GRM final project
Interactive multi-label segmentation of RGB-D images - Diebold et al.

Martin BROSSET

1 Introduction

Image segmentation is a cornerstone challenge in computer vision, aiming to divide images into meaningful regions applicable across a broad spectrum of fields. Given the dependency of what constitutes "meaningful regions" on the application, fully automated segmentation methods are often developed for highly specific tasks, learning to identify particular objects within predefined contexts. However, the pursuit of versatile segmentation tools has led to the development of interactive segmentation methods. These methods leverage direct user input, such as scribbles (see Figure 1), to guide the segmentation process, offering a tailored approach to various segmentation needs.

This project report delves into the realm of interactive RGB-D image segmentation, enhancing current methodologies by incorporating depth information alongside traditional RGB data. By building upon existing frameworks that utilize spatially varying color distributions, this work [1] proposes two novel extensions: treating depth as an additional color channel and evolving color distributions from a planar to a volumetric variation. Such advancements allow for more precise segmentation in complex scenes where objects of similar color present challenges under difficult lighting conditions. This approach not only addresses the limitations of current segmentation techniques but also showcases the significant role of depth-sensing technology in improving segmentation outcomes.

2 Presentation of the method

2.1 Problem statement and approach via convex relaxation

Let an image be represented by a mapping $I : \Omega \to \mathbb{R}^d$, where $\Omega \subset \mathbb{R}^2$ is the domain of the image and d is the dimensionality of the color space, with $d = 3$ for RGB and $d = 4$ for RGB-D images.
The segmentation process aims to partition Ω into n pairwise disjoint regions $\{\Omega_i\}_{i=1}^n$, each corresponding to a meaningful part of the image. The partitioning can be formulated as an energy minimization problem where the energy functional E is given by:

$$E(\Omega_1, ..., \Omega_n) = \frac{1}{2}\sum_{i=1}^{n} \mathrm{Per}(\Omega_i) + \lambda \sum_{i=1}^{n} \int_{\Omega_i} f_i(x)dx,$$

with $\mathrm{Per}(\Omega_i)$ representing the perimeter of region Ω_i. This is minimized to encourage segmentation boundaries that align with image edges. The perimeter may be computed using either a Euclidean metric or an edge-dependent metric represented by non negative function g often taken as $g(x) = \exp(-\gamma \nabla I(x))$.
Here, $f_i(x)$ symbolizes the appearance model i.e. the data term, and λ is a parameter tuning the regularization term's influence.

To address the issue of nonconvexity in the minimization problem, a convex relaxation technique is employed. This involves the representation of the disjoint regions Ω_i by indicator functions v_i, where $v_i(x) = 1$ if $x \in \Omega_i$ and zero otherwise. Using the total variation (TV) of the indicator functions, the energy functional can be reformulated [6] as:

$$E(v_1, ..., v_n) = \frac{1}{2}\sum_{i=1}^{n} \int_{\Omega} g(x)|Dv_i|dx + \lambda \sum_{i=1}^{n} \int_{\Omega} v_i(x) f_i(x)dx \tag{1}$$

where $|Dv_i|$ denotes the distributional derivative of v_i, relating to the image's edges.
To find the optimal segmentation $(v_1, ..., v_n)$, we minimize E subject to the constraints that $v_i(x) \in \{0, 1\}$ and $\sum_i v_i(x) = 1$ for all $x \in \Omega$.
When carefully chosen, the functions f_i and g imbue the segmentation task with convex properties, notwithstanding the non-convexity introduced by the integer constraints on v_i. The authors propose circumventing this by relaxing these constraints, permitting $v_i(x)$ to take on any value in the interval [0, 1]. This relaxation transforms the problem into a convex one, which can then be resolved with relative ease.

