Copyright © 1999 Springer Verlag.
Reprinted from (*Pattern Analysis and Applications*, M. Egmont-Petersen, E. Pelikan. "Detection of bone tumours in radiographs using neural networks," Vol. 2, No. 2, pp. 172-183, 1999, Copyright Springer Verlag), with permission from Springer Verlag.

This material is posted here with permission of Springer Verlag. Single copies of this article can be downloaded and printed for the reader's personal research and study.

For more information, see the Homepage of the journal Pattern Analysis and Applications:

http://www.dcs.ex.ac.uk/paa

or Springer Verlag

http://www.springer.de

Comments and questions can be sent to: michael@cs.uu.nl

Pattern Analysis & Applications (1999)2:172–183
© 1999 Springer-Verlag London Limited

Detection of Bone Tumours in Radiographic Images using Neural Networks

M. Egmont-Petersen[1] **and E. Pelikan**[2]

[1]*Division of Image Processing, Department of Radiology, Leiden University Medical Centre, Leiden, The Netherlands;* [2]*Scientific Technical Department, Philips Medical Systems, Hamburg, Germany*

Abstract: We develop an approach for segmenting radiographic images of focal bone lesions possibly caused by bone tumour. A neural network is used to classify individual pixels by a convolution operation based on a feature vector. We design eight features that characterise the local texture in the neighbourhood of a pixel. Four of the features are based on co-occurrence matrices computed from the neighbourhood. The true class label of the pixels in the radiographs are obtained from annotations made by an experienced radiologist. Neural networks and self-organising feature maps are trained to perform the segmentation task. The experiments confirm the feasibility of using a feature-based neural network for finding pathologic bone changes in radiographic images. An analysis of the eight features indicates that the presence of edges and transitions, the complexity of the texture, as well as the amount of high frequencies in the texture, are the main features discriminating (soft) tissue from pathologic bone, the two classes most likely to be confused.

Keywords: Bone tumours; Feature selection; Feed-forward neural network; Quality assessment; Self-organising feature map; Texture

1. INTRODUCTION

Neural networks have been developed for various classification tasks in image processing and computer vision. Among others, neural networks were trained to diagnose malignant melanoma [1] and to perform optical character recognition [2–6]. An optical network has been built for face recognition [7], a (hardware) RAM neural network has been trained for classification of danger labels [8], and neural networks have been trained for recognition of marker points in remote sensing images [9,10].

In our line of research, we are developing an approach for automatically screening radiographs for focal bone lesions, in particular bone tumours. These tumours constitute less than one percent of all tumours in Germany [11]. Because of their low incidence, a radiologist in a nonspecialised clinic will typically see a radiograph with indications of a bone tumour two or three times in his/her career [12]. Until now, relatively little research has been done to automatically locate morphological structures in radiographs of bone tumours (see Sorgel et al [13]). Therefore, we aim at developing an approach that can be used for screening radiographs for pathologic bone changes that are caused by bone tumours or other bone lesions with a similar appearance.

In this paper, our objective is to investigate the feasibility of automatically identifying pathologic bone changes in radiographs but not to discriminate between the several types of bone tumours. The main feature discriminating normal from pathologic bone is the local textural appearance. We therefore developed a segmentation approach based on local features that are invariant to translation and rotation. The image is segmented by two statistical classifiers – a feed-forward neural network and a self-organising feature map – into four categories. The classifiers obtain as input a set of features that characterises textural properties of a local neighbourhood around a pixel.

This paper is organised as follows. First, the medical background is briefly discussed. In Section 3, we compare different types of classifiers and motivate why we chose the neural network and the self-organising feature map for the segmentation task. We also introduce the translation- and rotation-invariant features. It is then described how the neural network and the self-organising feature map are trained to perform the segmentation task. Subsequently, the results are presented and discussed.

Received: 4 June 1998
Received in revised form: 15 September 1998
Accepted: 11 December 1998

2. MEDICAL BACKGROUND

Once a focal bone lesion is detected, highly specialised radiologists establish the differential diagnosis by looking for the presence and absence of several so-called morphological structures. In the 1960s, Lodwick [14] made a taxonomy of focal bone lesions (including bone tumours), and characterised different morphological structures which indicate pathologic bone changes. Since then, others have extended the framework proposed by Lodwick [15,16]. Besides age, some questions address the location of the lesion and the type of bone. Based on this information, the specialised radiologist establishes the correct differential diagnosis among the 42 possibilities. However, to a radiologist who is not specialised in the diagnosis of bone tumours, the morphological structures associated with bone tumours appear as unexpected patterns – abnormal textures – in the radiograph. Moreover, nonspecialised radiologists are normally unaware of the relation between age, the appearance of the unexpected textural patterns and the relation with specific bone tumours. As we focus on the screening application, our task is to aid the nonspecialised radiologist in locating pathologic bone changes. More specifically, we want to segment the pixels in a radiograph into one of the four categories: *healthy bone*, *soft tissue*, *pathologic bone* and *image background*. In our application, a segmentation approach is needed that is invariant to changes in: translation, rotation and image contrast. Variations in scale are small in our image material.

3. SEGMENTATION APPROACH

Define an image as a two-dimensional function[1] $f(x,y)$ of the coordinates $x \in 1,\ldots,x_{\max}$ and $y \in 1,\ldots,y_{\max}$. Segmentation entails partitioning an image into regions that are coherent with respect to some criterion. Image segmentation can also be seen as a classification task in which each pixel is assigned to one among a number of categories. We decided to train a statistical classifier to perform the segmentation task. Two important choices needed to be made: which type of statistical classifier to use and whether to apply the classifier directly on the image data or on a set of derived features.

3.1. Choosing a Classifier

Classification entails assigning a class label to an object (pixel) based on an n-dimensional feature vector $\boldsymbol{e}$. A statistical classifier implements a mapping from the n-dimensional feature space to the c-dimensional probability class space

$$h: \boldsymbol{e} \in \mathcal{R}^n \rightarrow \boldsymbol{z} \in (0,1)^c \quad (1)$$

Let $p(\boldsymbol{e}|\omega_j)$ denote the n-dimensional class-conditional Probability Density Function (PDF) of the n features for class j, $j = 1,\ldots,c$. In general, classifiers partition the feature space into disjoint regions R_j, $j = 1,\ldots,c$.

A variety of statistical pattern classifiers exist: the Bayesian (minimal error rate) classifier, the k-nearest neighbour classifier, linear and quadratic discriminants, feed-forward neural networks (MLP), the self-organising feature map (SOM), induction trees, etc. In choosing among these classifiers for our application, the error rate obtained is solely one among several criteria that should be taken into account. It is important to make the best trade-off between error rate, the computational cost associated with building and applying the classifier, how much computer memory it requires as well as the transparency of the classifier to developers and end users (for a discussion see Sklansky and Vriesenga [10]). The criteria are listed in the top row of Table 1.

Although the Bayesian classifier guarantees the minimal error rate, one needs to know the type of distribution of the features to build it. A possible approximation to the class-conditional densities of the features is to use kernel functions [17]. When the class-conditional PDFs of the features can be characterised by normal densities, however, linear or quadratic discriminants will result in Bayes classifiers [18]. The asymptotic error rate of the nonparametric k-nearest neighbour classifier is bounded by about twice the Bayes error rate [19]. For feed-forward neural networks with one hidden layer, it has been proven that they can implement any discriminant when provided with enough hidden nodes [20,21]. When the size of the training set goes to infinity, the output of a network with a sufficient number of hidden nodes approaches the Bayesian posterior probabilities [22]. For the self-organising feature map, it has been shown that the weight vectors connecting the inputs with the nodes in the map specify the centres of clusters that cover the input space and the point density function of these centres tends to approximate the probability density function of the feature space [23]. Also the induction trees built with NPPA [24] or ID3 [25] are capable of modelling any discriminant when no limits are posed on the number of branches, i.e. the depth of the trees.

