EM BASED APPROXIMATION OF EMPIRICAL DISTRIBUTIONS WITH LINEAR COMBINATIONS OF DISCRETE GAUSSIANS

Ayman El-Baz

Bioengineering Department
University of Louisville
Louisville, KY, USA

Georgy Gimel'farb

CS Dept., Tamaki Campus
University of Auckland
Auckland, New Zealand

ABSTRACT

We propose novel Expectation Maximization (EM) based algorithms for accurate approximation of an empirical probability distribution of discrete scalar data. The algorithms refine our previous ones in that they approximate the empirical distribution with a linear combination of discrete Gaussians (LCDG). The use of the DGs results in closer approximation and considerably better convergence to a local likelihood maximum compared to previously involved conventional continuous Gaussian densities. Experiments in segmenting multi-modal medical images show the proposed algorithms produce more adequate region borders.

Index Terms— Linear combination of discrete Gaussians, modified expectation maximization algorithm.

1. INTRODUCTION

Approximation of empirical probability distributions of scalar measurements with mixtures of probability models is in wide use in advanced data analysis [2, 6]. Statistical decision making based on such models is now a common practice in astronomy, physics, remote sensing, medical imaging and many other application areas where data sets can be extremely large. In many cases, e.g. in multi-modal images, each prominent peak, or mode of the mixed marginal distribution of signals relates to a particular object-of-interest (or class of signals). Then the approximation pursues the goal of separating individual classes from their mixture in order to use these models to classify the signals, e.g. segment the objects. The basic problem essential for precise data classification is to accurately model not only each peak itself but also the behavior of signals of each class between the peaks. This is because borders of each object relate often to intersections of tails of the individual class distributions. Of course, generally no accurate classification can be achieved by using only a mixed marginal probability distribution by itself. Nonetheless, such rough data classification or clustering techniques are of practical interest in many important application problems, e.g., for automated screening of multi-modal medical images obtained by computer tomography or magnetic resonance imaging.

This paper introduces refined versions of our previous Expectation Maximization (EM) based algorithms [3]. The versions accurately approximate an empirical marginal probability distribution of discrete scalar data with a linear combination of discrete Gaussians (LCDG), the latter notion being defined in Section 2 below. The linear combination involves both positive and negative Gasussians so that it approximates empirical data more accurately than a conventional mixture of only positive components [4, 7]. The main advantage of LCDG model is that it fits better the discrete empirical distribution than more conventional continuous Gaussian densities and their linear combinations in [3].

Historically, the first EM algorithm for estimating parameters of probability mixtures appeared in the late nineteen sixties [8] (see also [9]). But this technique received its current name and became very popular only a decade later after it was successfully applied to a general problem of parameter estimation from an incomplete data in [1], and many EM-algorithms exist today to find the maximum likelihood parameter estimates for mixtures of probability distributions [5].

2. LCDG MODEL

Let $\mathbf{F} = [f(q) : q \in \mathbf{Q} = \{0, 1, \ldots, Q-1\}; \sum_{q=0}^{Q-1} f(q) = 1]$ be an empirical probability distribution of discrete Q-ary signals q. We define a discrete Gaussian (DG) with the mean μ and variance σ^2 as the distribution $\mathbf{\Psi}_\theta = [\psi(q|\theta) : q \in \mathbf{Q}; \sum_{q=0}^{Q-1} \psi(q|\theta) = 1]$ such that $\psi(0|\theta) = \Phi_\theta(0.5)$, $\psi(q|\theta) = \Phi_\theta(q+0.5)) - \Phi_\theta(q-0.5))$ for $q = 1, \ldots, Q-2$, and $\psi(Q-1|\theta) = 1 - \Phi_\theta(Q-1.5)$. Here, $\Phi_\theta(\ldots)$ is the cumulative Gaussian probability function with a shorthand notation $\theta = (\mu, \sigma^2)$ for its mean and variance. The LCDG model $\mathbf{P}$ of the distribution $\mathbf{F}$ has C_p positive and C_n negative DGs: $p(q) = \sum_{r=1}^{C_\mathrm{p}} w_{\mathrm{p},r}\psi(q|\theta_{\mathrm{p},r}) - \sum_{l=1}^{C_\mathrm{n}} w_{\mathrm{n},l}\psi(q|\theta_{\mathrm{n},l})$ with the restricted positive weights $\mathbf{w} = [w_{\mathrm{p},.}, w_{\mathrm{n},.}]$:

$$\sum_{r=1}^{C_\mathrm{p}} w_{\mathrm{p},r} - \sum_{l=1}^{C_\mathrm{n}} w_{\mathrm{n},l} = 1 \tag{1}$$

Under a fixed number $C = C_\mathrm{p} + C_\mathrm{n}$ of the DGs, the model parameters are the weights $\mathbf{w} = \{w_c; c = 1, \ldots, C\}$ and characteristics of the individual DGs $\mathbf{\Theta} = \{\theta_c : c = 1, \ldots, C\}$. Probability distributions form a proper subset of all the LCDGs under the additional restriction $p(q) \geq 0$ that automatically holds for mixtures with no negative DGs. Just as in [3], we ignore this restriction because our goal is only to closely approximate the empirical distribution $\mathbf{F}$. We also assume that the numbers C_p and C_n of the components of each type are known after an initialization and do not change during the EM-based refinement of the model parameters. The initialization provides also the starting parameters $\mathbf{w}^{[0]}$ and $\mathbf{\Theta}^{[0]}$.

Assuming statistical independence of the mixed signals, the optimal model parameters are found by the EM-based maximization of the log-likelihood of the empirical data:

$$L(\mathbf{w}, \mathbf{\Theta}) = \sum_{q \in \mathbf{Q}} f(q) \log p(q) \tag{2}$$

To estimate the parameters of the LCDG model, we modified the conventional EM algorithm for estimating parameters of normal mixtures [9] to account for the DGs with alternating signs as shown in Section 3. Because this modification is sensitive to its starting

1-4244-1437-7/07/$20.00 ©2007 IEEE

state, a close initial LCDG-approximation of the empirical distribution is built by a sequential EM-based algorithm presented in Section 4.

3. EM BASED REFINEMENT OF THE LCDG

A local maximum of the log-likelihood in Eq. (2) is given with the EM process [3,9]. Let $p^{[m]}(q) = \sum_{r=1}^{C_\mathrm{p}} w_{\mathrm{p},r}^{[m]} \psi(q|\theta_{\mathrm{p},r}^{[m]}) - \sum_{l=1}^{C_\mathrm{n}} w_{\mathrm{n},l}^{[m]} \psi(q|\theta_{\mathrm{n},l}^{[m]})$ denote the current LCDG at iteration m. Relative contributions of each signal $q \in \mathbf{Q}$ to each positive and negative DG at iteration m are specified by the respective conditional weights

$$\pi_\mathrm{p}^{[m]}(r|q) = \frac{w_{\mathrm{p},r}^{[m]} \psi(q|\theta_{\mathrm{p},r}^{[m]})}{p_{\mathbf{w},\mathbf{\Theta}}^{[m]}(q)}; \;\pi_\mathrm{n}^{[m]}(l|q) = \frac{w_{\mathrm{n},l}^{[m]} \psi(q|\theta_{\mathrm{n},l}^{[m]})}{p_{\mathbf{w},\mathbf{\Theta}}^{[m]}(q)} \qquad (3)$$

such that the following constraints hold:

$$\sum_{r=1}^{C_\mathrm{p}} \pi_\mathrm{p}^{[m]}(r|q) - \sum_{l=1}^{C_\mathrm{n}} \pi_\mathrm{n}^{[m]}(l|q) = 1; \; q = 0, \ldots, Q-1 \qquad (4)$$

The EM process iterates the following two steps until the changes of the log-likelihood become small:

E– step $[m+1]$: Find conditional expectations of the parameters $\mathbf{w}^{[m+1]}$, $\mathbf{\Theta}^{[m+1]}$ using the fixed weights of Eq. (3) for the step m as conditional probabilities, and

M– step $[m+1]$: Find the latter weights by maximizing $L(\mathbf{w}, \mathbf{\Theta})$ under the fixed parameters $\mathbf{w}^{[m+1]}$, $\mathbf{\Theta}^{[m+1]}$.

