows an example multimodal image document.

6.2 Image retrieval

Because GFM is a multimodal image retrieval method, our Lucene and Essie implementations support three distinct retrieval paradigms. In addition to multimodal retrieval, these search strategies also include text-based and content-based approaches. We present implementation details related to each of these uses of GFM in the remainder of this section.

Because one of the primary objectives of our current work is to demonstrate that visual information can be used to improve upon a competitive text-based approach, it is important that our textual baseline be a state-of-the art retrieval method. For automatically generating queries, our textual baseline first organizes the textual description of a topic into the well-formed clinical question (i.e., PICO[13]) framework (Richardson et al. 1995) following the method described by Demner-Fushman and Lin (2007). Accordingly, it extracts from the topic UMLS concepts related to problems, interventions, age, anatomy, drugs, and image modality. In addition to automatically expanding these extracted concepts using the UMLS synonymy, it also expands identified modalities using a thesaurus manually constructed by Demner-Fushman et al. (2008) based on the RadLex (Langlotz 2006) ontology.[14] Our textual baseline then constructs a disjunctive query consisting of all the expanded terms. To ensure the early precision of our retrieval results, the textual baseline weights term occurrences in image captions and article titles more than occurrences in other text-based fields. It also requires that any modality terms identified in the query occur in a retrieved image's caption or mentions. Finally, in order to improve recall, our textual baseline pads the initially retrieved results with images retrieved using the verbatim topic description as query.

We refer to the use of our GFM implementations for content-based image retrieval as content-based GFM. Content-based GFM is an approximation of a typical CBIR system that represents images using the content-based features shown in Table 1 and compares them using Euclidean distance. In contrast with our textual baseline, content-based GFM only searches the fields of our indices that correspond to content-based features and only processes the example images of a multimodal topic to construct a query. For automatically generating queries, content-based GFM first concurrently extracts the content-based features for all the example images in a topic. It then maps the extracted features to code words using a default code word expansion factor of one ($\varepsilon = 1$). Finally, content-based GFM constructs a disjunctive query consisting of the mapped code words for all example images, enabling it to retrieve images visually similar to any of the examples.

The last search paradigm our GFM implementations support is multimodal image retrieval, and we refer to this use as multimodal GFM. Multimodal GFM is the combination of our textual baseline with content-based GFM. It searches all the fields of our indices, and it processes both a topic's textual description as well as its example images to construct a multimodal query. Based on our preliminary experiments, multimodal GFM weights the images' text-based features significantly more than their content-based features and uses a default code word expansion factor of two ($\varepsilon = 2$).

[13] PICO is a mnemonic for structuring clinical questions in evidence-based practice and represents Patient/Population/Problem, Intervention, Comparison, and Outcome.

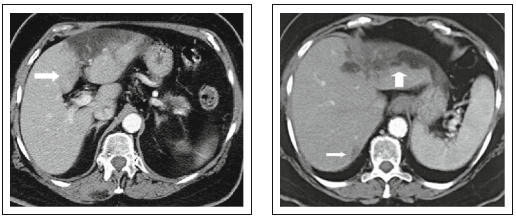
[14] RadLex is a unified ontology of radiology terms, many of which are not included in the vocabularies contained in the current release of the UMLS.

7 Results

In this section we present experimental results for the evaluation of our GFM implementations. Because GFM seeks to improve retrieval precision by providing a practical means of efficiently incorporating visual information into a predominately text-based image retrieval strategy, we discuss here results for both retrieval time and performance. In our evaluation, we demonstrate that, although GFM makes use of content-based image features, it requires a retrieval time that is roughly equal to that of a traditional text-based retrieval system, providing evidence that GFM is capable of indexing large-scale image collections. We also show that our multimodal and content-based GFM implementations achieve statistically significant improvements in retrieval precision on both the 2010 and 2012 ImageCLEF collections.

Before presenting our complete set of results, we show in Fig. 5 example retrieval results obtained with our Lucene GFM implementation for a topic taken from the 2010 ImageCLEF medical retrieval track data set. Depicted are (1) the textual description of the topic and its example images, (2) relevance scores and retrieval times obtained with our three retrieval approaches, and (3) the top five ranked images retrieved using each method. The compared retrieval methods include the textual baseline (TB) as well as content-based and multimodal GFM (CB-GFM and M-GFM, respectively). The results in Fig. 5 show that for this topic and among our methods, multimodal GFM achieves the best performance by successfully combining and improving upon our text-based and content-based approaches. Since GFM utilizes traditional inverted file indices for indexing and retrieval, the search latency achieved by content-based and multimodal GFM is comparable to that of the textual baseline. Note that we first introduced this particular multimodal topic when describing Fig. 1. We explore these and additional results in more detail in Sects. 7.1 and 7.2.

Example Images



Textual Description

Topic 6: CT images containing a fatty liver

Retrieval Results

Method	bpref	MAP′	P′@10	Time (ms)
TB	0.4116	0.3176	0.6000	93
CB-GFM	0.0229	0.0024	0.0000	31
M-GFM	0.4229	0.3905	0.4000	156

Retrieved Images

Rank	TB	CB-GFM	M-GFM
1.			
2.			
3.			
4.			
5.			

Fig. 5 Example retrieval results for topic six of the 2010 ImageCLEF medical retrieval track data set. Relevance scores are given for the 1,000 highest ranked images with metrics including binary preference (bpref), judged mean average precision (MAP′), and judged precision-at-ten (P′@10). Results are shown for content-based global feature mapping (CB-GFM), multimodal global feature mapping (M-GFM), and the textual baseline. All reported times are the lowest of ten retrieval runs and are given in milliseconds

7.1 Retrieval time

Table 2 shows the retrieval time required by our Lucene GFM implementation for each topic taken from the 2010 ImageCLEF medical retrieval track data set. Retrieval times are reported in milliseconds and reflect the lowest search latency obtained from ten retrieval runs. We report the lowest retrieval time for each method—as opposed to the average—to avoid including the cost of any other processes running on our evaluation system that may have preempted the retrieval process. In addition to content-based and multimodal GFM, we provide for comparison the retrieval times for the textual baseline as well as the brute-force CBIR approach (BF-CBIR). For brute-force CBIR, all visual descriptors were loaded to memory prior to timing; therefore, the search latencies reported for this method only reflect the time needed for first computing Euclidean distances between the query and collection descriptors and then sorting these distances. For all other methods, the reported latencies include the time required for generating code words and parsing the queries in addition to retrieving the images. The time needed for extracting content-based features from the example query images is not included our reported search latencies.