Building a Bayesian classifier with kernel densities requires much memory and is moderately complex to use. Relatively little computation and computer memory are needed to build, use and store linear and quadratic discriminants. The k-nearest neighbour classifier is simple to build from a training set but requires immense computation and a large memory when used for classifying cases (although algorithms that speed up the search process have been developed). Building feed-forward neural networks and self-organising feature maps, on the other hand, is computationally complex because convergence of the learning algorithms is often slow. However, once optimally tuned, applying these classifiers requires little computation and computer memory. Induction algorithms like NPPA and ID3, by virtue of their recursive structure, require little computation in use. Because induction trees partition the feature space along the n axes that span the feature space, such trees may contain many branches which again requires much computer memory.

[1] Henceforward, a capital letter X denotes a matrix, a bold letter $\mathbf{y}$ a column vector. $x_{k,i}$ denotes the kth element in column i in X. The ith element in vector $\mathbf{y}$ is denoted by y_i. A function is in the main text rendered by $f(\cdot)$. Finally, $p(x)$ denotes the probability density function of variable x.

Table 1. The criteria along which the different statistical classifiers are compared

Classifier	Error rate	Comp. compl. (training)	Comp. complexity (use)	Required memory (use)	Transparency
Bayes (kernels)	Asymp. minimal	Small	Moderate	Very large memory	Not transparent
Lin./quad. discriminant	Minimal (Gauss. distrib.)	Small	Very small	Small memory	Transparent
k-near. neigh	Two times minimal	Very small	Very complex	Very large memory	Transparent
MLP	Asym. minimal	High	Small	Small memory	Not transparent
SOM	Unknown	High	Small	Small memory	Not transparent
Induc. trees	Unknown	Moderate	Small	Small to large memory (probl. dep.)	Transparent (but complex, probl. dep.)

As with respect to transparency, a Bayesian classifier based on kernels is complex to comprehend. The functioning of linear and quadratic discriminants, on the other hand, is well-understood. Also, the *k*-nearest neighbour classifier is transparent in the sense that the *k* nearest training examples in relation to a vector can be recalled upon request and displayed to the user. The feed-forward neural network, however, is often regarded as a black box [26,27], albeit techniques exist which can map the net to a more transparent classifier [27,28] as well as techniques that characterise certain properties of the network explicitly, e.g. by estimating the contribution of the input nodes to the classification of a case [29], computing a quality profile of a neural network [30] or estimating the importance of the features used by the neural network [31].

In digital image processing, a high performance and a fast computation when applying the classifier are both mandatory requirements. Our segmentation task is complex and a classifier that can approximate any discriminant function is needed. Moreover, since trained neural networks require only little memory and are efficient in use, we chose to train feed-forward neural networks to perform the segmentation task. We also wanted to investigate whether an unsupervised clustering algorithm would result in the same segmentation result as the (supervised) feed-forward neural network. Therefore, a self-organising feature map was trained and each node in the map associated with a particular class after training had terminated.

3.2. Feature Space

After having chosen the classifiers that will be used, we need to specify which information should be provided as input. The morphological structures we seek are small compared to the size of a whole radiograph. The smallest morphological structure that had been demarcated by the involved radiologist had an area of about 25 pixels. As (local) texture is the most important salient feature distinguishing pathologic bone from the other three classes, segmentation could, for instance, be performed by convolving the radiographic image with a neural network that obtains as input the intensities of a quadratic window with, for example, 5×5 pixels. As in our application the network should be invariant to rotation, the training images would then have to been rotated by a randomly chosen multiple of θ degrees (e.g. $\theta = 360/12$). Such a rotation inevitably entails bi-linear interpolation and the resulting texture becomes smoothed. Consequently, the amount of high frequencies with a possible discriminative power is reduced which might lead to a decrease in performance. To avoid this problem, we experimented with different rotation-invariant features that were provided as input to the neural network in the form of a feature vector $\boldsymbol{e}(x,y)$.

It is important to choose features that result in a good discrimination between the four classes. In the following, we define eight features that characterise the texture of a small neighbourhood around a pixel [32]. Each feature $o \in \{1,\ldots,8\}$ can be seen as an operation on the original image, $O(f(x,y))$. More textural features such as gradient and edge operators had been evaluated in preliminary experiments [33–35]. However, the best performance was obtained by combining the features presented here.

The radiographs we work with vary in contrast and brightness. To provide the neural network with an absolute indication of the grey level of a pixel, we included the feature $O_1(f(x, y))$ which performs a histogram equalisation of the image

$$\eta = \mathbf{g}[f(x,y)], \quad \mathbf{g}[i] = \left(\frac{\sum_{(x',y')} f(x',y') \leq i}{\sum_{(x',y')} f(x',y')} 255 \right) \tag{2}$$

with $\mathbf{g}$ denoting the new look-up table. η is the only global feature provided to the classifier. Two other features that contributed to the discrimination between the four classes in earlier experiments [34] were O_2 unsharp masking (7×7 kernel) and O_3 a median filter (7×7 kernel). Unsharp masking is defined as

$$\nu = f(x,y) + \{f(x,y) - \mu[f,(x,y)]\} \tag{3}$$

with $\mu[f,(x,y)]$ denoting the median filter applied on a 7×7 window from f and (x,y) the coordinates of its central pixel. Unsharp masking amplifies high frequencies in the image. It accentuates the (diverse) trabecula structures of healthy and pathologic bone. The third feature O_3 is the median

filter $\mu[f,(x,y)]$ which passes low frequencies, i.e. the median grey level.

In the 1970s, Haralick [36] introduced the concept of a co-occurrence matrix to characterise and discriminate different types of texture. Earlier experiments on our image material had indicated that measures based on co-occurrence matrices computed from a small neighbourhood provide neural networks with a high discriminatory power [37]. Define the co-occurrence matrix $M(l)$ as

$$m_{i,j} = \rho(f(\boldsymbol{p}) = i, f(\boldsymbol{q}) = j), \quad d(\boldsymbol{p},\boldsymbol{q}) = l \tag{4}$$

with $\rho(a,b)$ being the correlation between a and b, and $\boldsymbol{p} = (x,y)^{\mathrm{T}}$ and $\boldsymbol{q} = (x',y')^{\mathrm{T}}$ coordinate pairs of pixels in the quadratic window from which the correlation between pixels is estimated. The function $d(\boldsymbol{p},\boldsymbol{q})$ computes the Euclidean distance between the two coordinate pairs; l specifies the required distance. The parameter l is directly related to the frequency spectra of the textures one wants to discriminate. The best results on our image material was obtained with the distance l set to 1 [37].

One also needs to optimise the size of the window from which $M(l)$ is computed. On the one hand, the larger the sample used to estimate the correlation measures in $M(l)$, the smaller is their variance. On the other hand, larger windows are more likely to include different types of textures which biases the correlation measures. In general, one has to find the optimal trade-off between variance and bias which is problem dependent. As the minimally required size of the window depends on the frequency characteristics of the textures, we computed two-dimensional auto correlation functions (AKFs) from different Fourier transformed subimages of soft tissue, healthy and pathologic bone. The major difference between the AKFs of healthy and pathologic bone is the relative magnitude of frequency components with periods between 5 and 10 pixels. Moreover, preliminary experiments in which textural features based on co-occurrence matrices were computed using various window sizes confirmed that the four classes are best separated with features computed from windows with sizes varying from 7×7 to 11×11 pixels [35]. Adding textural features computed on coarser scales did not improve the segmentation result.

