Interactive Energy Minimizing Segmentation Frameworks

Von der Fakultät für Elektrotechnik und Informatik

der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des akademischen Grades

Doktor-Ingenieur

(abgekürzt: Dr.-Ing.)

genehmigte

Dissertation

von

Dipl.-Math. Björn Scheuermann

geboren am 16. Februar 1983 in Gehrden.

2014

Referent:	Prof. Dr.-Ing. B. Rosenhahn
Korreferent:	Prof. Dr.-Ing. V. Roth
Vorsitzender:	Prof. Dr.-Ing. J. Ostermann
Tag der Promotion:	13.08.2014

Acknowledgement

The time working at the Institut für Informationsverarbeitung (TNT), Leibniz Universität Hannover, to earn a Doctor of Engineering degree has been one of the greatest times in my life. There is so much I learned on image segmentation, computer vision, doing good research and so on. Working towards a Doctor degree means lots of hours of work, setbacks and sleepless nights. But it also means meeting great people, traveling to conferences, moments of excitement and having a lot of fun with my colleagues after hours. All this would not have been possible without the help, guidance, support and criticism of many people I have gotten to know.

I owe very much to my supervisor Prof. Dr.-Ing. Bodo Rosenhahn for giving me the opportunity to work in such an awesome group. I like to thank him for guiding, inspiring and challenging me. His vision and his leadership largely contributed to making this thesis a success.

I also like to thank Prof. Dr.-Ing. Volker Roth for being my second supervisor and Prof. Dr.-Ing. Jörn Ostermann, the chair of my defense committee.

A special thank goes to all my colleagues at the TNT. You all contribute to this great and fun place to work! I would particularly like to mention my officemate Kai Cordes, my best man Florian Baumann, Stephan Preihs, Matthias Reso and Hanno Ackerman. I 'm glad to have found you as good friends.

A Special thanks also goes to Matthias Schuh and the secretaries, Mrs. Brodersen, Mrs. Bank and Mrs. Jaspers-Göring. I appreciate for all your commitment and valuable assistance.

Finally, a very special thanks goes to my family. To my parents, Anita and Uwe Scheuermann for enabling me to study and supporting me in everything I do. And to my wife: I am deeply grateful for the love of my wife Meret and her continuously backing me.

This dissertation is dedicated to my loved grandparents.

Contents

1 Introduction **1**

2 Segmentation Using Energy Minimization Techniques **8**
2.1 Image Segmentation 9
2.2 Segmentation as Energy Minimization 13
2.3 Interactive Segmentation 15
2.4 Benchmarks 18

3 Background **21**
3.1 Segmentation Using a Variational Framework 22
3.1.1 Chan-Vese Energy Functional 22
3.1.2 Probabilistic View 30
3.2 Segmentation by Discrete Energy Minimization 31
3.2.1 Discrete Energy Model 31
3.2.2 Probabilistic View 34
3.3 Feature Fusion using Dempster's Theory of Evidence 37
3.3.1 Dempster's Theory of Evidence 38
3.3.2 Relation To Classical Probability Theory 41
3.3.3 Examples 43

4 Dempster's Theory of Evidence for Variational Image Segmentation **48**
4.1 Energy Function including Dempster's Theory of Evidence 49
4.1.1 Defining Appropriate Mass Functions 53
4.1.2 Experimental Results 54
4.2 Interactive Variational Image Segmentation 59
4.2.1 Integration of User Constraints 60
4.2.2 Experimental Results 65
4.3 Discussion 66

5 Discrete Energy Minimization with Dempster's Theory of Evidence **69**
5.1 SlimCuts: High Resolution Image Segmentation 71
5.1.1 Constructing *Slim Graph* 74
5.1.2 *Slim Graphs* for Simplified User Interaction 79

5.1.3 Experiments . . . 81
5.2 Efficient Pixel Grouping with Dempster's Theory of Evidence . . . 85
5.2.1 Variable Grouping based on Dempster's Theory of Evidence . 87
5.2.2 Experiments . . . 92
5.3 Discrete Energy Function including Dempster's Theory of Evidence . 97
5.3.1 Feature Fusion using Dempster's Theory of Evidence . . . 99
5.3.2 Experimental Results . . . 103
5.4 Discussion . . . 105

6 Conclusion **109**
6.1 Summary . . . 110
6.2 Contributions . . . 111
6.3 Possible Directions for Future Work . . . 112

A Appendix **114**
A.1 Building the Euler Lagrange Equation . . . 115
A.2 Application - N-View Human Silhouette Segmentation . . . 116

B List of Publications **124**
B.1 Refereed Publications . . . 125
B.2 Patents . . . 127

Bibliography **128**

Abbreviations

2D	two dimensional
3D	three dimensional
BG	background region
BK-algorithm	Boykov-Kolmogorov augmenting paths algorithm
bpa	basic probability assignment (mass function)
CRF	conditional random field
DS	Dempster-Shafer theory of evidence
FG	foreground region
GC	graph cut
GHz	gigahertz
GMM	Gaussian mixture model
GPU	graphics processing unit
KL	Kullback-Leibler divergence
LS	level set
MAP	maximum a posteriori
MRF	Markov random field
MST	minimum-weight spanning tree
PDE	partial differential equation
pep, pixel	picture element, image point
RGB-D	rgb-color and depth image pair
ToF	Time-of-Flight

Symbols and Notation

$Bel(A)$	belief function
$\mathcal{B}$	background seeds
$c(e)$	capacity function
w_{ij}	weight/capacity of edge $e = (i,j)$
p_i	conditional probability $p(I(x) \mid \mathcal{L}_x = i)$
Σ	covariance matrix
C	boundary curve in a variational framework
$\mathcal{C}$	cut of a graph
D	depth image
$\delta(z)$	dirac measure
dist(i,j)	euclidean distance between pixels i and j
div	divergence operator
$\otimes$	Dempster's rule of combination
$\emptyset$	empty set
$E(\mathcal{L})$	energy function
$E(\varphi)$	energy functional
$\langle \cdot \rangle$	expectation over the image
$I(x)$	feature vector at location x
f	flow in a network graph
Ψ	frame of discernment or hypotheses set
K_σ	gaussian kernel with standard deviation σ
G	graph
$H(z)$	heaviside function
A	hypothesis, $A \in \wp(\Psi)$
I	image
I_j	feature channel j, if $I = (I_1, \ldots I_k)$
Ω	image domain
$\mathcal{L}$	image labeling
Ω_i	image region i

$[\cdot]$	indicator functionl
δ^{KL}	Kullback-Leibler divergence
$\delta^{sym.KL}$	symmetric Kullback-Leibler divergence
$\mathcal{L}_x$	label of pixel x
$\varphi(x)$	level set function
$\log(p(x))$	log-likelihood
$m(A)$	mass function or basic probability assignment
μ	mean or expected value
∇	nabla or del operator
$\mathcal{N}$	neighborhood system of a pixel
$\mathcal{O}$	object seeds
$\tau_{i,j}$	pairewise term/potential
∂	partial derivative
x	image pixel
$Pl(A)$	plausibility function
$\wp(\Psi)$	power set
2^{Ψ}	power set
$\propto$	proportional relation
$\mathcal{E}_G$	set of edges of graph G
$\mathcal{E}$	set of edges between image pixels
$\mathcal{V}$	set of image pixels
$\mathcal{U}$	set of user seeds
$\mathcal{V}_G$	set of vertices of graph G
sign	sign function
$\tilde{G}$	Slim Graph
σ	standard deviation
J	structure tensor
Δt	time step used to solve a PDE
τ_i	unary term/potential
$L(x)$	user defined labeling
G'	variable grouping of graph G
γ	weighting factor (graph cuts)
λ	weighting factor (level sets)
ν	weighting factor (level sets)

Abstract

Image segmentation is the process of partitioning an image into at least two regions. It is one of the fundamental research areas in computer vision and has been widely studied in the last years. A popular application in movie post production build upon image segmentation is the integration of virtual objects. Because of many aspects in real-world scenarios image segmentation is a very challenging task.

The segmentation problem can be formulated as an energy minimization problem. Active contours or level set representations are an efficient way to find minimas of a continuous energy functionals. In the discrete domain the problem can be formulated using probabilistic models like Markov or conditional random fields. The maximum a posteriori solution of such a model corresponds to the discrete optimization of an appropriate energy function that can be solved using graph cuts.

Usually, all those methods use Bayes' theorem to combine different features, arising from color distribution, texture and scale information or additional sensors like depth information. Therefore, the features are assumed to be statistically independent and the joint probability is given by the product rule. Due to inaccurate and incomplete or conflictive features this fusion can lead to unsatisfactory segmentation results.

The first part of this dissertation addresses the problem of feature combination by proposing to use Dempster's theory of evidence to fuse the information. This theory of evidence offers an elegant and intuitive way to fuse information from different feature channels and offers an alternative to Bayes' theory. In contrast to Bayes' theory, Dempster's theory allows to explicitly model inaccuracy and uncertainty of features at the same time. Thus it provides a way to incorporate the reliability of a feature. In this dissertation the classical energy minimizing frameworks are extended by means of this theory. Experiments on (interactive) image and video segmentation will demonstrate the properties and advantages of using Dempster's theory of evidence for image segmentation.

The second problem addressed in this thesis relates to the efficient minimization of the discrete energy function using graph cuts. It has been shown, that these methods are impracticable for high-resolution images or video sequences due to their running time and memory requirements. Two methods are presented to reduce the problem size. The first method reduces the underlying graph while maintaining the maximum flow property. The second method groups similar variables using the terms of the energy function itself and Dempster's theory of evidence. Experiments will show that these methods are able to drastically reduce the problem size and thus the runtime of the graph cut algorithm itself.

Keywords: interactive image segmentation, video segmentation, energy minimizing methods, Dempster's theory of evidence, feature fusion

Kurzfassung

Die Bildsegmentierung beschreibt den Prozess der Unterteilung eines Bildes in mindestens zwei Bereiche. Im Gebiet Computer Vision zählt die Bildsegmentierung zu den fundamentalen Problemen. Eine beliebte Anwendung findet sich in der Film-Postproduktion, wo virtuelle Objekte in Filmszenen integriert werden. Aufgrund unterschiedlicher Aspekte in realen Szenarien ist die Bildsegmentierung eine sehr anspruchsvolle Aufgabe.

Die Segmentierung kann mathematisch als Energieminimierungsproblem formuliert werden. Level-Set-Methoden bieten eine effiziente Möglichkeit, solch eine kontinuierliche Energiefunktion zu minimieren. Im diskreten Fall kann das Problem mit probabilistischen Modellen wie Markov oder Conditional Random Fields formuliert werden. Die Maximum-a-posteriori Lösung eines solchen Modells entspricht der diskreten Optimierung einer geeigneten Energiefunktion, die durch Berechnung des minimalen Schnittes eines Graphen bestimmt werden kann.

In der Regel kombinieren alle diese Methoden mit Hilfe des Theorem von Bayes verschiedene Merkmale, z.B. Farbverteilungen, Textur- und Skalen-Informationen oder Daten von Tiefensensoren. Die Merkmale werden als statistisch unabhängig angenommen und die Gesamtwahrscheinlichkeit ergibt sich dann aus der Produktregel. Durch unvollständige oder ungenaue Merkmale kann diese Art der Fusion zu unbefriedigenden Segmentierungen führen.

Der erste Teil der Arbeit befasst sich mit der Fusion verschiedener Merkmale. Es wird vorgeschlagen Dempster's Theorie der Evidenzen zur Fusion der Informationen zu nutzen. Im Gegensatz zur Bayeschen Theorie ermöglicht Dempster's Theorie explizit die zeitgleiche Modellierung von Ungenauigkeit und Unsicherheit. So bietet sie eine Möglichkeit, die Glaubwürdigkeit eines Merkmals miteinzubeziehen. In dieser Dissertation werden die klassischen energieminimierenden Verfahren durch diese Theorie erweitert. Experimente demonstrieren die Eigenschaften und Vorteile von Dempster's Theorie der Evidenzen in der Bildsegmentierung.

Im zweiten Teil wird die diskrete Minimierung durch minimale Schnitte in einem Graphen betrachtet. Durch hohe Laufzeiten und Speicheranforderungen sind diese Verfahren nicht für hochauflösende Bilder oder Videosequenzen geeignet. Im Rahmen dieser Arbeit werden zwei Verfahren vorgestellt, die die Problemgröße reduzieren. Das erste Verfahren verkleinert den zugrunde liegenden Graphen ohne den maximalen Fluss zu verändern. Das zweite Verfahren gruppiert ähnliche Variablen über die Terme der Energiefunktion selbst und Dempster's Theorie der Evidenzen. Experimente zeigen, dass beide Verfahren die Problemgröße und damit die Laufzeit des Graph-Cut Algorithmus drastisch reduzieren können.

Stichworte: interaktive Bildsegmentierung, Videosegmentierung, energieminimierende Verfahren, Dempster's Evidenztheorie, Merkmalsfusion

Chapter 1

Introduction

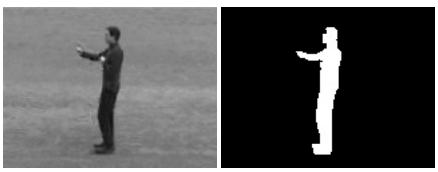
(a) (b) (c)

Figure 1.1: Example of an interactive binary segmentation: Given an image, the user roughly marks foreground (blue) and background (red) regions with scribbles (a). An interactive segmentation system labels each pixel as either foreground (white) or background (black) (b); in (c) the two regions are visualized using the boundary between.

Motivation

This thesis deals with the problem of feature fusion for interactive binary image segmentation. Given some prior information on the object and the background of an image (usually rough user scribbles), the goal of an interactive segmentation system is to label each pixel in the image as either foreground or background. The result is a dense labeling of the image, see Figure 1.1.

A crucial step of every segmentation system is the choice of features used to solve the problem. Usually, features like gray values, color and texture are available. In some scenarios additional information like motion (in video sequences) or depth (e.g. from a Time-of-Flight camera) is existent. Considering the example in Figure 1.2 color and depth information, so-called RGB-D image pairs, are given. Using only the color features to segment the person yields an insufficient segmentation result in the first example, since foreground and background color features are indistinguishable. Utilizing the available depth information instead, yields a visually good segmentation, since the depth histograms of foreground and background are well separated. Choosing another RGB-D image pair from the same video sequence (2nd example in Figure 1.2), the situation changes considerably. The depth histograms of foreground and background are overlapping, resulting in an insufficient segmentation of the person. On the contrary, color information is much more significant to



segment the person and yields a good segmentation.

This example clarifies, that the choice of features depends on the application and the scenario. In some situations one feature is more discriminative and in the next situation another feature is better suited. Thus, using all the available information can help to improve the overall performance of a segmentation system. State-of-the-art methods, e.g. energy minimizing approaches, usually assume statistically independent features and use the traditional Bayesian framework to fuse the available information. Due to the product law, this approach tends to favor features with low support to a region. Because of this tendency inaccurate, incomplete or conflictive features with low support have an immoderate influence on the segmentation result.

A generalization of the Bayesian theory is given by Dempster's theory of evidence. It allows to explicitly model inaccuracy and uncertainty information at the same time and to describe conflicts in the information fusion process. In this thesis, Dempster's theory of evidence is integrated into an energy minimizing segmentation systems to improve the segmentation quality by a more intuitive and elegant feature fusion.

Besides the feature selection and fusion, the runtime of an interactive segmentation system is another important aspect. Since the user has the ability to refine a segmentation result by providing additional information, the lag between user interaction and computed segmentation should be as small as possible. In case of a discrete energy minimizing segmentation framework, this thesis proposes two novel algorithms that reduce the complexity of the graph and thus the runtime. Therefore, pixels are grouped based on a defined similarity. The first algorithm proposed, simplifies the underlying graph without changing the maximum flow property. The second algorithm uses the terms of the energy function to simplify the graph and approximate the original segmentation.

Contributions

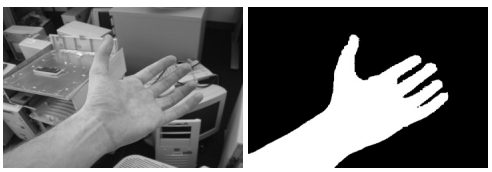
The main contributions of this thesis can be divided into two parts:

- Application of Dempster's theory of evidence for image segmentation
- Graph simplification and variable grouping for image segmentation

In the following, both types of contributions are shortly reviewed.

Application of Dempster's theory of evidence for image segmentation:

Continuous and discrete energy minimizing approaches for the problem of binary image segmentation have become very popular in the last decades. In this thesis, those frameworks are extended by means of Dempster's theory of evidence to fuse information arising from different feature channels, e.g. different sensors. Dempster's



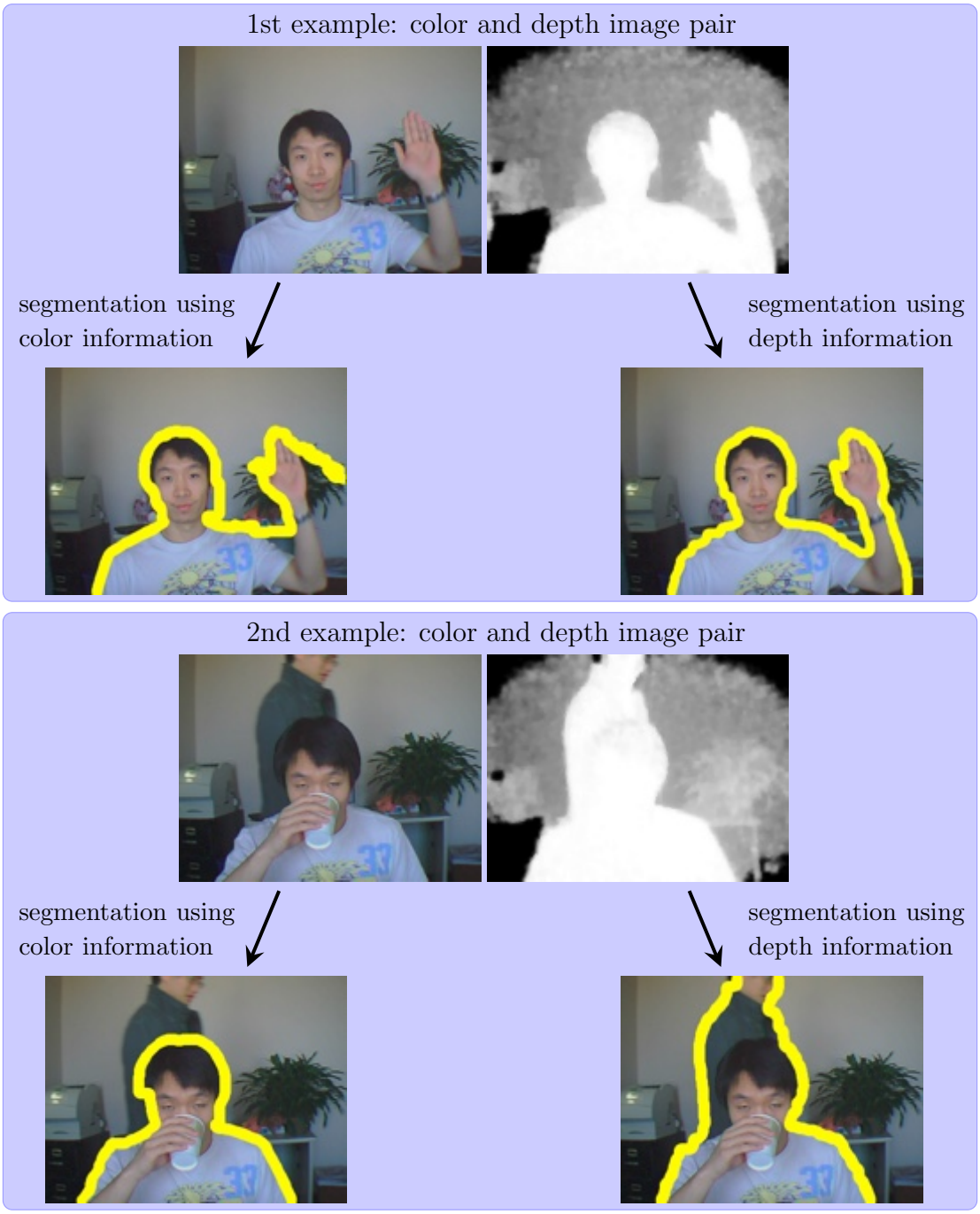


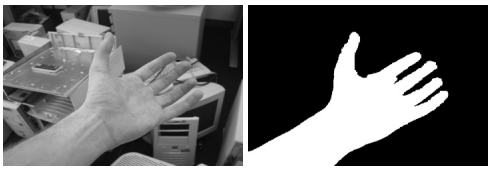
Figure 1.2: Example segmentation results using different features. In the first row of each example a color and depth (RGB-D) image pair is given. The second row shows the segmentation results using either color or depth information only. Obviously, in the first example the depth information is more reliable and produces a visually better segmentation result. The second image pair shows an example where the color information is more reliable. Images are taken from the ToFCut dataset [WZYZ10].

theory of evidence can be described as a generalization of the Bayesian approach. It allows to model inaccuracy and uncertainty of the information and thus provides an elegant way to fuse information and offers an alternative to the classical probabilistic Bayesian model. A prerequisite for Dempster's theory of evidence are so-called mass functions. Appropriate mass functions modeling the available information are proposed and integrated in both, the continuous segmentation framework and the discrete segmentation framework. Furthermore, Dempster's theory of evidence is used to extend the continuous level set approach by means of user interaction. Thus, a complete user interactive segmentation framework is developed and the experimental results show, that this framework outperforms similar approaches in terms of segmentation accuracy and user effort. To summarize the contributions:

- appropriate mass functions modeling the information are defined
- integration of Dempster's theory of evidence in the continuous level set framework
- extension of the level set framework by means of user interaction based on Dempster's theory of evidence
- an example application, combining the level set framework with other segmentation systems is given
- integration of Dempster's theory of evidence in the discrete graph cut framework

Graph simplification and variable grouping for image segmentation:

Discrete energy minimizing approaches based on graph cuts are a widely used technique to solve binary segmentation problems. However, the complexity of these algorithms is a crucial drawback since each pixel in an image corresponds to an unknown variable of the energy function. Roughly speaking, that means that graph cut approaches are not able to efficiently segment high-resolution images or video sequences due to their running time and memory requirements. That is why it is important to reduce the complexity of the algorithm. This thesis proposes two algorithms reducing the complexity by simplifying the graph or grouping the variables of similar pixels. The first approach, the so-called *SlimCuts*, proposes to contract *simple edges*. Those edges are efficient to find and contracting them reduces the graph and thus the problem size. It is proven that the maximum flow is preserved by contracting *simple edges*, which means that the segmentation result is not changed. Since the amount of graph reduction is limited by the number of *simple edges* a second algorithm for variable grouping is proposed. In this algorithm the similarity of neighboring pixels (or neighboring groups of pixels) is measured using Dempster's theory of evidence. Therefore the unary and pairwise potentials of the energy functions are interpreted as information on the similarity and appropriate mass functions are defined and fused within the framework of Dempster's theory of evidence. The experiments show that the amount of reduction is drastic, while the changes in the



segmentation result are negligible. To summarize the contributions:

- *SlimCuts*: a graph simplification method that maintains the maximum flow property
- a variable grouping algorithm based on Dempster's theory of evidence

Structure of the thesis

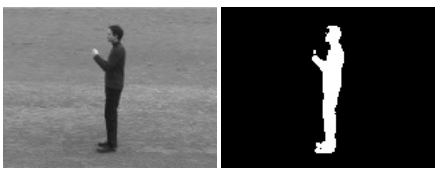
The outline of this thesis is as follows.

Chapter 2: In Chapter 2, the basic terms of (interactive) image segmentation are introduced and reviewed. The problem is mathematically defined and the general components of a segmentation system are given. Different energy minimizing approaches are stated and categorized. Furthermore, the workflow of an interactive, energy minimizing image segmentation system is summarized. The chapter concludes with a brief review of typical benchmarks and performance measures used to evaluate segmentation frameworks.

Chapter 3: In this chapter, the used energy minimizing segmentation frameworks, namely level sets and graph cuts are described and reviewed. The level set framework minimizes an energy function in the continuous domain by propagating a curve in normal direction. In contrast to the level set framework, the graph cut framework minimizes an objective function in the discrete domain. Therefore, the energy function is represented by a network graph and the minimum cut of this graph yields a minimum of the objective function. The connection of both frameworks is clarified by a probabilistic interpretation. Finally, the idea of fusing available features for image segmentation using Dempster's theory of evidence is explained. This theory is reviewed and the relation to classical probability theory is given.

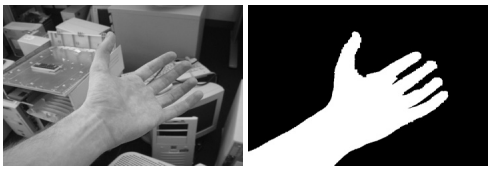
Chapter 4: The level set framework described in the previous chapter is extended by means of Dempster's theory of evidence. Therefore, appropriate mass functions are defined and the joint mass is included in the energy function. The mass functions are based on the image likelihoods similar to the classical approach. In contrast to the classical Bayesian approach, Dempster's theory of evidence offers a sound and intuitive way to model inaccuracy and uncertainty of a feature. Based on this novel framework, user constraints are integrated to develop an interactive, level set based image segmentation system. Experiments on real and synthetic images demonstrate the properties and advantages of the proposed frameworks. The chapter concludes with a review and discussion of the proposed variational frameworks. Previous versions of this chapter appeared in [SR10] and [SR11a].

Chapter 5: In Chapter 5, the discrete segmentation framework graph cut is extended by means of Dempster's theory of evidence to fuse color and depth information. Beforehand, the chapter concentrates on reducing the complexity of the



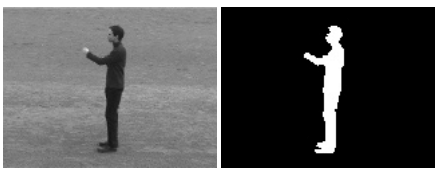
approach by simplifying the underlying network graph. Simplifying the graph means a reduction of unknown variables and thus a more efficient segmentation framework. First, the so-called *SlimCut* approach is introduced. This approach contracts *simple edges* so that the maximum flow property is maintained, which means that the segmentation result does not change. In a second approach Dempster's theory of evidence is used to group similar pixels to one variable. Thus, the graph can be simplified, while the segmentation results stay comparable. The experimental results of both approaches show that the number of variables can be reduced dramatically. In case of the variable grouping framework, the results show that the changes in the segmentation results are negligible. Parts of this chapter appeared earlier in [SR11b, SSR12, SGR13].

Chapter 6: The last and concluding chapter of this dissertation summarizes the whole work and lists the main contributions. At the end of the chapter some promising future developments are discussed.



Chapter 2

Segmentation Using Energy Minimization Techniques



In this chapter, the terms segmentation and interactive segmentation (as they have been used in this work) are defined and related work is shortly reviewed. At the end of the chapter typical benchmarks, that are used to analyze the quality of a segmentation algorithm, are summarized.

2.1 Image Segmentation

Image segmentation is the process (or the result of the process) of partitioning an image into at least two regions. More precisely, the process of assigning a label to every pixel in an image is called image segmentation. Thus, image segmentation is a special kind of classification or labeling problem. Pixels having the same label share characteristics with respect to a given uniformity criterion. Therefore, similar pixels are grouped based on this criterion. Typical uniformity criteria for image segmentation are pixel-intensities, color, texture, depth or motion. In computer vision, the problem of segmenting an image has been studied for decades and now it is one of the most widely studied problems [Sze10]. Formally, the segmentation problem is defined such that the image domain Ω is partitioned into K (disjoint) regions Ω_i:

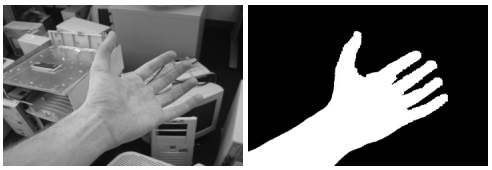
$$\Omega = \bigcup_{i=1}^{K} \Omega_i \,, \qquad (\text{and } \Omega_i \cap \Omega_j = \emptyset \; \forall \, i \neq j) \tag{2.1}$$

In this thesis, unless otherwise stated, the segmentation problem is restricted to the special case $K = 2$ which is referred to as binary segmentation. The two subsets $\Omega_{1,2}$ are denoted foreground (FG) and background (BG). An example of visualizing a binary segmentation result is given in Figure 2.1. The case $K > 2$ is denoted as multi-label segmentation.

The general components of a segmentation system are visualized in Figure 2.2. After defining the segmentation problem (i.e. binary or multi-label), the application needs to be specified. This can either be task dependent, object dependent, or image dependent. The next decisions to be made are:

- which features are available and useful to solve the problem
- how these features can be statistically modeled to fit the requirements

Features used in this thesis are gray values, color, texture and depth. Those histogram features can be modeled using Parzen estimates or Gaussian mixture models. Having different features available, the feature fusion is a crucial step. This thesis proposes to use Dempster's theory of evidence to fuse different features. Since this theory allows to model inaccuracy and uncertainty explicitly, the feature fusion becomes more intuitive compared to the feature fusion in classical probability theory. After all these requirements are determined, the segmentation concept that will solve the problem needs to be chosen. For Instance, this could be a simple thresholding



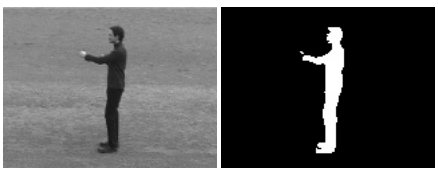
(a) image (b) segmentation mask (c) segmentation boundary

Figure 2.1: Visualizing segmentation results. Given the input image in (a), the image domain is partitioned into foreground (FG) and background (BG). The segmentation result is typically illustrated with a FG/BG mask (b) or by the boundary separating FG and BG (c).

or an energy minimizing approach using a variational or a discrete formulation. Due to their capabilities, in this thesis, energy minimizing approaches are used to solve the segmentation problem.

From a user point of view, segmentation systems can be categorized into supervised, interactive (semi-supervised), and unsupervised systems [Sze10]. Unsupervised segmentation systems segment the image according to some priors that hold for every image. These systems run fully automatic and are typically designed for a special task, e.g. segmentation of anatomical structures in medical images [PFK+05, CSFB08]. Hence, unsupervised segmentation systems are useful to extract similar regions/objects from a large database of images. In supervised segmentation systems the user has to manually assign a label to each pixel of an image. Even though, this is very time consuming, for some applications it is necessary that the user has the complete control over the segmentation result.

This thesis focuses on interactive segmentation systems utilizing energy minimizing approaches. Interactive segmentation systems try to combine the advantages of the two other systems. Typically, some image regions are roughly marked by the user and utilized to learn a prior, e.g. the appearance of the marked regions. This prior is then used to segment the remaining image regions [BJ01, CFRA07].