This block relaxation process is converging to a local maximum of the likelihood in Eq. (5). It is easily shown by using the constraints of Eq. (4) as unit factors and rewriting the log-likelihood of Eq. (2) in the equivalent form:

$$\begin{aligned} L(\mathbf{w}^{[m]}, \mathbf{\Theta}^{[m]}) = &\sum_{q=0}^{Q} f(q) \left[\sum_{r=1}^{C_\mathrm{p}} \pi_\mathrm{p}^{[m]}(r|q) \log p^{[m]}(q)\right] \\ &- \sum_{q=0}^{Q} f(q) \left[\sum_{l=1}^{C_\mathrm{n}} \pi_\mathrm{n}^{[m]}(l|q) \log p^{[m]}(q)\right] \qquad (5) \end{aligned}$$

Let us replace the term $\log p^{[m]}(q)$ in the first and the second brackets, respectively, with the equal terms which follow from Eq. (3): $\log w_{\mathrm{p},r}^{[m]} + \log \psi(q|\theta_{\mathrm{p},r}^{[m]}) - \log \pi_\mathrm{p}^{[m]}(r|q)$ and $\log w_{\mathrm{n},l}^{[m]} + \log \psi(q|\theta_{\mathrm{n},l}^{[m]}) - \log \pi_\mathrm{n}^{[m]}(l|q)$. At the E-step, the expected weights

$$w_{\mathrm{p},r}^{[m+1]} = \sum_{q \in \mathbf{Q}} f(q) \pi_\mathrm{p}^{[m]}(r|q); \; w_{\mathrm{n},l}^{[m+1]} = \sum_{q \in \mathbf{Q}} f(q) \pi_\mathrm{n}^{[m]}(l|q)$$

follow also from the conditional Lagrange maximization of the log-likelihood in Eq. (5) under the restriction of Eq. (1). The expected parameters of each DG are also the conventional unconditional MLEs that stem from the maximization of the log-likelihood after each difference of the cumulative Gaussians is replaced with its close approximation by the Gaussian density (below "c" stands for "p" or "n", respectively):

$$\begin{aligned} \mu_{\mathrm{c},r}^{[m+1]} &= \tfrac{1}{w_{\mathrm{c},r}^{[m+1]}} \textstyle\sum_{q \in \mathbf{Q}} q \cdot f(q) \pi_\mathrm{c}^{[m]}(r|q) \\ (\sigma_{\mathrm{c},r}^{[m+1]})^2 &= \tfrac{1}{w_{\mathrm{c},r}^{[m+1]}} \textstyle\sum_{q \in \mathbf{Q}} \left(q - \mu_{\mathrm{c},i}^{[m+1]}\right)^2 \cdot f(q) \pi_\mathrm{c}^{[m]}(r|q) \end{aligned}$$

The M-step performs the conditional Lagrange maximization of the log-likelihood of Eq. (5) under the Q restrictions of Eq. (4) and results in the same weights $\pi_\mathrm{p}^{[m+1]}(r|q)$ and $\pi_\mathrm{n}^{[m+1]}(l|q)$ as in Eq. (3) for all $r = 1, \ldots, C_\mathrm{p}$; $l = 1, \ldots, C_\mathrm{n}$ and $q \in \mathbf{Q}$. This modified EM-algorithm is valid until the weights are strictly positive but the iterations should be terminated when the log-likelihood of Eq. (5) begins to decrease.