2.2 Choices of the components

The crux of developing a robust segmentation method is to determine the functions f_i that guide the segmentation with high fidelity.
The authors' method invoke the log-likelihood of the estimated probability distribution motivated by maximum a-posteriori probability (MAP) estimate for the computation of f_i:

$$f_i(x) = -\log \hat{\mathcal{P}}(I(x), D(x), x \mid u(x) = i)$$

with $I(x)$ the pixel color and $D(x)$ the pixel depth used to leverage every information available. Thus the author have used gaussian kernels with different standard deviation (i.e. bandswidth) to model the joint probability distribution :

$$\hat{\mathcal{P}}(I(x), D(x), x \,|\, u(x) = i) :$$
$$\frac{1}{m_i} \sum_{j=1}^{m_i} k_{\rho_i}(X - X_{ij}) k_\sigma(I(x) - I(x_{ij})) k_\tau(D(x) - D(x_{ij}))$$

where k_{p_i} is the distance kernel, k_σ is the color kernel, k_τ is the depth kernel, $X = (x, D(x))$ is the 3D position of pixel x and the x_{ij} are the m_i user's scribbles pixels for class i.
The maximum a posteriori estimation is composed of three kernels—distance, color, and depth—to capture the multidimensional similarity between user-specified scribbles and image pixels, aiming to accurately segment an image by considering spatial proximity, color consistency, and depth alignment for enhanced delineation of regions. Let's detail them in a bit more detail.

2.2.1 Distance kernel k_{ρ_i}

The distance kernel within the maximum a posteriori framework weights the influence of each user-provided scribble based on its spatial proximity to a given pixel. This kernel diminishes the impact of scribbles that are farther away from the pixel in question, thereby promoting the segmentation of the pixel to be more influenced by nearby scribbles, which are likely to belong to the same segment or object in the image. Essentially, it enforces spatial contiguity in the segmentation process, under the assumption that pixels closer to a scribble are more likely to share the same label as that scribble.
The bandwidth for the distance kernel, $\rho_i(X)$, is selected based on the planar back-projection using the normalized depth as a third dimension :

$$\rho_i(X) = \alpha \min_{j=1,\dots,m_i} \|X - X_{ij}\|$$

This formula balances the influence of scribbles on pixel X in proportion to their proximity in 3D space. It allows adaptative influence, a variable bandwidth adapts to the density of the scribbles. In regions with sparse scribbles, the bandwidth increases, which allows the scribbles to exert influence over a larger area, ensuring that all parts of the image are affected by the user input. Conversely, in regions with a higher density of scribbles, the bandwidth decreases, sharpening the influence of scribbles and allowing for more detailed segmentation.

2.2.2 Color kernel k_σ

The color kernel evaluates the similarity between the color of each pixel and the scribbles, weighting the segmentation based on color consistency to enhance segment distinction based on visual similarity.

2.2.3 Depth kernel k_τ

The color kernel evaluates the similarity between the depth of each pixel and the scribbles, weighting the segmentation based on depth consistency to enhance segment distinction based on this additional information.

2.3 Active Scribbles

One of the two new ideas proposed in the paper under scrutiny is the the notion of active scribbles. The authors proposed to distinguish for each pixel x active and inactive scribbles.
Active scribbles are determined by analyzing the proximity of user-provided annotations to each pixel. For a pixel x, scribbles within a certain threshold, specifically three times the distance to the closest scribble, are considered active. The rule of 0.8 is applied as a criterion: if less than 80% of the scribbles are active for a pixel, then the algorithm calculates separate probability densities for active and inactive scribbles. This is expressed mathematically as:

$$\hat{\mathcal{P}}(x) = \begin{cases} \hat{\mathcal{P}}_{\text{all scribbles}}(x), & \text{if active scribbles} \geq 80\% \\ 0.8 \cdot \hat{\mathcal{P}}_{\text{active}}(x) + 0.2 \cdot \hat{\mathcal{P}}_{\text{inactive}}(x), & \text{otherwise} \end{cases} \tag{2}$$

Where $\hat{\mathcal{P}}_{\text{active}}(x)$ is the probability density based on active scribbles and $\hat{\mathcal{P}}_{\text{inactive}}(x)$ is based on inactive scribbles. This strategy enables the method to focus on regions with a higher concentration of scribbles—often corresponding to more intricate parts of the image—while allowing a broader influence for sparsely annotated areas, ensuring a nuanced segmentation that caters to the varied complexities within an image.

