In Proceedings Artificial Intelligence in Medicine, pp. 339-348, LNAI 2101, Springer, 2001.

Systematic Construction of Texture Features for Hashimoto's Lymphocytic Thyroiditis Recognition from Sonographic Images

Radim Šára[1], Daniel Smutek[2], Petr Sucharda[2], and Štěpán Svačina[2]

[1] Center for Machine Perception, Czech Technical University, Prague, Czech Republic
sara@cmp.felk.cvut.cz

[2] 1st Medical Faculty, Charles University Prague, Czech Republic
smutek@cesnet.cz, {petr.sucharda, stepan.svacina}@lf1.cuni.cz

Abstract. The success of discrimination between normal and inflamed parenchyma of thyroid gland by means of automatic texture analysis is largely determined by selecting descriptive yet simple and independent sonographic image features. We replace the standard non-systematic process of feature selection by *systematic feature construction* based on the search for the separation distances among a clique of n pixels that minimise conditional entropy of class label given all data. The procedure is fairly general and does not require any assumptions about the form of the class probability density function. We show that a network of weak Bayes classifiers using 4-cliques as features and combined by majority vote achieves diagnosis recognition accuracy of 92%, as evaluated on a set of 741 B-mode sonographic images from 39 subjects. The results suggest the possibility to use this method in clinical diagnostic process.

1 Introduction

Hashimoto's lymphocytic thyroiditis, one of the most frequent thyropathies, is a chronic inflammation of the thyroid gland [21]. The inflammation in the gland changes the structure of the thyroid tissue. These changes are diffuse, affecting the entire gland, and can be detected by sonographic imaging.

The advantages of using sonographic imaging are obvious. However, in clinical praxis the assessment of the diffuse processes is difficult [16, 19] and the diagnosis is made only qualitatively from the size of the gland being examined, from the structure and echogenicity of its parenchyma, and from its perfusion. In making an overall evaluation of a sonogram, the physician uses her/his clinical experience without giving any quantifiable indexes which are reproducible.

Early studies of automatic texture analysis of thyroid gland [10] were limited to the comparison of grey-level histograms of different diagnoses. Later works involved mainly localised changes (e.g., nodules, tumours and cysts) in thyroid tissue [6, 12, 13]. Our final goal is quantitative assessment of the diffuse processes associated with chronic inflammations. This is facilitated by recent developments

in imaging technology that considerably improved the quality of sonograms, mainly of subsurface organs such as thyroid gland.

This paper focuses on diagnosis recognition problem rather than on finding quantitative indexes measuring the degree of inflammation. Classification is a much simpler task, since it requires less training data. If properly chosen, the classification methods generalise to regression methods.

In the first stage of our project we showed that local texture properties of sonographic B-mode images measured by the first-order texture statistic are independent of the location in the image and thus are suitable for tissue classification [15]. Further step of our research focused on the selection of a subset of co-occurrence matrix features suitable for classification [17]. In [18] we tried to use a large set of texture features based on co-occurrence matrices combined with features proposed by Muzzolini et al. [14]. As there was no way to systematically explore all possible features to find the best-suited, we turned our attention towards methods that do not require classical features.

This paper is concerned with the following principal questions: *What are the simplest texture features that are most efficient in distinguishing between normal tissue and the chronic inflammation process in thyroid gland by means of texture analysis? Can these features be found in a systematic way?*

The approach we take avoids the standard heuristic and non-systematic process of descriptive feature selection. The texture feature of our choice is nothing but a small subset of untransformed image pixels. All such possible features differ in how many pixels are involved. They are all parameterised by the separation vectors among the pixels. The search for the optimal feature can thus be done in a systematic way. It is implemented as a simple exhaustive search for the optimum separation distances in a clique of s sampling pixels. The search is performed in a space of dimension $2(s-1)$ and minimises the conditional entropy between class label and data. This *systematic feature construction* procedure is fairly general and does not require any assumptions about the form of the class probability density function.

2 Systematic Feature Construction

On the image grid we define a system of data sampling variables by a set of translation rules according to [8]. Each data sampling variable holds the value of an individual image pixel. Their range is thus a discrete set. The translation rules define the mutual positional relationship among the sampling variables and can be represented by a rectangular mask as illustrated in Fig. 1. Let C be class label variable and D be a system of s data sampling variables. We say certain vector $\mathbf{d}$ of dimension s is a data sample if it is obtained by placing a sampling mask at some position in the image and reading out the image values in the order defined by the order of the corresponding translation rules. The mask must only be placed such that no image pixel is used more than once in this data collection procedure. The position of the mask shown in Fig. 1 defines a data sample of $\mathbf{d} = [3, 7, 5]$.