The results depicted in Table 2 show that content-based and multimodal GFM require a retrieval time that is roughly comparable to that of a traditional text-based information retrieval system. In addition, it shows that for the 2010 ImageCLEF collection, brute-force CBIR takes approximately two orders of magnitude longer to retrieve the highest-ranked images than the comparable content-based GFM. The fact that brute-force CBIR requires so much more time compared to the others is not surprising—this is a naïve retrieval

Table 2 Retrieval time

Topic	BF-CBIR	CB-GFM	TB	M-GFM
1	2,155	31	188	234
2	2,171	0	15	31
3	2,172	31	0	62
4	2,140	46	62	125
5	1,093	0	78	93
6	2,140	31	93	156
7	2,140	46	62	140
8	2,124	15	0	47
9	2,140	31	62	109
10	2,155	31	63	109
11	2,156	15	31	78
12	2,156	47	0	78
13	2,171	47	0	78
14	2,155	31	15	78
15	2,156	46	15	78
16	2,156	15	0	46
All	2,086	29	43	96

Topics are taken from the ImageCLEF 2010 medical retrieval track data set. Results for brute-force CBIR are given as a comparison. For content-based GFM, $\varepsilon = 1$ and $p = 2$, whereas for multimodal GFM, $\varepsilon = 2$ and $p = 2$. All reported times are the lowest of ten retrieval runs and are given in milliseconds

BF-CBIR Brute-force content-based image retrieval, *CB-GFM* content-based global feature mapping, *TB* textual baseline, *M-GFM* multimodal global feature mapping

strategy. However, brute-force CBIR remains a remarkably common approach used for searching small-to-medium sized collections. Table 2 demonstrates that the efficiency of inverted file indices is easily obtainable for the content-based and multimodal image retrieval paradigms, which is especially significant for managing large image collections.

The retrieval times presented in Table 2 vary slightly for each topic. For the GFM-based results, the variation in retrieval time generally reflects differences in the length of the queries and the number of images the queries retrieve. Longer queries require additional time to parse, and queries that retrieve many images require more time to score the results. The length of a query depends on the length of the topic's textual description as well as the number of example images it has. The number of images retrieved for each topic depends on the queries. For image-based queries, this is related to the number of feature vectors in each cluster. For example, a query containing a cluster word representing many images will result in more images being scored because the cluster word is more common within the collection. For the brute-force approach, the number of scored images and the length of the feature vectors remains constant across the topics.

7.2 Retrieval performance

Having demonstrated that content-based and multimodal GFM achieve response times comparable to what is obtained by our textual baseline, we now show that GFM is capable of improving upon the average retrieval precision of existing methods. In doing so, we also demonstrate the effectiveness of code word expansion and show that intracluster image ranking—the relative ranking of images mapped to identical sets of code words—is improved when indexing a sufficient number of features or feature partitions.

7.2.1 Multimodal retrieval

Tables 3 and 4 show the retrieval results obtained by our multimodal Lucene and Essie GFM implementations for each topic taken from the 2010 and 2012 ImageCLEF medical retrieval track data sets. For comparison, we also include in the tables the results obtained by our textual baseline and the multimodal systems that achieved the highest average bpref at the ImageCLEF evaluations. Taken together, these results show that (1) our textual baseline is statistically indistinguishable from the best performing systems evaluated at ImageCLEF and that (2) the performance of multimodal GFM is significantly better than that of our textual baseline. The observed improvement in retrieval precision is especially encouraging because it is consistent across two different data sets and GFM implementations and, as we saw in Table 2, requires a negligible increase in retrieval latency over our textual baseline.

For the 2010 ImageCLEF results shown in Table 3, we see that our Lucene implementation of multimodal GFM achieved a statistically significant increase in both MAP′ (10.03 %, $p = 0.02$) and bpref (7.46 %, $p = 0.02$) compared to our textual baseline. Although the average P′@10 obtained by multimodal GFM is also greater than that of our textual baseline, this improvement did not reach the level of statistical significance ($p < 0.05$). These results show that incorporating a limited amount of visual information into the retrieval process can provide a slight but consistent performance improvement over text-based retrieval.

The potential for multimodal GFM to create synergistic combinations of text-based and content-based features is perhaps best demonstrated by topic 9. Topic 9 is about MR images of papilledema (swelling of the optic disc) and both of the provided MR images are

Table 3 Multimodal retrieval results for ImageCLEF 2010

Topic	Textual baseline			Multimodal GFM			Best ImageCLEF 2010		
	bpref	MAP′	P′@10	bpref	MAP′	P′@10	bpref	MAP′	P′@10
1	0.4560	0.4237	0.2000	**0.4731**	**0.4489**	0.3000	0.4112	0.3658	**0.6000**
2	**1.0000**	**1.0000**	**0.1000**	**1.0000**	**1.0000**	**0.1000**	0.0000	0.0000	0.0000
3	0.1774	0.1930	0.1000	0.1943	0.1969	0.2000	**0.5848**	**0.5226**	**0.4000**
4	0.1285	0.1301	0.3000	0.1380	0.1405	0.3000	**0.1701**	**0.1608**	**0.4000**
5	0.1855	0.0994	**0.5000**	**0.1924**	**0.2201**	**0.5000**	0.0918	0.1334	0.3000
6	0.4116	0.3176	0.6000	**0.4229**	**0.3905**	0.4000	0.2398	0.2272	**0.7000**
7	**0.5000**	**0.5026**	**0.2000**	**0.5000**	**0.5026**	**0.2000**	0.2500	0.2643	0.1000
8	0.0000	0.0055	0.0000	0.0000	0.0048	0.0000	**1.0000**	**1.0000**	**0.1000**
9	0.3819	0.4049	0.5000	0.7778	0.7674	0.8000	**0.8889**	**0.8846**	**0.9000**
10	**0.6599**	**0.6587**	**0.8000**	**0.6599**	**0.6587**	**0.8000**	0.6485	0.6367	0.7000
11	0.1744	**0.1652**	0.6000	0.1744	**0.1652**	0.6000	**0.1872**	0.1293	**0.7000**
12	**0.3569**	**0.3860**	**0.7000**	**0.3569**	**0.3860**	**0.7000**	0.2081	0.2241	0.6000
13	**0.1494**	**0.1140**	**0.5000**	**0.1494**	**0.1140**	**0.5000**	0.0547	0.0167	0.2000
14	**0.7857**	**0.8010**	**1.0000**	**0.7857**	**0.8010**	**1.0000**	0.6110	0.5875	0.7000
15	0.7678	0.7320	**1.0000**	**0.7679**	**0.7321**	**1.0000**	0.7995	0.5623	0.6000
16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0004**	0.0000
All	0.3834	0.3708	0.4437	**0.4120***	**0.4080***	**0.4625**	0.3841	0.3572	0.4375

