BOUNDARY AWARE IMAGE SEGMENTATION WITH UNSUPERVISED MIXTURE MODELS

Thorsten Wilhelm, Christian Wöhler

Image Analysis Group, Technical University Dortmund,
Otto-Hahn-Str. 4, 44227 Dortmund, Germany

ABSTRACT

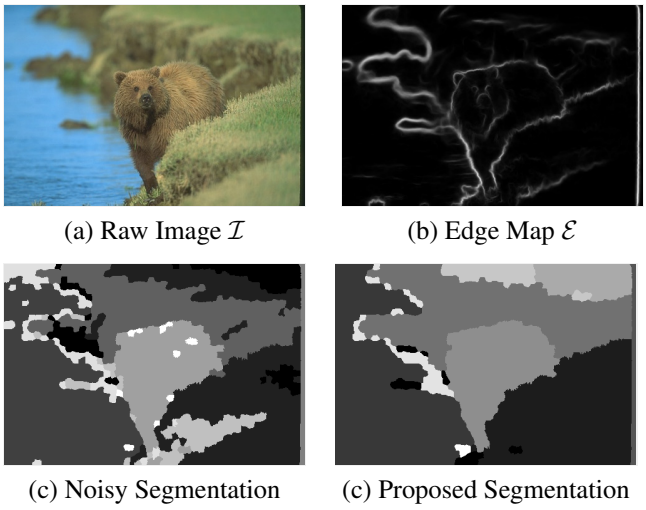
Recent image segmentation methods focus mostly on the topic of semantic segmentation and are trained in a supervised fashion. This work proposes a novel and unsupervised bayesian segmentation method, which includes the edges of an image as part of the model. This reduces typical noise patterns in unsupervised segmentation and increases the overall capability of the segmentation. Two ways are proposed to include edges. One way is a passive edge model which grades the segmentation according to a precomputed edge map, and a second variant where this method is used in conjunction with an active edge movement scheme. Both methods are tested on a publicly available dataset, compared to methods from the literature, and the benefit of including edges is emphasized.

***Index Terms*—** Segmentation, Bayesian, Unsupervised, Edges

1. INTRODUCTION

Image Segmentation is a commonly used tool in a vast range of computer vision tasks. It ranges from medical image analysis [1] to autonomous driving [2]. Recently, a lot of work has been achieved by training Convolutional Neural Networks (CNNs) on semantic segmentation tasks. However, to the best of our knowledge no neural network is able to perform reasonably well on general purpose segmentation tasks like the Berkeley Segmentation Database (BSD500) [3], which does not offer a vast amount of semantically annotated training images. In contrast to the BSD500 other datasets like Microsoft Common Objects in Context (COCO) [4] offer such required semantic labels to perform supervised learning on a segmentation task. However, acquiring a vast amount of annotated images always comes at a high cost. At this point unsupervised methods become vital, because they do not need annotations to perform reasonably well.

In this work the focus resides on unsupervised image segmentation, namely mixture models. This means, no region labels or any knowledge about the appearing objects or scenes are used to aid the algorithm. Furthermore, the number of different regions is inferred from the data on a per image basis.



(a) Raw Image $\mathcal{I}$ (b) Edge Map $\mathcal{E}$

(c) Noisy Segmentation (c) Proposed Segmentation

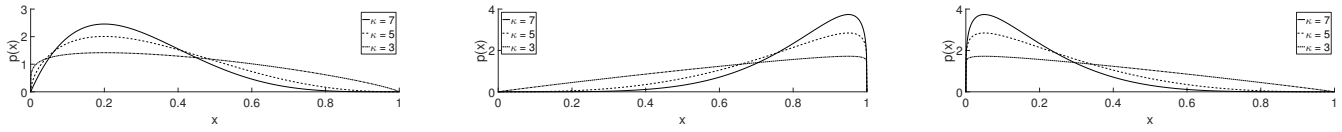
Fig. 1. From upper left to lower right: Raw image $\mathcal{I}$, edge map $\mathcal{E}$, noisy segmentation, and segmentation by proposed method. Due to the integration of edges within the modelling process a solution less prone to segmentation noise is achieved. Raw image taken from BSD500 [3].

