A shape-dependent variability metric for evaluating panel segmentations with a case study on LIDC data

Stephen Siena[a], Olga Zinoveva[b], Daniela Raicu[c], Jacob Furst[c], and Samuel Armato III[d]

[a]University of Notre Dame, Notre Dame, IN, 46556;
[b]Harvard University, Cambridge, MA, 02138;
[c]DePaul University, Chicago, IL, 60604;
[d]University of Chicago, Chicago, IL, 60637

ABSTRACT

The segmentation of medical images is challenging because a ground truth is often not available. Computer-Aided Detection (CAD) systems are dependent on ground truth as a means of comparison; however, in many cases the ground truth is derived from only experts' opinions. When the experts disagree, it becomes impossible to discern one ground truth. In this paper, we propose an algorithm to measure the disagreement among radiologist's delineated boundaries. The algorithm accounts for both the overlap and shape of the boundaries in determining the variability of a panel segmentation. After calculating the variability of 3788 thoracic computed tomography (CT) slices in the Lung Image Database Consortium (LIDC), we found that the radiologists have a high consensus in a majority of lung nodule segmentations. However, our algorithm identified a number of segmentations that the radiologists significantly disagreed on. Our proposed method of measuring disagreement can assist others in determining the reliability of panel segmentations. We also demonstrate that it is superior to simply using overlap, which is currently one of the most common ways of measuring segmentation agreement. The variability metric presented has applications to panel segmentations, and also has potential uses in CAD systems.

Keywords: LIDC, variability, panel segmentation, lung nodule, reference truth

1. INTRODUCTION

Lung cancer is the leading cause of cancer-related deaths in the world. Quality computer-aided segmentations of lung nodules help computer-aided classification and diagnosis systems remove noise from images, increasing their effectiveness. Currently two problems make the segmentation of lung nodules difficult. The first problem involves the variety of lung nodules. While some nodules are smooth and well-defined, others can be non-solid, or next to lung vessels and other healthy structures. Distinguishing between healthy tissue and the nodule is a challenge that even radiologists sometimes struggle. This challenge leads to the second problem: the absence of a satisfactory ground truth.[1]

The Lung Image Database Consortium (LIDC) contains series of computed tomography (CT) scans with contours provided by 4 radiologists.[2] In many cases, the differences in segmentations may be trivial, but in other cases, the radiologist outlines show significant disagreement. Because of this, different research groups use different combinations of the panel segmentation to formulate a ground truth, so results are not easily comparable.

There has been some work on radiologist variability in image segmentation, and the LIDC data in particular. Armato et al. discuss the variability in nodule classification among the radiologists,[1] as well as how different "truths" obtained from the four radiologists affect the perceived performance of other segmentations.[3,4] Picking a single favorable "truth" can also help superficially improve results.[5] In addition, there are algorithms that attempt to reconcile the differences between segmentations to get one "ground truth."[6,7] The problem with the algorithms is that once the "ground truth" is established, the differences in the segmentations from which the ground truth was derived are lost. Presently there is no suitable metric for the variability of contours on a single

Further author information:
S.A.S.: E-mail: ssiena@nd.edu, Telephone: 1.203.305.7606

Medical Imaging 2010: Computer-Aided Diagnosis, edited by Nico Karssemeijer, Ronald M. Summers,
Proc. of SPIE Vol. 7624, 762416 · © 2010 SPIE · CCC code: 1605-7422/10/$18 · doi: 10.1117/12.844639

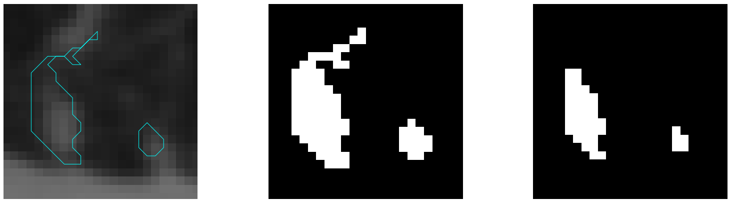


Figure 1. *Left* Outlines provided by a single radiologist for a slice with separate regions for the same nodule. *Center* Area marked if outlines are included. *Right* Area marked if outlines are not included. Note that a large protrusion, a significant feature, is completely lost when the outlines are not included.

