On Reproducibility of Ultrasound Image Classification

Martin Švec[1], Radim Šára[1], and Daniel Smutek[2]

[1] Center for Machine Perception, Department of Cybernetics,
Faculty of Electrical Engineering, Czech Technical University Prague,
Technická 2, 166 27 Praha 6, Czech Republic
{xsvecm,sara}@cmp.felk.cvut.cz
http://cmp.felk.cvut.cz
[2] Charles University Prague, 1st Medical Faculty,
3rd Department of Medicine,
128 08 Praha 2, Czech Republic
smutek@cesnet.cz

Abstract. Ultrasound B-mode images of thyroid gland were previously analyzed to distinguish normal tissue from inflamed tissue due to Hashimoto's Lymphocytic Thyroiditis. This is a two-class recognition problem. Sensitivity and specificity of 100% was reported using Bayesian classifier with selected texture features. These results were obtained on 99 subjects at a fixed setting of one specific sonograph, for a given manual thyroid gland segmentation and sonographic scan orientation (longitudinal, transversal). To evaluate the reproducibility of the method, sensitivity analysis is the topic of this paper. A general method for determining feature sensitivity to variables influencing the scanning process is proposed. Jensen Shannon distances between modified and unmodified inter- and intra-class feature probability distributions capture the changes induced by the variables. Among selected features, the least sensitive one is found. The proposed sensitivity evaluation method can be used in other problems with complex and non-linear dependencies on variables that cannot be controlled. [1]

1 Introduction

Hashimoto's lymphocytic thyroiditis (LT), one of the most frequent thyroid disorders, is a chronic inflammation of the thyroid gland. This disease changes the structure of the tissue. Changes are diffuse (they affect the entire gland) and can be detected by sonographic imaging. Information extracted from images by computers may provide additional support for diagnostic hypothesis.

Automatic recognition of LT has been attempted based on textural image features [12, 14]. Classification was done with features selected by a search pro-

[1] This work has been supported by the Grant Agency of the Czech Academy of Sciences under project 1ET101050403 and by the Czech Ministry of Health under project NO/7742-3.

cedure out of 129 features. The optimal features achieved sensitivity and specificity[2] of 100% in a cross-validation experiment on an independent set of 18 subjects [14].

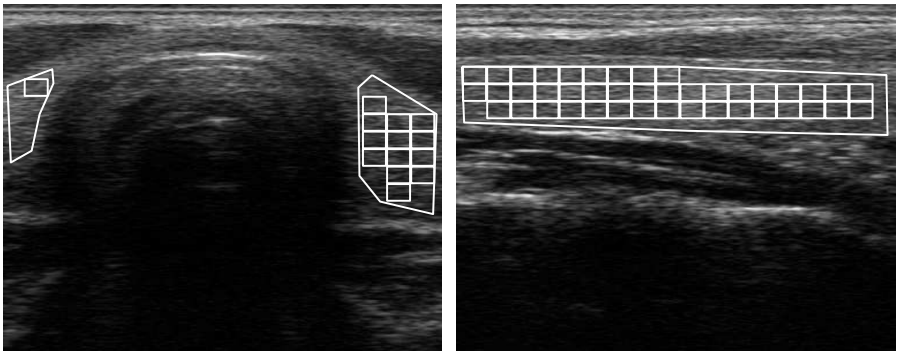
Although high success rate was achieved, the results were limited to one particular setting of one specific sonograph. This has been recognized as the most important obstacle to bringing the method to online clinical practice. The reproducibility issue is a long-standing problem in similar quantitative methods [4]. In relevant works, parameter settings were adjusted for optimal visualization [8], fixed to have standardized conditions [11, 6], or kept at values normally used in clinical practice [3]. Chan [1] tried to tackle this problem by changing the gain setting during the experiment and capturing for each gain at least five images of the object. Mojsilovic [9] removed the mean of each image in order to eliminate effects of unequal ultrasound gain settings.

The goal of this paper is to *quantify reproducibility* of features used previously [14]. Reproducibility is the possibility to achieve the same classification results under different sonograph setting, different gland delineations in the manual segmentation step (depending on physician's knowledge and experience), and different scan orientation (longitudinal or transversal). The proposed analysis is general enough to be applied to other data interpretation problems involving complex and non-linear dependencies on variables that cannot be controlled. The rest of this paper is structured as follows. Texture features are described and sensitivity analysis method is proposed in Sec. 2. Sec. 3 describes a feature sensitivity experiment and results. Discussion follows in Sec. 4 and conclusions are given in Sec. 5.

