Abstract ID: 15708 Title: Evaluation of maximum likelihood ground truth and performance of readers stratified by aggressiveness from the Lung Image Database Consortium (LIDC) study

Innovation

The novelty of this work is the application of expectation-maximization and maximum-likelihood techniques, which are well known to the biostatistics community, to the unique task of determining a reasonable and objective ground-truth for nodule detection. While this work builds upon the important paper by Warfield, Zou and Wells, describing the STAPLE method, it represents a distinct change in implementation to address the immediate problem of deducing an optimal nodule **detection** ground-truth from multiple readers' assessments in the absence of the known ground-truth. The task addressed by Warfield *et al*. required an implementation of STAPLE specifically for the nodule **segmentation** task. Ross *et al*. have since applied STAPLE in the **segmentation** task to an early LIDC subset, consisting of 41 scans.

The unique utility of this work is that it is applied in this paper to the large LIDC outcomes data for which no previous analysis has been successful in extracting a viable detection ground-truth. Indeed, although portions of the LIDC database have become available publically since 2006, there is yet no full publication (although several IEEE SPIE Medical Imaging conference abstracts have been produced) applying CAD detection to the LIDC data, largely because of the inability to overcome the ground-truth problem. No prior publication deals with the task of extracting an objective ground-truth for detection using STAPLE from the LIDC or any other tumor image database. The LIDC investigators have published several papers looking at the variability in detection and segmentation performance amongst the readers of the study, but none have ventured to propose an optimal detection ground-truth from their data.

Magnitude of the ground-truth problem for the LIDC

These differing reader interpretations complicate the task of using the LIDC datasets for defining the ground truth for subsequent training and evaluation of CAD performance. Ochs *et al*. detailed the differences in 'ground-truth' when using different LIDC subsets based on the number of consensing readers and/or nodule size[1]. In a more detailed analysis on an LIDC subset of 25 scans, Armato *et al*. found that, for any one expert reader amongst the four in their ground-truth panel, detection sensitivity ranged from 51-83%, depending on which radiologist and which ground-truth metric was used[2]. They therefore concluded that, "even experienced thoracic radiologists may not perform well when measured against the 'truth' established by other experienced thoracic radiologists."[2] Notably, it is conceivable that one of the LIDC experts achieved 100% sensitivity and specificity, yet his or her sensitivity performance would have been assessed at only 51-83% by the subjective ground-truth. By extension, applying this LIDC ground-truth to train the next generation of radiologists, or the latest CAD algorithm, will inevitably lead both humans and computers to misidentify 17-49% of nodules. Due in large part to the problem in establishing an LIDC ground-truth, only a few detection studies have been published involving the LIDC data[3]. The majority limited their focus to nodules > 3 mm[4] and many further limited to those large nodules identified as such by all four LIDC readers[5,6].

EM-ML formulation

The EM-ML algorithm applied in this paper follows closely that of STAPLE, save for interpretation of ground-truth and reader performance evaluated on a per-nodule rather than per-pixel basis. Only a cursory description is provide here and the reader is referred to the original paper by Warfield[7]. For $p_r^{(k)}$ representing the percent of true-positive detections (sensitivity performance) for reader *r* estimated at iteration *k*; $q_r^{(k)}$ representing analogously the reader's percent of true-negative detections (specificity performance); and D_{rn} the decision by reader *r* of whether nodule *n* is a true nodule, the equation for the estimate of the conditional probability of the true segmentation, $W_n^{(k)}$, for each nodule, *n*, is constructed through the intermediate parameters, $a_n^{(k)}$ and $b_n^{(k)}$ as:

$$a_n^{(k)} = W_n^{(k-1)} \prod_{r:D_{rn}=1} p_r^{(k)} \prod_{r:D_{rn}=0} \left(1-p_r^{(k)}\right)$$

$$b_n^{(k)} = \left(1-W_n^{(k-1)}\right) \prod_{r:D_{rn}=0} q_r^{(k)} \prod_{r:D_{rn}=1} \left(1-q_r^{(k)}\right) \qquad (1)$$

$$W_n^{(k)} = \frac{a_n^{(k)}}{a_n^{(k)}+b_n^{(k)}}$$

Here $W_n^{(k)}$ can be interpreted as the weight, or probability, of the nodule *n* being true, at iteration *k*, with a value ranging from 0.0 to 1.0. Given $W_n^{(k)}$, it is straightforward to compute expressions for the reader performance metrics, $p_r^{(k)}$ and $q_r^{(k)}$ that are derived from the definitions of sensitivity and specificity given earlier:

$$p_r^{(k)} = \frac{\sum_{n:D_{rn}=1} W_n^{(k-1)}}{\sum_n W_n^{(k-1)}}$$

$$q_r^{(k)} = \frac{\sum_{n:D_{rn}=0} \left(1 - W_n^{(k-1)}\right)}{\sum_n \left(1 - W_n^{(k-1)}\right)} \qquad (2)$$

Table 1: Number of nodule detections and optimized performance metrics for each reader, where each reader is ranked by aggressiveness in identifying nodules. The same performance metrics were obtained for a variety of initial ground-truth estimates, and nodule weighting schemes, and for initial reader performance values (p and q) of 0.9.

Ranked Reader	1	2	3	4
# of detections	2505	2119	1712	1272
Sensitivity (p)	0.9840	0.9668	0.9183	0.7179
Specificity (q)	0.2227	0.5305	0.8126	0.9106

Literature Cited

1. Ochs R, Kim HJ, Angel E, et al. Forming a reference standard from LIDC data: impact of reader agreement on reported CAD performance. In: *Proceedings of SPIE*. San Diego, CA, USA; 2007.
2. Armato SG, Roberts RY, Kocherginsky M, et al. Assessment of radiologist performance in the detection of lung nodules: dependence on the definition of "truth". *Acad Radiol*. 2009;16(1):28-38.
3. Ozekes S, Osman O. Computerized lung nodule detection using 3D feature extraction and learning based algorithms. *J Med Syst*. 2010;34(2):185-194.
4. Messay T, Hardie RC, Rogers SK. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. *Med Image Anal*. 2010;14(3):390-406.
5. Golosio B, Masala GL, Piccioli A, et al. A novel multithreshold method for nodule detection in lung CT. *Med Phys*. 2009;36(8):3607-3618.
6. Opfer R, Wiemker R. Performance Analysis for Computer Aided Lung Nodule Detection on LIDC Data. In: *Proceedings of SPIE*. San Diego, CA, USA; 2007.
7. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*. 2004;23(7):903-21.