ELSEVIER

Image and Vision Computing 19 (2001) 679–690

image AND vision COMPUTING

www.elsevier.com/locate/imavis

Bone tumor segmentation from MR perfusion images with neural networks using multi-scale pharmacokinetic features

A.F. Frangi[a,*], M. Egmont-Petersen[b], W.J. Niessen[a], J.H.C. Reiber[b], M.A. Viergever[a]

[a]*Image Sciences Institute, University Medical Center, Utrecht, Room E01.334, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands*
[b]*Division of Image Processing, Department of Radiology, Leiden University Medical Center, 2300 RC Leiden, The Netherlands*

Received 7 April 2000; accepted 18 December 2000

Abstract

Bone tumor segmentation and the distinction between viable and non-viable tumor tissue is required during the follow-up of chemotherapeutical treatment. Monitoring viable tumor area over time is important in the ongoing assessment of the effect of preoperative chemotherapy. In this paper, features derived from a pharmacokinetic model of tissue perfusion are investigated. A multi-scale analysis of the parametric perfusion images is applied to incorporate contextual information. A feed-forward neural network is proposed to classify pixels into viable, non-viable tumor, and healthy tissue. We elaborate on the design of a cascaded classifier and analyze the contribution of the different features to its performance. Multi-scale blurred versions of the parametric images together with a multi-scale formulation of the local image entropy turned out to be the most relevant features in distinguishing the tissues of interest. We experimented with an architecture consisting of cascaded neural networks to cope with uneven class distributions. The classification of each pixel was obtained by weighting the results of five bagged neural networks with either the mean or median rules. The experiments indicate that both the mean and median rules perform equally well.

Keywords: Bone tumor; Perfusion; Pharmacokinetic modeling; Classification; Classifier combination; Bagging; Cascade; Multi-scale image analysis

1. Introduction

Segmentation of bone tumor entails the distinction between viable and non-viable tumor tissue and is required for the follow-up of chemotherapeutical treatment. Monitoring the volume change of viable tumor can help in assessing the success of chemotherapy, or it may lead to abort its application. It is well known that most malignant tumors are highly vascularized tissues. Distinction between viable and non-viable tumor in post-chemotherapy studies can only be performed by dynamic Magnetic Resonance (MR) perfusion imaging using an intravenous contrast tracer [1,2]. The dynamic MR-signal associated with each physical pixel characterizes perfusion properties of the tissue under study.

In this paper, we present a feature-based neural network for segmentation of dynamic MR-images. A pharmacokinetic model of the tissue perfusion is used to obtain a compact representation of the perfusion properties of each pixel. This allows us to reduce the MR image sequence into three parametric images. Furthermore, concepts of linear scale-space [3,4] are used to analyze the parametric images at several scales in order to incorporate contextual (spatial) information into the segmentation approach. Finally, a classifier is used to combine both temporal and spatial information.

Haring et al. [5] have previously presented a method for image segmentation based on Kohonen networks using multi-scale features. This method, however, performed the multi-scale analysis on the original images since it was devised for static examinations. Our method performs a similar multi-scale analysis but takes advantage of a physically-motivated model of tissue perfusion. Another distinction is that our classifier is based on a supervised feed-forward neural network. Class labels used in the training and testing sets were inferred from post-operative histological studies. Although invasive, this technique is regarded as the gold standard given its high spatial resolution (a histologic slice is about 5 μm thin while a two-dimensional slice in a MR-perfusion acquisition can be 8 mm thick).

Given the uneven distribution of the different tissue types in our application, a two-stage cascaded classifier architecture was designed. A hierarchical classification between healthy and tumor tissue and, within the latter, between viable and non-viable tumor, is thereby obtained. This architecture is inspired by the decision-making process followed by a radiologist when confronted with this type of images.

* Corresponding author.

0262-8856/01/$ - see front matter
PII: S0262-8856(01)00044-0

The paper is structured as follows. In Section 2, multi-scale image features are introduced which are based on a pharmacokinetic model of the perfusion process. Section 3 describes the type of neural network architecture that we have adopted and the quality measures that we have used in designing the classifier. Section 4 describes our experiments in order to design an optimal classifier with our features. Finally, the paper is concluded in Section 5 with a discussion and issues for future research.

2. Methods

Our starting point is an MR image sequence, $s(x,y;t)$, that characterizes the perfusion properties of each pixel, i.e. the up-take and secretion of a blood tracer over time. In the next two sections, we indicate how this image sequence can be reduced to three parametric images summarizing the perfusion properties of the tissue, and how to build a multi-scale feature vector based on the parametric images so as to incorporate spatial information.

2.1. Pharmacokinetic features

The presence of the MR-tracer (e.g. gadopentetate dimeglumine or Gd-DTPA) in tissue causes magnetic field fluctuations which result in shorter relaxation times and consequently a higher intensity of the T1-weighted MR-signal (for details see Ref. [6]). The relation between the concentration of tracer, $\mathscr{C}$, and signal intensity is, in general, approximated by a linear function $s(x,y) = s_0(x,y) + \beta\mathscr{C}$ with $s_0(x,y)$ the signal intensity in the absence of tracer and β a proportionality factor. The exchange of blood (and tracer) between the blood compartment (vessels) and the extracellular water compartment can be characterized by a three-compartment pharmacokinetic model introduced by Tofts [7]. We use a simplification of Tofts' model [8]. The simplified two-compartmental model has three basic parameters: wash-in rate m_1, wash-out rate m_2, and the maximal enhancement a. The concentration of the tracer in blood, $\mathscr{C}_b$, and in the extracellular compartment, $\mathscr{C}_e$, satisfy the following differential equations:

$$V_b \frac{d\mathscr{C}_b}{dt} = -k_1(\mathscr{C}_b - \mathscr{C}_e) - k_2\mathscr{C}_b \tag{1}$$

$$V_e \frac{d\mathscr{C}_e}{dt} = k_1(\mathscr{C}_b - \mathscr{C}_e) \tag{2}$$

where V_b and V_e are the volumes of blood plasma and extracellular water, respectively, k_1 the transfer rate from the blood to the extracellular space, and k_2 the transfer rate of the contrast bolus until renal secretion (normally $k_1 \gg k_2$).

For the extracellular (tumor) compartment, these differential equations yield the following solution [8]

$$\mathscr{C}_e(x,y;t) \propto s_0(x,y) + a(e^{-m_2 t} - e^{-m_1 t}) + \epsilon(x,y) \tag{3}$$

which contains the wash-in rate m_1, and wash-out rate m_2 in the tissue, the maximal enhancement a, and the signal intensity before the tracer has arrived $s_0(x,y)$. The wash-in and wash-out rates can be computed as

$$m_1 = 2\frac{PS}{V_e v_e} \tag{4}$$

$$m_2 = k_1 \frac{V_b + V_e}{V_b V_e} \tag{5}$$

where $0 \le v_e \le 1$ is the fraction of heavily vascularized tissue, P the permeability coefficient between the extracellular and the blood compartments, and S the area of the leaking capillaries [8].