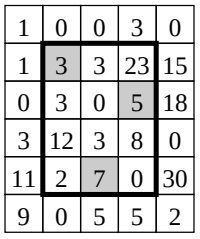


Fig. 1. A system of three sampling variables (shaded) defined on image grid and represented by the sampling mask (thick rectangle). The triple has a base variable corresponding to translation rule $(0, 0)$ (upper left corner of the mask) and two more variables defined by translation rules $(3, 1)$ and $(1, 2)$, respectively. Numbers are image values

Conditional entropy $H(C \mid D)$ tells us how much information in bits is missing in all image data about the class we want to determine:

$$H(C \mid D) = -\sum_{i=1}^{n} p(C, D) \log p(C \mid D)\,, \tag{1}$$

where n is the number of samples collected from all data, $p(\cdot)$ is probability, and $\log(\cdot)$ is the dyadic logarithm. If $H(C \mid D) = 0$ the data contain unambiguous information about the class, i.e. there is some (unknown) function f such that $C = f(D)$. If $H(C \mid D) = H(C)$ the data contain no information about the class. It is not difficult to prove that

$$B = \frac{H(C \mid D)}{c \log c} \tag{2}$$

is an upper bound on Bayesian classification error (of a sample), where c is the number of classes. No classifier can achieve worse error. For the definition of Bayesian error see, e.g. [4].

A sampling system is equivalent to object feature in the classical recognition literature [4]. There is no known efficient way to systematically generate all possible object features. In sampling systems the situation is quite different. Since a sampling system of s variables is defined by $2(s-1)$ parameters, it is possible to find an optimum system for a given recognition problem. We suggest the optimality of a sampling system should be measured by B. It is easy to evaluate B for given discrete data in $O(s\,n \log n)$ time, where n is the data size. The advantage of using entropy is that the resulting optimal sampling system is not biased by any systematic or random artefacts (patterns) in sonographic images.

3 Experiments

The principal goal of the experiments was not to design an efficient classifier with good generalisation properties, it was merely to estimate the sufficiency of our data for the classification task, estimate the smallest number of sampling variables needed, and obtain a good estimate of the classification accuracy that can be achieved. By classification we mean automatic diagnosis recognition given a set of sonographic scans collected for an individual.

We collected data from 37 subjects, 17 of them had normal thyroid (class N) and 20 had Hashimoto's lymphocytic thyroiditis (class LT). The diagnosis was confirmed by clinical examination, elevation of level of antibodies against thyroid gland and by fine needle aspiration biopsy, which are standard diagnostic criteria. Typically, 10 images of the longitudinal cross-section of the left lobe and 10 images for the right lobe were acquired. Twenty images per subject were found sufficient to suppress data acquisition noise. We did not distinguish between the left and right lobe scans in this experiment, since the observed changes are supposed to be diffuse, affecting the entire gland.

Sonographic imaging system (Toshiba ECCO-CEE, console model SSA-340A, transducer model PLF-805ST at frequency 8 MHz) was used to acquire input images. The system settings were kept constant during data collection. We reduced the original 8-bit image resolution to 5 bits. Reductions to 4 up to 6 bits showed very similar results but the computational time needed to compute B differed significantly, especially for large sampling systems.

Typical images of our two classes at full resolution are shown in Fig. 2. Note that the variability of the LT class is much greater. This is due to the fact that chronic inflammations of thyroid gland can be divided into several nosologic units [9].

