Bayesian Non-parametric Image Segmentation with Markov Random Field Prior

Ehsan Amid

Department of Information and Computer Science
Aalto University, 02150 Espoo, Finland
ehsan.amid@aalto.fi

Abstract. In this paper, a Bayesian framework for non-parametric density estimation with spatial smoothness constraints is presented for image segmentation. Unlike common parametric methods such as mixtures of Gaussians, the proposed method does not make strict assumptions about the shape of the density functions and thus, can handle complex structures. The multiclass kernel density estimation is considered as an unsupervised learning problem. A Dirichlet compound multinomial (DCM) prior is used to model the class label prior probabilities and a Markov random field (MRF) is exploited to impose the spatial smoothness and control the confidence on the Dirichlet hyper-parameters, as well. The proposed model results in a closed form solution using an expectation-maximization (EM) algorithm for maximum a posteriori (MAP) estimation. This provides a huge advantage over other models which utilize more complex and time consuming methods such as Markov chain Monte Carlo (MCMC) or variational approximation methods. Several experiments on natural images are performed to present the performance of the proposed model compared to other parametric approaches.

Keywords: Multiclass kernel density estimation, Dirichlet compound multinomial distribution, Markov random field prior, Image segmentation.

1 Introduction

Image segmentation techniques usually require some prior information about the regions of interest through a human input to produce satisfactory results [1]. However, in many cases, providing the prior knowledge about the present objects or segments to the system is infeasible. Despite the performed research during many decades, fully unsupervised image segmentation is still a challenging problem due to the fact that there is no clear objective measure about how a particular segment can be considered as meaningful. Moreover, parametric methods e.g. Gaussian mixture models (GMM) [5] which aim to fit predefined density functions to the existing distribution of the data may fail to capture the underlying structure due to poor assumptions or complex density shapes [3]. In these situations, an unsupervised method which can infer the required

J.-K. Kämäräinen and M. Koskela (Eds.): SCIA 2013, LNCS 7944, pp. 76–84, 2013.

information only using the data itself without any severe assumptions that may restrict the estimation process can be highly advantageous.

Along with different density estimation methods, spatial smoothness constraints have been used in image segmentation problems to take into account the local commonality of the location while grouping the data together [2]. The performance of the segmentation technique highly depends on using this a priori knowledge about the image that the adjacent pixels presumably belong to the same cluster. These constraints can be imposed on the prior probability of the class labels [4] or alternatively, can be considered in a more meaningful sense on the hyper-parameters of the prior distribution of the mixing portions through a Markov random field (MRF) distribution [2]. In this paper, a Bayesian framework for unsupervised image segmentation is presented which automatically infers the required information about the regions through the local contextual analysis of the data as well as considering its global distribution, at the same time. Additionally, the proposed method allows incorporation of some partial prior information on the assignment of the points through the posterior class label probabilities.

The remainder of the paper is organized as follows: The non-parametric density estimation method is presented in Section 2. Section 3 describes the Dirichlet compound multinomial distribution on the class label variables. Spatial smoothness constraints is given in Section 4 and in Section 5, MAP-EM estimation algorithm for the proposed method is presented. A discussion on the feature extraction for image segmentation is given in Section 6. Section 7 contains the experimental results of the proposed method on a set of natural images and finally, Section 8 discusses the results and concludes the paper.

2 Non-parametric Density Estimation

In non-parametric density estimation, only a few assumptions are made about the shape of the distribution generating the data. The idea is to consider a probability bump around each data point using a kernel function [5]. With assumption of an i.i.d. dataset, the probability density function in a particular point in the space is then, estimated by summing up the effects of the surrounding data points. Smoother density functions can be obtained by considering smoother kernel functions. A common choice is the Gaussian kernel function defined by

$$\mathcal{K}_{\sigma}(\mathbf{x}, \mathbf{x}') = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right) \tag{1}$$

