Classification Improvement by Segmentation Refinement: Application to Contrast-Enhanced MR-Mammography

Christine Tanner[1], Michael Khazen[2], Preminda Kessar[2], Martin O. Leach[2], and David J. Hawkes[1]

[1] Imaging Sciences, Guy's Hospital, King's College London, UK
[2] Clin. MR Section, Inst. of Cancer Research & Royal Marsden NHS Trust, Sutton, UK
christine.tanner@kcl.ac.uk

Abstract. In this study we investigated whether automatic refinement of manually segmented MR breast lesions improves the discrimination of benign and malignant breast lesions. A constrained maximum a-posteriori scheme was employed to extract the most probable lesion for a user-provided coarse manual segmentation. Standard shape, texture and contrast enhancement features were derived from both the manual and the refined segmentations for 10 benign and 16 malignant lesions and their discrimination ability was compared. The refined segmentations were more consistent than the manual segmentations from a radiologist and a non-expert. The automatic refinement was robust to inaccuracies of the manual segmentation. Classification accuracy improved on average from 69% to 82% after segmentation refinement.

1 Introduction

The development of computer aided diagnostic systems for MR mammography relies on the collection of ground truth information of the breast lesion's image position and extent. Currently, a radiologist's segmentation is the accepted gold standard for this definition. Manual segmentation is, however, very labour-intensive and prone to inaccuracies. The workload of radiologists often prohibit building large annotated databases.

While humans can rapidly perceive image objects, the exact definition of the object boundary is time consuming, especially for contrast-enhanced image sequences. Fully automatic segmentation, on the other hand, has proven more difficult than expected. We therefore aim to develop a semi-automatic method that reduces the segmentation workload significantly, that is also applicable for weakly enhancing structures and that can readily be applied to registered images.

A few automatic and semi-automatic segmentation algorithms have been proposed for the extraction of breast lesion from dynamic contrast-enhance MR images. Lucas-Quesada et al. [1] recommended segmentation of MR breast lesions by manually thresholding the similarity map, generated from the normalized cross-correlation between the time-intensity plot of each voxel and a reference plot derived from a small user defined region of interest (ROI). This method compared favourably to a multispectral analysis method, where 20 to 30 manually selected lesion voxels were used to generate a lesion cluster in the 2D pre- to post-contrast intensities space by means of the k-nearest

C. Barillot, D.R. Haynor, and P. Hellier (Eds.): MICCAI 2004, LNCS 3216, pp. 184–191, 2004.

neighbour algorithm. Note that the multispectral analysis method may have been disadvantaged by exploiting only limited data from the temporal domain, while the temporal correlation method was dependent on a single reference enhancement curve. Jacobs et al. [2] employed a k-means related clustering algorithm for extracting 4D feature vectors of adipose, glandular and lesion tissue from T1- and T2-weighted images and 3D fat-suppressed T1-weighted pre- and post-contrast images. Lesion classification was based on the angular separation from the adipose feature vector. Adipose and glandular reference feature vectors were provided by the user. Extraction of lesion outlines was not attempted. Fischer et al. [3] clustered the intensity enhancement profiles employing self-organizing Kohonen maps. The cluster results were shown to the user for interrogation of the dynamic sequences. No segmentation was attempted.

Our segmentation refinement method aims to extract the most probable lesion object of the contrast-enhanced MR image sequence from the data provided by the manual segmentation and prior knowledge about the segmentation process. The problem was posed as a two-class classification problem where the training data were provided by the manual segmentation. The segmentation decision was based on a constrained *maximum a-posteriori probability* (MAP) estimation in order to account for the imbalanced number of lesion and background voxels. The class conditional probability density functions were directly estimated from the temporal domain of the data samples. Sparsely sampled distributions were avoided by reducing the temporal dimensions with principle component analysis.