slice. Such a metric would provide insight into the consistency of the outlines. Consistency is crucial, because errors in manual or computer segmentations are less critical so long as the mistakes are uniform.[8, 9]

There are many methods for evaluating a segmentation given a ground truth. Some of the common methods include volumetric overlap error and relative volume difference,[10] both of which are 3D measures that can also be extended to 2D segmentations. However, these methods depend on a single segmentation that can be regarded as the ground truth. In the LIDC, there are multiple segmentations that have equal claim to being a ground truth, and there is no intuitive way to extend these metrics to situations like the LIDC. In addition, in situations in which there are more than two segmentations, there is no one way to define overlap.

To alleviate this problem, we propose a measure of the radiologists' variability that is based on the level of disagreement according to the percent area of overlap and the shape of each radiologist's segmentation. This variability measure can be used to measure the consistency of the radiologists' segmentations, thus quantifying the validity of the panel segmentation. The algorithm can be extended to any number of segmentations, so it is more versatile than overlap. Beyond panel segmentations, the variability measure has applications for judging CAD systems.

2. MATERIALS AND METHODS

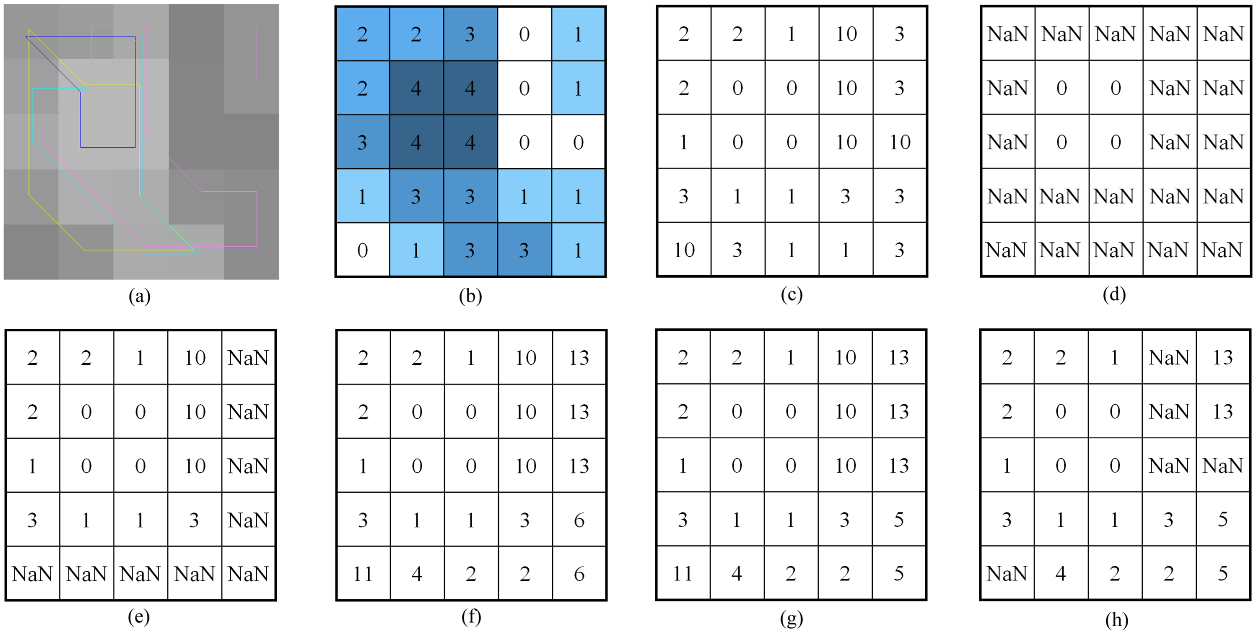
2.1 Materials

The LIDC dataset contains 400 series of CT scans. Of these, 315 series from 313 patients contain a total of 921 distinct nodules. Each series of scans was presented to four radiologists who were directed to outline any nodules they found that were between 3 mm and 30 mm in diameter. The radiologists were instructed to draw their outline so that only pixels contained within the outline, but not the outline itself, were part of the nodule. However, in this work we include the coordinates given in the LIDC annotation files. If not, features of many segmentations are completely or nearly completely lost (see example in Figure 1). The radiologist outlines are the closest approximation to ground truth that is available for the LIDC data.