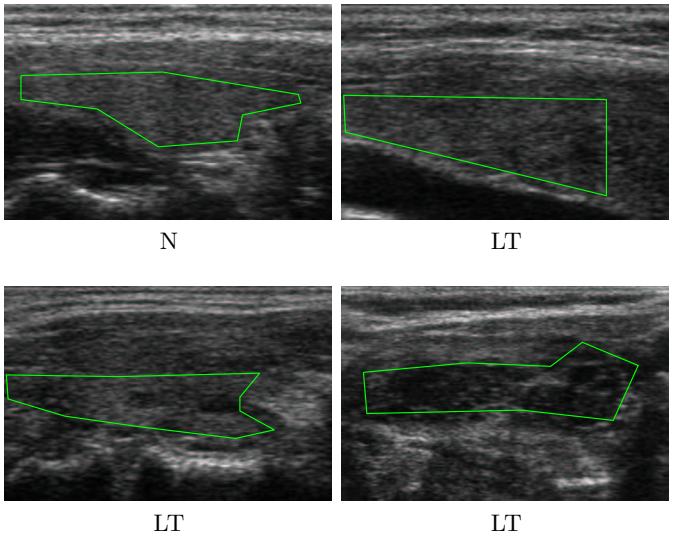


Fig. 2. Typical longitudinal cross sections of the thyroid gland in four different subjects. Skin is the topmost white structure. The classes are N (normal) and LT (Hashimoto's lymphocytic thyroiditis). The outline roughly delineates the gland and is the region from which data was considered. Image size is 250 × 380 pixels

Since the changes in the gland are diffuse it is possible to use global textural characteristics within each image region corresponding to the thyroid gland tissue. Automatic segmentation of such regions is complex and is not the subject of this paper. An interactive tool was used to delineate the boundary of the gland. See Fig. 2 for the result of this manual segmentation.

3.1 Evaluation of B from Data

For a given sampling system, the maximum possible number of non-overlapping samples was collected from the thyroid region in each image. Each individual sample was assigned class label (N, LT). A collection of such samples from all images of a subject formed a data set representing that subject's thyroid. All data from all subjects together with the corresponding class labels in each sample formed one large data-system with s data variables D and one class variable C. This data system was then reduced to discrete probability function by aggregating equal states and estimating the probability $p(C, D)$ of each of the states as its relative frequency in the dataset [8]. The probability $p(C, D)$ was subsequently used to evaluate B using (2).

3.2 Full Search in Low-Dimensional Space

For given s we first constructed optimal sampling system by exhaustive search for the minima of B over the space of $s-1$ separation variables. We have chosen only vertical and horizontal separations since these are the two natural independent image directions: Namely, the vertical direction is the ultrasonic wave propagation direction. In our experiments we always found a unique global minimum for B for given s. Within the separation search range of $[0, 30]^s$ there were other minima with close values of B as well, however.

Once the optimum sampling system was found, the corresponding probability density functions $p(C \,|\, N)$ and $p(C \,|\, LT)$ were used in a network of weak Bayes classifiers whose structure is shown in Fig. 3. All classifiers use the same two probability density functions. Each of them classifies *one* independent (i.e. non-overlapping) texture sample defined by the sample mask. Their outputs are then combined by majority vote to determine the final class label (i.e. the most probable subject diagnosis). The number of classifiers used to determine one final class label per subject is not constant and depends on the number of data available for each subject (the number of images and the size of the thyroid region). It generally decreases with increasing s as there are usually less independent samples that can be collected. In our case it varied between about 250 000 for $s = 1$ and 15 000 for $s = 8$.

For the networked classifier we evaluated the re-substitution error R, which measures the ability of features to describe the entire training dataset, and the leave-one-out error L, which is related to the generalisation capability of the classifier. Both errors are evaluated on the set of 37 subjects, not just individual images or samples (the resolution of the experiment is thus $100/37 \approx 2.7\%$). See [4] for the definition of R and L. Note that B is an upper bound on *sample* recognition error, not the diagnosis classification error. In a properly designed classifier, however, smaller sample recognition error results in smaller object recognition error [4] (as long as it is smaller than 0.5).

The experimental results are shown in Tab. 1. Each sampling system S_s is represented by the set of translation rules. Along with the Bayesian error upper bound B, the leave-one-out classification error L, and the re-substitution error

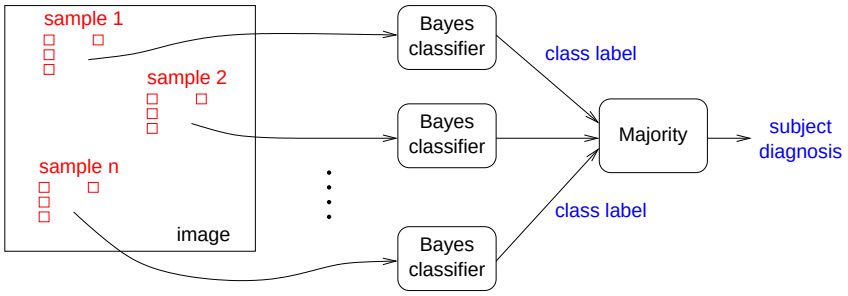


Fig. 3. Classifier structure

R we also show false negatives F^- (inflamed state not recognised), and false positives F^+ (normal state misclassified). It holds that $L = F^- + F^+$. We can see that the system for $s = 4$ is the first one in complexity that is able to fully describe the training set (its re-substitution error is zero). We can also see that the false negative rate F^- is always higher than the false positive rate F^+. This is probably related to the fact that the LT class has richer structure than the N class. The simplest explanation is that the LT class is represented by several subclasses in our data and has thus multi-modal probability distribution function. We observed this property in our earlier experiments as well [18].

