Contents

22

Systematic Evaluations and Ground Truth

Jayashree Kalpathy-Cramer and Henning Müller

[1] Oregon Health & Science University, Portland, OR kalpathy@ohsu.edu
[2] Business Information Systems, University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland henning.mueller@hevs.ch

Abstract. Every year, we see the publication of new algorithms for medical image analysis including segmentation, registration, classification and retrieval in the literature. However, in order to be able to translate these advances into clinical practice, the relative effectiveness of these algorithms needs to be evaluated.

In this chapter, we begin with a motivation for systematic evaluations in science and more specifically in medical image analysis. We review the components of successful evaluation campaigns including realistic data sets and tasks, the gold standards used to compare systems against, the choice of performance measures and finally workshops where participants can share their experiences with the tasks and explain the various approaches. We also describe some of the popular evaluation campaigns in retrieval, classification, segmentation and registration techniques. We describe the challenges in organizing such campaigns including the acquisition of databases of images of sufficient size and quality, establishment of sound metrics and ground truth, management of manpower and resources, motivation of participants, and the maintenance of a friendly level of competitiveness among participants. We conclude with lessons learned over the years of organizing campaigns, including successes and road-blocks.

22.1 Introduction

Medical images are being produced in ever-increasing quantities as a result of the digitization of medical imaging and advances in imaging technology in the last two decades. The assorted types of clinical images are critical in patient care for diagnosis and treatment, monitoring the effect of therapy, education and research. The previous chapters have described a number of techniques used for medical image analysis from 3D image reconstruction to segmentation and registration to image retrieval. The constantly expanding set of algorithms being published in the computer vision, image processing, machine learning and medical image analysis literature underscores the need for sound evaluation methodology to show progress based on the same data and tasks.

It has been shown that many of these publications provide limited evaluation of their methods using small or proprietary data sets, making a fair comparison of the performance of the proposed algorithm with previous algorithms difficult [1, 2]. Often, the difficulty in obtaining a high quality data sets with ground truth can be an impediment to computer scientists without access to clinical data. We believe that newly proposed algorithms must be compared to the existing state-of-the-art using common data sets with application-specific, validated metrics before they are likely to be incorporated into clinical applications. By providing all participants with equal access to realistic tasks, validated data sets (including ground truth), and fora for discussing results, evaluation campaigns can enable the translation of superior theoretical techniques to meaningful applications in medicine.

22.2 Components for Successful Evaluation Campaigns

Evaluation is a critical aspect of medical image analyses and retrieval. In the literature, many articles claim superior performance compared to previously-published algorithms. However, in order to be able to truly compare and contrast the performance of these techniques, it is important to have a set of well-defined, agreed-upon tasks performed on common collections using meaningful metrics. Even if the tasks are very different from a technical standpoint (segmentation vs. retrieval, for example), their evaluation can share many common aspects. Evaluation campaigns can provide a forum for more robust evaluations and equitable comparisons between different techniques.

22.2.1 Application and Realistic Task

First of all, the goal of the algorithms being evaluated must be well understood. A technique such as image segmentation is useful in many clinical areas; however, to perform a thorough evaluation of such algorithms, one must keep the ultimate application in mind. For example, consider the following two segmentation tasks: tumor segmentation to monitor a response to cancer therapy, and anatomical segmentation of the brain from an fMRI study. The nature of each task informs the choice of the optimal evaluation metric. In the first case, missing portions of the tumor (and thereby under-estimating its size) can have serious consequences, and therefore penalties for under-segmentation might be more appropriate. In other applications, however, under-segmentation and over-segmentation may be considered to be equally inconvenient.

An image retrieval system used for performing a systematic review might have different goals than a system used to find suitable images for a lecture or scientific presentation. In the first case, the goal might be to find every relevant article and image, while in the second case a single image that meets the search need might be sufficient. For some applications accuracy might be more important while for those being used in real-time, speed can be critical.

Evaluation campaigns are usually geared toward a specific clinical application. For instance, the Medical Image Computing and Computer Assisted Intervention (MICCAI) grand challenges for segmentation [3] target very specific tasks (e.g., segmentation of prostate, liver etc.). The goal for the image retrieval task in the Cross Language Evaluation Forum (CLEF)[3], medical retrieval campaign (ImageCLEF 2009) is to retrieve images from the medical literature that meet information needs of clinicians [4].

Once the overall goal of the algorithm has been well understood, it is important to identify a set of realistic, meaningful tasks towards that goal. For evaluating an image retrieval system this might consist of a set of reasonable search topics (often derived from user studies or log file analyses [5–7]). For the evaluation of a registration algorithm, an appropriate task might be to register structures in an atlas to equivalent structures in a set of patients. For segmentation challenges the task might be to segment normal anatomical organs (e.g., lung, liver, prostate, vasculature) or abnormalities (e.g., lung nodule, liver tumor, lesion). Classification tasks might include classifying radiographs based on the anatomical location [8], or separating voxels in the brain into white, gray matter and Cereborspinal Fluid (CSF) in Magnetic Resonance Imaging (MRI) data [9]. The number and scale of these tasks (how many topics, how many structures for how many different patient studies, etc.) must be carefully chosen to support the derivation of statistically meaningful metrics.

22.2.2 Collections of Images and Ground Truth

In order to perform a fair comparison of different algorithms, ideally all techniques must be compared on the same database or collection of images. Additionally, these data must be of a sufficient variety, so as to encompass the full range of data found in realistic clinical situations.

Often, computer scientists wishing to evaluate state-of-the-art algorithms do not have access to large amounts of clinical data, thereby limiting the scope of their evaluations. In general, getting access to the large collections necessary for a robust evaluation has been challenging, even for researchers associated with clinical facilities due to issues of cost, privacy and resources.

Recently, there has been a growing trend towards making databases of images available openly towards the goal of promoting reproducible science. Many governmental agencies, including the National Institutes of Health (NIH) in the United States have funded initiatives like the Lung Imaging Database Consortium (LIDC) [10] and the Alzheimer's Disease Neuroimaging Initiative (ADNI)[4] [11] that create well-curated collections of images and clinical data. These collections are typically anonymized to preserve patient privacy, and openly available to researchers. These and other similar initiatives

[3] `http://www.clef-campaign.org/`
[4] http://www.loni.ucla.edu/ADNI

foster collaboration between groups across the world and researchers from different domains including clinicians, imaging scientists, medical physicists and computer scientists.

The issue of attaining ground truth or a *gold standard* continues to be challenging. For most applications, the best references are manually generated, and therefore their construction is an extremely time consuming and resource-intensive task. However, often the absolute truth is unknown or unknowable. For instance, it would be quite difficult to absolutely verify the registration of a brain atlas to the MRI of a patient. Similarly, in order to evaluate the performance of segmentation algorithms, experts usually manually delineate the Regions of Interest (ROI). However, the *true* segmentation of a tumor that is not physically resected may never be definitively established.

