Automatic seed initialization for the expectation-maximization algorithm and its application in 3D medical imaging

M. LYNCH*, D. ILEA, K. ROBINSON, O. GHITA and P. F. WHELAN

Vision Systems Group, Dublin City University, Dublin 9, Ireland

Statistical partitioning of images into meaningful areas is the goal of all region-based segmentation algorithms. The clustering or creation of these meaningful partitions can be achieved in number of ways but in most cases it is achieved through the minimization or maximization of some function of the image intensity properties. Commonly these optimization schemes are locally convergent, therefore initialization of the parameters of the function plays a very important role in the final solution. In this paper we perform an automatically initialized expectation-maximization algorithm to partition the data in medical MRI images. We present analysis and illustrate results against manual initialization and apply the algorithm to some common medical image processing tasks.

Keywords: Image segmentation; Medical imaging; Statistical pattern analysis; Expectation-maximization

1. Introduction

Segmentation has been a key goal in imaging research for a number of decades. The applications of robust techniques for object classification in images are extensive, none more so than in the rapidly advancing field of medical imaging [1,2]. With the introduction of faster and more powerful imaging devices the amount of data produced makes it impractical for experts to manually segment objects of interest. The need for more automated methods of segmentation is evident. Medical scanners, such as MRI, utilize the metaphysical response of the body's organs to create an image. This response is tissue-dependent and therefore the resultant image is comprised of almost homogenous regions which are representative of organs, tissues or fluids in the body.

Region-based methods [2] are used to segment the image, normally using no *a priori* information. The most basic form of region-based segmentation is thresholding. Thresholding techniques create a binary image of pixels above and below a user-defined threshold value. Thresholding does not take into account the structure or connectivity of the points that it segments and the threshold value is seldom automatically determined. Segmentation results can sometimes be filled with holes or ragged edges, which in a crude way can be eliminated with a combination of morphological operators [3,4]. In medical imaging, thresholding is not widely used without advanced preprocessing steps due to its sensitivity to noise. More complex statistical methods, such as clustering, join pixels of similar intensities to create a segmentation of structures in the image. All statistical based classification methods [5–9] aim to optimize the results based on an initialization. This initialization is commonly chosen randomly, and as a consequence results are not reproducible, do not take advantage of inherent patterns in the data or may be initialized on outliers. Methods for automatic initialization of clusters have been proposed in the literature [10–12]. Al-Daoud and Roberts [10] proposed two methods, the first of which picks points randomly in evenly spaced cells across the entire histogram of the data and reduces the number until the required seeds are found. The second method tries to optimize the sum of squares of the distances from the cluster centres. Mitra *et al.* [11] describe a *rough-set* initialization provided by graph-theoretic methods. Khan and Ahmad [12] assumed a normal distribution over the data attributes and divided the normal distribution curve into equal percentile cells. The

*Corresponding author. Email: lynchm@eeng.dcu.ie seeds are chosen as the midpoints of the interval of each of these partitions.

In this paper we present a novel algorithm that automatically initializes the seeds used in statistical based classification algorithms. The advantage over previous implementations is that it is reproducible, robust and easy to implement. The algorithm firstly selects a large number of possible partitions, using peaks (local maxima) in the intensity histogram, which are evenly distributed over the data. The algorithm then performs an iterative clustering of these peaks, using their histogram heights and greyscale difference until the optimal number of seeds is reached. To verify the results from the initialization, the seeds picked were used as the initial estimates for a segmentation using the expectation-maximization (EM) algorithm. The segmentation results are given for both 2D and 3D data and common applications of segmentation in cardiac, brain and whole-body MRI are also presented.

2. EM algorithm

The EM algorithm [6,13] attempts to classify data using a soft membership function as a weighted sum of a number of Gaussian distributions called a Gaussian Mixture Model (GMM). The generation of this GMM is achieved through an expectation-maximization technique, which aims to find