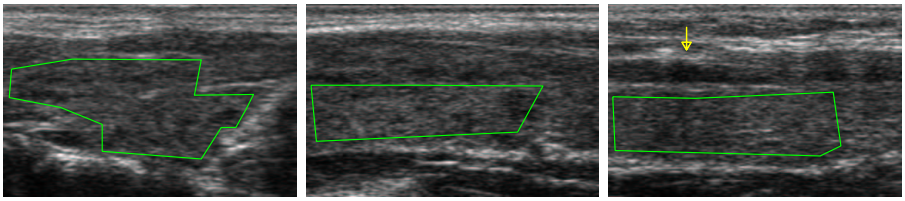
Table 1. Recognition results for optimal sampling systems for given s. The experiment resolution is 2.7%

s	system S_s	B	L	R	F^-	F^+
1	$(0,0)$	25.3%	**8.1%**	8.1%	8.1%	0%
2	$(0,0)$, $(11,0)$	19.9%	**10.8%**	8.1%	5.4%	5.4%
3	$(0,0)$, $(11,0)$, $(0,34)$	16.6%	**13.5%**	10.8%	8.1%	5.4%
4	$(0,0)$, $(7,0)$, $(15,0)$, $(0,32)$	14.2%	**8.1%**	0%	5.4%	2.7%

We can see that the leave-one-out error L increases initially and then drops back to the 8.1% level. This suggests an important part of information about class is present in higher-dimensional features. The relatively good results in leave-one-out error L for the simplest sampling system may be due to the fact that the simplest classifier has a good generalisation property. This observation is consistent with reports of successful classification of sonographic textures using one-dimensional histograms [6, 10].

Figure 4 shows images for the three misclassified cases using sampling system S_4. The sonographic texture in the two false negatives still exhibits some non-uniformity suggesting the model S_4 may not be sufficient. The false positive image (far right) seems to be influenced by acoustic shadows below the superficial hyperechogenic structures (arrow). In all three misclassifications, the assigned class label probability was near to the undecidability level of 0.5, namely it was

0.48 for the two LT cases and 0.498 for the N case. There were three more class label probabilities below 0.6 in the correctly classified diagnoses. Thus, with S_4 the total of just six cases belonged to the correct class with probability smaller than 0.6. For comparison, with S_1 there were nine such cases.



LTs misclassified as Ns (F^-) N misclassified as LT (F^+)

Fig. 4. Images misclassified using S_4

Sampling systems larger than $s = 4$ were not considered because of prohibitively long computational time needed to search through the high-dimensional space. For higher dimensions we used approximate search as described next.

3.3 Approximate Search in High-Dimensional Space

In this experiment we did not search the entire parameter space to find the optimum sampling variable configuration. Instead, a higher-order system $\widehat{S}_{n+1}$ was constructed from $\widehat{S}_n$ by adding one optimal translation rule at a time, while the translation rules in $\widehat{S}_n$ remained fixed. To distinguish systems found by optimal and sub-optimal search we denote them as S_i and $\widehat{S}_i$, respectively.

Each new translation rule was found by minimising a one-parametric bound $B(x)$ for system $\widehat{S}_{n-1} \cup (x, 0)$, where x was the only free variable. The variable range was limited to the interval of $[0, 30]$. Since the full search experiment results suggest that sampling variables defined by vertical translations are more relevant for our recognition task and since it is also the principal direction in the sonographic image, we used only vertical translations in this experiment. The search begun with the simplest system $\widehat{S}_1 = \{(0, 0)\}$. The order in which the optimal translation rules were added was $(11, 0)$, $(17, 0)$, $(5, 0)$, $(20, 0)$, $(14, 0)$, $(7, 0)$, $(2, 0)$. The largest system we tested was thus $\widehat{S}_8$ comprising of all eight translation rules.