2 Methods

Given a sonographic B-mode image, a two-class classification problem is considered in the previous work [12, 14]: distinguishing healthy tissue (denoted here as N) from tissue changed due to Hashimoto's Lymphocytic Thyroiditis (denoted as LT). The classification is done on textural features computed from a set of fixed-size rectangular regions referred to as texture samples, as shown in Fig. 1. The non-overlapping samples are obtained from a manually segmented thyroid gland. In previous work, optimally performing features from the set of 129 candidates consisted of Haralick's texture features [5] and Muzzolini's spatial features [10] were automatically searched for. Their performance was measured as Bayes classifier error. The classifier was learned on a training set of 81 patients and classification error was evaluated on an independent test set of 18 subjects. Three one-dimensional features turned up, each for a different texture sample size, i.e. F2 for 41×41, F6 for 31×31 and F7 for 21×21 texture samples (features are named consistently with previous work [14]; the size is given in pixels). The selected features achieved sensitivity and specificity of 100% in a cross-validation experiment on the independent set of 18 subjects. The principal parameters of

[2] Sensitivity is the proportion of subjects with disease who have a positive test result, specificity is the proportion of subjects without disease who have negative test result.



(a) Transversal scan (b) Longitudinal scan

Fig. 1. Sonographic images with manually segmented thyroid gland and covered by rectangular texture samples

the sonograph[3] were fixed in the study: the gain of 92, medium sensitivity by depth, maximal acoustic power, frequency of 8MHz, repetition rate of 19Hz, and maximum spatial resolution of 4cm. All details concerning data acquisition and processing are given in [14].

Features used in the current sensitivity analysis are the selected features F2, F6 and F7, as described above. Feature probability distributions (FPD) for each class were estimated by histogramming. Optimal (the least bias and variance) histogram resolution according to Scott's rule was used [13]. The idea of the sensitivity analysis is to quantify the changes of these histograms under various modifications of the data acquisition.

A suitable statistic for this purpose is a divergence measure between two feature probability distributions. Jensen-Shannon divergence (JS) is used as a semi-definite, additive and symmetric measure. Let X be the range of a discrete random variable and let p_1 and p_2 be two probability distributions over X. The JS is defined in terms of discrete Shannon entropy $H(p)$ of a probability distribution function p as

$$JS(p_1, p_2) = H\left(\frac{p_1 + p_2}{2}\right) - \frac{H(p_1) + H(p_2)}{2} \,. \tag{1}$$

A detailed description is given in [7] and recommendations for practical usage in [2].

Using JS the sensitivity is measured by comparing the inter-class difference d_I (difference between N and LT class) and the within-class difference $d_{W|N}$ or $d_{W|LT}$ (difference between FPD and changed FPD, for given class N or LT). Changes in FPD are given by different 1) sonograph gain setting; 2) thyroid

[3] Toshiba ECCO-CEE, console model SSA-340A, transducer model PLF-805ST.

gland segmentation; and 3) scan orientation according to the following diagram

$$\begin{array}{ccc} & \overset{d_{W|N}}{\longleftrightarrow} & \\ FPD|N & & FPD'|N \\ d_I \updownarrow & & \\ & \underset{d_{W|LT}}{\longleftrightarrow} & \\ FPD|LT & & FPD'|LT \end{array} \tag{2}$$

where $FPD|N$ stands for 'FPD given class N' and $FPD'|N$ stands for the same under changed conditions. The inter-class distance d_I was measured on the full dataset (99 subjects) under the standard gain and the standard segmentation. If d_W is higher than d_I, the feature is sensitive on given changes and reproducibility can not be achieved.

3 Experiments and Results

Sensitivity of the features to sonograph gain setting was determined by computing the distance $d_{W|N}$ between the N class FPDs under the standard gain of 92 and the respective distributions obtained from a set of images from one subject (class N) under two other gain settings (90, 94). The result was compared to the inter-class distance d_I.

