Cross Entropy: A New Solver for Markov Random Field Modeling and Applications to Medical Image Segmentation

Jue Wu and Albert C.S. Chung

Bioengineering Program and Computer Science Department,
Lo Kwee-Seong Medical Image Analysis Laboratory,
The Hong Kong University of Science and Technology, Hong Kong

Abstract. This paper introduces a novel solver, namely cross entropy (CE), into the MRF theory for medical image segmentation. The solver, which is based on the theory of rare event simulation, is general and stochastic. Unlike some popular optimization methods such as belief propagation and graph cuts, CE makes no assumption on the form of objective functions and thus can be applied to any type of MRF models. Furthermore, it achieves higher performance of finding more global optima because of its stochastic property. In addition, it is more efficient than other stochastic methods like simulated annealing. We tested the new solver in 4 series of segmentation experiments on synthetic and clinical, vascular and cerebral images. The experiments show that CE can give more accurate segmentation results.

1 Introduction

As the core concept of Bayesian image analysis, Markov random field (MRF) theory has aroused great interest in the field of medical image processing. Researchers have applied MRF to various applications such as image enhancement [1], object detection [2], data modeling [3], tissue classification (segmentation) [4], etc. The reasons why MRF modeling has so many successful applications in medical image analysis are that it can easily incorporate spatial interaction and convert a problem in image processing or computer vision into a mathematical optimization problem by means of maximum *a posteriori* (MAP). The former property allows us to consider contextual constraints in images and the latter makes many delicate problems computationally tractable.

The most important part of the MRF modeling is the problem formulation in which we specify the posterior probability to be maximized or the energy function to be minimized. However, it remains a major open problem in MRF theory to optimize the objective function. Due to the large number of pixels in usual images, the configuration space of MRF in image analysis is huge. This makes the brute force methods to search for the optima infeasible in practice. Actually, it is proved that obtaining the global optimum of an arbitrary objective function is NP-hard [5]. Therefore, it has been an active research topic to design a "good" solver for MRF models over the past two decades. The goodness of a

J. Duncan and G. Gerig (Eds.): MICCAI 2005, LNCS 3749, pp. 229–237, 2005.

solver lies in whether it can efficiently find a local optimum which is as optimal as possible (e.g., the lower energy, the better).

One of the earliest efforts to optimize MRF objective functions was made by Kirkpatrick *et al.* [6] who proposed the solver, simulated annealing (SA). SA can guarantee to converge to a global minimum as long as the temperature is decreasing slowly enough which makes SA too slow for practical use especially for clinical data. Another pioneering work was done by Besag [7], where the iterated conditional modes (ICM) was presented. This is a fast solver at the cost that it finds local optima in a neighborhood where only one site label is allowed to change. After those two methods, quite a few solvers were introduced [8], such as mean field approximation (MFa), relaxation labeling (RL), graduated non-convexity (GNC), etc. Recently two efficient and fairly accurate solvers, belief propagation (BP) [9] and graph cuts (GC) [5], were proposed. These two solvers are now often used for MRF models because they give good accuracy in an efficient way, which means they can find "global" optima within a rather large neighborhood while maintaining acceptable time complexity. However, since the perfect solver is not existing unless P = NP, there is still space to get more accurate results. Moreover, BP and GC are not applicable to all types of objective functions. They obtain their accuracy at the cost of function form restrictions. For example, standard BP is only proper for pairwise MRFs and generalized BP is either not for all functions [9]. So we cannot solve by BP such MRF models as FRAME [10] or multi-level logistic (MLL) with more than two sites in a clique [8]. The same situation occurs for graph cuts because GC will work only when the energy function is regular [11]. These may considerably limit the usage of the two popular solvers.

In this paper, we proposed a new simple stochastic solver for MRF modeling, called cross entropy (CE). This idea is originated from the field of operations research to simulate rare events [12]. This paper combines the idea of CE with MRF theory for the first time and applies the whole model to medical image segmentation. The CE solver is a general and stochastic optimization method that can be applied to any kind of MRF formulation. Unlike BP and GC, CE makes no assumption on the form of the objective function so it is able to solve more complicated MRF models. The efficient CE solver is completely insensitive to initialization and more importantly, as a stochastic method, CE tends to find more global optimum than deterministic solvers like BP. This statement is supported by the experiments on four sets of synthetic and clinical data, which shows CE has higher segmentation accuracy than BP. It is believed that the accuracy of BP is comparable to that of GC [13]. Although CE is stochastic, it is efficient but expends more time than deterministic solvers like BP and GC.