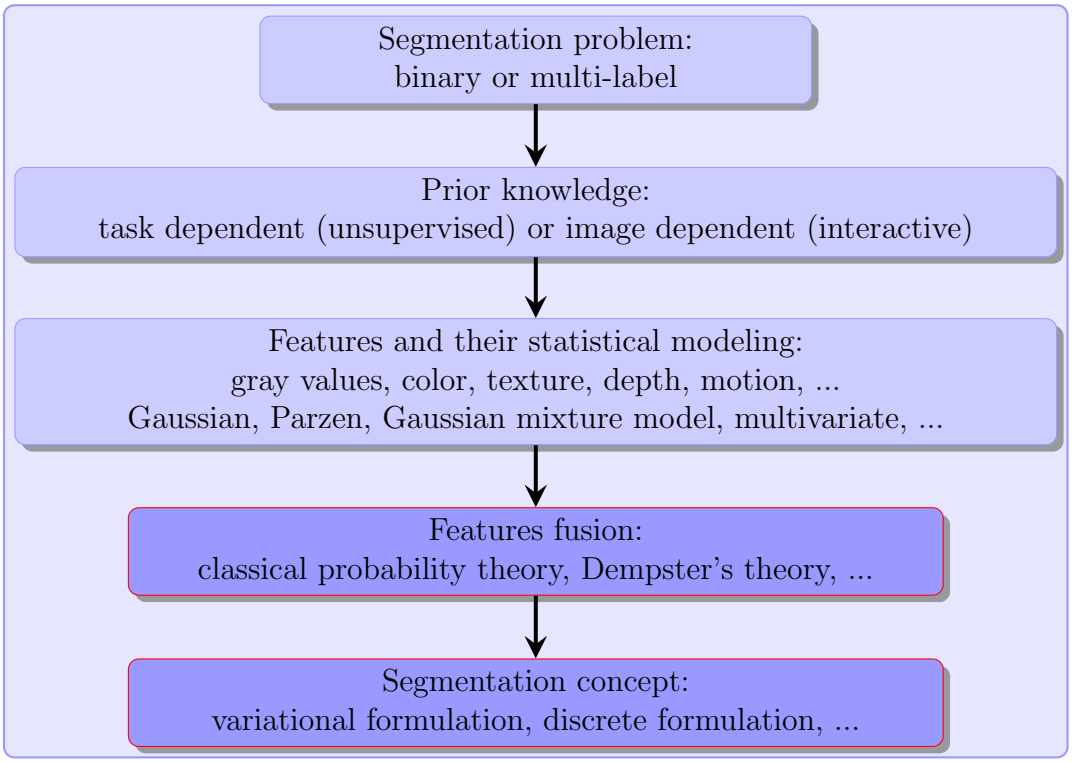
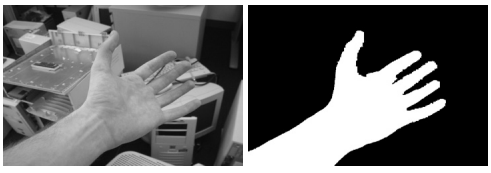


Figure 2.2: General components of a segmentation system including problem and feature definition, feature modeling as well as concepts to fuse features and algorithm to solve the segmentation problem. The contributions of this thesis can be arranged in the feature fusion and segmentation concept stage.

Well known algorithms for unsupervised image segmentation include thresholding [SS04], region merging [Sze10], mean-shift [CM02], k-means clustering [Sze10], and many more. Even though the focus of this thesis is feature fusion for interactive segmentation approaches, a short overview of basic concepts and algorithms follows. This is motivated by the fact that there exist many similarities and the methods share some common concepts.

The most simple methods are pixel based segmentation schemes. The decision whether a pixel belongs to a specific region is taken separately for every pixel in the image domain. Well-established is the segmentation by thresholding the image based on some uniformity criterion [MM03, SS04]. Incorporating spatial information is one possibility to enhance segmentation performance, even if those methods in general work on single pixels.

An elementary assumption for image segmentation is that regions in an image contain pixels that are similar with respect to some similarity criterion. Thus, it can



also be interpreted such that there exists a dissimilarity at the edges of two adjacent image regions. This dissimilarity can be detected using image derivatives that build the basis for edge-based segmentation schemes. A popular method finding those edges is the Canny edge detector proposed by Canny [Can86]. Sonka et al. [SHB07] give an extensive overview of segmentation schemes, that use such edge detectors to find connected contours. An example segmentation scheme, based on those edge detectors, is the Hough transformation and its extensions [HP60, IK88].

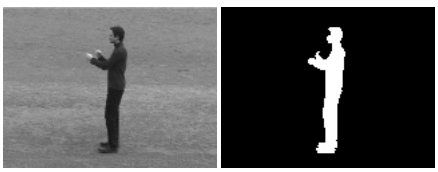
In contrast to edge-based approaches, region-based segmentation schemes try to find and exploit the similarity criterion directly. Well known examples are region merging, the watershed approach [VS91, Beu91] and the split-and-merge approach [HP74].

Some more advanced segmentation schemes use a combination of edge-based and region-based techniques, more complex and higher-dimensional feature spaces or additional shape-constraints, e.g. such that resulting regions are spatially connected and somehow compact. The mean-shift segmentation approach, proposed by Comaniciu and Meer [CM02], is such an approach. It uses a feature space composed of color and spatial information to cluster the image into similar regions.

Shape based segmentation approaches use additional a priori information about the rough shape of the object that needs to be segmented [HP60, PFK$^+$05]. This information restricts the segmentation result to be close to a given shape. Typically, shape constraints are combined with other methods to improve segmentation results [RP02, PRR02, CSS03, CZ05, CSB08].

As mentioned earlier, the selection of image features is very important to solve the segmentation problem. Most of the aforementioned algorithms use single features, e.g. the gray value or color, to define the similarity criterion for segmentation. If other features, like texture or depth, are available or useful, the combination of those features is as important as their selection. Traditionally, the different features are assumed to be independent, see Chapter 3.1. Thus, classical probability theory can be used to fuse features. In contrast to others, this thesis proposes to use Dempster's theory of evidence to fuse information arising from different feature channels. This theory allows to model inaccuracy and uncertainty explicitly. Thus, the feature fusion using Dempster's theory of evidence becomes much more reasonable.

Earlier works on image segmentation using Dempster's theory of evidence, that are somehow related to this thesis have been presented in [MM03, CSFB08, CSFB09, AO07, RZ02]. These works combine the evidence theory with either a simple thresholding [MM03], a decisional procedure [CSFB08], a fuzzy clustering algorithm [CSFB09], a region merging algorithm [AO07] or a k-means clustering algorithm [RZ02]. All cited works use Dempster's theory of evidence to fuse the information arising from three color channels. In contrast to these works the proposed approaches combine the evidence theory with a variational (or discrete) segmentation



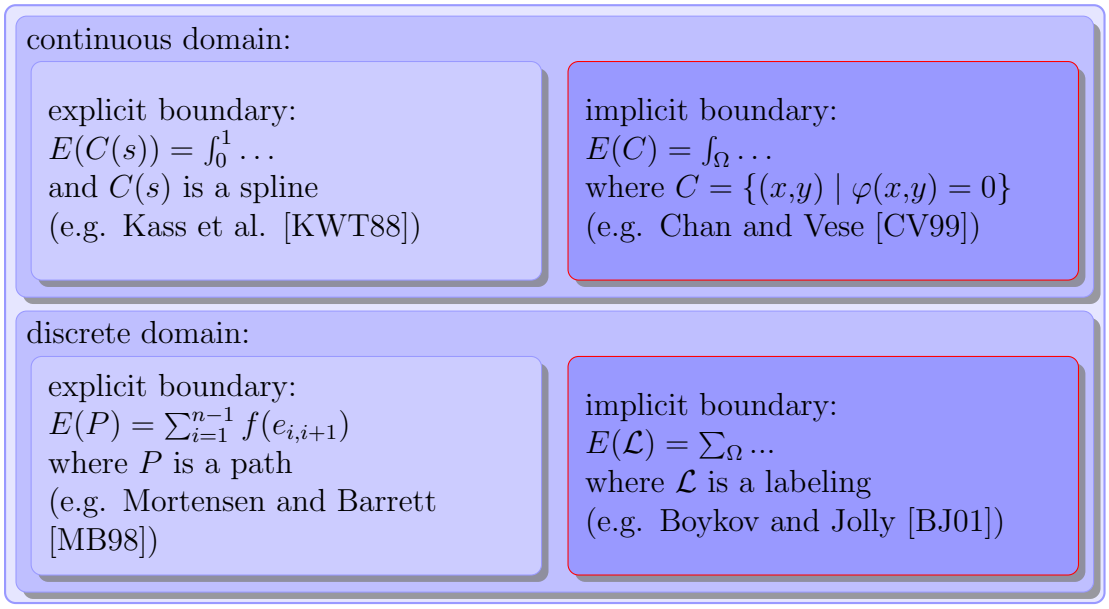


Figure 2.3: Overview of energy minimizing segmentation approaches. Segmentation approaches can be divided into four categories based on the domain of the objective function (discrete or continuous) and the boundary representation (explicit or implicit).

framework, including statistical modeling of regions, the respective Euler-Lagrange equations and a smoothness term.

The aforementioned algorithms are useful, but it has been shown that they do not succeed in many complex situations. Recent state-of-the-art algorithms often minimize an appropriate objective function that takes into account region properties and some regularization term, e.g. boundary length. Those methods are summarized in the following chapter.

2.2 Segmentation as Energy Minimization

The problem of image segmentation has been formalized in 1985 by Mumford and Shah as the minimization of an objective function [MS85]. In 1988 Kass, Witkin and Terzopoulos [KWT88][1] used a parametric curve to minimize an objective function. As shown by many papers and textbooks on image segmentation using energy minimizing frameworks [OS88, MB98, CV99, CSV00, SM00, CV01, BJ01, OP03, FH04] there has been a lot of progress in the last decades.

[1] This breakthrough is one of the most cited papers in computer vision.

Energy minimizing segmentation approaches can be divided into four categories, see Figure 2.3. They are divided based on the boundary representation, which can be explicit or implicit, and on the domain of the objective function, which can be defined in the discrete or continuous domain. Most methods, where the objective function is defined in the continuous domain, use variational methods (e.g. Euler-Lagrange equation, partial differential equations, total variation) to minimize the objective function. On the other hand, if the objective function is defined in the discrete domain (on pixel level), graph-based methods (e.g. shortest path, graph cuts, method by Felzenszwalb and Huttenlocher) are used for minimization.

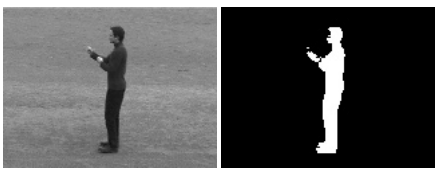
In the continuous domain

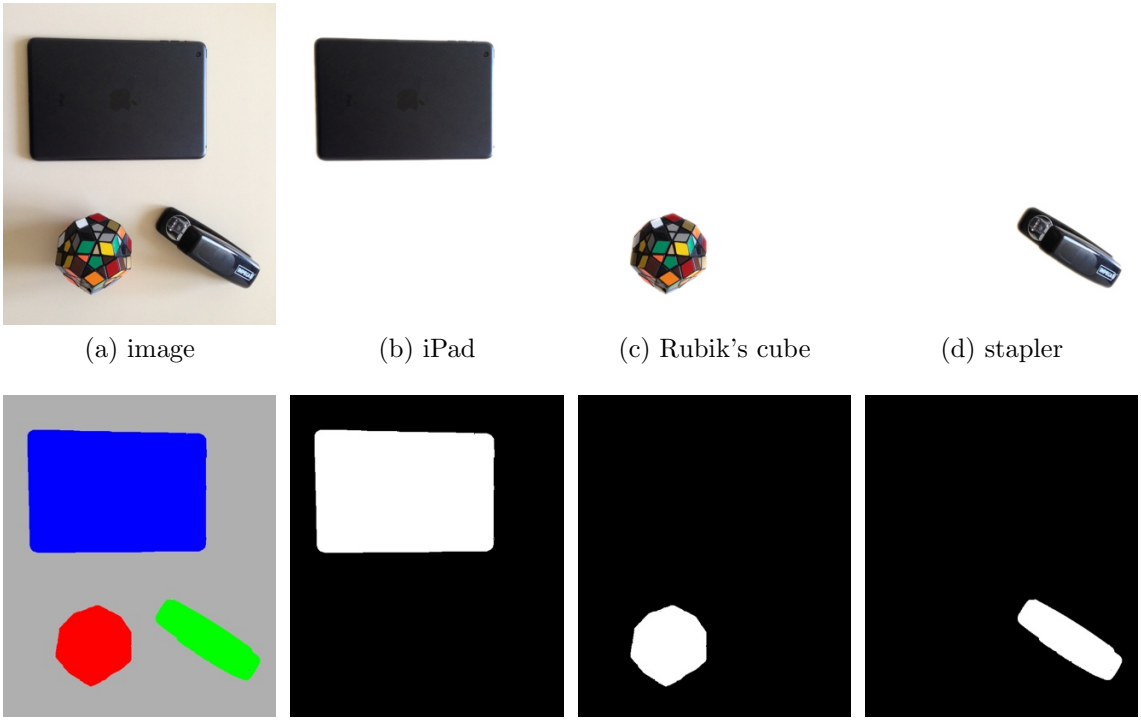
- ... using an explicit boundary representation (snakes and active contours): The curve, dividing two regions, is typically given as a parametric curve, e.g. a spline function. The objective function is defined over the curve and usually includes the image gradient under the curve and regularization terms defined on the curve, e.g. the length of the curve. The minimization of the objective function leads to an evolving curve. A limitation of these methods is that it is hard to change the topology of an evolving parametric curve, [KWT88].
- ... using an implicit boundary representation (level sets): The curve, describing object boundaries is typically embedded as the zero level set of a higher-dimensional function. The objective function is defined in the image domain and usually includes visual cues like object/background color and boundary length [CV99, CSV00, CV01] or curvature and gradient information [OS88, OP03].

In the discrete domain

- ... using an explicit boundary representation (dynamic programming and path-based): Typically, a path between given points on the boundary of an object is searched by minimizing an objective function. Analogously to the continuous domain, the objective function includes the length of the path and the image gradient along the path. A very well known example is intelligent scissors, proposed by Mortensen and Barrett [MB98].
- ... using an explicit boundary representation (graph cuts): Typically the object is separated from the background by cutting a (network) graph. The objective function is defined on nodes of the graph (pixels, image domain) and similarly includes cues like color, texture and boundary length [BJ01].

This thesis concentrates on methods using an implicit boundary representations. The objective function of these methods, is typically defined in the image domain. Therefore, it is straight forward to take into account model-specific visual cues and contextual information in order to segment a particular object of interest. These methods are explained in more detail in Chapter 3.





(a) image (b) iPad (c) Rubik's cube (d) stapler

(e) multi-label (f) iPad (g) Rubik's cube (h) stapler

Figure 2.4: Examples of different segmentation results to show the ambiguity of the problem. (a) input image; (b - c) possible binary segmentations; (d) multi-label segmentation mask; (e - f) binary segmentation masks (foreground depicted in white).

2.3 Interactive Segmentation

Fully automated (unsupervised) image segmentation still is an unsolved problem and only possible for very specific tasks. On the other hand, manual pixel labeling is accurate but very time consuming. Hence manually assigning each pixel in the image a corresponding region label is unacceptable. Furthermore, already the binary segmentation problem is highly ambiguous (see Figure 2.4). In (b - d) and (f - h) binary segmentation problems are solved to obtain the iPad, Rubik's cube or the stapler from a natural scene (a). A possible multi-label segmentation problem ($K = 4$) was solved in (e). Here each color represents a region. In the given example all regions are spatially connected, but this is not required by the definition given in Equation (2.1).

Since the segmentation problem is highly ambiguous and at the end task dependent, a useful segmentation system needs some task specific prior to solve the

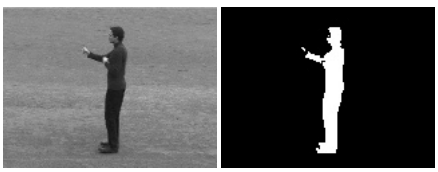
problem. Else it is likely useless for many applications. Here the user, that needs to solve the segmentation problem comes into the process loop. The user has to provide prior information to restrict the solution space. The prior information used in this thesis is a user input marking foreground and background. Typically this prior information comes from user scribbles (strokes), a rectangle around the foreground object or a so-called lasso initialization that roughly marks object edges. Other possible priors could be connectivity priors, shape priors, texture priors or object class priors [CZ05, CRD07].

The role of the user is a fundamental aspect of interactive segmentation systems. In many applications, e.g. medical image analysis, it is very important that the user maintains control over the segmentation result. Therefore an optimal interactive segmentation framework involves the user with very little effort but full control. Thus it allows segmenting objects quickly and accurately with only minor supervisions. Common and well known methods for interactive image segmentation are intelligent scissors, proposed by Mortensen and Barrett [MB98], seeded region growing, proposed by Adams and Bischof [AB94, Zuc76], magic wand, proposed by Adobe [Inc02], level sets [CSV00], geodesic active contours [CKS97], weighted total variation [BElV$^+$07, UPT$^+$08], graph cuts [BJ01, RKB04] and geodesic segmentation [Toi96, BS07, PS07].

The present thesis focuses on feature fusion for interactive, energy minimizing image segmentation to improve segmentation results. Interactive image segmentation using level sets by Chan and Vese [CSV00] and Cremers et al. [CFRA07] and using graph cuts by Boykov and Jolly [BJ01], respectively, build the basis for this thesis. Therefore, their methods and related work are reviewed in more detail in Chapter 3.

Figure 2.5 summarizes the general workflow of an interactive segmentation system. Before the segmentation takes place, the user needs to provide prior information on the regions. Based on this prior information statistical models, e.g. Gaussian distributions, are learned for each region. An interactive segmentation algorithm aims to search for the best segmentation according to the statistical models. If the segmentation result is not accurate enough the user can refine the segmentation by providing additional priors, e.g. by marking more object/background regions. Chapter 4 shows how Dempster's theory of evidence is used to include such prior information in a variational segmentation framework.

Yet another important aspect of an interactive segmentation framework is the runtime. It is crucial to get a fast feedback (segmentation result) based on given prior information to see if additional priors are necessary to satisfactory solve the task. For a human user it would not be acceptable to wait several minutes or hours for a single segmentation that might be not satisfying. Many papers address this problem for variational segmentation methods. In [AS95], Adalsteinsson and Sethian proposed to evaluate the curve only in a so-called narrowband to decrease



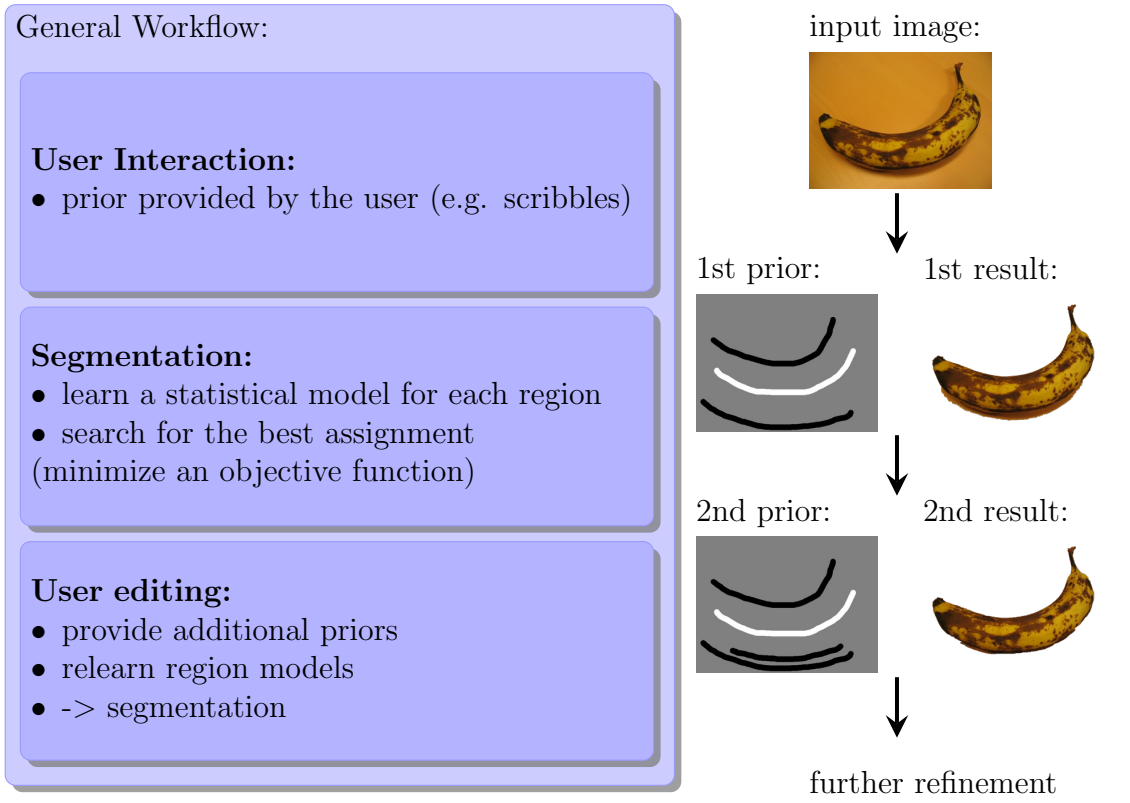


Figure 2.5: General interactive segmentation workflow.

computational complexity. Different data structures, e.g. sparse fields, octaves or run-length encoded curves, have been shown to speedup existing segmentation approaches [Whi98, Bri03, Str99, LGF04, HNB+06]. To further improve the complexity Sethian [Set96, Set99] and Osher and Fedkiw [Set96] proposed to combine the level set method and the fast marching method. GPU implementations of variational approaches have been shown to outperform traditional algorithms in terms of computational time [LCW03, CLW04, RPSM10].

Besides the improvements on variational segmentation approaches there has also been a lot of progress in the discrete domain. Existing works either improve the augmenting path algorithm by Boykov and Kolmogorov [BJ01, BK04], that is a widely used algorithm in computer vision, or push-relabel algorithms, that are parallelizable [DB08]. Besides those approaches to develop more efficient algorithms for the general maximum flow / minimum cut problem researches are also trying to reduce the size of the labeling problem itself by grouping variables [LSTS04, WBC+05, LSS09, KNKY11] or multi-scale approaches [PB99, SG06, KLR10]. Chapter 5 will show how Dempster's theory of evidence is used to speed up a discrete segmentation framework, by reducing the problem size efficiently, in order to enhance usability.

2.4 Benchmarks

Performance evaluation for (interactive) segmentation systems is very important for developing a real world application. Traditionally, computer vision algorithms are evaluated on a benchmark with preselected training and test data. Typical segmentation benchmarks are the Berkeley segmentation benchmark (BSD500 or BSD300)[2,3] and the Grab Cut benchmark, provided by Microsoft[4] [RKB04]. The Berkeley segmentation benchmark [MFTM01, AMFM11] consist of 300 (500) images and ground truth segmentations. As stated by the authors, the main goal of this dataset is to provide an empirical basis for research on image segmentation and boundary detection. Therefore most of the images are chosen to fit into the problems multi-region segmentation or boundary detection. Since this thesis deals with the problem of (interactive) binary segmentation the Microsoft Grab Cut dataset is used in most experiments. It consists of 50 real world images and corresponding ground truth segmentations, see Figure 2.6 (a) and (b). This dataset was especially designed to measure the performance of binary segmentation algorithms.

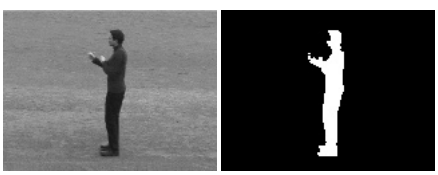
The performance is typically given as the Hamming distance of the segmentation result and the ground truth segmentation [BRB+04]:

$$\epsilon = \frac{\text{no. misclassified pixels}}{\text{no. pixels in unclassified region}}, \tag{2.2}$$

where the denominator is the number of pixels that are not classified by the user. Other performance measures used in this thesis are *Precision*, *Recall* and F_1-*measure* [MO10, MFTM01].

Therefore a fixed set of user-interactions is commonly used to initialize an interactive segmentation system and simulate the human user. In this work the set of interactions provided by the Microsoft dataset (rectangle or lasso trimaps) and our own brush stroke trimaps, see Figure 2.6 (c) are used for initialization. These trimaps provide a priori information about the object, background and unclassified regions to simulate the user. Furthermore the proposed methods are evaluated with a user study measuring the number of user interactions needed for satisfactory segmentation results. In order to evaluate the methods on synthetic textured images the Prague texture segmentation data-generator and benchmark [HM08][5] is used. Segmenting video sequences is an additional application of the proposed methods. Therefore videos sequences from the KTH action dataset [SLC04][6], videos provided

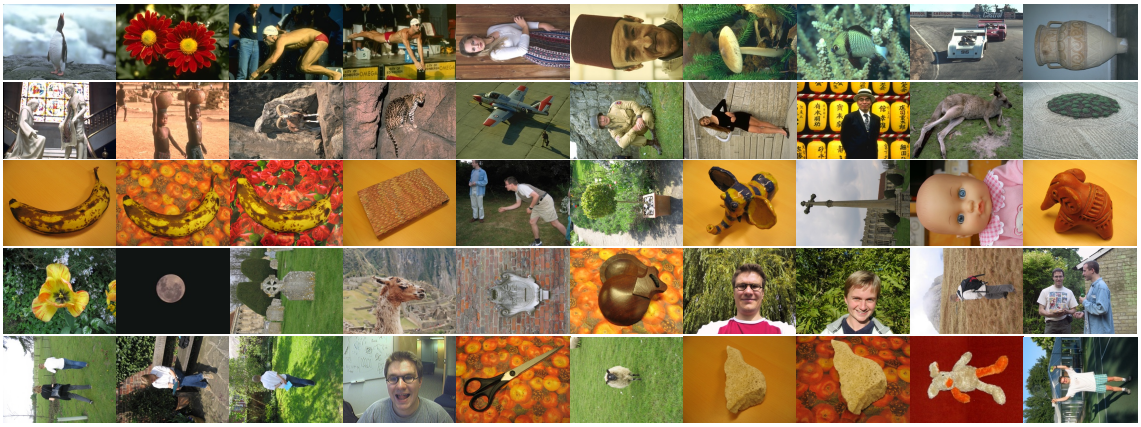
[2] http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/
[3] http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/resources.html
[4] http://research.microsoft.com/en-us/um/cambridge/projects/visionimagevideoediting/segmentation/grabcut.htm
[5] http://mosaic.utia.cas.cz
[6] http://www.nada.kth.se/cvap/actions/



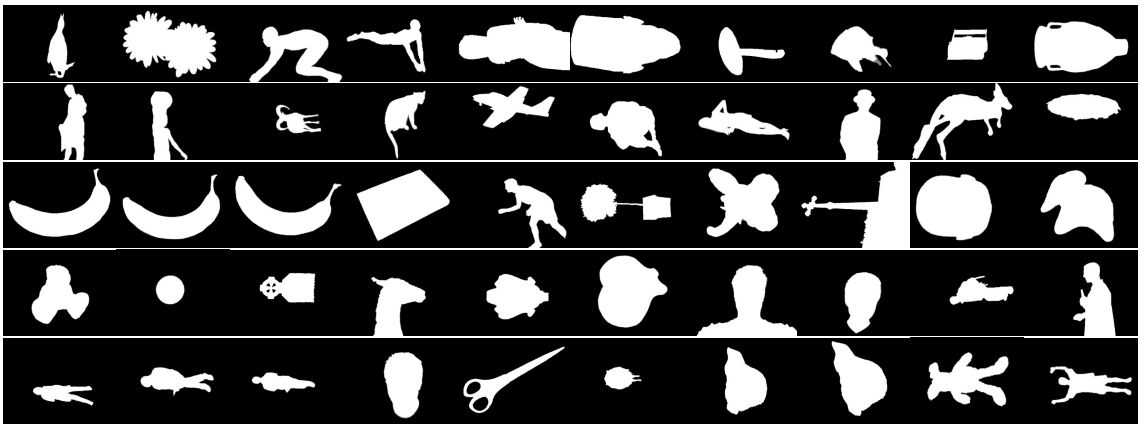
by Sand and Teller [ST06][7] and the ToFCut dataset including depth information [WZYZ10][8] are used for evaluation.

A critical point, that all the benchmark images share, is the small image resolution. Typically, the size of these images is less than 0.5 megapixels. In contrast, nowadays digital cameras are able to take images with more than 20 megapixels. Thus, it is an important aspect to evaluate the segmentation performance and runtime on such high-resolution images. Therefore, images found on the web with resolutions up to 26 megapixels are further used for evaluation. A similar aspect is to evaluate the segmentation performance and runtime on resource-limited systems. For that, benchmark images with up to 2.5 megapixels are evaluated on Apple's iPhone 4.

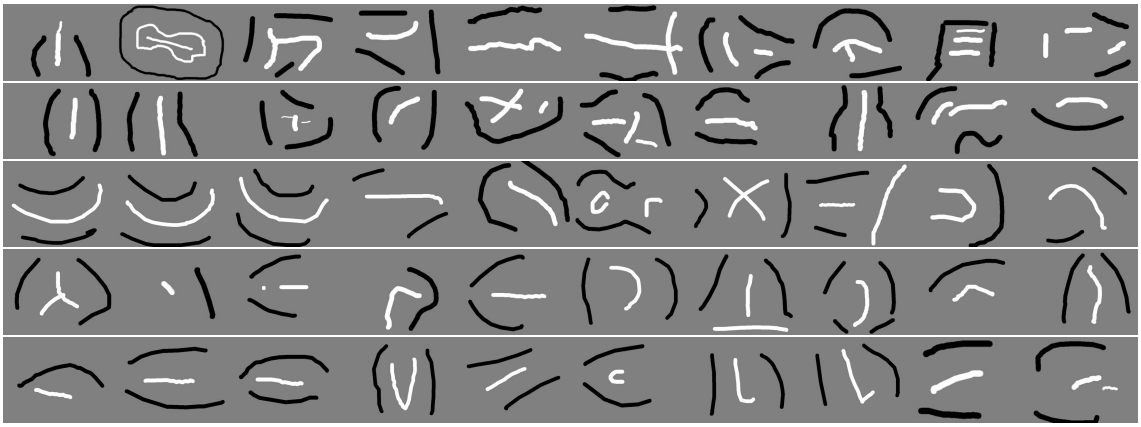
[7] http://rvsn.csail.mit.edu/pv/
[8] http://vis.uky.edu/%7Egravity/Research/ToFMatting/ToFMatting.htm



(a) images

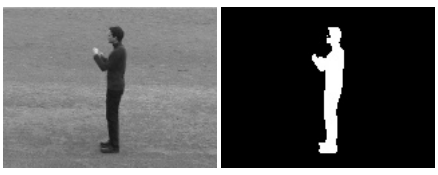


(b) ground truth segmentations



(c) user stroke initializations

Figure 2.6: Images (a) and ground truth segmentations (b) from the Microsoft GrabCut Benchmark [RKB04]. Since the benchmark does not include stroke initializations our own initializations are used(c). Note: Some images are cropped for a smooth illustration. In the experiments the original images are used.