We used four texture measures derived from the co-occurrence matrix $M(l)$ to characterise the local texture [36]:

Second angular moment (O_4)

$$\zeta = \sum_i \sum_j m_{i,j}^2 \tag{5}$$

Inverse difference moment (O_5)

$$\iota = \sum_i \sum_j \frac{1}{1 + (i-j)^2} m_{i,j} \tag{6}$$

Contrast (O_6)

$$\kappa = \sum_i \sum_j (i-j)^2 m_{i,j} \tag{7}$$

and Entropy (O_7)

$$\epsilon = \sum_i \sum_j m_{i,j} \ln(m_{i,j}) \tag{8}$$

The feature O_4 – the second angular moment – is the sum of squares of the correlation measures in the co-occurrence matrix. It can distinguish areas in which pixels with a distance l are correlated, i.e. a systematic texture, from areas where the pixels are randomly distributed, e.g. due to white noise. The feature O_5 – the inverse difference moment – is a weighted sum of the correlation measures in the matrix. Entries close to and on the diagonal obtain high weights whereas entries with a large diagonal distance in M are assigned small weights. The measure ι can distinguish low frequent textures from textures in which pixels (with a distance l) often have very different intensities. Feature O_6 – contrast – measures almost the opposite textural property of feature O_5, although in the formula for κ diagonal entries are assigned the weight zero. Feature O_7 – entropy – measures the variation among the correlation measures in the co-occurrence matrix. It is a measure for the complexity of the texture.

The last feature $O_8(f(x,y))$ is the intensity of pixel (x,y) itself, i.e. the identity operation.

4. EXPERIMENTS

We tested our feature-based segmentation approach on radiographs of different bone tumours including a ewing- and an osteo-sarcoma. The input to the neural-net classifier is the 8-dimensional feature vector $\boldsymbol{e}(x,y)$, the output the 4-dimensional vector $\boldsymbol{z}(x,y)$. The features unsharp masking and the median filter had a window size of 7×7, the co-occurrence matrices were estimated from 11×11 windows. Segmentation is obtained by convolving the image with the neural network. Each input vector $\boldsymbol{e}(x,y)$ is classified according to the winner takes all rule:

$$class(\boldsymbol{z}(x,y)) = \begin{cases} j: & \forall l \neq j: z_l(x,y) < z_j(x,y) \\ \varnothing: & else \end{cases} \tag{9}$$

with $\boldsymbol{z}(x,y) = h(\boldsymbol{e}(x,y))$. Figure 1 illustrates how the neural network segments a radiograph based on the feature vector.

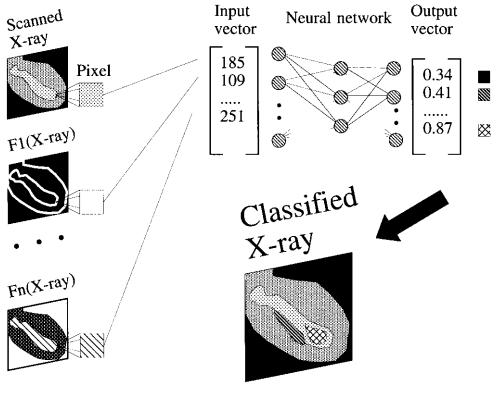

Fig. 1. Schematic representation of the segmentation approach. The original image is transformed by a number of image operations. These provide the classifier (MLP or SOM) with neighbourhood information of the pixel that is to be classified.

4.1. Correct Segmentation from Radiologist

We obtained a set of about 200 radiographs with focal bone lesions from the department of radiology at the RWTH-Aachen. In addition to this, we added a smaller set of images with normal findings, i.e. reference images. The radiographs were digitised with a Dupont laser scanner with a depth of 10 bit and a maximum matrix size of 2048×1684. For our experiments we have chosen a core set of 20 radiographs containing the most important types of bone lesions. An experienced radiologist was asked to specify the class membership of regions inside each radiographic image. Based on the annotations, which were made on a transparent folio, a digital mask was composed indicating the class membership of the pixels in each of the 20 images. However, in most of the radiographs the class membership of only a subset of the pixels was known as, due to uncertainty of the radiologist, not all segments were associated with a specific class.

4.2. Experiment with Radiograph of Brown Tumour

First, we tested our approach on a real radiograph with large segments of healthy and pathologic bone. In this experiment, we trained a neural network to segment the radiograph of a brown tumour as specified by the annotations of the radiologist. A training set was composed by choosing at random 8000 vectors from the regions associated with a specific class (about 3% of the image).

Different neural networks with 4–8 hidden nodes were trained with the backpropagation algorithm for 25,000 cycles, learning rate = 0.0001, momentum = 0.5, off line learning. A class label is assigned to each pixel according to Eq. (9). The best results were obtained using the network with six hidden nodes. When used to segment the areas of which the correct class membership was known, this network obtained a correctness of 0.9301. Figure 2 shows the radiograph (left), the annotation mask obtained from the radiologist (centre) and the obtained segmentation result (right). In the mask image, darkest grey indicates the segment that is not assigned to any class by the radiologist (unknown).

Segments with a low class-conditional correctness were soft tissue, 0.8427, and pathologic bone, 0.9472 (see Table 2). From this contingency table, we computed the class-conditional bias measures given the true class label u as defined in Egmont-Petersen et al [30]. We used the bias term (Eq. (39) in Egmont-Petersen et al [30]) to indicate whether class u is biased *towards* or *away from* class $t(\neq u)$ while taking the prior class distribution into account. The bias analysis indicates that the class tissue is biased towards the class pathologic bone (see Table 2). The two classes healthy and pathologic bone are both biased towards tissue. So pixels from the classes healthy and pathologic bone tend to be misclassified as tissue. On the other hand, pixels truly belonging to the class tissue are primarily misclassified as healthy or pathologic bone.

4.3. Experiment with Radiograph of an Osteosarcoma

In a second experiment, we wanted to estimate the discriminative performance of our neural approach on a test image.

Table 2. Contingency table of the feed-forward neural network used for segmenting the brown tumour image

	True class label MLP			
	Backg.	Tissue	H. bone	P. bone
Backg.	23411	0	0	0
Tissue	0	29513	7	3858
H. bone	0	2531	2784	0
P. bone	1	2979	13	69150
Total	23412	35023	2804	73008
Correctness	1.0000	0.8427	0.9929	0.9472

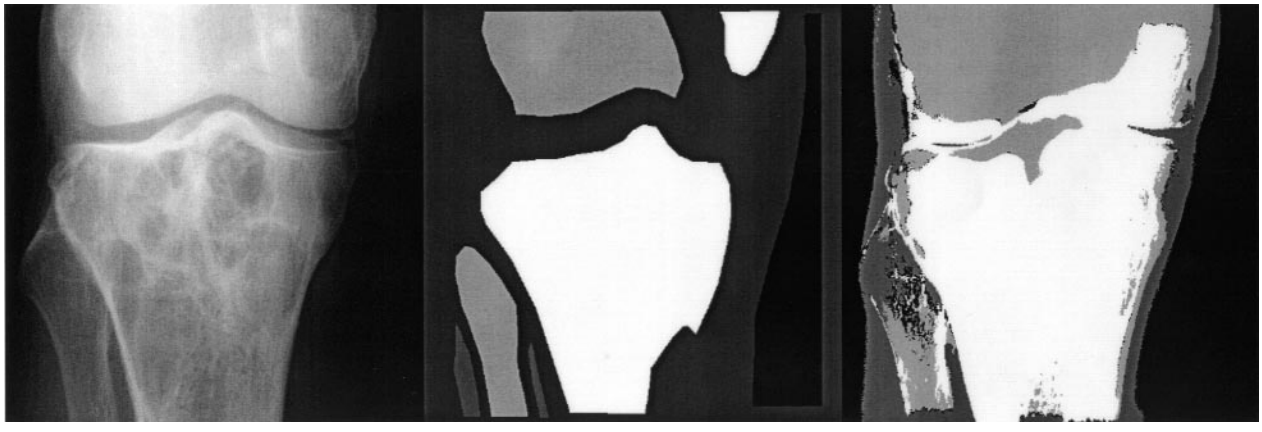

Fig. 2. Radiograph of knee with brown tumour (left). The middle image contains the mask as obtained from the expert radiologist. The grey level of a pixel in the mask indicates its class membership. Black is background, dark grey is soft tissue, light grey is healthy bone and white is pathologic bone. Part of the mask indicates pixels of which the class membership is unknown (darkest grey). The right image is the segmentation result obtained with the neural network.

