Medical Engineering & Physics 26 (2004) 71–86

Medical Engineering & Physics

www.elsevier.com/locate/medengphy

Segmentation of magnetic resonance images using a combination of neural networks and active contour models

Ian Middleton [a], Robert I. Damper [b,*]

[a] *Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA*

[b] *Image, Speech and Intelligent Systems (ISIS) Research Group, Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK*

Received 7 January 2003; accepted 21 July 2003

Abstract

Segmentation of medical images is very important for clinical research and diagnosis, leading to a requirement for robust automatic methods. This paper reports on the combined use of a neural network (a multilayer perceptron, MLP) and active contour model ('snake') to segment structures in magnetic resonance (MR) images. The perceptron is trained to produce a binary classification of each pixel as either a *boundary* or a *non-boundary* point. Subsequently, the resulting binary (edge-point) image forms the external energy function for a snake, used to link the candidate boundary points into a continuous, closed contour. We report here on the segmentation of the lungs from multiple MR slices of the torso; lung-specific constraints have been avoided to keep the technique as general as possible. In initial investigations, the inputs to the MLP were limited to normalised intensity values of the pixels from an (7×7) window scanned across the image. The use of spatial coordinates as additional inputs to the MLP is then shown to provide an improvement in segmentation performance as quantified using the effectiveness measure (a weighted product of precision and recall). Training sets were first developed using a lengthy iterative process. Thereafter, a novel cost function based on effectiveness is proposed for training that allows us to achieve dramatic improvements in segmentation performance, as well as faster, non-iterative selection of training examples. The classifications produced using this cost function were sufficiently good that the binary image produced by the MLP could be post-processed using an active contour model to provide an accurate segmentation of the lungs from the multiple slices in almost all cases, including unseen slices and subjects.

Keywords: Magnetic resonance imaging; Image segmentation; Neural networks; Active contour models

1. Introduction

Segmentation has been defined [1:p. 347] as the process of: "dividing the image into regions that … correspond to structural units in the scene or distinguish objects of interest". It is a necessary first step in the visualisation and interpretation of many complex images, such as those typically encountered in medical imaging. In this area, fully automatic and robust segmentation techniques would have an enormous beneficial impact on clinical practice and research, by decreasing dramatically the manual effort which must otherwise be devoted to this task. Not only are medical images themselves inherently complex, but acquisition must also recognise practical needs to limit radiation dose, scan time, etc., so that image quality is often compromised. Given this, deployment of conventional image-processing techniques has not so far led to a robust fully automatic solution usable in a range of clinical settings, although semi-automatic systems do exist.

Semi-automatic segmentation has been used extensively in nuclear medicine based on thresholding or gradient techniques: both two- and three-dimensional techniques have been described [2]. Medical image-processing systems such as ANALYZE (from the Mayo Foundation, Rochester, MN) also include tools for partially automating manual segmentation. Fully automatic segmentation is possible in specific instances, such as thresholding to identify bone in computer tomography

* Corresponding author. Tel.: +44-1073-594577; fax: +44-1073-594498.

E-mail addresses: ianmid@microsoft.com (I. Middleton); rid@ecs.soton.ac.uk (R.I. Damper).

1350-4533/$30.00
doi:10.1016/S1350-4533(03)00137-1

(CT) images [3]. Various new approaches look to have considerable potential for automatic segmentation in more general applications (e.g. [4,5]) although their use in clinical practice has yet to be proven. Thus, segmentation remains the "image-processing bottleneck" [6].

The method proposed here is developed and illustrated on the practical problem of segmenting lung outlines from magnetic resonance (MR) images of the thorax. It consists of two stages. First, a neural network (multilayer perceptron, MLP) trained in supervised fashion is used to classify each pixel of the MR image into *boundary* and *non-boundary* classes, so producing a binary, edge-point image. Second, to compensate for classification errors, the edge-point images are then post-processed using an active contour model, or 'snake' [7,8]. In this way, the edge-point image acts as the external energy function for the snake. A similar combination of MLP classifier and active contour model has previously been used to locate the interior contour of the brain from MR images of the head [9]. However, the initial classification achieved by the neural network in that work was relatively poor and required a rather complex model-based active contour technique (using a stochastic decision mechanism based on a Gibbs sampler) to extract the final boundary. In this work, we aim to produce a sufficiently good initial classification to be able to use a simple and standard snake as the post-processor.

Early results for the classification stage, using data from a single subject and a restricted number of slices, were reported by Middleton and Damper [10], and showed good segmentation of the lung boundaries in a given MR image of the torso. Unfortunately, however, generalisation to other (unseen) slices and subjects was very much worse. We have since shown that an elastic net [11] modified to give robustness against initial classification errors, can be used to extract the region of the lungs very effectively from some of these classifications [12]. However, many of the results from the classification stage were too poor for the lungs to be accurately identified in this way. In the present work, several modifications and improvements have been made to our earlier work, which allow a much more accurate classification to be achieved from unseen MR data from different slices and different subjects. In particular, a novel cost function is proposed that simplifies the process of selecting the training data. Further, automatic exclusion of some pixels from the training set leads to dramatic improvement in classification. Consequently, the lungs can now be successfully segmented from the vast majority of available images using a standard active contour model, for which purpose we use the Cohen snake [13].

The remainder of this paper is structured as follows. The next two sections describe the images used (Section 2) and review alternative segmentation techniques (Section 3). The purpose of the latter section is to illustrate the difficulties in segmenting MR images using conventional image-processing techniques and to motivate the use of a neural network. The initial approach to classification using an MLP is then described in Section 4. We then detail our method of quantifying the quality of the segmentation (Section 5). Section 6 presents preliminary results of MLP classification using the standard squared-error cost function in training the network, and details the steps that were found necessary to achieve reasonable results. In Section 7, we define a new cost function for training based on the measure used to quantify segmentation performance. Section 8 describes MLP training using the new cost function, and Section 9 presents classification results using this new function. Post-processing using the snake and the results of the final segmentation are given in Section 10, and Section 11 concludes.

2. Image data and labelling

MR imaging is a non-invasive method of acquiring precise anatomical information in the form of three-dimensional data sets (see [14,15] for good introductory treatments). The data used here consist of transverse slices of the thorax obtained from a 0.5 T MR machine from 13 subjects. All were healthy volunteers, identified hereafter by a two letter abbreviation of the subject's name. Fig. 1 shows two examples of slices from subject AC. The lungs are clearly visible in these images as two large, low intensity regions within the torso. The lung regions are similar in intensity to the background because both are air-filled. The other low intensity regions correspond to the great blood vessels. These have a low intensity because moving blood gives virtually no MR signal.

The slices in Fig. 1 are shown with the anterior aspect of the subject uppermost and with the left side of the subject on the right of the image (in accordance with medical convention). This view is used for the presentation of all images in the present work. The slices are numbered from zero upwards, i.e. slice 0 is the lowermost slice. Each data set is composed of approximately 35 slices, and each slice is composed of (256×256), 2-byte integer values corresponding to the T_1-weighted value of individual pixels [14].

Some applications have used multi-spectral data (i.e. T_1-weighted, T_2-weighted and proton density data) since this offers a greater potential for discriminating between different tissues (e.g., [16,17]). In the present work, only T_1-weighted images were obtained, because of the prohibitive cost of collecting multi-modal data for a research project. In addition, multi-spectral data sets typically have a lower resolution than single-channel data sets because of limits on acquisition time. T_1-weighting was preferred over T_2-weighting and proton

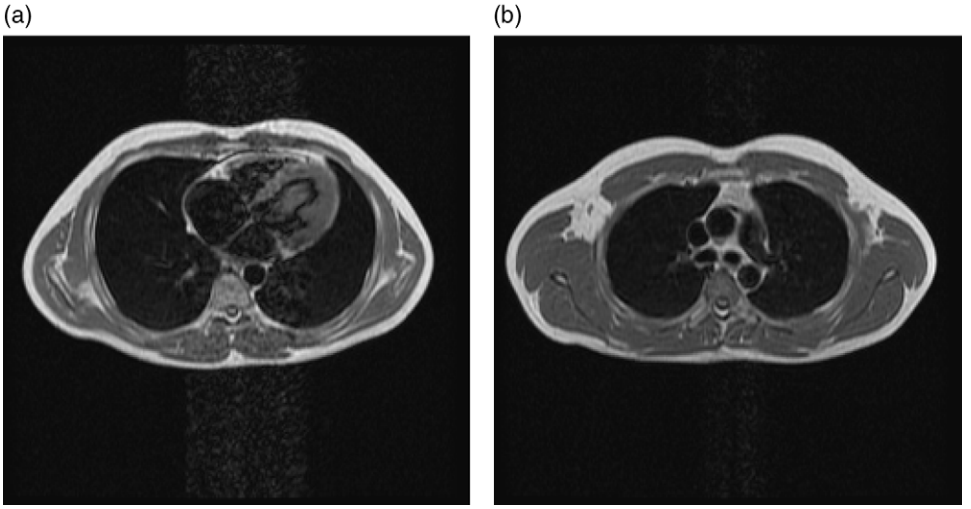

Fig. 1. Typical MR images of the torso. (a) Slice 12 of subject AC. (b) Slice 21 of subject AC.

density for several reasons, but mostly because T_1-weighting gives better contrast for present purposes.