Additionally, even if there theoretically exists an "objective truth", experts often disagree on what constitutes that truth. In cases with more than one human rater, these questions of inter-observer agreement make the creation of a gold standard difficult. By providing segmentation in the form of annotations of lung nodules by four independent raters, the LIDC database exemplifies this difficulty in obtaining ground truth. Recent research has demonstrated that all four raters agreed on the presence of a nodule at a given location in only approximately 40% of the cases [12].

The problem is not limited to segmentation gold standards. When evaluating the effectiveness of information retrieval systems, relevance judgments are typically performed by domain experts. However, the kappa-measures (used to quantify inter-observer agreement) between experts in relevance judgment tasks often indicate significant levels of disagreement as to which documents count as "relevant". The concept of relevance as applied to images is particularly problematic, as the relevance of a retrieved image can depend on the context in which the search is being performed. An additional source of judgment difficulty is that domain experts tend to be more strict than novices [4], and so the validity of their judgments for a particular task may depend on the nature of the intended users.

22.2.3 Application-Specific Metric

The choice of metrics should depend on the clinical goal of the algorithm being evaluated. In classification tasks, *error rate* is often the metric of choice. However, if the cost of a miss (e.g., missed detection of a lung nodule) is high, a non-symmetric measure of cost can be used. For registration and segmentation, measures related to volumetric overlap or surface distances can be used. If the goal of an image retrieval system is to find a few good images to satisfy the information need, early precision might be a good measure. On the other hand, if the goal of the task is to find every relevant image in the database, recall-oriented measures might be better suited.

In most evaluation campaigns, the evaluation measures are specified at the outset. Often, a single measure that combines different aspects of the evalua-

tion is preferred, as this makes comparisons between participants straightforwardly (see Sec. 22.3).

22.2.4 Organizational Resources and Participants

Evaluation campaigns are usually conducted on a voluntary basis as funding for such efforts can be hard to obtain. Organizing such campaigns can be quite resource and time-intensive as the organizers need to acquire databases of images of sufficient size and quality, establish sound performance metrics and ground truth, provide the tabulation of the results, potentially organize the publications of the proceedings and motivate participation by balancing competitiveness with a friendly spirit of collaboration and cooperation.

Having a diverse set of loyal participants is a hallmark of a good evaluation campaign. Often, significantly larger number of groups register for and obtain data to evaluation campaigns than actually submit results and participate in the workshops. It is important to strive to increase the number of actual participants as the collaborative atmosphere, as found in the evaluation campaigns engenders strides in the field by enabling participants to leverage each other's techniques. One of the challenges of organizing an evaluation campaign is providing tasks that are appropriate for research groups with varying levels of expertise and resources. If the task is too challenging and requires massive computing resources, participation by groups without access to such facilities can be limited. On the other hand, if the task is regarded as being too trivial, the sought-after participation by the leading researchers in the area can be difficult to attract. Options explored by some of the campaigns include providing multiple tasks at different levels, providing baseline runs or systems that can be combined in a modular fashion with the participants' capabilities (ImageCLEF) or providing the option of submitting both fully automatic and semi-automatic runs. Participants can generally be motivated by the opportunity to publish, by providing access to large collections of images that they might otherwise not have access to, as well as the spirit of the competition.

Many evaluation campaigns (see Sect. 22.4) organize workshops at the end of the evaluation cycle where participants are invited to present their methods and participate in discussions. They are often, but not exclusively, held in conjunction with larger conferences.

These workshops are an important part of evaluation cycle and can be a great opportunity for researchers from across the globe to meet face-to-face in an effort to advance their fields. In addition to the technical aspects, the workshops also provide a chance for participants to provide feedback to the organizers about the collections, the nature of the task as well as the level of difficulty and organizational issues. They also provide a forum where participants can offer suggestions for future tasks, collections, and metrics. Furthermore, an in-person workshop is an excellent opportunity to recruit new organizers, thereby aiding the sustainability of the campaign.

22.3 Evaluation Metrics and Ground Truth

This section describes several of the commonly used performance metrics in medical imaging tasks including registration, segmentation and retrieval.

22.3.1 Registration

One of the first steps in the evaluation of the performance of registration algorithms is simply a visual check. This can be accomplished using image fusion in which one image is overlaid on top of the other with partial transparency and potentially different colors. Alternatively, the images can be evaluated using a checkerboard pattern.

At is has been introduced in Chapter **??**, the intensities of the registered images can be used as metric [13]. The rationale behind this approach is that the better the registration performance, the sharper the composited image is expected to be as the registered image will be closer to the target image. With respect to the template image j, the intensity variance is given as

$$\mathrm{IV}_j(x) = \frac{1}{M-1}\sum_{i=1}^{M}(T_i(h_{ij}(x)) - \mathrm{ave}_j(x))^2, \tag{22.1}$$

where $\mathrm{ave}_j(x) = \frac{1}{M}\sum_{i=1}^{M} T_i(h_{ij}(x))$ denotes the average, $h_{ij}(x)$ is the transformation from image i to j and M is the number of images being evaluated.

Other methods include comparing the forward and reverse transforms resulting from the registration. In a perfect situation, the forward transform would be the inverse of the reverse. The inverse consistency error measures the error between a forward and reverse transform compared to an identity mapping [13]. The voxel-wise Cumulative Inverse Consistency Error (CICE) is computed as

$$\mathrm{CICE}_j(x) = \frac{1}{M}\sum_{i=1}^{M}\|h_{ji}(h_{ij}(x)) - x\|^2, \tag{22.2}$$

where $\|\dot{\|}$ denotes the standard Euclidean norm. The CICE is a necessary but not sufficient metric for evaluating registration performance [13].

In addition, Christensen et al. [13] note that the transforms resulting from registration algorithms should satisfy the transitivity property. If H_{AB} is the transform from A to B, transitivity implies that $h_{CB}(h_{BA}(x)) = h_{CA}(x)$ or $h_{AC}(h_{CB}(h_{BA}(x))) = x\ \forall A, B, C\ T_i$ is the i^{th} image of the set and h_{ij} is the registration transform.

The Cumulative Transitive Error (CTE) is defined as

$$\mathrm{CTE}_k(x) = \frac{1}{(M-1)(M-2)}\sum_{i=1 i\neq 1}^{M}\sum_{j=1}^{M}\|h_{ki}(h_{ij}(h_{jk}(x))) - x\|^2 \tag{22.3}$$

Another common approach in registration is to define a structure in the initial image (e.g., in an atlas), register the initial image to the final image (e.g., actual patient image), and deform the structure using the resulting deformation field. If manual segmentation is available on the final image, then many of the metrics defined in the following subsection can be used to compare the manual segmentation to that obtained using registration of the atlas.

22.3.2 Segmentation

Image segmentation, the task of delineating an image into meaningful parts or objects, is critical for many clinical applications. One of the most challenging aspects in evaluating the effectiveness of segmentation algorithms is the establishment of ground truth against which the computer-derived segmentations are to be compared.