Due to the probabilistic nature of mixture models, uncertainty regarding the most probable label is visible by noise in the segmentation (see Fig. 1). In this work we propose to include a second input, such that boundaries in the segmentation shall only appear in places where an edge is probably visible. This improves current mixture models in two ways: first, the estimation process is now done with respect to edges, which should improve the overall capability of the segmentation algorithm, and second, the previously described segmentation noise is strongly reduced.

2. RELATED WORK

Mixture models are one way to perform unsupervised learning or clustering. The general form of mixture models is written as

$$p\left(\mathbf{X}|\mathbf{\Theta}\right) = \sum_{i=1}^{C} \pi_i \phi\left(\mathbf{X}|\mathbf{\Theta}_i\right) \tag{1}$$



(a) $\omega = 0.2$ (b) $\omega = 0.95$ (c) $\omega = 0.05$

Fig. 2. Influence of choosing an appropriate mode for the beta distributions of the edge model. If the mode ω is chosen as a value of the edge map (e.g. $\omega = 0.2$), too much probability mass is concentrated near zero (a), which means a low probability of an edge. This misbehaviour can be circumvented if the mode is chosen by the current segmentation.

with $\mathbf{X}$ as the input data, C as the number of mixture components, $\mathbf{\Theta}$ as the set of parameters, and $\phi(\cdot)$ as the probability density function of an arbitrary probability distribution. By changing $\phi(\cdot)$ different mixture models can be constructed. The most representative member of this class of distributions is the Gaussian mixture model (GMM) [5], which arises when $\phi(\cdot)$ is set to the Gaussian distribution. Other alternatives in the context of image analysis include mixtures of generalised hyperbolic distributions [6, 7] and various variants of Student's t-distribution [8, 6, 9].

Besides the aforementioned segmentation noise, mixture models suffer from the challenge to select an appropriate number of mixture components. The relation to the number of segments may be tried to be learned [7, 9], derived by cluster validations, or inferred from the provided image itself. One class of algorithms which are capable of doing this are Dirichlet processes.

2.1. Dirichlet Process Mixture Models

Dirichlet processes (DP) [10] can be used to model the influence of changes in the dimension of the model space. In the application scenario of mixture models the DP is used in conjunction with a mixture model and the resulting Dirichlet Process mixture model (DPMM) can then be used to model the uncertainty regarding the number of mixture components. Hence, the number of mixture components need not be known a-priori, but is inferred in a data driven fashion. An in-depth review of Dirichlet process models can for instance be found in [11].

The DPMM framework appears to be beneficial compared to other methods because of several reasons. First, the DPMM allows to model the uncertainty in the estimates of the number of mixture components. Second, the DPMM uses a coherent statistical framework with verified properties. Third, the DPMM does not rely on unjustified assumptions like common cluster validation criteria or any other heuristic to estimate the number of mixture components, and last, it solves the problem in a data driven way. Furthermore, the parameters of the mixture components are estimated as well. The DPMM can be written as

$$G \sim DP(G_0, \alpha) \tag{2}$$

$$p_i \sim G \tag{3}$$

$$x_i \sim F(p_i) \tag{4}$$

with G_0 as the base distribution, α as the dispersion parameter, p_i as the group level parameters, and x_i as the data points. Note that this is a generative framework, which means that this framework can be interpreted in two directions. In a top-down fashion where a data point is generated by sampling a value from the distributions, and in a bottom-up fashion, where the most suitable component is selected based on a given data point. Commonly, G_0 is chosen as a univariate Gaussian distribution with zero mean, and $F(\cdot)$ is chosen as a Gaussian distribution, which matches with the dimension of the input data. Typically, the parameters of a DPMM are estimated with Markov chain Monte Carlo (MCMC) [11]. However, the computational cost is high, due to the iterative sampling scheme. Therefore, a domain specific approximation technique using superpixels is used to reduce the computational burden [9].

2.2. Edges in Image Segmentation

Edges are treated in various ways in computer vision tasks. For instance by using active contour models (ACM) [12] or active shape models (ASM). However, these methods rely on a manual initialisation or on learned shape models. One way to include edges as an additional source of information within a probabilistic framework are described in [13], where Student's t-distribution is used to enforce spatial smoothness. Unfortunately, no method to infer the number of mixture components is presented.