2.4 Towards a primal-dual algorithm

It's been showed (in [6]) by transforming $g(x)|Dv_i|$ that the relaxed (1) minimization problem can be rewritten as :

$$\operatorname*{argmin}_{v_i \in \mathcal{B}} \sup_{\xi_i \in \mathcal{K}_g} \left\{ \sum_{i=1}^{n} \int_\Omega \lambda \, v_i(x) \, f_i(x) - v_i(x) \text{ div } \xi_i \, dx \right\}$$

With g-dependant set $\mathcal{K}_g$:

$$\mathcal{K}_g = \left\{ \xi \in C_c^1(\Omega, \mathbb{R}^2) \;\middle|\; |\xi(x)| \leq \frac{g(x)}{2}, \; x \in \Omega \right\}.$$

And convex relaxed simplex of bounded variation (BV) functions $\mathcal{B}$:

$$\mathcal{B} = \left\{ v \in BV(\Omega, [0,1])^n \;\middle|\; \sum_{i=1}^{n} v_i = 1 \right\}.$$

In our case, the $\xi_i \in C_c^1(\Omega, \mathbb{R}^2)$, smooth functions with compact support will play the role of dual variables. The method involves iteratively performing a projected gradient descent on the primal variables v_i,

paired with a projected gradient ascent on the dual variables ξ_i (cf Figure 2).
To solve this optimization problem, a primal-dual algorithm is used [3, 7, 6]:

Algorithm 1 Primal-Dual Algorithm for proposed segmentation method

1: Initialize ξ^0, v^0
2: **for** $t = 0, 1, 2, \ldots$ until convergence **do**
3: $\quad \xi_i^{t+1} \leftarrow \Pi_{\mathcal{K}_g}\left(\xi_i^t + \tau_d \nabla \bar{v}_i^t\right)$
4: $\quad v_i^{t+1} \leftarrow \Pi_{\mathcal{B}}\left(v_i^t + \tau_p(\ \mathrm{div}\ \xi_i^{t+1} - f_i)\right)$
5: $\quad \bar{v}_i^{t+1} \leftarrow v_i^{t+1} + (v_i^{t+1} - v_i^t)$
6: **end for**

With Π denoting the projection operator and the additional variable $\bar{v}_i$ is introduced as part of an acceleration technique known as over-relaxation, which can speed up the convergence of iterative algorithms like gradient descent. It works by taking a linear combination of the current and previous iterations to overshoot the current position, in the hope of getting closer to the minimum more quickly.
The difficulty of the algorithm lies thus in coding the $\Pi_{\mathcal{B}}$ operator. We find in the literature [2] how to code the projection onto a simplex.
Let's lastly notice that such an implementation is easily parallelizable.

3 Experimental results

In this section, we will evaluate the relevance of the proposed method. Our assessment will cover both quantitative metrics and qualitative observations, comparing the full method against variations that exclude the active scribbles concept and depth information.

3.1 Presentation of the dataset

The dataset analyzed was meticulously composed by the study's authors to align with the specific requirements of interactive RGB-D segmentation. Due to the absence of RGB-D data in prevalent benchmarks and the unsuitability of datasets like NYUv2, the authors opted for the Object Segmentation Database (OSD). The original OSD was enhanced with 16 additional challenging images captured using an RGB-D sensor, overcoming the original's limitations in complexity. The resulting set of 28 images, complete with carefully generated ground truth labels, offers a rigorous testing ground for the segmentation methods under review (see the dataset).
Figure 1 shows an extract from this dataset.

3.2 Algorithm implementation

Using the Primal-Dual Algorithm 1 we implement the segmentation with the parameters values given in the article.