Although the nodules have varying semantic characteristics, our focus is on the contours provided by the radiologists. We considered slices from the CT scans in which at least two radiologists had segmentations which overlapped. The few exceptions that were not included were situations in which the actual number of nodules was in dispute; in 2 cases the radiologists disagreed whether there were 2 small nodules or 1 large nodule present. The slices were classified independent of other slices in the same nodule by the maximum number of radiologists that included the same pixel. There were 1768, 1000, and 1020 slices with a region selected by all four radiologists, three radiologists, and just two radiologists, respectively. These groups will be referred to as the P_4, P_3, and P_2 classes. It is important to note that total number of unique slices is not 3788; in some cases, a single slice had two distinct nodules and these cases are actually double counted. Overall, there are 3641 unique CT slices. The data for all three classes were from 273 series from 271 patients containing 610 unique nodules.

2.2 Methodology

The proposed algorithm takes into account two factors - the *total proportion* of disputed pixels, and the *distance* of each disputed pixel from the region of the image that has full agreement among all segmentation. The total proportion of disputed pixels roughly correlates to overlap, and can adequately measure gross differences between



(a)

2	2	3	0	1
2	4	4	0	1
3	4	4	0	0
1	3	3	1	1
0	1	3	3	1

(b)

2	2	1	10	3
2	0	0	10	3
1	0	0	10	10
3	1	1	3	3
10	3	1	1	3

(c)

NaN	NaN	NaN	NaN	NaN
NaN	0	0	NaN	NaN
NaN	0	0	NaN	NaN
NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN

(d)

2	2	1	10	NaN
2	0	0	10	NaN
1	0	0	10	NaN
3	1	1	3	NaN
NaN	NaN	NaN	NaN	NaN

(e)

2	2	1	10	13
2	0	0	10	13
1	0	0	10	13
3	1	1	3	6
11	4	2	2	6

(f)

2	2	1	10	13
2	0	0	10	13
1	0	0	10	13
3	1	1	3	5
11	4	2	2	5

(g)

2	2	1	NaN	13
2	0	0	NaN	13
1	0	0	NaN	NaN
3	1	1	3	5
NaN	4	2	2	5

(h)

Figure 2. (**a**) Example of radiologist outlines. (**b**) Probability map. (**c**) Cost map using $R = 4$ and $k = 10$. (**d**) Variability matrix right after initialization. (**e-g**) Variability matrix after 1, 2, and 3 iterations. Note that two values in the bottom right change between the 2nd and 3rd iteration, despite already being set to a value. (**h**) Final variability matrix used in calculation of VI. $VI = 60$; $VI_n = 5.1064$

segmentations. Accounting for the distance improves the variability metric by capturing shape details such as spiculation, which can affect a diagnosis even if the total area it adds to the segmentation is small.

We first construct a probability map (pmap) that assigns each pixel a probability of belonging to the lung nodule by looking at the areas included in each of the contours (Figure 2b). Each value P(r, c) in the pmap equals the number of radiologists that selected the given pixel, where r is the pixel's row and c is the pixel's column.

Two additional matrices are constructed to calculate the variability metric. The first is the cost map C, which contains a cost for each pixel (Figure 2c). The cost varies inversely with P, so that

$$C(r,c) = \begin{cases} (R-1) \times \frac{max(P)-P(r,c)}{max(P)-1} & \text{if } P(r,c) > 0 \\ k & \text{if } P(r,c) = 0 \end{cases} \tag{1}$$