The parameters of the pharmacokinetic model of Eq. (3) can be fitted to the dynamic MR-signal of each pixel $s(x,y;t)$ by minimizing the residual error, $\epsilon(x,y)$, using non-linear regression. The set of values of a particular perfusion parameter corresponding to a 2D dynamic MR sequence, e.g. m_1, can be displayed as a so-called parametric image. Each of the three-perfusion parameters (a, m_1 and m_2) constitutes a feature characterizing the dynamic behavior of the tissues in a pixel.

2.2. Multi-scale spatial features

The pharmacokinetic model presented in the previous section can be seen as a transformation that maps the whole image sequence into three parametric images encoding the tissue perfusion process taking place in the tissue. This transformation is applied on a pixel-by-pixel basis, thereby neglecting any spatial relation.

In the presence of noise, neighboring pixels pertaining to the same tissue type can result in different estimates of the pharmacokinetic parameters. The effect of noise, however, can be reduced by taking into account spatial relations between pixels. One way of introducing spatial relations into the segmentation is to look at the parametric images at multiple scales. This can be done, within the framework of Gaussian scale-space [3], by blurring the image with a Gaussian kernel of increasing standard deviation. Image blurring introduces spatial correlation of the size of the kernel and helps to reduce noise. Since blurring may be seen as a zeroth order Gaussian derivative, one may also consider to incorporate extra, high-order information at each scale [9]. This could be specially useful in classifying boundary pixels which can be located on edges of different strength, depending on the type of tissue transition.

In general, we can represent an intensity function, $I(\mathbf{x})$, in a certain neighborhood of a point $\mathbf{x}_o$ by means of its Taylor expansion. If terms up to the second-order are kept, the intensity function can be locally approximated by

$$\mathscr{I}_\sigma(\mathbf{x}_o + \delta\mathbf{x}) \approx \mathscr{I}_\sigma(\mathbf{x}_o) + \delta\mathbf{x}^T\nabla\mathscr{I}_\sigma(\mathbf{x}_o) + \delta\mathbf{x}^T\mathscr{H}_\sigma(\mathbf{x}_o)\delta\mathbf{x} \tag{6}$$

where $\mathscr{I}_\sigma(\mathbf{x})$, $\nabla\mathscr{I}_\sigma(\mathbf{x})$, and $\mathscr{H}_\sigma(\mathbf{x})$ are the blurred versions of $I(\mathbf{x})$, its gradient vector and its Hessian matrix, respec-

tively. Eq. (6) explicitly states that the Taylor expansion is computed at a given scale, σ. Blurring is performed using an n-D Gaussian kernel

$$\mathscr{I}_\sigma(\mathbf{x}) = I(\mathbf{x}) * G(\mathbf{x};\sigma) \tag{7}$$

$$G(\mathbf{x};\sigma) = \frac{1}{\sqrt{2\pi\sigma^{2^n}}}\, \mathrm{e}^{(-\|\mathbf{x}\|^2/2\sigma^2)} \tag{8}$$

This formulation leads to operators that are regularized [4,10]. For instance, first-order derivatives can be computed by a convolution with the first derivative of the Gaussian kernel

$$\frac{\partial \mathscr{I}_\sigma(\mathbf{x})}{\partial x_i} = I(\mathbf{x}) * \frac{\partial G(\mathbf{x};\sigma)}{\partial x_i} \tag{9}$$

Eq. (6) locally describes the structure of the image up to second-order. For instance, image contrast is accounted for by the first term while the second and third terms account for edgeness and curvedness of the intensity function. In this description, *local* means that we incorporate knowledge of a circular region of approximately the scale of the Gaussian kernel. The differential parameters of Eq. (6) are dependent on the coordinate system that we use for their computation. One can obtain an invariant measure of first- and second-order structure by looking at, e.g., the norm of the gradient and the eigenvalues of the Hessian matrix. These features will be used since they convey the notion of edgeness and curvedness but are invariant up to a rigid transformation (rotation and translation).

Instead of directly using the pharmacokinetic parameters as features for the classifier, we will use a scale-space representation of each of the parametric images. A zeroth order Taylor expansion (only blurred contrast information) will be used for describing the wash-in and wash-out images. The maximal enhancement parametric image is the one providing the richest structural (anatomical) information so higher order information derived from this image could aid in discriminating the tissue type of boundary pixels. To explore the discriminative power of higher order information, we use a second-order Taylor expansion and compute the intensity, norm of the gradient and eigenvalues of the Hessian matrix. All these features are computed at multiple scales.

In order to incorporate *textural* information in the maximal enhancement parametric image, we compute an extra feature based on the entropy of the local image histogram

$$H_a(\mathbf{x},\sigma) = -\sum_{k=0}^{G-1} P(k)\log_2[P(k)]\big|_{\mathscr{N}(\mathbf{x},\sigma)} \tag{10}$$

where $P(k)$ stands for the probability that a pixel attains a value allocated to the kth histogram bin. The local histogram is discretized in G bins dividing the range of intensities in the whole image.[1] In order to reconcile the local nature of this measure with the multi-scale features computed under the Gaussian scale-space paradigm, we have modified the computation of the entropy in the following way: (i) at a given scale, σ, the histogram is computed using an image blurred with a Gaussian kernel of the same scale; (ii) the size of the window, in which the entropy is computed, is coupled to the scale parameter according to width $(\mathscr{N}(\mathbf{x},\sigma)) = \lceil 2\sigma \rceil + 1$ where $\lceil\cdot\rceil$ stands for integer ceiling operation. These two modifications can be interpreted as follows: (i) at a given scale, only details of similar size are kept; and (ii) the approximated support radius of the observational (Gaussian) window is σ.

To summarize, our feature vector entails seven components at each scale, σ:

$$\vec{\mathbf{f}}_\sigma = [a_\sigma, m_{1,\sigma}, m_{2,\sigma}, g_\sigma, \lambda_\sigma^0, \lambda_\sigma^1, h_\sigma] \tag{11}$$

where a_σ, $m_{1,\sigma}$, and $m_{2,\sigma}$ are the intensities of the blurred parametric images, g_σ, is the norm of the gradient of a_σ, and $|\lambda_\sigma^0| \le |\lambda_\sigma^1|$ are the two eigenvalues of the Hessian matrix of a_σ, and h_σ is the entropy of a_σ.

In our experiments, we have used four scales: 1, 2, 4 and 6 mm (resulting in 28 features). The smallest scale is on the order of the inner scale or pixel size in the original images.[2] The largest scale was selected small enough so that the features represent tissue properties and not the properties of the anatomy in which the tissue is embedded. This is important in our application where tumors can be located in various anatomical regions.