Metrics without Ground Truth

In real-life clinical images, establishing true segmentation often is difficult due to poor image quality, noise, non-distinct edges, occlusion and imaging artifacts. Physical and digital "phantoms" have been used to establish absolute ground truth; however, they do not contain the full range of complexity and variability of clinical images [14].

To avoid the use of phantoms, Warfield et al. [14] proposed the Simultaneous Truth and Performance Level Estimation (STAPLE) procedure, an expectation-maximization algorithm that computes a probabilistic estimate of true segmentation given a set of either automatically generated or manual segmentations. STAPLE has been used for establishing ground truth in the absence of manual segmentations as well as to provide a quality metric for comparing the performance of segmentation algorithms.

However, it should be pointed out that manual segmentations are not reproducible, i.e., they suffer from inter- as well as intra-observer variability, and hence, their usefulness in absolute evaluation of medical image segmentation is limited.

Volume-Based Metrics

Consider the case where the results of a segmentation algorithm are being compared to ground truth using binary labels (i.e., a label of "1" is given to a voxel that belongs to the object being segmented and a label of "0" otherwise). Let A indicate the voxels belonging to the object according to the segmentation under consideration (as determined by either another user or an automatic algorithm) and G refers to the ground truth (Fig. 22.1). A commonly used simple measure is based on the volumes enclosed by the respective segmentations. The Volumetric Difference (VD) [15] is defined as

$$\mathrm{VD} = \frac{V_a - V_g}{V_g} \times 100 \tag{22.4}$$

The Absolute Volumetric Difference (AVD) is the absolute value of the above measure. However, these measures do not take into account the spatial locations of the respective volumes, and hence have limited utility when used alone. Additionally, they are not symmetric.

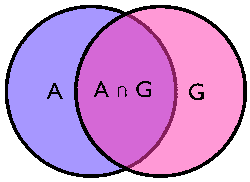


Fig. 22.1. Venn diagram. The diagram shows the intersection between the segmented label A and the gold standard G.

The Dice [16] and Jaccard coefficients [17] are the most commonly used measures of spatial overlap for binary labels. In both cases, the values for the coefficients range from zero (no overlap) to one (perfect agreement).

$$D = \frac{2\,|A \cap G|}{|A| + |G|} \times 100 \qquad J = \frac{|A \cap G|}{|A \cup G|} \times 100 \tag{22.5}$$

This is also sometimes known as the *relative overlap measure*. As all these measures are related to each other, typically only one or the other is calculated.

$$J = \frac{D}{2 - D} \tag{22.6}$$

The Dice coefficient has been shown to be a special case of the kappa coefficient [18], a measure commonly used to evaluate inter-observer agreement. As defined, both of these measures are symmetric, in that over- or under-segmentation errors are weighted equally. To characterize over- and under-segmentations in applications where these might be important (e.g., tumor delineation where the cost for missing the tumor is higher), false positive and false negative Dice measures can be used. The False Positive Dice (FPD) is measure of voxels that are labeled positive (i.e., one) by the segmentation algorithm being evaluated but not the ground truth and hence is a measure of over-segmentation. The False Negative Dice (FND) is a measure of the voxels that were considered positive according to the ground truth but missed by the segmentation being evaluated. Let $\bar{A}$ and $\bar{G}$ be the complements of the segmentation and the ground truth (i.e., they are the voxels labeled 0).

$$\mathrm{FPD} = \frac{2\,|A \cap \bar{G}|}{|A| + |G|} \times 100 \qquad \mathrm{FND} = \frac{2\,|\bar{A} \cap G|}{|A| + |G|} \times 100 \tag{22.7}$$

The above-mentioned spatial overlap measures depend on the size and shape of the object as well as the voxel size relative to the object size. Small differences

in the boundary of the segmentation can result in relatively large errors in small objects compared to large objects.

Additionally, the measures discussed above assume that we are comparing the results of one algorithm with one set of ground truth data. However, often there is either no ground truth available, or alternatively, manual segmentations from multiple human raters are available. In these cases, many approaches have been considered, ranging from fairly simplistic majority votes for the class membership of each voxel to the STAPLE algorithm mentioned above [14] or the Williams index [19].

The Williams index [19, 20] considers a set of r raters labeling a set of n voxels with one of l labels. D is the label map of all raters where D_j is the label map for rater j and D_{ij} represents the label of rater j for voxel i. Let $a(Dj, D_{ij})$ be the agreement between rater j and j_i over all n voxels. Several agreement measures can be used. The Williams index I_j as defined below, can be used to assess if observer j agrees at least as much with other raters as they agree with each other.

$$I_j = \frac{(r-2)\sum_{j' \neq j}^{r} a(D_j, D_{j'})}{2\sum_{j' \neq j}^{r} \sum_{j'' \neq j}^{j'} a(D_{j'}, D_{j''})} \tag{22.8}$$

All of the metrics discussed thus far have assumed that the class labels were binary, i.e. each voxel belonged to either the structure or the background. Although this has been the case historically and continues to be the predominant mode for classification, more recently, methods as well as probabilistic methods have required the use of partial labels for class membership. Crum et al. [21] discussed the lack of sufficient metrics to evaluate the validity of the algorithms in these cases. They proposed extensions of the Jaccard similarity measure, referred to as Generalized Tanimoto Coefficients (GTC) using results from fuzzy set theory. These overlap measures can be used for comparison of multiple fuzzy labels defined on multiple subjects.

Surface-Based Metrics

Unlike the region-based approaches, surface distance metrics are derived from the contours or the points that define the boundaries of the objects. The Hausdorff Distance (HD) is commonly used to measure the distance between point sets defining the objects. The HD (a directed measure as it is not symmetric) between A and G, $h(A, G)$ is the maximum distance from any point in A to a point in G and is defined as

$$h(A, G) = \max_{a \in A}(d(a, G)) \tag{22.9}$$

where $d(a, G) = \min_{G \in G} \|a - g\|$. The symmetric HD, $H(A, G)$ is the larger of the two directed distances, defined more formally as

$$H(A, G) = \max(h(A, G), h(G, A) \tag{22.10}$$

Table 22.1. Fourfold table. A 2x2 table for relevant and retrieved objects.

	Relevant	Not Relevant	
Retrieved	$A \cap B$	$\bar{A} \cap B$	B
Not Retrieved	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$	$\bar{B}$
	A	$\bar{A}$	

The Hausdorff distance, although commonly used, has a few limitations. It is highly susceptible to outliers resulting from noisy data. However, many variations of a more robust version of this measure have been used for applications in segmentation as well as registration.

Software Tools

The Valmet software tool, although no longer actively supported, incorporated many of these measures and has been used for evaluation and visualization of 2D and 3D segmentation algorithms [22]. It includes the measures: volumetric overlap (true and false positives, true and false negatives), probabilistic distances between segmentations, Hausdorff distance, mean absolute surface distance, and interclass correlation coefficients for assessing intra-, inter-observer and observer-machine variability. The software also enabled the user to visualize the results.