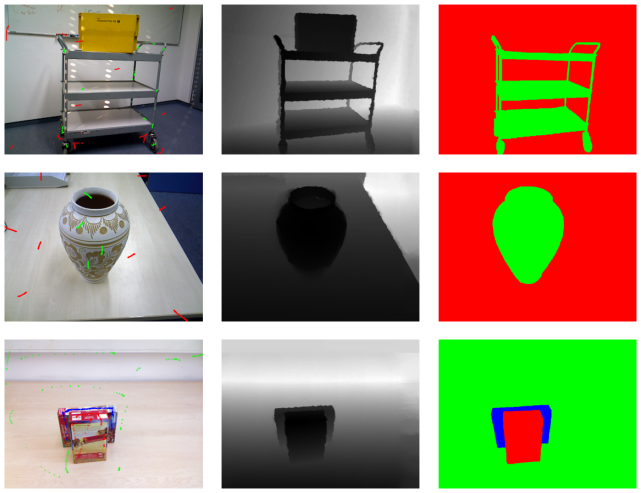

Figure 1: Example from the dataset. From left to right a) Color images with user's scribbles b) Depth information c) Ground truth labels

We continue the iterations until the difference between the energies of the primal and dual variables is inferior to a threshold as we can see in figure 2 :

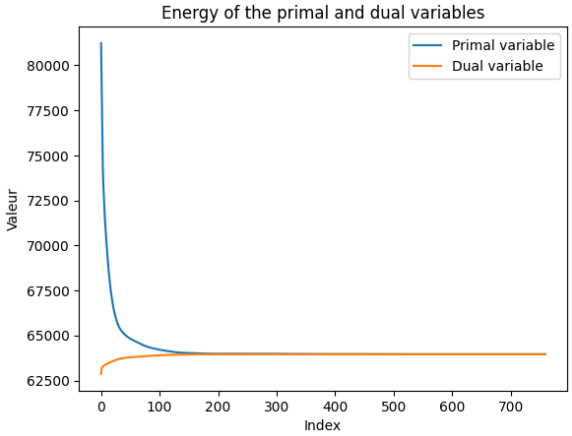

Figure 2: Evolution of the energies over the iterations

3.3 Comparison of performances

3.3.1 Qualitative performances

To assess the method's effectiveness, we first conduct a qualitative analysis of the segmentation results. This includes comparing outcomes obtained using the complete method, the method excluding the active scribbles feature, and the method without incorporating depth information across various examples that we see on figure 3 :

Visual inspection reveals that the full proposed method delivers superior results. Segmentation performed without utilizing the active scribbles (AS) feature produces acceptable outcomes but faces difficulties with intricate structures, as observed with the coat hanger and the trolley. This observation aligns

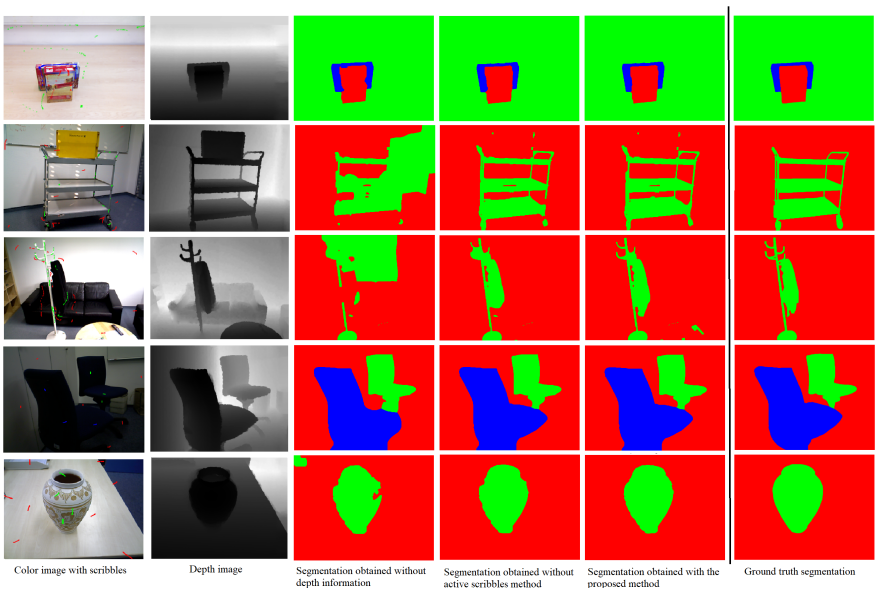

Figure 3: Visualisation of segmentation results

with the underlying active scribbles' principle of concentrating on areas with dense scribbles to navigate complex regions.
Conversely, segmentations lacking depth information fall short of optimal, showing that relying solely on color and distance kernels can yield significant inaccuracies, particularly with complex objects. They would require more scribbles to achieve satisfying results.