3. Tissue classification using neural networks

3.1. Choice of classifier

In principle, a large variety of statistical classifiers are available which could be trained to segment the dynamic MR-images. Statistical classifiers such as Bayes rule, discriminant analysis, k-nearest neighbor and feed-forward neural networks were compared theoretically along several criteria in Ref. [11]. It was argued that if a classifier is required that has a small error rate and is fast in application, a feed-forward neural network is a good choice. It has namely been proven that feed-forward neural networks with one layer of hidden nodes are capable of approximating any continuous function on a (compact) subset of $\Re$ when the number of hidden nodes can grow to infinity [12,13]. Furthermore, it has been shown that the output vector of feed-forward neural networks approximates the Bayes posterior probabilities $P(\omega_j|\mathbf{f})$, $j = 1,\ldots,c$, that the vector $\mathbf{f}$ belongs to the c classes [14].

One disadvantage of neural networks is that several factors besides the size and composition of the training set

[1] In our implementation we have used $G = 64$.

[2] Note, however, that in order to reduce partial volume effects, the computation of all features is performed on a resampled grid of 0.5 mm. See Section 4.2.

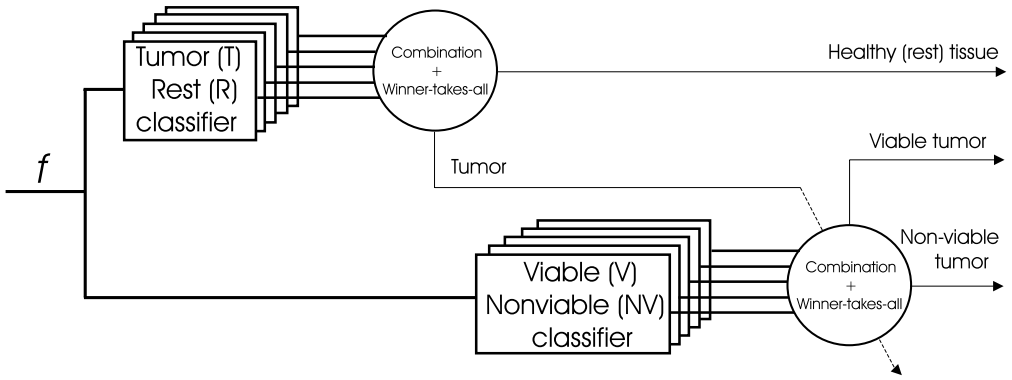

Fig. 1. Overview of the cascade classifier with five combined NN classifiers for each stage. After the parallel classifiers are combined with either the mean or median rule, a winner-takes-all decision is made at each stage. The cascaded approach resembles the decision-making strategy that a human observer would follow: first, a decision is made on whether tumor is present or not; in the presence of tumor, a second step decides whether the tumor is viable or non-viable.

influence the properties of a trained neural network: the number of hidden nodes, the chosen learning algorithm and its parameter settings, and the number of training cycles. Suited learning parameters are usually found during a pilot experiment whereas the optimal network topology needs to be found through further experimentation. The influence of different, randomly chosen initial weight configurations can be ruled out by so-called *bagging* where the classification of each pattern is obtained from a set of neural networks, e.g. by letting the networks vote which class label should be assigned or by summing the output vectors of the neural networks and choosing the class label with the maximal accumulated support. Kittler et al. [15] along with others [16,17], have shown that bagging classifiers leads to a better overall classification result. We have investigated this technique and implemented two alternative rules for combining the outputs of the concurrent classifiers: the *mean* and *median* rules [15].

Another issue known from practical experimentation with neural networks is that a poor performance may be obtained on classes with small prior probabilities. This is especially relevant if more than two classes have to be discerned. To cope with this technical problem, we developed a cascaded architecture. Fig. 1 shows the two-stage cascaded architecture based on feed-forward neural network classifiers. The first stage (TR classifier) has been trained to classify pixels into healthy tissue (+ background) and tumor. Only those pixels classified as tumor are feeded into the second stage (VNV classifier) which then classifies tumor pixels into viable and non-viable. In other words, the output of the VNV classifier is only taken into account when the conditional probability for the tumor class is higher than that for the healthy tissue class. The class-conditional probabilities of the tumor and healthy tissues are approximated by the outputs of the TR classifier. The cascaded approach is inspired by the way that a radiologist would analyze the MR-images. First, the tumor region is delineated and subsequently, this region is subdivided into its viable and non-viable parts. Moreover, a cascaded classifier allows independent assessment and design of each of the two stages [18].

3.2. Performance assessment

Performance assessment of classifiers can be based on quality measures such as correctness, ρ, the kappa statistic, κ, and their class-conditional variants [19]. The κ statistic and the correctness can be computed from the contingency table, $[C_{ij}]$[3], by

$$\kappa = \frac{\rho - p_e}{1 - p_e} \tag{12}$$

$$p_e = \sum_{i=1}^{c} C_{\cdot i} C_{i \cdot} / N^2 \tag{13}$$

$$\rho = \sum_{i=1}^{c} C_{ii} / N \tag{14}$$

where N is the total number of test patterns, and $C_{\cdot i}$ and $C_{i \cdot}$ stand for the ith column and row total sums, respectively. Compared to the *correctness* measure (ρ), the κ statistic is an overall performance measure that accounts for uneven class distributions. This statistic will be used to compare different network configurations. As with respect to the interpretation of the values of the κ statistic one can refer to standard tables provided in statistical books [20]. A value below 0.2 indicates poor agreement while values above 0.2 range from fair to excellent agreement.

3.3. Feature selection

Neural networks are often regarded as black boxes because a trained network is a highly parameterized vector function in which the individual weights — the parameters — do not express distinct stochastic properties of the training cases as, e.g. the estimated mean or variance do in parametric statistics. However, an important way of getting

[3] The contingency table, $\mathbf{C} = [C_{ij}]$, indicates how many samples of the ith class have been assigned to the jth label. Both i and j range from 1 to c, the number of classes.

some insight into the properties of neural networks is by feature selection [18,21].

In our application, we are interested in studying the added value of incorporating multi-scale features compared to single-scale image analysis. Our hypothesis is that multi-scale features would lead to a classifier being both more robust against noise and less dependent on the particular size of the structures under study (e.g. the tumor remnants). Therefore, the goal we pursue is different from more traditional feature selection problems in the sense that we study the decrease in performance when removing *groups* of features from the classifier. The features are divided into three categories: dynamic features ($A = \{a_\sigma, m_{1,\sigma}, m_{2,\sigma}\}$), differential features ($D = \{g_\sigma, \lambda_\sigma^0, \lambda_\sigma^1\}$), and entropy ($E = \{h_\sigma\}$). Subsequently, the performance of the classifier was analyzed for combinations of such sub-sets to gain insight into their discriminative power.