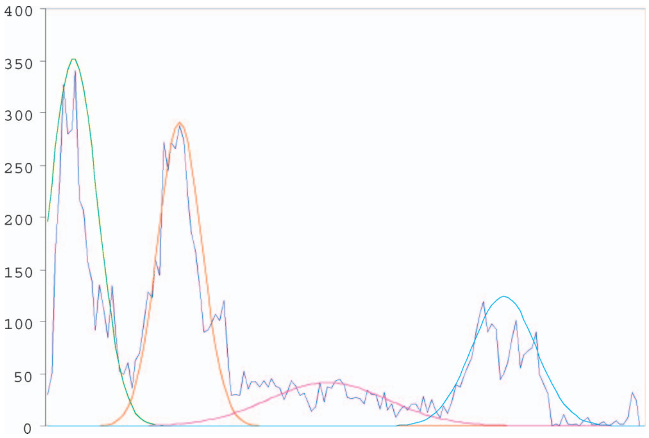
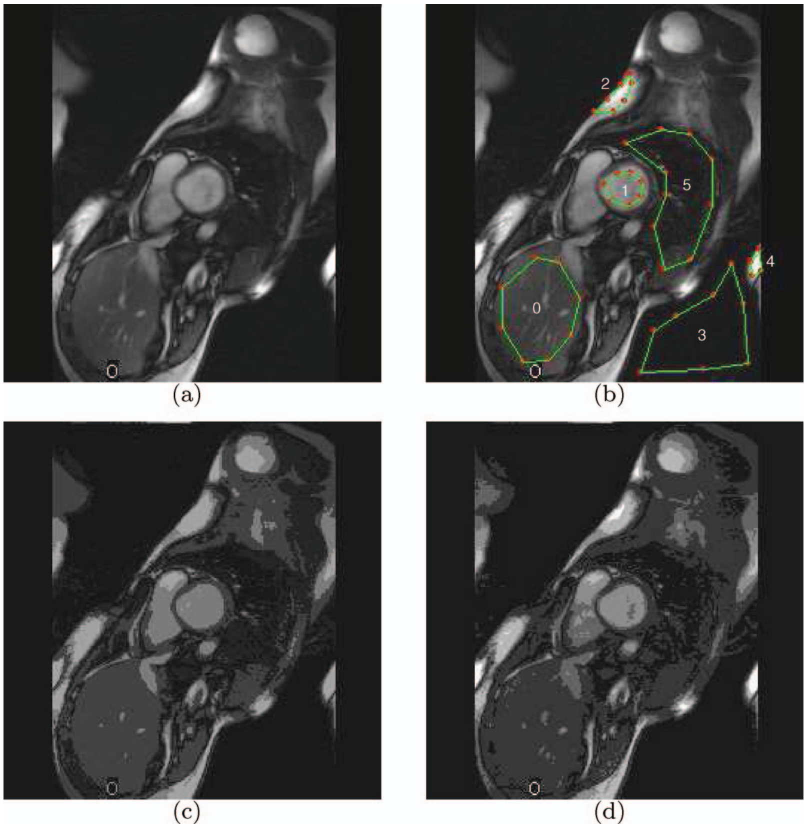


Figure 1. An illustration of the principle of signal intensity classification using a four-class Gaussian Mixture Model. (Scaled for illustration purposes.)



(a) (b) (c) (d)

Figure 2. Figures show the short axis view of cardiac MRI: (a) shows the original image, (b) indicates the manually selected areas, (c) represents the results after applying the EM using the manually picked initialization and (d) is the result after applying the automatic seed picking.

the maximum likelihood estimate for an underlying distribution from a given dataset when the data is incomplete. The basic idea of expectation-maximization is illustrated in figure 1.

The advantage of EM over the k-means clustering technique [8] is its ability to provide a statistical model of the data and its capability to handle the associated uncertainties. Consider the general case of a d-dimensional random variable $x=[x_1, x_2, x_3, \ldots, x_d]^T$ and suppose it follows a k-component finite mixture distribution. Its probability density function (pdf) could be written as:

$$p(x \mid \theta) = \sum_{m=1}^{k} \alpha_m p(x \mid \theta_m), \tag{1}$$

where α_m is the mixing parameter for each of the Gaussian distributions in the GMM and $\theta_m = \{\mu_m, \sigma_m\}$ are the parameters of the Gaussian distributions.

$$\alpha_m \geq 0, \text{ and } \sum_{m=1}^{k} \alpha = 1 \tag{2}$$

The algorithm is built on an iterative scheme and consists of two steps. The first, the E-step, calculates the expected log-likelihood function for the complete data, defined by Q using the estimates for the parameters $\hat{\theta}(t)$.

$$Q(\theta, \hat{\theta}(t)) \equiv E[\log p(X, Y \mid \theta) | X, \hat{\theta}(t)] \tag{3}$$

The second step, the M-step, uses the maximized values of this result to generate the next set of parameters.

$$\hat{\theta}(t+1) = \arg\max_{\theta} Q(\theta, \hat{\theta}(t)) \tag{4}$$

Table 1. Changes in cluster means in the whole body data. A: Manual μ; B: manual μ after EM; C: automatic μ; D: automatic μ after EM.

	A	B	C	D
$\mu(0)$	57.31914	55.2806	57	31.33457
$\mu(1)$	125.366	112.0961	137	125.284
$\mu(2)$	194.0437	151.1044	167	171.6872
$\mu(3)$	19.84193	16.74244	12	17.75531
$\mu(4)$	225.1899	112.8278	255	254.2933
$\mu(5)$	28.87568	28.43651	92	79.93145