There are several inherent problems in MR image segmentation. The main difficulty is the non-uniform nature of the MR signal intensity introduced by noise, physiological factors, partial volume effects and non-uniform radio frequency (RF) fields. The latter probably has the greatest influence on the intensity variations, and depends on a number of factors including the subject, slice orientation, RF coil design and pulse sequence [18]. Furthermore, there are particular difficulties involved in MR imaging of the lungs. These include cardiac and respiratory motion-induced artifacts that can hide fine structural detail [19]. An example of this can be seen in Fig. 1 (most clearly in Fig. 1(a)) as a central band of noise running from top to bottom of the image.

For (supervised) training and testing the neural network, we require some indication of 'ground truth' in the form of already-segmented, or labelled, images. Ideally, we would like to use unsupervised learning, yet the difficulty of medical image segmentation techniques is such that they "typically require some form of expert human supervision" [20:p. 437]. Here, outline images produced semi-automatically by an experienced radiologist have been used as the ground truth. The advantage of so doing is that "… it truly mimics the radiologist's interpretation, which realistically is the only valid truth available …" [18:p. 358]. On the other hand, we acknowledge the large inter-rater variability typically observed [21]. Indeed, somewhat ironically, this variability is one of the (several) motivations for the search for automatic methods.

Semi-automatic segmentation was done using the Mayo ANALYZE system. This involved growing connecting regions [22,23] from a defined seed position within a range of signal intensities, combined with manual tracing to prevent the region 'escaping' into non-lung areas. Once the region was correctly defined, all areas of the slice were erased leaving only the lung outline with non-zero intensity. Using this technique, it was possible to complete a segmentation in about 15 min, compared to at least 40 min for a completely manual process. However, the operator needed to check each slice, and frequent adjustments to both the seed position and the intensity range were necessary. A moderate amount of expert knowledge was needed, especially to distinguish between the lung and the great blood vessels (in so far as this was possible at all).

Fig. 2(a) shows a typical slice (number 18, subject AC) and Fig. 2(b) shows the corresponding segmented image produced semi-automatically as described above. In effect, this comprises a set of binary 1/0 labels for the MR image, and defines the target values for the supervised, back-propagation network training and/or assessment of network classification (see below). Hence, the task of the trained network is to produce a binary-

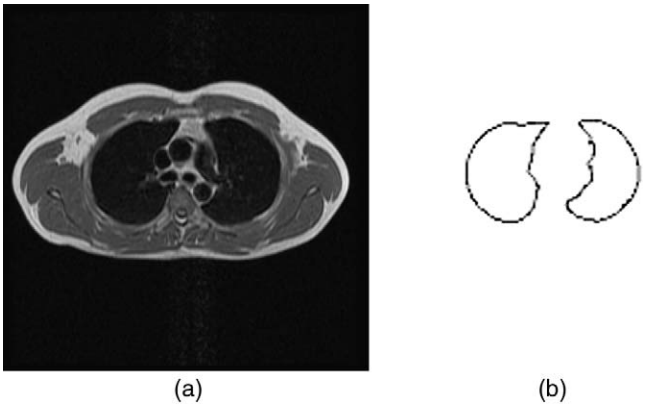

Fig. 2. (a) MR image: slice 18, subject AC. (b) Lung outlines produced semi-automatically (before manual correction) which act as 'ground truth' for network training and testing: same subject, slice and scale.

labelled outline like that in Fig. 2(b) when presented with *unseen* MR images like those in Figs. 1 and 2(a). These outline images are then post-processed by the active contour model (snake) to produce a closed representation of the lung boundary.

As the semi-automatically segmented images are subsequently used to act as ground truth for network training and testing, as well as for evaluation of the post-processing by the snake, it is worth considering how good they are. First, careful examination revealed inter-operator variability in the segmentations. Thus, only those produced by a single operator—a doctoral-level medical physicist specialising in MR imaging—have been used in this work. (See [24] for possible measures based on the opinions of multiple experts.) They were nonetheless found to contain a small number of errors, such as stray, misplaced or double boundaries—reflecting the difficulty of the labelling process. Any data used in network training were manually corrected by author I.M. but (because of the size of the task) this was not done for all data (including those used in evaluation). Note that Fig. 2(b) is depicted *before* manual correction with many errors clearly evident.

3. Alternative segmentation techniques

Difficulties such as those described above mean that standard image-processing techniques are often unable to segment MR images satisfactorily (unlike some other imaging modalities). For example, it is well documented that (unlike CT images) MR images cannot be segmented using histogram-based thresholding because of the non-uniform nature of the data [18,25]. To justify the approach taken here, various other standard image-processing techniques have been investigated for MR image segmentation.

One technique considered was to threshold a gradient-operated image in an attempt to identify the boundary of the target object(s)—here, the lungs. Not surprisingly, this was found to be inadequate, since the range of gradient values along the boundary of a target object generally overlap those along the boundaries of non-target objects. An alternative approach is to use shape matching to locate the boundary of a target object from an edge-detected image. The generalised Hough transform has been widely used for this purpose [26–29]. It can work in the presence of many non-target edges and tolerate any affine transformation of the target object. However, it can only be used where the basic shape of the target object (from which the observed shapes are produced by affine transformation) is well defined. The deformable nature of most internal structures means that this is not true for the majority of objects in medical images. This might suggest that an active contour would be a more suitable method of identifying the boundary of the target object. However, because there are many non-target boundaries in a typical image, it was found that an active contour had to be very carefully initialised to avoid it becoming trapped at non-target boundaries.

Ideally, a technique is needed that can distinguish the target boundary from the boundaries of other objects, yet is still deformable enough to represent the typical variations in shape that exist in medical images. This is achieved by the combination of MLP classifier and active contour model.

4. Initial classification using a neural network

Classification of each pixel of the image as either a boundary or non-boundary edge-point uses a multi-layer perceptron (MLP) trained on error back-propagation [30,31]. Since back-propagation is a supervised, gradient-descent technique, it requires labelled training data and some cost function which is differentiable, to give gradient information used in the search for a minimum-cost configuration of network connection weights. Initially, we have used the standard squared-error cost function. In this section, we consider the configuration of the MLP, the training regime for the network, and the classification rule used to produce a binary decision for the (continuous) network output.

4.1. Configuration of MLP

One advantage of the MLP as a classifier is that it can estimate the posterior probabilities required for Bayesian inference without the need for prior assumptions about the underlying probability distributions [32]. In addition, the MLP remains a popular neural classifier, which has been used successfully in a wide range of applications.

Fig. 3 shows the configuration of MLP used here. The output layer consists of a single node, whose activation determines the classification of the current input. There was a single hidden layer of 30 nodes. This was empirically determined, but at least one hidden layer was considered necessary since “MRI segmentation problems are nonlinearly separable and require multilayer networks” [33].