It is well known that statistical classifiers in general are sensitive to *peaking* [22]: excluding features with little discriminative power may lead to an increase in performance on a test set whereas one would expect the opposite or no effect on the generalization performance of the classifier. Previous experiments have, however, indicated that neural networks are insensitive to peaking [23], so including less informative features should not deteriorate their performance. It is not our purpose to investigate in depth any peaking effects. Instead, to get insight in the additional discriminative power of features, we have chosen to perform feature selection according to a backward search where groups of features are alternately excluded.

4. Experiments

4.1. Image acquisition

All MR-examinations were performed on a 0.5 T NT Gyroscan (Philips Medical Systems, Best, The Netherlands) using a surface coil. One, two or three sections were selected for T1-weighted dynamic contrast-enhanced imaging using a magnetization prepared imaging gradient recalled echo technique. The MR-images were acquired with a repetition time of 12 ms, an echo time of 5.7 ms and a prepulse delay time of 741 ms. The chosen flip angle was 30°. The field of view varied per patient depending on the size of the tumor (200–450 mm), and a 256×256 matrix was acquired. The slice thickness was 8 mm. An intravenous bolus injection of 0.2 ml per kg body weight of Gd-DTPA (Magnevist, Schering, Germany) was given followed by a saline flush. For each MR-section, 47 to 60 dynamic images were acquired with a temporal resolution of 3.3 s. Parametric images were computed from this image sequence using non-linear regression (cf. Eq. (3)) based on the Levenberg Marquardt algorithm [24].

All patients in our study had been treated with chemotherapy and underwent surgery a few days after their MR perfusion studies. A tumor resection was carried out, that was sent to histopathological analysis. During this analysis, a so-called macro-slice was made whose position and orientation correspond to that of one of the dynamic MR sections. Furthermore, areas with viable tumor cells were stained with Hematoxylin by a pathologist. Tumor cells bind the purple color pigment in Hematoxylin. Manual delineation of non-viable tumor was also specified by the pathologist. The histologic macro-slice is the gold standard that indicates the correct class label of each pixel in the MR section under study. The areas with viable tumor were delineated by segmenting the purple areas with viable tumor in the histologic macroslice, see Ref. [8].

4.2. Image pre-processing

The signal intensity in MR images is a relative measure for the local relaxivity of the tissue but it does not provide an absolute "unit" that can be used for cross-patient studies. Tissue with similar spin properties that is scanned at different examinations, or corresponding to different patients can have different intensities due to, e.g. differences in the amplifier gain or to the pixel dimensions. Therefore, to achieve a valid classification the features must be independent of such scaling factors. To accomplish this, we make the following observation. In clinical practice, perfusion images are interpreted by comparing the perfusion parameters in the pixels of interest with those of a region of interest in one of the feeding arteries. The perfusion level in those arteries provides an ideal (100%) perfusion. Since feeding arteries have the highest intensities in our MR-images, one is tempted to use the maximum signal intensity to normalize all the pixels. Such a normalization, however, would not be robust to image noise. In order to obtain a robust estimate of the maximum signal intensity, we chose the 95th percentile of the image histogram. Note that we only need to normalize the images of the maximal enhancement parameter — a in Eq. (3). The wash-in and wash-out parameters are independent of intensity rescalings. Finally, we note that if the maximal enhancement image is normalized in this way, all the differential features computed from it will also be normalized. (Fig. 2)

Pixel size can vary slightly across images of different patients.[4] In order to process all the images at a fixed pixel size, each image is initially resampled at 0.5 mm pixel size. This is performed based on *sinc* interpolation in order to reduce partial volume effects [25,26].

Features typically have different means and variances. Features with a large magnitude can overrule the discriminative power of small-magnitude features in neural networks. To avoid this problem and to expedite the training of the neural network, we rescaled all features to a normalized feature space. For each feature, f_i, the mean, $\bar{f}_i$, and standard deviation, SD_{fi}, of that feature was computed for

[4] In our training set, pixel size varied between 0.8 and 1.8 mm.

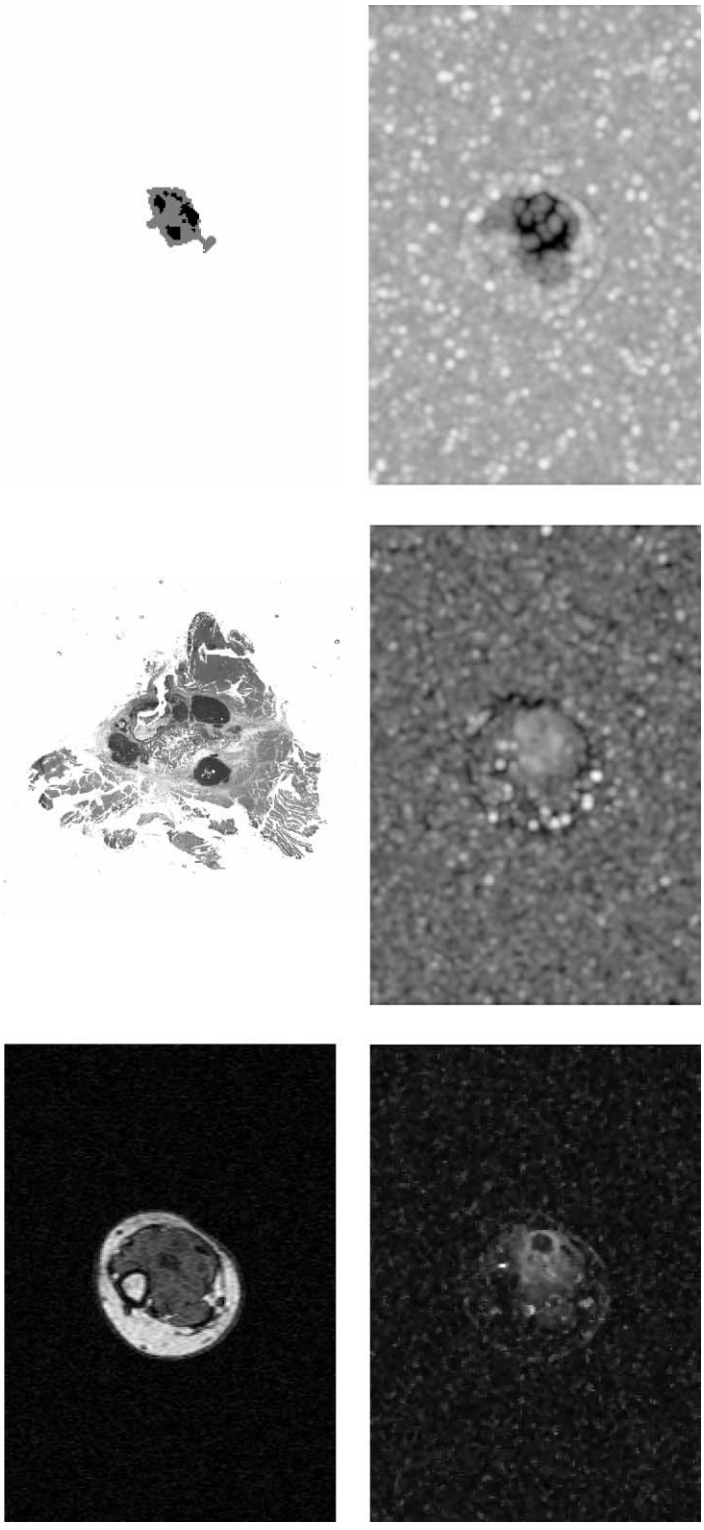

Fig. 2. Image material used in this paper. From left to right and top to bottom: pre-contrast T1-weighted MR image, histologic macro-slice before registration to the pre-contrast image, masks (grey indicates non-viable tumor and black the viable remnants), maximal enhancement (a), wash-in rate (m_1) and wash-out (m_2) rate images. Wash-in/out rate parametric images are displayed in logarithmic scale for improved visualization.