2 The Cross Entropy Method

The cross entropy (CE), also known as Kullback-Leibler cross entropy, has been used in combinatorial and multi-extremal optimization, and rare event simulation. Owing to its simplicity and accuracy, it has quite a few successful applications in operations research and machine learning [14]. In this section, we will present the CE method as a novel simple accurate solver for MRF modeling.

Consider the following general energy minimization problem of MRFs. Let $\mathcal{F}$ be the configuration space of MRF, F, and f is one configuration of F. The energy minimization of MRFs is formulated by

$$f^* = \arg\min_{f\in\mathcal{F}} E(f), \tag{1}$$

where $E(\cdot)$ is the energy function to be minimized and f^* is the wanted F configuration. CE method associates an estimation problem with the optimization problem (Eq. 1). We first define an indicator function $I_{\{\text{event}\}}$, which is equal to 1 when the event is true otherwise 0. Then, suppose $p(\cdot;v)$ is a family of discrete probability density functions (pdf) on $\mathcal{F}$ and v is its parameter. Let us estimate the following probability

$$\mathcal{P}_v(E(F) \le e) = \sum_x I_{\{E(f)\le e\}} p(f;v), \tag{2}$$

where $\mathcal{P}_v$ is the probability measure and F is a vector of configurations that has pdf $p(\cdot;v)$. If $e = \min_{f\in\mathcal{F}} E(f)$ and $p(\cdot;v)$ is a uniform density on $\mathcal{F}$, Eq. 1 and Eq. 2 are connected. Note that $\mathcal{P}_v(E(F) \le e)$ is typically $1/|\mathcal{F}|$, which is very small. This is similar to the situation of rare event simulation. Thus, we can borrow the idea of CE from rare event simulation to construct a multi-level optimization approach for MRF energy, where we generate a sequence of levels e_1,e_2,...,e_T and parameter vectors v_1,v_2,...,v_T such that e_T is close to the optimal e^* and v_T is the density that assigns high probability mass to the configuration which corresponds to a low energy.

Suppose m is the size of the label space of the MRF model and there are n sites altogether. The CE solver for MRF labeling can be described as follows.

CE Algorithm for MRF energy minimization

1. Set level $t = 1$ and the initial parameter vector $\boldsymbol{v}_0 = \{v_{0,1}, \ldots, v_{0,n}\}$. Each $v_{t,i} = \{v^1_{t,i}, \ldots, v^m_{t,i}\}$ is a vector with m elements for site i.

2. Generate a collection of samples $F_1, \ldots, F_N$ ($F = \{f_1, \ldots, f_n\}$ is one MRF configuration) from the density $p(\cdot;v)$ and compute the energy $E_i(F_i)$ for every $i \in \{1, \ldots, N\}$.

3. Sort all the $E_i(F_i)$ in a non-increasing order to $\{E_1, \ldots, E_N\}$. Then pick $e_t = E_{\lceil(1-\rho)N\rceil}$.

4. Use the samples $F_1, \ldots, F_N$ to update $\boldsymbol{v}_t$ by

$$v^j_{t,i} = \frac{\sum_{k=1}^N I_{\{E_k(F_k)\le e_t\}} I_{\{F_{ki}=j\}}}{\sum_{k=1}^N I_{\{E_k(F_k)\le e_t\}}}, \tag{3}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

5. If e_t remains unchanged for several iterations, go to step 6; else, set $t = t+1$ and go to step 2.

6. The final $E_N(F_N)$ of T-th iteration is the estimated minimal MRF energy. The corresponding configuration is embodied by the parameter vector $\boldsymbol{v}_T$, where

each element $v^j_{T,i}$ assigns most probability mass to a preferable label among m labels for site i. □

Function $p(\cdot; v)$ can be any kind of pdf but the simple m-point Bernoulli distribution is usually enough. This means each label j is randomly chosen for site i according to the probability of $v^j_{t,i}$. Two parameters need to be pre-defined, ρ and N. Usually, ρ is a small value between 1% and 10%. When the site number n is large, we tend to choose a large value of ρ. Regarding the sample size N, we set $N = cn$, where c is a constant and often between 1 and 10. Notice that there are other alternative stopping criteria, such as when the parameter $\boldsymbol{v}_t$ converges to a binary (0 or 1) vector.