Table 3. Contingency table of the feed-forward neural network used for segmenting the osteo-sacroma image

	True class label MLP			
	Backg.	Tissue	H. bone	P. bone
Backg.	7382	0	0	0
Tissue	0	67173	1505	1
H. bone	0	937	21797	4617
P. bone	0	1080	1852	53483
Total	7382	69190	25154	58101
Correctness	1.0000	0.9708	0.8665	0.9205

When it is to be used in clinical practice, the neural network should have a high probability of locating pathologic bone while misclassifying as few pixels as possible that belong to other classes. We chose four radiographs from among the 20 indicating different bone tumours. The radiologist was asked to make an annotation mask for each radiograph. We composed a training set consisting of 8000 pixels chosen by random from three of the images. Solely pixels were included of which the true class membership was known. Again, neural networks with 4–8 hidden nodes were trained with the backpropagation algorithm for 25,000 cycles, learning rate = 0.0001, momentum = 0.5, off line learning. A fourth radiograph was used for testing purposes (see Fig. 3 (upper left)). This radiograph shows an osteo-sarcoma, the bone tumour with the highest incidence in Germany. When tested on this image, the neural network that performed best on the training set (six hidden nodes) resulted in a correctness of 0.9375 which is comparable to the previous experiment.

The performance of the neural network when applying the winner takes all rule (Eq. (9)) is indicated in Table 3. It is clear that healthy and pathologic bone most difficult to discern. This is also seen in the last four images in Fig. 3 (upper right and lower row), which show the output activation of the neural network corresponding to each of the four classes (background, tissue, healthy and pathologic bone). These output images can be used to indicate areas that might contain pathologic bone.

We also computed a ROC-curve for this test image (see Fig. 4). The ROC-curve indicates the trade-off between true and false positive (pathologic bone) pixels when varying a threshold τ. For each value of τ, the fractions of true positive and false positive pixels were computed according to $o_j - \max(o_i) > \tau$, $i \neq j$, with j the index of the class

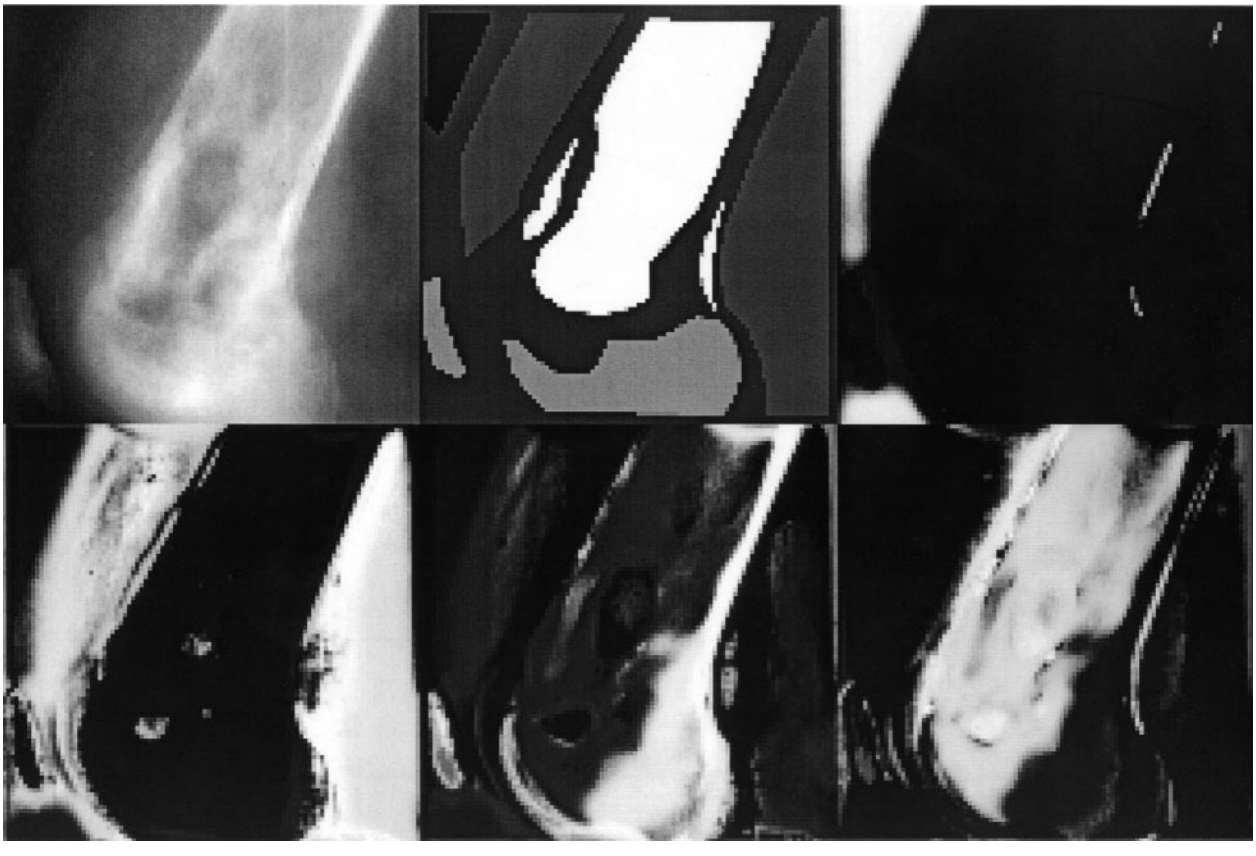

Fig. 3. Radiograph of osteo-sarcoma of a 20 year old patient. From upper left to lower right: The radiograph, hand-drawn annotation mask of the radiologist and visualisation of the activation of the output neurons for the four classes: background, soft tissue, healthy bone and pathologically changed bone.