4. SEQUENTIAL EM-BASED INITIALIZATION

We assume that the number of dominant modes K equal to the number of classes (objects) is known. To simplify the notation, let us consider the bi-modal case when the empirical distribution have only two separate dominant modes representing a desired object and its background, respectively. The algorithm below is easily extended to the general case of $K > 2$ dominant modes. Initially, each dominant mode is roughly approximated with a single DG, and deviations of the empirical distribution from the dominant two-component mixture are described with other components of the LCDG. Therefore, the model has the two dominant positive weights, say, $w_{\mathrm{p},1}$ and $w_{\mathrm{p},2}$ such that $w_{\mathrm{p},1} + w_{\mathrm{p},2} = 1$, and a number of "subordinate" weights of smaller absolute values such that $\sum_{r=3}^{C_\mathrm{p}} w_{\mathrm{p},r} - \sum_{l=1}^{C_\mathrm{n}} w_{\mathrm{n},l} = 0$.

The following sequential algorithm accurately estimates both the number of the non-dominant DGs and all the weights and parameters of the LCDG components:

1. Approximate the empirical mixed distribution $\mathbf{F}$ with the dominant mixture $\mathbf{P}_2$ of two DGs using the EM algorithm from Section 3 with only the positive weights (it closely resembles the conventional one in [9]).
2. Find all the deviations $\delta(q) = f(q) - p_2(q)$; $q \in \mathbf{Q}$, between $\mathbf{F}$ and $\mathbf{P}_2$ and split them into the positive $\mathbf{\Delta}_\mathrm{p} = [\delta_\mathrm{p}(q) = \max\{\delta(q), 0\} : q \in \mathbf{Q})$ and the negative $\mathbf{\Delta}_\mathrm{n} = [\delta_\mathrm{n}(q) = \max\{-\delta(q), 0\} : q \in \mathbf{Q}]$ parts such that $\delta(q) = \delta_\mathrm{p}(q) - \delta_\mathrm{n}(q)$.
3. Compute the factor $s = \sum_{q=0}^{Q-1} \delta_\mathrm{p}(q) \equiv \sum_{q=0}^{Q-1} \delta_\mathrm{n}(q)$ to scale the deviations up.
4. If the factor s is less than an accuracy threshold, terminate the algorithm and return the dominant model $\mathbf{P}_C = \mathbf{P}_2$.
5. Otherwise consider the scaled-up absolute deviations $\frac{1}{s}\mathbf{\Delta}_\mathrm{p}$ and $\frac{1}{s}\mathbf{\Delta}_\mathrm{n}$ as two new "empirical distributions" and use iteratively the EM algorithm from Section 3 with only the positive weights to find sizes (C_p, C_n) and parameters of mixtures of only positive DGs, $\mathbf{P}_\mathrm{p}$ and $\mathbf{P}_\mathrm{n}$, respectively, that approximate best the scaled-up deviations. Each size is found by sequential minimization of the total absolute error between the scaled-up deviation, $\mathbf{\Delta}_\mathrm{p}$ (or $\mathbf{\Delta}_\mathrm{n}$), and its mixture model, $\mathbf{P}_\mathrm{p}$ (or $\mathbf{P}_\mathrm{n}$) with respect to the number of the components.
6. Scale down the subordinate models $\mathbf{P}_\mathrm{p}$ and $\mathbf{P}_\mathrm{n}$ (i.e. scale down the weights of their components) and then add the scaled model $\mathbf{P}_\mathrm{p}$ and subtract the scaled model $\mathbf{P}_\mathrm{n}$ from the dominant mixture $\mathbf{P}_2$ in order to form the desired LCDG model $\mathbf{P}_C$ of the size $C = 2 + C_\mathrm{p} + C_\mathrm{n}$.