3 Experimental Results

In this section, we use CE and BP as MRF solvers to solve the same model for medical image segmentation and compare the results. We perform experiments on four sets of data of synthetic and clinical, vascular and cerebral images. Before that, we need to formulate the MRF energy function first.

3.1 MRF Formulation

We adopt the MAP-MRF framework for maximizing the posterior probability $P(X|Y) \propto P(X)P(Y|X)$, where X and Y are the labeling MRF and the observed data, respectively. We use multi-level logistics (MLL) and Gaussian distribution as the MRF prior and likelihood energy, which is one of the most often used models for medical image segmentation [8, 15, 16]. The MRF prior energy is expressed by

$$U(X) = \sum_{i \in S} \sum_{j \in N_i} I_{\{x_i \neq x_j\}}, \tag{4}$$

where N_i is the neighborhood of site i (we use the 4-neighborhood system) and S is the set of all sites. The likelihood energy is defined as

$$U(Y|X) = \sum_{i \in S} \sum_{j=0}^{m-1} \delta(x_i - j) \cdot \frac{(y_i - \mu_j)^2}{2\sigma^2}, \tag{5}$$

where if $x = 0$, $\delta(x) = 1$, else $\delta(x) = 0$, and μ_j represents the mean intensity of region j. In the comparisons below, all the parameters, e.g., μ, σ, are set the same for both solvers.

3.2 Synthetic Data

We first test the model on two sets of synthetic data, one simulates blood vessels following the style in [4] and the other is obtained from BrainWeb, a widely-used simulated brain database [17].

It is a binary segmentation problem to extract blood vessels. We synthesize alternate bars and circles to simulate the vessels and add various levels of Gaussian noise to produce corrupted images with different signal-to-noise ratios (SNR) (Fig. 1). The results of minimal energy values found by two solvers and segmentation errors are listed in Tab. 1. Regarding brain images, we segment the BrainWeb T1-weighted data into four classes {white matter, gray matter, cerespinal fluid, others} and perform several experiments with different levels of noise. Results are shown in Tab. 2. In both experiments, CE algorithm is repeated for 5 times to give means and standard deviations because it is stochastic. We do not repeat BP because it is deterministic. It can be found that the CE solver can reach lower energy than BP and the segmentation accuracy is also higher than BP for both synthetic images thanks to the ability of CE to find more global minima.

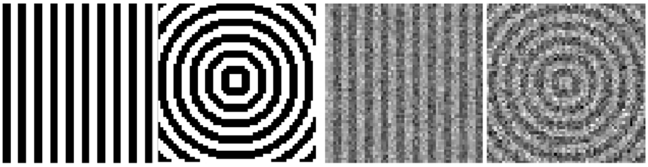

Fig. 1. Synthetic images for binary segmentation. The left two images are truth patterns and the right two images are corrupted images.

Table 1. Results of the MRF model for binary segmentation on synthetic images

		BP		CE	
image	SNR	minimum	error	minimum	error
bar (width=3)	2	2215505	18.65%	1656942±7707	9.01%±1.53%
	3	1848604	7.54%	1577403±17069	3.52%±0.69%
	4	1598662	2.25%	1523299±17271	1.44%±0.96%
	5	1526652	0.59%	1501826±2126	0.22%±0.28%
bar (width=6)	2	2101877	19.29%	1319552±4612	1.20%±0.37%
	3	1659417	8.30%	1242550±1923	0.59%±0.20%
	4	1408863	3.13%	1232195±2309	0.03%±0.06%
	5	1250346	0.68%	1216728±2223	0.07%±0.12%
circle (width=3)	2	2241872	17.58%	1721610±6633	11.45%±1.04%
	3	2000580	8.59%	1691143±3865	6.72%±0.54%
	4	1817165	4.49%	1672465±2259	3.37%±0.40%
	5	1774078	2.34%	1675028±926	1.91%±0.20%
circle (width=6)	2	2178246	18.65%	1348983±3880	4.45%±1.11%
	3	1715610	8.30%	1295023±1162	1.98%±0.25%
	4	1516673	4.39%	1278273±3337	1.86%±0.08%
	5	1380865	1.86%	1269815±634	1.37%±0.08%

Table 2. Results of the MRF model for multi-class segmentation on BrainWeb data

	BP		CE	
noise level	minimum	error	minimum	error
3%	348740	4.25%	211878.3 ± 2737.2	2.77% ± 0.08%
5%	488391	6.00%	257021.0 ± 1505.7	4.14% ± 0.06 %
7%	487083	6.58%	229624.5 ± 1184.7	3.57% ± 0.02%
9%	691409	11.65%	525039.5 ± 768.6	9.08% ± 0.01%

3.3 Clinical Data: Vascular and Cerebral Datasets

We then apply the model to clinical vascular and cerebral datasets, and compare the performance of the two solvers.