Chapter 3

Background

In this chapter the energy minimizing segmentation frameworks using level sets and graph cuts are described and recapitulated. A probabilistic interpretation of both frameworks is given to clarify the connection between both frameworks. At the end of the chapter, the proposed feature fusion using Dempster's theory of evidence is explained and the relation to Bayes' theory is explained.

3.1 Segmentation Using a Variational Framework

The variational segmentation framework used in this work is based on the works of [CV01, CRD07]. Using a level set representation of the curve C, describing object boundaries, has several well known advantages, e.g. the naturally given possibility to handle topological changes of the boundary curve. This is especially important if the object is partially occluded by another object or if the object consists of multiple parts.

3.1.1 Chan-Vese Energy Functional

In case of a binary segmentation, the level set function $\varphi : \Omega \to \mathbb{R}$ splits the image domain Ω into foreground and background regions FG, BG $\subseteq \Omega$ with $\mathrm{FG} \cap \mathrm{BG} = \emptyset$ and $\mathrm{FG} \cup \mathrm{BG} = \Omega$. Usually, φ is defined by a signed distance function, see Figure 3.1a, that holds:

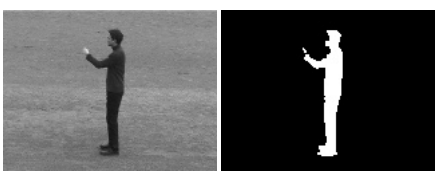
$$\varphi(x) = \begin{cases} \geq 0, & \text{if } \mathcal{L}_x \in \mathrm{FG} \\ < 0, & \text{if } \mathcal{L}_x \in \mathrm{BG} \end{cases} , \tag{3.1}$$

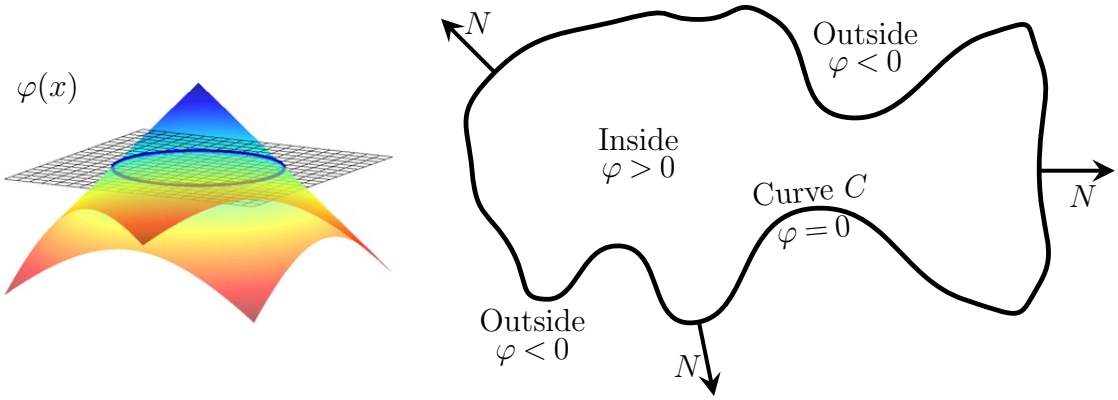
where $\mathcal{L} = \{\mathcal{L}_x \mid x \in \Omega \wedge \mathcal{L}_x \in \{\mathrm{FG}, \mathrm{BG}\}\}$ is the labeling. The zero-level line of the function $\varphi(x)$ represents the boundary C between the object, which is sought to be extracted, and the background.

$$C = \{x \in \Omega \mid \varphi(x) = 0\} . \tag{3.2}$$

Minimizing an objective function should propagate the curve C in normal direction, see Figure 3.1b

The basis segmentation framework in this thesis is the so-called *Chan-Vese energy functional* [CV01] for gray-scale images. Therefore, let $I : \Omega \mapsto \mathbb{R}$ be the image, a function that maps the image domain to the space of real numbers. Chan and Vese assume that the image is formed by two regions (FG and BG), that have a homogeneous but distinct gray value distribution. Furthermore it is assumed, that the distributions can be approximated by the mean values μ_{FG} and μ_{BG}. Using the Heaviside function H it is possible to indicate to which region a pixel belongs. It is





(a) (b)

Figure 3.1: (a): Level set representation of a curve. (b): Curve $C = \{x \in \Omega \mid \varphi(x) = 0\}$ propagating in normal direction.

defined by:

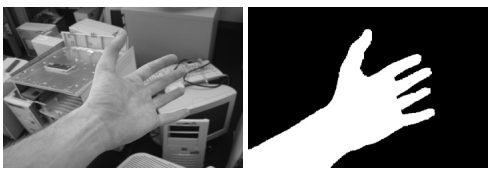
$$H(z) = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{if } z < 0\,, \end{cases} \qquad \nabla H(z) = \delta(z) = \begin{cases} 1, & \text{if } z = 0 \\ 0, & \text{else }, \end{cases} \tag{3.3}$$

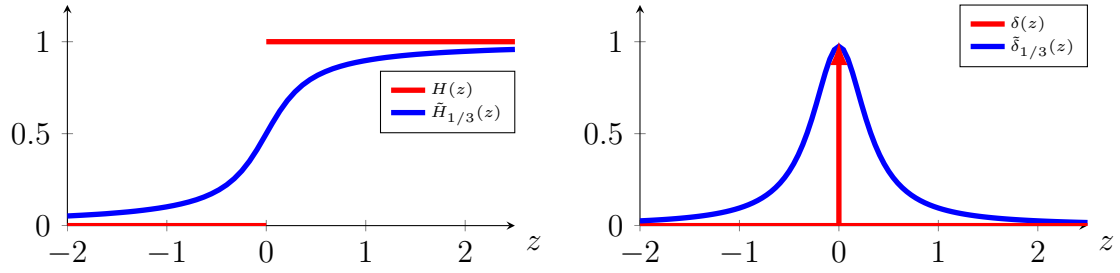
The derivative of the Heaviside function is the well known Dirac measure, that is used to indicate points on the curve C. Using the aforementioned notation, the Chan-Vese energy functional in its original form is given by:

$$\begin{aligned} E(\mu_{FG},\mu_{BG},\varphi) = \lambda_1 &\int_\Omega |I(x) - \mu_{FG}|^2 H(\varphi)\, dx \\ &+ \lambda_2 \int_\Omega |I(x) - \mu_{BG}|^2 (1 - H(\varphi))\, dx \\ &+ \nu_1 \cdot \int_\Omega |\nabla H(\varphi)|\, dx + \nu_2 \cdot \int_\Omega H(\varphi)\, dx\,. \end{aligned} \tag{3.4}$$

The first two terms are the *external energy*, taking into account the image data and the other two terms are the *internal energy*, that acts directly on the curve. Note: For the sake of convenience $\varphi(x)$ is shorten by φ.

Minimizing the external energy with $\lambda_1 = \lambda_2$ and $\nu_1 = \nu_2 = 0$, results in the best labeling by minimizing the squared distances to the mean values of foreground and background. In other words: a pixel x is labeled as foreground, if $|I(x) - \mu_{FG}|^2 < |I(x) - \mu_{BG}|^2$. Ignoring the internal energy, the energy minimization is a special case of the k-means clustering with $k = 2$.





(a) Heaviside function (b) Dirac measure

Figure 3.2: Visualization of the Heaviside function and its derivative, the Dirac measure (red) and the corresponding regularized functions (blue) with $\epsilon = 1/3$.

The internal energy is used for regularization by minimizing the length of the curve C and the area inside the curve:

$$Length(C) = \int_\Omega |\nabla H(\varphi)|\, dx\,, \qquad Area(inside(C)) = \int_\Omega H(\varphi)\, dx\,. \tag{3.5}$$

Keeping the level set function φ fixed and minimizing the energy $E(\mu_{FG},\mu_{BG},\varphi)$ (see Equation (3.4)) with respect to μ_{FG} and μ_{BG} yields:

$$\mu_{FG} = \frac{\int_\Omega I(x)H(\varphi(x))\, dx}{\int_\Omega H(\varphi(x))\, dx} \text{ and } \mu_{BG} = \frac{\int_\Omega I(x)(1-H(\varphi(x)))\, dx}{\int_\Omega (1-H(\varphi(x)))\, dx}\,, \tag{3.6}$$

if $\int_\Omega H(\varphi(x))\, dx > 0$ and if $\int_\Omega (1-H(\varphi(x)))\, dx > 0$. In the end, this comes down to compute the mean value of FG and BG.

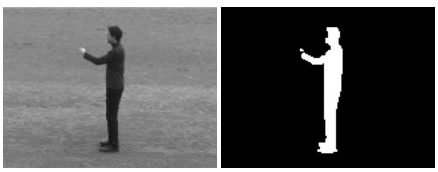
In order to minimize the energy with respect to φ, regularized versions of the Heaviside function and the Dirac measure are considered. They need to hold:

$$(i)\ \lim_{z\to -\inf} \tilde{H}(z) = 0\,, \quad (ii)\ \lim_{z\to \inf} \tilde{H}(z) = 1\,, \quad (iii)\ \tilde{H}(0) = 0.5\,. \tag{3.7}$$

One family of such regularizations can be defined by:

$$\tilde{H}_\varepsilon(z) = \frac{1}{2} + \frac{1}{\pi}\arctan\frac{s}{\varepsilon}\,. \tag{3.8}$$

For $\varepsilon \to 0$, $\tilde{H}(z)$ converges to the Heaviside function $H(z)$. An important property of this approximation is that it acts on all level curves, what makes it more robust to local minima [CV01]. A comparison for $\varepsilon = 1/3$ is visualized in Figure 3.2. In the following, $H(z)$ and $\delta(z)$ denote their regularized versions.



Minimizing the energy functional with respect to φ while keeping μ_{FG} and μ_{BG} fixed can be performed by solving the corresponding Euler-Lagrange equation. The Euler-Lagrange equation is a necessary condition for an extremum of the energy functional. The energy functional has the form:

$$E(\varphi) = \int_{\Omega} \mathcal{L}(\varphi, \nabla\varphi)\, dx \tag{3.9}$$

and the corresponding Euler-Lagrange equation is given by:

$$\frac{dE}{d\varphi} = \frac{\partial \mathcal{L}}{\partial \varphi} - \frac{\partial}{\partial x_1}\frac{\partial \mathcal{L}}{\partial \varphi_{x_1}} - \frac{\partial}{\partial x_2}\frac{\partial \mathcal{L}}{\partial \varphi_{x_2}} = 0\,. \tag{3.10}$$

Inserting the energy function from Equation (3.4) or building the derivatives of the corresponding terms of the energy function, respectively, leads to:

$$\frac{dE}{d\varphi} = \delta(\varphi)\left[\lambda_1 |I(x) - \mu_{FG}|^2 - \lambda_2 |I(x) - \mu_{BG}|^2 + \nu_2 - \nu_1 \cdot \operatorname{div}\left(\frac{\nabla\varphi}{|\nabla\varphi|}\right)\right], \tag{3.11}$$

where div is the divergence (see Appendix A.1 for more details).

Parameterizing the descent direction by an artificial time [CV01], the Euler-Lagrange equation leads to the following partial differential equation (PDE):

$$\frac{\partial\varphi}{\partial t} = -\frac{dE}{d\varphi} = \delta(\varphi)\left[\lambda_2 |I(x) - \mu_{BG}|^2 - \lambda_1 |I(x) - \mu_{FG}|^2 - \nu_2 + \nu_1 \cdot \operatorname{div}\left(\frac{\nabla\varphi}{|\nabla\varphi|}\right)\right], \tag{3.12}$$

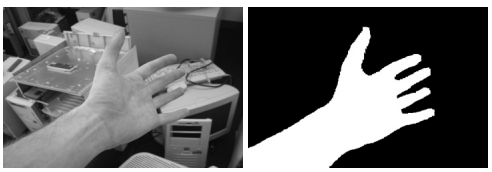
that can be solved using numerical methods. The final curve C, describing the boundary between foreground and background, is computed by alternating between the gradient descent step and the parameter optimization of μ_{FG} and μ_{BG} [CRD07]. Starting with some (manually given) initial contour φ^0, this is an initial value problem. As a consequence, the quality of the segmentation process is limited by the initial curve. A block diagram of the algorithm is given in Figure 3.3. Given a curve φ^n, the parameters μ_{FG} and μ_{BG} are updated according to Equation (3.6). Next, the curve is evolved according to the following discretization and linearization of Equation (3.11), which is in principal the well known Euler method (see [CV01]):

$$\varphi^{n+1} = \varphi^n + \Delta t \cdot \delta(\varphi)\left[\lambda_2 |I(x) - \mu_{BG}|^2 - \lambda_1 |I(x) - \mu_{FG}|^2 - \nu_2 + \nu_1 \cdot \operatorname{div}\left(\frac{\nabla\varphi}{|\nabla\varphi|}\right)\right], \tag{3.13}$$

where Δt is the time step. To improve segmentation results, more advanced techniques can be used to solve the partial differential equation. E.g. in [SR09], a 2nd order Runge-Kutta method, that outperforms the Euler method, is proposed.

Equation (3.4) is a special case of the piecewise constant Mumford Shah model [MS89]:

$$\int_{FG} |I(x) - c_{FG}|^{\beta}\, dx + \int_{BG} |I(x) - c_{BG}|^{\beta}\, dx + \nu \cdot (|\partial FG| + |\partial BG|)\,, \tag{3.14}$$



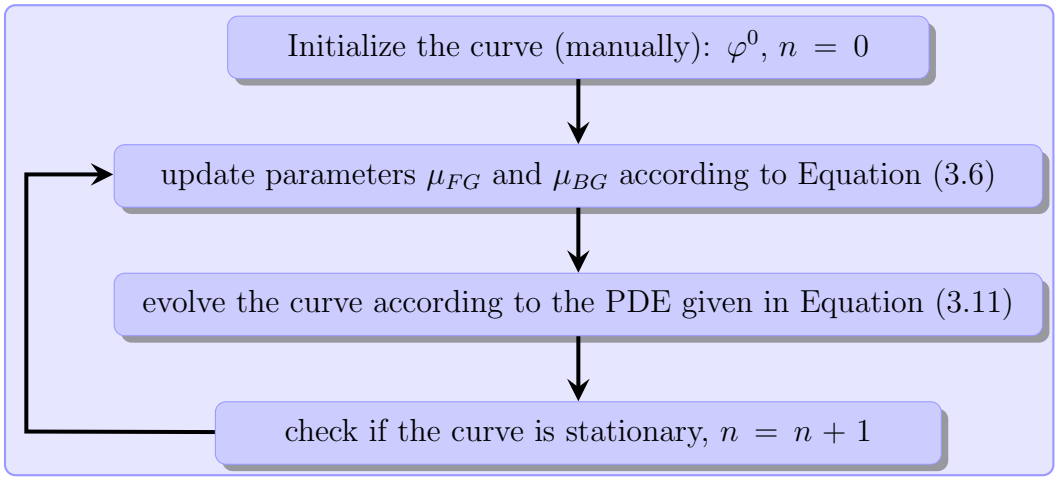


Figure 3.3: General workflow of the level set segmentation framework.

with $\beta = 2$ and $|\partial FG|$ represents the length of the boundary of the region FG. Keeping c_{FG} and c_{BG} fixed, the Mumford Shah model for image segmentation has the form of a Pott's model [Pot52].

One general assumption of the original Chan-Vese energy functional is, that the two regions can be modeled by two disjoint, unimodal distributions. If this assumption is violated, e.g. by overlapping and/or multimodal distributions, the segmentation will easily fail.

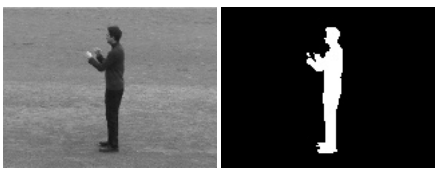
A natural extension is given by the following energy function:

$$\begin{aligned} E(\varphi) = &- \int_\Omega H(\varphi) \log p_1 + (1 - H(\varphi)) \log p_2 \, dx \\ &+ \nu \int_\Omega |\nabla H(\varphi)| \, dx \,, \end{aligned} \tag{3.15}$$

where $\nu \geq 0$ weights the influence of the internal energy, $H(s)$ is a regularized Heaviside function and p_1 and p_2 are conditional probability densities of FG and BG:

$$p_1 = p_1(I(x)) = p(I(x) \mid \mathcal{L}_x = FG) \,, \quad p_2 = p_2(I(x)) = p(I(x) \mid \mathcal{L}_x = BG) \tag{3.16}$$

Given the two conditional probabilities p_1 and p_2, the total a posteriori probability is maximized by minimizing the first term of Equation (3.15), i.e. pixels are assigned to the most probable region according to Bayes' rule. Minimizing the second term penalizes the length of the contour. It can be interpreted as a priori knowledge and acts as a smoothing term. In comparison to the original energy functional, see Equation (3.4), the regularization using the area inside the curve is omitted, since it reduces to a constant shrinking bias (see Equation (3.13) and [CRD07]).



Similarly, minimization of the energy functional (3.15) can be performed by solving the Euler-Lagrange equation with respect to φ [CV01]. This leads to the following partial differential equation:

$$\frac{\partial \varphi}{\partial t} = \delta(\varphi) \left(\log \left(\frac{p_1}{p_2} \right) + \nu \operatorname{div} \left(\frac{\nabla \varphi}{|\nabla \varphi|} \right) \right) , \tag{3.17}$$

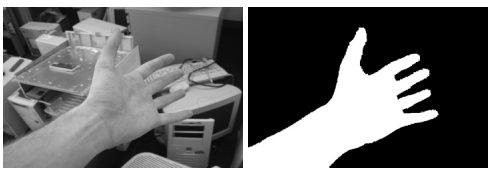
Again, starting with some initial contour φ^0 and given the conditional probability densities p_1 and p_2 an initial value problem has to be solved. The idea of curve evolution by iteratively solving the partial differential equation is clarified in Figure 3.4.

As a consequence, the quality of the segmentation process is limited by the initial contour and the way the foreground and background probabilities p_1 and p_2 are modeled. For the variational segmentation approach in this thesis, the nonparametric Parzen estimates [RBD03], which is a well known histogram-based method, is used. This Model is chosen since, compared to multivariate Gaussian Mixture Models (GMM), it leads to similar results without the need of estimating model parameters. An example, showing the difference between both models for a gray value image is given in Figure 3.5.

Other possibilities to model the probability densities given the image cues are, e.g., a Gaussian density with fixed standard deviation [CV01] or a generalized Laplacian [HS05]. In scenes with complex objects, shadows, and highlights, where differences between the object and the background are often only locally visible, a local Gaussian probability density that varies with the position $x \in \Omega$ in the image can be used[KB03, RBW07].

For the proposed method it is necessary to extend this segmentation framework from gray-scale images to feature vector images $I = (I_1, \ldots, I_m)$. This extension is straight forward and has been applied to the Chan-Vese model [CSV00]. It is assumed that the channels I_j are independent. Thus, the conditional probability density $p_i(I(x))$ of foreground or background is the product of the separated conditional probabilities $p_i(I_j(x)) = p(I_j(x) \mid \mathcal{L}_x = i)$:

$$\begin{aligned} p_i(I(x)) &= p(I_1(x) \cap \ldots \cap I_m(x) \mid \mathcal{L}_x = i) \\ &= p(I_1 \mid \mathcal{L}_x = i) \cdot \ldots \cdot p(I_m \mid \mathcal{L}_x = i) \\ &= p_i(I_1(x)) \cdot \ldots \cdot p_i(I_m(x)) \,. \end{aligned} \tag{3.18}$$



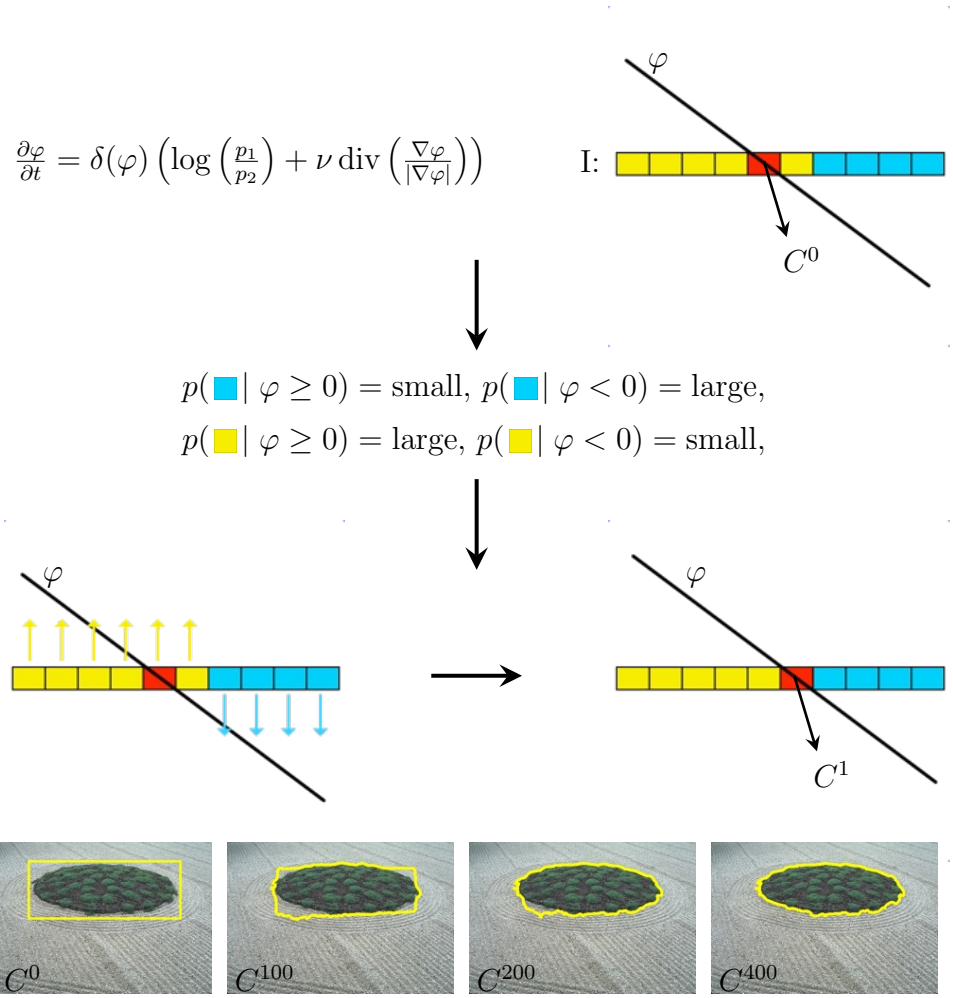
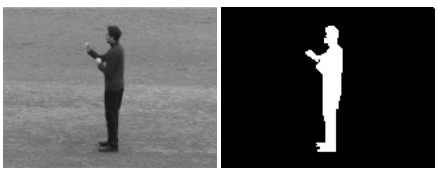


Figure 3.4: Example curve evolution using a level set representation. Upper part: Given the partial differential equation, an one dimensional image and an initial level set, the probabilities for each pixel are computed and the curve is evolved. In this example the probabilities for yellow (blue) pixels forces the level set function to move upwards (downwards). Thus the curve, depicted in red, evolves to the right. Lower part: Example of an evolving curve for a real image.



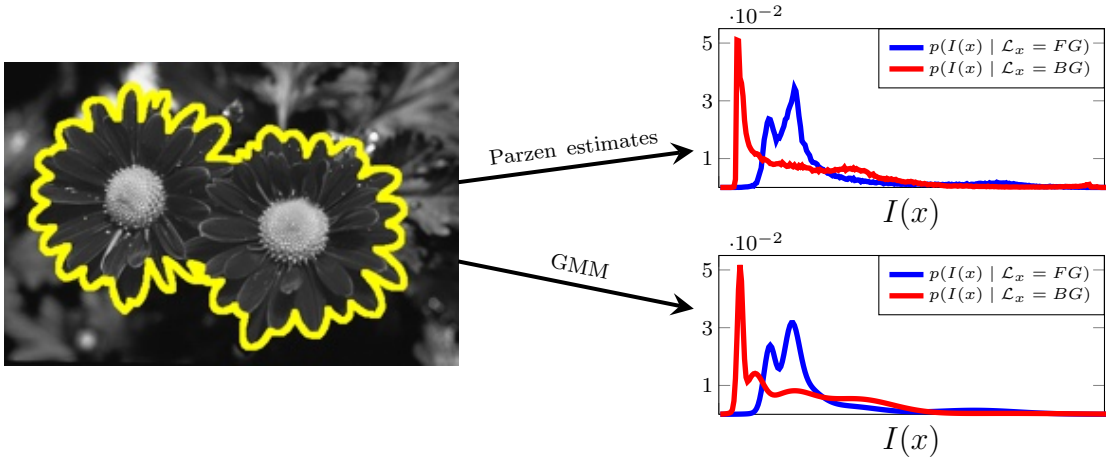


Figure 3.5: Comparison of the nonparametric Parzen estimates (top right) and a Gaussian Mixture Model with 5 kernels (bottom right) for foreground and background regions of the image on the left.

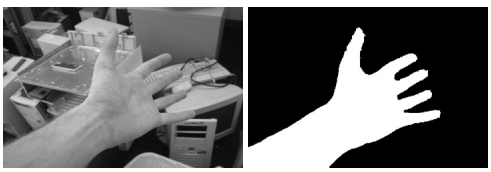
The Chan-Vese model (3.15) for vector images now reads:

$$\begin{aligned} E(\varphi) = &- \int_\Omega H(\varphi) \sum_{j=1}^{m} \log p_{1,j}\, dx \\ &- \int_\Omega (1 - H(\varphi)) \sum_{j=1}^{m} \log p_{2,j}\, dx \\ &+ \nu \int_\Omega |\nabla H(\varphi)|\, dx\,, \end{aligned} \tag{3.19}$$

where $p_{i,j} = p_i(I_j(x))$. Solving the corresponding Euler-Lagrange equation yields to the following partial differential equation:

$$\frac{\partial \varphi}{\partial t} = \delta(\varphi) \left[\sum_{j=1}^{m} \log \left(\frac{p_{1,j}}{p_{2,j}} \right) + \nu \operatorname{div} \left(\frac{\nabla \varphi}{|\nabla \varphi|} \right) \right]. \tag{3.20}$$

Because of their independency, the $p_{i,j}$ can be estimated for each region i and each channel j separately. In other words, for each pixel in the image the foreground probability $\sum_{j=1}^{m} \log p_{1,j}$ and the background probability $\sum_{j=1}^{m} \log p_{2,j}$ is computed over all feature channels j. Possible image features to be incorporated by means of this model are color, texture [BW06, RBD03], motion [CS05] or depth.



3.1.2 Probabilistic View

Besides the aforementioned derivation, the problem of image segmentation using level sets also has a probabilistic meaning. Given are the conditional probabilities p_1 and p_2, that are also referred to as likelihoods, modeling foreground and background statistics of an input image $I : \Omega \to \mathbb{R}$, see Equation (3.16). For the sake of simplicity only single valued images are considered. The goal of the level set segmentation approach can be interpreted as computing the most likely level set function φ separating foreground and background by maximizing the a posteriori distribution $p(\varphi \mid I)$. With Bayes' theorem this becomes:

$$p(\varphi \mid I) = \frac{p(I \mid \varphi)p(\varphi)}{p(I)} \propto p(I \mid \varphi)p(\varphi)\,, \tag{3.21}$$

since the evidence $p(I)$ is independent from the level set function φ. The first term in the product of Equation (3.21) allows to integrate the conditional probabilities p_1 and p_2 by defining:

$$p(I(x) \mid \varphi(x)) = \begin{cases} p_1(I(x))\,, \text{if } \varphi(x) \geq 0 \\ p_2(I(x))\,, \text{else} \end{cases} \quad . \tag{3.22}$$

Assuming, that the intensities at different spatial locations in an image are statistically independent, this leads to:

$$p(I \mid \varphi) = \prod_{x \in \Omega} \left[p(I(x) \mid \varphi(x))\right]^{dx}. \tag{3.23}$$

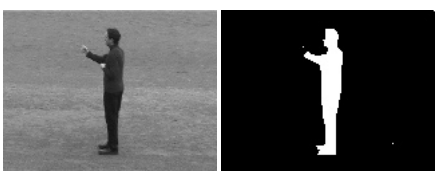
The bin value dx guarantees the correct continuum limit, see [CRD07, BC07].

The second term $p(\varphi)$ in the product of Equation (3.21) can be interpreted as prior knowledge on the embedding function, that could incorporate a priori more likely segmentations, e.g. smooth curves or special shapes. In the aforementioned Chan-Vese energy functional, this prior on the level set function is a constraint on the length of the curve:

$$p(\varphi) = \exp\left(-\nu \int |\nabla H(\varphi)| dx\right). \tag{3.24}$$

The particular choices of the conditional probability in Equation (3.23) and the prior probability in Equation (3.24) are due to the Chan-Vese segmentation model, see Equation (3.15). More sophisticated prior knowledge, e.g. knowledge about the shape or statistical shape priors can also be included [CRD07].

Maximizing the a posteriori probability $p(\varphi \mid I)$ is equivalent to minimizing its



negative logarithm. This leads to:

$$\begin{aligned}
&\max_{\varphi} p(\varphi \mid I) \propto \max_{\varphi} p(I \mid \varphi) p(\varphi) \\
\Leftrightarrow\ &\min_{\varphi} -\log\left[p(I \mid \varphi)\right) - \log(p(\varphi)] \\
=\ &\min_{\varphi} -\log\left(\prod_{x\in\Omega}\left[p(I(x) \mid \varphi(x))\right]^{dx}\right) - \log\left(\exp\left(-\nu\int|\nabla H(\varphi)|dx\right)\right) \\
=\ &\min_{\varphi} \sum_{i\in\{FG,BG\}} \int_i -\log\left(p_i(I(x))\right)\, dx + \nu\int|\nabla H(\varphi)|dx\,,
\end{aligned} \tag{3.25}$$

which is equivalent to the Chan-Vese energy functional.

3.2 Segmentation by Discrete Energy Minimization

An important drawback of the aforementioned variational segmentation approach using level sets is the dependence of the segmentation result from the initial contour. This is caused by the gradient descent approach to minimize the energy function. Since an image is defined in the discrete domain, another straightforward idea is to formulate the energy function in the discrete domain.

Due to the works of Boykov and Jolly [BJ01], Rother et al. [RKB04] and many others [GPS89, KZ04, BRB+04, KT05, Li09] image segmentation by discrete energy minimization using graph cuts became a powerful and widely used framework. The main advantage of the graph cut framework is the ability to globally minimize a certain set of energy functions. The discrete segmentation framework used in this thesis directly builds upon the framework proposed by Boykov and Jolly [BJ01].

3.2.1 Discrete Energy Model

The discrete energy $E : \mathcal{L}^n \rightarrow \mathbb{R}$ for the problem of binary image labeling can be written as the sum of unary τ_i and pairwise potentials $\tau_{i,j}$

$$E(\mathcal{L}) = \sum_{i\in\mathcal{V}} \tau_i(\mathcal{L}_i) + \sum_{(i,j)\in\mathcal{E}} \tau_{i,j}(\mathcal{L}_i,\mathcal{L}_j)\,, \tag{3.26}$$

where $\mathcal{V}$ corresponds to the set of all image pixels and $\mathcal{E}$ is the set of all edges between pixels in a defined neighborhood $\mathcal{N}$. For the problem of binary image segmentation, which is addressed in this thesis, the labeling $\mathcal{L}$ consists of foreground (FG) and a background (BG) labels.