The final mixed LCDG-model P_C has to be split into the K LCDG-submodels $P_{[k]} = [p(q|k) : q \in \mathbf{Q}]$, one per class $k = 1, \ldots, K$. This is done by associating each subordinate DG with a particular dominant term as to minimize the expected misclassification rate. Let us illustrate the association principle by the bi-modal case where the two dominant DGs have the mean values μ_1 and μ_2 such that $0 < \mu_1 < \mu_2 < Q - 1$. If all the subordinate DGs are ordered by their mean values, then those with the mean values smaller than μ_1 and greater than μ_2 relate to the first and second class, respectively. The DGs with the mean values in the range $[\mu_1, \mu_2]$ are associated with the classes by simple thresholding, the components with the means below the threshold, t, belonging to the the first class. The chosen threshold minimizes the misclassification rate $e(t)$:

$$e(t) = \sum_{q=0}^{t-1} p(q|2) + \sum_{t}^{Q-1} p(q|1) \qquad (6)$$

Figure 1 shows the initial approximation of the bi-modal empirical distribution of $Q = 256$ grey levels over a typical DC-MRI (Dynamic Contrast-Enhanced Magnetic Resonance Imaging) slice of human abdomen. The dominant modes represent the brighter kidney area and its darker background, respectively. After the additive and subtractive parts of the absolute deviation are approximated with the DG mixtures, the initial mixed LCDG-model consists of the 2 dominant, 4 additive, and 4 subtractive DGs, that is, $C_{\mathrm{p}} = 6$ and $C_{\mathrm{n}} = 4$. The LCDG models of each class are obtained with $t = 78$ ensuring the best class separation.

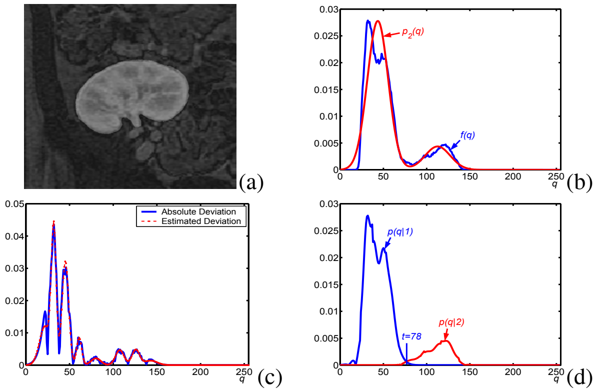


Fig. 1. Initial LCDG model of the bi-modal empirical grey level distribution: the DC-MRI slice (a), its empirical grey level distribution approximated with the dominant mixture of the DGs (b), the scaled-up absolute deviation of the approximation, (c) approximation error for the scaled absolute deviation as a function of the number of the subordinate Gaussians and its LCDG model (c), and the LCDG model of each class (d) for the best separating threshold $t = 78$.

5. EXPERIMENTS AND CONCLUSIONS

Figure 2 presents the final LCDG model obtained by refining the above initial one using the modified EM-algorithm introduced in Section 3. First 37 iterations of the algorithm increase the log-likelihood of Eq. (5) from -6.90 to -4.49, and the convergence to the log-likelihood maximum are considerably more stable than with our previous algorithm in [3] involving linear combinations of continuous Gaussian densities. The resulting segmentation has an error of 1.26% with respect to the expert's region map. Figure 3 shows more segmentation results obtained by the proposed algorithm.

Figure 4 shows one more example, namely, the LCDG approximation of a 4-modal empirical grey level distribution for a CTA (computed tomography angiography) image. The classes represent dark background and colon, liver and kidney, blood vessels, and bright bones, respectively, and the goal is to separate the blood vessels in spite of its large intersection with the second one and very low prior probability. The initialization returns the 13 components of the LCDG, and the 16 first iterations of the refinement before the process terminates increase the log-likelihood from -6.18 for the initial LCDG to -5.10 for the final one. The segmentation with the final LCDG-models of the classes has the error only about 0.59% with respect to the expert's map. Figure 5 shows more segmentation results obtained by the proposed algorithm.