The first dataset used is phase contrast magnetic resonance angiographic (PCMRA) images (size of 308 × 355) obtained from the University Hospital of Zurich, Switzerland. We test on 8 sets of data and the overall mean energy obtained by the CE solver is 174338.81 ± 765.84, which is smaller than the energy 177475.56 ± 769.93 obtained by BP. Fig. 2 shows one example of the experiments. Although we do not have the ground truth to compare the two solvers quantitatively, we still can find by visual inspection that the CE acquires a little better segmentation than BP which is consistent to the smaller energy values. BP may be trapped in those local minima caused by the noise which can be seen in the middle subfigure in Fig. 2 and CE can alleviate the problem although not completely solve it.

The second dataset used is acquired from the Internet Brain Segmentation Repository (IBSR) [18] which we can get some segmented data as ground truth. There are three labels {white matter, gray matter, others} and the image is 141 × 149 large. We again test the model on 8 datasets and calculate the energy and accuracy obtained by the two solvers, which is shown in Tab. 3. We repeat the CE algorithm for 5 times and calculate the means and standard deviations considering its stochastic property.

Results on both synthetic and clinical data show that the CE solver is able to find lower energy (which means more global) than BP. The improvement is usually from 5% to 45% and accuracy can be increased by 1 to 7 percentage points. It is worth pointing out that lower energy does not always give better segmentation results unless the formulation is good enough, but this is not the

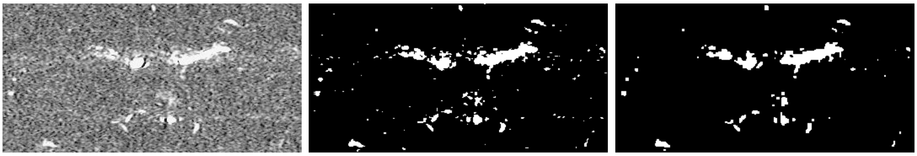

Fig. 2. The left is a region of interest of an original PCMRA image (contrast enhanced). The middle and right are the segmented images solved by BP and CE, respectively.

Table 3. Results of the MRF model for multi-class segmentation on IBSR brain data

	BP		CE	
dataset	minimum	error	minimum	error
1	19361	3.32%	15103 ± 39	2.29% ± 0.08%
2	28049	4.85%	24665 ± 24	3.70% ± 0.14%
3	34641	7.72%	30657 ± 28	6.87% ± 0.24%
4	31483	4.98%	25458 ± 12	2.94% ± 0.02%
5	22104	3.08%	18008 ± 15	1.80% ± 0.08%
6	35652	8.27%	31274 ± 59	5.50% ± 0.16%
7	37561	8.07%	31646 ± 39	6.23% ± 0.24%
8	29782	4.72%	25317 ± 24	2.86% ± 0.16%

goal of solvers, which should find energy as low as possible. The running time of CE algorithm for one of the clinical vascular images is around 300 minutes (BP needs about 15 minutes) and for one of the clinical brain images around 60 minutes (BP needs about 9 minutes) on a computer with 1.3GHz CPU and 500 MB memory. Decreasing the parameter ρ will reduce the required number of iterations and thus the computation time of CE at the cost of likely increasing the energy value found in the end.

4 Discussion and Conclusion

In general, the CE algorithm is very simple and easy to implement. It is an iterative procedure and in each iteration, a sequence of samples is generated according to a certain probability distribution. The method chooses one threshold of objective function value and just focuses on those samples whose performance (e.g., lower energy) is better than this threshold. Then CE updates the distribution parameters according to these good samples. This completes one iteration. We can see that CE is stochastic and no deterministic decision is made, which gives CE the ability to find more global optima than deterministic methods. Compared with other stochastic solvers like simulated annealing (SA), CE is obviously more efficient because it concentrates on a few high-performance samples among a large collection of random samples and quickly converges to states which have good performance. Moreover, the CE algorithm only evaluates the energy values and requires no specific energy form. Thus it can be applied to any type of objective functions. This makes CE a more general solver than BP and GC. Another advantage of CE lies in its insensitivity to initialization because CE's initialization is unchanged for any inputs, i.e., the parameter vector $\boldsymbol{v}_0$ is always set to uniform distribution and every element is $1/m$.