In the spatial domain, we observed that regions of non-enhancing tissue (like small heterogeneities, necrotic centres or fatty regions) were generally included in the manually segmented lesions. A MAP estimation solely based on the temporal domain would therefore lead to misclassifications. Instead, we rearranged the MAP estimation such that the ratio of prior class probabilities can be viewed as a threshold for the *likelihood ratio*. The segmentation process was then modelled by extract the biggest connected and filled lesion object for a given thresholded likelihood ratio map. The lesion candidate with the highest average a-posteriori probability that changed in size by less than a given limit was then selected as the most probable lesion. No assumptions were made about the edge strength or the shape or the enhancement profile of the lesion to avoid removing valuable information for the discrimination of benign and malignant lesions.

The aim of this study was two-fold. Firstly to assess the robustness and consistency of the segmentation refinement in comparison to the manual segmentations from an expert and a non-expert. Secondly, we compared the classification accuracy based on the segmentation refinement to that based on manual segmentation.

2 Materials

For this initial work we selected 18 patients from the symptomatic database of the UK multi-centre study of MRI screening in women at genetic risk of breast cancer (MARIBS) where patient motion was small enough allow interpretation of the images. The patients had in total 10 benign and 16 malignant histologically proven lesions. The images came from three centres of the MARIBS study and were all acquired according to the agreed protocol (3D gradient echo sequence on a 1.5T MR system with TR=12ms, TE=5ms,

flip angle=35^o, field of view of 340mm, 1.33x1.33x2.5mm^3 voxel size, coronal slice orientation, 90s acquisition time, 0.2mmol Gd-DTPA, see [4]).

The lesions were manually segmented by an experienced radiologist and a non-expert. The radiologist segmented the lesions by defining contours on the coronal slices of a selected post- to pre-contrast difference image. The radiologist had access to previous radiological reports to ensure that the correct lesion was segmented. Views of all original and all difference images of the dynamic sequence were provided. The non-expert segmented the lesions by employing region growing techniques from ANALYZE (Biomechanical Imaging Resource, Mayo Foundation, Rochester, MN, USA) with manual corrections if necessary. Generally, the same intensity threshold was applied to all slices while the seed voxel was moved. No information about the lesion location was provided to the non-expert. Eight missed lesions were segmented after comparison with the radiologist's segmentations. All manual segmentations were conducted without knowledge of the pathological results.

3 Methods

3.1 Data-Preprocessing

The image data was preprocessed by subtracting the mean of the two pre-contrast images from each post-contrast image. The sequence of subtracted images was then normalized to zero mean and unit variance for each 3D lesion ROI.

Many of the multispectral segmentation algorithms assume that the intensity distributions of the separate objects can be approximated with multivariate Gaussian distributions. There is, however, no reason to expect that the temporal data of MR mammograms conform to this assumption. Therefore we performed density estimations with Gaussian kernels and a bandwidth selected according to [5]. We reduced the dimensionality of the preprocessed data by principle component analysis to reduce sparseness.

3.2 Segmentation

The segmentation refinement aimed to extract the most probable connected lesion object of a 3D region of interest (ROI) for a given manual segmentation. The ROI was defined as the rectangular box extending the manual segmentation by 7mm in each direction. The problem was posed as a two-class classification problem where the training data was provided by the manual segmentation. The segmentation decision was based on the *maximum a-posteriori probability* (MAP) estimation in order to account for the unequal number of lesion and background voxels within the ROI.

Assuming equally likely image features x and taking the prior class probability $P(C_k)$ for class C_k into account, the most probable segmentation refinement is given by maximizing the a-posteriori probabilities, i.e. $argmax_k P(C_k|x) = P(x|C_k)P(C_k)$ where $P(x|C_k)$ was estimated from the manual segmentation. For a two class problem the discrimination function $y(x)$ can be written as

$$y(x) = \frac{P(x|C_1)}{P(x|C_2)} \quad \text{with} \quad x \in \begin{cases} C_1 & \text{if } y(x) > \theta \\ C_2 & \text{otherwise} \end{cases} \quad \text{where} \quad \theta = \frac{P(C_2)}{P(C_1)}. \quad (1)$$

Equation (1) emphasizes that the ratio of the prior probabilities act as a threshold (θ) on the *likelihood ratio*. Instead of estimating θ from the number of lesion and background voxels in the manual segmentation, we propose to use θ for implicitly incorporating prior knowledge about the segmentation process.