where C(r, c) is the cost of the pixel (r, c) based on its value P in the pmap. This ensures that pixels upon which there is less agreement contribute more to variability than those with higher agreement. The constant R is set to the number of raters including those who did not detect anything at all; in the case of the LIDC, $R = 4$. The value of k is set by user. In most cases, the value of k will not effect calculations. However, in situations where there are two disjoint regions of the same nodule within a slice (Figure 3, fourth from left), it is necessary to have a cost assigned to pixels identified as non-nodule by all radiologists. A larger value for k will penalize these pixels with disjoint regions more than a lower value. Other situations where the value of k will affect calculations is in situations where a long protrusion runs near the main region of the nodule (Figure 3, far right). In this case, the cost of pixels at the very top of the outline is determined by crossing through a region without any markings, with each pixel a cost of k. If the value of k were higher, the cost of those pixels would in actually be cheaper by traversing the longer portion of pixels outlined by the single radiologist than by crossing through the region with no markings, because the cost of those non-nodule pixels would be too high. Having a higher value of k would require the cost of those pixels to be computed by using the actual outlines provided by the radiologists, so in these cases, a higher k mor accurately reflects the variability of the outlines. To balance out the two possible scenarios, for all our calculations we set $k = 10$. We recommend that the absolute minimum value used is $k = R$, although higher values are probably more appropriate.

The second matrix is the variability matrix V. It is initialized with values of 0 for pixels that correspond to P(r, c) = max(P) in the pmap. The rest of the pixels are not assigned a numeric value (NaN). The matrix is then updated iteratively (Figure 2d-g) according to the cost map. For each pixel, the algorithm finds the lowest V as follows:

$$V(r,c) = \begin{cases} v^* + C(r,c) & \text{if } V(r,c) > v^* + C(r,c) \\ V(r,c) & \text{if } V(r,c) \leq v^* + C(r,c) \end{cases} \tag{2}$$

where V is the value of the current pixel (r, c) in the variability matrix, C is its cost from the cost map and $v*$ is the lowest value of the eight pixels surrounding (r, c) in the variability matrix. The matrix converges when the lowest values for all pixels have been found. All pixels in the variability matrix with value P(r, c) = 0 from the pmap are assigned NaN (Figure 2h), and are ignored in subsequent calculations.

We define the variability metric VI as the sum of all values in the variability matrix:

$$VI = \sum V(r,c) \tag{3}$$

However, this variability metric would easily be skewed by the nodule area; a larger nodule would have more pixels, and therefore more opportunities for the radiologists to disagree over individual pixels. We define the normalized variability index, VI_n, to take into account the differences in nodule area by dividing the variability index by the averge segmentation area so that

$$VI_n = \frac{VI}{\frac{\sum P(r,c)}{R}} \tag{4}$$

The average segmentation area of the radiologists includes radiologists that did not give any outline; they contribute an area of 0 to the average. This does not account for radiologists who considered the lesion to be a nodule of ¡3 mm, but since they do not provide outlines, they do not contribute to the variability of the outlines, so they are simply treated as not marking the nodule. Therefore, the average segmentation area corresponds to the sum of all values in the pmap divided by R.

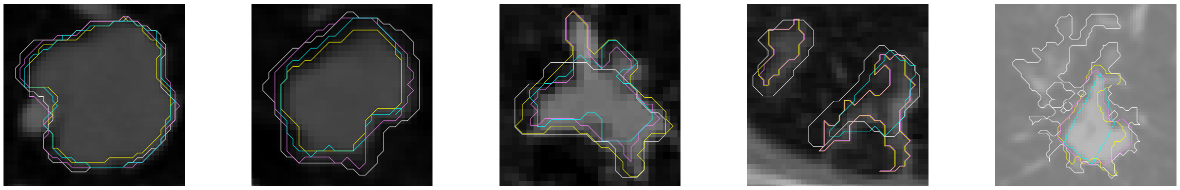


Figure 3. Left to right, pmaps with VI_n of 0.5227, 1.8198, 2,5977, 39.3927 and 42.5135

3. RESULTS

We analyzed the variability of the contours produced for lung nodule segmentation by four different radiologists. The VI_n for all 3788 slices ranged from 0 to 184.6935. Although the range for the four distributions is large, the first quartile (Q1) and third quartile (Q3) give a better representation of the range of values that VI_n produces (see Table 1). All the distributions contain a few outliers which have extremely high values of VI_n.

While the VI_n is intended to measure the variability of panel segmentations, specific values are not meant to have any intrinsic meaning. Rather, high values of VI_n can serve as flags for researchers. Visual inspection shows that relative values are sensible; relative values for nodules with higher and lower VI_ns correspond well to what humans would consider more or less variable panel segmentations.

