IEEE Inter. Conf. Image Processing, Atlanta, GA, Oct. 8-11, 2006.

MEASURING INTRA- AND INTER-OBSERVER AGREEMENT IN IDENTIFYING AND LOCALIZING STRUCTURES IN MEDICAL IMAGES

Mehul P. Sampat[1], Zhou Wang[2], Mia K. Markey[1], Gary J. Whitman[3], Tanya W. Stephens[3], Alan C. Bovik[1]

[1]The University of Texas at Austin, Austin, TX 78712, USA
[2]The University of Texas at Arlington, Arlington TX 76019, USA
[3]The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA

ABSTRACT

Inter- and intra-observer variability exists in any measurements made on medical images. There are two sources of variability. The first occurs when the observers identify and localize the object of interest, and the second happens when the observers make appropriate measurement on the object of interest. A number of statistical methods are available to quantify the degree of agreement between measurements made by different observers. However, little has been done to develop metrics for quantifying the variability in identifying and localizing the objects of interest prior to measurement. In this paper, we propose to use the complex wavelet structural similarity index (CW-SSIM) method to measure the variability in identifying and localizing structures on images. Performance comparisons using simulated images as well as real mammography images demonstrate the effectiveness and robustness of the CW-SSIM method.

1. INTRODUCTION

Detecting spiculated masses is critical for early detection of breast cancer, but it is challenging because of the variable appearance of these lesions. Our approach to computer-aided detection is evidence-based, *i.e.*, we use the physical properties of spiculated masses to set the parameters of the detection algorithm [1]. To the best of our knowledge no systematic study has been reported on the statistics of the physical parameters of these lesions. Thus, we are conducting studies in which experienced radiologists measure physical properties of spiculated masses, e.g., the length and width of spicules [2].

Inter- and intra-observer variability exists in any measurements on medical images. There are two important sources of observer variability in measurements of structures on medical images. Firstly, observers have to identify and localize the object of interest and secondly, they have to make the appropriate measurement on the object of interest. Several methods (e.g., Intra-class correlation (ICC), Bland-Altman method) are available for making statistical comparisons of observer measurements. While these methods have strong theoretical foundations and can provide an evaluation of the inter- and intra-observer agreement of measurements made by multiple individuals, they do not account for observer variability in the identification and localization of the objects under study.

To appreciate this issue, consider the following example. In our study, two radiologists (GJW, TWS) measured the properties of spiculated masses. GJW repeated the measurements after an interval of one week. Figure 1(a) shows the tracings made by two radiologists. Similarly, Figure 1(b) shows the two sets of tracings made by the same radiologist on two different occasions. Intuitively, one would expect a reader to agree more with himself in the task of identifying and localizing the spicules than with another individual. By visual inspection, we can see that the intra-observer agreement (Figure 1(a)) is greater than the inter-observer agreement (Figure 1(b)). However, a visual inspection of the overlay of tracings only provides a subjective, qualitative assessment of the observer agreement.

This work aims to develop automatic algorithms that can provide *objective* and *quantitative* evaluations of the observer variability in object identification and localization. A number of related metrics have been proposed previously, but were mainly devoted to image segmentations. Two of the most widely used metrics are the Dice coefficient [3] and the Jaccard coefficient [4]. For our purpose, one common problem with these approaches is that they are sensitive to small spatial translations and rotations. This is an undesirable property because when people are asked to trace or make measurements of linear structures in medical images it is very likely that the tracings in the two (or more) evaluations will be slightly misaligned, even though these tracings maybe intended to represent the same structure. The goal of this study is to develop new metrics that can effectively measure observer variability without being unduly sensitive to very small perturbations. In particular, we propose to use the complex wavelet structural similarity (CW-SSIM) index, which was originally proposed for general-purpose image quality assessment and pattern recognition [5]. We believe it is a good candidate for our purpose because it provides a measure of structural similarity between images and is robust to small geometric distortions.

2. METRICS

2.1. Dice similarity coefficient

The Dice similarity coefficient (DSC) is a simple and intuitive metric [3]. The DSC was selected for this study because it is commonly used in medical imaging studies to quantify the degree of overlap between two segmented objects, e.g., [6,7]. Let $Seg1$ and $Seg2$ represent two binary segmentations of an object made by two experts. A pixel that belongs to the segmented object is labeled one and zero otherwise. Then the DSC ($DSC \in$ [0,1]) is defined as follows:

$$DSC = 2 \times [n(Seg1\ \cap\ Seg2)] \,/\, [n(Seg1)\ +\ n(Seg2)] \tag{1}$$

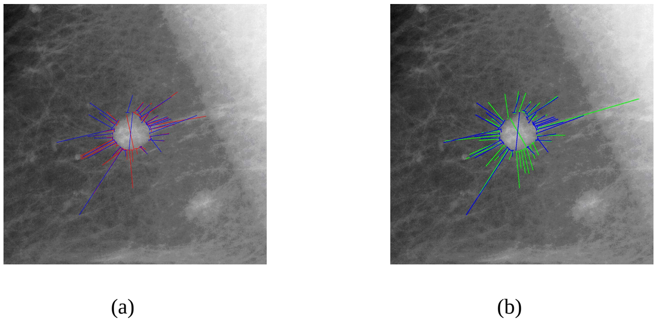
where $n(Seg1 \cap Seg2)$ denotes the number of pixels that are non-zero in both images. This can viewed as a measure of overlap between the two segmentations. $n(Seg1)$ and $n(Seg2)$ represent the number of non-zero pixels in images $Seg1$ and $Seg2$ respectively. Thus, for example if the two segmentations overlap completely then the DSC = 1 and DSC = 0 if there is no overlap. It is generally accepted that a DSC value of greater that 0.7 denotes good agreement [7].

2.2. Complex Wavelet - SSIM

Recently, Wang *et al.* proposed the structural similarity (SSIM) index for the prediction of human preferences in evaluating image quality [5, 8]. The underlying idea of this approach is that the human visual system (HVS) is highly adapted to extract structural information from the visual scene and thus a measure of structural similarity should provide a good estimate of the perceived image quality. It has been demonstrated that the SSIM index is successful in predicting the quality of images degraded with a wide variety of distortion types and levels. In [5], this approach was extended to the complex wavelet domain, and the resulting complex wavelet SSIM (CW-SSIM) index has proven to be more robust than the baseline SSIM index for geometric image distortions. The CW-SSIM method uses the phase information of the coefficients in the complex wavelet domain. It is based on the belief that the structural information of image features is mostly contained in the relative phase patterns of wavelet coefficients [5]. To compute the CW-SSIM metric for two images, we first compute the complex wavelet transform of those images. Let $\mathbf{c}_x = \{c_{x,i} | i = 1, ..., N\}$ and $\mathbf{c}_y = \{c_{y,i} | i = 1, ..., N\}$ be the two sets of coefficients extracted at the same spatial location in the same wavelet subbands of the two images being compared, respectively. The CW-SSIM metric is defined as:

$$\tilde{S}(\mathbf{c}_x, \mathbf{c}_y) = \frac{2\,|\sum_{i=1}^{N} c_{x,i}\, c^*_{y,i}| + K}{\sum_{i=1}^{N} |c_{x,i}|^2 + \sum_{i=1}^{N} |c_{y,i}|^2 + K} \tag{2}$$

Here c^* denotes the complex conjugate of c and K is a small positive constant. The CW-SSIM index ranges from a value of 0 to 1, where 1 denotes perfect similarity between two images.



(a)

(b)

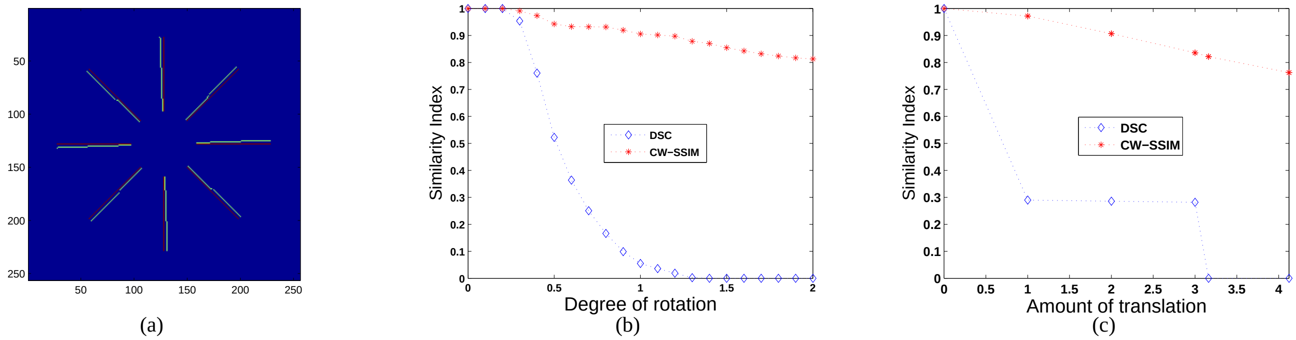
Fig. 1. The measurements made by GJW and TWS which are overlaid on the original image. Figure 1(a) shows the two set of measurements made by the radiologists GJW. The second set of measurements was made after an interval of one week. Whereas, Fig. 1(b) shows the measurements made by the two radiologists.