Assuming that one connected lesion was manually segmented per ROI we firstly extracted for a given threshold θ the biggest connected object. Thereafter we applied morphological closing and hole filling operations to model the observation that manually segmented lesions generally include non-enhancing regions. A set of lesion candidates was then generated by varying the threshold θ. Assuming that the manual segmentation is similar in size to the actual lesion object, we selected all candidates that had a volume change of less than a certain percentage compared to the manual segmentation. Of this subset we finally choose the object with the maximum average a-posteriori probability for the whole lesion. We tested ten threshold variations, namely **MAP**: $\theta = V(C_2)/V(C_1)$ with volume $V(C_k)$ estimated from input segmentation, **Tp**: connected filled lesion that changed volume by less than $p\%$ while maximizing the average posterior probability, tested for $p \in \{0, 10, 20, 30, 40, 50, 60\}$, **Tmax**: lesion with maximal average posteriori probability and **ML**: maximum likelihood decision $\theta = 0$.

Coarse input segmentations were simulated by approximating the manual segmentation by an ellipse on each 2D slice. The sensitivity to the size of the initial segmentation was assessed by changing the size of these ellipses by $s\%$ for all cases or by randomly selecting a size change $s\%$, with $s \in \{-33, -20, 0, 25, 50\}$.

Segmentations were compared by means of the overlap measure O = V($A \cap B$) / V($A \cup B$) where A and B are two segmented lesion regions; $A \cap B$ ($A \cup B$) are the intersection (union) of region A and B; and $V(C)$ is the volume of region C.

3.3 Feature Extraction

The size of our dataset limits the number of feature candidates that can reasonably be assessed. We therefore restricted ourselves to the 10 least correlated features of 27 previously reported 3D features used in this context [6,7,8]. These were selected by hierarchical clustering the feature vectors derived from the radiologist's segmentation according to their correlation. The 10 selected features were the following: *Irregularity* was determined by $irr = 1 - V_{in}/V$ where V_{in} is the volume within the effective radius $(3V/(4\pi))^{1/3}$. *Eccentricity* was determined by $ecc = \sqrt{a^2 - b^2}/a$ where $2a$ ($2b$) was the longest (shortest) axis, of the ellipsoid approximating the lesion shape. *Rectangularity* was defined as $rec = V/V_{rec}$ where V_{rec} is the volume of the smallest enclosing rectangular box. *Entropy of Radial Length Distribution* was calculated as $erl = \sum_{n=1}^{20} P_n log(P_n)$, where P_n is the probability that a surface vertex has an Euclidean Distance to the lesion's centre of gravity that lies within the nth increment of the distribution. *Peripheral-Central* and *Adjacent-Peripheral Ratio* were derived from partitioning the lesion and its immediate surrounding into 4 equally sized regions (central, middle, peripheral, adjacent) with boundaries kept in similar shape as the lesion itself. The ratios were given by $pcr = MITR(peripheral)/MITR(central)$ and $apr = MITR(adjacent)/MITR(peripheral)$ where $MITR$ is the maximum intensity time ratio as defined in [4]. *Slope Factor* m was derived from non-linearly fitting the general saturation equation $I_a/((T_{1/2}/t)^m + 1)$ to the average intensity difference

$I_t - I_0$ of the lesion where I_t is the intensity at time t after contrast infusion and I_0 is the pre-contrast intensity. *Texture parameters* were derived from the average *Spatial Gray-Level Dependence* matrix (one voxel distance, average of 9 directions) of the first post-contrast image intensities with values scaled from 1 to 40. *Spatial Correlation* is defined as $cor = \sum\sum(i - \mu_x)(j - \mu_y)P(i,j)/(\sigma_x\sigma_y)$ where $\mu_x = \sum iP_x(i)$, $\mu_y = \sum jP_y(j)$, $\sigma_x^2 = \sum(i - \mu_x)^2P_x(i)$, $\sigma_y^2 = \sum(j - \mu_x)^2P_x(j)$. *Angular Second Moment* is given by $asm = \sum\sum[P(i,j)]^2$. *Difference Average* was calculated by $dia = \sum\sum kP_{x-y}(k)$.