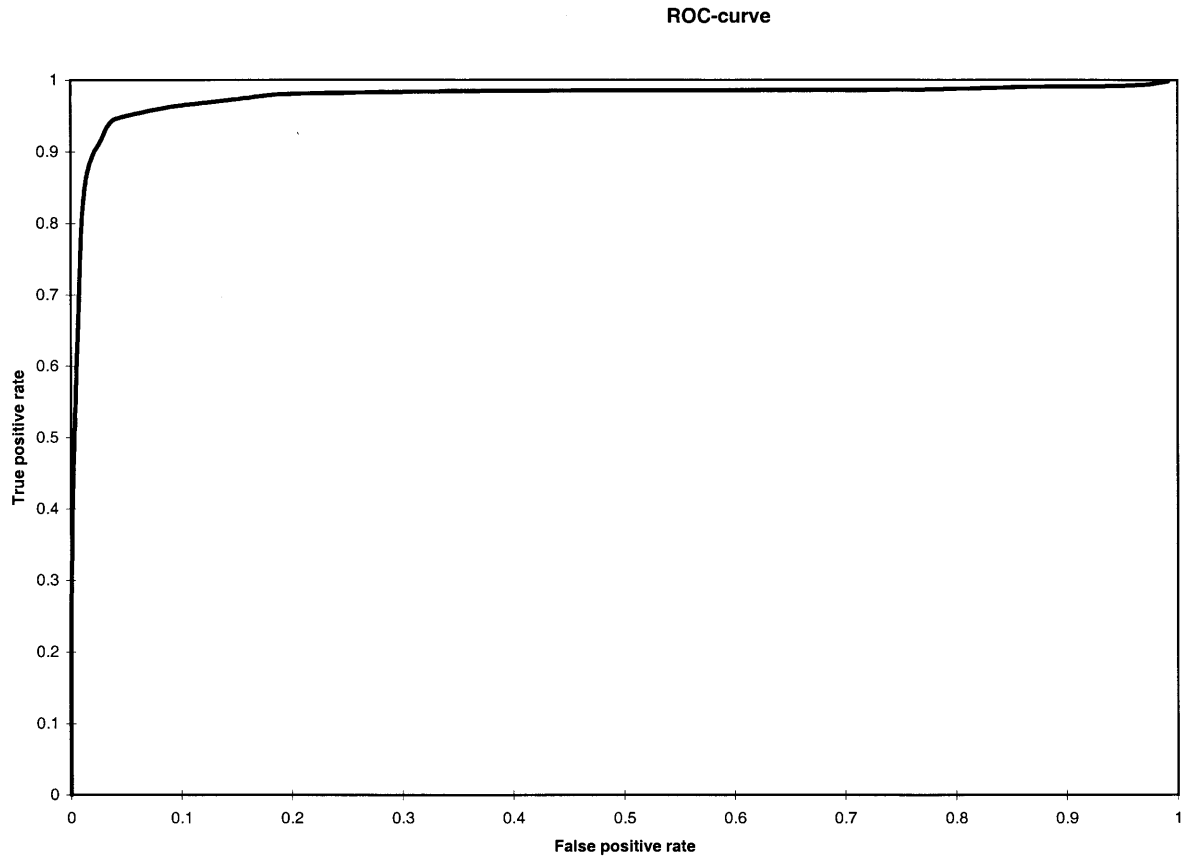

Fig. 4. ROC-curve computed on the test image in Fig. 3. True positive rate indicates the fraction of correctly classified truly pathologic pixels for a certain threshold value, the false positive rate indicates the fraction of pixels classified as pathologic bone but belonging to one of the other three classes.

pathologic bone. Pixels that were not associated with any class in the mask were not included in the computation. The ROC-curve shows that it is possible to locate pathologic bone while keeping the fraction of false positive pixels very small.

The second experiment shows that the chosen features are robust to variation introduced by using more or less randomly chosen radiographs indicating bone tumours. It also shows that only a relatively small set of training samples is necessary to obtain a well performing neural-net classifier.

4.4. Experiment with Synthetic Radiograph

We wanted to test our algorithm on image material of which the class label was known for all pixels with a high degree of confidence. Moreover, the algorithm should be capable of identifying pathologic bone in radiographs showing different types of bone tumours. Therefore, we composed a synthetic radiograph by 'cutting' rectangular regions (subimages) in eight different radiographs inside areas annotated by the radiologist [38]. The subimages were 'pasted' into an empty image (see Fig. 5 (left)). The two subimages showing pathologic bone were taken from a ewing-sarcoma and an osteo-sarcoma, the two bone tumours with the highest incidence. The synthetic image contained 492×492 pixels, 256 intensities.

We composed a training set by choosing at random 8000 vectors (about 3% of the image). Each vector $\boldsymbol{e}(x,y)$ was used as input to a neural-net classifier and a self-organising feature map. A number of different feed-forward neural networks were trained with the same settings as in the previous experiments. Figure 5 contains the synthetic radiograph (left), the segmentation result from the neural network (centre) and from the self-organising feature map (right).

The self-organising feature map was trained according to a two-phased algorithm: self-organisation followed by convergence as proposed by Kohonen [39]. During training, a mapping is established from the n-dimensional feature space onto the 2-dimensional grid S called the *feature map* consisting of $20 \times 30 = 600$ nodes. After 10,000 cycles the learning process was terminated and a class label assigned to each node $s_{a,b}$ in the feature map. The most frequent class label among the training vectors of which the Euclidean distance to node $s_{a,b}$ is smaller than to any other node in the map determines the class label of $s_{a,b}$. Topological nodes that did not obtain a class label were temporarily assigned to the special class 'unknown'. Subsequently, these nodes $s_{a,b}$ were assigned the same class label as the input

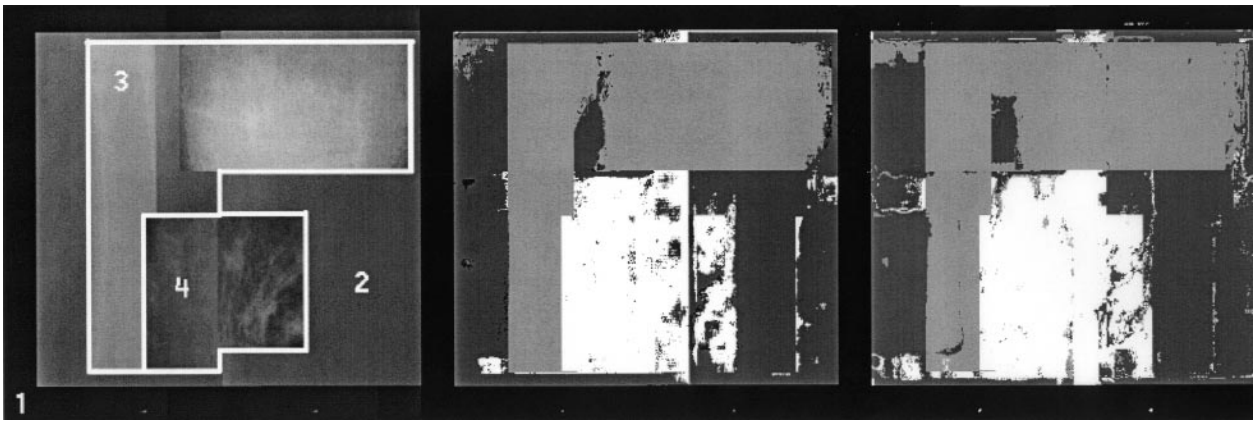

Fig. 5. Synthetic radiograph (left) and the segmentation results (MLP middle and SOM right). The numbers 1–4 represent the classes background, soft tissue, healthy and pathologic bone. The grey level of a pixel in the central and right images indicates the class membership as assigned by the classifier. Black is background (1), dark grey is soft tissue (2), light grey is healthy bone (3) and white is pathologic bone (4).

vector with the smallest Euclidean distance.

Also in this experiment, the best result was obtained with a network with six hidden nodes. It obtained an overall correctness of 0.8983 based on all pixels in the synthetic image (see Table 4). Also in this image, tissue, pathologic and, to a lesser extent, healthy bone are difficult to discern. The class background is biased towards the two classes with the darkest average intensity: tissue and pathologic bone. The class tissue is biased towards the classes healthy and pathologic bone which are again biased towards tissue.

The self-organising feature map resulted in a lower class-conditional correctness on the class tissue as compared with the feed-forward neural network, whereas pathologic bone obtained a higher class-conditional correctness. In fact, the difference between the number of pixels correctly classified as tissue plus pathologic bone is about the same for the two classifiers, $67{,}679 + 28{,}639 = 96{,}318$ for the neural network and $65{,}904 + 30{,}401 = 96{,}305$ for the self-organised feature map. Which of the two classifiers are chosen, the same performance is obtained.

4.5. Experiment with Reduced Feature Sets

The previous experiments showed that our segmentation approach resulted in a good performance. However, we have no knowledge of which features are important for discriminating the four classes. The correlation matrix (Table 5) was computed from the data in the synthetic image. It shows that some of the features are highly dependent.

We also computed the eigenvectors of the correlation matrix. They reflect the relative variances of the principal components of the feature distribution (a principal component analysis is often used for feature extraction). Figure 6 shows the accumulated variance as a function of the number of principal components. It illustrates that three components explain 97% of the variation in the data set. However, what we do not obtain from the correlation matrix and the principal component analysis is information about which features contribute most to the discriminative performance of the classifier.

