A regularization-based approach for unsupervised image segmentation

Aleksandar Dimitriev, Matej Kristan

Faculty of Computer and Information Science, University of Ljubljana
{*ad7414@student, matej.kristan@fri*}*.uni-lj.si*

Abstract

We propose a novel unsupervised image segmentation algorithm, which aims to segment an image into several coherent parts. It achieves this by first over-segmenting the image into several hundred superpixels, which are then iteratively joined on the basis of color and texture information, while simultaneously being regularized by a Markov random field (MRF). In each iteration, similar superpixels are merged, until only a few coherent labels remain in the image. The algorithm was tested on a standard evaluation data set, where it performs on par with state-of-the-art algorithms in term of precision and greatly outperforms the state of the art by reducing the oversegmentation of the object of interest.

1 Introduction

Image segmentation is a popular problem for which many techniques have been developed [1, 2]. The purpose of image segmentation is to partition an image into a number of differently labelled segments that correspond to some meaningful parts of the image. The number of segments, or labels, in the image can run from as little as two, to several hundred, in which case, the segments can be said to correspond to superpixels [3], a small group of pixels that is similar, e.g. in color or texture.

A significant challenge in unsupervised segmentation is determining the number of labels, classes, to decompose the image into. Our contribution is the iterative approach that first clusters hundreds of thousands of pixels into a few hundred superpixels, from which color and texture features are obtained. In the intialization stage, a classifier is learned for each superpixel. Then the classifiers are used to re-classify all superpixels. The classifier learning and superpixel classification iterated and classifiers with poor support in the data are removed from the classifier set. Thus the number of classes gradually reduces until the classifier set reaches a balance. The spatial consistency of segmentation is enforced by a Markov random field that regularizes the labelling process during the classifier learning. See Figure 1 for an example of superpixels, the pair-wise potentials in the Markov random field, and the final regularized segmentation.

Figure 1: From left to right: input image, an example of MRF pairwise consistency encoded by color similarity, and final segmentation.

2 Related Work

In the context of image segmentation, the use of superpixels as a preliminary step has recently emerged [4]. In addition, bottom-up aggregation has been proven to yield good segmentations [5]. Ren and Shakhnarovich [6] use a superpixel oversegmentation merged in a bottom-up fashion. Like most hierarchical agglomeration methods, however, and unlike ours, it is greedy and suffers from error propagation, where incorrectly merged regions are propagated. Joulin et al. [7] also use bottom-up superpixel agglomeration, but rely on co-segmentation, i.e. simultaneously segmenting multiple images with the same foreground object.

Arbelaez et al. [4] tackle the problem of image segmentation through contour detection, but the countours are not always valid segmentations because they are not necessarily closed. In the context of MRFs and superpixels, the most related work is Rantalankila et al. [8], which uses segmentation for single-object proposal, but it only labels the regions into foreground or background using a supervised graph-cut algorithm. Also related is a Gaussian mixture model for segmentation [9] that uses a hidden Markov random field, but does not use superpixels, whereas [10] uses conditional random fields to encode structural information, as well as a SVM, in addition to superpixel segmentation as preliminary step, but it is an object detection framework.

3 Methods

The task of segmenting an image can be formulated as assigning a label to each pixel. The number of labels, however, is unknown a priori. An iterative approach can therefore be applied, that starts with labeling every pixel

with its own label and then gradually reducing the number the number of labels. Our approach is a two-stage approach composed of a pre-segmentation stage and followed by an iterative segmentation stage, both of which are described next, as well as the segmentation algorithm.

3.1 Pre-segmentation

Many regions of the image are visually similar and will likely have been assigned the same label in the final segmentation. This means that such neighboring pixels can be grouped, thus pre-segmenting the image. The image is over-segmented into a few hundred coherent groups of pixels called superpixels. This can be thought of jump-starting the iterative merging of regions, since merging the pixels in the beginning with a hundred superpixels instead of a million pixels reduces the computational cost of any graph-based methods that operate on the (super) pixel level, at the expense of having slightly less-refined boundaries. The algorithm used in this preliminary step is called Simple Linear Iterative Clustering (SLIC) [3].