Table 1. Statistics for VI_n distributions

	Mean	Std. Dev.	Min	Q1	Median	Q3	Max
All Slices	3.5290	6.6422	0	1.3333	2.1131	3.5294	184.6935
P_2	5.0361	9.4441	0	1.8292	2.8725	4.8765	148.9670
P_3	4.0001	7.5591	0.3333	1.5372	2.3846	3.9617	184.6935
P_4	2.3943	2.8509	0.1007	1.1216	1.6388	2.5979	42.5135

One aim of the VI_n measurement is not only to measure the differences in the outlines, but also to account for detection. Therefore, VI_n should have lower overall values for slices in which all four radiologists provided outlines. This holds true between the P_4, P_3, and P_2 distributions (Fig. 4), based on the statistics shown in Table 1. However, depending on what the VI_n is being used for, it may not be completely useful to compare values between different classes. Completely ignoring P_2, because half of the raters did not even detect a nodule, could be the correct approach depending on the application. However, for slices within any class, the relative values have significance.

An oddity in the data are the slices for which the VI_n equaled 0. In some cases, two radiologists shared identical segmentations. In cases where there were more than two radiologists providing an outline, these situations can go by unnoticed, because the third (or fourth) radiologist could have an outline that differed to provide variability. When only two radiologists provided an outline, the two identical segmentations would produce a VI_n of 0, because the only outlines that exist have no difference. This occurs in 21 slices, and does not drastically affect the distribution.

The transformation from VI to VI_n is meant to ensure that there was no bias when considering the size of the nodule. Simply dividing the VI in the manner described was effective enough. The highest correlation between the VI_n for any distribution and nodule size was roughly .12, corresponding to the P_2 class. The correlation for all slices was less than .03, so VI_n effectively negates the nodule size when providing a value for the variability.

Overall, the VI_n values were low for the large majority of slices. P_4 slices are arguably the most important, because CAD systems do not have to worry about the uncertainty of the radiologists' detection, only the exact outlines provided by the radiologists. Most slices in the P_4 distribution had relatively low values, and even 2.5979 (Q3) does not indicate a large amount of disagreement (see Fig. 3, middle). This is not always the case though, and that is where VI_n can be valuable. For some slices, the VI_n is extremely high, which indicates that the panel segmentation is less reliable. Perhaps at some point the VI_n could be so high that a CAD system cannot be expected to match the radiologists results, or there may be less value in trying to simulate the radiologists when they cannot agree very much.

In this way, the VI_n can identify cases which researchers may simply want to consider with caution when testing their system. Alternatively, a user could choose to remove a single radiologist's outline. In the pmap in Fig. 3 on the far right, if the white outline is removed, the VI_n drops from 42.5135 to 4.9132. How exactly VI_n is used is left to the user, but options like this are certainly available.

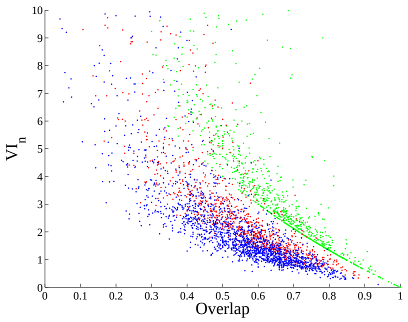


Figure 4. Scatterplot showing relation between overlap and VI_n. Approximately 95% of data is shown ($VI_n < 10$). P_4 - blue, P_3 - red, P_2 - green. Overlap defined as $max(P)$.

It is also important to consider the standard deviation of a distribution of VI_n values. Lower standard deviations would be more desirable, regardless of the average VI_n. If the VI_n was higher but had a very low standard deviation, this would indicate that the variability between slices would be lower. This would be significant, because the amount of disagreement between the radiologists would be predictable. With the standard deviations of each class being higher than the average (both mean or median) in every distribution, each new slice considered is an unknown; it may have a near consensus outline, or the four radiologists may have very different segmentations. If the standard deviation was lower, the VI_n of new slices and nodules with similar characteristics could be reliably predicted. In addition, a lower standard deviation might allow for the possibility that each slice has the same differences between each of the 4 outlines (perhaps 1 radiologist always has a larger outline than the other 3). In such a case, using the intersection of 3 out of 4 radiologists may give a consistent "ground truth" for the LIDC data. As the data actually is, there is no simple method to derive a consistent hard outline of each nodule from the radiologists' outlines. Using the same method for all slices to derive a "ground truth" from the data in the LIDC would likely eliminate different features from each set of radiologist outlines.