Bold entries signify the best scores obtained for each topic and metric

Topics are taken from the ImageCLEF 2010 medical retrieval track data set, and relevance scores are given for the 1,000 highest ranked images per topic. Results for the single best performing multimodal system at ImageCLEF 2010 are given as a comparison. Results for the text-based component of the multimodal global feature mapping approach are given as a baseline. For multimodal GFM, $\varepsilon = 2$ and $p = 2$

GFM Global feature mapping, *bpref* binary preference, *MAP′* judged mean average precision, *P′@10* judged precision-at-ten

* Statistically significant ($p < 0.05$) with respect to Textual Baseline

of the head. For this topic, our textual baseline achieved a bpref of 0.3819, which is consistent with its average bpref over all sixteen topics (0.3834). However, multimodal GFM dramatically improved upon this result by obtaining a bpref of 0.7778. We will see in Sect. 7.2.2 that when configured for performing content-based retrieval, GFM is unable to retrieve a single relevant image for topic 9, demonstrating that it is through the combination of features that multimodal GFM improves performance. The inability of content-based GFM to retrieve relevant images for this topic may be due to the dissimilarity of the two example images: one image depicts a sagittal view of the head whereas the other shows a coronal view. Although both topic images are used for constructing a query, the lack of a singular way in which to visually describe the concept likely contributes to content-based GFM's retrieval of many irrelevant images. For topics such as this one, the use of semantic information provided by text-based features significantly improves the performance of GFM.

The 2012 ImageCLEF results depicted in Table 4 show that, like our Lucene implementation, our Essie implementation of multimodal GFM also achieved a statistically significant increase in MAP′ (2.61 %, $p = 0.02$) compared to our textual baseline. However, we did not find its improvement in bpref or P′@10 to be statistically significant. The

Table 4 Multimodal retrieval results for ImageCLEF 2012

Topic	Textual baseline			Multimodal GFM			Best ImageCLEF 2012		
	bpref	MAP′	P′@10	bpref	MAP′	P′@10	bpref	MAP′	P′@10
1	0.3469	0.3282	0.4000	**0.3696**	**0.3437**	**0.6000**	0.3469	0.3282	0.4000
2	**0.3719**	**0.3578**	**0.5000**	**0.3719**	**0.3578**	**0.5000**	**0.3719**	**0.3578**	**0.5000**
3	0.1390	0.0890	**0.2000**	**0.1575**	**0.1003**	**0.2000**	0.1390	0.0890	**0.2000**
4	**0.4401**	**0.5234**	**0.8000**	**0.4401**	**0.5234**	**0.8000**	**0.4401**	**0.5234**	**0.8000**
5	**0.0797**	**0.0685**	**0.2000**	**0.0797**	**0.0685**	**0.2000**	**0.0797**	**0.0685**	**0.2000**
6	**0.3787**	0.3529	**0.6000**	0.3491	**0.3654**	0.4000	**0.3787**	0.3529	**0.6000**
7	**0.1653**	**0.2791**	**0.2000**	**0.1653**	**0.2791**	**0.2000**	**0.1653**	**0.2791**	**0.2000**
8	0.0000	0.0276	0.0000	0.0000	**0.0688**	**0.1000**	0.0000	0.0276	0.0000
9	**0.2500**	**0.4500**	**0.2000**	**0.2500**	**0.4500**	**0.2000**	**0.2500**	**0.4500**	**0.2000**
10	**0.1626**	0.1425	**0.3000**	**0.1626**	**0.1720**	**0.3000**	**0.1626**	0.1425	**0.3000**
11	0.6491	0.6834	**1.0000**	**0.7047**	**0.7178**	**1.0000**	0.6491	0.6834	**1.0000**
12	**0.6626**	**0.6594**	**0.8000**	**0.6626**	**0.6594**	**0.8000**	**0.6626**	**0.6594**	**0.8000**
13	**0.8067**	**0.7808**	**0.9000**	**0.8067**	**0.7808**	**0.9000**	**0.8067**	**0.7808**	**0.9000**
14	0.0114	0.0073	**0.5000**	**0.0260**	**0.0118**	**0.5000**	**0.0260**	**0.0118**	**0.5000**
15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
16	**0.2222**	**0.1667**	**0.1000**	**0.2222**	**0.1667**	**0.1000**	**0.2222**	**0.1667**	**0.1000**
17	0.0000	**0.0400**	0.0000	0.0000	0.0376	0.0000	0.0000	**0.0400**	0.0000
18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
19	**0.1667**	0.1719	**0.1000**	**0.1667**	**0.1780**	**0.1000**	**0.1667**	0.1719	**0.1000**
20	**0.2000**	**0.2407**	**0.1000**	**0.2000**	**0.2407**	**0.1000**	**0.2000**	**0.2407**	**0.1000**
21	**0.4207**	**0.4180**	**1.0000**	**0.4207**	**0.4180**	**1.0000**	**0.4207**	**0.4180**	**1.0000**
22	**0.1053**	**0.1227**	**0.2000**	**0.1053**	**0.1227**	**0.2000**	**0.1053**	**0.1227**	**0.2000**
All	0.2536	0.2686	0.3682	**0.2573**	**0.2756***	**0.3727**	0.2542	0.2688	0.3682

Bold entries signify the best scores obtained for each topic and metric

Topics are taken from the ImageCLEF 2012 medical retrieval track data set, and relevance scores are given for the 1,000 highest ranked images per topic. Results for the single best performing multimodal system at ImageCLEF 2012 are given as a comparison. Results for the text-based component of the multimodal global feature mapping approach are given as a baseline. For multimodal GFM, $\varepsilon = 2$ and $p = 6$

GFM Global feature mapping, *bpref* binary preference, *MAP′* judged mean average precision, *P′@10* judged precision-at-ten

* Statistically significant ($p < 0.05$) with respect to both Textual Baseline and Best ImageCLEF 2012

overall trend for our Essie implementation on the 2012 ImageCLEF collection is similar to that of our Lucene implementation on the 2010 data set, with multimodal GFM providing a slight but consistent improvement in performance for many of the topics. We did observe a decrease in bpref and P′@10 on topic 6, though. Although the average performance of the three systems shown in Table 4 differ somewhat, their similarity is not a coincidence: the best performing system at the 2012 ImageCLEF evaluation was an earlier implementation of multimodal GFM . In addition to reducing the overall weight Essie allocates to the content-based fields of our indices when scoring documents, our current implementation of multimodal GFM also dynamically reduces the weight given to image code words for topics the ImageCLEF organizers categorized as "Semantic" topics.