3. EDGE MODELS

In this work two ways to include edges are proposed. A passive edge model (PEM) and a passive edge model in conjunction with an active edge movement scheme (AEM), where the boundary between two segments is shifted. The former rates a change in the segmentation according to an edge model, whereas the latter actively proposes to modify the current

boundaries of the segmentation in a beneficial way. Note that every pixel has its own edge model which is independent of the neighbouring pixels.

3.1. Passive Edge Model

The proposed PEM weights each resulting edge in the current segmentation of the Markov chain according to a pre-defined model distribution. The definition of this model is described in the following paragraph.

The edge model should fulfil three requirements. First, it should give a high probability to data points where an edge is probably present and this edge is captured by the current segmentation. Second, a high probability should be given to data points where an edge is unlikely and this absence is captured by the current segmentation. And last, a low probability is given to those parts where the current segmentation does not match the probable edge map.

First, an edge map $\mathcal{E}$ is computed with the method provided in [14]. The values of the resulting edge map reside in the range of $x \in [0, 1]$. A natural choice in this domain is the beta distribution having the probability density function (pdf)

$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}. \quad (5)$$

The parameters are denoted by α and β, and $\Gamma(\cdot)$ is the gamma function. The influence of the parameter choices are depicted in Fig. 2. The mode of the distribution ω equals

$$\omega = \frac{\alpha - 1}{\alpha + \beta - 2}. \quad (6)$$

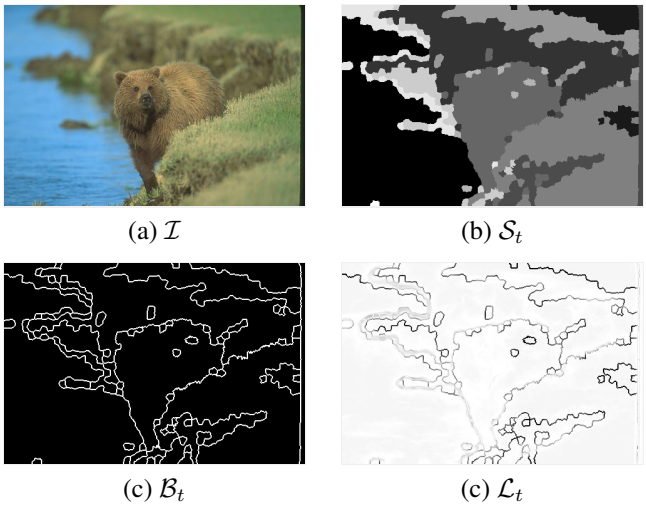
It is then possible to use a different parametrization of Eq. 5 where the parameters α and β are expressed in terms of the mode ω and a concentration parameter $\kappa = \alpha + \beta$:

$$\alpha = \omega(\kappa - 2) + 1, \quad (7)$$
$$\beta = (1 - \omega)(\kappa - 2) + 1. \quad (8)$$

With the help of this parametrisation it is easier to define an appropriate edge model, because the values of α and β can now be chosen according to the mode of the distribution. Due to the flexibility of the beta distribution the edge model can now be built according to the aforementioned requirements. In contrast to common practice, the input of the edge model is not the current segmentation, but the edge map $\mathcal{E}$.

With the help of the current segmentation $\mathcal{S}_t$ the resulting boundary mask $\mathcal{B}_t$ at iteration t is computed. Note that $\mathcal{B}_t$ is binary. Therefore, it is rather difficult to compare $\mathcal{E}$, where all values are from the unit interval, with $\mathcal{B}_t$. As can be seen in Fig. 1(b) not all edges are equally probable and not all pixels around an edge are equally probable.