3. METHODS

3.1. Data Description

Two sets of data were used to compare the DSC and CW-SSIM metrics. The metrics were compared on a set of simulated images and images of measurements made by experienced radiologists. The first set consisted of simulated set of binary images. To generate this set, a binary image was created (Fig.2(a)) and this was then rotated and translated by different amounts to generate the simulated data-set. The original image was rotated from 0.1 to 2 degrees in increments of 0.1 degrees. Note that since the amount of rotation that was applied was very small, the linear segments are very close to each other. These images model, the case when two readers measure linear structures on images and although they maybe measuring the same structure their measurements are off by a few pixels. The original image was also translated by 0 to 4 pixels.

The second of images for this study were obtained from the Digital Database for Screening Mammography (DDSM) [9]. The DDSM is the largest publicly available dataset of digitized mammograms and a set of 12 images containing a single lesion each were randomly selected from those scanned with a single digitizer. The radiologists (GJW and TWS) marked the structures of interest on the images and measured the lesion properties. To compute an estimate of the intra-observer agreement, GJW repeated the process. The analysis was conducted on regions-of-interest (ROIs) using the ROI Manager plugin of NIH ImageJ. Figures 3(a) and 3(b) show the observer tracings along the length of spicules. The binary images were created by assigning a value of one to the pixels marked by the radiologists and zero otherwise. To implement the CW-SSIM index for the comparison of images,



(a) (b) (c)

Fig. 2. In Fig 2(a) the original simulated test image is shown in red. The original image was rotated from 0.1 to 2 degrees in increments of 0.1 degrees. The image that was obtained after rotation by 2 degrees is shown in green in Fig. 2(a). Similarly, the original image was also translated by different distances. The original image was compared to each of the rotated images and translated and for each pair of images the DSC and the CW-SSIM metrics were computed. Figures 2(b) and 2(c) show the effect of rotation and translation on the two metrics. The DSC is sensitive to small rotations and spatial translations, whereas the CW-SSIM metric is robust to these transformations.

we first decompose the images using a complex version of a 4-scale, 8-orientation steerable pyramid decomposition [10]. The CW-SSIM indices are then computed locally using a sliding 7x7 window.

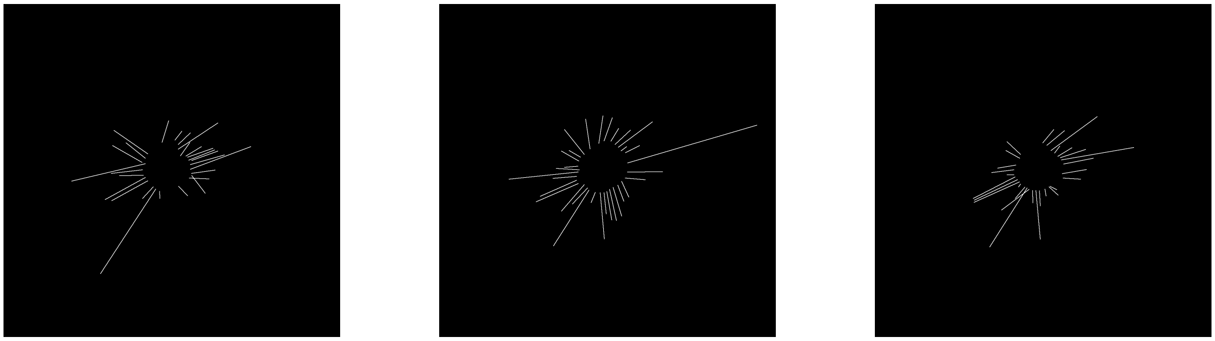
4. RESULTS

For the simulated data, the original image was compared to each of the rotated and translated images and the DSC and the CW-SSIM metrics were computed. The effect of very small rotations and translations on the two metrics is shown in Figs. 2(b) and 2(c) respectively. We see that while DSC is sensitive to small rotations and translations, the CW-SSIM metric is robust to these transformations. Table 1 shows the values obtained for the DSC and CW-SSIM metrics for each of the mammography images. By visual inspection, we can see that the intra-observer agreement is more than the inter-observer agreement, which is intuitive since one would expect a reader to agree more with himself than with another individual. However, although one can see considerable agreement in the spicules outlined by the two radiologists (Figs. 1 and 3) the DSC metric fail to capture this fact and very low values are obtained for each pair of images. (Note that as mentioned earlier a DSC value of greater than 0.7 is considered to denote good agreement). In contrast, we see that the CW-SSIM values are much higher and agree much better with expectations than the DSC values. Secondly, they further agree with expectations, as the CW-SSIM values are much greater for intra-observer agreement than the corresponding inter-observer agreement values for 10 out of the 12 pairs of images.