Table 5 Content-based retrieval results for ImageCLEF 2010

Topic	Brute-force CBIR			Content-based GFM			Best ImageCLEF 2010		
	bpref	MAP′	P′@10	bpref	MAP′	P′@10	bpref	MAP′	P′@10
1	0.0107	0.0015	0.0000	**0.0219**	**0.0026**	**0.1000**	0.0000	0.0003	0.0000
2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0000	**0.0017**	0.0000	**0.0008**	0.0000	0.0000	0.0000	0.0002	0.0000
4	0.0000	0.0163	0.0000	**0.1096**	**0.0730**	**0.2000**	0.0000	0.0099	0.0000
5	**0.0003**	0.0000	0.0000	0.0000	**0.0009**	0.0000	0.0000	0.0000	0.0000
6	0.0105	**0.0031**	0.0000	**0.0229**	0.0024	0.0000	0.0145	0.0019	**0.1000**
7	0.0000	0.0421	0.0000	0.0000	0.0000	0.0000	**0.1875**	**0.1258**	**0.1000**
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
11	0.0182	**0.0048**	**0.1000**	**0.0343**	0.0042	0.0000	0.0174	0.0008	0.0000
12	0.0000	0.0003	0.0000	**0.0056**	**0.0008**	0.0000	0.0000	0.0000	0.0000
13	0.0000	0.0028	0.0000	**0.0469**	**0.0097**	**0.2000**	0.0137	0.0021	0.0000
14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
15	0.0563	0.0127	0.1000	**0.0633**	**0.0281**	**0.3000**	0.0526	0.0025	0.0000
16	0.0143	0.0035	0.0000	**0.0226**	**0.0045**	0.0000	0.0012	0.0019	0.0000
All	0.0069	0.0055	0.0125	**0.0204***	0.0079	**0.0500**	0.0179	**0.0091**	0.0125

Bold entries signify the best scores obtained for each topic and metric

Topics are taken from the ImageCLEF 2010 medical retrieval track data set, and relevance scores are given for the 1,000 highest ranked images per topic. Results for the single best performing content-based system at ImageCLEF 2010 are given as a comparison. For content-based GFM, $\varepsilon = 1$ and $p = 2$

CBIR Content-based image retrieval, *GFM* global feature mapping, *bpref* binary preference, *MAP′* judged mean average precision, *P′@10* judged precision-at-ten

* Statistically significant ($p < 0.01$) with respect to brute-force CBIR

7.2.2 Content-based retrieval

Although GFM is primarily intended to be a practical means of performing multimodal image retrieval, we can also evaluate its use for content-based retrieval. Table 5 shows retrieval results obtained by our Lucene implementation of content-based GFM for each topic taken from the 2010 ImageCLEF medical retrieval track data set. For comparison, we also include in the table results obtained by brute-force CBIR and the content-based system that achieved the highest average bpref at the 2010 ImageCLEF evaluation.

The results depicted in Table 5 show that content-based GFM achieved a statistically significant increase in bpref (195.65 %, $p < 0.01$) over brute-force CBIR, but it did not perform significantly better than the brute-force method in terms of MAP′ or P′@10. Thus, the performance of content-based GFM is comparable with that of the best system at the 2010 ImageCLEF evaluation. This result is especially interesting because these methods are conceptually similar: they both utilize the same set of content-based visual descriptors and compare these descriptors with Euclidean distance. However, our Lucene implementation of content-based GFM additionally applies *tf-idf* term weighting to the image code words. As we described in Sect. 4.2.2, because $tf = 1$ for all code words in an image's surrogate document, code words with a greater *idf* are weighted more heavily. Thus, code

words corresponding to clusters containing a smaller number of visual descriptors are given more weight than code words mapped to clusters of larger sizes. This difference favors images that are more unique within the collection, and it contributes to the average increase in retrieval performance obtained by content-based GFM. Finally, the increase in bpref is also significant because, as was shown in Table 2, the response time of content-based GFM is a fraction of that required by brute-force CBIR.

Immediately apparent from the results depicted in Table 5 is the poor performance of content-based retrieval in relation to the text-based and multimodal strategies shown in Tables 3 and 4. However, it is well-known that CBIR generally does not perform as well as textual methods for literature-based image retrieval tasks such as those encountered through participation in the ImageCLEF evaluations (Müller et al. 2010a). Because the performance of CBIR systems can be so poor, it is not surprising that the community has had difficulty developing multimodal retrieval strategies that improve upon the performance of text-based approaches. In this regard, multimodal GFM is significant for its ability to consistently demonstrate an increase in performance over our textual baseline.