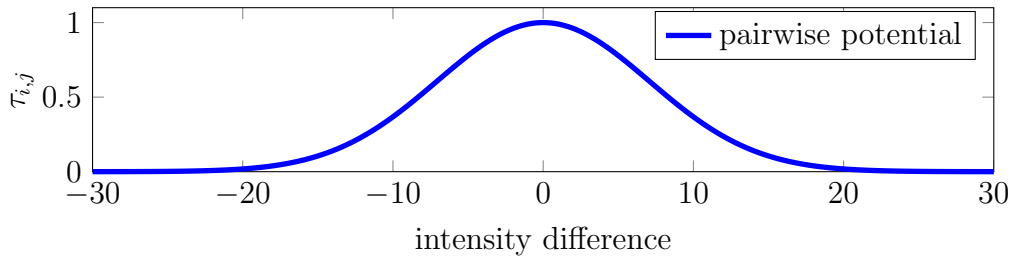


Figure 3.6: The pairwise potential $\tau_{i,j}(\mathcal{L}_i,\mathcal{L}_j)$ given in Equation (3.30) for $\mathcal{L}_i \neq \mathcal{L}_j$, $\beta = 1/100$, $\gamma = 1$ and pixels with a distance of one.

Note: In the literature, e.g. [KZ04, BKR11], typically the unary and pairwise potentials are defined by φ_i and $\varphi_{i,j}$ or D_i and $V_{i,j}$. Since in this thesis φ is already used for the level set function and D for the depth image, τ_i and $\tau_{i,j}$ are used instead.

In comparison to the variational approach, the unary potential represents the external energy and the pairwise potential the internal energy, that is used for regularization.

The unary potential τ_i is given as the negative log-likelihood of a learned foreground/background model. Typically, a standard Gaussian mixture model (GMM) $p(\cdot)$ [RKB04] is used. Thus, the unary potential is defined by

$$\tau_i(\mathcal{L}_i) = -\log p(I(i) \mid \mathcal{L}_i = S)\,, \tag{3.27}$$

where S is either foreground or background. In [BJ01, RKB04] it is assumed that such likelihoods are known a priori or learned directly from user labeled pixels, so-called seeds. The set of seeds is denoted by:

$$\mathcal{U} = \{x \in \mathcal{V} \mid x \text{ marked as } FG\} \cup \{x \in \mathcal{V} \mid x \text{ marked as } BG\} = \mathcal{O} \cup \mathcal{B}\,, \tag{3.28}$$

that impose hard constraints on the segmentation result:

$$\forall\, x \in \mathcal{O} \Rightarrow \mathcal{L}_x = FG \;\wedge \forall\, x \in \mathcal{B} \Rightarrow \mathcal{L}_y = BG\,. \tag{3.29}$$

The pairwise potential $\tau_{i,j}$ takes the form of a contrast sensitive Ising model, defined by:

$$\tau_{i,j}(\mathcal{L}_i,\mathcal{L}_j) = \gamma \cdot \mathrm{dist}(i,j)^{-1} \cdot [\mathcal{L}_i \neq \mathcal{L}_j] \cdot \exp(-\beta \|I(i) - I(j)\|^2)\,. \tag{3.30}$$

Here $I(i)$ describes the feature vector (e.g. color) of pixel i and $\mathrm{dist}(i,j)$ is the Euclidean distance of the pixels i and j. $[\cdot]$ is the indicator function and the parameter γ specifies the impact of the pairwise function. The indicator function is defined by

$$[\mathcal{L}_i \neq \mathcal{L}_j] = \begin{cases} 1, & \text{if } \mathcal{L}_i \neq \mathcal{L}_j \\ 0, & \text{else} \end{cases} \tag{3.31}$$

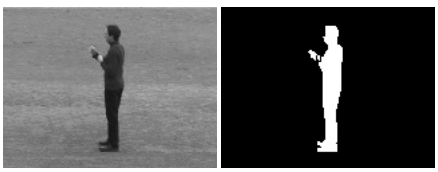


Table 3.1: Edge weights of a network graph according to Boykov and Jolly [BJ01] to represent the energy function (3.26).

edge	weight (cost)	for
$\{x,y\}$	$\tau_{x,y}(\mathcal{L}_x, \mathcal{L}_y)$ with $\mathcal{L}_x \neq \mathcal{L}_y$	$\{x,y\} \in \mathcal{E}$
$\{x,S\}$	$\tau_x(BG)$	$x \in \mathcal{V} \wedge x \notin \mathcal{O} \cup \mathcal{B}$
	K	$x \in \mathcal{O}$
	0	$x \in \mathcal{B}$
$\{x,T\}$	$\tau_x(FG)$	$x \in \mathcal{V} \wedge x \notin \mathcal{O} \cup \mathcal{B}$
	0	$x \in \mathcal{O}$
	K	$x \in \mathcal{B}$
$K = 1 + \max_{x \in \mathcal{V}} \sum_{\substack{y:\{x,y\} \in \mathcal{E} \\ \wedge\, \mathcal{L}_x \neq \mathcal{L}_y}} \tau_{x,y}(\mathcal{L}_x, \mathcal{L}_y)$		

and allows to capture gradient information only along the segmentation boundary. The constant β includes the feature variance of the image and is just as defined as in [BJ01, RKB04]:

$$\beta = \left(2 \cdot \left\langle (I(i) - I(j))^2 \right\rangle\right)^{-1} , \tag{3.32}$$

where $\langle (I(i) - I(j))^2 \rangle$ denotes expectation over the image. For traceability the pairwise potential is illustrated in Figure 3.6. In contrast to the level set approach the pairwise potential is contrast sensitive, whereas the internal energy in Equation (3.15) does not include gradient information. A small γ leads to a strong unary term whereas a large γ leads to a weak unary term.

Using the defined unary and pairwise functions, the energy (3.26) is submodular. A function $E : \{0,1\}^n \to \mathbb{R}$ is submodular if and only if, for all label assignments $\mathcal{L}_1, \mathcal{L}_2 \in \{0,1\}^n$, the function satisfies the condition (see [BKR11]):

$$E(\mathcal{L}_1) + E(\mathcal{L}_2) \geq E(\mathcal{L}_1 \vee \mathcal{L}_2) + E(\mathcal{L}_1 \wedge \mathcal{L}_2) . \tag{3.33}$$

For the energy function in Equation (3.26) that has an arity of 2, this simplifies to the condition:

$$E(1,0) + E(0,1) \geq E(1,1) + E(0,0) . \tag{3.34}$$

It is easy to see that the energy function (3.26) fulfills this condition.

Represented as a (network) graph, globally minimizing the submodular energy function corresponds to the problem of finding the minimum cut in the graph [BJ01, KZ04].

The graph $G = (\mathcal{V}_G, \mathcal{E}_G)$, representing the energy function, consists of a set of vertices $\mathcal{V}_G$ and a set of edges $\mathcal{E}_G \subseteq \mathcal{V}_G \times \mathcal{V}_G$. Analogously to [BJ01] the set of vertices is the set of pixels unified with two special vertices, the so-called terminals,

denoting the source S and the sink T of the network. Thus, $\mathcal{V}_G = \mathcal{V} \cup \{S,T\}$, where $\mathcal{V}$ is the set of image pixels.

The set of edges $\mathcal{E}_G$ consists of two different types of edges. First, the so-called neighboring links (n-links), that is the set of edges between neighboring pixels $\mathcal{E}$. Second, there is an edge between each pixel and the source and sink, respectively. These are the so-called terminal links (t-links). Thus, $\mathcal{E}_G = \mathcal{E} \cup \{(S,p),(p,T) \mid p \in \mathcal{V}\}$. The capacities $c(e)$ of all edges $e \in \mathcal{E}_G$ are defined according to Boykov et al. [BJ01] so that the graph represents the energy function. The capacities of the n-links represent the pairwise potential and the capacities of the t-links correspond to the unary potential. They are defined according to Table 3.1.

A cut $\mathcal{C} \subset \mathcal{E}_G$ of the graph G is a set of edges so that $G \setminus C = G_S \cup G_T$, where $S \in G_S$ and $T \in G_T$. That is, removing the edges $\mathcal{C}$ from the graph G partitions the graph into two disjoint sets G_S and G_T separating source and sink. Such a cut is also referred to as s-t-cut. The cost of a cut is defined as the sum of all edge capacities, whose endpoints belong to different sets:

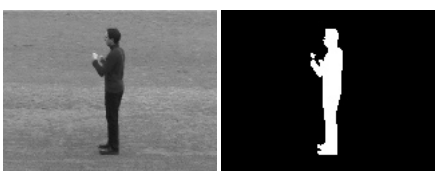
$$|\mathcal{C}| = \sum_{e \in \mathcal{C}} c(e) = \sum_{\substack{(i,j) \in \mathcal{E}_G \\ i \in G_S, j \in G_T}} c(i,j)\,. \tag{3.35}$$

Ford and Fulkerson [FF56] and Elias et al. [EFS56] showed independently, that the problem of finding the minimum cut is equivalent to the problem of finding the maximum flow between S and T. Since the minimum energy state of (3.26) corresponds to the minimum cut of the graph G, standard maximum flow algorithms [BVZ01, BK04, DB08] can be used to solve the labeling problem. Given the maximum flow of a graph, a so-called residual graph is implied. The residual graph is given by all edges that are not saturated by the maximum flow. The minimum cut and thus the labeling is determined by a simple reachability test on the residual graph. Figure 3.7 summarizes this approach and shows some exemplary segmentation results.

An extension to classical graph cuts is the GrabCut algorithm proposed by Rother et al. [RKB04]. This algorithm iteratively uses graph cuts to minimize the energy function and based on the segmentation result, the foreground and background models are refined/relearned.

3.2.2 Probabilistic View

Similarly to the variational energy minimization (see Chapter 3.1.2), the discrete energy minimization can also be interpreted from a probabilistic point of view. Minimizing the discrete energy given in Equation (3.26) is also known as the maximum a posteriori estimation of a Markov Random Field (MAP-MRF problem) [BVZ98, Li09, BKR11]. Markov Random Fields were first introduced into computer vision by



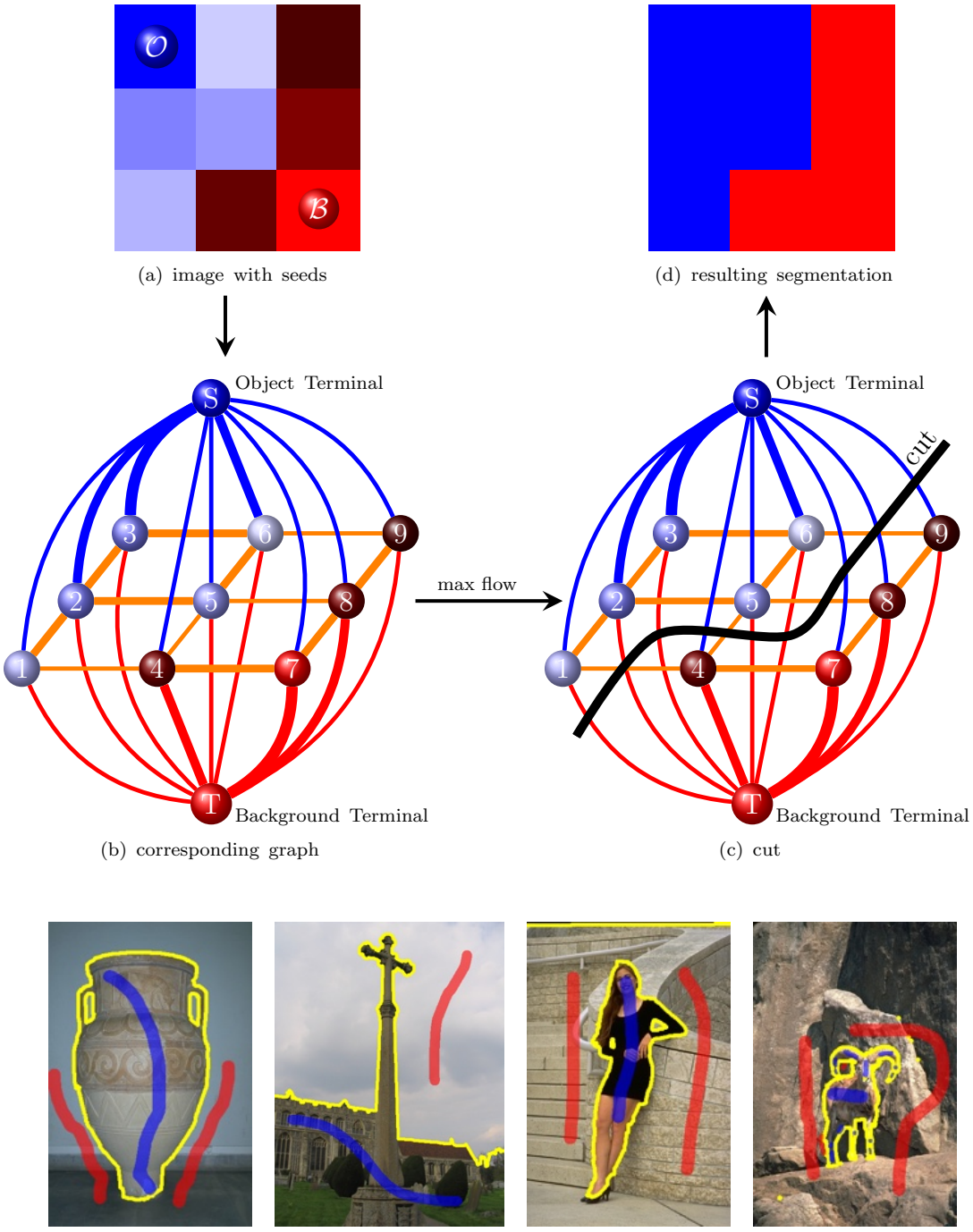


Figure 3.7: Upper part: General workflow of the graph cut framework explained by taking an example 3×3 image (referring to [BFL06]). The seeds are $\mathcal{O} = \{3\}$ and $\mathcal{B} = \{7\}$. The weight of an edge, defined by the boundary term or regional term, is reflected by the edge's thickness. By computing the minimum cost cut a global optimal segmentation is defined. Lower part: Some exemplary results using graph cuts. Foreground and background seeds are visualized in blue and red, respectively.

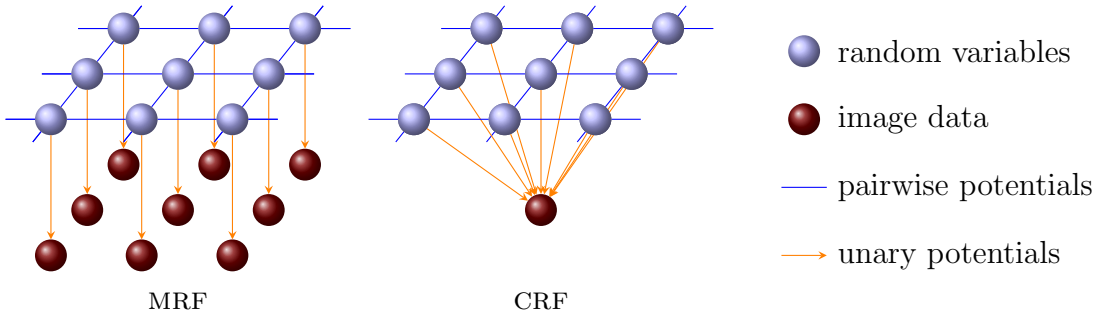


Figure 3.8: Commonly used MRF and CRF modeling the segmentation problem. The random field consists of random variables corresponding to the pixels in an image. If the unary term takes the form of an Ising model, the segmentation problem is modeled as an MRF since the pairwise potentials are independent from the image data. For a contrast sensitive Ising model, modeling the pairwise term, the problem is modeled as a CRF.

Geman and Geman [GG84] and Greig et al. [GPS89] were the first discovering graph cut algorithms from combinatorial optimization by using a graph cut algorithm for restoration of binary images. The problem was formulated as a MAP-MRF problem that required minimization of an energy function similar to (3.26).

A Markov Random Field (MRF) consist of:

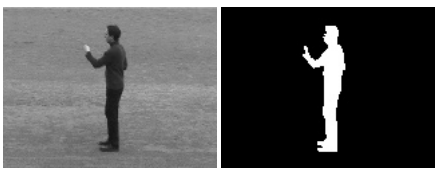
- a set $\mathcal{V} = \{1, \ldots, n\}$ of sites (in this case pixels)
- a neighborhood system $\mathcal{E} = \{\mathcal{N}_p \mid p \in \mathcal{V}\}$, where each $\mathcal{N}_p$ is a set of pixels describing the neighbors of site (pixel) p
- a field (or set) of random variables $\mathcal{L}^n = \{\mathcal{L}_p \mid p \in \mathcal{V}\}$
- a joint distribution $p(\mathcal{L}^m = \mathcal{L}) > 0$

Each random variable $\mathcal{L}_p$ takes a value of the label set $l = \{FG, BG\}$. $p(\mathcal{L}^m = \mathcal{L})$ is the probability of the joint event $\mathcal{L}^m = \mathcal{L}$ where $\mathcal{L} = \{\mathcal{L}_p \mid p \in \mathcal{V}\}$ is a configuration of $\mathcal{L}^n$. In order to fulfill the Markovian property, the random field must satisfy:

$$p(\mathcal{L}_p \mid \mathcal{L}_{\mathcal{V}-\{p\}}) = p(\mathcal{L}_p \mid \mathcal{L}_{\mathcal{N}_p}) \, \forall p \in \mathcal{V} \text{ and } p(\mathcal{L}) > 0 \, \forall \mathcal{L} \in \mathcal{L}^n \, . \tag{3.36}$$

A Conditional Random Field (CRF) can be seen as a MRF globally conditioned on the image data. A graphical model of a pairwise MRF and a pairwise CRF is given in Figure 3.8. The Hammersley-Clifford theorem [HC71] states that the conditional distribution over the random variables of the CRF is a Gibbs distribution:

$$p(\mathcal{L} \mid I) = \frac{1}{Z} \exp\left(- \sum_{c \in C} \tau_c(\mathcal{L}_c) \right) , \tag{3.37}$$



where Z is a normalizing factor and C is the set of all cliques[1] [LMP01]. The corresponding Gibbs energy is given by:

$$E(\mathcal{L}) = -\log p(\mathcal{L} \mid I) - \log Z = \sum_{c \in C} \tau_c(\mathcal{L}_c) \tag{3.38}$$

The most probable or maximum a posteriori labeling $\mathcal{L}^*$ of the random field is defined as

$$\mathcal{L}^* = \arg\max_{\mathcal{L} \in \mathcal{L}^n} p(\mathcal{L} \mid I) \tag{3.39}$$

and can be found by minimizing Equation (3.38). For the binary segmentation problem only up to pairwise clique potentials are nonzero. Thus, the Gibbs energy can be written in the form:

$$E(\mathcal{L}) = \sum_{i \in \mathcal{V}} \tau_i(\mathcal{L}_i) + \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{N}_i} \tau_{i,j}(\mathcal{L}_i, \mathcal{L}_j)\,, \tag{3.40}$$

which is equivalent to the energy given in Equation (3.26).

3.3 Feature Fusion using Dempster's Theory of Evidence

The probabilistic interpretation, of image segmentation using variational (see Chapter 3.1.2) or discrete energy minimization (see Chapter 3.2.2), clarified that it is very important how to model the conditional probabilities or likelihoods, respectively. Due to this question it is important how to fuse likelihoods arising from different feature channels. The classical probability theory assumes independent feature channels and uses Bayes' theorem to get the joint probability by multiplying the likelihoods, see Chapter 3.3.2. In contrast, Dempster's theory of evidence is proposed to fuse information from different feature channels (likelihoods). The basic ideas and all necessary notations are illustrated in the following chapter.

Dempster's theory of evidence, also called Dempster-Shafer theory of evidence or short evidence theory, was first introduced in the late 60s by A.P. Dempster [Dem68], and formalized in 1976 by G. Shafer [Sha76].

This theory is often described as a generalization of the Bayesian theory. It allows to explicitly model inaccuracy and uncertainty information at the same time and to describe conflicts in the information fusion process. In the classical Bayesian framework a probability x for a hypothesis Ψ_1 directly leads to a $(1-x)$ probability that refuses the hypothesis. Mathematically this can be expressed by $p(\Psi_1) + p(\overline{\Psi_1}) = 1$, the so-called additivity rule which results directly from the axioms of Kolmogorov

[1] A clique is a set of random variables x_c that are conditionally dependent on each other

[SV05]. Using different feature channels for the problem of image segmentation the validity of this connectivity is limited. The learned models for different hypotheses in the segmentation framework are only approximations. Thus, making a statement on the occurrence of a hypothesis Ψ_2 from the learned model of hypothesis Ψ_1 is questionable. In fact the probability $(1 - x)$ only represents the inaccuracy of the model, that will be interpreted as the uncertainty.

Using Dempster's theory of evidence to explicitly model uncertainty, conflicting information for a pixel get another meaning. The occurrence of a conflict in that sense means, that different models argue for different labels (hypotheses) for a given pixel.

In the present work the properties of Dempster's theory of evidence to model uncertainty and account for conflicts to fuse different feature channels are utilized. Section 3.3.1 introduces the theoretical background of Dempster's theory of evidence and with the help of simple examples the differences to classical probability theory are emphasized.

The following introduction to Dempster's theory of evidence is based on the paper of G. Shafer [Sha76].

3.3.1 Dempster's Theory of Evidence

The basic idea of the evidence theory is to define a so-called mass function on a hypotheses set Ψ, also called frame of discernment. The hypotheses set Ψ is composed of n single mutually exclusive subsets Ψ_i, which is symbolized by:

$$\Psi = \{\Psi_1, \Psi_2, \ldots, \Psi_n\}, \quad \text{with } \Psi_i \cap \Psi_j = \emptyset \, \forall \, i \neq j \,. \tag{3.41}$$

For the problem of image segmentation the frame of discernment is the set of possible regions, e.g. for a binary segmentation:

$$\Psi = \{\text{FG, BG}\}, \tag{3.42}$$

where FG/BG denotes foreground/background respectively. The power set of Ψ, 2^Ψ or $\wp(\Psi)$ describes the set of hierarchically ordered hypotheses. For the example of binary image segmentation this means: $\wp(\Psi) = \{\emptyset, \text{FG, BG}, \{\text{FG, BG}\}\}$. Thereby the power set includes the impossible hypothesis, the empty set ($\{\emptyset\}$), all hypotheses from the frame of discernment (Ψ_i) and all disjoint combinations of these elements.

Assuming a hypotheses set composed of three single mutually exclusive subsets Ψ_i, the power set consists of seven hypotheses, see first column in Table 3.2.

In order to express a degree of confidence for each element $A \in \wp(\Psi)$ of the power set, an elementary function $m(A)$ is defined. The function $m(A)$ is the so-called *mass function*, or *basic probability assignment* (bpa). Similar to the Bayesian

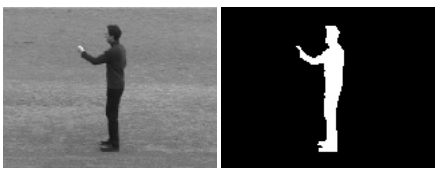


Table 3.2: Example showing the basic elements of Dempster's theory of evidence.

$A \subseteq \wp(\Psi)$	$m(A)$	$Bel(A)$	$Pl(A)$	$Pl(A) - Bel(A)$
$\emptyset$	0	0	0	0
Ψ_1	0.30	0.30	0.55	0.25
Ψ_2	0.25	0.25	0.48	0.23
Ψ_3	0.15	0.15	0.37	0.22
Ψ_1, Ψ_2	0.08	0.63	0.85	0.22
Ψ_1, Ψ_3	0.07	0.52	0.75	0.23
Ψ_2, Ψ_3	0.05	0.45	0.7	0.25
Ψ_1, Ψ_2, Ψ_3	0.10	1.00	1.00	0.00

theory each hypothesis is assigned a probability. Furthermore, in Dempster's theory of evidence each element of the power set is assigned a probability. The function $m : \wp(\Psi) \rightarrow [0,1]$ is a mass function if it fulfills the following conditions:

$$\begin{aligned} &(i) \quad m(\emptyset) = 0 \\ &(ii) \quad \sum_{A \subseteq \Psi} m(A) = 1\,. \end{aligned} \tag{3.43}$$

In this context, $m(A)$ can be interpreted as the belief strictly placed on hypothesis A. Compared to a classical probability function, the totality of the belief is not only distributed on simple classes, but also on composed classes. The modeling shows the impossibility to dissociate several hypotheses, which characterizes the principal advantage of the evidence theory. Besides it is possible to assign a probability to exact one hypothesis or on a set of hypotheses without considering their complement. The assignment of the remaining probability to Ψ can than be interpreted as a degree of ignorance or uncertainty:

$$m(\Psi) = 1 - \sum_{A \subset \Psi} m(A)\,. \tag{3.44}$$

An element $A \in \Psi$ with $m(A) > 0$ is called a focal element. For the aforementioned example with three single mutually exclusive subsets, a possible mass function is defined in Table 3.2.

From the basic probability assignment m, a *belief function* $Bel : \wp(\Psi) \rightarrow [0,1]$ and a *plausibility function* $Pl : \wp(\Psi) \rightarrow [0,1]$ can be defined as

$$Bel(A) = \sum_{A_n \subseteq A} m(A_n)\,, \quad Pl(A) = \sum_{A \cap A_n \neq \emptyset} m(A_n)\,, \tag{3.45}$$

with $A_n \in \wp(\Omega)$. $Bel(A)$, that is, the mass of A itself plus the mass attached to all subsets of A, is interpreted as the total belief committed to hypothesis A. $Bel(A)$

then is the total positive effect the body of evidence has on a value being in A. It quantifies the minimal degree of belief in hypothesis A.

A particular characteristic of Dempster's theory of evidence (one which makes it different from classical probability theory) is that if $Bel(A) < 1$, then the remaining evidence $1 - Bel(A)$ needs not necessarily refute A (i.e., support its negation $\overline{A}$). That is, the so-called additivity rule needs not to be true, e.g. $Bel(A) + Bel(\overline{A}) \leq 1$. Some of the remaining evidence may be assigned to propositions which are not disjoint from A, and hence could be plausibly transferable to A in the light of new information. This is formally represented by the plausibility function $Pl(A)$ (see Equation (3.45)). $Pl(A)$ is the mass of hypothesis A and the mass of all sets which intersect with A, i.e. those sets which might transfer their mass to A. It is the extent to which the available evidence fails to refute A. It quantifies the maximal degree of belief in hypothesis A.

The relation between mass function, belief function and plausibility function is described by:

$$m(A) \leq Bel(A) \leq Pl(A) \quad \forall A \in \wp(\Omega)\,. \tag{3.46}$$

The belief function and the plausibility function define the lower and upper bound of the interval $[Bel(A), Pl(A)]$. The width of this interval can also be interpreted as the uncertainty of hypothesis A. Table 3.2, clarifies the relations between all these functions with a simple example.

Dempster's rule of combination

The Dempster-Shafer theory of evidence has one important operation, *Dempster's rule of combination*, for pooling evidence from a variety of sources. This rule combines two independent sets of basic probability assignments defined over the same frame of discernment. It derives the shared belief between two mass functions and ignores all the conflicting belief through a normalization factor. Let m_1 and m_2 be two mass functions associated with two independent bodies of evidence defined over the same frame of discernment. Mathematically, the combination or joint mass m of these two sets of mass functions is defined by:

$$\begin{aligned} m(\emptyset) &= 0 \\ m(A) &= m_1(A) \otimes m_2(A) \\ &= \frac{\sum\limits_{B \cap C = A} m_1(B) m_2(C)}{1 - K}\,, \text{with } K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)\,. \end{aligned} \tag{3.47}$$

The normalization factor K can be interpreted as conflict between the two mass functions. Dempster's rule of combination computes a measure of agreement between two mass functions concerning various propositions from a common frame of

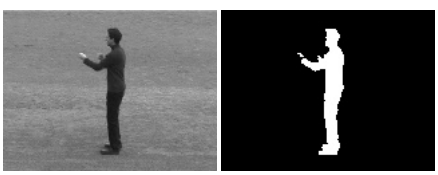


Table 3.3: Example of combing two mass functions into one body of evidence. Light gray cells denote the evidence of the joint mass for hypothesis Ψ_1 and dark gray the evidence of hypothesis Ψ.

$\otimes$		$m_1(A)$			
	$A =$	Ψ_1	Ψ_2	Ψ
$m_2(B)$	B =		0.4	0.2	0.4
	Ψ_1	0.6	0.24	0.12	0.24
	Ψ_2	0.2	0.08	0.04	0.08
	Ψ	0.2	0.08	0.04	0.08

discernment. Since Dempster's rule of combination is associative, fusion of information coming from more than two feature channels is straight forward.

The following example will demonstrate the idea of feature fusion using Dempster's rule of combination. Assume the following two mass functions $m_{1,2}$ defined on the frame of discernment $\Psi = \{\Psi_1, \Psi_2\}$.

$$\begin{aligned} m_1(\Psi_1) = 0.4\,, \quad m_1(\Psi_2) = 0.2\,, \quad m_1(\Psi) = 0.4\,, \\ m_2(\Psi_1) = 0.6\,, \quad m_2(\Psi_2) = 0.2\,, \quad m_2(\Psi) = 0.2\,. \end{aligned} \tag{3.48}$$

The combination using Dempster's rule of combination can be clarified by a so-called combination table, that is given in Table 3.3. For the hypothesis Ψ_1 this yields (light gray cells):

$$\begin{aligned} m(\Psi_1) &= \frac{m_1(\Psi_1)m_2(\Psi_1) + m_1(\Psi_1)m_2(\Psi) + m_1(\Psi)m_2(\Psi_1)}{1 - m_1(\Psi_1)m_2(\Psi_2) - m_1(\Psi_2)m_2(\Psi_1)} \\ &= \frac{0.24 + 0.08 + 0.24}{1 - 0.08 - 0.12} = 0.7\,. \end{aligned} \tag{3.49}$$

Overall, fusing m_1 and m_2 yields the joint mass m:

$$m(\Psi_1) = 0.7\,, \quad m(\Psi_2) = 0.2\,, \quad m(\Psi) = 0.1\,. \tag{3.50}$$

Other possibilities to combine different mass functions are given by Yager's rule [Yag87], that does not use a normalization of the result or Inagaki's rule [Ina91], a parameterized class of combination rules. A survey of these and other combination rules is presented in [SF02].