where σ^2 is the variance and can be chosen separately for each point based on local analysis of the data [10]. In case of multiple classes, the density function of a class is estimated by using only the points belonging to that class. Let $\mathbf{X} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N\}$ be the set of D-dimensional observations (features) generated by K independent processes, $C^j, j = 1, 2, \ldots, K$ and $\mathbf{z}^i$ be the K-dimensional binary latent random variable of point $\mathbf{x}^i$ in which all the elements are equal to zero except a particular element $z^i_j = 1$, which represents the class label of the point. The probability density function of class C^j can be estimated by

$$\hat{f}_j(\mathbf{x}) = \frac{1}{N_j} \sum_{i=1}^{N} z_j^i \mathcal{K}(\mathbf{x}, \mathbf{x}^i) \tag{2}$$

where $N_j = \sum_{i=1}^{N} z_j^i$ and it is used to have a valid probability density function. In case of an incomplete dataset, the class labels are not available explicitly and thus, z_j^i in equation is replaced by its expected value, $E[z_j^i]$.

The expected value of the indicator variable z_j^i under the posterior distribution is given by

$$E[z_j^i] = \frac{\pi_j^i \hat{f}_j(\mathbf{x}^i)}{\sum_{k=1}^{K} \pi_k^i \hat{f}_k(\mathbf{x}^i)} = \gamma(z_j^i) \tag{3}$$

in which, $\pi_j^i = P(z_j^i = 1)$ is the prior probability that $\mathbf{x}^i$ belongs to class C^j. Thus, $\gamma(z_j^i)$ is just the responsibility of j_{th} process for data point $\mathbf{x}^i$ [5]. Substituting z_j^i in (2) with its expected value (3), we may have

$$\hat{f}_j(\mathbf{x}) = \frac{1}{\tilde{N}_j} \sum_{i=1}^{N} \gamma(z_j^i) \mathcal{K}(\mathbf{x}, \mathbf{x}^i) \tag{4}$$

as the density estimate, where N_j in (2) is replaced by $\tilde{N}_j = \sum_{i=1}^{N} \gamma(z_j^i)$. Additional prior information on true class labels of some particular data points (e.g. labeled by a human observer) can be incorporated by considering the responsibility of the corresponding process for those points equal to 1.

3 Dirichlet Compound Multinomial Distribution

For each pixel (feature vector) $\mathbf{x}^i = \left[x_1^i, x_2^i, \ldots, x_D^i\right]^T$, the class label $\mathbf{z}^i$ is taken to be a random variable having a multinomial distribution with a set of parameters $\xi^i = \{\xi_1^i, \xi_2^i, \ldots, \xi_K^i\}$, where K is the number of classes. By definition,

$$P(\mathbf{z}^i|\xi^i) = \frac{M!}{\prod_{j=1}^{K} (z_j^i)!} \prod_{j=1}^{K} (\xi_j^i)^{z_j^i} \tag{5}$$

with

$$\xi_j^i \geq 0, \quad \sum_{j=1}^{K} \xi_j^i = 1, \quad i = 1, 2, \ldots, N \tag{6}$$

Therefore, $\mathbf{z}^i$ is considered as the outcome of M realizations of a multinomial process with K possible outcomes. The conjugate prior distribution on ξ^i follows a Dirichlet distribution with parameters $\alpha^i = \{\alpha_1^i, \alpha_2^i, \ldots, \alpha_K^i\}$

$$P(\xi^i|\alpha^i) = \frac{\Gamma(\sum_{j=1}^{K} \alpha_j^i)}{\prod_{j=1}^{K} \Gamma(\alpha_j^i)} \prod_{j=1}^{K} (\xi_j^i)^{(\alpha_j^i - 1)} \tag{7}$$

where $\alpha_j^i \geq 0$ and $\Gamma(.)$ is the gamma function. The sum of the hyper-parameters $\alpha_0^i = \sum_{j=1}^K \alpha_j^i$ indicates the level of confidence in the prior information about the processes generating the data. Larger values of α_0^i yields lower variance among different realizations of the parameter $\xi^i, i = 1, 2, \ldots, N$ [11].