7.2.3 Code word expansion

Table 6 shows retrieval results obtained by our Lucene implementation of content-based GFM with varying code word expansion factors for each topic taken from the 2010

Table 6 Usefulness of query expansion for content-based global feature mapping

Topic	$\varepsilon = 1$				$\varepsilon = 2$				$\varepsilon = 3$			
	bpref	ret	rel_ret	Time	bpref	ret	rel_ret	Time	bpref	ret	rel_ret	Time
1	**0.0219**	0.27	96	31	0.0195	0.27	98	31	0.0100	0.27	99	32
2	0.0000	0.15	100	0	0.0000	0.17	100	0	0.0000	0.18	100	15
3	0.0000	0.25	95	31	**0.0065**	0.26	97	31	0.0000	0.26	97	46
4	**0.1096**	0.31	100	46	0.1040	0.32	100	47	0.0983	0.32	100	47
5	0.0000	0.25	100	0	0.0000	0.25	100	0	0.0000	0.25	100	0
6	**0.0248**	0.29	100	31	0.0232	0.30	100	31	0.0225	0.30	100	46
7	0.0000	0.34	100	46	0.0000	0.34	100	47	0.0000	0.34	100	47
8	0.0000	0.26	100	15	0.0000	0.27	100	15	0.0000	0.29	100	31
9	0.0000	0.35	100	31	0.0000	0.35	100	31	0.0000	0.35	100	46
10	0.0000	0.27	100	31	0.0000	0.27	100	31	0.0000	0.27	100	31
11	0.0657	0.36	100	15	**0.0786**	0.37	100	32	0.0668	0.37	100	46
12	0.0056	0.27	95	47	**0.0175**	0.27	95	46	0.0031	0.27	100	47
13	**0.0469**	0.28	100	47	0.0332	0.28	100	46	0.0215	0.28	100	46
14	0.0000	0.24	100	31	0.0000	0.24	100	47	0.0000	0.24	100	47
15	0.4045	0.27	99	46	0.4521	0.27	100	47	**0.4906**	0.27	100	47
16	0.0226	0.35	100	15	0.0285	0.37	100	16	**0.0488**	0.37	100	31
All	0.0439	0.28	99	29	**0.0477**	0.29	99	31	0.0476	0.29	100	38

Bold entries signify the best scores obtained for each topic and metric

Topics are taken from the ImageCLEF 2010 medical retrieval track data set. For each content-based GFM approach, $p = 2$. All reported times are the lowest of ten retrieval runs and are given in milliseconds

bpref Binary preference, *ret* percentage of total images retrieved, *rel_ret* percentage of relevant images retrieved

ImageCLEF medical retrieval track data set. We include in the table results obtained without code word expansion ($\varepsilon = 1$), with an expansion factor of two ($\varepsilon = 2$), and with an expansion factor of three ($\varepsilon = 3$). For each topic and expansion factor, we report the bpref obtained by content-based GFM, the number of images retrieved as a percentage of the total number of images in the collection (ret), the number of relevant images retrieved as a percentage of the total number of images relevant to the topic (rel_ret), and the time taken in milliseconds to obtain the results. Unlike the relevance scores presented in Tables 3, 4 and 5, here we report results for all retrieved images—instead of only the one thousand highest ranked images—to clearly demonstrate the impact of code word expansion on image ranking.

Table 6 demonstrates that, although the performance of content-based GFM is low, a code word expansion factor of one obtains nearly all relevant images by retrieving less than one percent of the total number of images in the collection. While increasing the expansion factor results in the retrieval of additional relevant images for some topics, it does not significantly improve retrieval precision, and in some cases it actually worsens performance. For example, an expansion factor of three allows content-based GFM to retrieve several additional relevant images for topic 1 compared with no code word expansion, but it decreases the bpref of topic 1 from 0.0219 to 0.0100. The limited effectiveness of code word expansion provides evidence that the *k*-means++ algorithm, despite its policy of rigid cluster membership, is already successful at producing a clustering of content-based features adequate for our retrieval experiments. Because the number of images in each cluster is small, increasing the expansion factor does not significantly affect the number of images retrieved as a percentage of the total number of images in the collection. However, code word expansion causes a modest increase in response time because a larger number of images must be scored, and longer queries require additional time to parse.

7.2.4 Intracluster ranking

Figure 6 shows the average number of images retrieved by our Lucene implementation of content-based GFM at each retrieval rank under various configurations. Because GFM represents with a single code word all images whose visual descriptors for a given feature lie within the same cluster, it lacks the ability to discriminate among images mapped to the

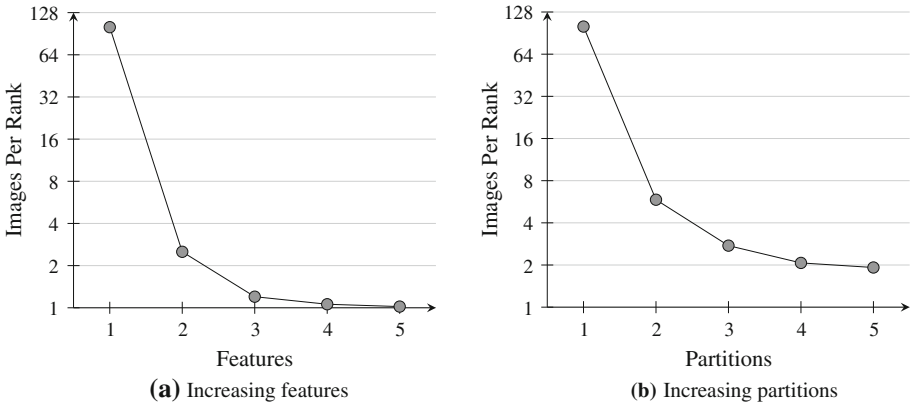


Fig. 6 Average number of images at each retrieval rank as the number of indexed features (**a**) is increased and, for a single feature, as the number of indexed subspace partitions (**b**) is increased

same code word. For example, with a query consisting of a single code word, content-based GFM will retrieve all images whose visual descriptors are in the cluster the code word represents. However, because each of the retrieved images is given the same score, their relative similarity to the example query image is lost, and any ranking of the images is meaningless. We seek to explain this behaviour by presenting in Fig. 6 the number of images given the same score by content-based GFM averaged over all retrieval ranks for all sixteen topics taken from the 2010 ImageCLEF medical retrieval track data set.

In Fig. 6a, we show the average number of images retrieved by content-based GFM per rank as the number of indexed features is increased from one to five. The results show that as we increase the number of features describing the images, we drastically decrease the average number of images retrieved per rank. Thus, increasing the number of indexed features improves the ability of content-based GFM to discriminate among visually similar images. In Fig. 6a, the number of feature partitions is one ($p = 1$), and the number of images per rank is averaged over all possible combinations of the given number of features. For example, the average number of images retrieved per rank with three features (1.20) is averaged over all sixteen ImageCLEF 2010 topics, over all retrieval ranks, and over all possible combinations of three features. Among five total features there are ten possible combinations of three features.