3.4 Classification

Single features were combined using stepwise linear discriminate analysis. The ability of combined features to discriminate between benign and malignant lesions was quantified by receiver operating characteristics (ROC) analysis. The ROC curve is defined by the fraction of false positives (1-specificity) to the fraction of true positives (sensitivity) for various thresholds on the decision criterion. The classification performance was summarized by the area under the ROC curve (A_{ROC}) for leave-one-out tests on a per-lesion basis. The statistical significance of the difference between ROC curves was tested using the ROCKIT program from Metz et al. [9].

4 Results

4.1 Data Preprocessing

The background (lesion) distributions of the original data were statistically significantly different (Lilliefors test, 5% level) from Gaussian normal distributions in 100% (67%) of all cases. The first two principle components of the preprocessed enhancement curves describe on average 98% of the variation in the data (range [91,100]%). The background (lesion) distributions of these two component were statistically significantly different from Gaussian normal distributions in 98% (69%) of all cases.

4.2 Segmentation

Fig. 1 illustrates for two neighbouring example slices the 3D segmentation refinement. It can be observed that the MAP refinement (contours in row 2) of the initial coarse segmentation (row 1) underestimated the extent of the lesion when compared with the radiologist's segmentation (row 6). The ML refinement (row 5) overestimated the lesion. The T0 (row 3) and Tmax (row 4) refinements produced almost identical results that are reasonable improvements to the radiologist's segmentation.

The average lesion size was statistically significantly bigger (paired t-test, $P<0.001$) for the radiologist's segmentation than for the non-expert (4.11ml vs 2.77ml). The average overlap between the two manual segmentations improved statistically significantly (paired t-test, $P<0.01$) after T0 refinement for all scenarios (from 59% to [64,68]%).

Fig. 2 shows how thresholding the likelihood ratio and changing the size of the initial segmentations affected the overlap measure O. Applying a threshold to keep

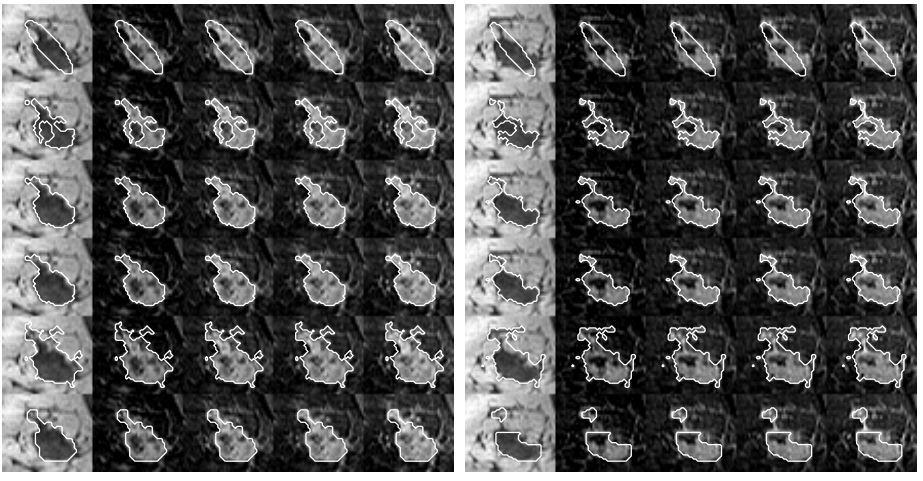

Fig. 1. Two example slices showing each (left to right) mean pre-contrast image and difference images after subtracting the mean pre-contrast image from 1st, 2nd, 3rd or 4th post-contrast image. Overlayed contours show (top to bottom) initial segmentation (E-20%), refinements of E-20% by MAP, T0, Tmax or ML criterion (see section 3.2); and radiologist's segmentation.