3.3.2 Relation To Classical Probability Theory

Analogously to Dempster's theory of evidence, a hypothesis set Ψ composed of n single mutually exclusive subsets Ψ_i is assumed (see Equation (3.41)). Every element

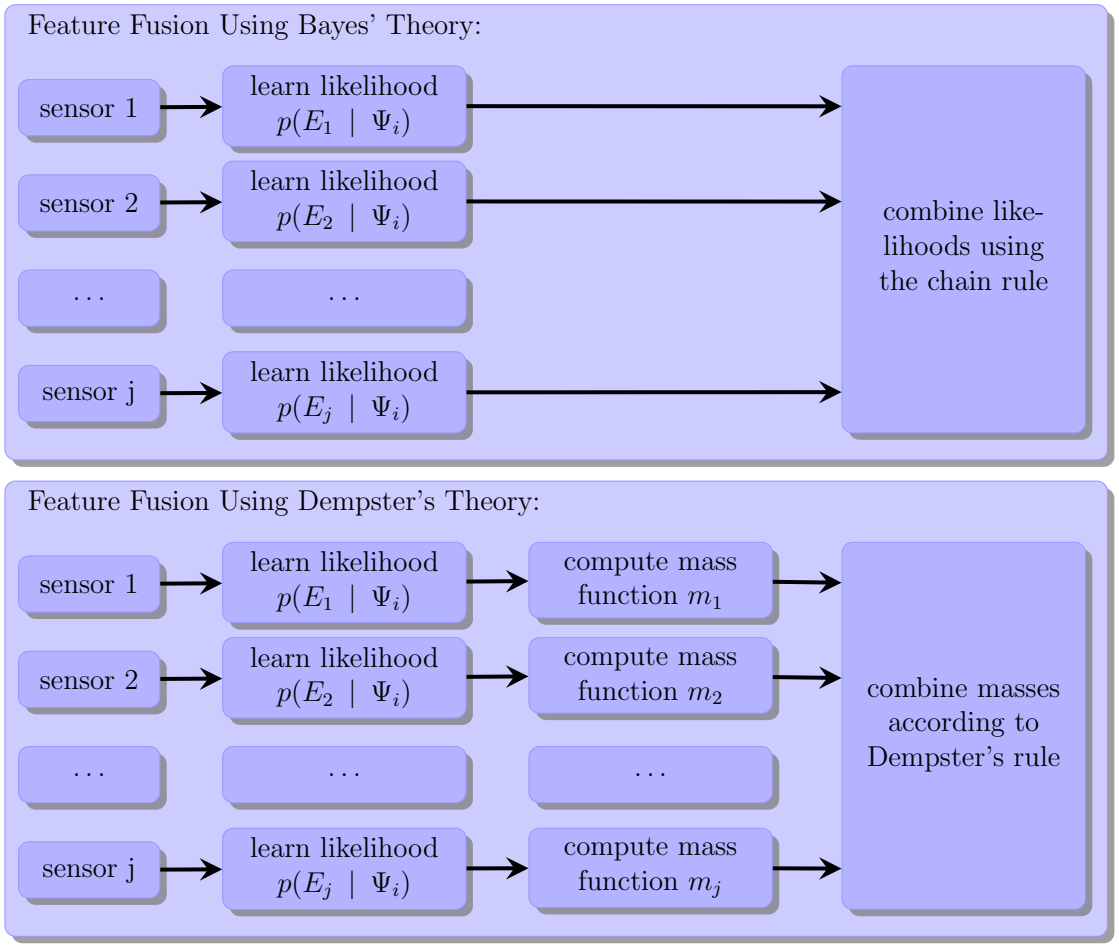


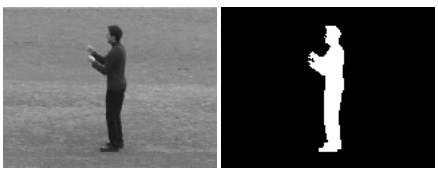
Figure 3.9: Comparison of feature fusion according to Bayes and Dempster-Shafer. Using the same feature models this thesis proposes to use mass functions and Dempster's rule of combination to fuse information from different feature channels.

Ψ_i, e.g. FG and BG can be the result of a sensor E. Using Bayes' theorem yields the conditional probability

$$p(\Psi_i \mid E) = \frac{p(E \mid \Psi_i) \cdot p(\Psi_i)}{\sum_{k=1}^{n} p(E \mid \Psi_k) \cdot p(\Psi_k)}, \tag{3.51}$$

where $p(\Psi_i \mid E)$ denotes the a posteriori probability of hypothesis Ψ_i constrained to sensor E, $p(\Psi_i)$ the a priori probability of hypothesis Ψ_i, $p(E \mid \Psi_i)$ the likelihood of the measurement of sensor E constrained to Ψ_i. The normalization factor (the denominator) is the so-called evidence. Using multiple sensors E_j, Equation (3.51) can be extended to

$$p(\Psi_i \mid E_1 \cap \ldots \cap E_j) = \frac{p(E_1 \cap \ldots \cap E_j \mid \Psi_i) \cdot p(\Psi_i)}{\sum_{k=1}^{n} p(E_1 \cap \ldots \cap E_j \mid \Psi_k) \cdot p(\Psi_k)}, \tag{3.52}$$



which denotes the joined probability of hypothesis Ψ_i constrained to the sensor measurements. Assuming statistically independent sensors, which is usually done in segmentation schemes, the joined probability simplifies to:

$$\begin{aligned} p(\Psi_i \mid E_1 \cap \ldots \cap E_j) =& \frac{p(E_1 \cap \ldots \cap E_j \mid \Psi_i) \cdot p(\Psi_i)}{\sum_{k=1}^{n} p(E_1 \cap \ldots \cap E_j \mid \Psi_k) \cdot p(\Psi_k)} \\ =& \frac{p(E_1 \mid \Psi_i) \cdot \ldots \cdot p(E_j \mid \Psi_i) \cdot p(\Psi_i)}{\sum_{k=1}^{n} p(E_1 \cap \ldots \cap E_j \mid \Psi_k) \cdot p(\Psi_k)}\,, \end{aligned} \tag{3.53}$$

using the chain rule that implies

$$p(E_1 \cap \ldots \cap E_j \mid \Psi_i) = p(E_1 \mid \Psi_i) \cdot \ldots \cdot p(E_j \mid \Psi_i)\,, \tag{3.54}$$

if the sensor measurements are statistically independent. The basic fusion strategies of Bayes and Dempster's theory are depicted in Figure 3.9.

3.3.3 Examples

The following examples will show the differences of combining feature channels with Dempster's rule of combination and classical probability theory. Two independent models p_1 and p_2, that describe the probability for a pixel x belonging to foreground are assumed. They are given by $p_1(\mathrm{FG}) = 0.8$ and $p_2(\mathrm{FG}) = 0.6$. Using Bayesian theory this directly implies probabilities that refuse the hypothesis foreground: $p_1(\mathrm{BG}) = 0.2$ and $p_2(\mathrm{BG}) = 0.4$. Since both models are independent it follows that the combined probability p is given by[2]: $p(\mathrm{FG}) \approx 0.86$ and $p(\mathrm{BG}) \approx 0.14$.

Using Dempster's theory of evidence in the same scenario the following mass functions are defined:

$$\begin{aligned} &m_1(\mathrm{FG}) = 0.8\,, \quad m_1(\mathrm{BG}) = 0\,, \quad m_1(\{\mathrm{FG},\, \mathrm{BG}\}) = 0.2\,, \\ &m_2(\mathrm{FG}) = 0.6\,, \quad m_2(\mathrm{BG}) = 0\,, \quad m_2(\{\mathrm{FB},\, \mathrm{BG}\}) = 0.4\,. \end{aligned} \tag{3.55}$$

That means that, in contrast to Bayesian theory, the probability $1 - p_i(\mathrm{FG})$ is assigned to the disjoint combination of our hypotheses, and quantifies the uncertainty. Using Dempster's rule of combination to compute the joint mass m is visualized in Table 3.4. The joint mass m is then given by:

$$m(\mathrm{FG}) = 0.92\,, \quad m(\mathrm{BG}) = 0\,, \quad m(\{\mathrm{FG},\, \mathrm{BG}\}) = 0.08\,. \tag{3.56}$$

Thus, the combination leads to a strong belief of foreground and some uncertainty while, in contrast to the Bayesian approach, the belief in background is zero. The difference between the Dempster's theory of evidence and the Bayesian framework

[2]The a priori probabilities are assumed to be equally distributed.

Table 3.4: Example one of combing two mass functions into one body of evidence. Light gray cells denote the evidence of the joint mass for the hypothesis FG and dark gray the evidence of hypothesis $\{FG,BG\}$.

$\otimes$		$m_1(A)$			
		$A =$	FG	BG	FG, BG
$m_2(B)$	B =		0.8	0.0	0.2
	FG	0.6	0.48	0.0	0.12
	BG	0.0	0.0	0.0	0.0
	FG, BG	0.4	0.32	0.0	0.08

Table 3.5: Example two of combing two mass functions with different support for one hypothesis into one body of evidence. Light gray cells denote the evidence of the joint mass for the hypothesis FG and dark gray the evidence of hypothesis $\{FG,BG\}$.

$\otimes$		$m_1(A)$			
		$A =$	FG	BG	FG, BG
$m_2(B)$	B =		0.7	0.0	0.3
	FG	0.2	0.14	0.0	0.06
	BG	0.0	0.0	0.0	0.0
	FG, BG	0.8	0.56	0.0	0.24

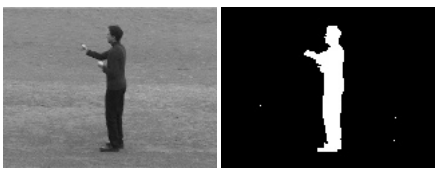
become more clearly in the next example where one model gives very low support to hypothesis FG. The two models are now assumed to be: $p_1(\text{FG}) = 0.7$ and $p_2(\text{FG}) = 0.2$. In classical probability theory the second model implies probability of 0.8 for the pixel to be labeled as background (which is even higher than the probability of model one for foreground). The combined probability is now given by: $p(\text{FG}) \approx 0.37$ and $p(\text{BG}) \approx 0.63$. This means that the combination of two probabilities modeling the foreground results in a high probability for that pixel to be background, which is not intuitive at all.

Using Dempster's rule of combination with mass functions m_1 and m_2 defined analogously to Equation (3.55) the joint mass is given by (see Table 3.5):

$$m(\text{FG}) = 0.76\,, \quad m(\text{BG}) = 0\,, \quad m(\{\text{FG, BG}\}) = 0.24\,. \tag{3.57}$$

Compared to the combination in classical probability theory this result is more intuitive. Given two models describing the probability of a pixel labeled as foreground and combine them with Dempster's theory of evidence results in a probability for foreground and a degree of ignorance given by $m(\{\text{FG, BG}\})$.

The last examples will show how Dempster's theory of evidence will be used to



give more support to significant features in a segmentation framework. Therefore, two foreground and background models estimated on the gray values of an image are assumed. The foreground models $p_{1,1}(I(x))$ and $p_{1,2}(I(x))$ and background models $p_{2,1}(I(x))$ and $p_{2,2}(I(x))$ for gray values $I(x) \in [0,255]$ of pixels $x \in \Omega$ are given by Gaussian distributions:

$$\begin{aligned} p_1\left(I(x) \mid \mathcal{L}_x = FG\right) &= \mathcal{N}(200,10)\,, \quad p_2(I(x) \mid \mathcal{L}_x = FG) = \mathcal{N}(170,10)\,, \\ p_1(I(x) \mid \mathcal{L}_x = BG) &= \mathcal{N}(100,40)\,, \quad p_2(I(x) \mid \mathcal{L}_x = BG) = \mathcal{N}(140,40)\,. \end{aligned} \tag{3.58}$$

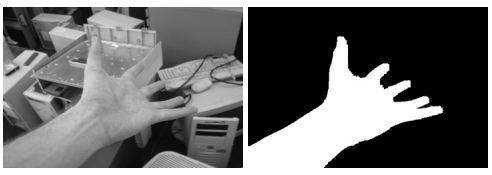
According to the probability distributions the mass functions are defined by:

$$\begin{aligned} &m_1(FG) = m_1(I(x) \mid \mathcal{L}_x = FG) = p_1(I(x) \mid \mathcal{L}_x = FG)\,, \\ &m_1(BG) = m_1(I(x) \mid \mathcal{L}_x = BG) = p_1(I(x) \mid \mathcal{L}_x = BG)\,, \\ &m_1(FG,BG) = m_1(I(x) \mid \mathcal{L}_x = FG,BG) = 1 - (m_1(FG) + m_1(BG)) \\ &m_2(FG) = m_2(I(x) \mid \mathcal{L}_x = FG) = p_2(I(x) \mid \mathcal{L}_x = FG)\,, \\ &m_2(BG) = m_2(I(x) \mid \mathcal{L}_x = BG) = p_2(I(x) \mid \mathcal{L}_x = BG)\,, \\ &m_2(FG,BG) = m_2(I(x) \mid \mathcal{L}_x = FG,BG) = 1 - (m_2(FG) + m_2(BG))\,. \end{aligned} \tag{3.59}$$

Here $m_1(I(x) \mid \mathcal{L}_x = FG)$ describes the mass of pixel x having label FG and $m_1(I(x) \mid \mathcal{L}_x = FG,BG)$ is the mass of pixel x having label FG or label BG. This is the inaccuracy of the model. The combinations using classical probability theory are:

$$\begin{aligned} p(I(x) \mid \mathcal{L}_x = FG) &= p_1(I(x) \mid \mathcal{L}_x = FG) \cdot p_2(I(x) \mid \mathcal{L}_x = FG)\,, \\ p(I(x) \mid \mathcal{L}_x = BG) &= p_1(I(x) \mid \mathcal{L}_x = BG) \cdot p_2(I(x) \mid \mathcal{L}_x = BG)\,. \end{aligned} \tag{3.60}$$

Using Dempster's theory of evidence the combination of m_1 and m_2 leads to the



joint mass m given by:

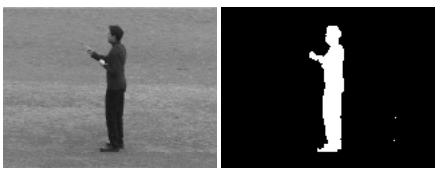
$$
\begin{aligned}
&m(I(x) \mid \mathcal{L}_x = FG) = m_1(I(x) \mid \mathcal{L}_x = FG) \otimes m_2(I(x) \mid \mathcal{L}_x = FG) \\
&= \frac{\sum\limits_{B \cap C = FG} m_1(B) m_2(C)}{1 - K}, \text{ with } K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \\
&= \frac{m_1(FG) m_2(FG) + m_1(FG) m_2(FG,BG) + m_1(FG,BG) m_2(FG)}{1 - (m_1(FG) m_2(BG) + m_1(BG) m_2(FG))}
\end{aligned}
$$

$$
\begin{aligned}
&m(I(x) \mid \mathcal{L}_x = BG) = m_1(I(x) \mid \mathcal{L}_x = BG) \otimes m_2(I(x) \mid \mathcal{L}_x = BG) \\
&= \frac{\sum\limits_{B \cap C = BG} m_1(B) m_2(C)}{1 - K}, \text{ with } K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \\
&= \frac{m_1(BG) m_2(BG) + m_1(BG) m_2(FG,BG) + m_1(FG,BG) m_2(BG)}{1 - (m_1(FG) m_2(BG) + m_1(BG) m_2(FG))}
\end{aligned}
$$

$$
\begin{aligned}
&m(I(x) \mid \mathcal{L}_x = FG,BG) = m_1(I(x) \mid \mathcal{L}_x = FG,BG) \otimes m_2(I(x) \mid \mathcal{L}_x = FG,BG) \\
&= 1 - (m(I(x) \mid \mathcal{L}_x = FG) + m(I(x) \mid \mathcal{L}_x = BG)) .
\end{aligned}
\tag{3.61}
$$

The differences are visualized in Figure 3.10. Here it becomes clear that the joint mass favors mass functions with high support, whereas classical probability theory supports small probabilities. This property enlarges the area of supported foreground in this example.

A more drastically example is shown in Figure 3.11. If the two foreground models are conflicting, the combination with Bayes leads to a joint distribution that is very small. Thus, the fusion using Dempster's theory is much more intuitive.



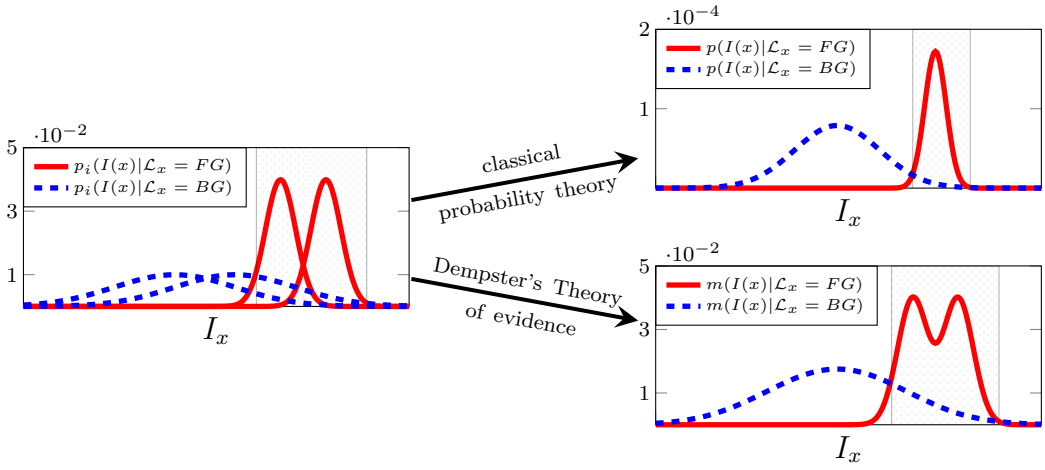


Figure 3.10: Example comparing the results of a feature fusion using classical probability theory and Dempster's theory of evidence. It can be seen, that the feature fusion with Dempster' theory enlarges the area of supported foreground (gray dotted area) when comparing it to classical fusion.

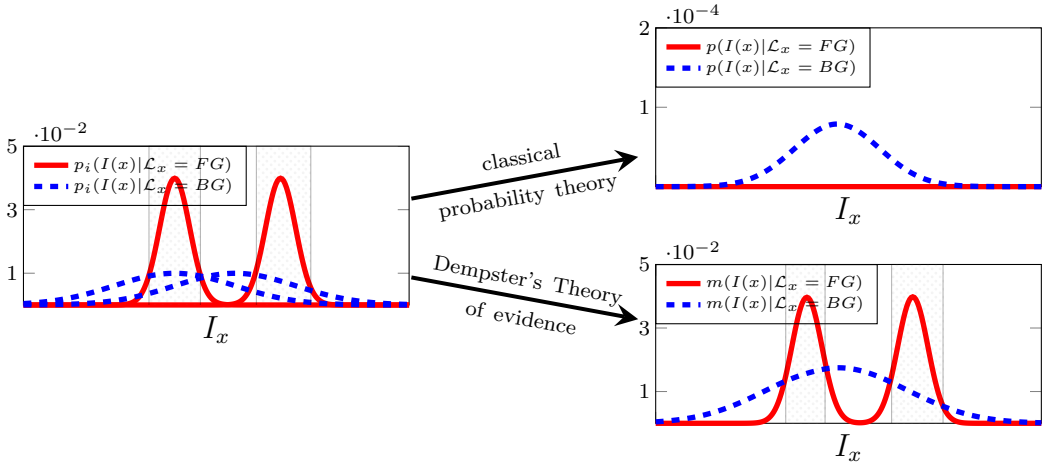
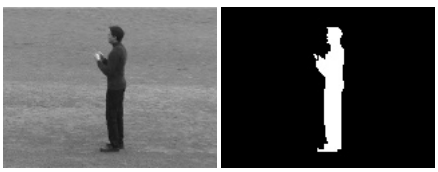


Figure 3.11: Example feature fusion of two conflicting models. It can be seen from the supported foreground area (gray dotted area), that Dempster's theory of evidence leads to a more intuitive result.

Chapter 4

Dempster's Theory of Evidence for Variational Image Segmentation



Bayes Dempster-Shafer Bayes Dempster-Shafer

Figure 4.1: Segmentation results comparing the probabilistic Bayesian model or Dempster's theory of evidence to fuse separated RGB color models.

In many image segmentation scenarios, multiple information from different sensors are available. Examples include: (i) separated color information, e.g. RGB-Color, HSV-Color or CieLab-Color [ZY96]; (ii) Texture information, e.g. the nonlinear structure tensor or Gabor filtering [RBD03, CRD07] ; (iii) motion information [CS05]; (iv) shape knowledge [MSV95];(v) depth information [KCB+05] or (vi) thermographic information. Thus, the need for an accurate and elegant fusion method is given. As stated earlier, Dempster's theory of evidence provide such a feature fusion and compared to the Bayesian fusion it is easy to model inaccuracy and uncertainty within this framework.

This chapter shows how to fuse the available information using mass functions and how the uncertainty of a sensor (feature channel) is modeled. After the feature fusion, the joint mass is integrated into a variational energy minimizing framework. Furthermore, the framework is extended by means of user interactions and other segmentation methods are integrated as additional information sources. This chapter is directly based on the publications [SR10, SR11a].

4.1 Energy Function including Dempster's Theory of Evidence

To introduce the proposed method, using the Dempster-Shafer evidence theory to fuse information arising from different feature channels, the following example is considered. Let $I : \Omega \to \mathbb{R}^2$ be a vector image with two feature channels ($I(x) =$

$(I_1(x),I_2(x)))$ and φ^t with $t \geq 0$ a contour dividing the image into foreground and background. To minimize the energy functional (3.19), foreground and background likelihoods need to be learned for both feature channels. E.g., for a full Gaussian density the mean $\mu_{i,j}$ and standard deviation $\sigma_{i,j}$ (its variance is therefore $\sigma_{i,j}^2$) need to be learned. Thus, the likelihoods are defined by:

$$p(I_j(x) \mid \mathcal{L}_x = i) = p_{i,j}(I(x)) = \frac{1}{\sqrt{2\pi}\sigma_{i,j}} e^{-\frac{(I_j(x)-\mu_{i,j})^2}{2\sigma_{i,j}^2}}, \tag{4.1}$$

for $i,j \in \{1,2\}$, where $i = 1$ defines foreground probabilities and $i = 2$ background probabilities, j characterize the feature channel and $\mathcal{L}_x \in \{FG,BG\}$ denotes the label of pixel x. Using the Bayesian model and disregarding the smoothness term in Equation (3.19), a pixel is defined as foreground if (see Equations (3.19) and (3.18))

$$\begin{aligned} & \sum_{i=1}^{2} \log\left(\frac{p_{1,i}}{p_{2,i}}\right) > 0 \\ \Leftrightarrow\ & \log(p_{1,1}) + \log(p_{1,2}) > \log(p_{2,1}) + \log(p_{2,2}) \\ \Leftrightarrow\ & \log(p_{1,1} \cdot p_{1,2}) > \log(p_{2,1} \cdot p_{2,2}) \\ \Leftrightarrow\ & p_{1,1} \cdot p_{1,2} > p_{2,1} \cdot p_{2,2} \\ \Leftrightarrow\ & p(I(x) \mid \mathcal{L}_x = FG) > p(I(x) \mid \mathcal{L}_x = BG)\,. \end{aligned} \tag{4.2}$$

In other words, the joint probability for foreground needs to be bigger than the joint probability for background.

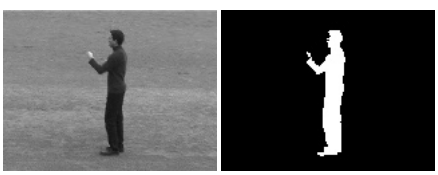
The proposed method uses the Dempster-Shafer theory of evidence to fuse the information coming from different feature channels. Therefore, appropriate mass functions $m_i : \wp(\Psi) \to [0,1]$ for each feature channel i have to be defined. In case of a two-phase segmentation, the frame of discernment becomes $\Psi = \{FG, BG\}$. Dempster's rule of combination, see Equation (3.47), is used to fuse the two bodies of evidence. The joined mass function $m = m_1 \otimes m_2$, represents a measure of agreement between both mass functions.

For a two-phase segmentation, the total belief committed to a focal element Ψ_i is equal to the belief strictly placed on Ψ_i. Thus it holds:

$$Bel(\Psi_i) = m(\Psi_i) \text{ for } \Psi_i \in \{FG, BG\} \text{ and } Bel(\Psi) = 1\,. \tag{4.3}$$

Using the total belief committed to a focal element and disregarding the smoothing term, a pixel should be foreground if

$$Bel(FG) > Bel(BG) \Leftrightarrow \log\left(\frac{Bel(FG)}{Bel(BG)}\right) = \log\left(\frac{m(FG)}{m(BG)}\right) > 0\,. \tag{4.4}$$



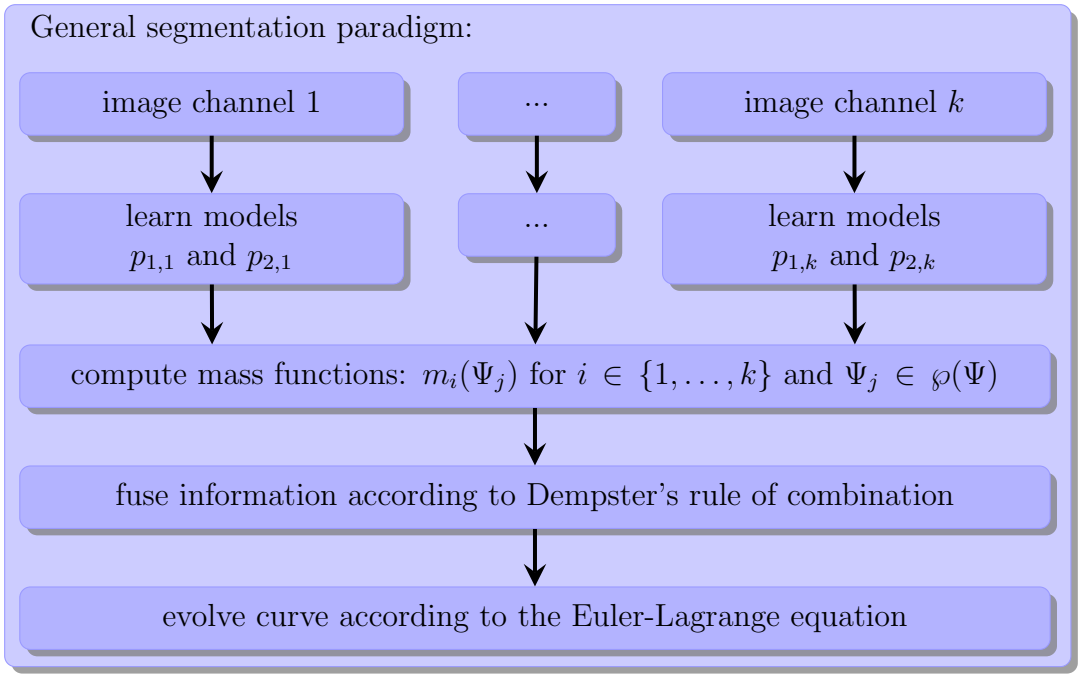


Figure 4.2: Proposed image segmentation paradigm based on combining the variational framework and Dempster's theory of evidence.

In other words, a pixel should be foreground if the total belief committed to foreground is bigger than the total belief committed to background.

In the same context, the plausibility function $Pl(\Psi_i)$, which quantifies the maximal degree of belief of a hypothesis Ψ_i should define a pixel as foreground if

$$\log\left(\frac{Pl(FG)}{Pl(BG)}\right) = \log\left(\frac{m(FG)+m(\Psi)}{m(BG)+m(\Psi)}\right) > 0\,. \tag{4.5}$$

Obviously, the following relation holds:

$$\log\left(\frac{Pl(FG)}{Pl(BG)}\right) > 0 \Leftrightarrow \log\left(\frac{Bel(FG)}{Bel(BG)}\right) > 0\,, \tag{4.6}$$

and the uncertainty $m(\Psi)$ can be interpreted as a smoothing term.

Therefore, the proposed method uses the total belief committed to foreground or background regions. To define appropriate mass functions, the learned conditional likelihoods $p_{i,j}$ are used (the same likelihoods are also used in the Bayesian approach). The proposed mass functions are explained in more detail in the following section. Figure 4.2 shows the general segmentation paradigm to evolve a contour combining the variational framework and Dempster-Shafer evidence theory.

Assuming k different image channels, first a foreground $p_{1,j}$ and a background $p_{2,j}$ model is learned for each channel $j \in \{1, \dots k\}$. Based on this learned conditional likelihoods for each channel, appropriate mass functions are computed. The k different mass functions are fused using Dempster's rule of combination, see Equation (3.47). The joint mass m is included in the energy function and finally, the curve is evolved according to minimize the Euler-Lagrange equation.

The next step includes the joint mass of all feature channels in the energy function by replacing the conditional likelihoods (joint probabilities). In general, let $I : \Omega \Rightarrow \mathbb{R}^k$ be a vector image $(I(x) = (I_1(x), \dots, I_k(x)))$ with k feature channels. Assuming independent feature channels and using the Bayesian model, the total a posteriori probability of a region is the product of the separated conditional probabilities (see Equations (3.18) and (3.53)). With Dempster's theory of evidence, the information arising from the k feature channels is fused using Dempster's rule of combination, resulting in the joint mass:

$$m = m_1 \otimes m_2 \otimes \dots \otimes m_k \,, \tag{4.7}$$

where m_j models the mass of feature channel j. Now, the mass committed to a region is used as the conditional likelihood of a region. Thus, the new energy functional of the proposed method is given by:

$$\begin{aligned} E(\varphi) = &- \int_\Omega H(\varphi) \log m(FG)\, dx - \int_\Omega (1 - H(\varphi)) \log m(BG)\, dx \\ &+ \nu \int_\Omega |\nabla H(\varphi)|\, dx \end{aligned} \tag{4.8}$$

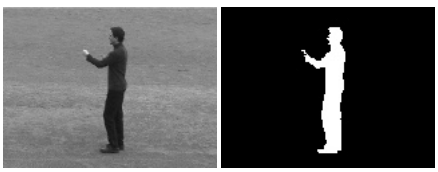
and yields the following Euler-Lagrange equation:

$$\frac{\partial \varphi}{\partial t} = \delta(\varphi) \left[\log \left(\frac{m(FG)}{m(BG)} \right) + \nu \operatorname{div} \left(\frac{\nabla \varphi}{|\nabla \varphi|} \right) \right] . \tag{4.9}$$

From a probabilistic point of view, the conditional likelihood is now defined by the joint mass instead of the joint probability.

As shown in Chapter 3.3, the advantage of Dempster's theory of evidence is the fact, that it supports large masses (or probabilities) whereas the Bayesian model tends to supports small probabilities. One advantage of the Bayesian model is that a pixel is assigned to Ψ_i, if $\exists k$ with $p_{j,k} \approx 0$ for $i \neq j$. This advantage can easily be integrated in our framework but it contradict our interpretation of supporting large likelihoods.

Other properties of the Bayesian model, e.g., to include shape priors [RP02], can be transferred straightforward to the proposed framework by adding terms to the new energy functional (4.8).