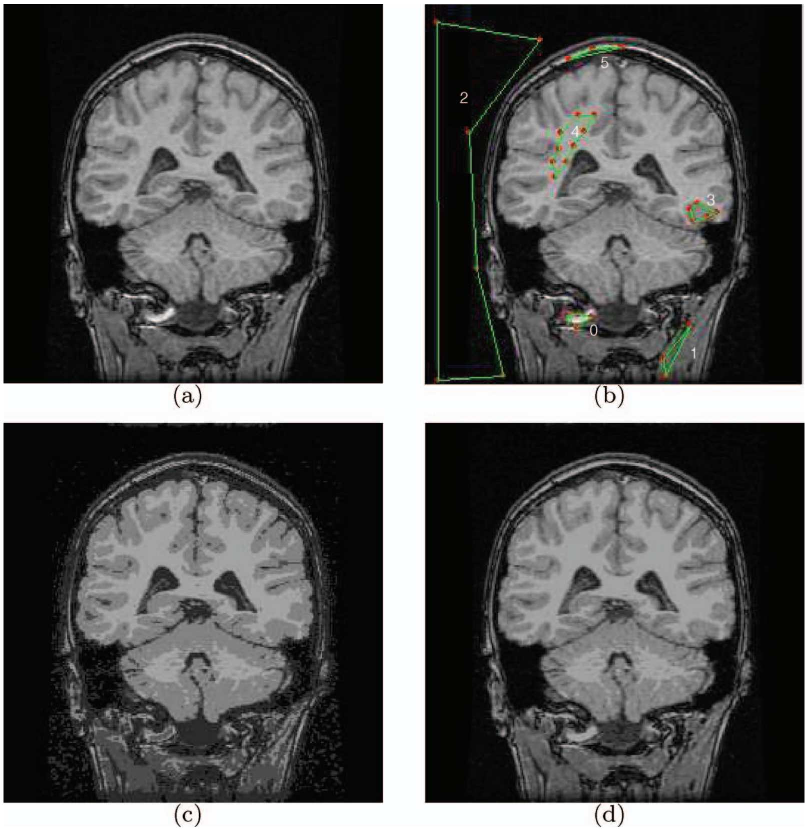


Figure 3. Figures show a coronal slice from a brain MRI: (a) shows the original image, (b) indicates the manually selected area, (c) represents the results after applying the EM using the manually picked initialization and (d) is the result after applying the automatic seed picking.

The algorithm iterates between equations (3) and (4) until convergence is reached. It is important to note that local convergence of the EM algorithm is assured [6,14,15].

The updates for the parameters for the GMM are the mixture values α_m and the parameters of the Gaussian distributions $\theta_m = \{\mu_m, \sigma_m\}$. These can be calculated from equations (5), (6) and (7).

$$\alpha_m^{\text{new}} = \frac{1}{N}\sum_{m=1}^{k} p(m \mid x_i, \hat{\theta}(t)) \tag{5}$$

$$\mu_m^{\text{new}} = \frac{\sum_{m=1}^{k} xi\, p(m \mid x_i, \hat{\theta})}{\sum_{m=1}^{k} p(m \mid x_i, \hat{\theta})} \tag{6}$$

$$\sigma_m^{\text{new}} = \frac{\sum_{m=1}^{k} p(m \mid x_i, \hat{\theta})(x_i - \mu_m^{\text{new}})(x_i - \mu_m^{\text{new}})^{\mathrm{T}}}{\sum_{m=1}^{k} p(m \mid x_i, \hat{\theta})} \tag{7}$$

2.1. Seed generation

This paper proposes a novel approach to initialization of cluster centres based on histogram analysis. A histogram

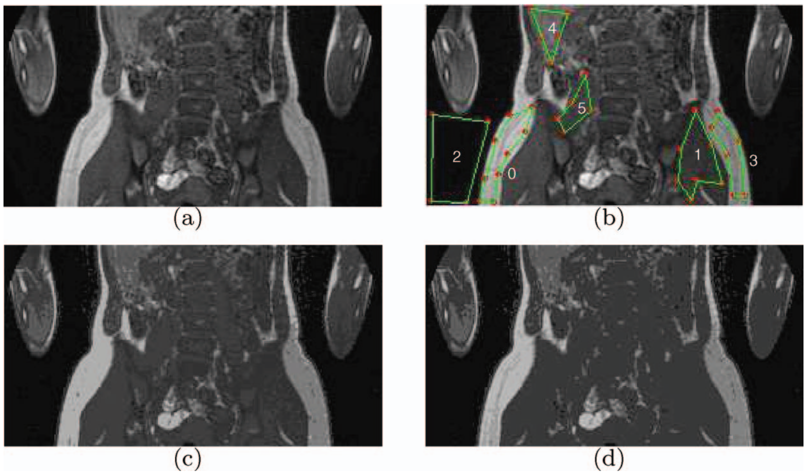


Figure 4. Figures show a coronal slice from a section of a full body MRI: (a) shows the original image, (b) indicates the manually selected areas, (c) represents the results after applying the EM using the manually picked initialization and (d) is the result after applying the automatic seed picking.