22.3.3 Retrieval

Information Retrieval (IR) has a rich history of evaluation campaigns, beginning with the Cranfield methodology in the early 60's [23] and the System for the Mechanical Analysis and Retrieval of Text (SMART) [24], to more recent Text Retrieval Conference (TREC)[5] campaigns [25].

Precision, Recall and F-Measure

Precision and recall are two of the most commonly used measures for evaluation retrieval systems, both for text and images. Precision is defined the fraction of the documents retrieved that are relevant to the user's information need. For binary relevance judgments, precision is analogous to positive predictive value. Consider a 2×2 table for relevant and retrieved objects (Tab. 22.1) where A is the set of relevant objects and B is the set of retrieved objects

$$\text{precision} = \frac{\text{relevant documents retrieved}}{\text{retrieved documents}} \qquad P = \frac{|A \cap B|}{|B|} \tag{22.11}$$

Precision is often calculated for a given number of retrieved objects. For instance P_{10} (precision at 10) is the number of relevant objects in the first ten

[5] `http://trec.nist.gov/`

objects retrieved. Recall, on the other hand, is the ratio of the relevant objects retrieved to the total number of relevant objects in the collection

$$\text{recall} = \frac{\text{relevant documents retrieved}}{\text{relevant documents}} \qquad R = \frac{|A \cap B|}{|A|} \tag{22.12}$$

Recall is equivalent to sensitivity. It is important to note that recall does not consider the order in which the relevant objects are retrieved or the total number of objects retrieved.

A single effectiveness measure, based on both precision and recall was proposed by van Rijsbergen [26]

$$E = 1 - \frac{1}{\alpha/P + (1-\alpha)/R} \tag{22.13}$$

where α denoting a fraction between zero and one can be used to weigh the importance of recall relative to precision in this measure.

The weighted F-score (F-measure) is related to the effectiveness measure as $1 - E = F$

$$F = \frac{1}{\alpha/P + (1-\alpha)/R} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \tag{22.14}$$

where $\beta^2 = \frac{1-\alpha}{\alpha}$ and $\alpha \in [0,1]$, $\beta^2 \in [0,\infty]$.

In the balanced case where both precision and recall are weighted equally, $\alpha = 1/2$ and $\beta = 1$. It is commonly written as F_1, or $F_{\beta=1}$. In this case, the above equation simplifies to the harmonic mean

$$F_{\beta=1} = \frac{2PR}{P+R}$$

However, α or β can be used to provide more emphasis to precision or recall as values of $\beta < 1$ emphasize precision, while values of $\beta > 1$ emphasize recall.

Average Precision

Overall, precision and recall are metrics based on the set of objects retrieved but not necessarily the position of the relevant objects. Ideal retrieval systems should retrieve the relevant objects ahead of the non-relevant ones. Thus, measures that consider the order of the returned items are also important. Average precision, defined as the average of the precisions computed for each relevant item, is higher for a system where the relevant documents are retrieved earlier.

$$\text{AP} = \frac{\sum_{r=1}^{N}(P(r) \times \text{rel}(r))}{\text{number of relevant documents}} \tag{22.15}$$

where r is the rank, N the number retrieved, rel() a binary function on the relevance of a given rank, and $P()$ precision at a given cut-off rank.

In evaluation campaigns with many search topics, the Mean Average Precision (MAP) is a commonly used measure. The MAP is the mean of the average precisions for all the search topics and is meant to favor systems that return more relevant documents at the top of the list. However, the maximum MAP that a system can achieve is limited by its recall, and systems can have very high early precision despite having low MAP.

Software

Trec_eval, a software package created by Chris Buckley[6] is commonly used for retrieval campaigns. This package computes a large array of measures including the ones specified above [27]. The ideal measure depends on the overall objective, but many information retrieval campaigns, both text-based (TREC) and image-based (ImageCLEF) use MAP as the lead metric but also consider the performance of early precision.

22.4 Examples of Successful Evaluation Campaigns

22.4.1 Registration

Image registration is another critical aspect of medical image analysis. It is used to register atlases to patients, as a step in the assessment of response to therapy in longitudinal studies (serial registration), and to superimpose images from different modalities (multi-modal registration). Traditionally, rigid and affine techniques were used for registration. More recently, deformable or non-rigid registration techniques have been used successfully for a variety of application including atlas-based segmentation, and motion tracking based on 4D CT. The evaluation of non-rigid registration can however be quite challenging as there is rarely ground truth available.

The original Retrospective Registration Evaluation Project (RREP) and the more recent Retrospective Image Registration Evaluation (RIRE)[7] are resources for researchers wishing to evaluate and compare techniques for CT-MR and PET-MR registration. The "Vanderbilt Database" is made freely available for participants. Although the "truth" transforms remain sequestered, participants can choose to submit their results on-line, enabling them to compare the performance of their algorithms to those from other groups and techniques.

The Non-rigid Image Registration Evaluation Project (NIREP)[8] is an effort to "develop, establish, maintain and endorse a standardized set of relevant benchmarks and metrics for performance evaluation of nonrigid image registration algorithms". The organizers are planning to create a framework to

[6] `http://trec.nist.gov/trec_eval/`
[7] `http://www.insight-journal.org/RIRE/index.php`
[8] `http://www.nirep.org/index.php?id=22`

evaluate registration that does not require ground truth by utilizing a diverse set of metrics instead. The database consists of 16 annotated MR images from eight normal adult males and eight8 females acquired at the University of Iowa. The metrics that are currently implemented include: squared intensity error, relative overlap, inverse consistency error and transitivity error.

22.4.2 Segmentation

MICCAI Grand Challenges are the most prominent of the evaluation events for segmentation. In 2007, a Grand Challenge workshop was held in conjunction with MICCAI to provide a forum for researchers to evaluate their segmentation algorithms on two anatomical sites, liver and caudate, using a common data sets and metrics. This popular workshop has continued to grow with three and four different sub-tasks in 2008 and 2009, respectively.

MICCAI Segmentation in the Clinic: A Grand Challenge

The liver is a challenging organ for CT-based segmentation as it lies near other organs that are of similar density. Additionally, in the case of diseases there can be significant non-homogeneity within the liver itself, adding to the challenge. The MICCAI Grand Challenge Workshop was one of the most prominent efforts to provide an opportunity for participants to compare the performance of different approaches to the task of liver segmentation. Twenty studies were provided as training data, while ten studies were used for the testing and an additional ten were used for the on-site portion of the evaluation. Participants were allowed to submit results from both completely automated techniques as well as interactive methods.

