Sensitivity Analysis for Reproducibility of Ultrasound Image Classification *

Martin Švec, Radim Šára
Center for Machine Perception
Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University Prague
Technicka 2, 166 27 Praha 6
Czech Republic

Daniel Smutek
Charles University Prague
1st Medical Faculty
3rd Department of Medicine
128 08 Praha 2
Czech Republic

e-mail:{xsvecm,sara}@cmp.felk.cvut.cz,
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 optimal texture features. These results were obtained on 99 subjects at a fixed setting of the sonograph, for a given manual thyroid gland segmentation and sonographic scan type (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. It is shown there are stable features insensitive to small sonograph gain changes and gland segmentation. Features computed from transversal scans are less sensitive because they have greater inter-class distance than features computed from the longitudinal scans. The proposed sensitivity evaluation method can be used in other problems with complex and non-linear dependencies on variables that cannot be controlled.

1 Introduction

Hashimoto's lymphocytic thyroiditis (LT), one of the most frequent thyroid disorders, is a chronic inflammation of the thyroid gland. The incidence rate of LT is higher in women

*This work has been supported by the Czech Ministry of Health (project NO/7742–3) and by the Czech Ministry of Education (project MSM 210000012).

(0.35%) than in men. This disease changes the structure of the tissue. Changes are diffuse (they affect the entire gland) and can be detected by sonographic imaging. Images from the sonographic tool are used by physicians to determine diagnosis of the patient. Not all information present in the sonographic image is accessible to the naked eye. 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 [15, 17]. Classification was done with optimal features selected by a search procedure out of 129 features. The optimal features achieved sensitivity and specificity [1] of 100% in a cross-validation experiment on an independent set of 18 subjects [17].

Although high success rate was achieved, the results were limited to one particular setting of the sonographic tool. 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 [5]. In relevant works, parameter settings were adjusted for optimal visualization [11], fixed to have standardized conditions [14, 7], 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 [12] removed the mean of each image in order to eliminate effects of unequal ultrasound gain settings.

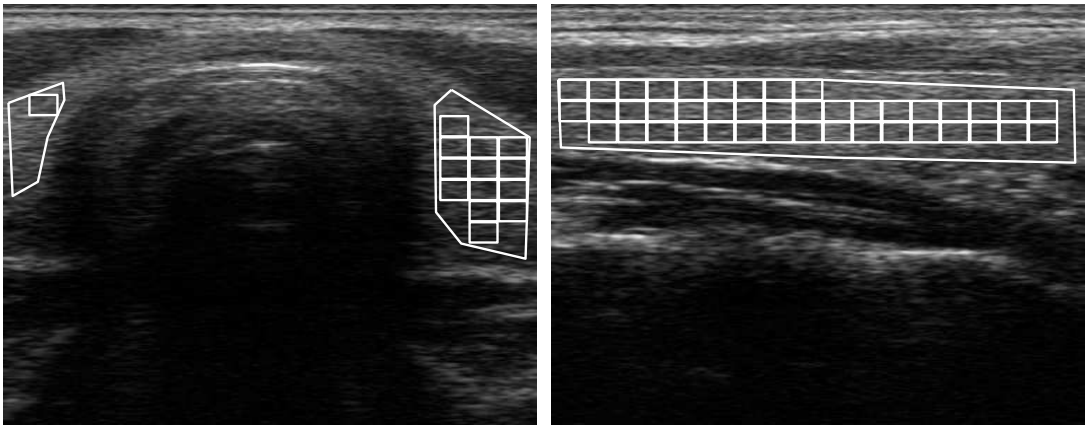
The goal of this paper is to *quantify reproducibility* of optimal features used previously [17]. Reproducibility is the possibility to achieve the same classification results under different sonograph setting, different gland delineations in the manual segmentation step (depending on physician's knowledge and experience, see Fig. 2), and different scan type (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 image, a two-class classification problem is considered in the previous work [15, 17]: 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. Automatic segmentation of thyroid gland in sonographic images is difficult and was not subject of this work. Optimally performing features from the set of 129 candidates including Haralick's texture features [6] and Muzzolini's spatial features [13] were automatically searched for. Their performance was measured as Bayes classifier error. The classifier

[1] 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

Figure 1: Sonographic images with manually segmented thyroid gland and covered by automatically selected rectangular texture samples.

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 optimal features turned up, each for a different optimal 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 our previous work [17]; the size is given in pixels). The optimal features achieved sensitivity and specificity of 100% in a cross-validation experiment on the independent set of 18 subjects. The principal parameters of the sonograph[2] 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 [17].