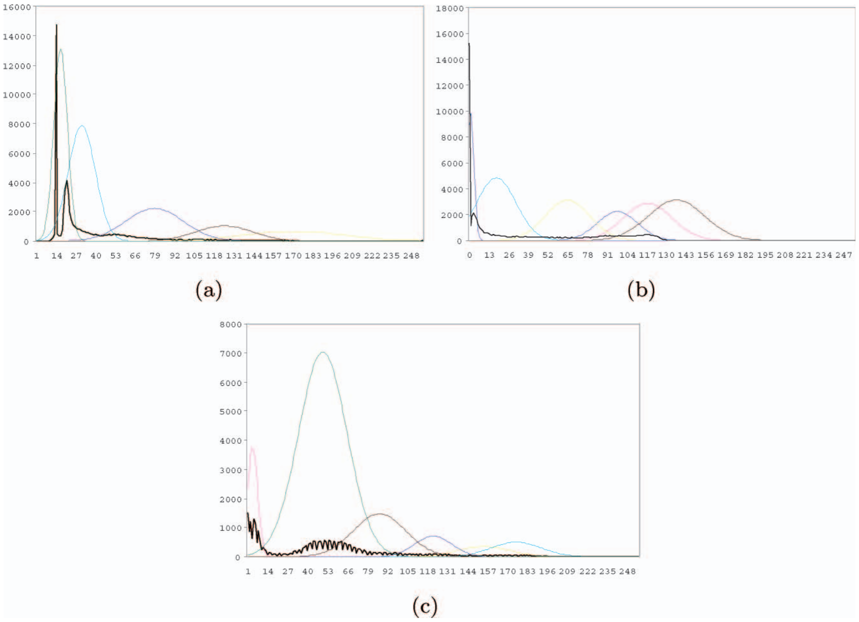


Figure 5. Histograms of the data with the associated scaled GMM after the application of our automatically seeded EM segmentation (results from figures 2(d), 3(d) and 4(d)).

Table 2. Changes in cluster means in the whole body data. A: Manual μ; B: manual μ after EM; C: automatic μ; D: automatic μ after EM.

	A	B	C	D
$\mu(0)$	164.6	123.922	116	117.66
$\mu(1)$	131.18	120.03	96	97.8356
$\mu(2)$	2.3	2.03	13	2.07
$\mu(3)$	66.59	33.01	44	27.48
$\mu(4)$	90.1	94.49	73	70.836
$\mu(5)$	164.21	194.81	153	140.6223

Table 3. Changes in cluster means in the whole body data. A: Manual μ; B: manual μ after EM; C: automatic μ; D: automatic μ after EM.

	A	B	C	D
$\mu(0)$	170.92	169.4365	183	178.41
$\mu(1)$	42.29	44.45	52	50.484
$\mu(2)$	3.84	4.177	5	4.27
$\mu(3)$	123.61	118.868	151	153.720
$\mu(4)$	95.35	82.99	124	121.496
$\mu(5)$	57.2	55.897	92	85.687

of the image data is constructed, n_j, where n is the number of pixels contained in the bin with value j. This histogram is then divided into M evenly distributed bins. This value M is manually set, typically to a higher number than the number of perceived relevant regions in the image. For the images shown in this paper, the value of M was set experimentally to 25. From each bin, the highest peak in the histogram is assigned to a seed centre, C_m.

$$C_m = \arg\max_j (n_j) \tag{8}$$

These M seed centres are then clustered together using their closeness in the greyscale space and their heights n_j until the desired number of seeds, k, is reached. The clustering is an iterative process where clusters are joined together by evaluating the Euclidean distance between the cluster centres.

3. Results

The described scheme was applied to gated MRI short-axis images of the heart, MRI coronal brain slices and a section from a whole body MRI showing the lower abdomen. In order to illustrate the validity of the automatic seed selection algorithm, the results are compared against those