Table 4. Contingency tables of the feed-forward neural network and the self-organising feature map for the synthetic radiograph

	True class label							
	MLP				SOM			
	Backg.	Tissue	H. bone	P. bone	Backg.	Tissue	H. bone	P. bone
Backg.	52955	1540	526	1188	52519	604	0	3
Tissue	10	67679	4979	2493	86	65904	5687	1736
H. bone	0	1210	68184	180	0	2374	68437	360
P. bone	0	8750	3731	28639	360	10297	3296	30401
Total	52965	79179	77420	32500	52965	79179	77420	32500
Correctness	0.9998	0.8548	0.8807	0.8812	0.9916	0.8323	0.8840	0.9354

Table 5. Correlation matrix of the eight features computed from all pixels in the synthetic radiograph

O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8
1.00	0.98	0.98	0.40	0.73	–0.18	0.40	0.98
	1.00	0.99	0.41	0.73	–0.21	0.41	1.00
		1.00	0.41	0.74	–0.21	0.41	1.00
			1.00	0.82	–0.08	0.91	0.41
				1.00	–0.09	0.83	0.74
					1.00	–0.13	–0.21
						1.00	0.41
							1.00

A method developed by Egmont-Petersen [40] (chapter 3) was used to identify features with a low discriminative power. Subsequently, neural networks were trained with reduced feature sets. The developed method ranks the n features according to their relative discriminative power for the classification of a vector $\boldsymbol{e}(x,y)$. The discriminative power of a feature is defined as the probability that the vector would obtain *another* class label if the feature would be observed again while keeping the other feature values fixed. For each vector in the training set, the relative importance of each feature is determined by ranking the n features according to their discriminative power and summing these ranks per feature across all vectors in the training set.

We used a special test set consisting of 996 vectors for feature assessment. Thirty neural networks, each with six hidden nodes but a unique initial weight configuration, were trained with 8000 vectors from the synthetic radiograph. After training, the eight features were ranked according to their discriminative power. The four features with the lowest overall discriminative power – the original image, second angular moment, contrast and histogram equalisation – were removed from the training and test sets. Subsequently, 30 new networks with six hidden nodes were trained and the performance computed on the test set. Again, the discriminative power of the four features was computed, the least important feature – median – was removed and 30 new neural networks were trained. Finally, the inverse difference moment was removed and 30 new neural networks were trained.

The performance of the 30 neural networks based on all eight features was on the test set 0.891 (±0.028). The results from the feature assessment in Table 6, indicate that some features have a higher discriminative power than others. When the four features with the smallest discriminative power were removed (rank 5–8) and 30 new neural networks were trained, the performance slightly increased 0.903 (±0.004), a phenomenon known as peaking. Note that the average correctness measures obtained from the networks based on four and eight features are not significantly different. When the feature Median was removed and 30 new networks were trained, a performance of 0.890 (±0.007)

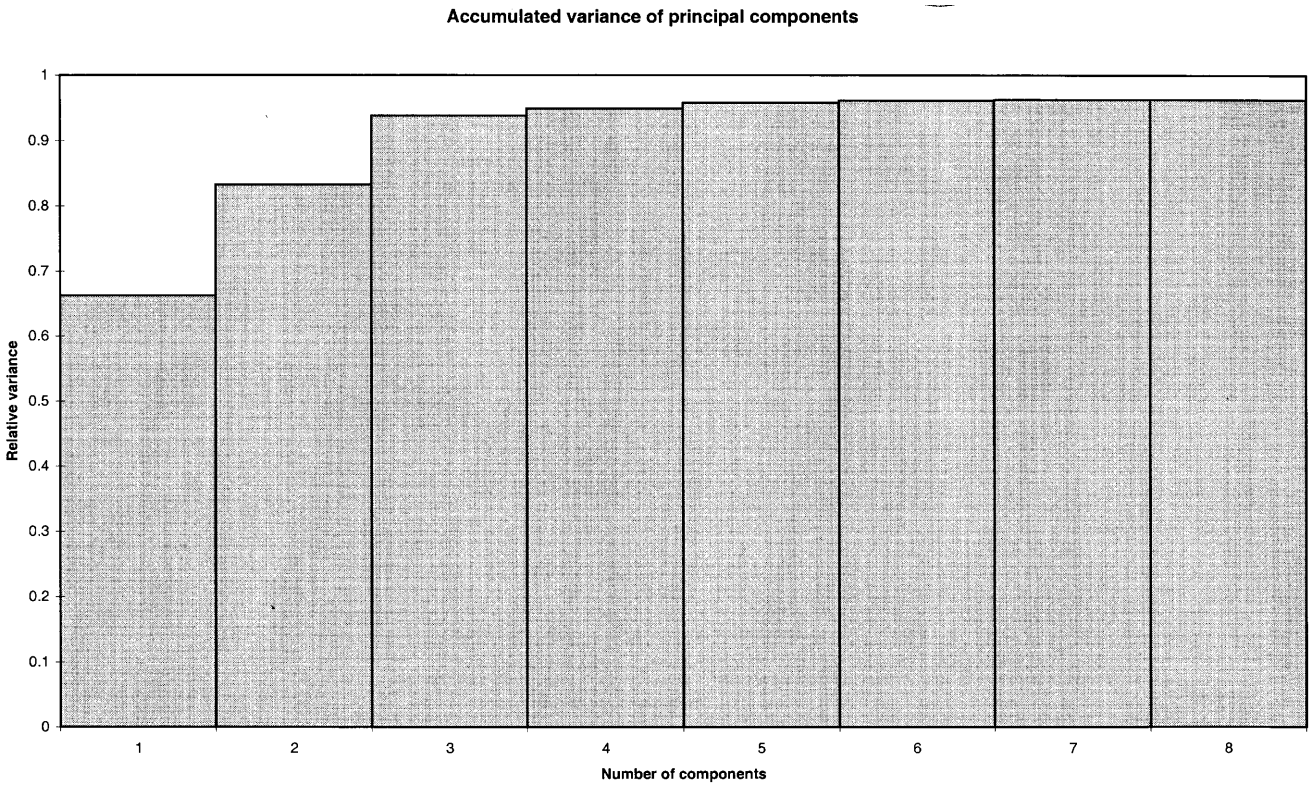

Fig. 6. Curve showing the accumulated variance explained by an increasing number of principal components. The eigenvalues were computed from the correlation matrix (Table 5).

Table 6. The ranks of the features are averaged over 30 neural networks. The first eight rows contain the average ranks of all features for each of the four classes and the average overall rank computed while taking into account the prior class distribution. A low rank is associated with an important feature. The next four rows contain similar ranks for the 30 networks trained with four features, the last three rows contain the ranks for the networks trained with 3 features. The last column indicates the rank assigned to each feature

Avg. rank	Backg.	Tissue	H. bone	Pat. bone	Overall	Rank
Hist.	4.50	4.68	4.51	4.78	4.61	5
Ush.	4.50	3.69	4.48	3.30	4.02	2
Med.	4.50	3.74	4.48	2.98	3.99	1
Sam.	4.50	5.13	4.52	5.07	4.82	7
Idm.	4.50	4.70	4.50	4.56	4.58	4
Cont.	4.50	4.64	4.50	5.78	4.75	6
Entr.	4.50	4.13	4.49	3.99	4.28	3
Orig.	4.50	5.30	4.53	5.54	4.95	8
Ush.	2.50	2.46	2.49	3.01	2.58	3
Med.	2.50	2.50	2.49	2.99	2.59	4
Idm.	2.50	2.67	2.52	2.06	2.48	2
Entr.	2.50	2.37	2.50	1.93	2.35	1
Ush.	2.00	1.77	1.98	2.24	3.92	2
Idm.	2.00	2.28	2.02	1.96	4.18	3
Entr.	2.00	1.95	2.00	1.79	3.89	1

was obtained. Reducing the feature set further led to a significant drop in performance, 0.858 (±0.029). The graph in Fig. 7 shows the average correctness of the neural networks as a function of the number of features.

The results in Table 6 show that when one or more features are removed, the remaining features do not keep the same mutual rank. This phenomenon is caused by dependencies between the features and can be observed also for the Bayes classifier [31]. With respect to the segmentation task, the three most important features are unsharp masking which accentuates the trabecula structure, the inverse difference moment which distinguishes low from high frequent textures and entropy which measures the complexity of the texture. Note that, according to the principal component analysis, three components could explain 97% of the variation in the feature set.