If an edge has a low probability, say around 0.2, and a beta distribution with mode $\omega = 0.2$ and a reasonably high concentration is placed at this edge, the corresponding probabilities will be low (see Fig. 2(a)), because B_t is binary and the resulting value is one if an edge is present at this position. Therefore, it would be highly improbable that an edge is accepted at this position, which contradicts the prior belief of 0.2. Since this behaviour is undesirable, $\mathcal{E}^i$ is not a good choice as the mode of the edge model at pixel i. However, we define the edge model the other way round. To stick with our example of a weak edge with probability 0.2, we now choose the value of the current boundary $\mathcal{B}_t^i$ as the mode of the distribution, which results in the distributions depicted in Fig. 2(b) and (c). If $\mathcal{B}_t^i = 0$, a high probability score can be achieved. However, if $\mathcal{B}_t^i = 1$, a lower, but still reasonable amount of score can be gained by setting this pixel to an edge pixel. Therefore, $\mathcal{B}_t^i$ is chosen as the mode and the concentration parameter is set to $\kappa = 2.5$. This model is termed passive edge model and is used in conjunction with a mixture model in order to compute the posterior distribution of the segmentation.



(a) $\mathcal{I}$ (b) $\mathcal{S}_t$ (c) $\mathcal{B}_t$ (c) $\mathcal{L}_t$

Fig. 3. From upper left to lower right: Raw image $\mathcal{I}$, current segmentation $\mathcal{S}_t$, boundary image $\mathcal{B}_t$, and likelihood of edge model $\mathcal{L}_t$. Raw image taken from BSD500 [3].

3.2. Active Edge Movement

An active edge movement appears to be beneficial compared to a passive evaluation because by moving the border of the segmentation the label of several superpixels can be changed at once, which only happens by chance with an PEM. AEM should speed up the convergence and allow the Markov chain to surpass local minima, where the passive edge model may get stuck because it is not probable that complete borders are moved by chance. To select an appropriate border the current border image $\mathcal{B}_t$ is evaluated with the passive edge model.

As a first step each appearing border needs to get an ap-

propriate unique label. This is done by imposing two constraints. First, the line must be connected by a four point neighbourhood, and second the adjacent labels of the line have to be equal. After the labelling of all appearing lines in $\mathcal{B}_t$ a score for each line can be computed from the passive edge model. The scores are then normalized and transformed to the unit interval. Afterwards one of the lines is randomly selected based on its inverse score. This leads to the fact that lines with the lowest scores are updated more frequently in the Markov chain.

After a border is selected two new segmentations are proposed. Since the current border has two adjacent labels A and B, the first variant sets the adjacent superpixels of the border to A and the second variant to B. Therefore, the border is either shrunk or expanded. Finally, the posterior distribution of the resulting segmentations are computed and one of the two segmentations is accepted or declined in a common Metropolis-Hastings acceptance step [15].

4. EXPERIMENTS

Following [9], for all experiments 1500 SLIC superpixels [16] are used and the most probable point inside every superpixel is used to describe this superpixel. This amounts to a reduction in the computational burden of roughly 99.99% on the BSD500. Moreover, the position and the colour values in the LAB colour space are used as features. A mixture of multiple scaled t-distributions [17] is chosen as mixture distribution as a trade-off between model flexibility and complexity. The parameters of the mixture model are estimated with DRAM [18].

Two commonly used measures are presented to access the quality of the achieved segmentation and compare the proposed method to methods from the literature. According to [3], the Probabilistic Rand Index (PRI) computes the overlap between two segmentations as

$$\mathrm{PRI}(A, G) = \frac{1}{T} \sum_{i<j} \left[c_{ij} f_{ij} + (1 - c_{ij})(1 - f_{ij}) \right], \quad (9)$$

where c_{ij} indicates if pixel i and j have identical labels in A and G, and f_{ij} is the corresponding probability of this event. In contrast, VoI measures the difference in terms of the average conditional entropy between two segmentations A and G, defined in [3] as

$$\mathrm{VoI}(A, G) = H(A) + H(G) - 2I(A, G) \quad (10)$$

with entropy $H(\cdot)$ and mutual information $I(\cdot)$.

The results of the proposed segmentation method on the BSD500 are presented in Table 1. Note that the reported measures are given at the optimal data set scale (ODS) [3] if the method is supervised, which implies that parameters like thresholds or the number of mixture components are calibrated on a training set for the entire test set. However, the proposed method is unsupervised and therefore not calibrated or trained on a training set.