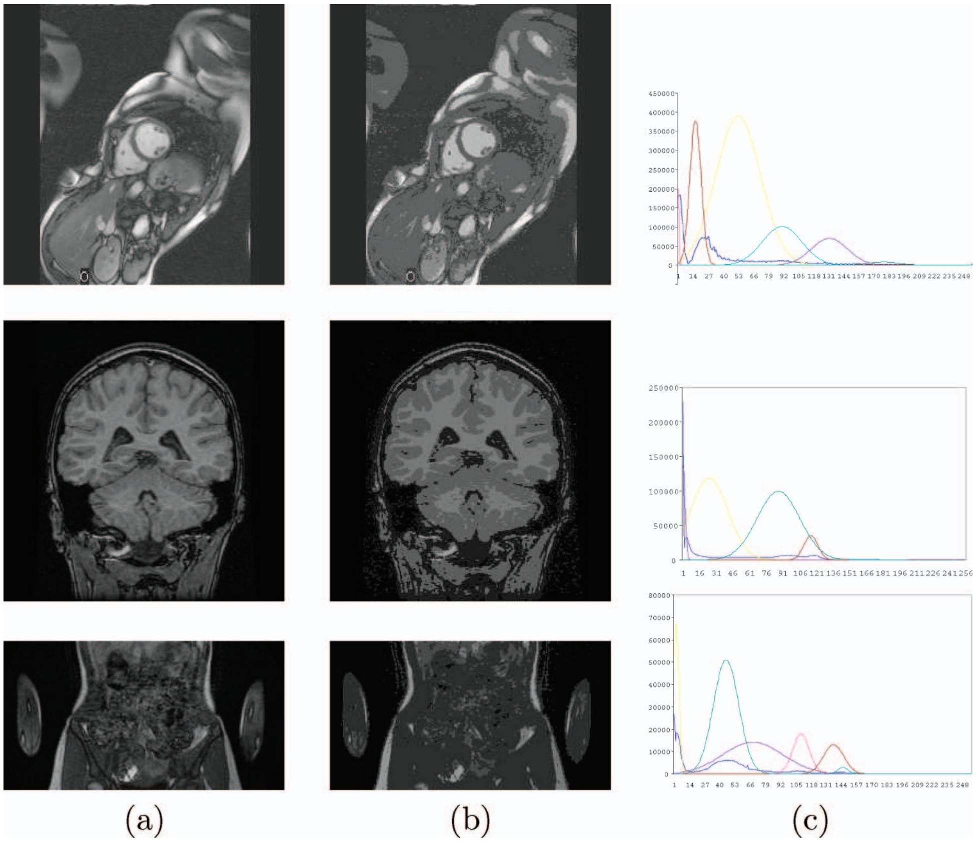


Figure 6. 3D space partitioning using EM: images show (a) a single slice of a 3D dataset from the original volume, (b) after segmentation with the EM algorithm and (c) shows the associated histogram of the data with scaled GMM included.

obtained when the cluster means and variances are manually extracted from the image. An example of this is shown in figure 2(b), which shows the areas in the image that were selected manually for use as the initialization of the EM algorithm. A visual comparison of the segmentation after initialization using these manually selected regions against the results obtained after the automatic seed selection detailed in §2.1 can be seen in figure 2(c) and 2(d).

From figure 2 and table 1, it is clear that using the automatic seed initialization gives a better distribution of initial seeds across the data. Table 1 presents the manually selected means of the Gaussian distributions and automatically selected means using the method described above. Also, the Gaussian means after the EM algorithm has been applied are presented.

To evaluate the performance of the described algorithm, the EM segmentation algorithm is applied to each of the