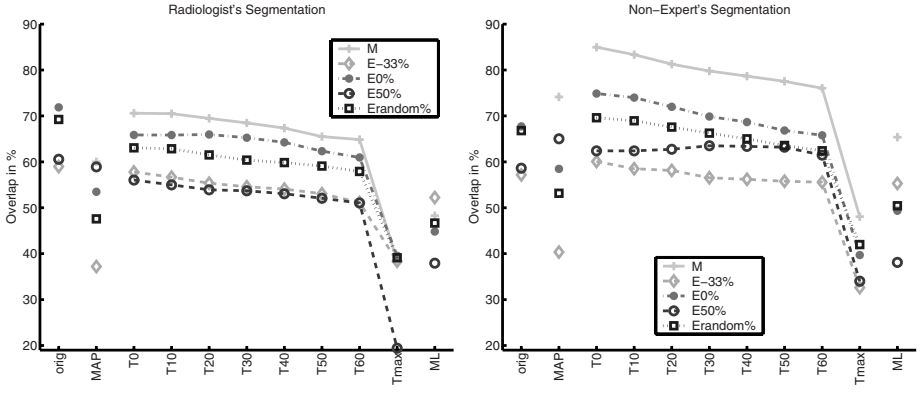

Fig. 2. Average overlap of refined and manual segmentation for (left) radiologist's and (right) non-expert's segmentation. Values along x-axis represent initial overlap (orig) and overlap after refinement with 10 threshold variations (MAP, T0-60, Tmax, ML) described in 3.2. Refinement was tested for manual segmentation (M) or for ellipsoidal approximations of manual segmentation that either all changed in size by $s\%$ (E$s\%$) or that randomly changed in size (Erandom%).

volume changes to a minimum (T0) provided on average the biggest overlap of all refinement strategies apart from input scenario E50%. Refinements of the ellipsoidal approximations (E$s\%$) produced generally smaller mean overlaps than refinements of the manual segmentations (M). The maximal possible average overlap when thresholding

the likelihood ratio was 73% and 86% for the two manual segmentations. Optimal thresholding of the temporal correlation maps (created with respect to the most enhancing 5x5 ROI within the lesion) resulted in a maximal mean O of 60% for both manual segmentations.

4.3 Classification

We compared the classification performance of features derived from manually segmented lesion to that from refined segmentations. A combination of 3 refinement strategies (MAP, T0, ML) and 5 initial segmentations (M, E-33%, E0%, E50%, Erandom%) was tested (see section 3.2). Tp ($p > 0$) and Tmax refinements were not assessed because their overlaps were either very similar to T0 or were very small. The assessment was based on the area under the ROC curve (A_{ROC}) created from leave-one-out tests.

The best feature extracted from the radiologist's segmentation was *Texture Correlation* with an A_{ROC} of 0.57. After segmentation refinement, the best feature changed to *Peripheral-Central Ratio* (pcr) with average A_{ROC} values of 0.40, 0.66, 0.74 for MAP, T0 and ML, respectively. The best feature for the non-expert's segmentation was already pcr (A_{ROC}=0.69) and remained so after refinement (mean A_{ROC} 0.42, 0.71, 0.82 for MAP, T0 and ML). The pcr mean values of benign and malignant lesions were statistically significantly different (pooled t-test, 5% level) in 90% of all input scenarios after T0 and ML refinement.

To avoid overfitting the data, we combined not more than two features during stepwise linear discriminate analysis. This resulted in A_{ROC} of 0.57 and 0.70 for features extracted from the radiologist's and the non-expert's segmentation, respectively. Classification based on the refined segmentations provided A_{ROC} values between 0.57 and 0.89, of which none was statistically significantly worse (ROCKIT, 5% level) than the results of the manual segmentation. The best results were achieved with the ML refinement strategy. It produced in all cases the highest A_{ROC} (mean 0.80, range [0.69,0.89]). For half of the initial segmentations it was statistical significantly better at the 5% level than the manual segmentations. The second best results were produced by T0 with A_{ROC} values between 0.53 to 0.76 (mean 0.70). MAP was on average not better than the manual segmentation (mean 0.64, range [0.52,0.77]).