In Fig. 6b, we show the average number of images retrieved by content-based GFM per rank as the number of feature partitions is increased from one to five. Similar to Fig. 6a, this figure demonstrates that as we partition the visual descriptors of the images' content-based features into an increasing number of lower-dimensional vectors, we quickly decrease the average number of images retrieved per rank. Thus, increasing the number of feature partitions also improves the ability of content-based GFM to discriminate among visually similar images. Because increasing the number of feature partitions only impacts intracluster rankings, recall-based measures that are computed over all retrieved images, such as MAP, are generally not sensitive to feature partitioning. However, metrics computed on partial ranked lists, such as P@10, may be affected by the number of feature partitions. In Fig. 6b, the number of features representing each images is one, and the number of images retrieved per rank for a given number of partitions is averaged over all sixteen ImageCLEF 2010 topics, over all retrieval ranks, and over all five image representations consisting of a single content-based feature.

8 Conclusion

The images found within biomedical articles are sources of essential information to which we must provide efficient access. Not surprisingly, various image retrieval strategies have been proposed for use in the biomedical domain. Unfortunately, although they demonstrate considerable empirical success, traditional text-based image retrieval methods are often unable to retrieve images whose relevance is not explicitly mentioned in the article text. Additionally, content-based retrieval methods are unable to produce meaningful results for many literature-based information needs because visual similarity can be a poor indicator of image relevance. Due to the limitations of these unimodal strategies, practical retrieval techniques capable of fusing information from multiple modalities is desirable.

Global feature mapping (GFM) is a multimodal strategy for retrieving images from biomedical articles. The approach seeks to improve upon the precision of text-based image retrieval methods by providing a practical and efficient means of incorporating a limited amount of visual information into the retrieval process. GFM operates by (1) grouping the

global content-based features extracted from an image collection into clusters, (2) assigning images alphanumeric code words indicative of the clusters in which their features reside, (3) indexing a combination of image code words and descriptive text using a text-based information retrieval system, and (4) searching the image index using textual queries derived from multimodal topics.

We evaluated the performance of GFM on the 2010 and 2012 ImageCLEF medical retrieval track data sets. Our multimodal retrieval approach utilizing GFM demonstrated a statistically significant improvement in mean average precision over our text-based strategy, a baseline retrieval method competitive with the best performing systems evaluated at the ImageCLEF forums. Additionally, when configured for performing content-based retrieval, our approach outperformed the highest ranked content-based systems.

Although GFM's improvements in retrieval precision were small, its performance validates our intuition that visual similarity can play a small yet significant role in multimodal literature-based image retrieval tasks. Key to its success were GFM's use of an inexact representation of content-based features and its weighting of these features, in conjunction with image-related text, according to an underlying text-based retrieval model. The advantages of these two qualities are perhaps best demonstrated by the comparison of content-based GFM to brute-force CBIR, where GFM outperformed the brute-force method using the same set of features and the same similarity metric for clustering. Because we did not evaluate our particular choice of content-based features, it remains to be seen if the use of a more sophisticated image representation would result in similar improvements in retrieval precision.

To demonstrate GFM's practicality, we implemented it in two information retrieval systems: a general purpose system based on the vector space retrieval model and a biomedical system that utilizes a probabilistic retrieval model. We obtained statistically significant improvements using both systems. As further evidence of its practicality, we demonstrated that the response time of our multimodal approach is comparable to that of our text-based strategy. Owing to its empirical success, we have incorporated GFM into OpenI, a biomedical image retrieval system currently indexing over one million images from the articles included in the open access subset of PubMed Central.

Acknowledgments The authors would like to thank Dr. Md. Mahmudur Rahman and Srinivas Phadnis for extracting and preparing the content-based and text-based features of the images used in this work. This work is supported by the intramural research program of the U.S. National Library of Medicine, National Institutes of Health, and by an appointment to the NLM Research Participation Program administered by the Oak Ridge Institute for Science and Education.

References

Alpkocak, A., Ozturkmenoglu, O., Berber, T., Vahid, A. H., & Hamed, R. G. (2012). DEMIR at ImageCLEFMed 2011: Evaluation of fusion techniques for multimodal content-based medical image retrieval. In *Working notes for the CLEF 2011 workshop*.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms*, pp. 1027–1035.

Atrey, P., Hossain, M., El Saddik, A., & Kankanhalli, M. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems, 16*(6), 345–379.

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., & Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research, 3*, 1107–1135.

Beatty, M., & Manjunath, B. (1997). Dimensionality reduction using multi-dimensional scaling for content-based retrieval. In *Proceedings of the international conference on image processing*, pp. 835–838.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM, 18*(9), 509–517.

Bezdek, J. C., Pal, M. R., Keller, J., & Krisnapuram, R. (1999). *Fuzzy models and algorithms for pattern recognition and image processing*. Norwell: Kluwer.

Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 127–134.

Brin, S. (1995). Near neighbor search in large metric spaces. In *Proceedings of the 21th international conference on very large data bases*, pp. 574–584.

Caicedo, J. C., Moreno, J. G., Niño, E. A., & González, F. A. (2010). Combining visual features and text data for medical image retrieval using latent semantic kernels. In *Proceedings of the international conference on multimedia information retrieval*, pp. 359–366.

Callan, J. P., Croft, W. B., & Harding, S. M. (1992). The INQUERY retrieval system. In A. M. Tjoa, & I. Ramos (Eds.), *Database and expert systems applications* (pp. 78–83). Vienna: Springer.

Chang, E., Goh, K., Sychay, G., & Wu, G. (2003). CBSA: Content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology, 13*(1), 26–38.

Chang, S. F., Sikora, T., & Puri, A. (2001). Overview of the MPEG-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology, 11*(6), 688–695.

Chatzichristofis, S. A., & Boutalis, Y. S. (2008a). CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In A. Gasteratos, M. Vincze, & J. K. Tsotsos (Eds.), *Proceedings of the 6th international conference on computer vision systems, Lecture Notes in Computer Science* (Vol. 5008, pp. 312–322). Berlin: Springer.

Chatzichristofis, S. A., & Boutalis, Y. S. (2008b). FCTH: Fuzzy color and texture histogram: A low level feature for accurate image retrieval. In *Proceedings of the 9th international workshop on image analysis for multimedia interactive services*, pp. 191–196.

Ciaccia, P., Patella, M., & Zezula, P. (1997). M-tree: An efficient access method for similarity search in metric spaces. In *Proceedings of the 23rd international conference on very large data bases*, pp. 426–435.