4. DISCUSSION

Currently overlap is frequently used to measure how similar segmentations are. This becomes a problem in cases where there are more than two segmentations being compared, such as the LIDC. To calculate it, "overlap" can be defined in one of a few different ways, either as pixels where $P(\mathrm{r, c}) = 4$ or $P(\mathrm{r, c}) \geq 3$ or 2. Even this is not a great solution; if the overlap is defined as $P(\mathrm{r, c}) = 4$, all P_3 and P_2 slices would have no overlap.

Even when considering slices in P_2, where overlap would be best suited (because there are only two segmentations), VI_n still provides information that overlap cannot. Figure 5 shows three slices that share nearly the same overlap, but have very different segmentation features. The VI_n has a lower value for the slice with two outlines of roughly the same shape (Fig. 5, top left), while the VI_n is much higher when the shape is drastically different (Fig. 5, top right). The last slice (Fig. 5, top middle) has two outlines with different shapes, but lands somewhere between the two other slices in terms of VI_n, despite having nearly the same overlap.

Overlap still has value though when used alongside VI_n. Just as there are slices with the same overlap and different values of VI_n, there are different values for overlap for slices that have nearly the same VI_n. Using the two together might be able to provide more insight into how the radiologists disagree than using either metric alone. Just as differences are visible in pmaps with different VI_n but the same overlap, similar differences can be seen when the VI_n is relatively constant while overlap varies (Fig. 5, bottom).

VI_n has a number of potential applications. The work done here shows its use in measuring variability in panel segmentations. In addition to panel segmentations used to derive a reference truth, it has value in evaluating CAD systems. For hard segmentations, where the CAD would output a segmentation similar to a single radiologist, and soft segmentations, which would output a segmentation closer to an ensemble of radiologists (like the LIDC), VI_n has different uses.

VI_n can be used to validate hard segmentations, which are more commonly found in literature than soft segmentations. Some papers use previous work to compare to their own CAD systems,[9] or they choose one of the four radiologists to compare with their system.[11] Instead of comparing segmentations by using some form of overlap, VI_n could be used instead, or as a second metric. One could simply compute the VI_n by adding in an additional segmentation, but this would not accurately reflect the effectiveness, because the number of raters would change. Instead, replacing a radiologist with the CAD system and calculating the VI_n would provide feedback on whether or not the CAD system decreases the amount of uncertainty in the panel segmentation or not. A good CAD system could have a lower average VI_n when substituting out each radiologist one at a time than the four original radiologists' outlines.

While VI_n cannot really be used to validate a soft segmentation, it still has value in evaluating a CAD system that produces a series of soft segmentations. A soft segmentation associates a probability ranging from 0 to 1 for each pixel it considers. If these probabilities are grouped into a discrete number of possible probabilities, it becomes very similar to a panel segmentation. For instance, a soft segmentation that outputs probabilities of 100, 75, 50, and 25% would mimic the LIDC ($R = 4$). Once the pmap is created for the soft segmentation, the VI_n can be calculated. The difference is that a single soft segmentation producing different levels of confidence

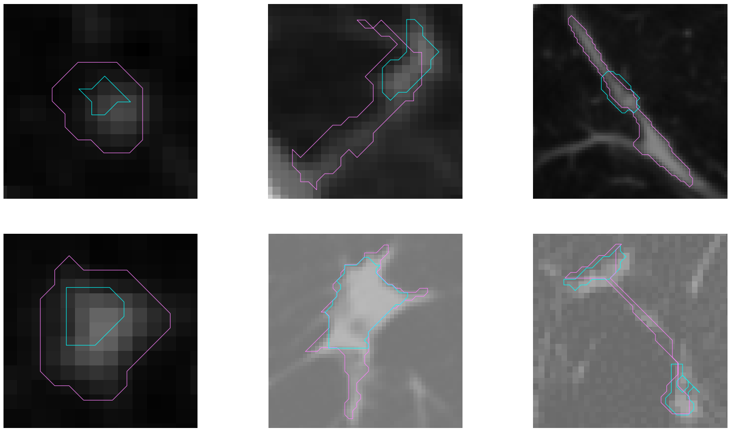


Figure 5. *Top* Three P_2 slices with nearly the same overlap (from left, 22.45%, 22.46%, and 20.64%), but large differences in VI_n (from left, 11.6, 35.0059, and 81.4449). *Bottom* Three P_2 slices with nearly the same VI_n (from left, 12.3711, 12.5101, and 12.5896), but large differences in overlap (from left, 27.63%, 44.62%, and 67.71%).

does not really indicate "variability" or disagreement as much as it indicates how uncertain the boundaries of the segmentation are.