5 Conclusion

We have shown that the refinement of manual segmentations based on thresholding the likelihood ratio map can significantly improve the discrimination of benign and malignant breast lesions from contrast-enhanced MR images. Simplification of the lesion delineation by 2D ellipses and change of lesion size before refinement did not lead to inferior classification results. The consistent classification success of the maximum-likelihood refinement strategy was surprising, given its low overlap measures and apparent overestimation of the lesions extent, and requires further investigations.

The overlap between a manual segmentation and its refinement was on average significantly higher for the non-expert. This is likely due to region growing being more similar

to the refinement technique than manual outlining. Classification results improved to a similar level for both manual segmentations after segmentation refinement.

Thresholding optimized for maximal overlap provided higher results for the likelihood map than for the temporal correlation map. This could be because the probabilistic approach removed the dependency on a single reference enhancement curve.

Computerized segmentation methods are generally evaluated against radiologist's manual segmentations. It is, however important to assess the effects of the segmentation on the ultimate goal, in this case the ability to discriminate benign and malignant MR breast lesion. To our knowledge, such a study has not been published, apart from evaluating the enhancement characteristics of region subsampling methods [10,11].

Classification of MR breast lesion based on step-wise linear discriminant analysis of extracted features from lesion segmentations has been reported previously [6,7,8]. These studies achieved classification accuracies of 72%, 79% (without leave-one-out tests) and 87%, respectively, when combining two features. Our classification results were on the lower end when based on features from the manual segmentations (69%) but improved to 78% and 85% after maximum-likelihood segmentation refinement. In future work, we will study how much segmentation refinement and registration improves classification for a large dataset.

Acknowledgements. CT acknowledges funding from EPSRC (MIAS-IRC). The image data were provided by MARIBS [4]. This study and MK are supported by the MRC (G9600413).

References

1. F. A. Lucas-Quesada et al., "Segmentation Strategies for Breast Tumors from Dynamic MR Images," *J Magn Reson Imaging*, vol. 6, p. 753, 1996.
2. M. A. Jacobs et al., "Benign and Malignant Breast Lesions: Diagnosis with Multiparametric MR Imaging," *Radiology*, vol. 229, p. 225, 2003.
3. H. Fischer et al., "Local Elastic Matching and Pattern Recognition in MR Mammography," *Int J Imaging Syst Technol*, vol. 10, p. 199, 1999.
4. J. Brown et al., "Magnetic Resonance Imaging Screening in Women at Genetic Risk of Breast Cancer: Imaging and Analysis Protocol for the UK Multicentre Study," *Magn Reson Imaging*, vol. 18, p. 765, 2000.
5. B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman Hall, 1986.
6. S. Sinha et al., "Multifeature Analysis of Gd-Enhanced MR Images of Breast Lesions," *J Magn Reson Imaging*, vol. 7, p. 1016, 1998.
7. K. G. A. Gilhuijs et al., "Computerized Analysis of Breast Lesions in Three Dimensions using Dynamic Magnetic-Resonance Imaging," *Med Phys*, vol. 1, p. 1647, 1998.
8. L. I. Sonoda, *Classification of Lesions in Magnetic Resonance Images of the Breast*. PhD thesis, King's College London, 2003.
9. C. E. Metz et al., "Maximum Likelihood Estimation of Receiver Operating Characteristics Curves from Continuously-Distributed Data," *Statistics in Medicine*, vol. 17, p. 1033, 1998.
10. S. Mussurakis et al., "Primary Breast Abnormalities: Selective Pixel Sampling on Dynamic Gadolinium-Enhanced MR Images," *Radiology*, vol. 206, p. 465, 1998.
11. G. P. Liney, P. Gibbs, C. Hayes, M. O. Leach, and L. W. Turnbull, "Dynamic Contrast-Enhanced MRI in the Differentiation of Breast Tumors: User-Defined Versus Semi-automated Regions-of-Interest Analysis," *J Magn Reson Imaging*, vol. 10, p. 945, 1999.