The results of the search are shown in Tab. 2. The first system that is able to fully describe the training set, $\widehat{S}_5$, has one variable more than the fully descriptive system in the previous experiment. This shows that the approximate search is quite efficient and finds systems close to those that are optimal. Note that the leave-one-out error is larger in this set of classifiers because of the false negative component of the error. This suggest the horizontal direction plays some role after all and should be accounted for as well.

Table 2. Recognition results for approximately optimal sampling systems $\widehat{S}_s$

system	B	L	R	F^-	F^+
$\widehat{S}_1 = S_1$	25.3%	**8.1%**	8.1%	8.1%	0.0%
$\widehat{S}_2 = S_2$	19.9%	**10.8%**	8.1%	5.4%	5.4%
$\widehat{S}_3$	17.3%	**10.8%**	8.1%	8.1%	2.7%
$\widehat{S}_4$	15.1%	**10.8%**	2.7%	8.1%	2.7%
$\widehat{S}_5$	12.2%	**10.8%**	0.0%	8.1%	2.7%
$\widehat{S}_6$	7.2%	**10.8%**	0.0%	8.1%	2.7%
$\widehat{S}_7$	2.8%	**8.1%**	0.0%	2.7%	5.4%
$\widehat{S}_8$	0.7%	**8.1%**	0.0%	0.0%	8.1%

The leave-one-out recognition accuracy remains at the 92% level for systems above 6 sampling variables. Larger systems would probably result in overfitting [4]. For completeness, in S_8, the total of four cases belonged to the correct class with probability below 0.6 (cf. the results reported in the previous section).

Note that the trend of the F^- error is exactly opposite to that of the F^+. The $F^- = 0$ means that the Hashimoto's lymphocytic thyroiditis was always recognised and $F^+ = 0$ means that normal condition was always recognised. This means that low-order features capture the N class well, while the high-order features capture well the structure of the LT class. The residual errors of 8.1% suggest the existence of partial class overlap in feature space, which is in agreement with observations made by Mailloux et al. [10].

Our next goal is to collect more data to see whether the achieved classification accuracy is low due to the small size of our dataset, due to large class overlap, or due to poor discriminability of our features.

3.4 Comparison with Other Sonographic Texture Recognition Results

We are not aware of any published results that would discriminate between normal and chronically inflamed thyroid parenchyma based solely on sonographic texture analysis. However, other types of tissue were discriminated successfully as shown in the following brief overview.

Hirning et al. report 85% success in detection of nodular lesions in thyroid using the 90-percentile of one-dimensional histogram as a feature [6]. Muller et al. report 83.9% diagnostic accuracy in distinguishing malignant and benign thyroid lesions based on three texture features [13]. Ariji et al. report 96.9% diagnostic accuracy in the diagnosis of Sjogren's syndrome in parotid gland using a combination of spectral feature and standard deviation features [1]. Cavouras et al. report 93.7% classification accuracy for distinguishing normal and abnormal livers (cirrhosis, fatty infiltration) using twenty two features [2]. Horng et al. report 83.3% classification accuracy for distinguishing normal liver, hepatitis and cirrhosis using two-dimensional histograms defined on sonographic image gradient [7]. Mojsilovic et al. report 92% classification accuracy in detecting

early stage of liver cirrhosis using wavelet transform [11]. Sujana et al. report up to 100% classification accuracy in distinguishing liver lesions (hemangioma and malignancy) using run-length statistics and neural network [20]. Chan reports 94% classification accuracy in images of unspecified tissue [3].

Although the results of these methods are not comparable since they were applied to very different tissues and the classification error was evaluated using different statistical methods, the overview suggests our features have good discriminatory power as compared to features used in similar recognition problems using B-mode sonographic images.

4 Conclusions

Our results show that it is possible to achieve good discrimination (92%) of chronically inflamed thyroid tissue from normal thyroid gland by means of low-dimensional texture feature vector. The vector is constructed from a sample of just four pixel values. The four pixels are separated by certain translation vectors that are found by a simple optimisation procedure. Of all possible feature vectors it minimises the conditional entropy of class label given all collected data *and* is the smallest-dimension vector that achieves zero re-substitution error. This optimality property guarantees the best utilisation of available data and the best generalisation properties of a classifier using the features. Note that our features have no explicit intuitive meaning as standard features often do.

The results suggest that the information related to diagnosis can be extracted well from high-quality sonographic images. There is thus a possibility to calculate quantitative characterisation of the chronic inflammatory processes in thyroid gland, which the human visual system is not capable to achieve. In clinical diagnostic process this characterisation enables objective reproducibility of sonographic findings. This facilitates the assessment of changes in the thyroid tissue in follow-up examinations of the same subject made by different physicians.