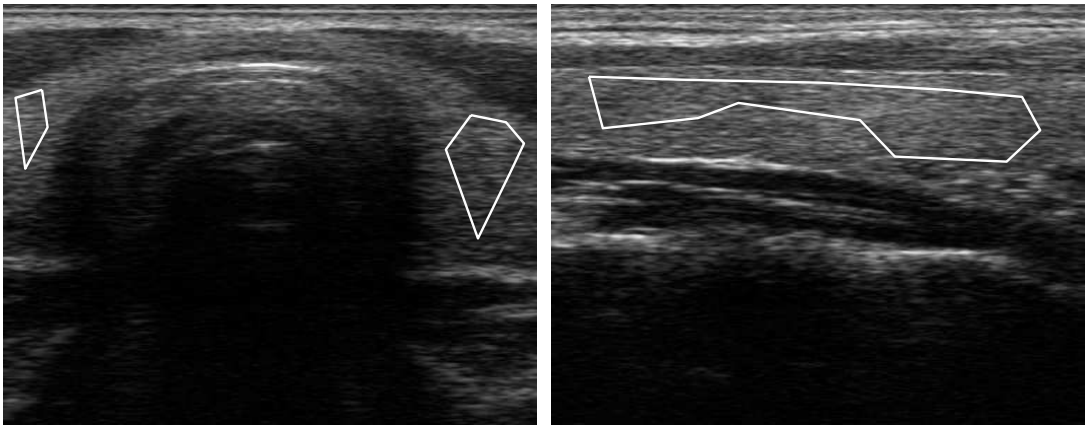
Features used in the current sensitivity analysis are the optimal features F2, F6 and F7, as described above. Feature probability distributions for each class were estimated by histogramming. Optimal (the least bias and variance) histogram resolution according to Scott's rule was used [16]. The idea of the sensitivity analysis is to quantify the changes of these histograms under various modifications of the data acquisition and processing pipeline that produced them.

A suitable statistic for this purpose is a divergence measure between two feature probability distributions. For this purpose, Kullback-Leibler distance (KL) is often used. Let X be the range of a discrete random variable and let p_1 and p_2 be two probability distributions over X. Kullback-Leibler distance is then defined as

$$KL(p_1, p_2) = \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{p_2(x)}. \tag{1}$$

The KL distance is semi-definite, additive but not symmetric [4]. It is undefined if $p_2(x) = 0$ and $p_1(x) \neq 0$, hence to compute this divergence it is recommended to exclude such occur-

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



(a) Transversal scan (b) Longitudinal scan

Figure 2: Segmentation by a different physician of the same images as in Fig. 1.

rences from the sum [8]. Divergence measure that does not require this fix is called Jensen-Shannon divergence (JS). Jensen-Shannon divergence is symmetric. These are the reasons we will use it in this work. The JS is defined in terms of discrete Shannon entropy $H(p)$ of a pdf p as

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

A detailed description is given in [10] and recommendations for practical usage in [2]. Comparison of KL and JS measures is given in [9].

The feature probability distribution (FPD) required in computing JS is estimated by the optimal histogramming. Using JS the sensitivity is measured by comparing the inter-class difference d_I in FPD and the within-class difference $d_{W|N}$ or $d_{W|LT}$ between feature probability distribution and the same distribution changed under a different 1) sonograph gain setting; 2) thyroid gland segmentation; and 3) scan type according to the following diagram

$$\begin{array}{ccc} & \overset{d_{W|N}}{\longleftrightarrow} & \\ F|N & & F'|N \\ d_I \updownarrow & & \\ & \underset{d_{W|LT}}{\longleftrightarrow} & \\ F|LT & & F'|LT \end{array} \tag{3}$$

where $F|N$ stands for 'probability distribution of feature F given class N' and $F'|N$ 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. The within-class distances $d_{W|N}$ were measured between FPD of the class N and FPD obtained from a set of class N images processed under modified conditions. Similarly for $d_{W|LT}$. The FPDs were computed from all longitudinal and transversal scans combined.

Table 1: The JS distances between optimal features under the gains of 90 and 94 and the same features under the standard gain of 92 according to Eq. (3). 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

3 Experiments and Results

Sensitivity of the optimal features to the 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 optimal 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 optimal feature selection.

Finally, the influence of the scan type on the optimal feature vectors was assessed. Given the optimal feature vector 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 type separately according to the following diagram

$$\begin{array}{ccc} & \overset{d^s_{W|N}}{\longleftrightarrow} & \\ F|N,\text{long} & & F'|N,\text{trans} \\ d_{I|\text{long}} \updownarrow & & \updownarrow d_{I|\text{trans}} \\ & \underset{d^s_{W|LT}}{\longleftrightarrow} & \\ F|LT,\text{long} & & F'|LT,\text{trans} \end{array} \tag{4}$$

For instance, the $d_{I|\text{trans}}$ is the distance between N and LT class in transversal scans.