Table 1
NN classifiers with different topologies. TR = tumor-rest classifier, VNV = viable-non-viable classifier. The kappa index, κ, and the correctness measure, ρ, were used to measure the agreement between the output on the test sets and the ground truth. The 95th confidence interval for the κ statistic was smaller than ± 0.012 (TR classifier) and ± 0.05 (VNV classifier) in all experiments. Numbers in bold indicate the topology selected for further experiments

		Number of hidden units											
		1	2	4	8	16	20	24	32	48	64	80	96
TR-mean	κ	0.887	0.922	0.939	**0.935**	0.944	0.952	0.951	0.951	–	–	–	–
	ρ	0.957	0.981	0.977	**0.975**	0.979	0.982	0.982	0.970	–	–	–	–
TR-median	κ	0.893	0.920	0.940	**0.938**	0.940	0.953	0.945	0.951	–	–	–	–
	ρ	0.960	0.981	0.977	**0.976**	0.977	0.982	0.979	0.970	–	–	–	–
VNV-mean	κ	0.481	0.536	0.572	0.580	0.574	0.607	0.604	0.612	**0.622**	0.639	0.610	0.623
	ρ	0.887	0.887	0.906	0.905	0.903	0.911	0.911	0.911	**0.912**	0.914	0.911	0.912
VNV-median	κ	0.520	0.502	0.546	0.578	0.589	0.597	0.604	0.608	**0.608**	0.644	0.608	0.630
	ρ	0.894	0.883	0.901	0.904	0.904	0.904	0.910	0.908	**0.908**	0.915	0.908	0.913

the tumor class (viable and non-viable tumor pooled together) in the training set. Finally, each feature in the feature vector, $\mathbf{f}$, is normalized according to: $f_i^n = (f_i - \bar{f}_i)/\mathrm{SD}_{f_i}$. Notice that the normalization parameters become constants of the classifier.

4.3. Training feed-forward neural network classifiers

Images corresponding to five patients were used to generate independent training and test sets. A sixth patient was kept aside of all experiments as a validation set.

A total of 10 000 patterns (pixels) were selected randomly from the five patients (2000 from each). Another set of 10 000 different patterns were left for testing purposes. Since the pixels belonging to the tumor class (viable and non-viable) are very scarce, they all have been included either in the training or the test sets. In order to train the TR classifier, the patterns of both viable and non-viable tumor were pooled together. When training the VNV classifier, only the tumor patterns were presented for both training and testing. Notice that this step is justified if the first classifier is very reliable (as shown in the examples). In this case, the second classifier will only have to know how to classify pixels that are belonging to the tumor class. Finally, both training and test patterns were randomized to eliminate any undesired stratification per patient.

After random initialization of the weights, the training of all neural networks was performed by the Scaled Conjugate Gradient[5] algorithm by Moller [27] using the Stuttgart Neural Network Simulator [28] (SNNS V 4.2). From pilot experiments, we conclude that this learning algorithm converges with a faster pace to a low mean square error (MSE) than standard back-propagation and that it is quite insensitive to the learning parameters. All networks were trained until either the relative change or the MSE was smaller than 0.1% over the last 100 iterations or a maximum of 5000 learning cycles was reached. Only occasionally, some of the networks converged to a poor local minimum (abnormally large final MSE). In those cases the weights were reinitialized and the network was retrained.

For each stage of the cascade (TR and VNV classifiers), five neural networks with equal topology were trained per experiment. The output vector of the five classifiers was combined according to the mean and median rules [15]. Final decision on the class label is made by applying a winner-takes-all rule on the combined output vector.

4.4. Selecting network topology

The first step is the selection of an optimal topology (number of hidden neurons) for the feed-forward neural network classifiers composing the first (TR) and second (VNV) stage of the cascade classifier. Table 1 shows the results of the experiments for the test set. The κ statistic and correctness indicate the performance of each stage separately. The first classifier (TR) reaches a very good performance ($\kappa > 0.92$) even for small a number of hidden neurons. With the only exception of the topology with only one hidden neuron, an increase in network complexity produces only a marginal increase in performance. We found eight hidden neurons a good compromise between complexity and performance. This topology for the TR classifier was used for further experiments. The second (VNV) classifier has a lower performance indicating the more complex nature of the classification task. We chose a topology with 48 hidden neurons for further experiments.

An interesting finding inferred from Table 1 is that no statistically significant difference (χ^2test, $p > 0.5$) in performance was found, in general, for the two combination rules of the combined classifiers at each stage. Another observation is that the correctness measure is less sensitive to improvements than the κ statistic. This is due to the uneven distribution of the classes. Therefore, a small improvement in classifying correctly viable tumor patterns affects the correctness, ρ, only marginally, which is clearly captured by the κ statistic.

[5] Learning parameters were the defaults provided by the SNN simulator: $\sigma_1 = 10^{-4}$, $\lambda_1 = 10^{-6}$, $\Delta_{\max} = 0.2$.

Table 2
Multi-scale versus single-scale features. The κ index, and the correctness measure, ρ, were used to measure the agreement between the output of the test sets and ground truth. The 95th confidence intervals for the κ statistic were smaller than ± 0.01 (TR classifier) and ± 0.06 (VNV classifier) in all experiments

		Scales (mm)							
		1	2	4	6	1 + 2	1 + 4	2 + 4	All
TR-mean	κ	0.859	0.909	0.929	0.907	0.914	0.941	0.956	**0.935**
	ρ	0.947	0.966	0.973	0.965	0.968	0.978	0.984	**0.975**
TR-median	κ	0.862	0.906	0.933	0.904	0.916	0.939	0.951	**0.938**
	ρ	0.948	0.965	0.975	0.964	0.969	0.977	0.982	**0.976**
VNV-mean	κ	0.403	0.530	0.531	0.493	0.554	0.594	0.563	**0.612**
	ρ	0.864	0.891	0.897	0.888	0.898	0.905	0.901	**0.911**
VNV-median	κ	0.428	0.532	0.548	0.508	0.559	0.598	0.556	**0.608**
	ρ	0.867	0.890	0.902	0.890	0.898	0.904	0.900	**0.908**

4.5. Single versus multi-scale analysis

Table 2 summarizes the performance of the TR (8 hidden neurons) and VNV (48 hidden neurons) classifiers when trained with all features computed at a single scale. We also computed the performance for some representative combinations of those scales. As was to be expected, working at a scale close to the pixel size is the least reliable. In fact, this is almost equivalent to working with no spatial information and noise will have a strong influence on the results. Increasing the scale does not always improve the results since too much blurring will lead to a poor localization of objects or will simply result in the disappearance of tumor remnants. It was found that if all scales are used, performance is better than when using information at a single scale.