3.2 Super-pixel feature description and classification

After oversegmenting the image, the next step is to use a descriptor to obtain discriminative features for each superpixel. The feature descriptor used in our algorithm is COLOR moments augmented Cumulative Histogram-based Image Local Descriptor (COLOR-CHILD) [11]. The color part contains the first, second, and third image moments of all three color channels, whereas the texture part includes information obtained from first and second-order derivatives. The color and texture features together comprise the 57-dimensional descriptor (9 color dimensions and 48-bin quantized histogram). These descriptors can readily be used to learn a classifier for each label. For example, assume we have M superpixels labelled by K labels. A classifier such as a one-versus-all support vector machine (SVM) can be learned for each class from the features extracted from the superpixels labelled by the same label, resulting in K SVMs. Each SVM is calibrated by a Platt calibration such that it outputs a probability of observing the feature, given the selected class. These classifiers can then be applied back to each superpixel. Each superpixel is assigned a class label of the SVM with the maximum probablity. In this way the the superpixels are re-labeled. But independent classification of the pixels will likely result in a noisy segmentation and regularization should be enforced.

3.3 Regularization of segmentation

To enforce regularization, we apply a Markov Random Field (MRF) on the superpixels. MRFs are commonly used as a way to encode spatial dependencies present between neighboring pixels in an image. They have found applications in image restoration, stereo vision, and segmentation [12]. Our approach uses MRFs to take advantage of the structural information present in an image, that would otherwise be unused. Each superpixel is a variable, with dependencies between superpixels that share a boundary.

The particular type of MRF applied here, and the corresponding procedure of energy minimization, is described in Kristan et al. [13]. Briefly, the energy function corresponding to the MRF is the following:

$$E = \sum_{i=1}^{M} \log p(\mathbf{f}_i, \theta) - \frac{1}{2}(E(\pi_i, \pi_{N_i}) + E(\mathbf{p}_i, \mathbf{p}_{N_i})), \quad (1)$$

where M is the number of superpixels, π_i denotes the i^{th} superpixel's prior probability distribution over the labels, π_{N_i} is a weighted sum of the priors of i's neighbors $\pi_{N_i} = \sum_{j \in N_i, j \neq i} \lambda_{ij} \pi_j$. The variables $\mathbf{p}_i$ and $\mathbf{p}_{N_i}$ are the corresponding posteriors and the neighborhood averages, similarly to the priors. The particular formulation of the MRF [13] treats the priors as well as posteriors as random variables and enforces an MRF on the prior and an MRF on the posterior.

We also need to specify the joint distribution $p(l_i, \theta) = p(l_i|\theta)p(\theta)$ in (1). We model $p(\mathbf{f}_i|\theta) = \sum_{k=1}^{K} p(\mathbf{f}_i|\theta_k)$ by K support vector machines (SVM) that are trained for each different label. Each SVM outputs a probability distribution over labels for each superpixel. In particular, the k-th Platt-calibrated SVM models the distribution $p(l_i|\theta_k)$. The posteriors $\mathbf{p}_i$ over each pixel are computed by minimizing the cost function (1) according to [13].

3.4 Iterative re-labeling and class reduction

Once the posteriors over superpixels are computed, each superpixel is assigned the label with maximal probability. The classes that receive no superpixels in one iteration are removed from the candidate classes. Then the remaining SVMs are re-learned from the labeled superpixels and the process or re-labeling with MRF constraint is repeated until the labeling converges. The iterative procedure is summarized in Algorithm 1.

Algorithm 1 Unsupervised MRF-based segmentation

```
Input: Image I
(f_i, l_i) ← color-child(spix(I)) ∀i
while ∃i : l_i ≠ l_{i−1} do
    for each label l do
        train a SVM and compute p(l_i, θ)
        minimize the energy function (1)
    for each superpixel i do
        //Assign each i-th superpixel to the MAP estimate of the label
        l_i = arg max_k p_ik
    remove labels with insufficient support
return labelled image I using l
```

4 Results

Our algorithm is evaluated on a standard data set [5] consisting of 100 color images which contain a single object of interest that usually occupies a majority of the image (Figure 2). The ground truth consists of 3 human subjects that segment each image into foreground and back-

ground. The task is to correctly infer the foreground region from the background pixels. The performance is measured by the F measure:

$$F = \frac{2PR}{P+R}, \tag{2}$$

where P and R denote precision and recall, which measure the fraction of the segment that contains foreground, and the fraction of the foreground that is contained by the segment, respectively. In addition to computing the F measure for each segment and reporting the best value as F_{single}, we also computed it for each combination of segments and report the highest value as F_{multi}. Finally, we assess the fragmentation of each method by counting the number of segments that comprise the combined F-measure as follows:

$$Frag_{object} = N - 1, \tag{3}$$

where N is the number of segments. Lower fragmentation, ideally zero, means that the object is represented by a single segment, whereas high $Frag_{object}$ implies over-segmentation.