Under the Bayesian framework, the probability of class label $\mathbf{z}^i$ given α^i is obtained by marginalizing over ξ^i

$$P(\mathbf{z}^i|\alpha^i) = \int_0^1 P(\mathbf{z}^i|\xi^i)P(\xi^i|\alpha^i)\mathrm{d}\xi^i, \quad i = 1, 2, \ldots, N \tag{8}$$

Substituting (5) and (7) into (8), and considering the property

$$\int_0^1 \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K (\xi_j)^{(\alpha_j - 1)} \mathrm{d}\xi = 1 \tag{9}$$

yields the following expression for label probabilities [11].

$$P(\mathbf{z}^i|\alpha^i) = \frac{M!}{\prod_{j=1}^K (z_j^i)!} \frac{\Gamma(\sum_{j=1}^K \alpha_j^i)}{\Gamma(\sum_{j=1}^K \alpha_j^i + z_j^i)} \prod_{j=1}^K \frac{\Gamma(\alpha_j^i + z_j^i)}{\Gamma(\alpha_j^i)} \tag{10}$$

with $i = 1, 2, \ldots, N$.

4 Spatial Smoothness Constraints

In case of a single image, the prior probability value of pixel $\mathbf{x}^i$ generated by process C^j is the outcome a Dirichlet compound multinomial process with only one realization [2]. Therefore, $M = 1$ in (10) and considering the property that $\Gamma(x + 1) = x\Gamma(x)$, the prior label probabilities for $\mathbf{x}^i$ become

$$\pi_j^i = P(z_j^i = 1|\alpha^i) = \frac{\alpha_j^i}{\sum_{k=1}^K \alpha_k^i}, \quad j = 1, 2, \ldots, K \tag{11}$$

The local smoothness constraints in the model is considered by assuming a Markov random field (MRF) [12] spatial prior on the set of all hyper-parameters $\mathbf{A} = \{\alpha^i\}, \quad i = 1, 2, \ldots, N$ of the model

$$P(\mathbf{A}) \propto \prod_{j=1}^K (\beta_j)^{-N} \exp\left[-\frac{1}{2} \frac{\sum_{i=1}^N \left((\alpha_j^i - \tilde{\alpha}_j^i)^2 + \sum_{m \in \mathcal{N}_i} (\alpha_j^i - \alpha_j^m)^2\right)}{\beta_j^2}\right] \tag{12}$$

The second term in the energy function imposes the local smoothness of the hyper-parameters in a neighborhood $\mathcal{N}_i$ of site α^i [2]. The first term controls the confidence of the model by modifying the value of α_j^i in each iteration of the EM algorithm. The higher values of posterior probability $\gamma(z_j^i) = P(z_j^i = 1|\mathbf{x}^i, \mathbf{A})$ tend to increase the α_j^i and vice versa. In this paper,

$$\tilde{\alpha}^{(t+1)} = \tilde{\alpha}^{(t)} \phi(\gamma(z_j^i)) \tag{13}$$

where $\tilde{\alpha}^{(t)}$ is the value of $\tilde{\alpha}$ in the previous iteration of the EM algorithm and $\phi(\gamma(z_j^i))$ is a monotonically increasing function of $\gamma(z_j^i)$. A reasonable choice of $\phi(.)$ is a shifted and scaled version of hyperbolic tangent function, because of its interesting properties and non-linearity. It has a large slope for values near the origin while its slope gets smaller for the larger values of the input. The function is shifted and scaled so that for values of posterior probability which equal to $\frac{1}{K}$ (uniform distribution), $\phi(\frac{1}{K}) = 1$ and $\tilde{\alpha}^{(t+1)} = \tilde{\alpha}^{(t)}$, since the posterior probability equals to the case having maximum entropy and does not provide any useful information for modifying the hyper-parameters. On the other hand, the maximum value of $\phi(\gamma(z_j^i))$, which occurs when $\gamma(z_j^i) = 1$, is chosen in such a way that it increases the $\tilde{\alpha}^{(t)}$ by 5%.