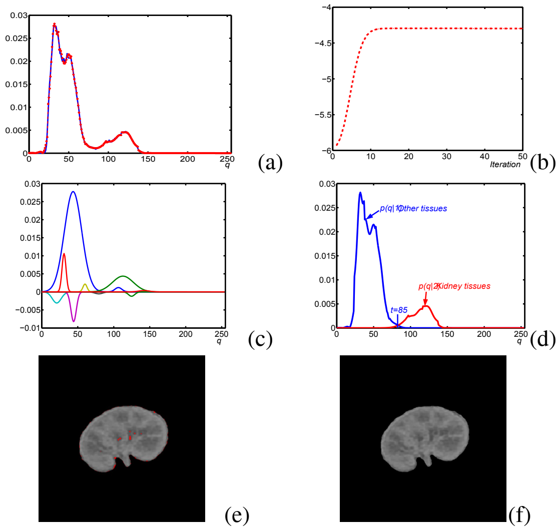
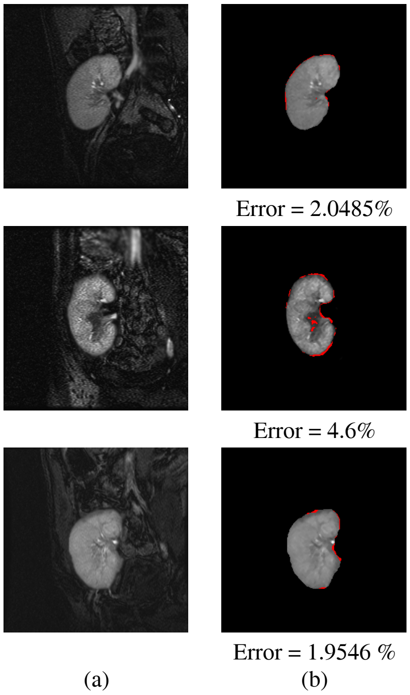


Fig. 2. Final 2-class LCDG model (a), log-likelihood changes at the EM-iterations (b), ten components of the final LCDG (c), the final LCDG model of each class for the best separating threshold $t = 85$ (d), the segmentation map (e) for Fig. 1(a), and the "ground truth" (f) produced by a radiologist. Errors w.r.t. the ground truth are highlighted by red color.


Error = 2.0485%

Error = 4.6%

Error = 1.9546 %

(a) (b)

Fig. 3. Segmentation of three other kidney DC-MRI images with our approach, (a) Original DC-MRI images and (b) Our segmentation results. Error w.r.t the ground truth are highlighted by red color.

These and other experiments with different multi-modal images show that the proposed EM-based techniques produce very accurate LCDG-models of empirical probability distributions of scalar signals, providing our initialization produces proper numbers of the

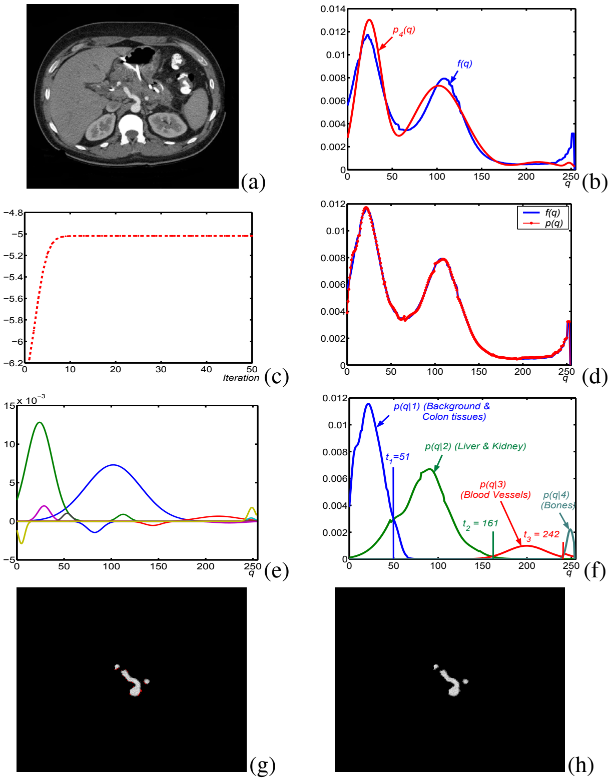


Fig. 4. Initial and final 4-class LCDG models: a CTA image (a), its 4-modal empirical grey level distribution approximated with the dominant 4-component DG mixture (b), log-likelihood changes at the EM iterations (c), the final mixed LCDG (d), its components (e), the class LCDGs (f), the blood vessel segmentation map (g), and the expert's map (h). Errors w.r.t. the ground truth are highlighted by red color.

additive and subtractive DGs. The computations are as simple as in the majority of conventional EM algorithms. The pixel-wise signal classification based on the final LCDG models of each class and combined with post-processing yields typically small segmentation errors with respect to the expert's maps (e.g. 0.005–4.89% for more than 532 CT, MRI, and CTA images). The post-processing involves simple Markov–Gibbs random field (MGRF) models of region maps with analytically estimated parameters.