The training data in the caudate part (33 data sets) were acquired from two different sites using different protocols: 18 healthy controls from the Internet Brain Segmentations Repository (IBSR)[9] from Massachusetts General Hospital and 15 studies consisting of healthy and pathological subjects from Psychiatry Neuroimaging Laboratory at the Brigham and Women's Hospital, Boston. The test data were studies from a challenging mix of ages (adult, pediatric, elderly), sites (Brigham and Women's Hospital, Boston, UNC's Parkinson research group, University of North Carolina at Chapel Hill, Duke Image Analysis Laboratory) and acquired along different axes (axial, coronal) The gold standard was established by manual segmentation of experts.

The organizers were interested in establishing a single score that combined many of the commonly used metrics for segmentation described above. They included volumetric overlap error (or Jaccard coefficient), the relative volume difference, average surface symmetric distance, root mean square surface distance and the maximum symmetric surface distance. This common score was provided for both the liver and the caudate cases. In addition, the caudate

[9] `http://www.cma.mgh.harvard.edu/ibsr/`

Table 22.2. MICCAI Grand Challenges. This is a suggestion of how to summarize the essential information from the MICCAI Grand Challenges ???

Year	Topic	Data Type	Training	Test	Ground Truth
2007	liver	MRI	20	10	manual
	caudate	MRI	33		manual
2008	lumen line	CTA	32 studies à 4 vessels		manual, 3 raters each
	MS lesion	MRI	20	25	manual
	liver tumor				
2009	prostate	MRI			
	head and neck	CT			
	left ventricle				

evaluation consisted of a test of reproducibility by providing a set of scans for the same subject on different scanners. The variability of the score across these scans was evaluated. The Pearson correlation coefficient between the reference and the segmentation volumes was another metric provided for the caudate set.

The organizers have continued to make available all the test and training data, enabling new algorithms to be evaluated against the benchmarks established in 2007. Furthermore, the success of the Grand Challenge in 2007 lead to the continuation of this endeavor in 2008 and 2009 with more clinically-relevant segmentation tasks [28, 29], including coronary artery central lumen line extraction in CT angiography (CTA), Multiple Sclerosis (MS) lesions, and others (Tab. ref).

Extraction of Airways from CT

The Extraction of Airways from CT (EXACT)[10] challenge was held as part of the Second International Workshop on Pulmonary Image Analysis in junction with MICCAI 2009. It provides participants with a set of 20 training CTs that had been acquired at different sites using a variety of equipment, protocols, and reconstruction parameters. Participants were to provide results of algorithms for airway extraction on the 20 test sets. The results were evaluated using the branch count, branch detection, tree length, tree length detected, leakage count, leakage volume and false positive rate. Fifteen teams participated in this task. The organizers noted that "there appears to be a trade off between sensitivity and specificity in the airway tree extraction" as "more complete trees are usually accompanied by a larger percentage of false positives". They also noted that the semi-automatic methods did not significantly outperform the automatic methods.

Volume Change Analysis of Nodules

Again performed in junction with MICCAI as part of the Second International Workshop on Pulmonary Image Analysis, the goal for the Volume

[10] `http://image.diku.dk/exact/information.php`

Change Analysis of Nodules (VOLCANO)[11] challenge was to measure volumetric changes in lung lesions longitudinally using two time-separated image series. This was motivated by the notion that measuring volumetric changes in lung lesions can be useful as they can be good indicators of malignancy and good predictors of response to therapy.

The images were part of the Public Lung Database provided by the Weill Cornell Medical College. 53 nodules were available such that the nodule was visible on at least three slices on both scans. These nodules were classified into three categories: 27 nodules ranging in diameter from 4-24mm visible on two 1.25 mm slice scans with little observed size change, 13 nodules ranging in size from approximately 8 to 30 mm, imaged using different scan slice thicknesses to evaluate the effect of slice thickness and 9 nodules ranging from 5-14 mm on two 1.25 mm scans exhibiting a large size change. The participants were provided with information to locate the nodule pairs. The participants were to submit the volumetric change in nodule size for each volume pair, defined as $\frac{(V_2-V_1)}{V_1}$ where V_1 and V_2 are the volumes of the nodule on the initial and subsequent scan.

22.4.3 Annotation, Classification and Detection

ImageCLEF IRMA

The automatic annotation task at ImageCLEFmed ran from 2005 until 2009 [30]. The goal in this task was to automatically classify radiographs using the Image Retrieval in Medical Applications (IRMA) code along for dimensions: acquisition modality, body orientation, body region, and biological system. The IRMA code is a hierarchical code that can classify radiographs to varying levels of specificity. In 2005, the goal was flat classification in to 57 classes while in 2006 the goal was again a flat classification into 116 unique classes. Error rates based on the number of misclassified images was used as the evaluation metric. In 2007 and 2008, the hierarchical IRMA code was used where errors were penalized depending on the level of the hierarchy at which they occurred. Typically, participants were provided 10,000-12,000 training images and were to submit classification for 1000 test images. In 2009[12] , the goal was to classify 2000 test images using the different classification schemes used in 2005-2008, given a set of about 12,000 training images.

Automatic Nodule Detection

Lung cancer is a deadly cancer, often diagnosed based on lung CT's. Algorithms for the automated Computer Aided Detection (CAD) for lung nodules are a popular area of research. The goal for the Automatic Nodule Detection (ANODE)[13] challenge in 2009 was the automated detection of lung nodules

[11] `http://www.via.cornell.edu/challenge/details/index.html`
[12] `http://www.imageclef.org/2009/medanno/`
[13] `http://anode09.isi.uu.nl/`

based on CT scans. The database consisted of 55 studies. Of these, five were annotated by expert radiologists and were used for training. Two raters (one expert and one trainee) reviewed all the scans, and a third rater was used to resolve disputes. The evaluation was based on a hit rate metric using the 2000 most suspicious hits. The results were obtained using Free-Response Receiver Operating Characteristic (FROC) curves.

Another effort towards the detection of lung nodules in the Lung Imaging Database Consortium (LIDC). The LIDC initiative provides a database of annotated lung CT images, where each image is annotated by four clinicians. This publicly available database enables researchers to compare the output of various Computer Aided Detection (CAD) algorithms with the manual annotations.

22.4.4 Information Retrieval

In information retrieval, evaluation campaigns began nearly fifty ago with the Cranfield tests [23]. These experiments defined the necessity for a document collection, query tasks and ground truth for evaluation, and set the stage for much of what was to follow. The SMART experiments [24] then further systematized evaluation in the domain. The role model for most current evaluation campaign is clearly TREC [25], a series of conferences that started in 1992 and has ever since organized a variety of evaluation campaigns in diverse areas of information retrieval. A benchmark for multilingual information retrieval is CLEF [31], which started within TREC and has been an independent workshop since 2000, attracting over 200 participants in 2009. In addition to its other components, CLEF includes an image retrieval track (called ImageCLEF) which features a medical image retrieval task [32].