5 MAP-EM Estimation

Considering the prior distribution defined in (12), yields the following MAP function to be maximized.

$$Q(\mathbf{A}|\mathbf{X},\mathbf{Z}) = \sum_{i=1}^{N}\sum_{j=1}^{K}\left\{ z_j^i \left[\log(\frac{\alpha_j^i}{\sum_{k=1}^{K}\alpha_k^i}) + \log(\hat{f}_j(\mathbf{x}^i))\right] + log(\beta_j^{-1}) - \frac{1}{2}\frac{\Delta}{\beta_j^2}\right\} \tag{14}$$

where

$$\Delta = (\alpha_j^i - \tilde{\alpha}_j^i)^2 + \sum_{m\in\mathcal{N}_i}(\alpha_j^i - \alpha_j^m)^2 \tag{15}$$

To maximize (14) at the M-step with respect to α_j^i, its partial derivative with respect to α_j^i should be set to zero. By setting $\frac{\partial Q}{\partial \alpha_j^i} = 0$, we have the following third degree polynomial equation in α_j^i

$$\begin{aligned}(\alpha_j^i)^3 &+ \left[A_{-j}^i - \frac{2\sum_{m\in\mathcal{N}_i}\alpha_j^m\tilde{\alpha}_j^i}{2|\mathcal{N}_i|+1}\right](\alpha_j^i)^2 \\ &- A_{-j}^i\left[\frac{2\sum_{m\in\mathcal{N}_i}\alpha_j^m + \tilde{\alpha}_j^i}{2|\mathcal{N}_i|+1}\right](\alpha_j^i) - \frac{\beta_j^2 A_{-j}^i\gamma(z_j^i)}{2|\mathcal{N}_i|+1} = 0\end{aligned} \tag{16}$$

where

$$A_{-j}^i = \sum_{\substack{k=1\\k\neq j}}^{K}\alpha_k^i, \quad i = 1, 2, \ldots, N \tag{17}$$

and $|\mathcal{N}_i| = 8$ for a 8-neighborhood.

Based on the constraints in (7), that is $\alpha_j^i \geq 0$, the constant term in (16) is negative, therefore the product of the roots is positive. This implies that for real valued roots, either all three roots should be positive or two of them be negative and the other one be positive. Since the quadratic term is positive, the sum of the roots should be negative and therefore, two of the roots are negative and one of them is positive. In case of a real root along with a pair of complex conjugate roots, the real root should be positive since the product of the roots is positive.

Finally, the update equation for $\beta_j, \quad j = 1, 2, \ldots, K$ can be obtained by setting $\frac{\partial Q}{\partial \beta_j} = 0$ which yields

$$\beta_j^2 = \frac{1}{N}\sum_{i=1}^{N}\left[(\alpha_j^i - \tilde{\alpha}_j^i)^2 + \sum_{m\in\mathcal{N}_i}(\alpha_j^i - \alpha_j^m)^2\right], \quad j = 1, 2, \ldots, K. \tag{18}$$

The overall procedure for MAP-EM estimation is summarized as follows:

- Initialize the posterior class label probabilities as well as the parameters α^i and $\tilde{\alpha}^i, \quad i = 1, 2, \ldots, N$ of MRF. Moreover, find the kernel values between each pair of pixels (features) in the image.
- Do until the MAP function (14) does not change significantly:
    - Find likelihood values $P(\mathbf{x}^i|C^j) = \hat{f}_j(\mathbf{x}^i), \quad j = 1, 2, \ldots, K$ for each point $\mathbf{x}^i$ using (4).
    - E-Step
        - Find posterior probabilities $\gamma(z_j^i)$ that pixel $\mathbf{x}^i$ belongs to class C^j.
        - Update $\tilde{\alpha}^i$'s using (13).
    - M-Step
        - Update the parameters of the Dirichlet distribution by only keeping the real non-negative roots of the polynomial equation (16).
        - Update the prior probabilities for the pixel labels (11).
        - Update the class variances $\beta_j^2, \quad j = 1, 2, \ldots, K$ (18).
- End

6 Feature Extraction for Texture Representation

Extracting the proper features to represent meaningful textural properties is an important step in image segmentation. In this paper, Blobworld features [9] were used because of their good performance in natural image processing.