5. DISCUSSION

We have developed a feature-based approach for segmentation of radiographs using a neural network and a self-organising feature map as classifiers. Between 90% and 95% of the pixels are assigned to the correct segments. The performance analysis indicates that especially soft tissue and pathologic bone are difficult to discern. This is partly due to the fact that these two segments have almost the same average intensity. The left and right subimages of segment 4 in the synthetic radiograph (Fig. 5, left) show that the appearance of pathologic bone varies much between radiographs. The segmentation results of both classifiers indicate that the right subimage is difficult to segment correctly because of its inhomogeneous texture. This problem could be remedied by grouping neighbouring pixels into the same segment. However, this might not improve the performance of the segmentation approach. One could also analyse the textures on more scales, e.g. by varying the parameter l that specifies the distance between pixels for which the correlation measures in the co-occurrence matrix are computed. This issue is left for further research. A multi-scale wavelet approach has shown interesting results on our image material [38], although the performance is poorer than that obtained with the features presented here.

In the synthetic image, edges occur which are not present in the original image material. These pixels could have been removed from the training set and their class membership considered unknown.

Our gold standard consists of annotations made by a radiologist who is specialised in diagnosing focal bone lesions. Because a high degree of experience is needed in this field, different (specialised) radiologists may disagree with respect to which parts of a radiograph indicate tissue, healthy and pathologic bone. We expect a certain inter-observer variation because a radiograph is a projection which, by virtue of the overlapping structures, remains difficult to analyse. The next step in our research is to initiate a larger study in which both the inter-observer variability is assessed and where the different types of focal bone lesions are being analysed by our algorithm with incidences that coincide with those of a nonspecialised clinic.

Another reason to expect a certain inter-observer variability is that the image modality used is not the optimal one for characterising tissue, which can better be investigated using MR-imaging. At the Leiden University Medical Centre, which is the Dutch centre for treating bone tumours, dynamic MR-imaging with the contrast medium Gadolinium is used to determine the size and malignancy of bone tumours. However, patients always present with a radiograph made in a peripheral hospital. Radiography is a relatively inexpensive technology that is available in almost every hospital in Europe and Northern America.

6. CONCLUSION

We have presented an approach for screening radiographic images of focal bone lesions for pathologic changes indicating bone tumour. A neural network and a self-organising feature map are used to classify individual pixels based on a feature vector. The feed-forward neural network was chosen because it generally results in low error rates, also when the features are not normally distributed and because the application of a neural network requires little computation. The features we used characterise different aspects of the texture in a small neighbourhood of a pixel. An analysis of the eight features indicates that the presence of edges and

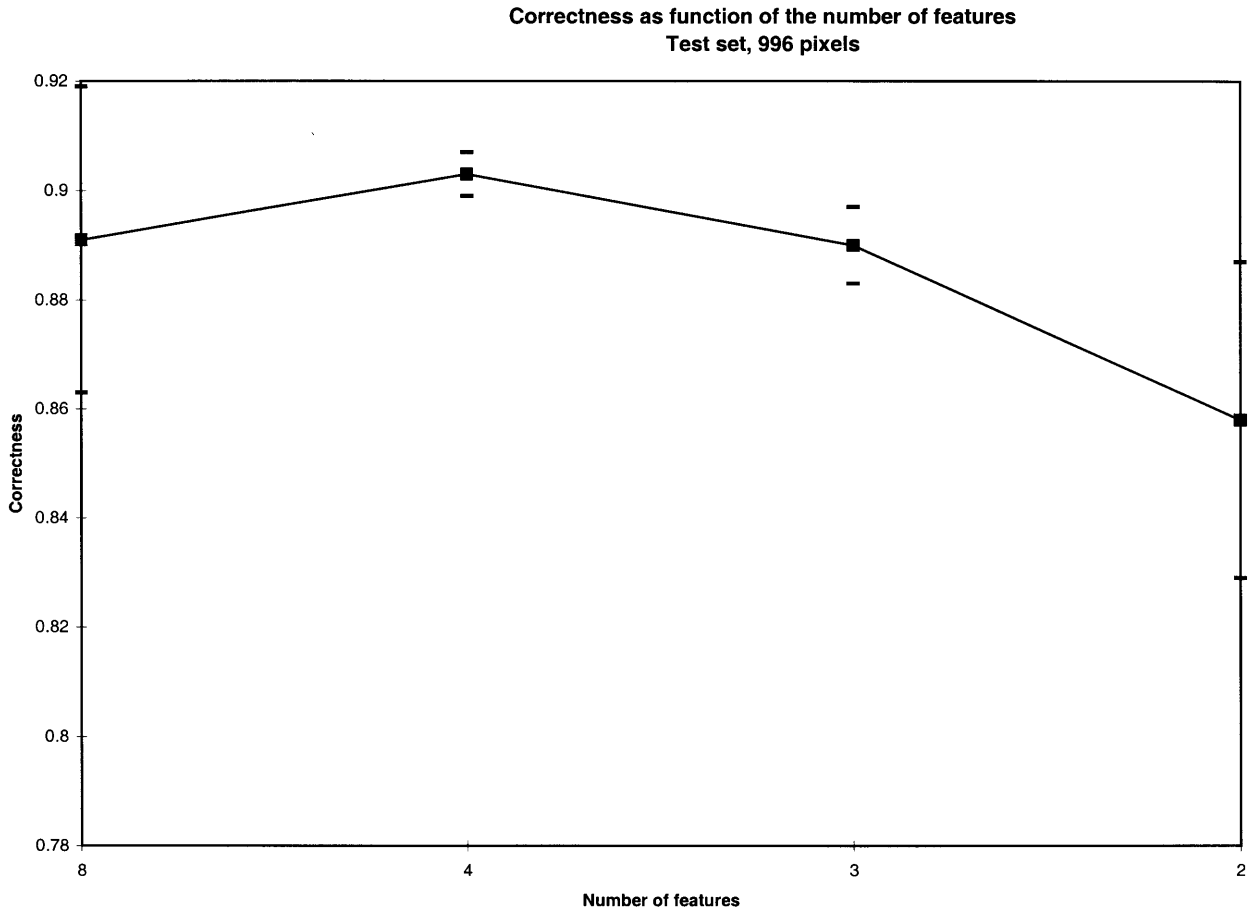

Fig. 7. Curve showing the average correctness and the confidence intervals ($\alpha = 0.01$) among the 30 neural networks when reducing the number of features.

transitions, the complexity of the texture as well as the amount of high frequencies in the texture are the main features discriminating (soft) tissue from pathologic bone, the two classes most likely to be confused. Our experiments indicate the feasibility of using a feature-based neural network for screening radiographic images for pathologic bone changes. However, more research is needed to ensure that the different appearances of pathologic bone are reliably recognised.

Acknowledgements

We wish to thank Dr B. Wein for providing the radiographs, the Department of Radiology at the University Hospital of the RWTH-Aachen for allowing us to use their DEC-alpha computer and Professor Dr T. Tolxdorff an Professor Dr K. Bohndorf for supporting this project. Part of the work was financed by DFG-project TO. 108/3–2 and by the Dutch Cancer Society, Grant RUL 97–1509. Frank Vogelsang is thanked for providing several images. Finally, we thank the reviewers for valuable comments and suggestions.