A similar method was used to assess the sensitivity of features to the thyroid gland segmentation. Three different boundary delineations for one subject of class N were drawn by another physician in addition to the one used in feature selection.

Finally, the influence of the scan orientation on the features was assessed. Given the feature, the distance between the longitudinal and transversal scans was measured in each class denoted here as $d^s_{W|N}$, $d^s_{W|LT}$, respectively. The larger of the two values $\max(d^s_{W|N}, d^s_{W|LT})$ was compared to the smaller of the two inter-class distances $\min(d_{I|\text{long}}, d_{I|\text{trans}})$ computed for each scan orientation separately according to the following diagram

$$\begin{array}{ccc} & \overset{d^s_{W|N}}{\longleftrightarrow} & \\ FPD|N, \text{long} & & FPD'|N, \text{trans} \\ d_{I|\text{long}} \updownarrow & & \updownarrow d_{I|\text{trans}} \\ & \underset{d^s_{W|LT}}{\longleftrightarrow} & \\ FPD|LT, \text{long} & & FPD'|LT, \text{trans} \end{array} \tag{3}$$

The results of the sensitivity analysis under gain change are shown in Tab. 1. In 21×21 texture samples the differences due to varying gain setting are consistently smaller than the inter-class difference d_I. In 31×31 and 41×41 samples the differences (shown in bold) are already comparable to d_I.

Tab. 2 shows that changes due to the different segmentations s_1, s_2, s_3 are bigger than the inter-class difference d_I for all three features. Again, the least influenced are the 21×21 texture samples.

In Tab. 3 results for different scan orientation (longitudinal, transversal) are shown. The inter-class distances (last two rows) for 31×31 and 41×41 samples

Table 1. The *JS* distances between selected features under the gains of 90 and 94 and the same features under the standard gain of 92 according to Eq. (2). The last row shows inter-class distances

	F7, 21×21	F6, 31×31	F2, 41×41
$d_{W\|N,\ \text{gain=90}}$	0.031	**0.573**	**0.542**
$d_{W\|N,\ \text{gain=94}}$	0.158	0.409	0.475
d_I	0.225	0.532	0.510

Table 2. The *JS* distances between selected features computed under different thyroid gland segmentations s_1, s_2, s_3 and the same features under the standard segmentation according to Eq. (2)

	F7, 21×21	F6, 31×31	F2, 41×41
$d_{W\|N,s_1}$	0.208	**0.701**	**0.850**
$d_{W\|N,s_2}$	0.151	**0.859**	**0.888**
$d_{W\|N,s_3}$	**0.390**	**0.904**	**0.844**
d_I	0.225	0.532	0.510

Table 3. The *JS* distances between selected features from individual scan orientations (first and second row) as compared to the inter-class distances in individual scan orientations (third and fourth row) according to Eq. (3)

	F7, 21×21	F6, 31×31	F2, 41×41
$d^s_{W\|N}$	0.033	0.042	0.021
$d^s_{W\|LT}$	**0.252**	0.092	0.014
$d_{I\|\text{trans}}$	0.478	0.276	0.409
$d_{I\|\text{long}}$	**0.184**	0.575	0.389

are consistently greater than the inter-scan distances (first two rows). Only in the 21×21 samples the two values (in bold) are comparable.

An example of feature probability distributions for N and LT tissue used in this analysis are shown in Fig. 2. We can see the *JS* distances capture the relative differences between the histograms well.

Since gain setting is the parameter that has the greatest influence on visual appearance of sonographic image, features were extracted for a wider gain range (82–100). Measurements were done on a grey-scale phantom[4]. The scatter plots of feature F7_{92} under gain 92 and F7_g under several other gains g are shown in Fig. 3. One data point corresponds to one texture sample (21×21). It can be seen that points make clusters and the mapping induced by the gain change is too complex to be mathematically described. Next we considered simple feature F0 defined as the mean value over the whole rectangular texture sample. Analogical plots to those in Fig. 3 using F0 are shown in Fig. 4.