To analyze the results of our method, we compare it to a number of state-of-the-art segmentation algorithms:

- Probabilistic Bottom-Up Aggregation and Cue Integration [5], denoted by PBACI. It gradually merges pixels into successively larger regions by taking into account intensity, geometry, and texture.
- Segmentation by weighted aggregation [14], denoted by SWA, which determines salient regions in the image and merges them into a hierarchical structure.
- Normalized cuts [15], denoted by N-cuts. It treats the problem of segmentation by computing multiple minimum normalized cuts on a pixel graph.
- Contour detection and hierarchical Image Segmentation [4], denoted by Gpb, which reduces the problem to contour detection and uses spectral clustering to combine local cues into a global framework.
- Mean shift [16], denoted by MS, a general mode-seeking algorithm on a non-parametric probability distribution, such as the color or intensity distribution.

The results of each method, including our own, can be seen in Table 1, which shows the average scores for all images in the data set. Our algorithm is comparable in terms of both variants of the F measure. The advantage of our approach is very apparent in fragmentation, where it significantly outperforms the state-of-the-art, which means that it correctly identifies the object with an average of 1.4 segments, whereas all other methods over-segment it.

It should be noted that there is an inverse relationship between the F measure, specifically F_{multi} and the fragmentation. If a method has high fragmentation, meaning the foreground object is made up of several segments, it is natural to assume that they cover it better than a method that only produces one segment, but the ground truth has only one segment, which should be preferred. Therefore the advantage of our method is correctly delineating the object in the image as being comprised of a single segment. This is because similar superpixels are identified as having the same label early in the iterative process and we are only left with a few segments. A few examples of the segmentation produced by our algorithm are shown in Figure 2.

Table 1: Results of single and multi-segment coverage on the dataset (95% confidence).

Method	F_{single}	F_{multi}	$Frag_{object}$
Ours	0.72 ± 0.01	0.84 ± 0.01	$\mathbf{0.40} \pm 0.03$
PBACI	$\mathbf{0.86} \pm 0.01$	0.87 ± 0.02	1.66 ± 0.30
SWA	0.76 ± 0.02	0.86 ± 0.01	2.71 ± 0.33
N-cuts	0.72 ± 0.02	0.84 ± 0.01	2.12 ± 0.17
Gpb	0.54 ± 0.01	$\mathbf{0.88} \pm 0.02$	7.20 ± 0.68
MS	0.57 ± 0.02	$\mathbf{0.88} \pm 0.01$	11.08 ± 0.96

5 Conclusion

An unsupervised iterative segmentation algorithm was proposed. The results show that the algorithm is comparable to the state-of-the-art in precision and recall, and also outperforms the state-of-the-art by more often correctly identifying the segments belonging to a single object.

Future work will involve comparing different classifiers instead of SVMs, which were chosen for their robustness in high-dimensional data. The segment labelling is also binary (hard), so using soft-labelling, where a segment would belong to different labels simultaneously, should also be explored. The pairwise MRF energy term, i.e. the edge weight between neighhboring superpixels is dependent on color similarity, but could also be extended to texture. The parameters of the method will be analyzed in a greater detail and other pre-segmentation approaches will be explored instead of superpixels. We will also consider a hierarchical approach in which the segmentation presented in this work acts as a prior on pixel-level segmentation, which is expected to further improve the segmentation quality. Lastly, saliency detection, the task of determining the important regions of an image, could benefit from our approach as a preliminary step.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "From contours to regions: An empirical evaluation," in *CVPR*, pp. 2294–2301, IEEE, 2009.

[2] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE TPAMI*, vol. 22, no. 8, pp. 888–905, 2000.

[3] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE TPAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