4.1.1 Defining Appropriate Mass Functions

Given the novel energy function, see Equation (4.8), including the joint mass for different feature channels, appropriate mass functions m_j for each feature channel need to be defined. Intuitively, the likelihoods for foreground and background, learned from the image, should be used to define the mass functions. Therefore, a simple definition would be:

$$\begin{aligned} &m_j(\mathcal{L}_x = i) = p(I_j(x) \mid \mathcal{L}_x = i) = \frac{p_{i,j}(I(x))}{p_{FG,j}(I(x)) + p_{BG,j}(I(x))} \text{ for } i \in \{FG,BG\}\,, \\ &m_j(\Psi) = m_j(\emptyset) = 0\,. \end{aligned} \tag{4.10}$$

This definition totally ignores the possibility of modeling uncertainty and inaccuracy by setting $m_j(\Psi) = 0$. Furthermore, a fusion with Dempster's rule of combination would be equivalent to Bayesian probability fusion, since the joined mass m fulfills:

$$m(FG) > m(BG) \Leftrightarrow p(I(x) \mid \mathcal{L}_x = FG) > p(I(x) \mid \mathcal{L}_x = BG)\,. \tag{4.11}$$

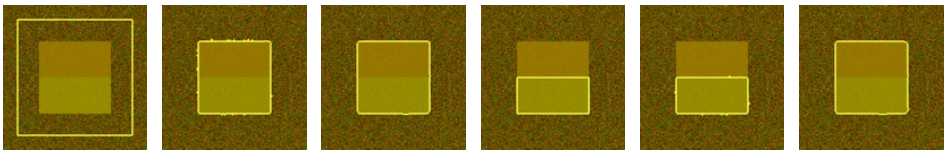
The idea is to use the union of the conditional likelihoods $p(I_j(x) \mid FG \cup BG) = p(I_j(x) \mid FG) + p(I_j(x) \mid BG)$ as a measurement of uncertainty. This means, that the mass function of a feature channel will have a large uncertainty if both conditional likelihoods are small. Thus, the influence of such a mass function will be minimized. The proposed mass function m_j of feature channel j is defined by:

$$\begin{aligned} &m_j(\emptyset) = 0\,, \\ &m_j(FG) = m_j(\mathcal{L}_x = FG) = p(I_j(x) \mid \mathcal{L}_x = FG)\,, \\ &m_j(BG) = m_j(\mathcal{L}_x = BG) = p(I_j(x) \mid \mathcal{L}_x = BG)\,, \\ &m_j(\Psi) = 1 - (p(I_j(x) \mid \mathcal{L}_x = FG) + p(I_j(x) \mid \mathcal{L}_x = BG))\,. \end{aligned} \tag{4.12}$$

In practice, the defined mass functions fulfill Eq. (3.43): $\sum_{A \subseteq \Psi} m(A) = 1$. For the case that the sum of likelihoods is bigger than one, $p(I_j(x) \mid \mathcal{L}_x = FG) + p(I_j(x) \mid \mathcal{L}_x = BG) > 1$, $m_j(\Psi)$ is set to zero and the other terms are normalized to one.

To include prior knowledge about the accurateness of a feature channel, a weighting parameter λ_j can be integrated to control the influence of a mass function m_j to the joint mass m. The mass function of feature channel j is then defined by:

$$\begin{aligned} &m_j(\emptyset) = 0\,, \\ &m_j(\Psi) = \frac{\lambda_j \cdot (1 - (p(I_j(x) \mid \mathcal{L}_x = FG) + p(I_j(x) \mid \mathcal{L}_x = BG)))}{K}\,, \\ &m_j(FG) = (1 - m_j(\Psi)) \cdot \frac{p(I_j(x) \mid \mathcal{L}_x = FG)}{p(I_j(x) \mid \mathcal{L}_x = FG) + p(I_j(x) \mid \mathcal{L}_x = BG)}\,, \\ &m_j(BG) = (1 - m_j(\Psi)) \cdot \frac{p(I_j(x) \mid \mathcal{L}_x = BG)}{p(I_j(x) \mid \mathcal{L}_x = FG) + p(I_j(x) \mid \mathcal{L}_x = BG)}\,, \end{aligned} \tag{4.13}$$



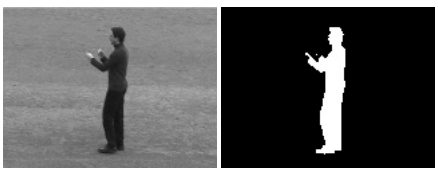
(a) (b) (c) (d) (e) (f)

Figure 4.3: (a) and (d): Two different initializations for the segmentation methods, (b) and (e): final segmentations using the Bayesian model, (c) and (f): final segmentations using Dempster-Shafer evidence theory. While the proposed method converges to almost identical results, the Bayesian model get stuck in two different local minima, due to one inadequate foreground histogram. Evaluating the energy of the results shows that the energy of (e) is bigger than the other ones, which means that this is a local minimum.

where K is an additional normalization term that can be used to force $\sum_j m_j(\Psi) = 1$, so that the sum of modeled uncertainty is one. The parameters $\lambda_j \in [0,1]$ for each channel j are small, if the channel is accurate and high otherwise. Choosing $\lambda_j = K = 1$ for all feature channels leads to the same definition given in Equation (4.12). Integrating prior knowledge can be useful in situations where a sensor is more meaningful than another one. An example is shown in Chapter 5.3, where depth and color information are fused using Dempster's theory of evidence. For the following experiments no prior knowledge is assumed.

4.1.2 Experimental Results

Chapter 4.1 introduced a segmentation method that integrates the Dempster-Shafer theory of evidence into a variational segmentation framework. To show the effect of the new approach, some results of the proposed method are presented and compared to segmentation results using the traditional Bayesian framework to fuse information arising from different feature channels. The method is evaluated on real images taken from the Berkeley segmentation dataset [MFTM01] as well as on synthetic textured images from the Prague texture segmentation data-generator and benchmark [HM08]. A 2nd order Runge-Kutta method was used to solve the partial differential equation and minimize the proposed energy functional, since it has been shown that it outperforms the Euler method [SR09]. In most of the experiments the CieLab-Color channels are used. This implies that, besides the experiments on the textured images, no additional information such as texture or shape priors are used in the experiments. To demonstrate the advantages of the proposed framework texture features such as the nonlinear structure tensor are additionally used for the experiments on the synthetic textured images.



A situation in which the advantages of the proposed method become apparent is the synthetic example shown in Figure 4.3. While the segmentation using the Bayesian model computes different segmentation boundaries for the two initializations, the proposed methods converges to almost identical results for both initializations. The reason for the different segmentation results with the initialization in Figure 4.3e is an inadequate foreground statistic for one of the two feature channels (red and green). For the upper part of the object, that has to be segmented, this inadequate feature channel (green) supports the foreground with very small probabilities. More formally:

$$\exists j \in \{1,2\} \mid p(I_j(x) \mid \mathcal{L}_x = FG) \approx 0 \quad \forall x \in A\,, \tag{4.14}$$

where A describes the upper part of the object. The Bayesian approach yields

$$p_1(I(x)) = \prod_j p(I_j(x) \mid \mathcal{L}_x = FG) \approx 0\,. \tag{4.15}$$

Because of the noisy background, these statistics have a larger standard deviation than the foreground statistics. For the example this results in

$$\begin{aligned} &p_{1,1}(I(x)) \in [0.024; 0.040]\,,\ p_{1,2}(x) \approx 0\,, \\ &p_{2,1}(I(x)) \in [0.006; 0.004]\,,\ p_{2,2}(x) \in [0.007; 0.006]\,, \end{aligned} \tag{4.16}$$

for $x \in A$, which results in

$$p_2(I(x)) = \prod_j p(I_j(x) \mid \mathcal{L}_x = BG) > p_1(I(x)) \quad \forall x \in A\,. \tag{4.17}$$

Apparently the first feature channel (red) is not considered, even if it supports the foreground with the highest probability. Using the mass functions defined in Equation (4.12) and fuse them with Dempster's rule of combination, this problem is solved elegantly, since in the aforementioned case it yields

$$Bel(FG) > Bel(BG) \quad \forall x \in A\,. \tag{4.18}$$

This may be explained by the fact, that the Dempster-Shafer evidence theory assigns more support to high probabilities whereas the Bayesian model more strongly supports small probabilities.

The same effect of the evidence theory can be observed in real images, where the Bayesian model does not segment parts of the object, because one of the feature channel is close to zero (e.g. Figure 4.1, or the tent in Figure 4.6). Conversely, the proposed method converges to a good segmentation in these regions because the other feature channels strongly support the foreground region. The effect of a feature channel close to zero can also be seen in Figure 4.5. Analyzing the upper

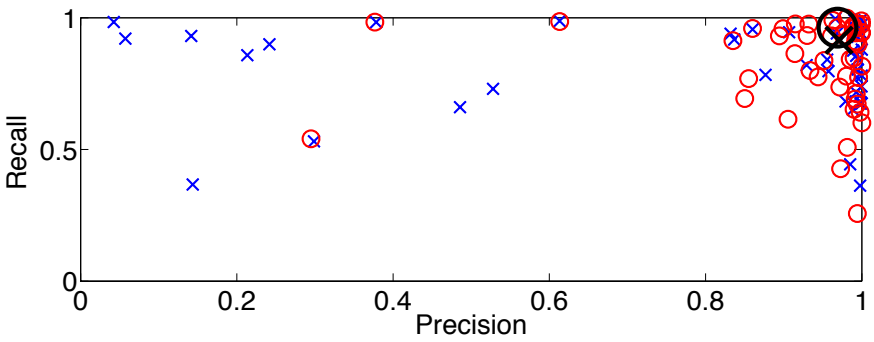


Figure 4.4: Precision-Recall-Diagram. Circles mark the performance using Dempster-Shafer theory of evidence, x marks the performance using Bayes. The black marks mark the mean performance (Dempster-Shafer: 0.93, 0.83; Bayes: 0.81, 0.82).

part of the flower, the proposed method converges to a good segmentation because of this effect. In some situation this effect can also lead to better results using the Bayesian approach (see lower left part). Visually the Bayesian model leads to a better segmentation. However, analyzing the statistical models the segmentation result of the proposed result is more meaningful.

To evaluate and demonstrate the impact of the novel feature fusion, 47 images from the Berkeley segmentation dataset [MFTM01] were chosen. The restriction to this small subset is due to practical reasons, because most of the other images are not suited for binary variational image segmentation. To measure the performance, *precision* and *recall* are calculated for each of the images, which are defined by:

$$Precision = \frac{|G \cap R|}{|G \cap R| + |R \setminus G|}, \quad Recall = \frac{|G \cap R|}{|G \cap R| + |G \setminus R|}, \tag{4.19}$$

where G is the ground-truth foreground object and the region R is the object-segment derived from the segmentation. A precision value near one means that only a few non object regions are segmented and a recall value near one means that most object regions are segmented. Thus, a perfect segmentation has a precision and recall value near one. The ground-truth foreground is the manual segmented foreground from [MO10], which is publicly available. For both frameworks the same manually selected initialization is used for each image, which was typically a rectangle inside the object. Manually initializations are used, since both methods find local minimizers of the given energy functional, thus a manual initialization yields more reliable results. The regularization parameter was chosen slightly different between the frameworks but remained the same for all images. The probability densities where modeled as nonparametric Parzen estimates [RBD03]. In these examples,

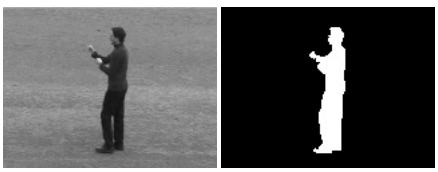


Figure 4.5: Image taken from the Berkeley dataset [MFTM01]. The left image shows the segmentation result using the Bayesian model to fuse the CieLab-Color channels, the right image shows the result using Dempster-Shafer theory of evidence to fuse the information. Differences in the segmentation result are highlighted.

modeling the densities as multivariate Gaussian-Mixture-Models for the Bayesian approach had no positive effect.

The performance of all images is shown in Figure 4.4. Using the Bayesian framework to fuse the information leads to a mean precision of 0.81 and a mean recall of 0.82 while the mean precision using the proposed Dempster-Shafer theory is 0.92 and the mean recall is 0.83.

Figures 4.1, 4.5 and 4.6 show some example segmentations on images from the Berkeley segmentation dataset. Using the proposed method to fuse the different information helps to segment and separate much better the semantically interesting and different regions by searching for features that support the foreground region or the background region. Analyzing the lower right Example of Figure 4.6 shows, that the Bayesian model segment parts of the swamp, because one channel does not support the background. However, in some situations the proposed model leads to slightly worse segmentations. E.g. analyzing the tail or the pecker of the bird, some parts are not segmented due to the smoothness term.

Example segmentations integrating texture features are given in Figure 4.7. The texture features used here are defined by the structure tensor:

$$J_\sigma = K_\sigma * (\nabla I \nabla I^T) = \begin{pmatrix} K_\sigma * I_{x_1}^2 & K_\sigma * I_{x_1} I_{x_2} \\ K_\sigma * I_{x_1} I_{x_2} & K_\sigma * I_{x_2}^2 \end{pmatrix} . \tag{4.20}$$

Thus, the structure tensor is defined by the spatial derivatives smoothed by a Gaussian kernel K_σ with standard deviation σ [DZ86, BGW91, CRD07].

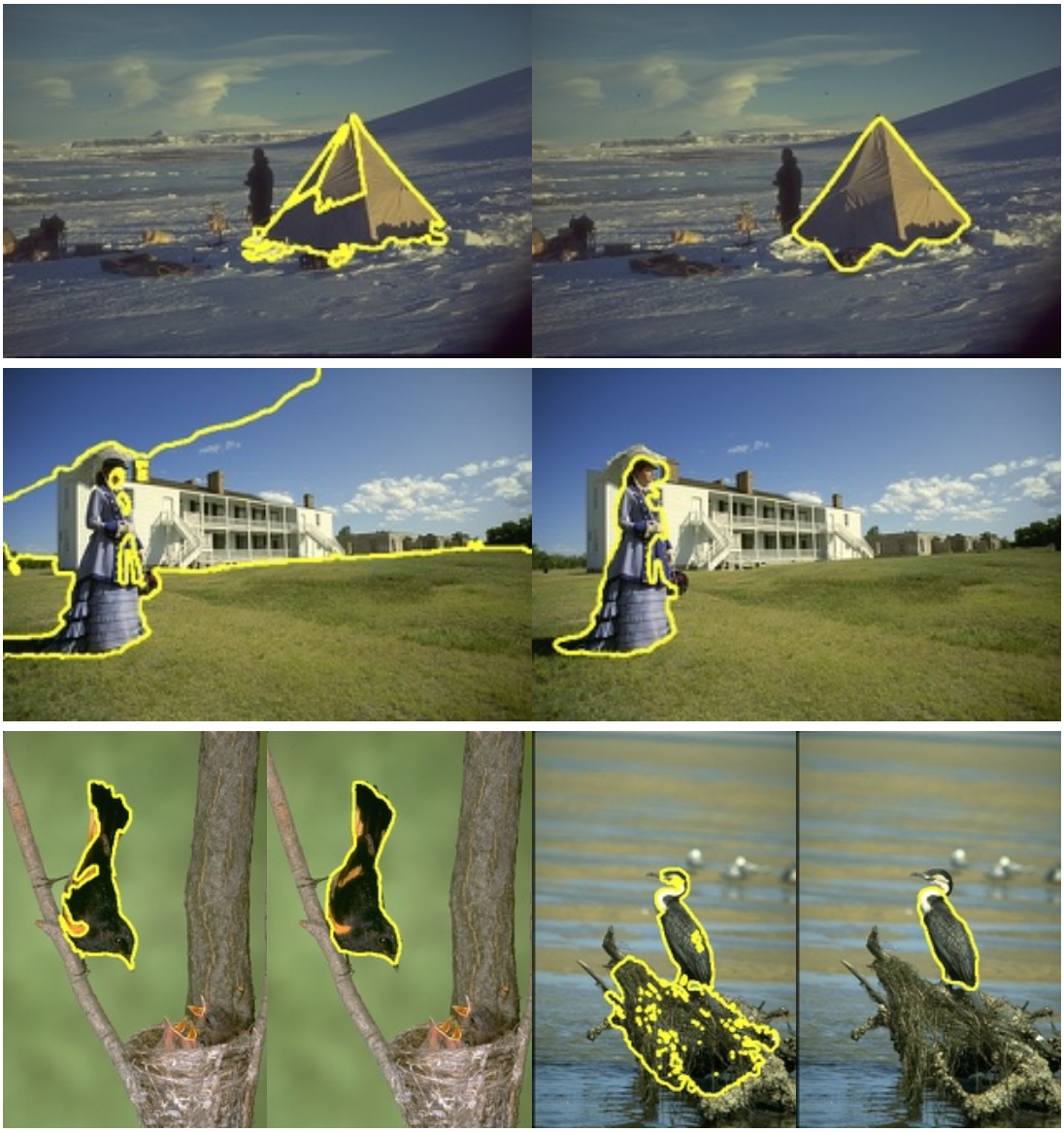
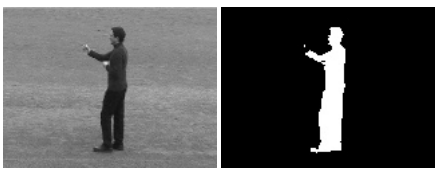


Figure 4.6: Images taken from the Berkeley dataset [MFTM01]. The left image shows the segmentation results using the Bayesian model to fuse the CieLab-Color channels, the right image shows the results using Dempster's theory of evidence to fuse the information. Each example shows that the evidence theory can help to improve the segmentation.



Bayes Dempster-Shafer Bayes Dempster-Shafer

Figure 4.7: Segmentation results using either the probabilistic Bayesian model or Dempster's theory of evidence to fuse color and texture information.

Again the proposed method separates the interesting regions much better than the Bayesian approach.

In [DPTA+13], Derraz et al. proposed to use a minimization scheme based on the Split Bregman method [GBO10] to globally optimize the energy functional. They used the same definition of mass functions given in Equation (4.12) and showed that for some images a global minimum outperforms the proposed local segmentation scheme. However, in some situations a global minimum does not yield the desired results. This is due to the fact, that the model assumptions are not always fulfilled. Therefore, additional user priors are suitable to be included in the segmentation scheme. The next chapter will show how such user priors, interpreted as hard constraints are integrated into the proposed variational segmentation method.

4.2 Interactive Variational Image Segmentation

Most existing level set methods [CV01, RP02, CRD07], e.g. the one proposed in the previous chapter, are not qualified as an interactive segmentation tool. The corresponding initial value problem propagates the region boundary to a local minimum of the continuous energy function without allowing the user to correct the final segmentation result. Thus, the only user interaction affecting the segmentation result is providing an initial curve, e.g. a rectangle around the object of interest. In contrast to the variational approaches, discrete energy minimization segmentation frameworks such as graph cut approaches [SM00, BJ01, RKB04] provide an elegant way to treat user interaction to guide or correct the segmentation process.

A simple rule-based reasoning is usually used to integrate user interaction into the segmentation process:

- if the user marks a pixel as an object, then it is forced to be object,

Figure 4.8: Segmentation results using the proposed user-interactive segmentation framework. The left example is a James Bond photo from the Internet[1]and the right one is from the Berkeley Segmentation Database [MFTM01].

- on the contrary a pixel is background if the user marks it as background,
- pixels that are not marked by the user are either foreground or background.

These so-called hard constraints can also be found in [BJ01, RKB04, CFRA07]. These methods learn foreground and background appearance based on the given user information, whereas variational approaches use the current curve to learn these models.

4.2.1 Integration of User Constraints

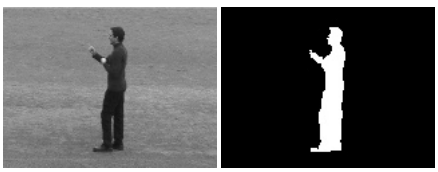
Level set methods for interactive organ segmentation have been proposed earlier by Cremers et al. [CFRA07]. They developed a statistical framework integrating user interactions. In addition to an initial boundary they provide a framework, where the user is able to mark object and background regions in terms similar to a shape prior. Thus the user can indicate which areas are likely to be part of the object or the background. However, the foreground and background appearances are not learned from the given user information. In contrast, the proposed framework based on Dempster's theory of evidence which actually also uses the appearance information contained in the user defined regions.

The evolving boundary is directly driven by the following three terms:

- the intensity information contained in the image [ZY96],
- the user labeling in terms similar to a shape prior [CFRA07] and
- the intensity information of the user labeling ([BJ01, SR11a]).

Due to these terms, user defined regions have a global influence on the segmentation, whereas other frameworks only allow local refinement of the segmentation (e.g. [CFRA07, BJ01]).

[1] `http://www.moviepilot.de/files/images/0005/1919/James_Bond_article.jpg`



Analogue to Cremers et al. [CFRA07] a given image $I : \Omega \to \mathbb{R}^n$ and a user input L, marking certain image locations as object or background regions, are assumed.

$$L : \Omega \to \{-1,0,1\}\,, \tag{4.21}$$

where the label values reflect the user input:

$$L(x) = \begin{cases} 1, & x \text{ marked as object,} \\ -1, & x \text{ marked as background,} \\ 0, & x \text{ not marked.} \end{cases} \tag{4.22}$$

This user-defined labeling $L(x)$ needs to be integrated as a constraint into the energy function (4.8). This is realized by the following novel energy function:

$$E(\varphi) = E_{image}(\varphi) + E_{curve}(\varphi) + \underbrace{E_{user}(\varphi)}_{new} \tag{4.23}$$

where the user term E_{user} consists of two different terms:

$$E_{user} = \nu_2 \cdot E_{user-shape} + E_{user-image}\,. \tag{4.24}$$

The first term of this *user-defined* energy function is defined similar to [CFRA07] by

$$E_{user-shape} = -\frac{1}{2}\int_\Omega L_\sigma(x)\text{sign}(\varphi(x))\,dx\,, \tag{4.25}$$

with a Gaussian smoothed label function

$$L_\sigma(x) = \int_\Omega L(x)K_\sigma(x)\,dx\,. \tag{4.26}$$

Thus, the label function is smoothed by a Gaussian kernel K_σ with standard deviation σ:

$$K_\sigma(x) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{(x_1-x_2)^2}{2\sigma^2}\right)\,. \tag{4.27}$$

This model has two free parameters ν_2 and σ which can be interpreted as follows. The parameter ν_2 provides the overall weight of the user interaction and determine how strongly the user input will affect the segmentation. The parameter σ defines the spatial range within which a point labeled as object or background will affect the segmentation. It can therefore be interpreted as a *brush size*.

Having in mind, that the sign function can be expressed using the Heaviside function $\text{sign}(\varphi(x)) = 2H(\varphi(x)) - 1$ the user-shape energy can be written in the form

$$E_{user-shape} = -\int_\Omega L_\sigma(x)H(\varphi(x))\,dx\,, \tag{4.28}$$

since the term $-\frac{1}{2}L_\sigma(x)$ is independent from $\varphi(x)$.

The novel second term $E_{user-image}$ of E_{user} is inspired by the image energy E_{image} and is defined as follows:

$$\begin{aligned} E_{user-image} = &\int_\Omega H(\varphi) \log m_{user}(FG)\, dx \\ &- \int_\Omega (1 - H(\varphi)) \log m_{user}(BG)\, dx\,. \end{aligned} \tag{4.29}$$

The mass function m_{user} is, in contrast to the mass function m_{image}, defined by the marked regions while the function m_{image} is defined by the image regions divided by the curve. Finally, the terms $E_{image}(\varphi)$ and $E_{curve}(\varphi)$ are given by (see Equation (4.8)):

$$\begin{aligned} E_{image}(\varphi) = &- \int_\Omega H(\varphi) \log m_{image}(FG)\, dx \\ &- \int_\Omega (1 - H(\varphi)) \log m_{image}(BG)\, dx\,, \\ E_{curve}(\varphi) = &\nu_1 \int_\Omega |\nabla H(\varphi)|\, dx \end{aligned} \tag{4.30}$$

Thus, $m_{image} = m_{image,1} \otimes \ldots \otimes m_{image,k}$ is the joint mass of the different image channels, that are given according to Equation (4.12):

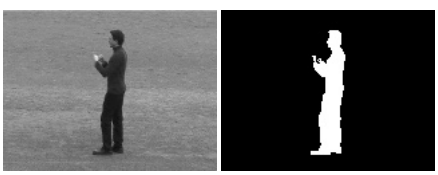
$$\begin{aligned} &m_{image,j}(\emptyset) = 0\,, \\ &m_{image,j}(FG) = m_{image,j}(\mathcal{L}_x = FG) = p_{image}(I_j(x) \mid \mathcal{L}_x = FG)\,, \\ &m_{image,j}(BG) = m_{image,j}(\mathcal{L}_x = BG) = p_{image}(I_j(x) \mid \mathcal{L}_x = BG)\,, \\ &m_{image,j}(\Psi) = 1 - (p_{image}(I_j(x) \mid \mathcal{L}_x = FG) + p_{image}(I_j(x) \mid \mathcal{L}_x = BG))\,. \end{aligned} \tag{4.31}$$

This is the same information used in the previous chapter, where the likelihoods $p(I_j(x) \mid \mathcal{L}_x = FG) = p_{image}(I_j(x) \mid \mathcal{L}_x = FG)$ are learned on basis of the current contour.

In contrast, $m_{user} = m_{user,1} \otimes \ldots \otimes m_{user,k}$ is directly learned from the user labeling L:

$$\begin{aligned} &m_{user,j}(\emptyset) = 0\,, \\ &m_{user,j}(FG) = m_{user,j}(\mathcal{L}_x = FG) = p_{user}(I_j(x) \mid \mathcal{L}_x = FG)\,, \\ &m_{user,j}(BG) = m_{user,j}(\mathcal{L}_x = BG) = p_{user}(I_j(x) \mid \mathcal{L}_x = BG)\,, \\ &m_{user,j}(\Psi) = 1 - (p_{user}(I_j(x) \mid \mathcal{L}_x = FG) + p_{user}(I_j(x) \mid \mathcal{L}_x = BG))\,. \end{aligned} \tag{4.32}$$

$E_{user-shape}$ can be interpreted as a user-defined shape prior, while $E_{user-image}$ takes the image information on the marked regions into account and can therefore be interpreted as an indicator for the appearance of a region.



Fusing the mass functions m_{im} and m_{user} contained in E_{image} and $E_{user-image}$ respectively, with Dempster's rule of combination yields an energy functional of the form:

$$\begin{aligned} E(\varphi) = & \underbrace{-\int_\Omega H(\varphi)\log m(FG)\,dx - \int_\Omega (1-H(\varphi))\log m(BG)\,dx}_{\text{data term + user defined term}} \\ & + \nu_1 \underbrace{\int_\Omega |\nabla H(\varphi)|\,dx}_{\text{curve constraint}} - \nu_2 \underbrace{\int_\Omega L_\sigma H(\varphi)\,dx}_{\text{user-shape}}, \end{aligned} \tag{4.33}$$

where the mass function $m = m_{image} \otimes m_{user}$ fuses the image data given by m_{image} and the user data given by m_{user} according to Dempster's rule of combination. Minimizing (4.33) using variational methods and gradient descent leads to the following partial differential equation:

$$\frac{\partial \varphi}{\partial t} = \delta(\varphi)\left[\log\left(\frac{m(FG)}{m(BG)}\right) + \nu_1 \mathrm{div}\left(\frac{\nabla\varphi}{|\nabla\varphi|}\right) + \nu_2 L_\sigma\right]. \tag{4.34}$$

The Dempster-Shafer theory of evidence is used to fuse the information since feature channels with low support have a lower influence on the evolving boundary. This is helpful because the user-defined regions can be very sparse, which means that the resulting channel-histograms can have regions where neither the object nor the background region is supported. Using the Bayesian framework for fusing this information would lead to small probabilities for both regions, ignoring all other feature channels, especially the image feature channels. With the proposed framework based on Dempster's rule of combination, this would be interpreted as uncertainty meaning that the other feature channels are not affected by this feature.

In contrast to the work of Cremers et al. [CFRA07] the proposed framework not merely provides an indication in terms of a shape prior for the segmentation, but actually uses the intensity information given by the user labeling. This information is further combined with Dempster's rule of combination, instead of multiplying the different probabilities, to represent inaccuracy and uncertainty. While the user labels in [CFRA07] principally have only local support to the evolving boundary and thus to the final segmentation, our framework allows global support for user defined regions. Figure 4.9 shows the proposed general interactive segmentation workflow. Given an initial curve without additional user information marking foreground or background, the segmentation follows the left hand side of the figure, that is the same framework proposed in the previous chapter. If the user provides additional information, e.g. in form of user strokes, likelihoods are learned and included in the minimization process (see the right hand side of the figure). An important property, of the proposed framework is the possibility to directly affect the evolution of the curve.

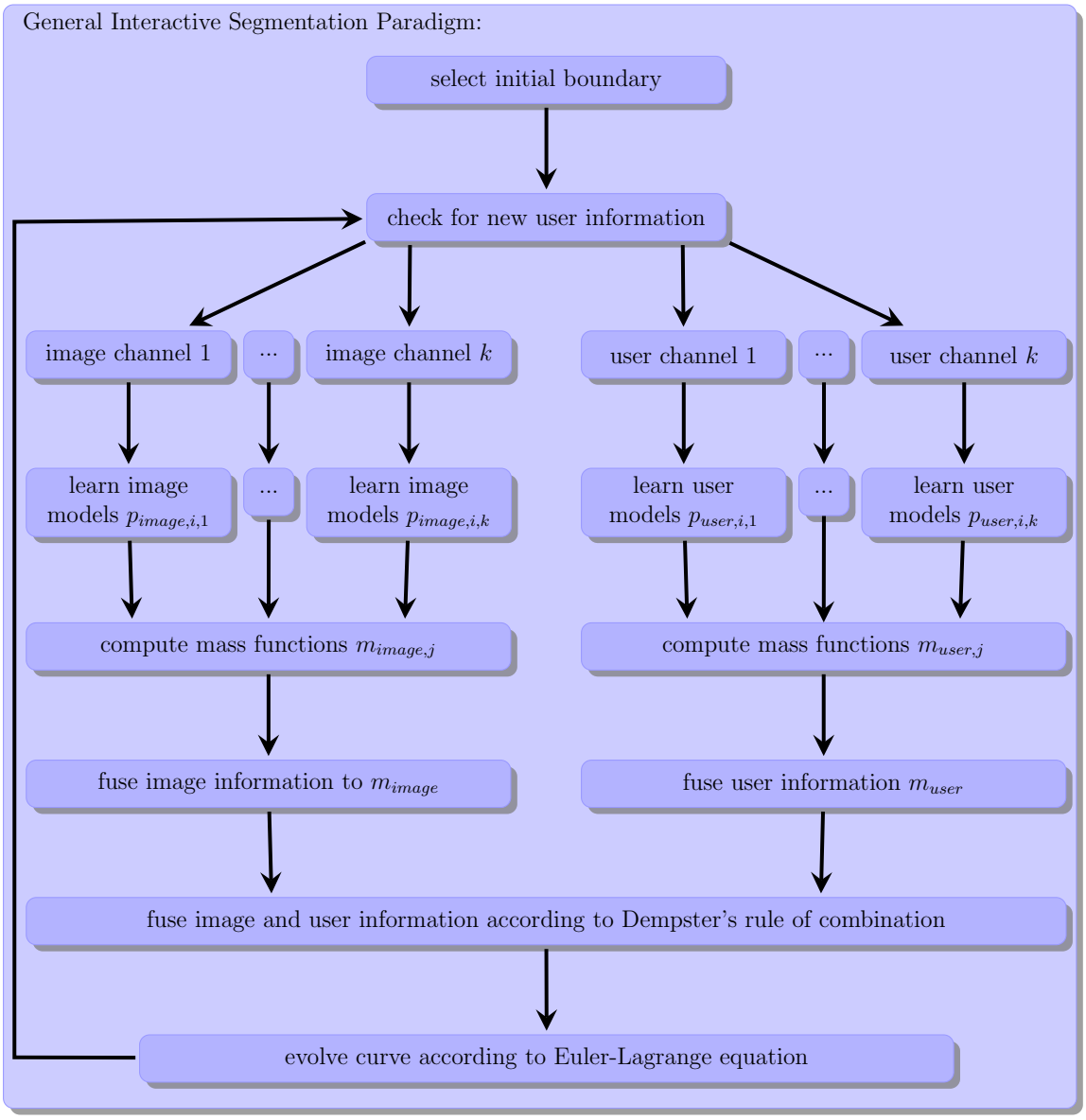


Figure 4.9: Proposed user interactive image segmentation paradigm based on combining the variational framework and Dempster-Shafer theory of evidence to fuse image and user given information. First the user has to select an initial curve; Without further user information, the segmentation follows our previously proposed method. During curve evolution, the user can provide additional information, that is included in the minimization process to refine the segmentation.