[4] Precision Small Parts Grey Scale Phantom Gammex 404GS LE

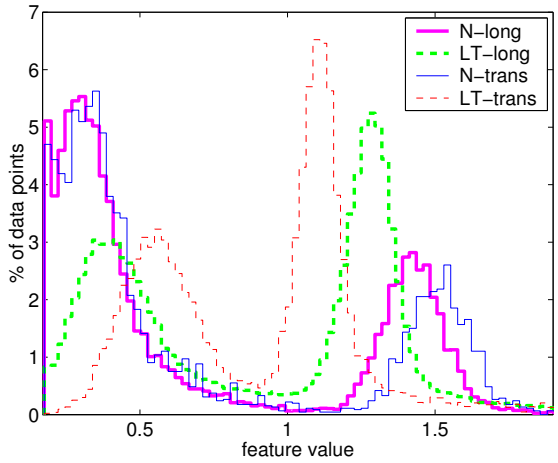


Fig. 2. Histograms for N and LT tissue (feature F7, 21×21 texture samples). We can see histograms for longitudinal scans are more similar than for transversal scans ($d_{I|\text{long}}$ is smaller than $d_{I|\text{trans}}$, see Tab. 3)

4 Discussion

Note that the image area of displayed thyroid tissue is much smaller in transversal scans than in longitudinal ones (see Fig. 1). This means that a smaller number of texture samples fit within the boundary of thyroid gland in a transversal scan as compared to a longitudinal scan. Larger texture samples do not cover the area of the transversal scans well. Therefore, substantial part of available information can be lost for feature construction process from transversal scans. Longitudinal scans provide greater amount of image data from a larger contiguous area of the gland tissue, therefore they should be more useful for automatic texture analysis. However, as can be seen from the last two rows of Tab. 3, distance between N and LT tissue is not always bigger for longitudinal scans than for transversal scans. This can be due to longitudinal artifacts in surrounding and examined tissue, e.g. muscle fibres or vessels. On the other hand we saw in Tab. 3 that inter-class distance is large in transversal scans when F7, 21×21 features are used and in longitudinal scans when F6, 31×31 features are used. Hence, the results could be improved by taking into account longitudinal and transversal images individually, e.g. by combining two classifiers, one using F7 on transversal scans and another using F6 on longitudinal scans.

There is high sensitivity to thyroid gland segmentation according to *JS* distance in larger samples (see Tab. 2). This can be related to sample placement method that leaves small areas along the boundaries uncovered by samples.

To guarantee reproducibility of results under different gain settings, transformation to recalculate features from arbitrary gain to standard gain should be found. From the results shown in Figs. 3,4 it follows that direct transformation of complex features is unfeasible but re-mapping of the raw image values prior

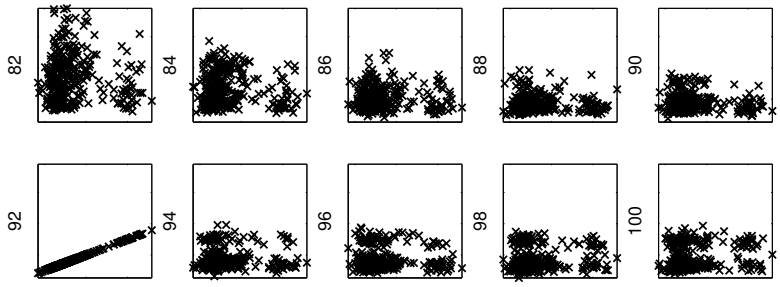


Fig. 3. Comparison of feature F7 under standard gain with F7 under gains of 82 to 100. Coordinates of each point are feature values for standard gain (92, x-axis) and under gain change (82-100, y-axis). 21×21 texture samples are used

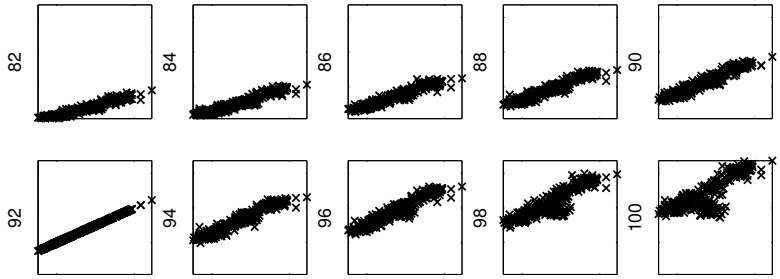


Fig. 4. Comparison of feature F0 under standard gain with F0 under gains of 82 to 100. Coordinates of each point are feature values for standard gain (92, x-axis) and under gain change (82-100, y-axis). 21×21 texture samples are used

to feature computation seems feasible. The results on F0 (see Fig. 4) reveal two components: one approximately linear and another random (see the cluster just below the linear cluster). The origin of this cluster is not known. Further analysis is necessary.