The study on the computational complexity of CE algorithm is still an open problem partly because CE can apply to all kinds of different applications and the necessary total number of iterations is different. For some applications like maxcut and partition problems [14], the theoretical complexity of CE is $\mathcal{O}(n^3 \ln n)$ and its empirical complexity is $\mathcal{O}(n \ln n)$. The CE solver is not as efficient as BP or GC. But, since the architecture of CE algorithm is inherently parallel and its steps are all simple, it has large potential to speed up. Regarding space complexity, CE can occupy a lot of space if we store all the samples. However, the problem can be solved by keeping a small portion of them since CE just uses the high-performance part.

In this paper, we have introduced a new MRF solver, namely cross entropy (CE), applied it to medical image segmentation and shown some advisable properties of it compared to an existing popular MRF solver, belief propagation (BP). CE algorithm is a general solver that can be applied to any type of MRF models and it is stochastic and iterative, which endows it with the capability to find more global optima than BP. This makes CE useful in medical image segmentation or any other tasks which can be formulated as energy minimization problems. However, CE requires more computation time and space than deterministic optimization methods. Since it has parallel architecture and is very simple in itself, CE has potential to be accelerated by parallel processing or algorithm optimization. This will be the major work to be done in the future.

References

1. N.Villain, Y.Goussard, J.Idier, M.Allain: Three-dimensional edge-preserving image enhancement for computed tomography. IEEE TMI **22** (2003) 1275–1287
2. Y.Chen, A.Amini: A map framework for tag line detection in spamm data using markov random fields on the b-spline solid. IEEE TMI **21** (2002) 1110–1122
3. M.Svensen, F.Kruggel, D.Cramon: Probabilistic modeling of single-trial fmri data. IEEE TMI **19** (2000) 25–35
4. A.C.S.Chung, J.A.Noble, P.Summers: Vascular segmentation of pcmra based on statistical mixture modeling and local phase coherence. IEEE TMI **23** (2004) 1490–1507
5. Y.Boykov, O.Veksler, R.Zabin: Fast approximate energy minimization via graph cuts. IEEE on PAMI **23** (2001) 1222–1239
6. S.Kirkpatrick, D.Gellatt, M.Vecchi: Optimization by simulated annealing. Science **220** (1983) 671–680
7. J.Besag: On the statistical analysis of dirty pictures. J. of the Roy. Stat. Soc. **Series B 48** (1986) 259–302
8. S.Z.Li, ed.: Markov Random Field Modeling in Image Analysis. Springer-Verlag Tokyo (2001)
9. J.S.Yedidia, W.T.Freeman, Y.Weiss: Understanding belief propagation and its generalizations. Technical Report, Mitsubishi Electric Research **TR-2001-22** (2001)
10. S.Zhu, Y.Wu, D.Mumford: Frame: filters, random fields, and minimax entropy towards a unified theory for texture modeling. In: IEEE CVPR. Volume 6. (1996) 686 – 693
11. V.Kolmogorov, R.Zabih: What energy functions can be minimized via graph cuts? IEEE Trans. PAMI **26** (2004) 147–159
12. R.Rubinstein: Optimization of computer simulation models with rare events. European Journal of Operations Research **99** (1997) 89–112

13. M.Tappen, W.Freeman: Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In: IEEE ICCV. (2003) 900–906
14. R.Rubinstein, D.Kroese, eds.: The Cross-Entropy Method: A Unified Approach to Cominatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer-Verlag, New York (2004)
15. K.Held, *et al*: Markov random field segmentation of brain mr images. IEEE TMI **16** (1997) 878–886
16. S.Ruan, *et al*: Brain tissue classification of magnetic resonance images using partial volume modeling. IEEE TMI **19** (2000) 1179–1187
17. BrainWeb: (http://www.bic.mni.mcgill.ca/brainweb/)
18. IBSR: (http://www.cma.mgh.harvard.edu/ibsr/)