Clinchant, S., Csurka, G., Ah-Pine, J., Jacquet, G., Perronnin, F., Sánchez, J., et al. (2010). XRCE's participation in Wikipedia retrieval, medical image modality classification and ad-hoc retrieval tasks of ImageCLEF 2010. In *Working notes for the CLEF 2010 workshop*.

Datta, R., Ge, W., Li, J., & Wang, J. Z. (2007). Toward bridging the annotation-retrieval gap in image search. *IEEE Multimedia, 14*(3), 24–35.

Datta, R., Joshi, D., Li, J., Wang, J.Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Compting Surveys 40*(2), 5:1–5:60.

Demner-Fushman, D., Antani, S., Simpson, M., & Thoma, G. (2008). Combining medical domain ontological knowledge and low-level image features for multimedia indexing. In *Proceedings of the language resources for content-based image retrieval workshop (OntoImage)*, pp. 18–23.

Demner-Fushman, D., Antani, S., Simpson, M., & Thoma, G. R. (2009). Annotation and retrieval of clinically relevant images. *International Journal of Medical Informatics, 78*(12), 59–67.

Demner-Fushman, D., Antani, S., Simpson, M., & Thoma, G. R. (2012). Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering* (to appear).

Demner-Fushman, D., & Lin, J. (2007). Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics, 33*(1), 63–103.

de Vries, A. P. (1999). *Content and multimedia database management systems*. PhD thesis, University of Twente.

de Vries, A. P., & Westerveld, T. (2004). A comparison of continuous vs. discrete image models for probabilistic image and video retrieval. In *International conference on image processing*, Vol. 4, pp. 2387–2390.

Duygulu, P., Barnard, K., de Freitas, J., & Forsyth, D. (2006). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In A. Heyden, G. Sparr, M. Nielsen, & P. Johansen (Eds.), *Computer vision—ECCV 2006, Lecture Notes in Computer Science* (Vol. 2353, pp. 349–354). Berlin: Springer.

Ferhatosmanoglu, H., Tuncel, E., Agrawal, D., & Abbadi, A. E. (2001). Approximate nearest neighbor searching in multimedia databases. In *Proceedings of the 17th international conference on data engineering*, pp. 503–511.

Gkoufas, Y., Morou, A., & Kalamboukis, T. (2011). Combining textual and visual information for image retrieval in the medical domain. *The Open Medical Informatics Journal, 5*(Suppl 1), 50–57.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. In *Proceedings of the ACM SIGMOD international conference on the management of data*, pp. 47–57.

Hamer, O. W., Aguirre, D. A., Casola, G., Lavine, J. E., Woenckhaus, M., & Sirlin, C. B. (2006). Fatty liver: Imaging patterns and pitfalls. *Radiographics, 26*(6), 1637–1653.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Proceedings of the fourth alvey vision conference*, pp. 147–151.

Helbich, T. H., Heinz-Peer, G., Eichler, I., Wunderbaldinger, P., Götz, M., Wojnarowski, C., Brasch, R. C., Herold, C. J.et al. (1999). Cystic fibrosis: CT assessment of lung involvement in children and adults. *Radiology, 213*(2), 537–544.

Herbrich, R., Graepel, T., & Campbell, C. (2001). Bayes point machines. *The Journal of Machine Learning Research, 1*, 245–279.

Hersh, W., Müller, H., & Kalpathy-Cramer, J. (2009). The ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging, 22*(6), 648–655.

Ide, N. C., Loane, R. F., & Demner-Fushman, D. (2007). Essie: A concept-based search engine for structured biomedical text. *Journal of the American Medical Informatics Association, 1*(3), 253–263.

Indyk, P. (2004). Nearest neighbors in high-dimensional spaces. In J. E. Goodman & J. O'Rourke (Eds.), *Handbook of discrete and computational geometry* (2nd ed., pp. 877–892). Boca Raton: CRC Press.

Indyk, P., & Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on the theory of computing*, pp. 604–613.

Jégou, H., Douze, M., & Schmid, C. (2011). Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(1), 117–128.

Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision, 45*(2), 83–105.

Kalpathy-Cramer, J., & Hersh, W. (2010). Multimodal medical image retrieval: Image categorization to improve search precision. In *Proceedings of the international conference on multimedia information retrieval*, pp. 165–174.

Kohonen, T. (2001). *Self-organizing maps, information sciences* (Vol. 30, 3rd ed.). Berlin: Springer.

Lacoste, C., Lim, J. H., Chevallet, J. P., & Le, D. (2007). Medical-image retrieval based on knowledge-assisted text and image indexing. *IEEE Transactions on Circuits and Systems for Video Technology, 17*(7), 889–900.

Lancaster, F. W., & Fayen, E. G. (1973). *Information retrieval on-line*. Los Angeles: Melville Publishing.

Langlotz, C. P. (2006). RadLex: A new method for indexing online educational materials. *Radiographics, 26*(6), 1595–1597.

Lavrenko, V., Manmatha, R., & Jeon, J. (2003). A model for learning the semantics of pictures. In *Proceedings of the seventeenth annual conference on neural information processing systems*, Vol. 16, pp. 553–560.

Li, J., & Wang, J. Z. (2008). Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(6), 985–1002.

Lim, J. H., & Chevallet, J. P. (2005). Vismed: A visual vocabulary approach for medical image indexing and retrieval. In G. Lee, A. Yamada, H. Meng, & S. Myaeng (Eds.), *Information retrieval technology, Lecture Notes in Computer Science* (Vol. 3689, pp. 84–96). Berlin: Springer.

Lindberg, D., Humphreys, B., & McCray, A. (1993). The unified medical language system. *Methods of Information in Medicine, 32*(4), 281–291.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, Vol. 2, pp. 1150–1157.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110.

Lux, M., & Chatzichristofis, S. A. (2008). LIRe: Lucene image retrieval: An extensible Java CBIR library. In *Proceedings of the 16th ACM international conference on multimedia*, pp. 1085–1088.

Mäenpää, T. (2003). *The local binary pattern approach to texture analysis—Extensions and applications*. PhD thesis, University of Oulu.

Miller, G. A. (1995). WordNet: A lexical database for English. *Commun ACM, 38*(11), 39–41.

Müler, H., Rosset, A., Vallée, J. P., & Geissbuhler, A. (2003). Integrating content-based visual access methods into a medical case database. In R. Baud, M. Fieschi, P. Le Beux, & P. Ruch (Eds.), *The new navigators: From professionals to patients, studies in health technology and informatics* (Vol. 95, pp. 480–485). Amsterdam: IOS Press.