The results of the sensitivity analysis for JS 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 scan type (longitudinal, transversal) are shown. The inter-class distances (last two rows) for 31×31 and 41×41 samples are consistently greater than the inter-scan distances (first two rows). Only in the 21×21 samples the two values (in bold) are comparable.

Table 2: The JS distances between optimal features computed under different thyroid gland segmentations s_1, s_2, s_3 and the same features under the standard segmentation according to Eq. (3).

	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 optimal features from individual scan types (first and second row) as compared to the inter-class distances in individual scan types (third and fourth row) according to Eq. (4).

	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
$d_{I\|\mathrm{trans}}$	0.478	0.276	0.409
$d_{I\|\mathrm{long}}$	**0.184**	0.575	0.389

The feature probability distributions for N and LT tissue used in this analysis are shown in Fig. 3, Fig. 4 and Fig. 5. We can see the JS distances capture the relative differences between the histograms well.

In the full version of this paper [18] sensitivity analysis was done based on both KL and JS distances. The results were similar.

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

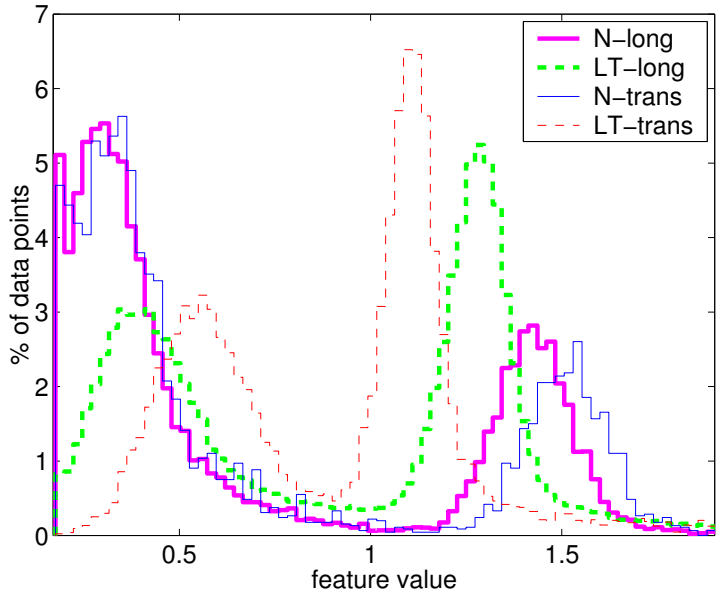


Figure 3: Histograms for N and LT tissue (feature F7, 21×21 texture samples). We can see histograms for longitudinal scans are more similar than for transversal scans, see also Tab. 3.

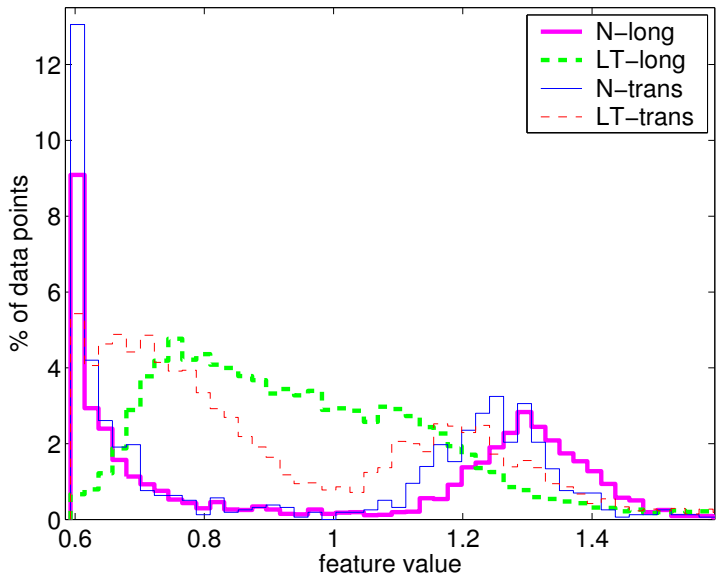


Figure 4: Histograms for N and LT tissue (feature F6, 31×31 texture samples).