5. DISCUSSION

In this paper, we have presented the use of the CW-SSIM to quantify the intra- and inter-observer agreement in the localization of structures in medical images. Testing on a simulated test image showed that the CW-SSIM metric is robust to rotations and translations whereas the popular DSC metric is quite susceptible to these transformations. Recently, Warfield *et al.* [11] proposed the STAPLE algorithm to simultaneously obtain a robust estimate of the true segmentation boundary and to compare the accuracy of various segmentation generators. It is difficult to compare the CW-SSIM and STAPLE algorithms because they were designed for different applications. As the STAPLE method was designed to determine the accuracy of segmentation, it penalizes segmentations that are off by even a few pixels. In comparison, if the goal is to trace and measure properties of linear structures (e.g.spicules, blood vessels) then it is highly likely that the measurements may not overlap completely and that a shift of a few pixels should not be penalized and CW-SSIM is ideal for this situation.

It is encouraging to observe that the CW-SSIM metric effectively capture trends that are expected based on visual inspection of the mammography images analyzed in this study. Notably, the within-observer agreement was consistently rated as higher than the between-observer agreement. Statistics for evaluating measurement agreement (e.g., ICC) can be interpreted in a task-independent manner to a large extent. For example, an ICC value of 0.7 is typically taken to indicate adequate agreement for any measurement task. However, it is more difficult to specify a general-cutoff on measures of agreement in structure localization in images such as the CW-SSIM metric. While some efforts have been made to define general cutoffs for measure such as DICE, it maybe be that these metrics will need to be interpreted in a context dependent manner.



(a)

(b) TWS’s First set of Measurement

(c) GJW’s Second set of Measurement

Fig. 3. This figure shows the measurements made by GJW and TWS for the spicule length only. Figures 3(a) and 3(b) show GJW’s and TWS’s first set of measurement respectively. Figure 3(c) shows GJW’s second set of measurements.

6. REFERENCES

[1] M. P. Sampat, G. J. Whitman, M. K. Markey, and A. C. Bovik, “Evidence based detection of spiculated lesions and architectural distortions,” in *SPIE Medical Imaging, Image Processing*, Feb 2005, vol. 5747, pp. 26–37.

[2] M. P. Sampat, G. J. Whitman, L. D. Broemeling, A. C. Bovik, and M. K. Markey, “Inter- and intra-observer variability in measuring properties of spiculated lesions on mammography,” in *Medical Imaging Perception Conference XI*, 2005.

[3] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945.

[4] P. Jaccard, “The distribution of flora in the alpine zone,” *New Phytologis*, vol. 11, pp. 37–50, 1912.

[5] Z. Wang and E.P. Simoncelli, “Translation insensitive image similarity in complex wavelet domain,” in *IEEE International Conference on Acoustics, Speech, and Signal Processin*, 2005, vol. 2, pp. 573–576.

[6] K. Zou, S. Warfield, A. Bharatha, et al., “Statistical validation of image segmentation quality based on a spatial overlap index.,” *Acad Radiol*, vol. 11, no. 2, pp. 178–189, Feb. 2004.

[7] A.P. Zijdenbos, B.M. Dawant, R.A. Margolin, and A.C. Palmer, “Morphometric analysis of white matter lesions in MR images: method and validation,” *IEEE Transactions on Medical Imaging*, vol. 13, no. 4, pp. 716–724, 1994.

[8] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[9] M. Heath, K. W. Bowyer, and D. Kopans, “Current status of the digital database for screening mammography,” *Digital Mammography*, pp. 457–460, 1998.

[10] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, and D.J. Heeger, “Shiftable multiscale transforms,” *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 587–607, 1992.

[11] S.K. Warfield, K.H. Zou, and W.M. Wells, “Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

Image No.	DSC (Intra-Obs)	DSC (Inter-Obs)	CW-SSIM (Intra-Obs)	CW-SSIM (Inter-Obs)
1	0.02	0.03	0.46	0.41
2	0.09	0.03	0.53	0.51
3	0.04	0.02	0.52	0.51
4	0.03	0.05	0.39	0.47
5	0.02	0.05	0.38	0.37
6	0.06	0.01	0.49	0.46
7	0.02	0.03	0.47	0.48
8	0.09	0.04	0.59	0.50
9	0.06	0.01	0.51	0.42
10	0.02	0.01	0.37	0.35
11	0.01	0.02	0.52	0.44
12	0.06	0.02	0.60	0.44

Table 1. Results of the intra- and inter-observer agreement for the CW-SSIM and the DSC metrics. The DSC (metric fails to quantify the intra- and inter-observer agreement for each pair of images. (A DSC value of greater than 0.7 is considered to denote good agreement). In contrast, for 10 of the 12 images the CW-SSIM metrics show that the intra-observer agreement is greater than the inter-observer agreement.