22.4.5 Image Retrieval

Image retrieval is a burgeoning area of research in medical informatics [33]. Effective image annotation and retrieval can be useful in the clinical care of patients, education and research. Many areas of medicine, such as radiology, dermatology, and pathology are visually-oriented, yet surprisingly little research has been done investigating how clinicians use and find images [6]. In particular, medical image retrieval techniques and systems are underdeveloped in medicine when compared with their textual cousins [34].

ImageCLEF[14], first began in 2003 as a response to the need for standardized test collections and evaluation forums and has grown to become today a pre-eminent venue for image retrieval evaluation. ImageCLEF itself also includes several sub-tracks concerned with various aspects of image retrieval; one of these tracks is the medical retrieval task. This medical retrieval task was first run in 2004, and has been repeated each year since.

[14] `http://www.imageclef.org/`

The medical image retrieval track's test collection began with a teaching database of 8,000 images. For the first several years, the ImageCLEF medical retrieval test collection was an amalgamation of several teaching case files in English, French, and German. By 2007, it had grown to a collection of over 66,000 images from several teaching collections, as well as a set of topics that were known to be well-suited for textual, visual or mixed retrieval methods.

In 2008, images from the medical literature were used for the first time, moving the task one step closer towards applications that could be of interest in clinical scenarios. Both in 2008 and 2009, the Radiological Society of North America (RSNA) made a subset of its journals' image collections available for use by participants in the ImageCLEF campaign. The 2009 database contained a total of 74,902 images, the largest collection yet. All images were taken from the journals Radiology and Radiographics, both published by the RSNA. The ImageCLEF collection is similar in composition to that powering the Goldminer[15] search system. This collection constitutes an important body of medical knowledge from the peer-reviewed scientific literature, and includes high quality images with textual annotations.

Images are associated with specific published journal articles, and as such may represent either an entire figure or a component of a larger figure. In either event, the image annotations in the collection contain the appropriate caption text. These high-quality annotations enable textual searching in addition to content-based retrieval using the image's visual features. Furthermore, as the PubMed IDs of each image's article are also part of the collection, participants may access bibliographic metadata such as the Medical Subject Headings (MeSH) terms created by the National Library of Medicine for PubMed.

A major goal of ImageCLEF has been to foster development and growth of multi-modal retrieval techniques: i.e., retrieval techniques that combine visual, textual, and other methods to improve retrieval performance. Traditionally, image retrieval systems have been primarily text-based, relying on the textual annotations or captions associated with images [35]. Several commercial systems, such as Google Images[16] and Yahoo! images[17], employ this approach. Although text-based information retrieval methods are mature and well-researched, they are limited by the quality of the annotations applied to the images. There are other important limitations facing traditional text retrieval techniques when applied to image annotations:

- image annotations are subjective and context sensitive, and can be quite limited in scope or even completely absent;
- manually annotating images is labor- and time-intensive, and can be very error prone;
- image annotations are very noisy if they are automatically extracted from the surrounding text; and

[15] `http://goldmier.arrs.org/`
[16] `http://images.google.com/`
[17] `http://images.yahoo.com/`

- there is far more information in an image than can be abstracted using a limited number of words.

Advances in techniques in computer vision have led to a second family of methods for image retrieval: Content-Based Image Retrieval (CBIR). In a CBIR system, the visual contents of the image itself are mathematically abstracted and compared to similar abstractions of all images in the database. These visual features often include the color, shape or texture of images. Typically, such systems present the user with an ordered list of images that are visually most similar to the sample (or query) image.

However, purely visual methods have been shown to have limitations and typically suffer from poor performance for many clinical tasks [36]. On the other hand, combining text- and image-based methods has shown promising results [37]

Several user studies have been performed to study the image searching behavior of clinicians [6, 38]. These studies have been used to inform the development of the tasks over the years, particularly to help ImageCLEF's organizers identify realistic search topics.

The goal in creating search topics for the ImageCLEF medical retrieval task has been to identify typical information needs for a variety of users. In the past, we have used search logs from different medical websites to identify topics [39, 40]. The starting point for the 2009 topics was a user study conducted at Oregon Health & Science University (OHSU) during early 2009. This study was conducted with 37 medical practitioners in order to understand their needs, both met and unmet, regarding medical image retrieval. During the study, participants were given the opportunity to use a variety of medical and general-purpose image retrieval systems, and were asked to report their search queries. In total, the 37 participants used the demonstrated systems to perform a total of 95 searches using textual queries in English. We randomly selected 25 candidate queries from the 95 searches to create the topics for ImageCLEFmed 2009. We added to each candidate query 2 to 4 sample images from the previous collections of ImageCLEFmed, which represented visual queries for content-based retrieval. Additionally, we provided French and German translations of the original textual description for each topic to allow for an evaluation of multilingual retrieval.

Finally, the resulting set of topics was categorized into three groups: 10 visual topics, 10 mixed topics, and 5 semantic topics. This classification was performed by the organizers based on their knowledge of the capabilities of visual and textual search techniques, prior experience with the performance of textual and visual systems at ImageCLEF medical retrieval task, and their familiarity with the test collection. The entire set of topics was finally approved by a physician. An example of a visual topic can be seen in Fig. 22.2 while that of a textual topic is shown in Fig. 22.2.

In 2009, we also introduced case-based topics [4] as part of an exploratory task whose goal was to generate search topics that are potentially more aligned

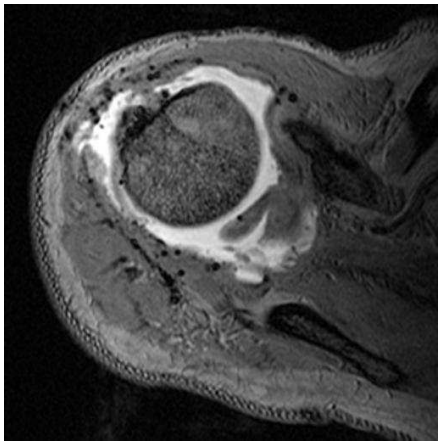
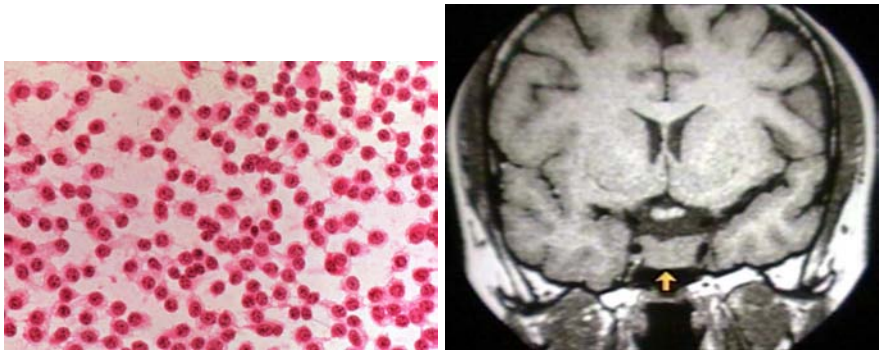