Initially, the inputs to the network consisted solely of the normalised intensity values of the pixels from the neighbourhood of the pixel to be classified. The size of this neighbourhood is defined by an $(m \times m)$ input window (where $m = 7$ in this work). This input window provides contextual information about the pattern of intensity values in the neighbourhood. Use of a context window is consistent with most image classification techniques in that only local image features are used (e.g., [9,34]). Preprocessing of the input data was deliberately limited to normalisation initially, since we wished to keep our approach generic and to avoid the

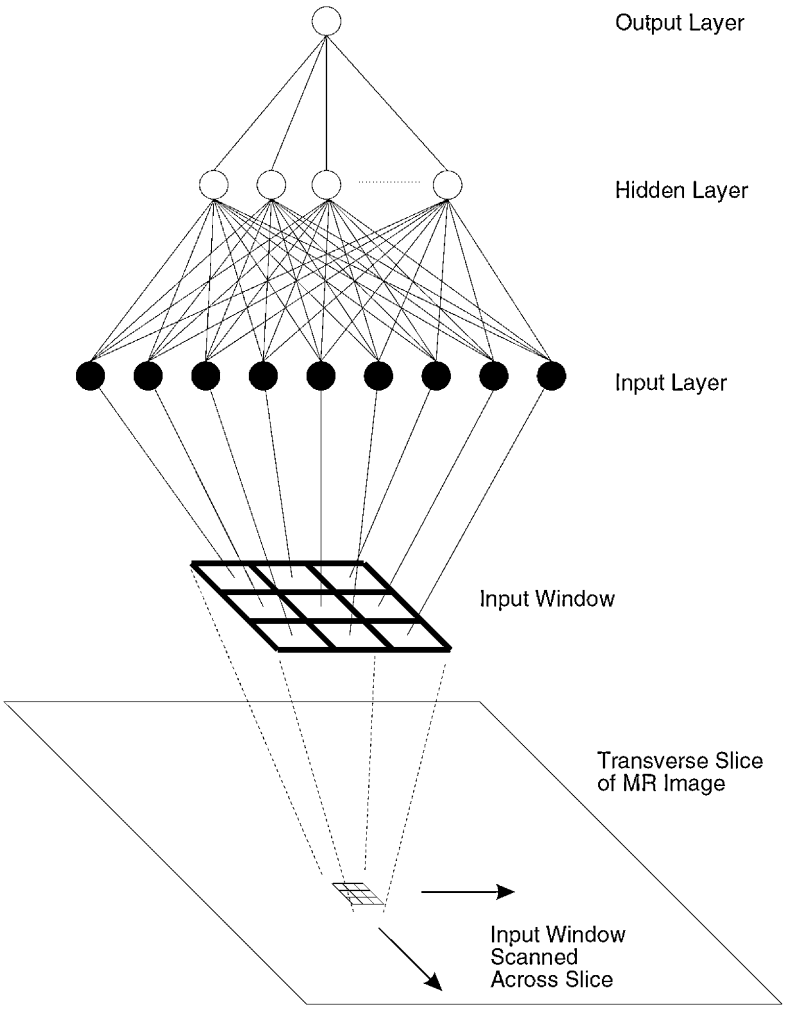

Fig. 3. Configuration of the MLP used to classify each pixel of an MR image as either a lung-boundary or non-lung-boundary example.

difficulty of determining an appropriate set of application-specific features (since a poor choice can have a substantial detrimental impact on performance [34]). Hence, our input scheme is simpler and more direct than that used by Chiou and Hwang [9], who explicitly calculated gradient information from a (9×9) window.

4.2. Training the network

The training data were labelled semi-automatically as detailed in Section 2. Using all the data in a selected slice for training is inappropriate as the data set would be extremely large. For $m = 7$, there are 3036 pixels at the image boundaries which cannot act as the centre of the window (because some of their 49 input cells will fall outside the boundaries of the image), so that there are $65{,}536 - 3036 = 62{,}500$ possible input/output pairs for each slice. (Although each of these has a unique centre pixel, they are not completely disjoint in the sense that many windows have overlapping context.) Consequently, training would be very slow. Furthermore, the boundary information would comprise only a relatively small fraction of the data set, as there are only some 500–600 lung-boundary pixels in the total of 62,500 (i.e. we would have a highly imbalanced training set). Consequently, the networks would be very likely to produce a constant, biased, negative ('no boundary') output regardless of input, corresponding to a false minimum of the squared-error cost function.

Initially, all the lung-boundary examples (i.e. all the (7×7) windows whose centre was a lung-boundary pixel) and an equal number of randomly selected non-lung-boundary examples were used for training. Hence, the size of the training set was approximately 1200 patterns. This is about 2% of the total number of possible windows but (because the windows are not disjoint) the network sees approximately 40% of the total number of pixels in the image in at least some position in a window. (This figure was obtained by averaging over several typical runs.)

The network was trained to produce a value of 1.0 for an input window with a target lung boundary at its

centre, and −1.0 otherwise. These values represent the extremes of the activation function used here (hyperbolic tangent). The cost function minimised during training was the standard squared error:

$$\zeta^2 = \sum_{s=1}^{S} \sum_{j=1}^{J_s} (o_j^s - t_j^s)^2 \quad (1)$$

where o_j^s is the network output for the window centred on the jth pixel of the sth slice and t_j^s is the labelled ('target') value for this input, J_s is the number of input windows for the sth slice, and S slices are used in training.

Learning and momentum rates for back-propagation learning were determined empirically. Results were found not to be especially sensitive to these parameters. We have used a learning rate of 0.02 and a momentum of 0.2. A fixed number of 1000 epochs of training was used throughout. This was found to be sufficient for the squared-error value to converge to a constant value.

It is well known that error back-propagation, in common with other gradient-descent search methods, is sensitive to initial conditions [35]. Hence, the networks used here were retrained several times (some 5–10) from different initial, random settings of the connection weights and the effect on results assessed. Although a few unusual results were obtained, the majority showed little variation in essential details. Hence, for clarity of presentation in what follows (e.g., to avoid having to give error bars for small numbers of trials), we choose to detail *typical* results.

4.3. Classification rule

Classification of the centre pixel of the input window during test on unseen data was simply determined by thresholding the network output. If the output was greater than the threshold, the input was classified as a lung boundary. Otherwise, it was classified as a member of the non-lung-boundary class. If the prior probabilities of each class are the same in both the training and test data, a threshold of 0.0 would be equivalent to assigning an input to the class with the highest posterior probability (as estimated from the network output). Performance is tested on the majority of the data from a given subject. Typically, this means the test data comprise 100 times more non-lung-boundary examples than lung-boundary examples. In contrast, the training sets used in this initial work consist of approximately equal numbers of lung-boundary and non-lung-boundary examples. This bias means that the prior probabilities of each class are not the same during training and testing. Thus, to assign an input to the class with the highest posterior probability, the threshold should be approximately 0.98 [32].

Clearly, a threshold of this value would minimise the number of misclassifications. However, our aim is to produce an accurate segmentation of the lungs from the classification produced by the MLP. When the threshold is set to minimise the number of misclassifications, the classified image tends to consist of just a very few correctly identified lung-boundary examples. This makes further processing to segment the lungs difficult. Consequently, a much lower threshold has been used: a value of 0.0 was chosen after some experimentation. From a risk minimisation stand-point, this is equivalent to assigning a cost to misclassifying a lung-boundary example that is 100 times greater than the cost of misclassifying a non-lung-boundary example [32].

5. Quantifying segmentation performance

To assess results, a method of measuring the accuracy of the segmentation techniques is required. This is a common problem in medical image segmentation [18]. Visual inspection is sometimes used to evaluate performance "since 'perfect' segmentations cannot be defined" [36:p. 341]. For instance, Brown et al. [37] assess the quality of chest CT segmentations through visual inspection by an experienced thoracic radiologist. Also, in the work of Chiou and Hwang [9], results are assessed impressionistically, by inspection, rather than quantitatively. (They could not have done otherwise since they did not label the complete contours in their images to provide a full picture of ground truth.) Visual inspection has not been used here for many reasons. Not only is it extremely subjective, we also wish to use our assessment measures as the basis of a cost function which will allow us to improve our network training methods. This makes quantitative measures—of the distance between the obtained segmentation and ground truth as defined by the semi-automatic segmentation (Section 2)—mandatory.

One potential quantitative measure of performance would be to use the rates of the two possible types of classification error. Taking a positive example to be a lung-boundary pixel, and a negative example to be a non-lung-boundary pixel, these are:

$$\text{False positive rate} = \frac{\text{No. of false positives}}{\text{No. of negative examples}}$$

$$\text{False negative rate} = \frac{\text{No. of false negatives}}{\text{No. of positive examples}}$$

There is, however, a considerable problem with these measures in the present work, stemming from the fact that there are many fewer lung-boundary pixels than non-lung-boundary pixels in each image slice. Hence, the denominators in the two cases are very different and the sensitivity of the two measures is incommensurate. Thus, a small change in the false positive error rate is

relatively more important than a comparable change in the false negative error rate, leading to difficulties in interpretation.