6.1 Blobworld Features

Blobworld features include a smoothed version of L*a*b* scale values as color features, as well as three texture features (polarity, anisotropy and contrast) [9]. Unlike the color which is a point property, texture representation requires local analysis of a pixel in its neighborhood. The process consists of finding the proper smoothing scale in a pixel neighborhood and then, calculating the texture features using the eigenvalues and first eigenvector of the second moment matrix in the corresponding scale.

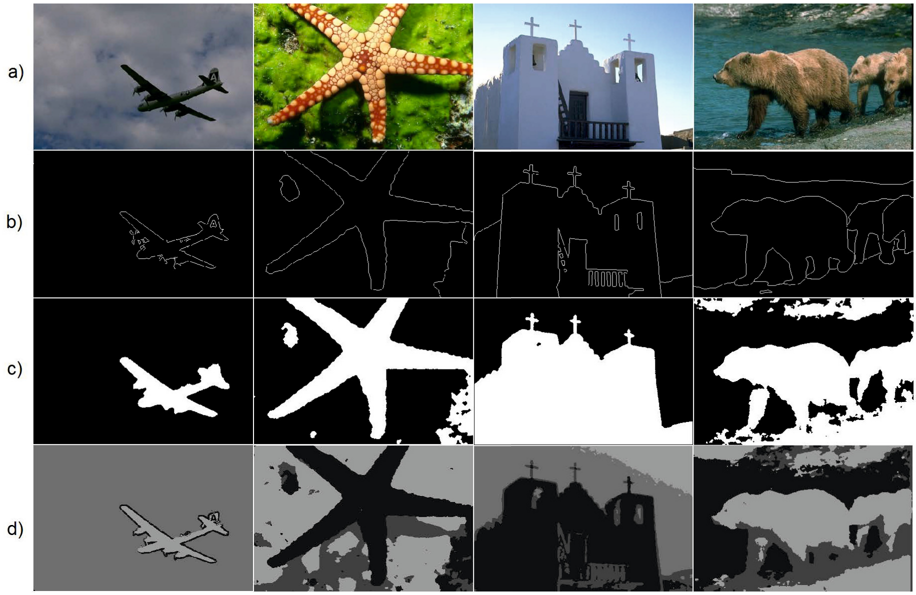

Fig. 1. Segmentation results: a) Original image, b) Ground truth, c) Results the of proposed method with $K = 2$, d) Results of the proposed method with $K = 4$

7 Experimental Results

Since the EM algorithm is sensitive to initialization, the experiments were performed several times using random initializations of the parameters and the termination criterion to be satisfied was defined as the percentage of change of the value of (14) between two consecutive iterations be less than the predefined threshold. Additionally, morphological smoothing operations were performed to remove small isolated components in the final stage.

The experiments on natural images were carried out on a subset of 50 images of Berkeley Segmentation Dataset [6, 7] which also contains the manual border annotation of the images as the ground truth for comparison purpose. Examples of segmentation results are shown in Fig. 1 and Fig. 2. As it can be seen, objects and background are separated accurately with respect to the ground truth. Moreover, due to the smoothing property, increasing the number of clusters does not simply produce larger number of isolated segments that are usually trivial and may only confuse the result. Additionally, by using Blobworld features, the boundaries between textured regions are maintained, as discussed in [2].

To obtain quantitative comparison of the results with conventional mixtures of Gaussians method, for a subset of 25 image containing meaningful object and background, the number of True Positive TP (classified as object by both the algorithm and the ground truth), True Negative TN (classified as non-object by both the algorithm and the ground truth), False Positive FP