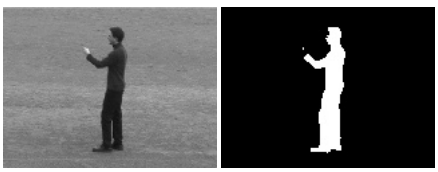


Table 4.1: Results of our user study. While the mean F_1-measure is comparable for all methods, the proposed method (highlighted in blue) required significantly fewer user interactions, comparing the average number of strokes.

Image	Graph Cuts	[CFRA07]	proposed Method
Lady Bug	4.33 str.	1.3 str. + 2 clicks	1.3 str. + 2 clicks
Eagle	9.33 str.	7 str. + 2 clicks	5.5 str. + 2 clicks
Bird	7.83 str.	2.3 str. + 2 clicks	1.83str. + 2 clicks
Flowers	7.17 str.	6.6 str. + 2 clicks	4 str. + 2 clicks
Soldier	9.33 str.	10.3 str. + 2 clicks	7.33 str. + 2 clicks
mean F_1	0.9623	0.9491	0.9547

4.2.2 Experimental Results

In this section experimental results of the proposed user-interactive segmentation framework are presented. Several qualitative results are shown in Figures 4.8 and 4.10 where the proposed method is compared to graph cuts [BJ01], see Chapter 3.2. This method was chosen for comparison since it is a widely used interactive segmentation method. The example images used for the experiments are natural images taken from the Berkeley segmentation dataset [MFTM01]. Furthermore the proposed framework is quantitatively compared to the user-interactive framework proposed in [CFRA07].

Additionally to the qualitative results, a user study was performed, where six persons segmented five real images with the proposed framework, the framework in [CFRA07] and the graph cut segmentation tool. In these moderately difficult examples (e.g. the soldier in Figure 4.11) the proposed framework needed significantly fewer user-interactions while the mean F_1 measure over all segmented images is comparable. The F_1 measurement used here is the harmonic mean of precision and recall calculated on the foreground pixels. It is defined by:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \,. \tag{4.35}$$

The quantitative results of our user study are summarized in Table 4.1. Figure 4.11 shows some qualitative results of the segmented images. Important to notice is the fact, that users tend to make longer strokes using graph cuts. Especially the two initial strokes are very large (see Figure 4.10 and 4.11) compared to the small strokes in our framework. The average stroke size (length) with the proposed method is approximately half of the stroke size within the graph cut framework. Although the proposed method is not implemented on the GPU, the total time for segmenting the images was almost the same for both methods. In addition the users

are able to guide the evolving boundary instead of changing the final segmentation, which is slightly more intuitive.

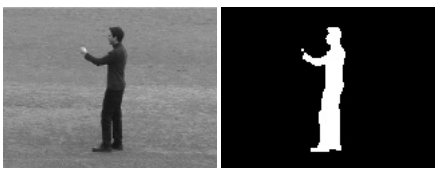
Note: The proposed method was not compared to other segmentation frameworks like e.g. GrabCut [RKB04], but the results presented in [RKB04] are very close to what the users achieved in the study.

4.3 Discussion

In Chapter 4 the Dempster-Shafer theory of evidence, as an extension of the Bayesian model, is proposed for the task of variational image segmentation based on level sets. Appropriate mass functions are defined and integrated in the energy minimization process. The main property of this theory, in contrast to Bayes, is to combine information arising from different feature channels by modeling inaccuracy and uncertainty at the same time. It therefore allows to fuse these information and resolve conflicts more intuitively. Several experiments on real and synthetic images demonstrated the properties and advantages of using the Dempster-Shafer evidence theory for image segmentation.

Furthermore the model including Dempster's theory of evidence is extended by means of user interaction, resulting in an intuitive interactive segmentation framework. The user interactions, in form of strokes marking foreground and/or background, are integrated elegantly into the framework to allow more precise segmentations. Therefore, a user-defined shape prior (local influence) and a user-defined image model (global influence) is defined and combined with the traditional framework. The impact of the proposed framework is demonstrated by several experiments on natural images and a user study, comparing the framework to other well known interactive segmentation frameworks. With the novel extensions the level set based interactive segmentation framework allows small user interactions having global influence on the evolving boundary. In comparison to graph cut, the presented framework needs significantly fewer user interactions to produce high-quality segmentations.

Overall, a segmentation framework is developed, that involves the user with very little effort but full control to allow segmenting objects in an image quickly and accurately. The proposed variational segmentation framework based on Dempster's theory of evidence was further extended to integrate available 3D information [FSRW10]. Therefore, a segmentation scheme using multi-view silhouette fusion is integrated into the proposed framework as additional information. A quick review of this approach can be found in the Appendix A.2.



initialization (2 clicks)

initialization (5 strokes)

segmentation result

segmentation result

user interactivity (4 additional strokes)

user interactivity (9 additional strokes)

segmentation result after user refinement (initialization (2 clicks), 1 foreground- and 3 background strokes)

segmentation result after user refinement (11 foreground- and 3 background strokes)

Figure 4.10: Segmentation results comparing the proposed interactive segmentation framework (left) and graph cut (right). The proposed segmentation algorithm needs significant fewer user interactions (red and blue strokes) to produce a slightly better result.

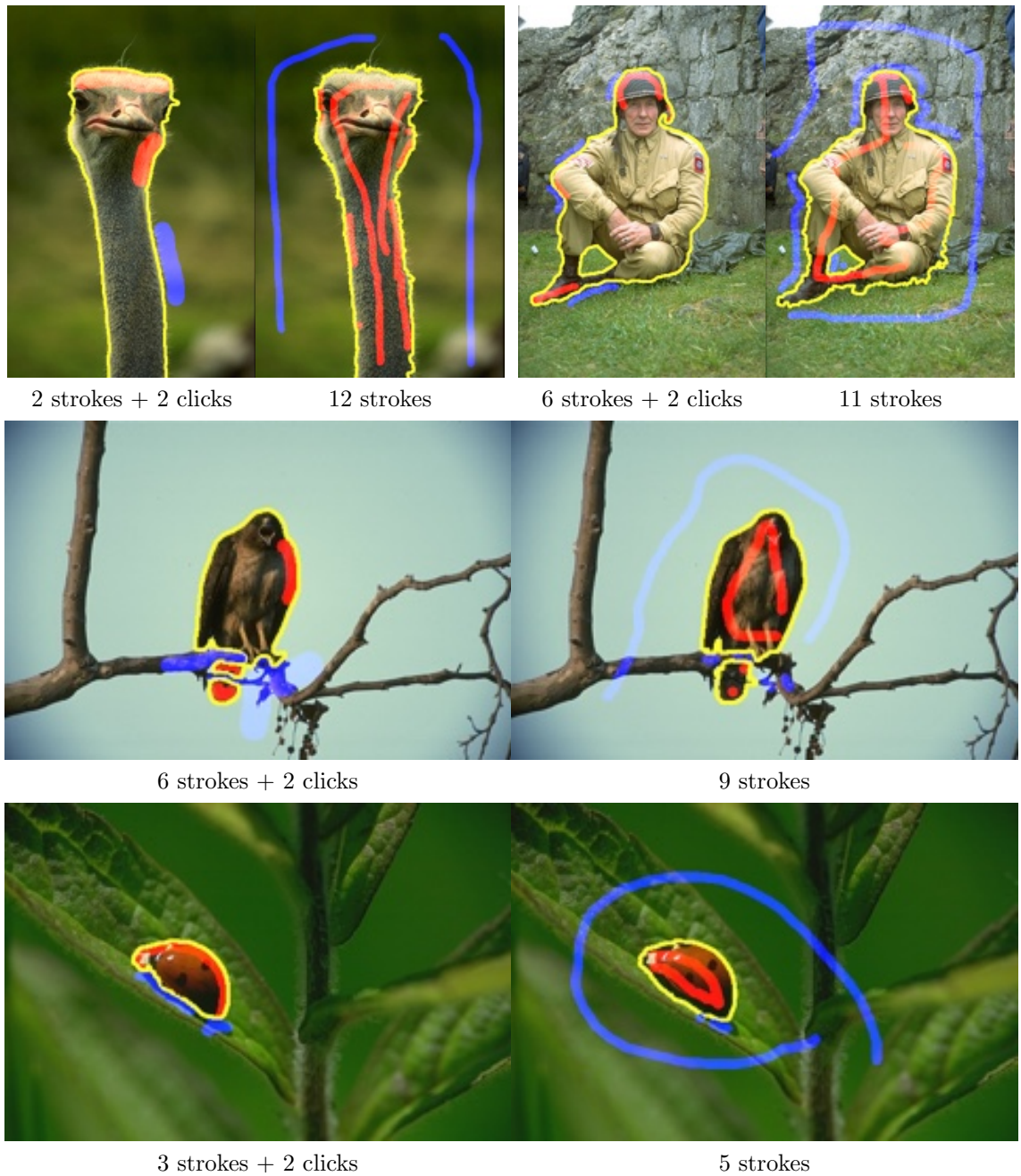
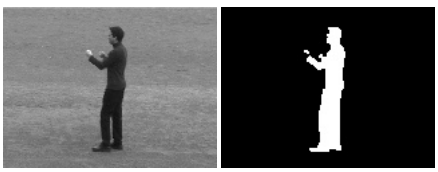


Figure 4.11: Segmentation results from the user study using the proposed variational framework (left) and graph cut (right). The yellow curves describe the segmentation boundaries, while the blue and red strokes mark the user defined regions.



Chapter 5

Discrete Energy Minimization with Dempster's Theory of Evidence

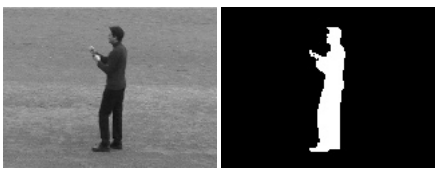
Since Boykov and Jolly [BJ01] presented their approach for interactive image segmentation, discrete energy minimization using graph cuts became a widely used framework for image segmentation. Besides variational approaches using level sets, see Chapter 4, graph cut approaches are often used for binary segmentation problems [GPS89, KZ04, BRB+04, KT05, Li09]. In the recent years, the graph cut approach was continuously extended, e.g. to solve multi-label problems [DB09].

However, an important drawback of the discrete optimization approach is their complexity. This means, that the approach by Boykov and Jolly is not able to efficiently segment high resolution images or video sequences due to their running time and memory requirements. The reason for this drawback is the correspondence between number of variables in the energy function and number of pixels in the image. Since the graph cut framework is an approach for interactive image segmentation it is necessary to have a small lag of time between user interaction and segmentation result. Therefore, it is important to reduce the complexity of the approach or the runtime to minimize the energy function.

The contributions of this chapter aim to reduce the complexity of the approach by simplifying the graph corresponding to the discrete energy function. Simplifying the graph means reducing the number of variables and thus reducing the lag of time between user interaction and the segmentation result. The first proposed approach, called *SlimCuts* simplifies the graph by contracting so-called *simple edges*. By contracting an edge in the graph the number of variables is reduced by one. An important characteristic of these *simple edges* is that they are efficiently to find and by contracting those edges, the maximum flow is preserved. Thus the segmentation result computed on the simplified graph is equal to the result computed on the original graph.

The amount of graph reduction of the *SlimCut* approach is limited by the number of *simple edges*. Therefore a second approach, that contract edges of similar pixels is proposed. Dempster's theory of evidence is used to fuse the terms of the discrete energy function to define a similarity weight for neighboring pixels. Similar pixels in the image are grouped by contracting the edge connecting the pixels in the underlying graph. The experimental results show that the number of variables can be reduced dramatically with only small changes in the segmentation result.

Finally, Dempster's theory of evidence is integrated into the discrete energy minimizing framework to fuse color and depth information in a novel way. The following chapters base directly on my publications [SR11b, SSR12, SGR13].



5.1 SlimCuts: High Resolution Image Segmentation

This chapter proposes an algorithm for interactive image segmentation using graph cuts. It can be used to efficiently solve labeling problems on high resolution images or resource-limited systems. The basic idea of the proposed algorithm is to simplify the original graph while maintaining the maximum flow properties. Thereby the segmentation result stays the same. The resulting *Slim Graph* can be solved with standard maximum flow/minimum cut-algorithms. It will be shown that the maximum flow/minimum cut of the Slim Graph corresponds to the maximum flow/minimum cut of the original graph. Experiments on image segmentation show that using the proposed graph simplification leads to a significant speedup and memory reduction of the labeling problem. Thus large-scale labeling problems can be solved in an efficient manner even on resource-limited systems.

Discrete optimization of energy minimization problems using maximum flow algorithms have become very popular in the fields of computer vision [BK04]. This has been driven by their ability to efficiently compute a global minimum of special energy functions occurring in computer vision problems. Examples for such energy functions include image segmentation, image restoration, dense stereo estimation and shape matching [BRB$^+$04, BK03, LB07].

In parallel to the improvement of discrete energy minimization algorithms [BVZ01, BK03, KT05, RKB04, BT08, Kom10], the resolution of single images and image sequences increased significantly. Compared to standard benchmark images, which have an approximate size of 120.000 pixels, nowadays commercial cameras capture images with many more pixels, e.g. more than 20 million. Since most energy functions for image segmentation or stereo matching contain one discrete variable per pixel, the minimization using maximum flow algorithms can be computationally extremely expensive.

Prior Work

Research on solving discrete optimization problems using maximum flow / minimum cut algorithms for applications in computer vision can be divided into the following approaches:

Augmenting paths: Due to the works of Boykov and Kolmogorov the so-called Boykov and Kolmogorov augmenting paths algorithm (BK-algorithm) [BJ01, BK04] is widely used for computer vision problems. This algorithm efficiently solves moderately sized 2D and 3D problems with low connectivity. In [SK10] Strandmark and Kahl proposed a parallel implementation of the BK-algorithm, where subproblems are iteratively solved on multiple cores or multiple machines.

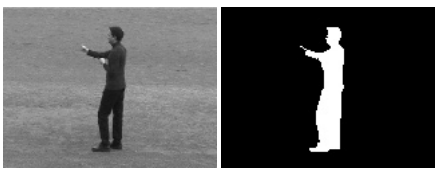
(a) (b) (c)

Figure 5.1: Example segmentation using Apple's iPhone 4. (a) shows the image with user scribbles marking foreground (red) and background (blue); (b) label map defined by the proposed Slim Graph. White and gray pixels denote fore- and background, black regions are unlabeled pixels; (c) final segmentation depicted in red.

Push-relabel: Most parallelized maximum flow/minimum cut algorithms have been focused on push-relabel algorithms [DB08]. These methods outperform the traditional BK-algorithm for huge and highly connected grids [BK04]. An algorithm that involves GPU processing is CUDA cuts, presented by Vineet and Narayanan [VN08]. In contrast to these algorithms the proposed SlimCuts does not use special hardware (multiple cores, GPU) to reduce the computational time.

Convex optimization: Formulating the maximum flow/minimum cut problem as a linear program is another promising approach to parallelize graph cuts. Assuming only bidirectional edges, the maximum flow problem can be reformulated as an l_1 minimization problem [BT08]. In [KSK+08] Klodt et al. used GPU based convex optimization to solve continuous versions of graph cuts. However, the speedup using a GPU is low compared to the BK-algorithm and the main advantage of continuous cuts is to reduce metrication errors due to discretization.

Multi-Scale: Besides the approaches to parallelize the maximum flow/minimum cut problem to outperform existing algorithms, multi-scale processing is an approach



to reduce memory and computational requirements of optimization algorithms. The idea to efficiently solve large scale optimization problem is to first solve the problem at low resolution using standard techniques [PB99]. The resulting low resolution labeling is refined on the high resolution problem in a following optimization step. Boundary banded algorithms [LSGX05, SG06] are examples for multi-scale image segmentation of high resolution images. Kohli et al. [KLR10] proposed an uncertainty driven multi-scale approach for energy minimization allowing to compute solutions close to the globally optimum. However, both approaches suffer from the problem that they are not able to efficiently recover from large scale errors present in the low resolution result.

Grouping of variables: Another simple and widely used technique merges variables of the energy function. The number of variables is reduced by segmenting the image into a small number of patches (called *superpixels*) [SM00, FH04]. Those groups are represented by a single variable in the energy function. This idea has been used for solving problem instances of object segmentation and stereo matching. Besides a number of well known image partitioning methods [FH04, CM02, LSK+09, VBM10], Kim et al. presented a method [KNKY11] where the terms of the energy function and the algorithm proposed by Felzenszwalb and Huttenlocher [FH04] are used to decide if two variables should be merged.

Graph sparsification: In the field of applied mathematics graph sparsification and graph simplification is an important matter. Karger and Stein proposed the Recursive Contraction Algorithm in [KS96]. The algorithm relies on the idea that the minimum cut is a small set of edges and a randomly chosen edge is unlikely to be in this set. They randomly contract edges and showed that the minimum cut is found with high probability. However, it is not guaranteed that the cut is optimal and it does not exist a fast prove that the minimum cut is found. Similarly Bencúr and Karger [BK96] and Spielman and Teng [ST04] proposed algorithms based on random sampling to approximate the minimum cut of a given graph. In contrast to these approximating algorithms, the proposed SlimCuts maintain the global optimal solution. In [CGK+97], Chekuri, Goldberg et al. developed heuristics to improve practical performance of minimum cut algorithms. They propose to use the Padberg-Rinaldi heuristic [PR90] to contract edges that are not in the minimum cut. Therefore they apply several so-called PR tests to identify those edges. In practice the PR tests are computationally too expensive for large graphs. In [HKR+01], Hogstedt et al. proposed a number of heuristically graph algorithms to simplify partitioning of distributed object applications. For the special case of an s-t minimum cut (two machine nodes) their condition for graph simplification preserves one minimum cut. In 2011, Lermé et al. proposed a similar approach for graph sparsification [LLM11]. Based on flow assumptions on regions, where the total flow on the boundary is analyzed, the graph simplified. In the experiments, the minimum cut was preserved. However they did not showed that the maximum

flow is preserved in all cases.

In contrast to these approaches, the edges to build a Slim Graph are efficient to find and contract and the simplification guarantees that all minimum cuts are preserved. Thus, the maximum flow is preserved.

Contribution

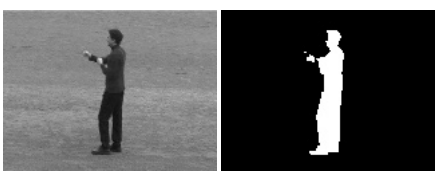
To efficiently solve large-scale labeling problems a so-called *Slim Graph* is computed by simplifying the original graph. Therefore, nodes that are connected by a so-called *simple edge* are identified and contracted, without changing the value of the maximum flow and the corresponding minimum cut. Nodes that are connected by a *simple edge* will have the same label in the final segmentation. Thus, they can be merged into a single node and the number of variables is reduced. The original graph is hence reduced to a *Slim Graph* without changing the minimum energy state. The proposed simplification can be applied to each of the aforementioned algorithms. Thus the proposed algorithm provides a general speedup and memory reduction without suffering from the problem of large scale errors or the use of special hardware, e.g. GPU or multiple processors.

Besides the speedup and memory reduction, the merged nodes can help the user to set the parameter included in the minimization problem and to guide the user. Visualizing the simplified graph reveals which areas of the image / graph are assigned to foreground or background because of the prior information. Even for high resolution images, this provides a fast feedback where further user strokes are necessary to satisfactory solve the segmentation problem.

The proposed algorithm is neither a parallelization of an existing algorithm nor a multi-scale approach to speedup and reduce the amount of memory of graph cuts. Hence, it does not suffer from the problems of these methods. In contrast to the works in the field of applied mathematics on graph sparsification and graph simplification, where the minimum cut is approximated, the proposed method guarantees that the value of the minimum cut is preserved. The given condition to test whether an edge is simple or not is computationally cheap and applicable for large-scale problems.

5.1.1 Constructing *Slim Graph*

In this section the construction of a so-called *Slim Graph* it explained. By merging or rather contracting nodes that are connected by *simple edges*, the original graph is reduced to a *Slim Graph*. First, these special *simple edges* are defined and it is proven that these edges are not part of any minimum cut. Based on this Lemma, it is proven that the minimum cut of a *Slim Graph* corresponds to the minimum



cut of the original graph and can be used to minimize the large-scale optimization problem.

Lemma 1 *Let $G = (\mathcal{V}_\mathcal{G}, \mathcal{E}_\mathcal{G})$ be a graph, A,$B \in \mathcal{V}_\mathcal{G}$, $e_{A,B} \in \mathcal{E}_\mathcal{G}$ the edge from node A to node B and $\mathcal{C}$ a minimum s-t-cut.*
If

$$c(e_{A,B}) > \sum_{\substack{i:e_{i,A}\in\mathcal{E}_\mathcal{G} \\ i\neq B}} c(e_{i,A}) \quad or \quad c(e_{A,B}) > \sum_{\substack{i:e_{B,i}\in\mathcal{E}_\mathcal{G} \\ i\neq A}} c(e_{B,i}) \tag{5.1}$$

then

$$e_{A,B} \notin \mathcal{C}. \tag{5.2}$$

Simply spoken: If the weight of one edge e of node A is bigger then the sum of all edges adjacent to A, then the minimum cut $\mathcal{C}$ does not contain the edge e.

Proof: Following Shannon, the value of the maximum flow is equal to the value of a minimum cut. The value of the maximum flow in G can be computed with the *augmenting path*-algorithm of Ford-Fulkerson [FF56]. Following this algorithm paths from s to t are searched and augmented, as long as there exists a path from s to t. Because of Equation (5.1) the edge $e_{A,B}$ will never become saturated, which means that the edge is not part of the minimum cut $\mathcal{C} \Rightarrow e_{A,B} \notin \mathcal{C}$. □

Those edges $e_{A,B} \in \mathcal{E}_\mathcal{G}$ fulfilling Equation (5.1) are called *simple edges*.

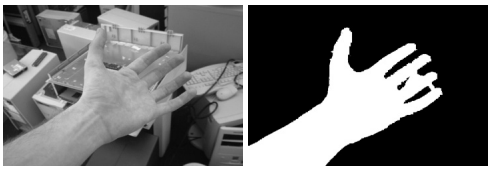
A similar Lemma was also given by Hogstedt et al. [HKR+01]. They defined a so-called dominant edge, with a stronger condition. Having a dominant edge e, there exists a minimum cut not containing this edge. In contrast, condition (5.1) results in a *simple edge* e that is not contained in any minimum cut. That means that all minimum cuts are preserved when contracting only *simple edges*.

Simplifying a graph:

With the following rules a graph is simplified and the number of variables of the maximum flow/minimum cut problem is reduced:
Let $G = (\mathcal{V}_\mathcal{G}, \mathcal{E}_\mathcal{G})$ be a graph with a *simple edge* $e_{A,B} \in \mathcal{E}_\mathcal{G}$ connecting nodes A,$B \in \mathcal{V}_\mathcal{G}$. Without loss of generality, let $e_{A,B}$ fulfill the left condition of Equation (5.1). Then the *Slim Graph* $\widetilde{G} = (\widetilde{\mathcal{V}}_\mathcal{G}, \widetilde{\mathcal{E}}_\mathcal{G})$ is constructed as follows:

Nodes: $\widetilde{\mathcal{V}}_\mathcal{G} = \mathcal{V}_\mathcal{G} \setminus \{A,B\} \cup \{AB\}$. That means nodes A and B are merged to node AB by contracting the edge $e_{A,B}$. The number of nodes in the Graph is reduced by one, since one new node is added and tow nodes are removed.

Edges: The following two cases are distinguished:



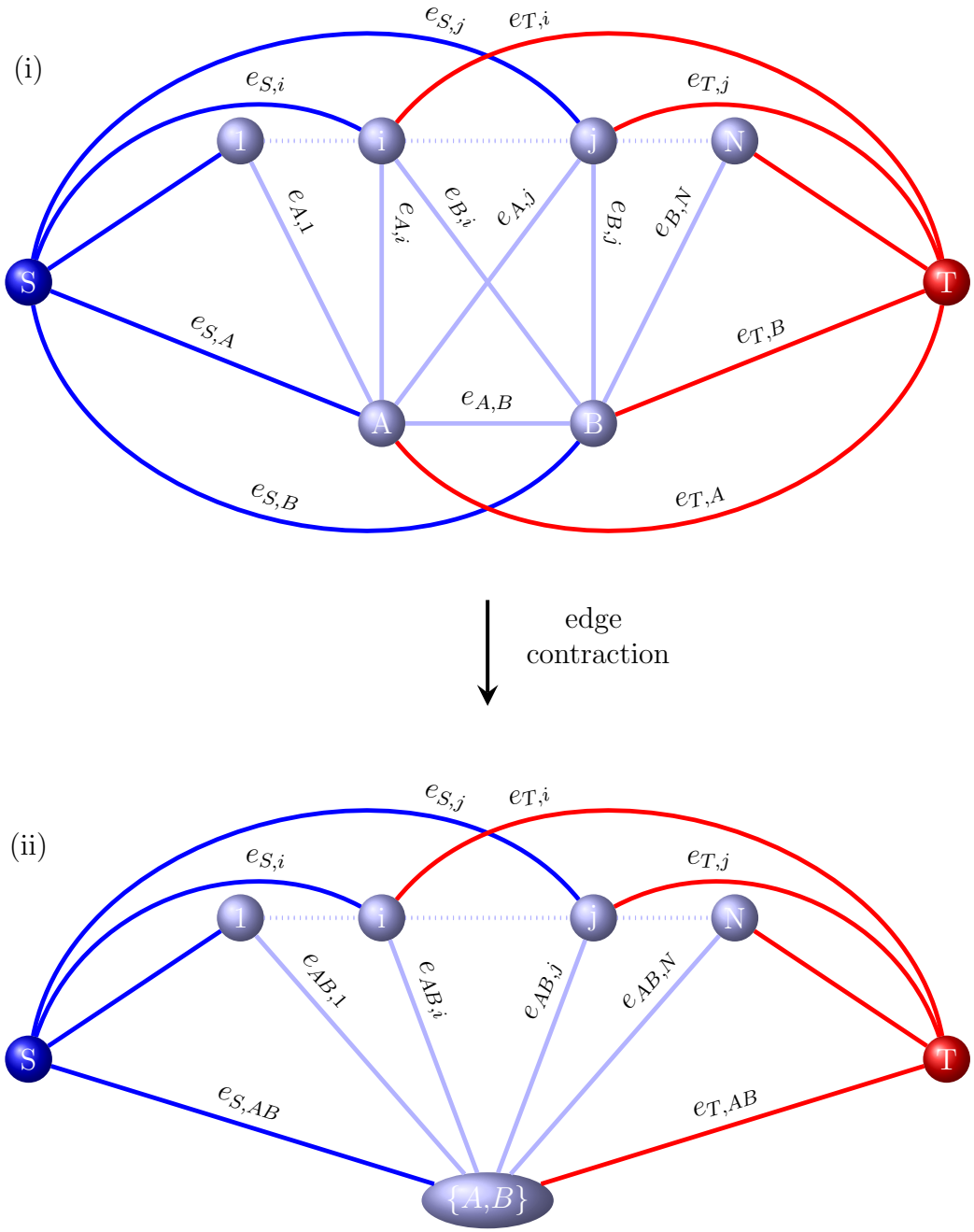
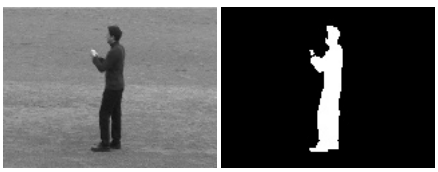


Figure 5.2: Example of contracting an edge to build a *Slim Graph* (*ii*) out of the original graph (*i*) because of a *simple edge* between nodes A and B. The given rules contract nodes A and B to a single node $\{A,B\}$, replace edges connected to one of the nodes and merge edges that are connected to both nodes.



(i) For all nodes $i \in \mathcal{V}_\mathcal{G}$ connected to exactly one of the nodes A or B: Without loss of generality, let $e_{i,A}$ be the edge connecting node i with node A ($e_{i,B} \notin \mathcal{E}_\mathcal{G}$). Then the edge $e_{i,A}$ is replaced by a new edge $e_{i,AB}$ with $c(e_{i,AB}) = c(e_{i,A})$. This operation does not change the number of edges.

(ii) For all nodes $i \in \mathcal{V}_\mathcal{G}$ connected to both of the nodes A and B with edges $e_{i,A}$ and $e_{i,B}$ or $e_{A,i}$ and $e_{B,i}$. The two edges are merged, resulting in a new edge $e_{i,AB}$ or $e_{AB,i}$ with the capacity $c(e_{i,AB}) = c(e_{i,A}) + c(e_{i,B})$ or $c(e_{AB,i}) = c(e_{A,i}) + c(e_{B,i})$. This operation reduces the number of edges by one.

Figure 5.2 shows the construction of a *Slim Graph*. Assuming a *simple edge* $e_{A,B}$, the nodes A and B are merged to a single node $\{A,B\}$. Edges connected to exactly one of the nodes are replaced by new edge. In the given example, these nodes are $1, \dots i-1, j+1, \dots N$. The edges $e_{A,1} \dots e_{A,i-1}, e_{B,j+1}, \dots e_{B,N}$ are replaced by new edges $e_{AB,1} \dots e_{AB,i-1}, e_{AB,j+1}, \dots e_{AB,N}$ without changing the capacities of these edges. Nodes that are connected to both A and B in this example are $i, \dots j$. For these nodes, the two edges $e_{A,h}$ and $e_{B,h}$ are merged to one edge $e_{AB,h}$ with capacity $c(e_{AB,h}) = c(e_{A,h}) + c(e_{B,h})$. The resulting *Slim Graph* has one node and $j - i$ edges less than the original graph.

The following theorem shows the connection of the minimum cut of a graph and its simplified *Slim Graph*.