For this reason, we have used *precision* and *recall* to assess the quality of classifications [38]. These measures were devised for assessing information retrieval methods where a similar problem is encountered in extracting sources of information from a large database, in that there are typically many orders of magnitude fewer target sources than irrelevant sources. These measures are defined as:

$$\text{Precision } (P) = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}, \quad 0 \leq P \leq 1 \tag{2}$$

$$\text{Recall } (R) = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}, \quad 0 \leq R \leq 1 \tag{3}$$

Recall measures the proportion of the positive examples that are correctly identified, and is therefore one minus the false negative rate. Precision measures the proportion of the nominated positive examples that are correct. Thus, unlike the false positive rate, it is not dominated by the large number of non-lung-boundary examples, most of which can be easily classified correctly.

Using two measures of performance has some advantages in that it gives a separate measure of the two error types. To determine if one technique outperforms another, however, it is useful to combine these two measures into a single measure of goodness. van Rijsbergen [38] proposed the *effectiveness* measure for this purpose, defined as:

$$\text{Effectiveness } (E) = 1 - \frac{PR}{(1-\alpha)P + \alpha R}, \quad 0 \leq E \leq 1 \tag{4}$$

where $\alpha = 1/(\beta^2 + 1)$ and E is β times more heavily weighted towards recall than precision. In this work, precision and recall are equally weighted (i.e. $\beta = 1.0$). Since E is an inverse measure of goodness, we will generally quote segmentation performance in terms of $F = (1-E)$ in what follows.

Finally, we note that only slices which are labelled by the MLP as having at least 400 lung-boundary pixels were considered to contain the lung(s). Slices producing less than this number were excluded from the quantification of segmentation performance.

6. Results using squared-error cost function

Initially, the MLP was trained to identify the lung boundaries on a single slice ($S = 1$) of an MR image of the torso using the squared-error cost function (1). This showed that the method of selecting the training data as described in Section 4.2, using all lung-boundary examples plus an equal number of randomly selected non-boundary examples, led to a very poor classification [10]. In classifying the interior contour of the brain with a similar MLP to that used here, Chiou and Hwang [9] also randomly (and manually) selected their training examples and obtained poor quality results (see their Fig. 3(b):p. 1410). However, they did not propose any method to improve the performance of their MLP. Instead, they used sophisticated, model-based post-processing to achieve an acceptable segmentation of the desired boundary. We have found, however, that an iterative approach to selecting the training data can lead to very substantial improvements in the accuracy of the classifications, as we now describe.

6.1. Iterative selection of training examples

The iterative procedure is initialised by training a network on the selected target slice(s) as described in Section 4.2. The trained network is then used to classify the selected target slice(s), and a small, randomly selected proportion (0.1) of the resulting erroneous classifications are used to augment the initial training set, and the network is then retrained on this new training set. This process repeats until the quality of the classification reaches an acceptable level. In effect, a priori knowledge about the classification problem is implicitly being used in the selection of the training examples. This means that fewer examples are needed than would otherwise be the case [39]. In fact, training sets were successfully used that were much smaller than suggested by Widrow's rule of thumb, namely, that the number of training examples should be more than 10 times the number of weights in the network [40,41].

Table 1 shows typical performance of the MLP on the 13 subjects studied using this iterative procedure. The training data for this network were selected from every third slice of subjects AC, CB and LP, provided there were more than a criterial number of lung-boundary points in the slice (actually 400). In testing, *all* windows were tested from *all* slices, but only those slices where the MLP produced more than 400 edge-points were used in assessing performance. (Our assumption was that the lungs were not present in the remaining slices. Examination of these slices showed this to be a very reasonable assumption.) It can be seen from the table a reasonably high level of both precision and recall is obtained on the subjects used in training (marked with an asterix). The average segmentation performance for these 3 subjects is $\bar{F} = 0.630$. The performance on the remaining 10 subjects is generally poorer, $\bar{F} = 0.547$. This is 13% below the figure for the subjects whose data was used in training. Average segmentation performance across all 13 subjects is $\bar{F} = 0.566$. It is noticeable that precision is

Table 1
Typical values of precision, recall and F for MLP trained on data from subjects AC, CB and LP (shown with an asterix) and tested on all 13 subjects. Iterative selection of training examples was used as described in the text. Average segmentation performance across all 13 subjects is $\overline{F} = 0.566$. Precision is noticeably and consistently lower than recall

Subject	Precision	Recall	F
AC∗	0.573	0.782	0.661
CB∗	0.529	0.714	0.608
DB	0.478	0.582	0.524
JT	0.530	0.691	0.600
LP∗	0.532	0.742	0.620
NC	0.478	0.640	0.547
NH	0.419	0.592	0.491
SB	0.555	0.663	0.604
SO	0.463	0.573	0.512
SS	0.467	0.601	0.526
ST	0.520	0.579	0.548
SW	0.428	0.572	0.489
TF	0.551	0.737	0.630

consistently poorer than recall; the reason for this will become apparent later.

This preliminary work also revealed (results not shown) that the performance of the MLP improved with the size of the input window, and that an input window of dimension at least $m = 7$ was necessary to distinguish the target boundary from the boundaries of non-target objects such as the great blood vessels [10]. Hence, as previously stated, a (7×7) input window has been routinely used—since it provides a good compromise between accuracy and computational expense. The quality of the results attained thus far also suggest that the network is able to learn to extract the necessary features for the classification from the normalised intensity values in the input windows used in training. This justifies the approach, although there is still room for improvement.

6.2. Using spatial inputs

Initially, for the reasons given in Section 4.1, we only used normalised intensity values as inputs to the MLP. We suspected, however, that a spatial input reflecting the position of the (7×7) window might help classification. To test this intuition, the input layer was extended to include the (x, y) coordinates of the centre pixel of the input window. The origin of this coordinate system was taken as the centre of the slice, and the x and y distances were normalised to the same range as the intensity values, $[-1, 1]$. The z coordinate was not used, since the position of the lung in this dimension varies considerably from subject to subject. (For example, in subject AC the lungs extend between slices 5 and 27, whereas in subject JT the lungs extend between slices 11 and 35.)

Table 2 shows typical performance of the MLP with spatial inputs. The performance is generally better than the network without spatial inputs (cf. Table 1). There is an improvement in average performance, $\overline{F}$, from 0.566 to 0.597 by adding spatial inputs. The improvement is better for the 3 subjects used in training than for the remaining 10 subjects. (In fact, for 2 of the latter 10 subjects, performance actually got worse.) Using the non-parametric Mann–Whitney U test [42:pp. 137–144], the improvement in average performance due to including spatial inputs is marginally significant ($z = 1.462$, $p = 0.0719$, one-tailed test). Because there was a significant improvement (albeit marginal), spatial inputs continue to be used in what follows. (One should remember that the Mann–Whitney test is rather stringent because it makes no assumptions about the distributions of the two data groups.)

6.3. Treatment of pixels adjacent to the boundary

An issue which arose during this work was the correct treatment of pixels adjacent to the boundary. On one hand, ground truth suggests that these must be treated as negative examples, and this is what has been done so far. On the other hand, the low resolution of the MR images themselves and the nature of the semi-automatic labelling procedure used to define ground truth strongly suggest that one-pixel accuracy may be unattainable. Indeed, windows centred on pixels adjacent to the boundary may be very like those actually centred on the boundary. In fact, Chiou and Hwang [9] designated pixels immediately adjacent to the target boundary as positive examples, in contradiction of our practice thus far.

To assess the effect that these problematic cases were having on precision and recall, performance results were recomputed for the network of Table 2 by excluding all boundary-adjacent pixels from both training and testing.

Table 2
Typical performance values for MLP with spatial inputs. There is an increase in average performance, $\overline{F}$, from 0.566 to 0.597 by adding spatial inputs, which is marginally significant ($p = 0.0719$)

Subject	Precision	Recall	F
AC∗	0.649	0.803	0.718
CB∗	0.641	0.746	0.689
DB	0.546	0.604	0.574
JT	0.566	0.675	0.616
LP∗	0.618	0.739	0.673
NC	0.510	0.714	0.595
NH	0.412	0.496	0.450
SB	0.621	0.640	0.630
SO	0.515	0.587	0.549
SS	0.512	0.622	0.561
ST	0.543	0.723	0.620
SW	0.356	0.514	0.421
TF	0.610	0.724	0.662

That is, boundary-adjacent pixels were removed from the original training set and the network retrained precisely as before (from the same initial start point). The trained network was then retested, again with the boundary-adjacent examples removed from the test set. The recomputed values of *P*, *R* and *F* are shown in Table 3. Since the pixels which have been excluded from the training and test sets were all previously considered to be negative examples, there can be no change either in true positive outcomes or in false negatives. Hence, recall remains unaltered. However, precision is increased very considerably, showing that many of the previous false positives were actually boundary-adjacent pixels. Overall, average segmentation performance goes up to $\overline{F}$ = 0.703. The difficulty of 'correctly' classifying these pixels explains the rather low precision obtained relative to recall in earlier work.