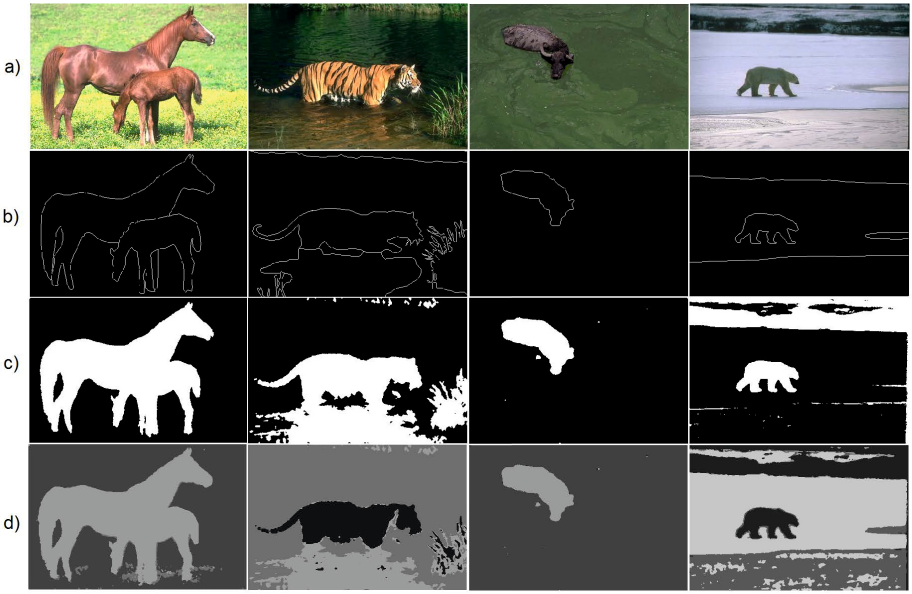

Fig. 2. Segmentation results: Continued

(non-object classified as object) and False Negative FN (object classified as non-object) pixels were stored. Then, recall and fall-out,

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{19}$$

and

$$\text{fall-out} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{20}$$

were calculated as a measure of segmentation quality, as suggested in [8]. Additionally, the overall accuracy

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{21}$$

was also considered for each image.

Table 1. Recall, fall-out and accuracy for 2-class segmentation

	Recall	Fall-out	Accuracy
GMM	78.55%	19.14%	80.89%
Proposed method	85.25%	14.77%	85.79%

The average values of recall, fall-out and accuracy over the subset of images for 2-class segmentation for both methods are shown in Table 1. As can be observed, the proposed method outperforms GMM in all the cases.

8 Conclusion

A fully unsupervised image segmentation method was presented in this paper. Kernel density estimation was performed using posterior class label probabilities of all data points and a MRF prior probability distribution was considered to impose spatial smoothness on the adjacent pixels. Update equations of the parameters were derived in a closed form solution in MAP-EM framework. The experimental results on natural images show that the proposed method outperforms the conventional GMM method in all cases.

References

1. Ding, L., Yilmaz, A., Yan, R.: Interactive Image Segmentation Using Dirichlet Process Multiple-View Learning. IEEE Transactions on Image Processing 21(4), 2119–2129 (2012)
2. Nikou, C., Likas, A.C., Galatsanos, N.P.: A Bayesian Framework for Image Segmentation With Spatially Varying Mixtures. IEEE Transactions on Image Processing 19(9), 2278–2289 (2010)
3. Andreetto, M., Zelnik-Manor, L., Perona, P.: Non-Parametric Probabilistic Image Segmentation. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, October 14-21, pp. 1–8 (2007)
4. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging 20(1), 45–57 (2001)
5. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
6. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings of the Eighth IEEE International Conference on Computer Vision, ICCV 2001, vol. 2, pp. 416–423 (2001)
7. The Berkeley Segmentation Dataset and Benchmark, `http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds`
8. Bowyer, K., Kranenburg, C., Dougherty, S.: Edge detector Evaluation Using Empirical ROC Curves. Computer Vision and Image Understanding 84(1), 77–103 (2001)
9. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(8), 1026–1038 (2002)
10. Zelnik-Manor, L., Perona, P.: Self-Tuning Spectral Clustering. In: Advances in Neural Information Processing Systems (NIPS), pp. 1601–1608 (2005)
11. Frigyik, B.A., Kapila, A., Gupta, M.R.: Introduction to the Dirichlet Distribution and Related Processes. UWEE Technical report (2010)
12. Li, S.Z.: Markov Random Field Modeling in Image Analysis. Springer, Tokyo (2001)