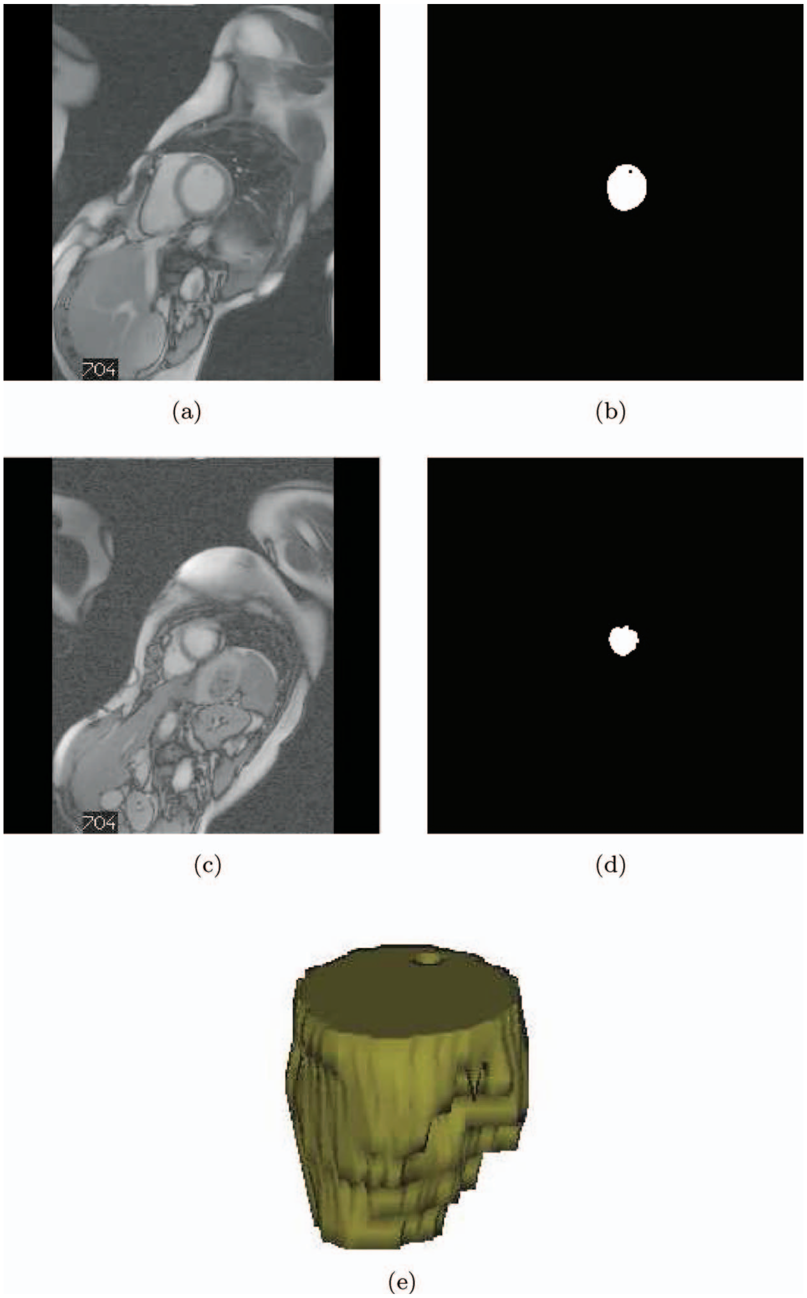


Figure 7. Images show slices 1((a) and (b)) and 4((c) and (d)) from the original volume (left) and with left ventricle blood cavity segmented (right) and (e) shows the rendered volume of the segmentation.

MRI datasets. As mentioned previously, the algorithm is locally convergent and therefore initialization of the algorithm is crucial to the final solution. A comparison is made between the results obtained using the automatically seeding process and the results obtained when the initial seeds for the EM segmentation are chosen manually. To achieve this, areas are selected in each of the images that attempt to represent the most significant regions. This is objective and related to the purpose of the segmentation but the overriding motivation is to pick regions that are clinically significant and also have a high degree of variation between regions. In each of the images given, six regions were manually selected. In these selected regions the mean pixel intensity values and the variance of the pixel intensity values are calculated. These manually selected values are used as the initial values of θ_m, where $1 \leq m \leq 6$ in the EM algorithm, and the mixing parameters α_m were each set to $\frac{1}{m}$.

Figure 2 illustrates the strategy applied to short axis images from a cardiac MRI study. The areas manually selected are shown in figure 2(b) and the resultant segmentation after applying the EM segmentation using these initial parameters is shown in figure 2(c). Figure 2(d) shows appropriate results after the automatic parameter selection; in particular the results show a better distribution within the greyscale distribution of the analysed image.

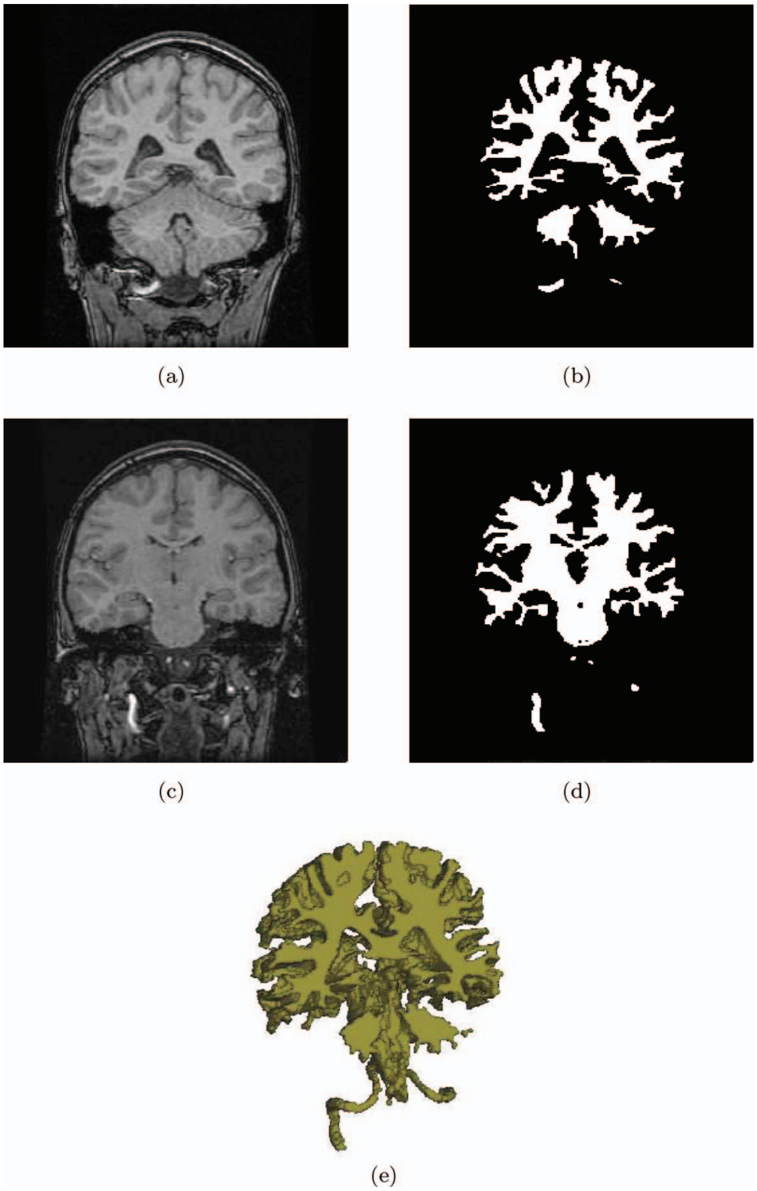


Figure 8. Images show slices 1((a) and (b)) and 14((c) and (d)) from the original volume (left) and with segmented white matter (right) and (e) shows the rendered volume of the segmentation.