Fig. 22.2. A sample image of a visual retrieval task. MR images of a rotator cuff

with the information needs of an actual clinician in practice. These topics were meant to simulate the use case of a clinician who is diagnosing a difficult case, and has information about the patient's demographics, list of present symptoms, and imaging studies, but not the patient's final diagnosis. Providing this clinician with articles from the literature that deal with cases similar to the case (s)he is working on (similar based on images and other clinical data on the patient) could be a valuable aide to creating differential diagnosis or identifying treatment options, for example with case-based reasoning [41]. These case-based search topics were created based on cases from the teaching file Casimage, which contains cases (including images) from radiological practice. Ten cases were pre-selected, and a search with the final diagnosis was performed against the 2009 ImageCLEF data set to make sure that there were at least a few matching articles. Five topics were finally chosen. The



(a) (b)

Fig. 22.3. A sample image of a textual retrieval task. Images of pituitary ademoma

diagnoses and all information about the chosen treatment were removed from the cases to simulate the aforementioned situation of a clinician dealing with a difficult diagnosis. However, in order to make the judging more consistent, the relevance judges were provided with the original diagnosis for each case.

During 2008 and 2009, relevance judgments were made by a panel of clinicians using a web-based interface. Due to the unfeasibility of manually reviewing 74,900 images for 30 topics, the organizers used a TREC-style pooling system to reduce the number of candidate images for each topic to approximately 1,000 by combining the top 40 images from each of the participants' runs. Each judge was responsible for between three to five topics, and sixteen of the thirty topics were judged multiple times (in order to allow evaluation of inter-rater agreement). For the image-based topics, each judge was presented with the topic as well as several sample images.

For the case-based topics, the judge was shown the original case description and several images appearing in the original article's text. Besides a short description for the judgments, a full document was prepared to describe the judging process, including what should be regarded as relevant versus non-relevant. A ternary judgment scheme was used, wherein each image in each pool was judged to be "relevant", "partly relevant", or "non-relevant". Images clearly corresponding to all criteria were judged as "relevant", images whose relevance could not be safely confirmed but could still be possible were marked as "partly relevant", and images for which one or more criteria of the topic were not met were marked as "non-relevant". Judges were instructed in these criteria and results were manually verified during the judgment process.

As mentioned, we had a sufficient number of judges to perform multiple judgements on many topics, both image-based and case-based. Inter-rater agreement was assessed using the kappa metric, given as:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}, \tag{22.16}$$

where $Pr(e)$ is the observed agreement between judges, and $Pr(a)$ the expected (random) agreement. It is generally accepted that a κ ¿ 0.7 is good and sufficient for an evaluation. The score is calculated using a 2x2 table for the relevances of images or articles. These were calculated using both lenient and strict judgment rules. Under the lenient rules, a partly relevant judgment was counted as relevant; under strict rules, partly relevant judgments were considered to be non-relevant. In general, the agreement between the judges was fairly high (with a few exceptions), and our 2009 overall average κ is similar to that found during other evaluation campaigns.

22.5 Lessons Learned

Conducting the ImageCLEF campaigns has been a great learning opportunity for the organizers. Most evaluation campaigns are run by volunteers with

meager resources. However, a surprising number of researchers willingly donate their data, time and expertise towards these efforts as they truly believe that progress in the field can only come as a result of these endeavors.

Participants have been quite loyal for the ImageCLEFmed challenge, an annual challenge that has been running since 2004. Many groups have participated for four or more years although each year sees newcomers, a welcome addition. A large proportion of participants are actually PhD students who obtain valuable data to validate their approaches. The participants have been quite cooperative, both at the workshops and during the year. They have provided baseline runs or allowed their runs to be used by other in collaborative efforts. Many of the new organizers were participants, thus ensuring a steady stream of new volunteers willing to carry on the mantle of those that have move away. By comparing the relative performance of a baseline run through the years, we have seen the significant advances being made in the field.

22.6 Conclusions

Evaluation is an important facet of the process of developing algorithms for medical image analysis including for segmentation, registration and retrieval. In order to be able to measure improvements resulting from new research in computer vision, image processing and machine learning when applied to medical imaging tasks, it is important to have established benchmarks against which their performance can be compared. Computer scientists are making huge strides in computer vision, image processing and machine learning, and clinicians and hospitals are creating vast quantities of images each day. However, it can still be quite difficult for the researchers developing the algorithms to have access to high quality, well curated data and ground truth. Similarly, it can also be quite difficult for clinicians to get access to state-of-the-art algorithms that might be helpful in improving their efficiency, easing their workflow and reducing variability.

Evaluation campaigns have provided a forum to bridge this gap by providing large, realistic and well annotated datasets, ground truth, meaningful metrics geared specifically for the clinical task, organizational resources including informational websites and software for evaluation and often workshops for researchers to present their results and have discussions. Examples of successful evaluation campaigns include ImageCLEFmed for medical image retrieval and annotation, the VOLCANO challenge to assess volumetric changes in lung nodules, the EXACT airway extraction challenge and the popular set of MICCAI segmentation grand challenges. Other efforts to provide publicly accessible data and ground truth include the LIDC set of images for the detection of chest nodules based on CTs, the CT and PET images from the ADNI initiative, and the RIRE and NIREP efforts to evaluate registration. Many of these efforts are continuing beyond the workshops by still enabling partic-

ipants to download data, submit results, evaluating and posting the results, thereby providing venues for the progress in the field to be documented.

References

1. Müller H, Müller W, Squire D. Performance evaluation in content–based image retrieval: overview and proposals. Pattern Recognit Lett. 2001;22(5):593–601.
2. Price K. Anything you can do, i can do better (no you can't). Computer Vis. 1986;.
3. Heimann T, Styner M, Ginneken B. 3D segmentation in the clinic: a grand challenge. Miccai. 2007;.
4. Müller H, Kalpathy-Cramer J, Eggers I. Overview of the CLEF 2009 medical image retrieval track. In: Working Notes of CLEF 2009 Corfu, Greece. 2009;.
5. Markkula M, Sormunen E. Searching for photos – journalists' practices in pictorial IR. The Challenge of Image Retrieval, A Workshop and Symposium on Image Retrieval Electronic Workshops in Computing Newcastle upon Tyne: The British Computer Society. 1998;.
6. Müller H, Despont-Gros C, Hersh W. Health care professionals' image use and search behaviour. In: Proceedings of the Medical Informatics Europe Conference (MIE 2006) IOS Press, Studies in Health Technology and Informatics Maastricht. 2006; p. 24–32.
7. Hersh W, Müller H, Gorman P. Task analysis for evaluating image retrieval systems in the ImageCLEF biomedical image retrieval task. In: Slice of Life conference on Multimedia in Medical Education (SOL 2005). 2005;.
8. Deselaers T, Deserno T, H M. Automatic medical image annotation in ImageCLEF 2007: Overview, results, and discussion. Pattern Recognit Lett. 2008;29(15):1988–95.
9. Davatzikos C, Xu F, An Y. Longitudinal progression of alzheimer's-like patterns of atrophy in normal older .... 2009;.
10. Armato S, McNitt-Gray M, Reeves A. The lung image database consortium (LIDC): An evaluation of radiologist variability in the identification of lung nodules on CT scans. Acad Radiol. 2007;14(11):1409–21.
11. Mueller S, Weiner M, Thal L. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (ADNI). Alzheimer's & dementia : the journal of the Alzheimer's Association. 2005;1(1):55–66.
12. Rubin G, Lyo J, Paik D. Pulmonary nodules on multi–detector row CT scans: performance comparison of radiologists and computer–aided detection radiology. 2005. 234;1(274).
13. Christensen G, Geng X, Kuhl J. Introduction to the non-rigid image registration evaluation project (NIREP). Lect Notes Computer Sci. 2006;.
14. Warfield S, Zou K, Wells W. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the .... IEEE Trans Med Imaging. 2004;.
15. Babalola K, Patenaude B, Aljabar P. Comparison and evaluation of segmentation techniques for subcortical structures in brain .... Proceedings of the 11th international conference on .... 2008;.
16. Dice L. Measures of the amount of ecologic association between species ecology. 1945;.