3.3.2 Quantitative performances

To draw broader and more reliable conclusions regarding the performances of the different segmentations, we compare for the different methods the dice similarity coefficient suggested in [8] :

$$DSC(S) = \frac{1}{n}\sum_{i=1}^{n}\frac{2|GT_i \cap S_i|}{|GT_i| + |S_i|}$$

This metric estimates the extent of overlap among the n regions within the segmentation. By aggregating across each region, it offers robustness to segmentations that may be unbalanced or contain very small objects, it is expressed in %.

Method	Mean DSC	Median DSC
Proposed method	92.65 ± 3.60	96.46
Without AS	92.03 ± 2.80	96.25
Without depth	85.73 ± 13.75	93.28

Table 1: Comparison of the methods' DSC $\pm$ std

The quantitative findings corroborate the qualitative assessments: the proposed methods demonstrate superior performance. The integration of depth data evidently enhances segmentation quality, as it introduces additional informative cues. Moreover, the active scribbles (AS) concept improves segmentation, albeit the improvement is marginally noticeable.

It's worth noting that the discrepancy between methods lacking depth data is less pronounced when examining median values rather than means. This is attributed to the fact that segmentation is satisfactory for certain items, like cereal boxes or armchairs as seen in Figure 3, but less effective for complex structures or objects that closely resemble their background with sparse scribbles. Consequently, the median is less influenced by these extremes.

3.4 Remarks and critics of the method

The presented results might be biased since the scribbles—crucial for segmentation—came from the researchers themselves. It's plausible that additional scribbles were applied to areas with initially poor results, improving the final output. Although the outcomes are significant, it's vital to acknowledge the extensive input required, including RGB-D images and carefully placed scribbles. Such intervention prompts questions about how the method would perform with fewer inputs.

Finally, in addressing the relaxed version of the problem, we encounter instances where the pixel values $v_i(x)$ may not be strictly binary. Our empirical findings indicate that for 96% of the pixels $x \in \Omega$ across all classes i, $(i = 1, \ldots, n)$, the relaxed solutions $v_i(x)$ were either below 0.001 or above 0.999. For the few other pixel, we threshold-ed to get a binary solution. The very low amount of pixels with values not close to 0 or 1 suggests that the segmentation we obtained approximates the true binary solution of problem 1 very closely.

Since the publishing of this paper, several method employing the same RGB-D and interactive data framework have emerged, featuring graph cut algorithms [4, 5] pretending to yield to better results.

References

[1] Julia Diebold, Nikolaus Demmel, Caner Hazırbas, Michael Moeller, and Daniel Cremers. Interactive multi-label segmentation of rgb-d images. Technical University of Munich, Germany, 2015.

[2] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l1-ball for learning in high dimensions. Stanford, Toyota, Google, 2008.

[3] Ernie Esser, Xiaoqun Zhang, and Tony Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. SIIMS, 2010.

[4] Jie Feng, Brian Price, Scott Cohen, and Shih-Fu Chang. Interactive segmentation on rgbd images via cue selection. Columbia University, Adobe Research, 2016.

[5] Ling Ge, Ran Ju, Tongwei Ren, and Gangshan Wu. Interactive rgb-d image segmentation using hierarchical graph cut and geodesic distance. In *Advances in Multimedia Information Processing*, pages 114–124. LNISA, 2015.

[6] Claudia Nieuwenhuis and Daniel Cremers. Spatially varying color distributions for interactive multi-label segmentation. Technical University of Munich, Germany, 2012.

[7] T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. ICCV, 2011.

[8] J. Santner, T. Pock, and H. Bischof. Interactive multi-label segmentation. ACCV, 2010.