Müller, H., Michoux, N., Bandon, D., & Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications—Clinical benefits and future directions. *International Journal of Medical Informatics, 73*(1), 1–23.

Müller, H., Clough, P., Deselaers, T., Caputo, B. (eds) (2010). *ImageCLEF: Experimental evaluation in visual information retrieval, the information retrieval series* (Vol. 32). Berlin: Springer.

Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Kahn, C.E., Jr., & Hersh, W. (2010b). Overview of the CLEF 2010 medical image retrieval track. In: *Working notes of CLEF 2010*.

Müller, H., de Herrara, A. G. S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., & Eggel, I. (2012). Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In: *Working notes for the CLEF 2012 workshop*.

National Center for Biotechnology Information. (2010). Entrez programming utilities help. http://www.ncbi.nlm.nih.gov/books/NBK25501/.

Ng, R. T., & Sedighian, A. (1996). Evaluating multidimensional indexing structures for images transformed by principal component analysis. In: I. K. Sethi, & R. C. Jain (Eds.), *Proceedings of SPIE, storage an retrieval for still image and video databases*, Vol. 2670, pp. 50–61.

Nowak, E., Jurie, F., & Triggs, B. (2006). Sampling strategies for bag-of-features image classification. In A. Leonardis, H. Bischof, & A. Pinz (Eds.), *Computer vision—ECCV 2006, Lecture Notes in Computer Science* (Vol. 3954, pp. 490–503). Berlin: Springer.

Pham, T. T., Maillot, N. E., Lim, J. H., & Chevallet, J. P. (2007). Latent semantic fusion model for image retrieval and annotation. In *Proceedings of the sixteenth ACM conference on information and knowledge management*, pp. 439–444.

Rahman, M., Antani, S., Long, R., Demner-Fushman, D., & Thoma, G. (2010). Multi-modal query expansion based on local analysis for medical image retrieval. In B. Caputo, H. Müller, T. Syeda-Mahmood, J. Duncan, F. Wang, & J. Kalpathy-Cramer (Eds.), *Medical content-based retrieval for clinical decision support, Lecture Notes in Computer Science* (Vol. 5853, pp. 110–119). Berlin: Springer.

Rahman, M. M., Antani, S., & Thoma, G. (2009). A medical image retrieval framework in correlation enhanced visual concept feature space. In *Proceedings of the 22nd IEEE international symposium on computer-based medical systems*.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., et al. (2010). A new approach to cross-modal multimedia retrieval. In *Proceedings of the international conference on multimedia*, pp. 251–260.

Richardson, W. S., Wilson, M. C., Nishikawa, J., & Hayward, R. S. (1995). The well-built clinical question: A key to evidence-based decisions. *ACP Journal Club, 123*(3), A12–A13.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The Earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision, 40*(2), 99–121.

Sakai, T. (2007). Alternatives to bpref. In *Proceedings of the 30th annual international acm sigir conference on research and development in information retrieval*, pp. 71–78.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Simpson, M., Rahman, M. M., Demner-Fushman, D., Antani, S., Thoma, G. R. (2009). Text- and content-based approaches to image retrieval for the ImageCLEF 2009 medical retrieval track. In *Working notes for the CLEF 2009 workshop*.

Simpson, M., Rahman, M. M., Phadnis, S., Apostolova, E., Demner-Fushman, D., Antani, S., et al. (2011). Text- and content-based approaches to image modality classification and retrieval for the ImageCLEF 2011 medical retrieval track. In *Working notes for the CLEF 2011 workshop*.

Simpson, M., Rahman, M. M., Singhal, S., Demner-Fushman, D., Antani, S., & Thoma, G. (2010). Text- and content-based approaches to image modality detection and retrieval for the ImageCLEF 2010 medical retrieval track. In *Working notes for the CLEF 2010 workshop*.

Simpson, M. S., You, D., Rahman, M. M., Antani, S. K., Thoma, G. R., & Demner-Fushman, D. (2012a) Towards the creation of a visual ontology of biomedical imaging entities. In *Proceedings of the annual symposium of the American medical informatics association (AMIA)*, (to appear).

Simpson, M. S., You, D., Rahman, M. M., Demner-Fushman. D., Antani, S., & Thoma, G. (2012b) ITI's participation in the ImageCLEF 2012 medical retrieval and classification tasks. In *Working notes for the CLEF 2012 workshop*.

Sivic, J., & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of the ninth IEEE international conference on computer vision*, Vol. 2, pp. 1470–1477.

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on information and knowledge management*, pp. 623–632.

Squire, D. M., Müller, W., Müller, H., & Pun, T. (2000). Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters, 21*(13–14), 1193–1198.

Srinivasan, G. N., & Shobha, G. (2008). Statistical texture analysis. In *Proceedings of world academy of science, engineering and technology*, Vol. 36, pp. 1264–9.

Tamura, H., Mori, S., & Yamawaki, T. (1978). Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics, 8*(6), 460–473.

Uhlmann, J. K. (1991). Satisfying general proximity/similarity queries with metric trees. *Information Processing Letters, 40*(4), 175–179.

Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval*. Cambridge: Digital Libraries and Electronic Publishing, The MIT Press.

Wang, C., Jing, F., Zhang, L., & Zhang, H. J. (2008). Scalable search-based image annotation. *Multimedia Systems, 14*(4), 205–220.

Yang, J., Jiang, Y. G., Hauptmann, A. G., & Ngo, C. W. (2007). Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on multimedia information retrieval*, pp. 197–206.

Yianilos, P. N. (1993). Data structures and algorithms for nearest neighbor search in general metric spaces. In *Proceedings of the fourth annual ACM-SIAM symposium on discrete algorithms*, pp. 311–321.

Zezula, P., Amato, G., Dohnal, V., & Batko, M. (2006). *Similarity Search: The metric space approach, advances in database systems* (Vol. 32). Berlin: Springer.

Zhou, X., Depeursinge, A., & Müller, H. (2010). Information fusion for combining visual and textual image retrieval in ImageCLEF@ICPR. In D. Ünay, Z. Çataltepe, & A. Aksoy (Eds.), *Recognizing patterns in signals, speech, images and videos, Lecture Notes in Computer Science* (Vol. 6388, pp. 129–137). Berlin/Heidelberg: Springer.