17. Jaccard P. The distribution of the flora in the alpine zone. New Phytologist. 1912;.
18. Zijdenbos A, Dawant B, Margolin R. Morphometric analysis of white matter lesions in MR images: method and validation. Med Imaging. 1994;13(4):716–24.
19. Williams G. Comparing the joint agreement of several raters with another rater. Biometrics. 1976;.
20. Martin-Fernandez M, Bouix S, Ungar L. Two methods for validating brain tissue classifiers. Lect Notes Computer Sci. 2005;.
21. Crum W, Camara O, Hill D. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans Med Imaging. 2006;25(11):1451–61.
22. Gerig G, Jomier M, Chakos A. Valmet: A new validation tool for assessing and improving 3D object segmentation. MICCAI 2001: Fourth International Conference on .... 2001;.
23. Cleverdon C. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Cranfield, USA: Aslib Cranfield Research Project. 1962;.
24. Salton G. The SMART Retrieval System, Experiments in Automatic Document Processing. Englewood Cliffs, New Jersey, USA: Prentice Hall. 1971;.
25. Voorhees E, Harmann D. Overview of the seventh text retrieval conference (trec–7). In: The Seventh Text Retrieval Conference Gaithersburg, MD, USA. 199; p. 1–23.
26. van Rijsbergen C. Evaluation. In: Information Retrieval Englewood Cliffs, New Jersey, USA. 1979; p. 112–23.
27. Voorhees E. Variations in relevance judgements and the measurement of retrieval effectiveness. Inf Process Manag. 2000;36:697–716.
28. Styner M, Lee J, Chin B. 3D segmentation in the clinic: a grand challenge ii: ms lesion. Segmentation. 2008;.
29. Hameeteman. Carotid lumen segmentation and stenosis grading challenge. The MIDAS Journal. 2009; p. 1–15.
30. Tommasi T, Caputo B, Welter P. Overview of the CLEF 2009 medical image annotation track;.
31. Savoy J. Report on CLEF–2001 experiments. In: Report on the CLEF Conference 2001 (Cross Language Evaluation Forum) Springer LNCS 2406. 2002; p. 27–43.
32. Müller H, Deselaers T, Kim E. Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: CLEF 2007 Proceedings vol 5152 of Lecture Notes in Computer Science (LNCS). 2008; p. 473–91.
33. Müller H, Michoux N, Bandon D. A review of content-based image retrieval systems in medicine–clinical benefits and future directions. Int J Med Inform. 2004;73(1):1–23.
34. Hersh W. Information retrieval — a health and biomedical perspective. Springer. 2003;.
35. Enser P. Pictorial information retrieval. J Doc. 1995;51(2):126–70.
36. Müller H, Kalpathy-Cramer J, Kahn Jr C. Overview of the ImageCLEFmed 2008 medical image retrieval task. 9th Workshop of the Cross-Language Evaluation Forum vol 5706 of Lecture Notes in Computer Science Aarhus. 2009; p. 500–10.

37. Kalpathy-Cramer J, Bedrick S, Hatt W. Multimodal medical image retrieval: OHSU at ImageCLEF 2008. Proceedings of Evaluating Systems for Multilingual and Multimodal Information Access–9th Workshop of the Cross-Language Evaluation Forum CLEF. 2008;.
38. Hersh W, Jensen J, Müller H. A qualitative task analysis for developing an image retrieval test collection. In: ImageCLEF/MUSCLE workshop on image retrieval evaluation. 2005; p. 11–6.
39. Müller H, Boyer C, Gaudinat A. Analyzing web log files of the health on the net honmedia search engine to define typical image search tasks for image retrieval evaluation. In: MedInfo 2007 vol 12 of IOS press, Studies in Health Technology and Informatics Brisbane, Australia. 2007; p. 1319–23.
40. Müller H, Kalpathy-Cramer J, Hersh W. Using medline queries to generate image retrieval tasks for benchmarking. In: Medical Informatics Europe (MIE2008). 2008; p. 523–8.
41. Aamodt A, Plaza E. Case–based reasoning: foundational issues, methodological variations, and systems approaches. Artiff Intell Commun. 1994;7(1):39–59.

List of Acronyms

ADNI	Alzheimer's Disease Neuroimaging Initiative
ANODE	Automatic Nodule Detection
AVD	Absolute Volumetric Difference
CAD	Computer Aided Detection
CBIR	Content-Based Image Retrieval
CICE	Cumulative Inverse Consistency Error
CLEF	Cross Language Evaluation Forum
CSF	Cereborspinal Fluid
CTA	CT angiography
CTE	Cumulative Transitive Error
EXACT	Extraction of Airways from CT
FND	False Negative Dice
FPD	False Positive Dice
FROC	Free-Response Receiver Operating Characteristic
GTC	Generalized Tanimoto Coefficients
HD	Hausdorff Distance
IBSR	Internet Brain Segmentations Repository
IR	Information Retrieval
IRMA	Image Retrieval in Medical Applications
LIDC	Lung Imaging Database Consortium,
MAP	Mean Average Precision
MeSH	Medical Subject Headings
MICCAI	Medical Image Computing and Computer Assisted Intervention
MRI	Magnetic Resonance Imaging
MS	Multiple Sclerosis
NIH	National Institutes of Health
NIREP	Non-rigid Image Registration Evaluation Project
RIRE	Retrospective Image Registration Evaluation
ROI	Regions of Interest
RREP	Retrospective Registration Evaluation Project

RSNA	Radiological Society of North America
SMART	System for the Mechanical Analysis and Retrieval of Text
STAPLE	Simultaneous Truth and Performance Level Estimation
TREC	Text Retrieval Conference
VD	Volumetric Difference
VOLCANO	Volume Change Analysis of Nodules

Index