These recomputed values must be interpreted with care. First and foremost, it is obviously impossible in actual practice to exclude test examples adjacent to the boundary when the boundary is itself unknown; it is the very thing we are trying to find in the test set. (We can do this for the training set, of course, but *not* for the test set.) Further, in no sense have we achieved an *improved* result since we have merely excluded from the test set difficult-to-classify negative pixels, which cannot do otherwise than improving precision while leaving recall unaltered. Rather, the exercise shows that the boundary-adjacent pixels are indeed especially problematic, and indicates that there might be an advantage in treating them differently. This will become important in the work described later.

Table 3
Performance values obtained by ignoring pixels adjacent to the target boundary in training and testing. (Subjects used for training are shown with an asterix.) Recall is unaffected but precision is considerably higher, leading to higher average performance ($\overline{F}$ = 0.703)

Subject	Precision	Recall	*F*
AC∗	0.831	0.803	0.817
CB∗	0.851	0.746	0.795
DB	0.818	0.604	0.695
JT	0.793	0.675	0.729
LP∗	0.860	0.739	0.795
NC	0.667	0.714	0.690
NH	0.664	0.496	0.568
SB	0.816	0.640	0.717
SO	0.713	0.587	0.644
SS	0.737	0.622	0.674
ST	0.710	0.723	0.716
SW	0.541	0.514	0.527
TF	0.831	0.724	0.773

7. Defining a new cost function

A potential advantage of the MLP classifier trained on squared error (Eq. (1)) or cross-entropy cost functions is that its output can be interpreted as an estimate of posterior probability [32:pp. 245–247]. However, the discussion in Section 4.3 indicated that this potential advantage is of limited value here, because of the imbalance of positive and negative examples in the test data. This suggests that there might be advantage to minimising during training a cost function more directly related to our measure of effectiveness, *E*, Eq. (4), which is a multiplicative combination of precision and recall, as defined via Eqs. (2) and (3), respectively.

The first step to defining a new cost function is to rewrite the expressions for precision and recall recalling that classification is actually determined using a threshold of 0.0:

$$P = \frac{\sum_i \delta_i H(o_i)}{\sum_i H(o_i)} \tag{5}$$

$$R = \frac{\sum_i \delta_i H(o_i)}{N} \tag{6}$$

Here, o_i is the network output for the input window centred on the *i*th pixel, $H(\ )$ is the Heaviside function equal to 1 when its argument is greater than or equal to 0 and equal to 0 otherwise, N is the number of lung-boundary examples and:

$$\delta_i = \begin{cases} 1 & \text{if } i \text{ is a lung–boundary pixel} \\ 0 & \text{otherwise} \end{cases}$$

Thus, precision and recall are not continuous in terms of the network output and so their derivatives cannot be calculated. Consequently, *E* as defined in Eqs. (2), (3) and (4) cannot be used as a cost function for network training since there is no gradient information on which to perform gradient descent. However, the Heaviside function can be approximated by the logistic function, $f(\ )$, since:

$$H(x) = \lim_{k\to\infty} f(k,x)$$

$$f(k,x) = \frac{1}{1 + \exp(-kx)}$$

Using this approximation in Eqs. (5) and (6), the expressions for precision and recall, and hence effectiveness, become continuous. The derivative of *E* with respect to the network output then becomes:

$$\frac{\partial E}{\partial o_j} \cong \frac{-((1-\alpha)P^2(\partial R/\partial o_j) + \alpha R^2(\partial P/\partial o_j))}{((1-\alpha)P + \alpha R)^2} \tag{7}$$

where

$$\frac{\partial P}{\partial o_j} \sim \begin{cases} \dfrac{kf(o_j)(1-f(o_j))\sum_i (1-\delta_i)f(o_i)}{(\sum_i f(o_i))^2} & \text{if } \delta_j = 1 \\ -\dfrac{kf(o_j)(1-f(o_j))\sum_i \delta_i f(o_i)}{(\sum_i f(o_i))^2} & \text{if } \delta_j = 0 \end{cases} \quad (8)$$

and

$$\frac{\partial R}{\partial o_j} \sim \begin{cases} \dfrac{kf(o_j)(1-f(o_j))}{N} & \text{if } \delta_j = 1 \\ 0 & \text{if } \delta_j = 0 \end{cases} \quad (9)$$

The closeness of the approximations in Eqs. (8) and (9) depends on the value of k, the scaling constant that determines the steepness of the logistic function. In this work, we use $k = 10$.

One important difference between the E cost function and standard cost functions is that the derivative $\partial E/\partial o_j$ depends on the performance of the network on *all* the examples in the training set, not just the current example, as can be seen from the explicit appearance of P and R in Eq. (7). This means the weight update rule will adapt to the current segmentation performance. For example, as the value of precision increases relative to the value of recall, the derivative of E becomes more weighted towards $\partial R/\partial o_j$. Similarly, as the value of recall increases relative to the value of precision, the derivative of E becomes more weighted towards $\partial P/\partial o_j$. This should result in classifications with a more equal balance between precision (or, strictly, $(1-\alpha)P$) and recall (or αR). This is intuitively reasonable; the goal is to minimise E and this measure was explicitly designed to require that (for $\alpha = 0.5$) a really low value can only be achieved by keeping P and R in balance (see later).

8. MLP training with the new cost function

In theory, the precision and recall should be recalculated each time the network is modified during training. For incremental learning, this would impose a severe computational burden, since weights are updated for each training example. If batch training was used, precision and recall would only have to be recalculated once per epoch. However, rather poor results were obtained using this batch method. Better results were obtained using incremental learning and approximate values of precision and recall—actually the value from the end of the previous epoch. This approximation is considered reasonable since precision and recall do not change significantly with each weight update.

It was quickly discovered that the training sets developed using the iterative technique described in Section 6.1 were unsuitable for use with the E cost function because they contain approximately equal numbers of positive and negative examples. This causes problems because E can be trivially 'minimised' with such a balanced training set by classifying all inputs as positive. In this situation, the number of true positives will be maximised and the number of false negatives will be zero. Hence, from Eq. (3), recall attains a maximum value of 1.0. Furthermore, although the number of false positives is also maximised, this number will only be approximately equal to the number of true positives (since the number of positive and negative examples are approximately equal). Thus, from Eq. (2), precision has a value of about 0.5 and, from Eq. (4), E has a value of about 0.33. This is a relatively low value for such a trivial solution and constitutes a false minimum with a wide basin of attraction during training. Consequently, training frequently became trapped in this false minimum.

Hence, when using the E cost function, training sets should contain at least an order of magnitude more negative examples than positive examples. In this case, classifying all inputs as positive would result in a very low value of precision, and consequently a very low value of E, so removing the false minimum. One way to ensure that the number of negative examples is much larger than the number of positive examples would be to use all the data from the slices used in training. This is possible when training on a limited number of slices, but is impractical for training on multiple slices since the size of the training set becomes too large. Further, a large proportion of the training data in these cases has a negligible effect on learning, because the network quickly learns to classify homogeneous regions (such as the background and lung interior in the torso images). Since these regions account for a large proportion of each slice, and the E cost function associates a negligible cost to correct classifications, the majority of the training data has a negligible effect after a few epochs. More precisely, the value of $f(o_j)$ (and therefore the value of $\partial E/\partial o_j$) is very small for output values on the correct side of the threshold used for determining classification. Thus, the weight modifications in such cases are very small.

Accordingly, we have removed pixels with an intensity gradient below a certain threshold from the training set, on the assumption that these constitute trivial examples. The intensity gradient was calculated as:

$$\text{grad}(x,y) = \sqrt{(I(x,y)-I(x+1,y))^2 + (I(x,y)-I(x,y+1))^2}$$

where $I(x, y)$ is the intensity of pixel (x, y) after the image was first smoothed by a (5×5) Gaussian filter with $\sigma = 0.5$. The threshold was determined empirically such that the majority of examples from the background, lung interior and other homogeneous regions were removed without excluding too many lung-boundary examples. This typically reduced the size of the training sets by a factor of about six; however, the MLPs could not then be guaranteed to classify correctly examples from the homogeneous regions excluded from training. Therefore, slices were preprocessed to remove low gradient examples before classification (just as in training). Thereafter, these were simply assumed to be negative (non-lung-boundary) pixels.