Our previous probability models with linear combinations of continuous Gaussian densities [3] have had similar low segmentation errors, too. But the LCDG models ensure the EM process has much more stable convergence to the log-likelihood maximum and suffers fewer accumulated numerical errors. Conventional normal mixtures of the same size and under the same post-processing yield up to ten times larger errors because some inter-class intervals are covered by single Gaussians. Because each such component combines tails of the two class distributions, the accurate separation of the class models becomes hardly possible.

6. REFERENCES

[1] A.P.Dempster, N. M.Laird, and D. B.Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. Royal Statistical Society*, Vol. 39B, pp. 1–38, 1977.

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley: N.Y., 2001.

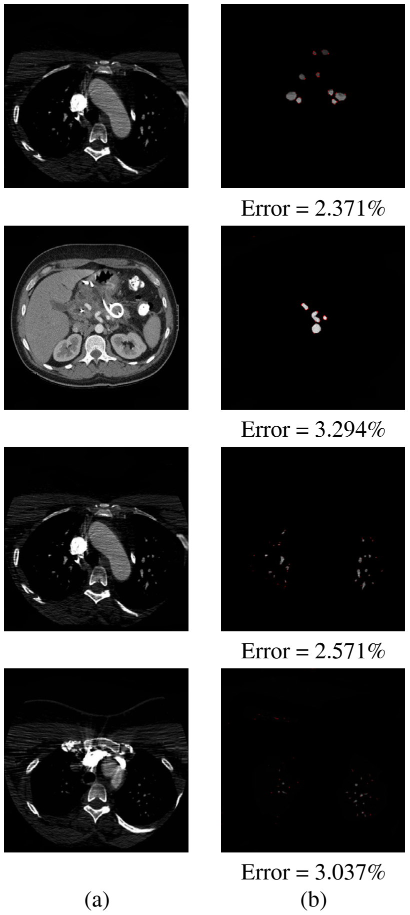


Fig. 5. Segmentation of four other CTA images with our approach, (a) Original CTA images and (b) Our segmentation results. Error w.r.t the ground truth are highlighted by red color.

[3] G.Gimel'farb, A.A.Farag, and A.El-Baz, "Expectation-Maximization for a linear combination of Gaussians", in *Proc. IAPR Int. Conf. Pattern Recognition, Cambridge, UK, 23–26 Aug. 2004*, IEEE CS Press: Los Alamitos, Vol. 3, 2004, pp. 422–425.

[4] A.Goshtasby and W.D.O'Neill, "Curve fitting by a sum of Gaussians", *CVGIP: Graphical Models and Image Processing*, Vol. 56, pp.281-288, 1999.

[5] C.J.McLachlan, *The EM Algorithm and Extensions*, Wiley: N.Y., 1997.

[6] N.R.Pal and S.K.Pal, "A review on image segmentation techiniques", *Pattern Recognition*, Vol. 26, pp.1277–1294, 1993.

[7] T.Poggio and F.Girosi, "Networks for approximation and learning", *Proc. IEEE*, Vol. 78, pp. 1481–1497, 1990.

[8] M.I.Schlesinger, "A connection between supervised and unsupervised learning in pattern recognition", *Kibernetika*, no. 2, pp. 81-88, 1968 [In Russian].

[9] M.I.Schlesinger and V.Hlavac, *Ten Lectures on Statistical and Structural Pattern Recognition*, Kluwer Academic: Dordrecht, 2002.