In the case of soft segmentation CAD systems, the standard deviation of a VI_n would indicate how consistently the boundaries of a soft segmentation are defined between different slices. If the system gave very well defined boundaries in some cases, but very indistinct boundaries in others, the standard deviation would be higher. Presumably the goal of a CAD system using this measure would desire a lower standard deviation. A standard deviation lower than the standard deviation of the panel segmentations could indicate that when given a new slice, the CAD system would produce a more predictable result than the panel of experts. However, VI_n, when calculated on a soft segmentation CAD system, does not actually account at all for the reference truth being used. Once again, using it in conjunction with overlap would provide the most information for evaluating CAD systems; overlap could give a general indication of the accuracy, and VI_n could give insight into details not conveyed by overlap.

5. CONCLUSION

While the radiologists often provide consistent segmentations, VI_n is able to quantify to what extent they disagree as a group. Therefore, it can show which panel segmentations are less reliable, which could help researchers identify troublesome cases before their own CAD systems are tested. It is better than overlap because it can be applied to any number of segmentations while still being calculated the same way every time. In addition, it provides values based on the unique shapes of each radiologist's outline in a panel segmentation, which overlap cannot accomplish.

In the future, extending VI_n to 3D would be valuable. Many CAD systems perform segmentations on entire nodules instead of individual slices, and ultimately the entire nodule is the object being characterized as malignant or benign, so having a single variability characterized for the entire nodule would be desirable. In addition to its applications to CAD systems, it may be beneficial to explore possible connections between the variability of the outlines to the semantic ratings given for nodules in the LIDC data.

ACKNOWLEDGMENTS

The research in this paper was supported by NSF award IIS-0755407.

REFERENCES

[1] Armato III, S. G. et al., "The Lung Image Database Consortium (LIDC): An evaluation of radiologist variability in the identification of lung nodules on CT scans," *Acad. Rad.* **14**, 1409–1421 (November 2007).

[2] Armato III, S. G. et al., "Lung Image Database Consortium: Developing a resource for the medical imaging research community," *Radiology* **232**, 739–748 (September 2004).

[3] Armato III, S. G. et al., "Assessment of radiologist performance in the detection of lung nodules: Dependence on the definition of "truth"," *Academic Radiology* **16**, 28–38 (January 2009).

[4] Revesz, G. et al., "The effect of verification on the assessment of imaging techniques," *Investigative Radiology* **18**, 194–198 (March/April 1983).

[5] Miller, D. P. et al., "Gold standards and expert panels: A pulmonary nodule case study with challenges and solutions," in [*SPIE Medical Imaging Conference*], (2004).

[6] Cholleti, S. R. et al., "Veritas: Combining expert opinions without labeled data," in [*20th IEEE International Conference on Tools with Artificial Inteligence*], (2008).

[7] Warfield, S. K. et al., "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging* **23**, 903–921 (July 2004).

[8] Reeves, A. P. et al., "On measuring the change in size of pulmonary nodules," *IEEE Transactions on Medical Imaging* **25**, 435–450 (April 2006).

[9] van Ginneken, B., "Supervised probabilistic segmentation of pulmonary nodules in CT scans," in [*MICCAI*], *LNCS* **4191**, 912–919, Springer Berlin / Heidelberg (2006).

[10] Heimann, T. et al., "Comparison and evaluation of methods of liver segmentation from ct datasets," *IEEE Transactions on Medical Imaging* **28**, 1251–1265 (August 2009).

[11] Opfer, R. and Wiemker, R., "A new general tumor segmentation framework based on radial basis function energy minimization with a validation study on LIDC lung nodules," in [*SPIE Medical Imaging Conference*], (2007).