Note that we have replaced the computationally expensive iterative construction of the training set (Section 6.1) which was necessary with the squared-error cost function by a very much simpler method. This is considered to be one of the advantages of the new cost function.

9. Classification results with the new cost function

Table 4 shows typical segmentation performance of a network trained using the E cost function with the (non-iterative) method of training set selection just described. To allow a fair comparison with earlier results, the network here was trained (with spatial inputs) on the same slices from subjects AC, CB and LP as used in producing the training data for the squared-error cost function (see Section 6.1).

These new results indicate that performance is comparable to that obtained using the squared-error cost function (cf. Table 2). The relevant values of $\overline{F}$ are 0.589 here and 0.597 previously. Using the Mann–Whitney U test, there is no significant difference ($z = 0.4359$, $p = 0.6630$, two-tailed test) between the two means. Although we have not obtained any improvement in average segmentation performance, it is striking that precision and recall are now very much closer, as expected from the discussion at the end of Section 7. According to the U test, there is no significant difference between precision and recall here ($z = 0.897$, $p = 0.3694$, two-tailed test). This is in marked contrast to the earlier results for the squared-error cost function, where precision is very significantly lower than recall ($z = 2.590$, $p = 0.0048$, one-tailed test).

The failure to achieve any improvement in spite of what we expected to be a superior cost function is, we think, primarily because the E cost function uses *all* the data for which the gradient threshold is exceeded, as described in the previous section. This is highly likely to include most or all of the examples adjacent to the ground truth boundary, which we know to be problematic (Section 6.3). Previously, using the iterative procedure to build the training set, only a small proportion (0.1) of erroneously classified examples were added back into the new training set. Thus, we believe the difficult-to-classify boundary-adjacent pixels are more heavily represented in the training set when using the E cost function.

There is, however, no problem in removing these problem cases from the *training* set since, in this case, the boundary is known. Of course, they cannot be removed from the test set(s) and have to be classified just like any other pixel whose intensity gradient exceeds the threshold for inclusion.

Table 5 shows typical performance of a network

Table 4
Typical performance values for MLP with spatial inputs and trained using the E cost function. Here, the average segmentation performance is $\overline{F} = 0.589$, comparable to the result using the squared-error cost function ($\overline{F} = 0.597$)

Subject	Precision	Recall	F
AC∗	0.718	0.741	0.729
CB∗	0.701	0.675	0.688
DB	0.588	0.547	0.567
JT	0.616	0.609	0.612
LP∗	0.682	0.690	0.686
NC	0.480	0.650	0.553
NH	0.460	0.517	0.487
SB	0.677	0.573	0.621
SO	0.573	0.497	0.532
SS	0.576	0.548	0.562
ST	0.461	0.657	0.542
SW	0.447	0.381	0.411
TF	0.684	0.658	0.671

Table 5
Typical performance values for MLP trained using the E cost function ignoring examples adjacent to the target boundary. Average performance is now $\overline{F} = 0.749$, an enormously significant increase on the value of 0.589 obtained by including boundary-adjacent pixels in the training set

Subject	Precision	Recall	F
AC∗	0.877	0.798	0.836
CB∗	0.898	0.771	0.830
DB	0.865	0.684	0.736
JT	0.863	0.684	0.764
LP∗	0.900	0.781	0.836
NC	0.767	0.653	0.705
NH	0.728	0.609	0.663
SB	0.829	0.643	0.724
SO	0.770	0.605	0.677
SS	0.826	0.666	0.738
ST	0.816	0.689	0.747
SW	0.629	0.556	0.590
TF	0.882	0.765	0.819

trained using the E cost function excluding examples adjacent to the target boundary. The training data were taken from every third slice of subjects AC, CB and LP that contained a significant number (>400) of target-boundary examples. It is clear that there is now a very significant improvement due to excluding boundary-adjacent pixels from the training set. Average performance, $\bar{F}$, increases from 0.589 to 0.749. This result is enormously significant according to the Mann–Whitney U test ($z = 4.08$, $p \sim 0$). In spite of the great improvement overall, we note that generalisation is not relatively better; performance on subjects not used for training is still 13% poorer than that on subjects providing the training data (as before).

Perhaps the most striking aspect of Table 5 is the remarkable increase in recall. We previously found that excluding boundary-adjacent pixels for training/test sets with the squared-error cost function led to a sizeable increase in precision (compare Tables 2 and 3), although this finding was neither practical (we cannot remove boundary-adjacent pixels from the test set without knowing the correct test result in advance) nor surprising (precision must increase when we exclude negative examples from the test set). Here, however, we have merely removed some of the training data, which is an entirely practical thing to do. The vastly improved results shown in Table 5 can, we believe, be attributed to (i) an increase in precision due to eliminating boundary-adjacent examples from the test set, plus (ii) the dynamics of training using the E cost function which act to keep precision and recall in balance, as seen earlier in this section, so yielding a very significant improvement in recall. As we shall show in the next section, these results are sufficiently good to produce an accurate segmentation of the lungs in nearly every slice of the 13 subjects after post-processing with a standard active contour model.

10. Post-processing with an active contour model

The initial segmentation by an MLP classifier produces an edge-point image of candidate boundary points which can never realistically give an acceptable closed contour. False negatives will lead to gaps in the contour (especially where the great blood vessels join the lungs and image evidence for a boundary is low or absent) and false positives will arise and need to be eliminated. Therefore, post-processing is required to close these gaps and to distinguish false positives from true positives.

An ideal candidate for this post-processing is an active contour model [7,8], or snake, since such models produce closed contour descriptions and their deformable nature is appropriate for most internal structures. In recent years, snakes have assumed great popularity in medical image segmentation [20,43–46].

In this work, a Gaussian smoothed version of the edge-point image produced by the MLP can act as the external energy of the active contour model. The model chosen here is a Cohen snake, of the kind that has been successfully used in a number of medical imaging applications (e.g., [13,47,48]). One of the main reasons for choosing the Cohen snake is its use of a deflationary normal force to drive the contour towards the target boundary. This means that the Cohen snake is much less sensitive to initialisation than the Kass et al. snake [7,13]. Consequently, the snake could simply be initialised as one of two large circles encompassing the potential region of either the left or right lung.

Two important modifications to the basic Cohen snake were used here. First, the snake was made scale invariant. This was done by recalculating the inter-unit distance (and therefore the pentadiagonal regularisation matrix) at the end of each iteration to take into account the snake's new length. A scale invariant internal energy term (for governing curvature) was also used. The scale invariance was introduced because of the difficulty of finding parameters for the snake that could cope with the variation in size of the lungs between different slices and subjects. The other important modification was the use of resampling. Without resampling, the distribution of discrete points representing the snake contour can become highly irregular [49].

Table 6 shows the accuracy of the segmentations achieved using the snake described above for each of the 13 subjects. The segmentations are based on processing typical output of the network trained using the E cost function with boundary-adjacent examples removed from the training set. Results are presented in terms of precision and recall, but (unlike earlier) these values are calculated from the region enclosed by the snake. That is, in Eqs. (2) and (3), a pixel which is inside the contour for both the test result and the ground truth is counted as a true positive. False positive pixels are inside the lung boundary found by the snake but outside the ground truth boundary; false negative pixels are outside the lung boundary found by the snake but inside the ground truth boundary. This method of quantifying the results is felt to be a fairer indication of performance, since a snake could give a very good segmentation of the lungs without being located *precisely* along the boundary indicated by the ground truth segmentation. (It could 'miss' by one pixel at all points.) Note, this approach could not have been used to assess the MLP output because this did not provide a closed boundary.

Individual performance values are presented for the left and right lungs, since they were segmented using snakes with slightly different parameters. This is sensible as the two lungs are separate objects, with slightly different anatomical characteristics. Average performance, $\bar{F}$, is 0.866 and 0.844 for the left and right lungs, respectively. We consider these results to be highly encour-

Table 6
Results of post-processing typical MLP output for each of the 13 subjects using a Cohen snake.