Figure 3 shows a coronal slice from a T1-weighted head MRI. Again the automatic segmentation method performs well in differentiating the white matter from the grey matter. Figure 4 shows a coronal slice from an abdominal section of a full body MRI.

The second measure of performance is given in figure 5, where the intensity histograms for each of the images shown in figures 2, 3 and 4 are plotted. Overlaid on these histograms are the resulting GMMs resulting from the EM segmentation using the automatic seeds. The Gaussian distributions are scaled for illustration purposes.

It is clear from tables 1, 2 and 3 that the described automatic seed picking algorithm demonstrates better performance when compared to the manual selection technique. This is evident from the lower differences between initialized seeds and the final values after optimization through the EM algorithm.

Most medical images obtained from MRI are 3D and in some cases 4D, but because the algorithm works on the data histogram (hence, intensity values) and is not dependent on spatial position, it can be applied equally successfully to any dimensioned data. This is illustrated in figure 6, where the algorithm is successfully applied in 3D MRI images. This aspect is examined further in §4, where the results are used in conjunction with a diffusion based filtering [16,17] to extract some clinically relevant regions from the images.

It is worth noting that statistical classification of pixels is a more appropriate way to segment medical images, as the standard region growing technique will fail to produce appropriate results in images that exhibit a low signal-to-noise ratio (SNR). Also, such medical images generally show good separation between significant regions. This is application-dependent so we will now look at some common medical applications.

4. Applications in medical imaging

One of the key indicators of cardiac health is left ventricle *ejection fraction*, a measure of the volume of blood pumped from the left ventricle with each heartbeat [18]. Cardiac cine MRI is a standard procedure where 3D volume images are acquired at gated temporal positions through the cardiac pumping cycle. Such images are frequently taken using gradient echo imaging, which exhibits a relatively high

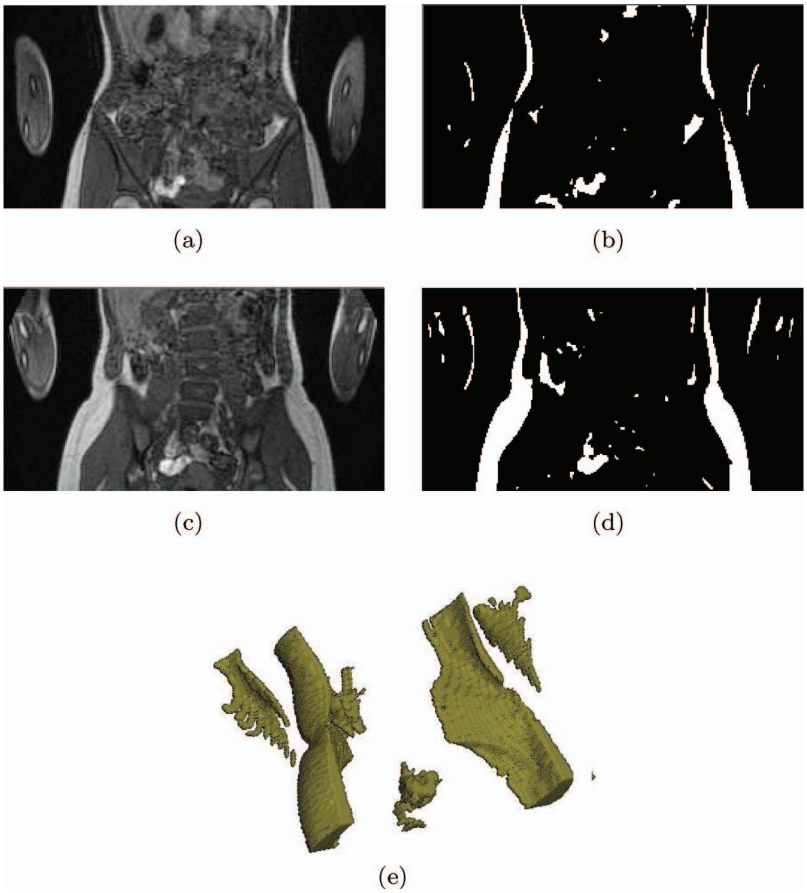


Figure 9. Images show slices 2((a) and (b)) and 6((c) and (d)) from the original volume (left) and with body fat segmented (right) and (e) shows the rendered volume of the segmentation.

differentiation between the blood and the myocardium. Figure 7 shows the end-diastole segmented left ventricle blood-pool after the application of the EM algorithm described in this paper to identify the left ventricle cavity. Figure 7(e) is a rendered volume of the blood pool, inside the cavity of the left ventricle when the muscle is at its end-diastole phase.

The classification of brain MRI white matter, grey matter, cerebrospinal fluid and in some cases lesions is a fundamental first step for surgical planning, radiotherapy planning and the identification of brain disease [19]. Illustrated in figure 8 is a segmentation of white matter of the brain.