	BSD500 ODS		
	PRI	VoI	supervised
gPb [3]	**0.83**	**1.69**	✓
NCutsMRF [19]	0.78	2.34	✓
MutualNN [20]	0.79	2.22	✓
t_{MS}MM [9]	0.82	2.22	✓
k-means [21]	0.67	2.32	-
agglomerative [5]	0.66	2.28	-
DPMM (this work)	0.80	2.39	-
PEM (this work)	**0.83**	1.94	-
AEM (this work)	**0.83**	1.90	-

Table 1. Evaluation of the segmentation accuracy of different algorithms on the BSD500 dataset [3]. Probabilistic Rand Index (PRI) and Variation of Information (VoI) are presented at the optimal data set scale (ODS) for the supervised methods, which are trained on training images with a known ground truth. Bold numbers indicate the best numbers. Although it is unsupervised, the performance of the proposed method is comparable to that of supervised state-of-the-art approaches.

DPMM indicates the performance of the initialisation without the edge model. PEM and AEM are the results with the respective edge model included. It is noteworthy that including the edge model greatly increases the performance and although the proposed method is unsupervised it is able to outperform most of the supervised methods from the literature in terms of PRI and VoI and is far superior to other unsupervised methods like k-means clustering [21] or agglomerative clustering [5], which are both used in conjunction with the Calinski-Harabasz criterion [22] to determine a suitable number of classes.

5. CONCLUSION

In this work a novel way to include edges as part of a mixture model for image segmentation is presented. The presented method is unsupervised. It determines a suitable number of segments on a per image basis with the help of a DPMM independently of training images. The proposed method provides encouraging results in comparison to supervised methods.

6. ACKNOWLEDGEMENTS

This work has been supported by the German Research Foundation (DFG) within project Wo1800/ 6-1 ('Partially Supervised Learning of Models for Visual Scene Recognition').

7. REFERENCES

[1] Li Wang, Yaozong Gao, Feng Shi, Gang Li, John H. Gilmore, Weili Lin, and Dinggang Shen, "Links: Learning-based multi-source integration framework for segmentation of infant brain images," *NeuroImage*, vol. 108, pp. 160 – 172, 2015.

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2011.

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *The elements of statistical learning*, vol. 1, Springer series in statistics Springer, Berlin, 2001.

[6] Sharon X Lee and Geoffrey J McLachlan, "Model-based clustering and classification with non-normal mixture distributions," *Statistical Methods & Applications*, vol. 22, no. 4, pp. 427–454, 2013.

[7] Thorsten Wilhelm and Christian Wöhler, "Flexible mixture models for colour image segmentation of natural images," in *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2016, pp. 598–604.

[8] Thanh Minh Nguyen and QM Jonathan Wu, "Robust student's-t mixture model with spatial constraints and its application in medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 31, no. 1, pp. 103–116, 2012.

[9] Thorsten Wilhelm and Christian Wöhler, "Improving bayesian mixture models for colour image segmentation with superpixels," in *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, 2017, pp. 443–450.

[10] Thomas S Ferguson, "A bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.

[11] Yee Whye Teh, "Dirichlet process," in *Encyclopedia of machine learning*, pp. 280–287. Springer, 2011.

[12] Michael Kass, Andrew Witkin, and Demetri Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[13] Giorgos Sfikas, Christophoros Nikou, and Nikolaos Galatsanos, "Edge preserving spatially varying mixtures for image segmentation," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–7.

[14] Piotr Dollár and C Lawrence Zitnick, "Fast edge detection using structured forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 8, pp. 1558–1570, 2015.

[15] W Keith Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[16] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[17] Cristina Tortora, Brian C Franczak, Ryan P Browne, and Paul D McNicholas, "A mixture of coalesced generalized hyperbolic distributions," *arXiv preprint arXiv:1403.2332*, 2014.

[18] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman, "Dram: efficient adaptive mcmc," *Statistics and Computing*, vol. 16, no. 4, pp. 339–354, 2006.

[19] Meng Tang, Dmitrii Marin, Ismail Ben Ayed, and Yuri Boykov, "Normalized cut meets mrf," in *European Conference on Computer Vision*. Springer, 2016, pp. 748–765.

[20] S. M. Abdullah, P. Tischer, S. Wijewickrema, and A. Paplinski, "Hierarchical mutual nearest neighbour image segmentation," in *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Nov 2016, pp. 1–8.

[21] Stuart Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[22] Tadeusz Caliński and Jerzy Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.