	Left Lung			Right Lung		
Subject	Precision	Recall	F	Precision	Recall	F
AC∗	0.973	0.867	0.917	0.932	0.856	0.892
CB∗	0.903	0.769	0.831	0.931	0.847	0.887
DB	0.866	0.795	0.829	0.921	0.835	0.876
JT	0.951	0.865	0.906	0.961	0.811	0.880
LP∗	0.973	0.776	0.863	0.959	0.881	0.918
NC	0.884	0.806	0.843	0.904	0.871	0.887
NH	0.901	0.906	0.903	0.870	0.787	0.826
SB	0.987	0.747	0.850	0.819	0.598	0.691
SO	0.886	0.715	0.792	0.886	0.624	0.732
SS	0.938	0.812	0.870	0.904	0.886	0.894
ST	0.966	0.884	0.923	0.874	0.880	0.877
SW	0.911	0.751	0.823	0.838	0.706	0.766
TF	0.953	0.866	0.907	0.934	0.780	0.850

Average performance, $\bar{F}$, is 0.866 and 0.844 for the left and right lungs respectively.

aging. The excellent performance and versatility of our method is illustrated in Fig. 4, which shows typical examples of the segmentations that can be achieved.

There are still a few slices that are poorly segmented, the majority being either at the top or the bottom of the lungs. Part of the problem is that these more extreme slices tend to be poorly represented in the training data, and are consequently more prone to misclassification. However, inadequacies in the snake are probably the main contributor to the poor results (especially for the lower slices). For example, at the lower end of the lungs, the cross-sectional shape can become very narrow. The Gaussian filter associated with the external energy function of the snake can cause such closely spaced contours to become blurred into a single energy minimum. Fig. 5 shows two examples of the effect this can have on the behaviour of the snake. In Fig. 5(a), part of the snake has collapsed onto a single contour. In Fig. 5(b), the snake has collapsed along the narrow region of the lung outline. This problem could be prevented using a Gaussian filter with a smaller value of σ, but this would also reduce the noise tolerance of the snake and lead to other slices being badly segmented.

The snake can also have trouble locating the boundaries at the top of the lung. In these cases, the cross-sectional shape of the lungs combined with any misclassifications can result in there being insufficient boundary points for the snake to consider the boundary a salient contour. Consequently, the snake collapses. The remaining poor segmentations are where gaps in the boundaries identified by the MLP are too large. In these cases, the deflationary force causes the snake to collapse through the gap in the boundary. Unfortunately such breaks in the boundary are unavoidable, since they generally correspond to regions where there is no consistent information for the MLP to identify the boundary correctly.

11. Conclusions

MR image segmentation is an important but inherently difficult problem in medical image processing. In general, it cannot be solved using straightforward, conventional image-processing techniques. The solution proposed here is to use a multilayer perceptron to form the external energy function for an active contour model ('snake'). Initial work used the conventional squared-error cost function for training the MLP. This showed that the MLP could classify the lung boundaries in MR images of the torso to a reasonable accuracy using only the intensity values from the (7×7) neighbourhood of the pixel to be classified, together with an iterative procedure to construct the training data set. It was also found that to ensure reasonable generalisation, training data had to be taken from several slices of different subjects. Spatial inputs were also found to result in a marginal improvement in segmentation performance.

By approximating precision and recall using logistic functions, it became possible to define a new cost function for training which is a close approximation to the measure (effectiveness, E) used to quantify segmentation performance. We showed that this new cost function acted to keep precision and recall in balance during gradient descent search (i.e. error back-propagation training). It also became possible to construct the training set automatically and non-iteratively, by using the local image gradient to exclude trivial, easy-to-classify examples.

Contrary to expectations, use of this new cost function

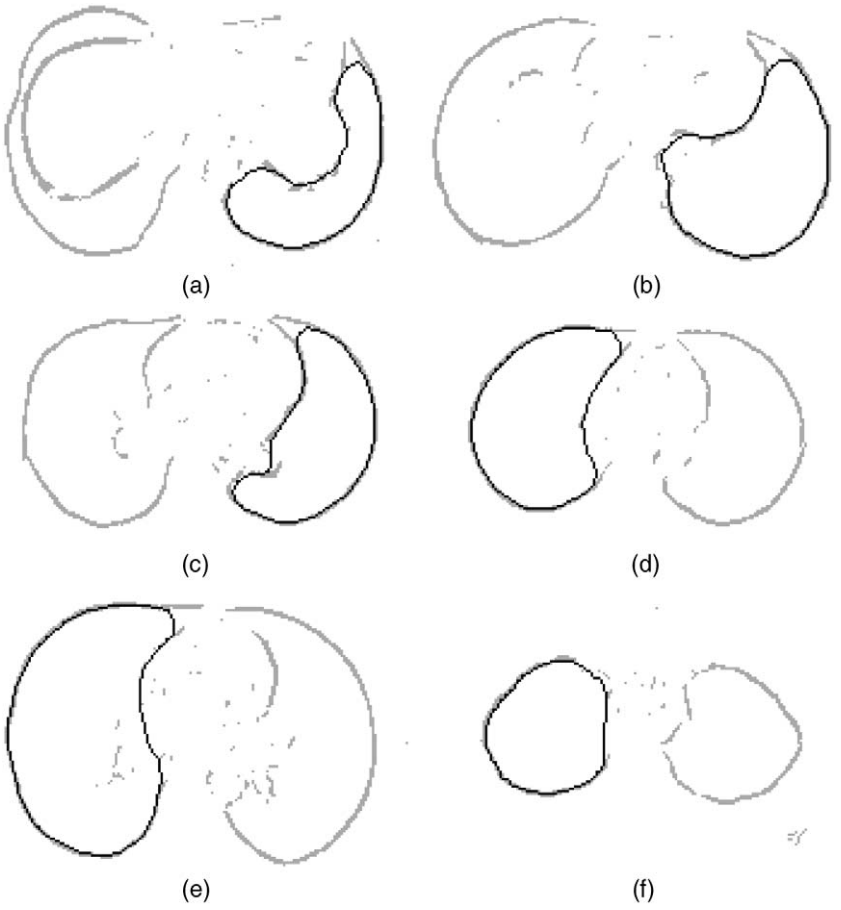

Fig. 4. Typical examples of the segmentation of the lungs. The final position of the snake is shown superimposed on the thresholded output of the network. (a) Slice 7 of subject AC: $P = 0.992$, $R = 0.873$, $F = 0.929$. (b) Slice 18 of subject NH: $P = 0.834$, $R = 0.980$, $F = 0.901$. (c) Slice 16 of subject AC: $P = 0.997$, $R = 0.903$, $F = 0.948$. (d) Slice 18 of subject TF: $P = 0.990$, $R = 0.946$, $F = 0.967$. (e) Slice 24 of subject JT: $P = 0.997$, $R = 0.943$, $F = 0.969$. (f) Slice 38 of subject DB: $P = 0.974$, $R = 0.936$, $F = 0.955$.

Fig. 5. Examples of the problems involved in segmenting the lower slices. Here, the cross-sectional shape of the lungs is too narrow for an accurate segmentation. (a) Slice 6 of subject AC: $P = 0.992$, $R = 0.778$, $F = 0.872$. (b) Slice 7 of subject AC: $P = 0.998$, $R = 0.468$, $F = 0.637$.

did not lead immediately to better performance than obtained with the earlier squared-error function. Investigation indicated that this was because problematic pixels adjacent to the lung boundary were more heavily represented in the training set. Fortunately, these boundary-adjacent pixels could be easily removed from the training set since the position of the boundary is known in these data. Exclusion of these problematic pixels resulted in an enormously significant improvement in performance ($p \sim 0$, Mann–Whitney U test).

Subsequent post-processing using a more-or-less standard Cohen snake achieved (in most cases) a very accurate closed-contour segmentation of the lungs for most slices. This is in contrast to previous work [12] in which we obtained good performance only for some slices using a non-standard modification to the Durbin–Willshaw elastic net [11]. Furthermore, many of the poor segmentations could be improved by modifying the active contour model while retaining generality. For example, the Cohen snake can be run on all the slices of a subject simultaneously and inter-slice constraints used to impose a degree of axial uniformity [48], as was also done by [12]. Provided the false positives are less well correlated between slices than the true positives, this modification should improve the snake's performance. It should be particularly effective with the torso images, since many of the poorly segmented slices are surrounded by good segmentations.

The technique presented here has shown a very encouraging level of performance for the problem of

lung segmentation in MR images of the torso. Efforts were made to reduce the amount of a priori knowledge used, so as to keep the method as generic as possible. This makes the approach worth serious consideration for further development as an automatic tool for image segmentation in medicine.

Acknowledgements

We are grateful to Dr. Liz Moore and Prof. John Fleming for supplying the MR images used here and for valuable advice in connection with this work. Liz Moore provided the semi-automatic labellings of the lung outlines.