In the first classifier, an optimal performance can be obtained for the combination of features at 2 and 4 mm. The difference with using all scales is small (although statistically significant, χ^2 test, $p < 0.05$). This is probably an example of peaking. Reducing the number of scales, however, would only lead to a slightly better performance and no reduction in computational cost is gained since all scales are required to obtain an optimal classification by the second classifier. We therefore opted for keeping all scales in both classifiers. These experiments confirm again that there is no preferred option regarding the combination rule (no significant statistical difference in performance, χ^2 test, $p > 0.5$). Both mean and median combination rules led to the same performance.

4.6. Assessing the contribution of different feature groups

In order to obtain a better understanding of the discriminative power of our features, we pursued an analysis similar to that described in the previous section, but now grouping the features by their type. We subdivided the features into three groups containing the spatio-temporal zeroth order features, $A = \{a_\sigma, m_{1,\sigma}, m_{2,\sigma}\}$, the first and second order features, $D = \{g_\sigma, \lambda_\sigma^0, \lambda_\sigma^1\}$, and the textural feature, $E = \{h_\sigma\}$. Subsequently, neural networks were trained with each sub-set and combinations thereof. Following the conclusions in the previous two subsections, all scales were used for each feature and the topology of the two stages was eight hidden neurons (TR) and 48 hidden neurons (VNV).

Table 3 shows the results of the experiments. For both stages (TR and VNV), most of the discriminative power is

Table 3
Discriminative power of different feature groups: blurred amplitude of maximal enhancement, and wash-in/out parametric images (A), norm of gradient and Hessian eigenvalues of maximal enhancement image (D), and entropy of maximal enhancement image (E). The kappa index was used to measure the agreement between the classification results on the test set as compared with the ground truth. The 95th percentile confidence interval for the κ statistic was smaller than ± 0.015 (TR classifier) and ± 0.05 (VNV classifier) in all experiments

		Groups of features						
		A	D	E	$A + D$	$A + E$	$D + E$	$A + D + E$
TR-mean	κ	0.934	0.757	0.664	0.938	**0.953**	0.842	0.935
	ρ	0.975	0.911	0.876	0.977	**0.982**	0.940	0.975
TR-median	κ	0.928	0.786	0.740	0.935	**0.943**	0.836	0.938
	ρ	0.978	0.920	0.899	0.976	**0.978**	0.938	0.976
VNV-mean	κ	**0.648**	0.465	0.189	0.574	0.599	0.520	0.612
	ρ	**0.916**	0.877	0.832	0.904	0.907	0.891	0.911
VNV-median	κ	**0.652**	0.457	0.178	0.574	0.586	0.521	0.608
	ρ	**0.916**	0.872	0.837	0.902	0.903	0.888	0.908

provided by the sub-set A (amplitudes of the pharmacokinetic parameters at multiple scales). This is particularly clear for the second classifier where adding the other subsets only leads to a decrease in performance (peaking). In the TR stage, addition of textural information (E) improves the classification.

From the previous experiments, we conclude that exclusion of the differential features can reduce the computational cost without compromising the performance of any of the two stages. Therefore, the optimal classifier will contain a first stage using the sets A and E, and a second stage with only amplitude features (A). Note that the cascade architecture allows selective exclusion of features at each stage.

4.7. Overall classifier performance and viable tumor area assessment

So far, we have analyzed the performance of the two classifiers independently. This has led to an "optimal" design (within our feature space) for each stage of the cascade. In this section, we report the κ statistic for the three-class cascade classifier on each of the MR-images of the five patients (I–V) analyzed. Additionally, we applied the classifier to a sixth patient (VI) whose patterns had neither been used for training nor testing.

Although the overall kappa is a good quality measure for assessing global performance of the three-class classifier, it is interesting to analyze the performance of the classifier in terms of the less probable class, viz. viable tumor. This class is of utmost clinical importance since its monitoring over time (i.e. after successive chemotherapy sessions) indicates the progress of the treatment.

It is possible to compute a class-conditional κ statistic, which is obtained by collapsing the 3×3 contingency table into a 2×2 (non-viable tumor + rest, pooled together) and using the formulas [12–14,29]. Table 4 compares the performance (κ_{ov}) of the optimal cascaded classifier (cf. Section 4.6) in the *whole* image for each patient. Overall label agreements fall in the range between fair and good [20]. Similar results are obtained for the viable tumor class-conditional agreement (κ_v). Fig. 3 shows the results of three patients (II, V and VI) together with the mask labels and the corresponding pre-contrast MR images.

The overall agreement is moderate for the validation patient (VI). However, the class-conditional κ statistic reflects low agreement for viable tumor. Although a thorough clinical evaluation still has to be performed, these preliminary experiments seem to suggest that this performance is correlated with the size and spatial distribution of the tumor remnants. Apparently, remnants with a section smaller than 100 mm^2 and/or sparsely distributed remnants are prone to area overestimation. Detection of such remnants of viable tumor is likely to be strongly affected by partial volume effects.

5. Discussion and conclusions

In this paper, a multi-scale feature-based neural network classifier for automatic segmentation of MR perfusion images of bone tumor is presented. The method classifies each pixel as either viable tumor, non-viable tumor or "rest" (healthy + background) tissue.

Features are computed in two steps. First, a pharmacokinetic model is used to summarize the temporal information in the perfusion sequence into three main parameters (a, m_1 and m_2). Subsequently, these parametric images are used to derive multi-scale features of contrast, edgeness, curvedness and texture. Given the (multi-scale) local support of these features, they encode spatial information not present in the original parametric images (which themselves contain solely temporal information).