An initial calibration (e.g., using a gray-scale phantom) and subsequent customization of the recognition tool could be another approach for solving the problem of reproducibility. The phantom would need to be specially designed to reproduce the statistical distribution of those features that were found to be optimal for the LT/N classification task. Whether this is feasible remains to be ascertained.

5 Conclusions

The sensitivity analysis shows that the results for 31×31 texture samples and 41×41 texture samples are sensitive to small changes in sonograph setting. Both are also sensitive to different gland segmentations. They are stable under transversal and longitudinal scans.

The 21×21 pixel samples are insensitive to different gain settings and their sensitivity to different gland segmentations is small. They can also distinguish scan orientation, since there is a significant difference between inter-class distances of longitudinal and transversal scans. Distance between N and LT tissue is bigger for transversal than for longitudinal scans.

For greater difference in sonograph parameter setting it will be necessary to remap raw image values by a corrective transformation. We believe that if features of small sensitivity are used subsequently, the classification results will be reproducible. The corrective transformation is a topic for ongoing work.

References

1. Chan, K.L.: Adaptation of ultrasound image texture characterization parameters. In Proc Int Conf IEEE Eng in Medicine and Biology, vol. 2, pp. 804–807, 1998
2. Dhillon, I., Manella, S., Kumar, R.: Information theoretic feature clustering for text classification. Tech. Rep. TR–02–17, Dept of CS, U of Texas at Austin, USA, 2002
3. Dixon, K.J., Vince, D.G., Cothren, R.M., Cornhill, J.F.: Characterization of coronary plaque in intravascular ultrasound using histological correlation. In Proc Int Conf IEEE Eng in Medicine and Biology, volume 2, pages 530–533, 1997
4. Garra, B.S., Krasner, B.H., Horii, S.C., Ascher, S., Mun, S.K., Zeman R.K.: Improving the distinction between benign and malignant breast-lesions: The value of sonographic texture analysis. Ultrasonic Imaging, 15(4):267–285, 1993
5. Haralick, R.M.: Statistical and structural approaches to texture. In Proc IEEE, volume 67, pages 786–804, 1979
6. Hirning, T., Zuna, I., Schlaps, D., Lorenz, D., Meybier, H., Tschahargane, C., van Kaick, G.: Quantification and classification of echographics findings in the thyroid gland by computerized B-mode texture analysis. Europ J Radiol, 9(4):244–247, 1989
7. Lin, J.: Divergence measures based on the Shannon entropy. IEEE Trans Inform Theory, 37(1):145–151, 1991
8. Mailloux, G., Bertrand, M., Stampfler, R., Ethier, S.: Computer analysis of echographic textures in Hashimoto disease of the thyroid. J Clinical Ultrasound, 14(7):521–527, 1986
9. Mojsilovic, A., Popovic, M., Sevic, D.: Classification of the ultrasound liver images with the $2N$ multiplied by 1-D wavelet transform. In Proc IEEE Int Conf Image Processing , volume 1, pages 367–370, 1996
10. Muzzolini, R., Yang, Y., Pierson, R.: Texture characterization using robust statistics. Pattern Recognition, 27(1):119–134, 1994
11. Pohle, R., von Rohden, L., Fisher, D.: Skeletal muscle sonography with texture analysis. In Proc Medical Imaging, vol. 3034 of Proc SPIE, pp. 772–778, 1997
12. Šára, R. Švec, M., Smutek, D., Sucharda, P., Svačina, Š.: Texture analysis of sonographic images for diffusion processes classification in thyroid gland parenchyma. In Proc Conf Analysis of Biomedical Signals and Images, pages 210–212, 2000
13. Scott, D.W.: Multivariate Density Estimation. John Wiley, 1992
14. Smutek, D., Šára, R., Sucharda, P., Tardi, T., Švec, M.: Image texture analysis of sonograms in chronic inflammations of thyroid gland. Ultrasound in Medicine and Biology, 29(11):1531–1543, 2003