The accurate measurement of body fat from whole-body MRI images is becoming an increasingly important metric, as high body fat level is recognized to play a significant role in a variety of serious health problems [20]. MRI is the modality of choice due to its repeatability and high spatial resolution. Figure 9 illustrates the results from one section of a whole-body MRI dataset where the fat tissue has being segmented out of the volume.

5. Conclusion

In this paper the implementation of an automatic seed picking algorithm to be used as the initialization of an expectation-maximization segmentation scheme is detailed. This segmentation technique is then applied to a variety of MRI datasets both in 2D and 3D. Statistical based classification of pixels is especially appropriate to MRI data, as traditional region growing and edge-based segmentation algorithms fail to produce accurate segmentation results when applied to medical datasets characterized by a low SNR. The EM algorithm shows robust and repeatable performance in the segmentations of heart, brain and abdominal images. The EM algorithm is locally convergent [6,14,15] so we have introduced an automatic seeding method that uses local maxima in the intensity histogram. The results are compared against a manual initialization, achieved by first manually selecting a region and then measuring the mean intensity values and variance in that region. The results of the manual initialization and the automatic initialization are shown after the application of the expectation-maximization algorithm. The methods shows appropriate results with respect to the greyscale values. From these results we can conclude that this approach offers robust, reproducible and accurate estimation of the initial parameters for the EM algorithm and the segmentation scheme described is capable of providing useful clinical measurements when applied to a large range of medical datasets.

References

[1] Clarke, L.P., Velthuizen, R.P., Camacho, M.A., Heine, J.J., Vaidyanathan, M., Hall, L.O., Thatcher, R.W. and Silbiger, M.L., 1995, MRI segmentation: methods and applications. *Magnetic Resonance Imaging*, **13,** 343–368.

[2] Pham, D.L., Xu, C. and Prince, J.L., 1998, A survey of current methods in medical image segmentation. Technical report. The John Hopkins University, Baltimore, MD 21218.

[3] Höhne, K.H. and Hanson, W.A., 1992, Interactive 3D-segmentation of MRI and CT volumes using morphological operations. *Journal of Computer Assisted Tomography*, **16,** 285–294.

[4] Soille, P., 2003, *Morphological Image Analysis*, 2nd ed (New York: Springer-Verlag).

[5] Hartigan, J.A. and Wong, M.A., 1979, Statistical algorithms: Algorithm AS 136: A *K*-means clustering algorithm. *Journal of Applied Statistics*, **28,** 100–108.

[6] Dempster, A., Laird, N. and Rubin, D., 1977, Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, **39,** 1–38.

[7] Jain, A.K., Murty, M.N. and Flynn, P.J., 1999, Data clustering: A review. *ACM Computing Surveys*, **31,** 264–323.

[8] Duda, R. and Hart, P., 1973, *Pattern Classification and Scene Analysis* (New York: Wiley).

[9] Jain, A.K. and Dubes, R.C., 1998, *Algorithms for Clustering Data* (New Jersey: Prentice-Hall).

[10] Al-Daoud, M.B. and Roberts, S.A., 1996, New methods for the initialization of clusters. *Pattern Recognition Letters*, **17,** 451–455.

[11] Mitra, P., Pal, S.K. and Siddiqi, M.A., 2003, Non-convex clustering using expectation maximization algorithm with rough set initialization. *Pattern Recognition Letters*, **24,** 863–873.

[12] Khan, S.S. and Ahmad, A., 2004, Cluster center initialization algorithm for *k*-means clustering. *Pattern Recognition Letters*, **25,** 1293–1302.

[13] Bishop, C.M., 1995, *Neural Networks for Pattern Recognition* (Oxford: Oxford University Press).

[14] Bilmes, J.A., 1998, A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden markov models. Technical Report TR-97-021, Berkeley, CA.

[15] Xu, L. and Jordan, M.I., 1996, On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, **8,** 129–151.

[16] Ghita, O., Robinson, K., Lynch, M. and Whelan, P.F., 2005, MRI diffusion-based filtering: A note on performance characterization. *Computerized Medical Imaging and Graphics*, **29,** 267–277.

[17] Perona, P. and Malik, J., 1990, Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12,** 629–639.

[18] Frangi, F., Rueckert, D. and Duncan, J.S., 2001, Three-dimensional modelling for functional analysis of cardiac images: a review. *IEEE Transactions on Medical Imaging*, **20,** 2–5.

[19] Zavaljevskia, A., Dhawan, A.P., Gaskil, W., Balld, M. and Johnsonb, J.D., 2000, Multi-level adaptive segmentation of multi-parameter MR brain images. *Computerized Medical Imaging and Graphics*, **24,** 87–98.

[20] Brennan, D., Whelan, P.F., Robinson, K., Ghita, O., Sadleir, R., O'Brien, J. and Eustace, S., 2005, Rapid automated measurement of body fat tissue distribution from whole body MRI. *American Journal of Roentgenology*, **185,** 418–423.