References

[1] Russ JC. The image processing handbook, 2nd ed. Boca Raton, FL: CRC Press, 1995.
[2] Fleming JS. Quantitative measurements for gamma camera images. In: Chandler ST, Thomson WH, editors. Mathematical techniques in nuclear medicine. York, UK: Institution of Physics and Engineering in Medicine and Biology; 1996. p. 21–46.
[3] Vannier MW, Hildebolt CF, Marsh JL, Pilgram TK, McAlister WH, Shackleford GD et al. Craniosynotosis: diagnostic value of 3D CT reconstructions. Radiology 1989;173(3):669–73.
[4] Yan MXH, Karp JS. An adaptive Bayesian approach to three-dimensional MR brain segmentation. In: Bizais Y, Barillot C, Di Paola R, editors. Proceedings of 14th International Conference on Information Processing in Medical Imaging. Dordrecht, The Netherlands: Kluwer Academic; 1995. p. 201–13.
[5] Cootes TF, Hill A, Taylor CJ, Haslam J. The use of active shape models for locating structures in medical images. In: Barrett HH, Gmitro AF, editors. Proceedings of 13th International Conference on Information Processing in Medical Imaging. Berlin, Germany: Springer Verlag; 1993. p. 33–47.
[6] Stiehl HS. 3D image understanding in radiology. IEEE Eng Med Biol Mag 1990;9(4):24–8.
[7] Kass M, Witkin A, Terzopoulos D. Snakes: active contour models. Int J Comput Vision 1987;1(4):321–31.
[8] Blake A, Isard M. Active contours: the application of techniques from graphics, control theory and statistics to visual tracking of shapes in motion. London, UK: Springer, 1998.
[9] Chiou GI, Hwang JN. A neural network-based stochastic active model (NNS-SNAKE) for contour finding of distinct features. IEEE Trans Image Process 1995;4(10):1407–16.
[10] Middleton I, Damper RI. Segmentation of magnetic resonance images of the thorax by back-propagation. In: Proceedings of IEEE International Conference on Neural Networks, Perth, Western Australia, vol. 5. 1995. p. 2490–4.
[11] Durbin R, Willshaw D. An analogue approach to the travelling salesman problem using an elastic net method. Nature 1987;326(6114):689–91.
[12] Damper RI, Gilson SJ, Middleton I. A semi-localized elastic net for surface reconstruction of objects from multislice images. Int J Neural Syst 2002;12(2):95–108.
[13] Cohen LD. On active contour models and balloons. Comput Vision Graphics Image Process 1991;53(2):211–8.
[14] Webb S, editor. The physics of medical imaging. Bristol, UK: Adam Hilger; 1988.
[15] Wright GA. Magnetic resonance imaging. IEEE Signal Process Mag 1997;14(1):56–66.
[16] Amartur SC, Piraino D, Takefuji Y. Optimization neural networks for the segmentation of magnetic resonance images. IEEE Trans Med Imaging 1992;11(2):215–20.
[17] Cline HE, Lorensen WE, Kikinis R, Jolesz F. Three-dimensional segmentation of MR images of the head using probability and connectivity. J Comput Assisted Tomogr 1990;14(6):1037–45.
[18] Clarke LP, Velthuizen RP, Camacho MA, Heine JJ, Vaidyanathan M, Hall LO et al. MRI segmentation: methods and applications. Magn Reson Imaging 1995;13(3):343–68.
[19] Berthezene Y, Revel D, Bendib K, Croisille P, Amiel M. MR imaging of the lungs: clinical applications and potential research. J Radiologie 1997;78(5):347–51 (in French).
[20] Boscolo R, Brown MS, McNitt-Gray MF. Medical image segmentation with knowledge-guided robust active contours. Radiology 2002;22(2):437–48.
[21] Özkan M, Dawant BM, Maciunas RJ. Neural-network-based segmentation of multi-modal medical images: a comparative and prospective study. IEEE Trans Med Imaging 1993;12(3):534–44.
[22] Mayo Foundation. Biomedical Imaging Resource, ANALYZE reference manual, version 7C, section III. Rochester, MN; 1995. p. 119–135.
[23] Robb RA, Barillot C. Interactive display and analysis of 3-D medical images. IEEE Trans Med Imaging 1989;8(3):217–26.
[24] Chalana V, Kim YM. A methodology for evaluation of boundary detection algorithms on medical images. IEEE Trans Med Imaging 1997;16(5):642–52.
[25] Rajapakse JC, Giedd JN, DeCarli C, Snell JW, McLaughlin A, Vauss YC et al. A technique for single-channel MR brain tissue segmentation: application to a pediatric sample. Magn Reson Imaging 1996;14(9):1053–65.
[26] Ballard DH. Generalizing the Hough transform to detect arbitrary shapes. Pattern Recog 1981;13(2):111–22.
[27] Illingworth J, Kittler J. A survey of the Hough transform. Comput Vision Graphics Image Process 1988;44(1):87–116.
[28] Leavers VF. Which Hough transform? Comput Vision Graphics Image Process 1993;58(2):250–64.
[29] Kassim AA, Tan T, Tan KH. A comparative study of efficient generalised Hough transform techniques. Image Vision Comput 1999;17(10):737–48.
[30] Rumelhart DE, Hinton GE, Williams R. Learning representations by back-propagating errors. Nature 1986;323(9):533–6.
[31] Chauvin Y, Rumelhart D, editors. Backpropagation: theories, architectures and applications. Hillsdale, NJ: Lawrence Erlbaum Associates; 1995.
[32] Bishop CM. Neural networks for pattern recognition. Oxford, UK: Clarendon Press, 1995.
[33] Bezdek JC, Hall LO, Clarke LP. Review of MR image segmentation techniques using pattern recognition. Med Phys 1993;20(4):1033–48.
[34] McNitt-Gray MF, Huang HK, Sayre JW. Feature selection in the pattern classification of digital chest radiographic segmentation. IEEE Trans Med Imaging 1995;14(3):537–47.
[35] Kolen J, Pollack JB. Back-propagation is sensitive to initial conditions. Technical Report TR-90-JK-BPSIC, Department of Computer and Information Science, Ohio State University, Columbus, OH; 1990.
[36] Haring S, Viergever MA, Kok JN. Kohonen networks for multiscale image segmentation. Image Vision Comput 1994;12(6):339–44.
[37] Brown MS, McNitt-Gray MF, Mankovich NJ, Goldin JG, Hiller J, Wilson LS et al. Method for segmenting chest CT image data using an anatomical model: preliminary results. IEEE Trans Med Imaging 1997;16(6):828–39.
[38] van Rijsbergen CJ. Information retrieval, 2nd ed. London, UK: Butterworth, 1979.

[39] Bose NK, Liang P. Neural network fundamentals with graphs, algorithms and applications. New York, NY: McGraw-Hill, 1996.
[40] Widrow B. Adaline and madaline—1963: plenary speech. In: Proceedings of 1st IEEE International Conference on Neural Networks, San Diego, CA, vol. 1. 1987. p. 143–58.
[41] Baum EB, Haussler D. What size net gives valid generalization? Neural Comput 1989;1(1):151–60.
[42] Siegel S, Castellan NJ. Nonparametric statistics for the behavioral sciences, 2nd ed. New York, NY: McGraw-Hill, 1988.
[43] McInerney T, Terzopoulos D. Deformable models in medical image analysis: a survey. Med Image Anal 1996;1(2):91–108.
[44] Liang JM, McInerney T, Terzopoulos D. Interactive medical image segmentation with United Snakes. In: Medical Image Computing and Computer-Assisted Intervention, MICCAI'99, Cambridge, UK. 1999. p. 116–27.
[45] McInerney T, Terzopoulos D. T-snakes: topology adaptive snakes. Med Image Anal 2000;4(2):73–91.
[46] Pardo XM, Carreira MJ, Mosquera A, Cabello D. A snake for CT image segmentation integrating region and edge information. Image Vision Comput 2001;19(7):461–75.
[47] Cohen I, Cohen LD, Ayache N. Using deformable surfaces to segment 3-D images and infer differential structures. Comput Vision Graphics Image Process 1992;56(2):242–63.
[48] Cohen LD, Cohen I. Finite-element methods for active contour models and balloons for 2-D and 3-D images. IEEE Trans Pattern Anal Mach Intell 1993;15(11):1131–47.
[49] Ranganath S. Contour extraction from cardiac MRI studies using snakes. IEEE Trans Med Imaging 1995;14(2):328–38.