Our approach indicates that neural networks are suitable for combining several types of information, in our case dynamic and spatial properties of tissue perfusion. By using techniques from feature selection, one is able to identify the best (sub)set of scales and features for a particular application at hand. However, our experiments with feature selection all indicate that peaking occurs when redundant features are being removed from the classifier. So the results of Hamamoto et al. that neural networks are insensitive to peaking could not be confirmed by our experiments [23]. Probably, peaking is caused by the high amount of correlation between the redundant features that are

Table 4
Optimal cascaded classifier performance for the MR-images corresponding to each of the six patients. Overall κ statistic, class-conditional κ for viable tumor and viable tumor area estimates. Viable tumor distribution was classified in three categories: compact (one nucleus), sparse (many small spots), and conglomerate (two or three compact regions). Patient I is completely cured according to the histological study. Mean/median refers to the type of combination rule

Patient	κ_{ov} (Mean/median)	κ_v (Mean/median)	A (mm^2)	$\hat{A}$ (mm^2) (Mean/median)	Slice orientation	Remnant distribution
I	0.38/0.34	0.00/0.00	0.0	12.4/11.0	Transversal	–
II	0.78/0.75	0.88/0.88	144.6	142.7/142.7	Sagittal	Compact
III	0.75/0.72	0.70/0.69	447.8	461.5/490.4	Coronal	Conglomerate
IV	0.42/0.31	0.36/0.32	95.8	228.7/272.0	Coronal	Sparse
V	0.66/0.64	0.72/0.70	247.2	254.1/259.6	Transversal	Conglomerate
VI	0.56/0.49	0.11/0.10	77.2	122.8/132.5	Sagittal	Sparse

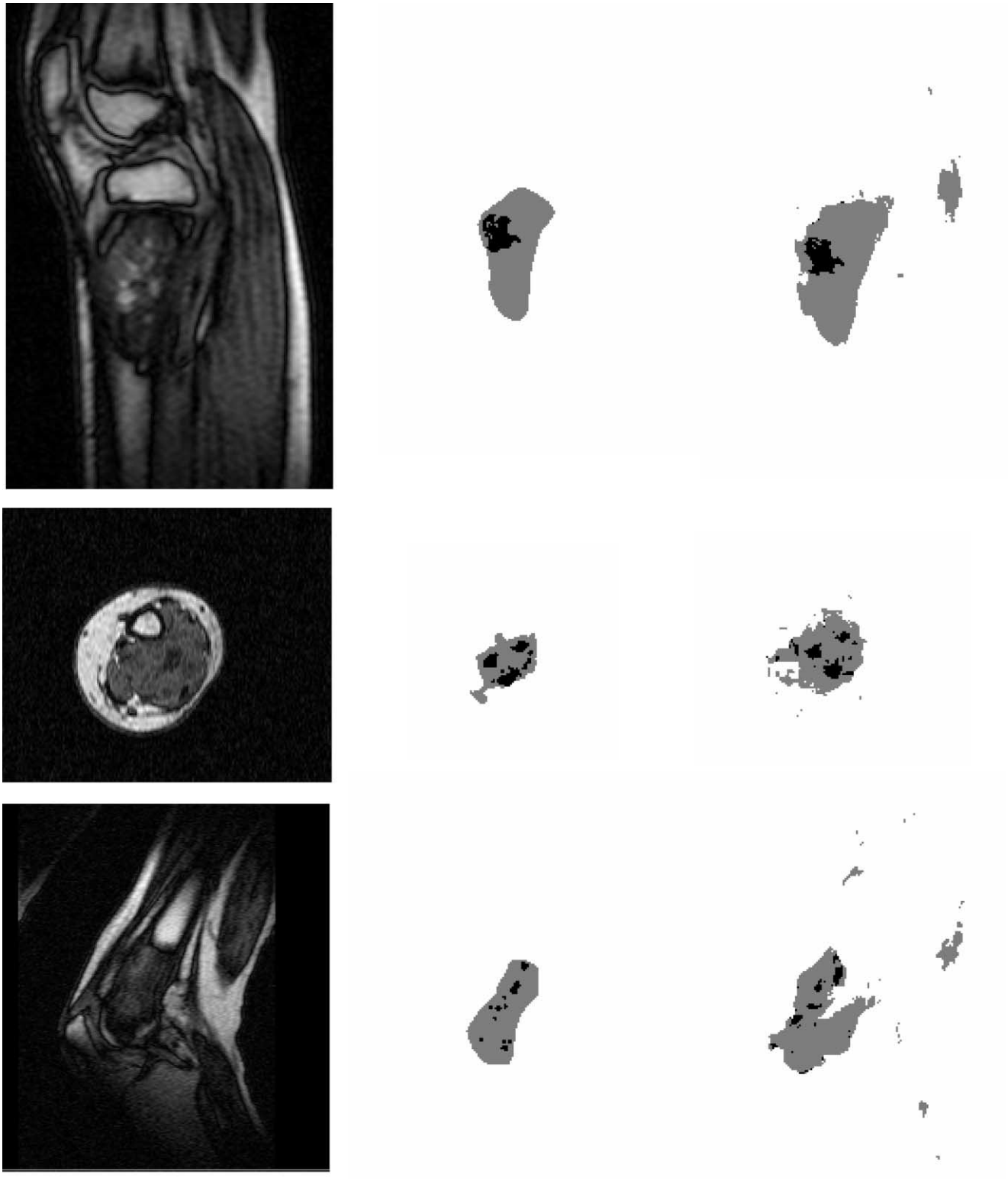

Fig. 3. Segmentation results. Left, pre-contrast images; center, histological masks; and right, neural network classification. First row corresponds to patient II, second row to patient V, and last row to validation patient VI.

removed (i.e. the same spatial derivatives computed at slightly different scales). Clearly, the possibility of peaking should be taken into account when neural networks are trained with sets of highly correlated, possibly redundant features.

A cascaded classifier architecture was implemented, which allows for flexible design of each individual stage. The κ statistic was used as a quality measure for the design of each stage of the classifier. This measure has the property

of being more sensitive to misclassification of patterns belonging to less probable classes compared to measures like the correctness. The use of this quality measure has guided the design of the optimal topology of the classifier and the assessment of how important different groups of features are. It was found that multi-scale analysis, which adds contextual information, leads to a better performance than single-scale analysis.

For each stage in the cascade, a combined classifier with five concurrent networks was implemented. Two different combination rules for the concurrent output vectors were compared: the mean and median rules. In general, there was no statistically significant difference in performance between them.

The results obtained in this work are encouraging since they represent a worst-case condition (large variability in imaging protocol). Moreover, moderate to good performance was obtained in segmenting tumor tissue from healthy tissue. For medium to large, compact remnants, estimation of viable-tumor area could be done with an average relative error of 3.0% (SD 4.1%) (patients II, III and V for both combination rules). However, distinction of viable tumor is not very accurate in the presence of small or very sparse remnants (cf. Table 4). Performance is in such cases deteriorated by a combination of partial volume effects and a possible discrepancy in orientation between the MR-section and the histologic macro-slice. Our training set was based on five patients who were examined in different anatomical regions, and with different slice orientations and pixel size. We expect that by using a fixed imaging protocol for different anatomical locations, anatomy-specific classifiers could be designed using the guidelines presented in this paper. This will lead to better classification performance for images acquired under the same protocol. Furthermore, the performance in classification of small and/or sparse tumor remnants may be improved using a higher resolution MR image acquisition. Confirmation of this hypothesis is a topic of future investigation.