Acknowledgements This research was supported by the Ministry of Health of the Czech Republic under project NB 5472–3. The first author was supported by the Czech Ministry of Education under Research Programme MSM 210000012.

We would like to thank Mohamed Bakouri who implemented an efficient Matlab toolbox for computations involving discrete probability functions.

References

1. Y. Ariji et al. Texture analysis of sonographic features of the parotid gland in Sjogren's syndrome. *AJR Am J Roentgenol*, 166(4):935–941, 1996.
2. D. Cavouras et al. Computer image analysis of ultrasound images for discriminating and grading liver parenchyma disease employing a hierarchical decision tree scheme and the multilayer perceptron neural network classifier. In *Medical Informatics Europe '97*, vol. 2, pp. 522–526, 1997.

3. K. Chan. Adaptation of ultrasound image texture characterization parameters. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2, pp. 804–807, 1998.
4. K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Computer Science and Scientific Computing. Academic Press, (2nd ed.), 1990.
5. R. M. Haralick and L. G. Shapiro. *Computer and Robot Vision*, vol. 1. Addison-Wesley, 1993.
6. T. Hirning et al. Quantification and classification of echographic findings in the thyroid gland by computerized B-mode texture analysis. *Eur J Radiol*, 9(4):244–247, 1989.
7. M.-H. Horng, Y.-N. Sun, and X.-Z. Lin. Texture feature coding method for classification of liver sonography. In *Proceedings of 4th European Conference on Computer Vision*, vol. 1, pp. 209–218, 1996.
8. G. J. Klir. *Architecture of Systems Problem Solving*. Plenum Press, 1985.
9. P. R. Larsen, T. F. Davies, and I. D. Hay. The thyroid gland. In *Williams textbook of endocrinology*. Saunders (9th ed), 1998.
10. G. Mailloux, M. Bertrand, R. Stampfler, and S. Ethier. Computer analysis of echographic textures in Hashimoto disease of the thyroid. *JCU J Clin Ultrasound*, 14(7):521–527, 1986.
11. A. Mojsilovic, M. Popovic, and D. Sevic. Classification of the ultrasound liver images with the $2N$ multiplied by 1-D wavelet transform. In *Proceedings of the IEEE International Conference on Image Processing*, vol. 1, pp. 367–370, 1996.
12. H. Morifuji. Analysis of ultrasound B-mode histogram in thyroid tumors. *Journal of Japan Surgical Society*, 90(2):210–21, 1989. In Japanese.
13. M. J. Muller, D. Lorenz, I. Zuna, W. J. Lorenz, and G. van Kaick. The value of computer-assisted sonographic tissue characterization in focal lesions of the thyroid. *Radiologe*, 29(3):132–136, 1989. In German.
14. R. Muzzolini, Y.-H. Yang, and R. Pierson. Texture characterization using robust statistics. *Pattern Recognition*, 27(1):119–134, 1994.
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 *Proceedings Conference Analysis of Biomedical Signals and Images*, pp. 210–212, 2000.
16. J. F. Simeone et al. High-resolution real-time sonography of the thyroid. *Radiology*, 145(2):431–435, 1982.
17. D. Smutek, R. Šára, M. Švec, P. Sucharda, and Š. Svačina. Chronic inflammatory processes in thyroid gland: Texture analysis of sonographic images. In *Telematics in Health Care – Medical Infobahn for Europe, Proceedings of the MIE2000/GMDS2000 Congress*, 2000.
18. D. Smutek, T. Tjahjadi, R. Šára, M. Švec, P. Sucharda, and Š. Svačina. Image texture analysis of sonograms in chronic inflammations of thyroid gland. Research Report CTU–CMP–2001–15, Center for Machine Perception, Czech Technical University, Prague, 2001.
19. L. Solbiati et al. The thyroid-gland with low uptake lesions—evaluation by ultrasound. *Radiology*, 155(1):187–191, 1985.
20. H. Sujana, S. Swarnamani, and S. Suresh. Application of artificial neural networks for the classification of liver lesions by image texture parameters. *Ultrasound Med Biol*, 22(9):1177–1181, 1996.
21. L. Wartfsky and S. H. Ingbar. *Harrisons Principles of Internal Medicine*, chapter Disease of the Thyroid, p. 1712. McGraw-Hill (12th ed), 1991.