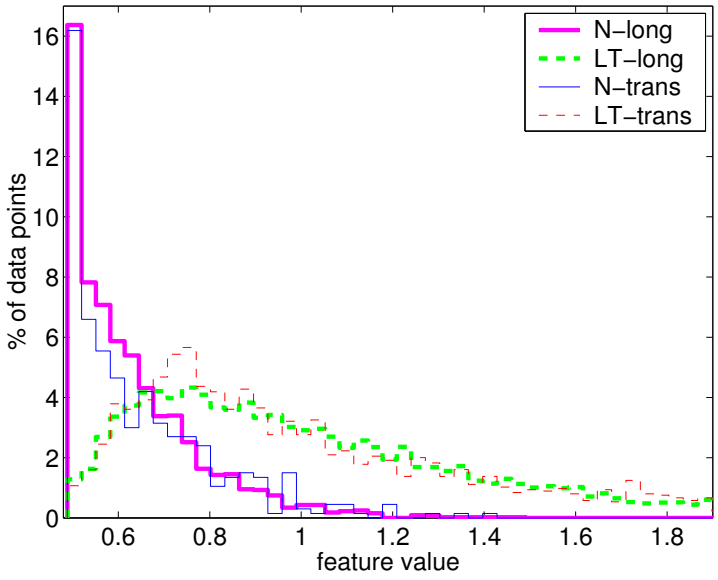


Figure 5: Histograms for N and LT tissue (feature F2, 41×41 texture samples).

are used and in longitudinal scans when F6, 31×31 features are used. Hence, we propose for future work to take into account longitudinal and transversal images individually, e.g. by combining two classifiers, one using F7, 21×21 on transversal scans and another using F6, 31×31 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 texture samples.

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 type, 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.

It has not been shown conclusively in this paper or elsewhere that reproducibility of arbitrary results can be generally achievable in principle or in practice. With regard to the results,

there is no feature stable under all examined changes all at once. However, for the transducer frequency and resolution used we can recommend 21×21 samples from transversal images as a good choice for classification thyroid gland images under different sonograph setting and gland segmentation. The question whether there is an optimal sample size with respect to both recognition success rate and reproducibility is left for further research.

References

[1] K.L. Chan. Adaptation of ultrasound image texture characterization parameters. In *Proc Int Conf IEEE Eng in Medicine and Biology*, volume 2, pages 804–807, 1998.

[2] I. Dhillon, S. Manella, and R. Kumar. Information theoretic feature clustering for text classification. Technical Report TR–02–17, Dept of CS, U of Texas at Austin, TX USA, 2002.

[3] K.J. Dixon, D.G. Vince, R.M. Cothren, and J.F. Cornhill. 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] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Willey, 2001.

[5] B.S. Garra, B.H. Krasner, S.C. Horii, S. Ascher, S.K. Mun, and 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.

[6] R.M. Haralick. Statistical and structural approaches to texture. In *Proc IEEE*, volume 67, pages 786–804, 1979.

[7] T. Hirning, I. Zuna, D. Schlaps, D. Lorenz, H. Meybier, C. Tschahargane, and G. van Kaick. Quantification and classification of echographics findings in the thyroid gland by computerized B-mode texture analysis. *Europ J Radiol*, 9(4):244–247, 1989.

[8] A. Korhonen and Y. Krymolowski. On the robustness of entropy-based similarity measures in evaluation of subcategorization acquisition systems. In *Proc Conf Natural Language Learning*, 2002.

[9] L. Lee. Measures of distributional similarity. In *Proc 37th Annual Meeting of the ACL*, pages 25–32, 1999.

[10] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Trans Inform Theory*, 37(1):145–151, 1991.

[11] G. Mailloux, M. Bertrand, R. Stampfler, and S. Ethier. Computer analysis of echographic textures in Hashimoto disease of the thyroid. *J Clinical Ultrasound*, 14(7):521–527, 1986.

[12] A. Mojsilovic, M. Popovic, and D. Sevic. 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.

[13] R. Muzzolini, Y. Yang, and R. Pierson. Texture characterization using robust statistics. *Pattern Recognition*, 27(1):119–134, 1994.

[14] R. Pohle, L. von Rohden, and D. Fisher. Skeletal muscle sonography with texture analysis. In *Proc Medical Imaging*, volume 3034 of *Proc SPIE*, pages 772–778, 1997.

[15] R. Šára, M. Švec, D. Smutek, P. Sucharda, and Š. Svačina. 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.

[16] D.W. Scott. *Multivariate Density Estimation*. John Wiley, 1992.

[17] D. Smutek, R. Šára, P. Sucharda, T. Tardi, and M. Švec. Image texture analysis of sonograms in chronic inflammations of thyroid gland. *Ultrasound in Medicine and Biology*, 29(11):1531–1543, 2003.

[18] M. Švec, R. Šára, and D. Smutek. Sensitivity of optimal texture features on presegmentation and parameters of sonographic image acquisition process. Research Report CTU–CMP–2003–22, Center for Machine Perception, Czech Technical University, 2003.