References

1. Ercal F, Chawla A, Stoecker WV, Lee H-C, Moss RH. Neural network diagnosis of malignant melanoma from color images. IEEE Trans Biomedical Engineering 1994; 41(9): 837–845
2. Bajaj R, Chaudhury S. Signature verification using multiple neural classifiers. Pattern Recognition 1997; 30(1): 1–7
3. Heikonen J, Mantynen M. A computer vision approach to digit recognition on pulp bales. Pattern Recognition Letters 1996; 17: 413–419
4. Huang K, Yan H. Off-line signature verification based on geometric feature extraction and neural network classification. Pattern Recognition 1997; 30(1): 9–17
5. Iftekharuddin KM, Schechinger TD, Jemili K, Karim MA. Feature-based neural wavelet optical character recognition system. Optical Engineering 1995; 34(11): 3193–3199
6. Itoh K. ID number recognition of X-ray films by a neural network. Computer Methods and Programs in Biomedicine 1994; 43: 15–18
7. Javidi B, Tang Q. Optical implementation of neural networks by the use of nonlinear joint transform correlators. Applied Optics 1995; 34(20): 3950–3962
8. Jørgensen TM, Christensen SS, Andersen AW. Detecting danger labels with RAM-based neural networks. Pattern Recognition Letters 1996; 17: 399–412

9. Kepuska VZ, Mason SO. A hierarchical neural network system for signalized point recognition in aerial photographs. Photogrammetric Engineering & Remote Sensing 1995; 61(7): 917–925
10. Sklansky J, Vriesenga M. Genetic selection and neural modelling of piecewise-linear classifiers. Int J Pattern Recognition and Artificial Intelligence 1996; 10(5): 587–612
11. Riede UN, Schaefer HE, Wehner H. Allgemeine und spezielle Pathologie. Georg Thieme, Stuttgart, 1989
12. Freyschmidt J, Ostertag H. Knochentumoren: Klinik – Radiologie – Pathologie. Springer-Verlag, Berlin, 1988
13. Sorgel W, Girod S, Szummer M, Girod B. Computer aided diagnosis of bone lesions in the facial skeleton. Proc Bildverarbeitung in der Medizin (Medical Image Processing), Springer-Verlag, 1998, pp 179–183
14. Lodwick GS, Reichhertz PL. Computerunterstutzte Diagnostik von Tumoren und tumorahnlichen Veranderungen des Knochens – Das begrenzte Bayes-Konzept. Rontgenblatt 1969; 22: 162–168
15. Bohndorf K, Pelikan E, Tolxdorff T, Zarrinnam D, Wein B, Gunther R. Computerassistierte Diagnose von Knochentumoren: Neue Entwicklungen. Zentralblatt Radiologie 1993; 147: 987
16. Reinus WR, Wilson AJ, Kalman B, Kwasny S. Diagnosis of focal bone lesions using neural networks. Investigative Radiology 1994; 29(6): 606–611
17. Hand DJ. Discrimination and Classification. Wiley, Chichester, 1981
18. Duda RO, Hart PE. Pattern Classification and Scene Analysis. Wiley, New York, 1973
19. Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE Trans Information Theory 1967; 13: 21–27
20. Funahashi K-I. On the approximate realization of continuous mappings by neural networks. Neural Networks 1989; 2: 183–192
21. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks 1989; 2: 359–366
22. Richard MD, Lippmann RP. Neural network classifiers estimate bayesian a posteriori probabilities. Neural Computation 1991; 3: 461–483
23. Kohonen T. Self-organizing maps: optimization approaches. Proc Artificial Neural Networks 1991; 981–990
24. Talmon JL. A multiclass nonparametric partitioning algorithm. Pattern Recognition Letters 1986; 4: 31–38
25. Quinlan JR. Induction of decision trees. Machine Learning 1986; 1(1): 81–106
26. Hart A, Wyatt J. Connectionist models in medicine: an investigation of their potential. Proc AIME-89. Springer-Verlag, 1989, pp 115–124
27. Serpico SB, Bruzzone L, Roli F. An experimental comparison of neural and statistical non-parametric algorithms for supervised classification of remote-sensing images. Pattern Recognition Letters 1996; 17: 1331–1341
28. Stassopoulou A, Patrou M, Kittler J. Bayesian and neural networks for geographic information processing. Pattern Recognition Letters 1996; 17: 1325–1330
29. Harrison RF, Marshall SJ, Kennedy RL. A connectionist aid to the early diagnosis of myocardial infarction. Proc AIME–91. Springer-Verlag, 1991, pp 119–128
30. Egmont-Petersen M, Talmon JL, Brender J, McNair P. On the quality of neural net classifiers. Artificial Intelligence in Medicine 1994; 6(5): 359–381
31. Egmont-Petersen M, Talmon J, Hasman A, Ambergen AW. Assessing the importance of features for multi-layer perceptrons. Neural Networks 1998; 11(4): 623–635
32. Pelikan E, Vogelsang F, Schulz B, Egmont-Petersen M, Bohndorf K, Tolxdorf T. Texturbasierte Segmentierung von Rontgenbildern mittels Multilayer-Perzeptron und Topologischer Karte. Proc DAGM (German conference on pattern recognition). Wien, 1994, pp 589–600
33. Vogelsang F, Pelikan E, Egmont-Petersen M, Tolxdorff T, Bohndorf K. Segmentierung von Rontgenbildern fokaler Knochenlasionen durch neuronale Netzwerke. Optimierung durch Quality Metrics und modifizierte Contribution Analysis. Proc Workshop on Neural Networks at the RWTH-Aachen. Aachen, 1993, pp 201–210
34. Vogelsang F, Pelikan E, Egmont-Petersen M, Tolxdorff T, Bohndorf K. Segmentierung von Rontgenbildern fokaler Knochenlasionen durch neuronale Netzwerke. Optimierung durch Quality Metrics und modifizierte Contribution Analysis. Proc DAGM (German conference on pattern recognition). Lubeck, 1993, pp 450–459
35. Nobis T. Berucksichtigung lokaler und glober Textureigenschaften durch Erweiterung des Konzepts der Grauwertubergangsmatrizen auf einen Multiskalenansatz. Medical Informatics and Biometrics, Aachen, Aachen University of Technology, 1994
36. Haralick RM, Shanmugam K, Dinstein I. Textual features for image classification. IEEE Trans Systems, Man and Cybernetics 1973; 3: 610–621
37. Weiler F, Pelikan E, Nobis T, Tolxdorff T, Bohndorf K. Texturbasierte Extraktion medizinischer Merkmale aus Filmrontgenbildern. Proc DAGM (German conference on pattern recognition). Lubeck, 1993, pp 460–467
38. Pelikan E. Texturorientierte Segmentierungsmethoden in der medizinischen Bildverarbeitung. Institute of Medical Informatics and Biometrics, Aachen, RWTH-Aachen, 1995
39. Kohonen T. Self-organized formation of topologically correct feature maps. Biological Cybernetics 1982; 43: 59–69
40. Egmont-Petersen M. Specification and Assessment of Methods Supporting the Development of Neural Networks in Medicine. Shaker Publishing, Maastricht, 1996

Dr Egmont-Petersen was born in Copenhagen, Denmark, in 1967. He received the combined BS and combined MS degrees in Computer Science/Business Administration from Copenhagen Business School in 1988 and 1990, respectively. He received his PhD degree in medical informatics from Maastricht University, The Netherlands, in 1996. He is currently associated with the Division of Image Processing, Department of Radiology, Leiden University Medical Centre, as postdoctoral researcher. Dr Egmont-Petersen is currently working with the quantification of perfusion in bone tumours based on dynamic MR-imaging. His main research interests include neural networks, statistical classifiers, feature selection, quality assessment of classifiers, non-linear filtering, digital image processing and invariant theory. He has published more than 25 papers in journals and conference proceedings.

Dr Erich Pelikan was born in Dusseldorf, Germany in 1965. He received the electronic engineering diploma in 1990 from Aachen Technical University. In 1995 he received his PhD degree from the Faculty of Mathematics and Computer Science of the Aachen Technical University for his work on image processing and segmentation of radiographs of bone tumours using neural classifiers. He is currently associated with the Department of Clinical Science, Philips Medical Systems, in Hamburg, Germany. Dr Pelikan takes care of the scientific cooperation between scientific medical institutions and Philips Medical Systems in the fields of image processing and digital communication. His main scientific interests include digital image processing, especially qualitative and quantitative analysis of virtual endoscopic sequences, radiology information systems and workflow techniques.

Correspondence and offprint requests to: Dr M. Egmont-Petersen, Division of Image Processing, Department of Radiology, C2-S, Leiden University Medical Centre, P.O.B. 9600, NL-2300 RC Leiden, The Netherlands. Email: michael@lkeb.azl.nl