Currently, patients are being scanned according to a 3D MR perfusion protocol. Extension of the method to three-dimensions is straightforward since all features are readily computed in n-D.

A thorough clinical evaluation is still required to establish the accuracy of viable tumor area estimates, the parameter of clinical relevance. Inclusion of more patients will allow to estimate more accurately the margins in tumor area under which the method can be regarded to be reliable.

Acknowledgements

We are grateful to Prof Dr J.L. Bloem (Department of Radiology) and Prof Dr P.C.W. Hogendoorn (Department of Pathology), both from the Leiden University Medical Center, for providing us with the patient material. This project was financially supported by a the Dutch Cancer Foundation (KWF), Grant RUL 97-1509 and by the Dutch Ministry of Economic Affairs, IOP Beeldverwerking Grant IBV-97009.

References

[1] H. van der Woude, J. Bloem, H. Holscher, N.M.A. A, H. Taminiau, J. Hermans, P.C. Falke, T.H. Hogendoorn, Monitoring the effect of chemotherapy in ewing's sarcoma of bone with MR imaging, Skeletal. Radiol. 23 (7) (1994) 493–500.

[2] H. van der Woude, J. Bloem, K. Verstraete, A. Taminiau, M. Nooy, P. Hogendoorn, Osteosarcoma and ewing's sarcoma after neoadjuvant chemotherapy: value of dynamic MR imaging in detecting viable tumor before surgery, Am. J. Roentgenol. 165 (3) (1995) 593–598.

[3] L.M.J. Florack, B.M. ter Haar Romeny, J.J. Koenderink, M.A. Viergever, Linear scale-space, J. Math. Image Vis. 4 (4) (1994) 325–351.

[4] J.J. Koenderink, The structure of images, Biol. Cybern. 50 (5) (1984) 363–370.

[5] S. Haring, M.A. Viergever, J.N. Kok, Kohonen networks for multiscale image segmentation, Image Vis. Comput. 12 (6) (1994) 339–344.

[6] U. Hoffmann, G. Brix, M. Knopp, T. Hess, W. Lorenz, Pharmacokinetic mapping of the breast: a new method for dynamic MR mammography, Magn. Reson. Med. 33 (4) (1994) 506–514.

[7] P.S. Tofts, B. Berkowitz, M.D. Schnall, Quantitative analysis of dynamic Gd-DTPA enhancement in breast tumours using a permeability model, Magn. Reson. Med. 33 (1995) 564–568.

[8] M. Egmont-Petersen, P.C.W. Hogendoorn, R. van der Geest, H.A. Vrooman, H.-J. van der Woude, J.P. Janssen, J.L. Bloem, J.H.C. Reiber, Detection of areas with viable remnant tumor in postchemotherapy patients with Ewing's sarcoma by dynamic constrast-enhanced MRI using pharmacokinetic modeling, Magn. Reson. Imaging 18 (5) (2000) 525–535.

[9] L.M.J. Florack, B.M.t Haar Romeny, J.J. Koenderink, M.A. Viergever, The Gaussian scale-space paradigm and the multiscale local jet, Int. J. Comput. Vis. 18 (1) (1996) 61–75.

[10] L.M.J. Florack, B.M. ter Haar Romeny, J.J. Koenderink, M.A. Viergever, Scale and the differential structure of images, Image Vis. Comput. 10 (6) (1992) 376–388.

[11] M. Egmont-Petersen, E. Pelikan, Detection of bone tumors in radiographic images using neural networks, Pattern Anal. & Appl. 2 (2) (1999) 172–183.

[12] K.-I. Funahashi, On the approximate realization of continuous mappings by neural networks, Neural Netw. 2 (3) (1989) 183–192.

[13] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Netw. 2 (1989) 359–366.

[14] M.D. Richard, R.P. Lippmann, Neural network classifiers estimate bayesian a posteriori probabilities, Neural Comput. 3 (4) (1991) 461–483.

[15] J. Kittler, M. Hatef, R. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Machine Intell. 20 (3) (1998) 226–239.

[16] L. Lam, C. Suen, Optimal combinations of pattern classifiers, Pattern Recogn. Lett. 16 (9) (1995) 945–954.

[17] A. Verikas, A. Lipnickas, K. Malmqvist, M. Bacauskiene, A. Gelzinis, Soft combination of neural classifiers: a comparative study, Pattern Recogn. Lett. 20 (4) (1999) 429–444.

[18] M. Egmont-Petersen, W. Dassen, J. Reiber, Sequential selection of discrete features for neural networks — a bayesian approach to building a cascade, Pattern Recogn. Lett. 20 (11-13) (1999) 1439–1448.

[19] M. Egmont-Petersen, J. Talmon, J. Brender, P. McNair, On the quality of neural net classifiers, Artif. Intell. Med. 6 (5) (1994) 359–381.

[20] D. Altman, Practical Statistics for Medical Research, Chapman & Hall, London, 1991.

[21] M. Egmont-Petersen, J. Talmon, A. Hasman, A. Ambergen,

Assessing the importance of features for multi-layer perceptrons, Neural Netw. 11 (4) (1998) 623–635.
[22] G. Huges, On the mean accuracy of statistical pattern recognizers, IEEE Trans. Inform. Theory 14 (1968) 55–63.
[23] Y. Hamamoto, S. Uchimura, S. Tomita, On the behavior of artificial neural network classifiers in high-dimensional spaces, IEEE Trans. Pattern Anal. Machine Intell. 18 (5) (1996) 571–574.
[24] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C: The art of Scientific Computing, 2nd ed., Cambridge University Press, Cambridge, 1992.
[25] Y.P. Du, D.L. Parker, W.L. Davis, G.C. Cao, Reduction of partial volume artifacts with zero-filled interpolation in three-dimensional MR angiography, J. Magn. Reson. Imaging 4 (5) (1994) 733–741.
[26] D.L. Parker, Y.P. Du, W.L. Davis, The voxel sensitivity function in fourier transform imaging: applications to magnetic resonance angiography, Magn. Reson. Med. 33 (2) (1995) 156–162.
[27] M.F. Moller, A scaled conjugate gradient algorithm for fast supervised learning, Neural Netw. 6 (1993) 525–533.
[28] A. Zell, N. Mache, T. Sommer, T. Korb, The SNNS Neural Network Simulator, GWAI-91: Fachtagung für künstliche Intelligenz, Springer, Berlin, 1991 (pp. 254–263).
[29] J.L. Fleiss, Statistical Methods for Rates and Proportions, Wiley Series in probability and mathematical statistics, 2nd ed., Wiley, New York, 1981.