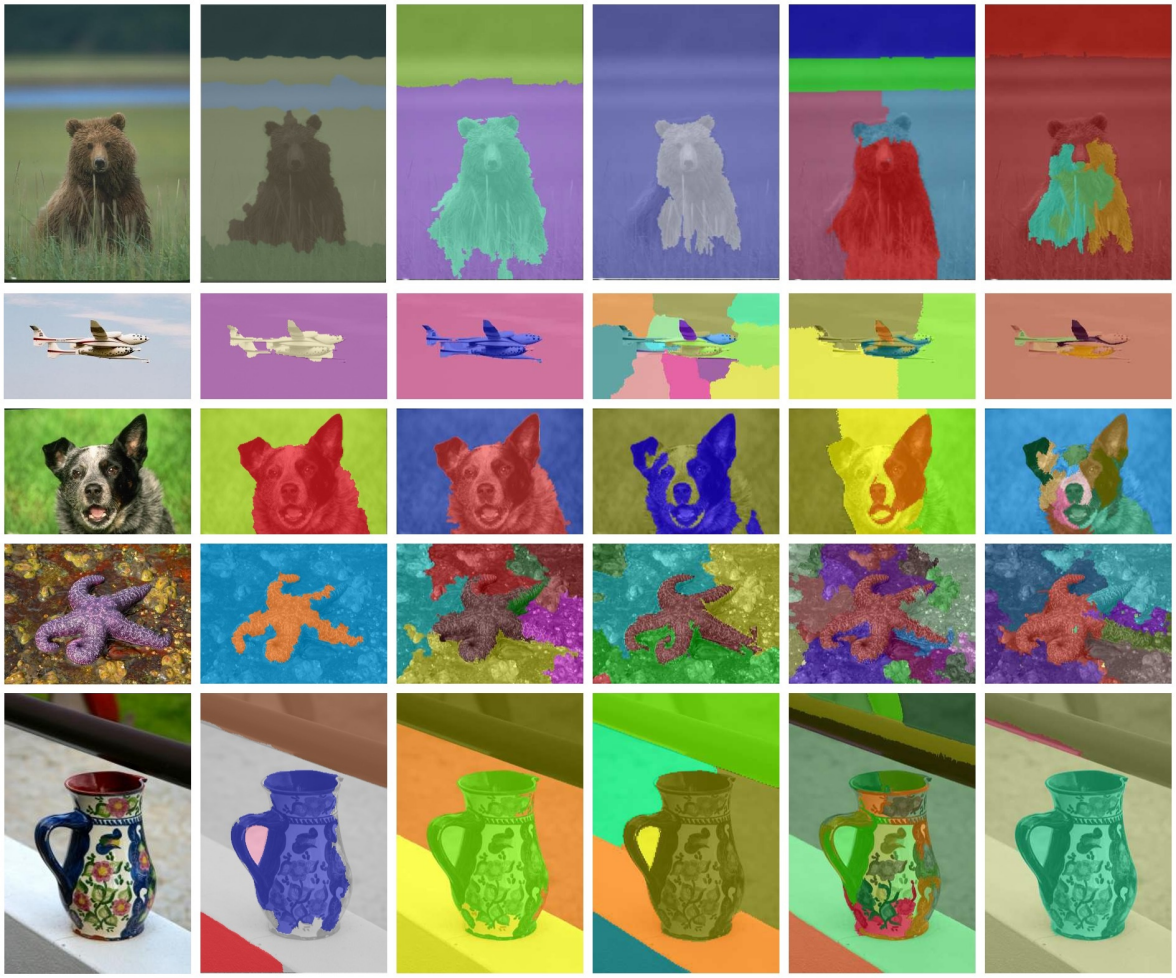

Figure 2: A few images from the dataset and different segmentations. From left to right: Original image, Our method, PBACI, SWA, Normalized cuts, Mean shift.

[4] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE TPAMI*, vol. 33, no. 5, pp. 898–916, 2011.

[5] S. Alpert, M. Galun, A. Brandt, and R. Basri, "Image segmentation by probabilistic bottom-up aggregation and cue integration," *IEEE TPAMI*, vol. 34, no. 2, pp. 315–327, 2012.

[6] Z. Ren and G. Shakhnarovich, "Image segmentation by cascaded region agglomeration," in *CVPR*, pp. 2011–2018, IEEE, 2013.

[7] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 1943–1950, IEEE, 2010.

[8] P. Rantalankila, J. Kannala, and E. Rahtu, "Generating object segmentation proposals using global and local search," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 2417–2424, IEEE, 2014.

[9] K. A. Tran, N. Q. Vo, T. T. Nguyen, and G. Lee, "Gaussian mixture model based on hidden markov random field for color image segmentation," in *Ubiquitous Information Technologies and Applications*, pp. 189–197, Springer, 2014.

[10] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *ICCV*, pp. 670–677, IEEE, 2009.

[11] S. H. Anamandra and V. Chandrasekaran, "Color child: A robust and computationally efficient color image local descriptor," in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pp. 227–234, IEEE, 2014.

[12] J. Wu and A. C. Chung, "A segmentation model using compound markov random fields based on a boundary model," *Image Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 241–252, 2007.

[13] M. Kristan, V. Sulic Kenk, S. Kovacic, and J. Pers, "Fast image-based obstacle detection from unmanned surface vehicles," *Cybernetics, IEEE Transactions on*, 2015.

[14] E. Sharon, M. Galun, D. Sharon, R. Basri, and A. Brandt, "Hierarchy and adaptivity in segmenting visual scenes," *Nature*, vol. 442, no. 7104, pp. 810–813, 2006.

[15] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *International journal of computer vision*, vol. 43, no. 1, pp. 7–27, 2001.

[16] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE TPAMI*, vol. 24, no. 5, pp. 603–619, 2002.