Theorem 1 *Let $G = (\mathcal{V}_\mathcal{G}, \mathcal{E}_\mathcal{G})$ be a graph, $A,B \in \mathcal{V}_\mathcal{G}$, $e_{A,B} \in \mathcal{E}_\mathcal{G}$ a simple edge connecting nodes A and B and f the maximum flow in Graph G. Since $e_{A,B}$ is a simple edge the Slim Graph $\widetilde{G} = (\widetilde{\mathcal{V}}_\mathcal{G}, \widetilde{\mathcal{E}}_\mathcal{G})$ can be computed by contracting the edge $e_{A,B}$. The value of the maximum flow $\widetilde{f}$ of graph $\widetilde{G}$ is equal to the value of the maximum flow in the original graph*

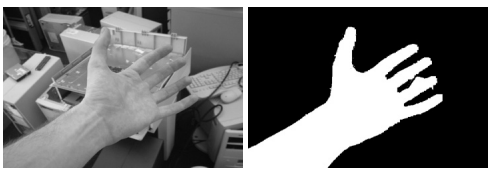
$$|f| = |\widetilde{f}|\,. \tag{5.3}$$

Simply spoken: By contracting a simple edge $e_{A,B}$, the maximum flow is preserved!

Proof: Following lemma 1 one knows that $e_{A,B} \notin \mathcal{C}$, where $\mathcal{C}$ is a minimum cut of G. This implies that the *simple edge* never becomes saturated. Therefore its capacity can be set to infinity without affecting the minimum cut or the maximum flow. It follows that both nodes A and B are in the same partition of $G \setminus \mathcal{C}$. W.l.o.g. let both nodes be in the partition connected to S. Hence the constructed node $AB \in \widetilde{\mathcal{V}}_\mathcal{G}$ in the graph $\widetilde{G} \setminus \mathcal{C}$ is connected to S. To prove the theorem one needs to show that:

(i) the minimum cut $\mathcal{C}$ of graph G implies a cut $\widetilde{\mathcal{C}}'$ in $\widetilde{G}$, with $|\mathcal{C}| = |\widetilde{\mathcal{C}}'|$.

(ii) the maximum flow f leads to a flow $\widetilde{f}'$ in $\widetilde{G}$, with the same value $|f| = |\widetilde{f}'|$



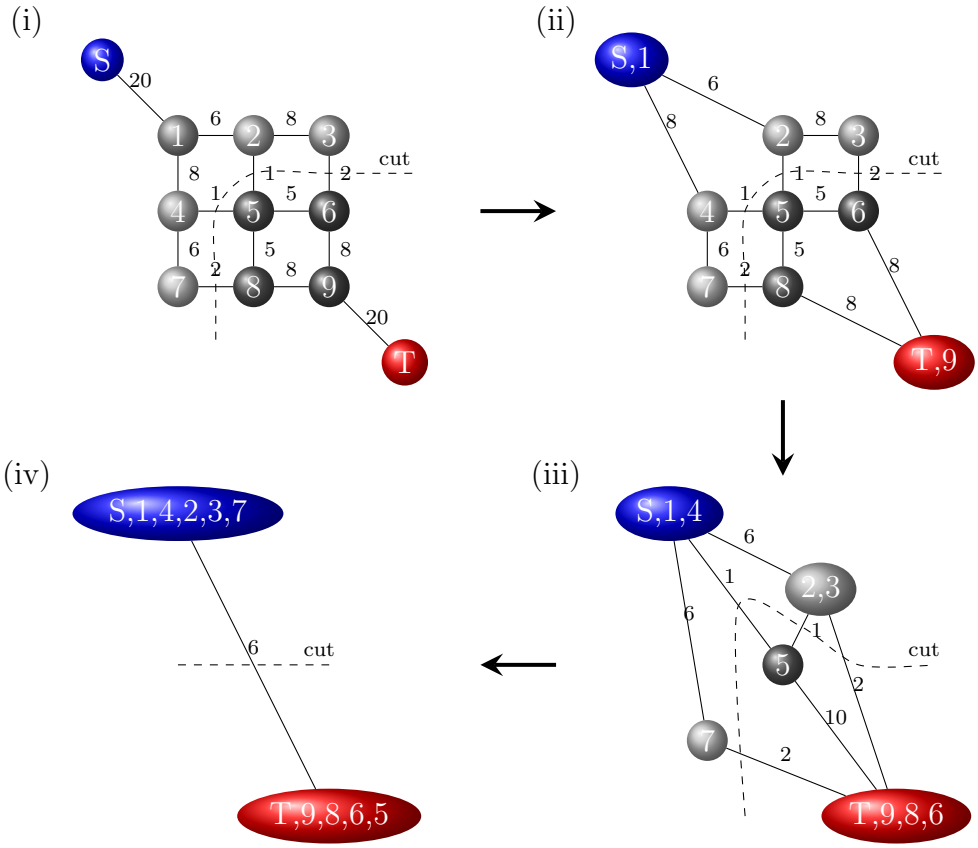
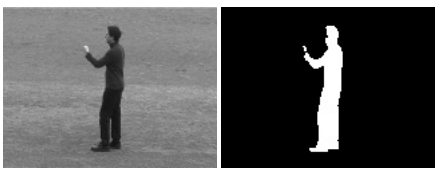


Figure 5.3: Example of simplifying a graph. (i) original graph with a minimum cut value of 6; (ii) Nodes S and 1 and nodes T and 9 are connected with a *simple edge* and merged respectively; (iii) Nodes S and 4, 2 and 3 and nodes T, 8 and 6 can be merged due to *simple edges* in one node respectively; (iv) shows the final *Slim Graph* for the example with a minimum cut value of 6. At each step of the construction the value of the maximum flow remains the same and also the final segmentation stays the same.



The first condition provides an upper bound for the value of the minimum cut in $\widetilde{G}$. On the other hand, the value of the flow $|\widetilde{f}'|$ in graph $\widetilde{G}$ provides a lower bound for the minimum cut. Since they are equal, the value of the maximum flow / minimum cut does not change in the *Slim Graph*.

Proof of (i): Let $i \in \mathcal{V}$ be nodes with a path to terminal node $T \in G \setminus \mathcal{C}$. Since A and B are connected to S, all edges $e_{A,i}$ and $e_{B,i}$ are part of the minimum cut $\mathcal{C}$. Defining $\widetilde{\mathcal{C}}'$ as follows implies a cut in the *Slim Graph* $\widetilde{G}$:

$$\begin{aligned}\widetilde{\mathcal{C}}' =& \{e_{i,j} \mid e_{i,j} \in \widetilde{\mathcal{E}}_{\mathcal{G}} \text{ and } e_{i,j} \in \mathcal{C}\} \\ &\cup \{e_{i,AB} \mid e_{i,A} \text{ or } e_{i,B} \in \mathcal{C}\}\end{aligned} \tag{5.4}$$

Due to the construction of the *Slim Graph* this definition leaves the value of the cut unchanged. Hence it holds $|\mathcal{C}| = |\widetilde{\mathcal{C}}'|$.

Proof of (ii): Let $i \in \mathcal{V}$ be a node and $p = (S, \ldots, A, i, \ldots, T)$ a path from S to T in G with flow $f(p)$. Following the construction of the *Slim Graph* the flow $f(p)$ is preserved by the path $\widetilde{p} = (S, \ldots, AB, i, \ldots, T)$ in $\widetilde{G}$. Hence the maximum flow of G implies a lower bound for the maximum flow in $\widetilde{G}$. □

Figure 5.3 shows how a *Slim Graph* can be constructed. By merging nodes that are connected by a *simple edge* the original graph (i) is simplified to the *Slim Graph* (iv). The value of the maximum flow and the minimum cut can be computed more efficiently on the new graph and remains identical. The labeling of the original graph is implicitly included in the labeling of the *Slim Graph*.

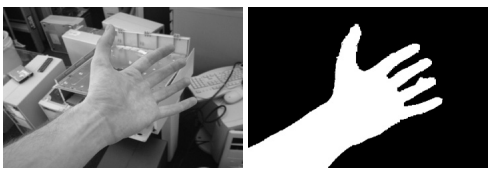
5.1.2 *Slim Graphs* for Simplified User Interaction

This section shows how the visualization of a *Slim Graph* can be integrated into the segmentation process to simplify user interactions and guide the user where to place additional strokes. Analyzing the original graph and the process of computing the *Slim Graph* shows, that *simple edges* exists most likely between pixel-nodes and terminal-nodes. Every pixel i that has been marked by the user or fulfill

$$-\log p(I(i) \mid \mathcal{L}_i = S) > \gamma \cdot \sum_{j \in \mathcal{N}_i} \text{dist}(i,j)^{-1} \cdot [\mathcal{L}_i \neq \mathcal{L}_j] \cdot \exp(-\beta \|I(i) - I(j)\|^2)\,, \tag{5.5}$$

where S is either foreground (FG) or background (BG), is connected to the corresponding terminal node by a *simple edge*. Visualizing these pixels in a label map results in a partial labeling with pixels labeled as foreground or background due to user marks or regional properties and unlabeled pixel.

Figure 5.4 shows an example image with user strokes and the label map coming from the *Slim Graph*. There are many pixels assigned to foreground or background due to their regional properties. Based on the given user input the final segmentation



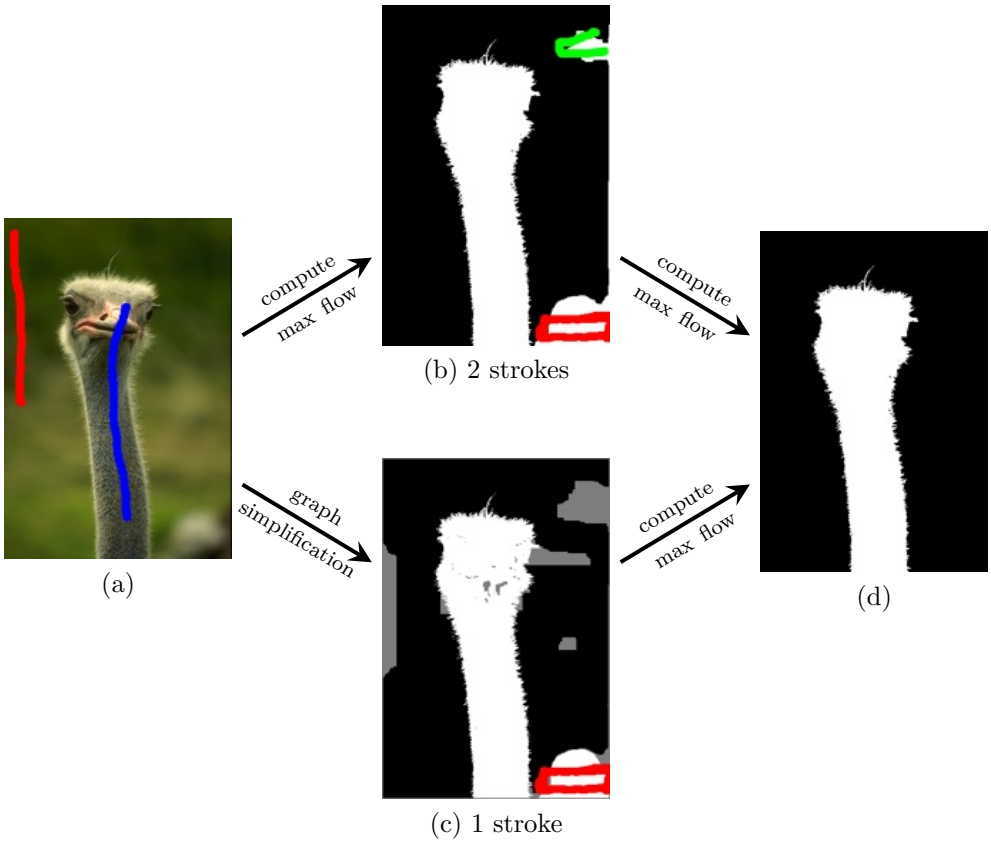
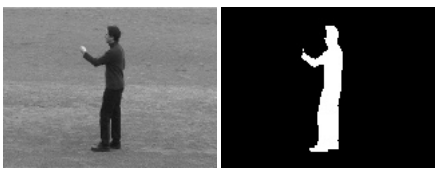


Figure 5.4: Example of utilizing the *Slim Graph* to simplify user interaction. (a) the original image with user scribbles; (b) resulting segmentation using graph cuts and possible additional user strokes to refine the segmentation (green and red); (c) the label map defined by the *Slim Graph*. White and black pixels denote fore- and background, gray pixels are unlabeled and one additional user strokes (red); (d) final segmentation result

will have two regions that are assigned a wrong label. To correct the segmentation the user has to mark these two regions or even one of them as background. That means the user has three options to affect the segmentation, shown in Figure 5.4b. In the label map coming from the *Slim Graph* there is exactly one region assigned a wrong label. That implies that the user has to mark this region as background to achieve a correct label map. In an optimal situation this user mark would also correct the labeling of the other region, leading to a good segmentation result with only one additional user mark. This situation is exemplarily shown in Figure 5.4c. Marking the wrong labeled region in the label map, guide the user to the desired segmentation 5.4d.



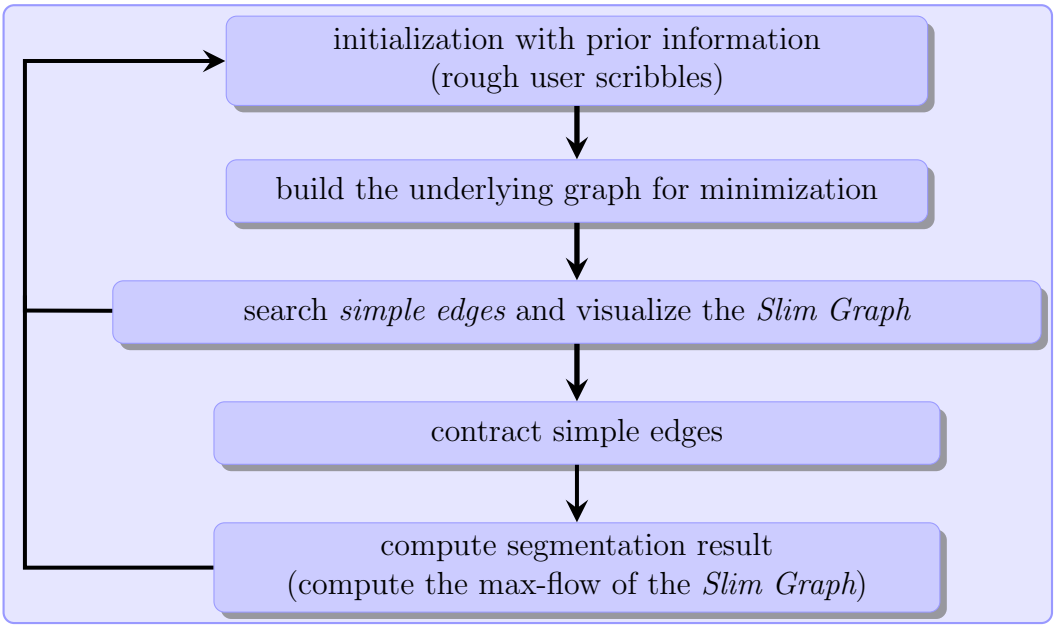


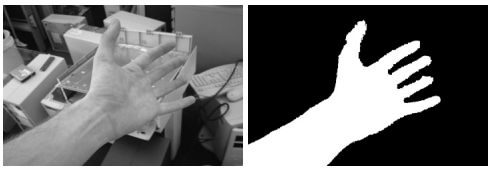
Figure 5.5: General workflow of the proposed *SlimCut* segmentation framework.

Using the proposed label map as additional information hints the user to place strokes in regions with high regional support. On the one hand this can lead to less user interactions for the problem of image segmentation and on the other hand, the label map can be computed very efficiently. That means, that it is much faster to start refining the label map of a high-resolution image than refining the segmentation result itself. In the given example using the *Slim Graph* for user guidance, the user interaction is reduced to one additional stroke. Furthermore, instead of computing the maximum flow two times it is required only once. The workflow of the proposed *SlimCut* segmentation framework is visualized in Figure 5.5.

5.1.3 Experiments

In this section, the proposed method is evaluated on small-scale images from the database used by Blake et al. [BRB+04] as well as on large-scale images with up to 26 million pixels found on the web. The images, trimaps and ground truth data is available online[1,2]. In the experiments the same energy function proposed by Blake et al. [BRB+04] and the same set of parameters are used. Since the proposed graph simplification was proven to not change the segmentation result segmentation results

[1] `http://research.microsoft.com/en-us/um/cambridge/projects/visionimagevideoediting/segmentation/grabcut.htm`
[2] `http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/`



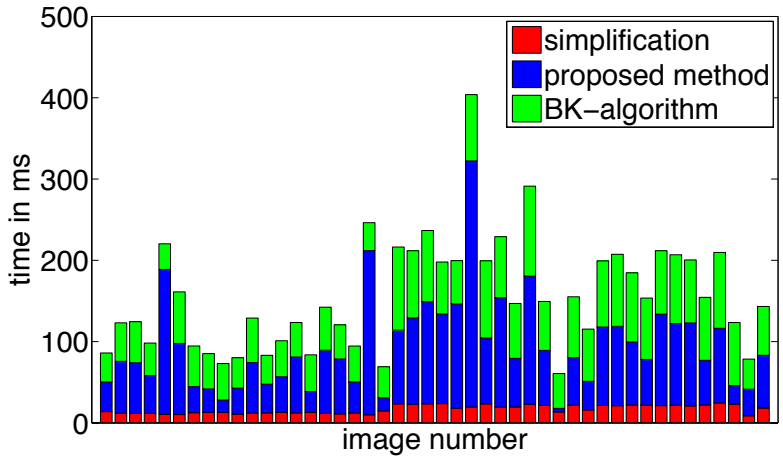


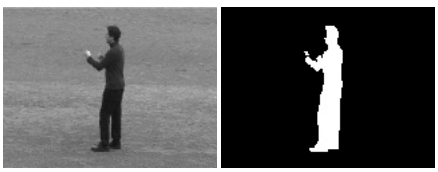
Figure 5.6: *Small-scale images:* Running time over 46 benchmark images [BRB+04] with image sizes between 481×321 and 640×480 pixels. The average speedup of the proposed method compared to BK-algorithm [BK04] on these small-scale problems is 40%. The maximum and minimum speedup is 70% and 14% respectively. The running time of the proposed method includes the reduction of the original graph.

are not shown and evaluated. Instead, the contribution is evaluated by comparing the computational time of the BK-algorithm with and without using the proposed *Slim Graphs*. All experiments were conducted on an Apple MacBook Pro with 2.4 GHz Intel Core i5 processor and 4GB Ram.

Experiments on small-scale images

Figure 5.6 shows the running times of Boykov's algorithm on the original graph and the *Slim Graph* and the running time of simplifying the graph. In the running time on the original graph the time creating the graph and computing the maximum flow is included. The times computing the capacities, histograms and learning the statistical models are excluded. The running time on the *Slim Graph* further includes the time for computing the *Slim Graph*. The experiments on the small scale images show that using *Slim Graphs* never affects the running time negatively and can significantly speedup the segmentation process.

As mentioned earlier, most *simple edges* exists between pixel-nodes and terminal-nodes. Since the weight of these edges is defined by the unary term, a second experiment on small scale images, comparing the effectiveness of the *Slim Graph* under weak vs. strong unary terms, was performed. Therefore the maximum flow for one image, three different trimaps (lasso, rectangle and user strokes) and varying parameter γ from 0 (strong unary term) to 100 (weak unary term), was computed. Figure



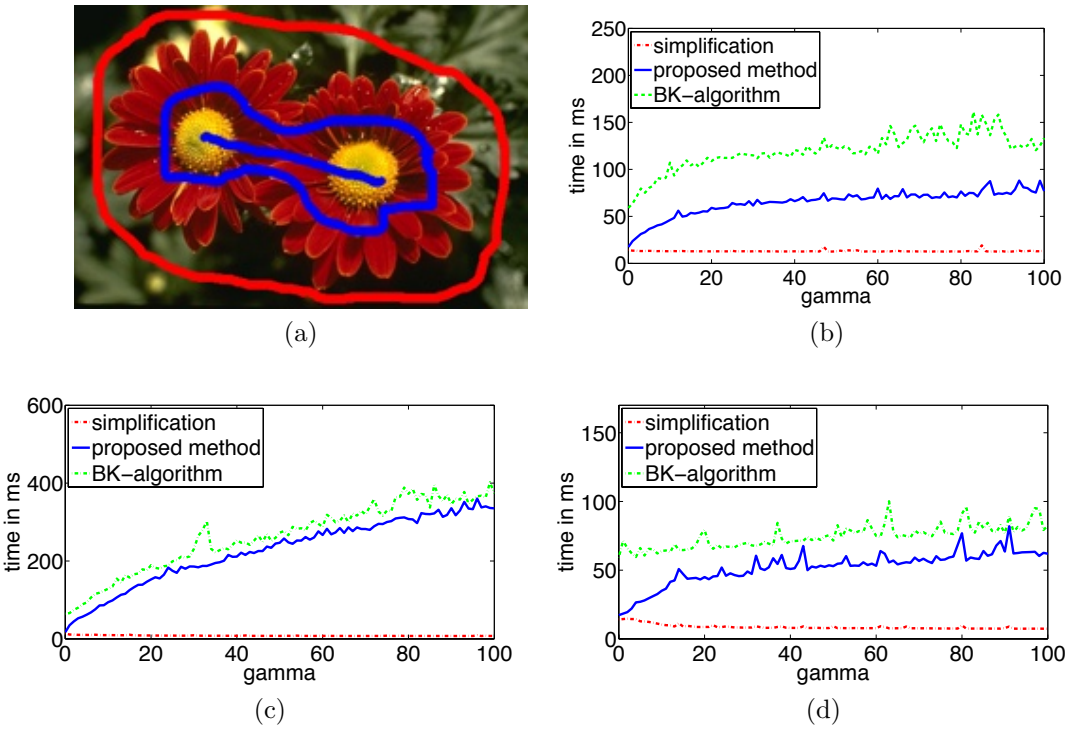


Figure 5.7: *Weak vs. Strong Unary Terms:* Running time over the flower image (a) with different trimaps and varying γ; (b) Lasso trimap around the flowers; (c) Rectangular trimap; (d) user strokes provided in (a). Using good initializations (b) and (c) the proposed algorithm performed significantly faster. Nevertheless, even with a poor initialization and a weak unary term we achieved a speedup.

5.7 shows the running times of this experiments. It turned out, that the speedup using *Slim Graphs* is highest using strong unary terms and trimaps separating fore- and background with small errors. Regardless, even with weak unary terms and poor initializations (e.g. rectangular-trimap) the proposed algorithm using the proposed *Slim Graph* performed faster and the running time was never affected negatively.

Experiments on high-resolution images

To evaluate the speed up of the proposed method on large-scale problems, high-resolution images with up to 26 MP were used. These images were down sampled to several image-sizes. As shown in Figure 5.8, solving the maximum flow problem on the *Slim Graphs* significantly speeds up the algorithm. This speed up is achieved by a large decrease of variables/nodes due to many *simple edges*. As already shown by Delong and Boykov [DB08] the problem of the BK-algorithm is that it becomes

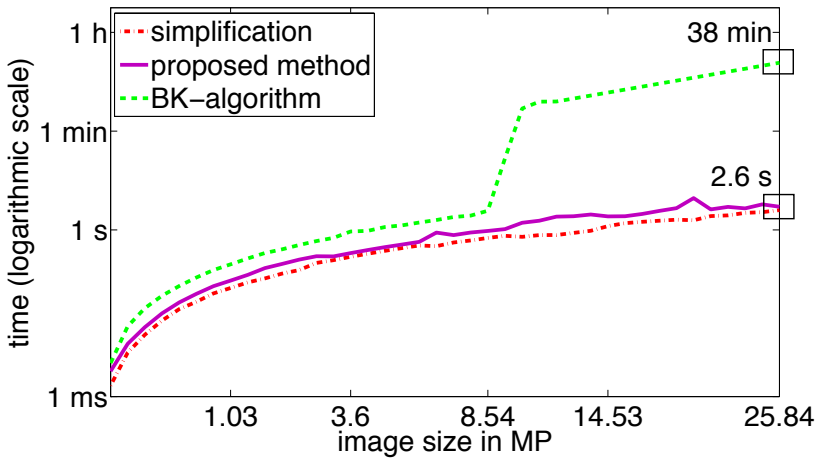
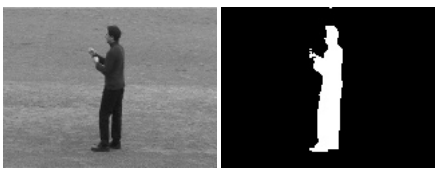


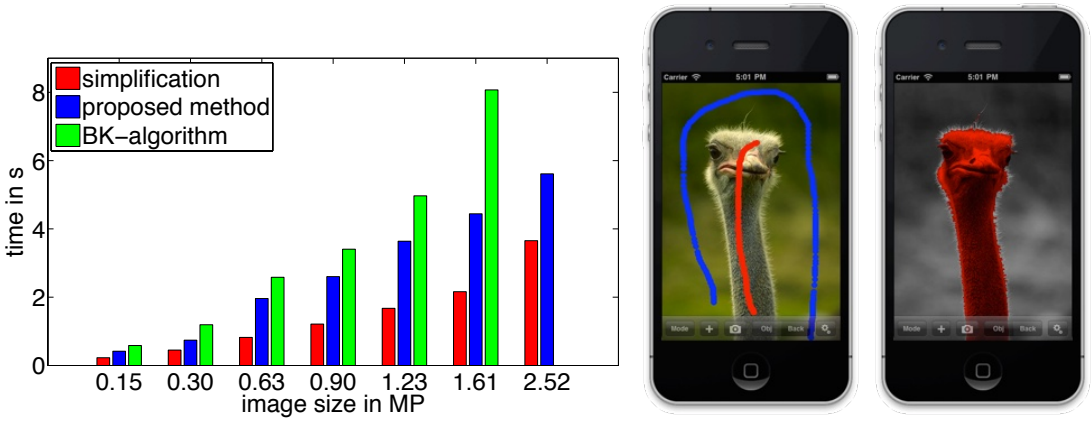
Figure 5.8: *High-resolution images:* Running time with one image and image sizes up to 25.84 MP. Up to an image size of 8.54 MP, the proposed method was approximately two times faster. For larger images, the BK-algorithm exceeded the physical memory so that the proposed method was approximately 877 times faster. On the original image size of 25.84 MP the computation time of the BK-algorithm was 38 minutes. The proposed method required 2.6 seconds. This time already includes the graph simplification step.

inefficient and unusable, if the graph does not fit into the physical memory. This can be observed in Figure 5.8 by the increasing running time at approximately 8.5 MP. Due to this limitation the algorithm is greatly extended by the proposed method. It has to be mentioned, that this problem is not fully solved by the *Slim Graph* but it can reduce the size of the data drastically so that *Slim Graphs* of high resolution images usually fit into the physical memory.

Experiments on resource-limited systems

The running time of the BK-algorithm was also measured on Apple's iPhone 4 with 512MB Ram. Therefore, 7 different sized benchmark images from 0.15 MP up to 2.52 MP were used. The average speedup of using *Slim Graphs* was approximately 32%. The limitations of the physical memory prohibited a comparison of larger images. The results of the experiments are shown in Figure 5.9a. Using *Slim Graphs* it is possible to segment images with up to 2.5 MP in 6 seconds on an iPhone 4, while using the original graph it is only possible to compute segmentations for images with up to 1.6 MP in approximately 8 seconds. The biggest speedup of approximately 45% was reached on an image with 1.61 MP, because the number of unlabeled nodes could be reduced from 1.61 million to 446951.





(a) (b)

Figure 5.9: *Resource-limited systems:* (a): Running time in seconds over 7 different sized images. The average speedup using *Slim Graphs* was approximately 36%. Running the BK-algorithm without using *Slim Graphs* was not possible on images bigger than 1.6 MP. The running time of the proposed method includes the graph simplification. (b): Example segmentation using Apple's iPhone 4 and a benchmark image.

5.2 Efficient Pixel Grouping with Dempster's Theory of Evidence

The basic idea of the SlimCut algorithm, proposed in the previous chapter, is to contract only simple edges so that the maximum flow and thus the segmentation result stays the same. Besides the property of maintaining the maximum flow, the effectiveness depends on the number of existing simple edges and in the worst case, no simple edges exist and the graph is not simplified at all. This effect can be observed in the experiments on small-scale images using different initializations. Having a rectangular initialization, which means imprecise prior models, the speedup is rather small.

In the following, another algorithm for image segmentation using graph cuts is proposed that aims to efficiently solve labeling problems on high resolution images or image sequences. The basic idea of this algorithm is to group large homogeneous regions to one single variable. Therefore, the appearance and the task specific similarity is combined using Dempster's theory of evidence to compute the basic belief that two pixels or groups will have the same label in the minimum energy state. In contrast to the SlimCut algorithm this grouping can lead to different segmentation results. Experiments on image and video segmentation show that the proposed

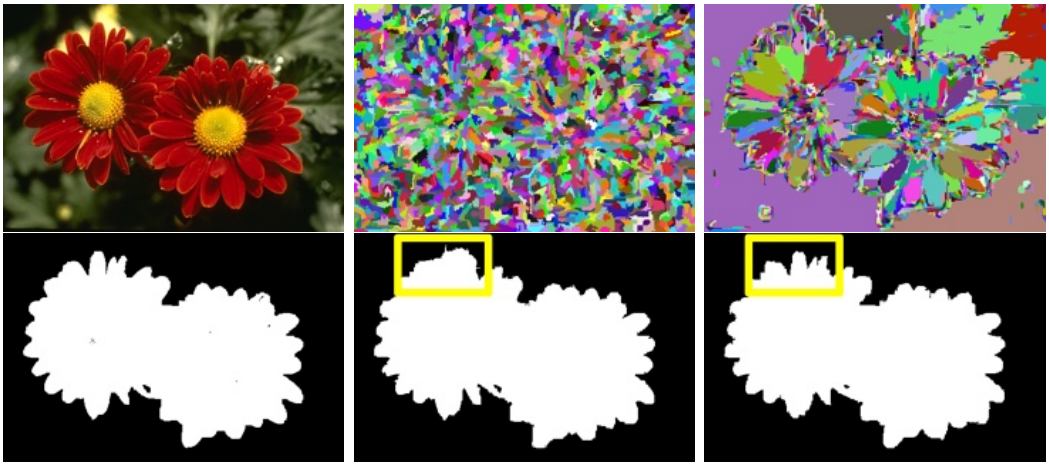
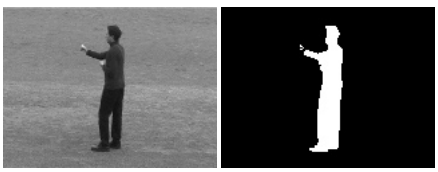


Figure 5.10: Variable grouping for image segmentation. First row: original image; variable grouping of [KNKY11] with a budget of 1%; proposed variable grouping (COMPACTEDGE) with a budget of 1%; Second row: corresponding segmentation results. Using the same budget the proposed grouping is semantically more meaningful and leads to a smaller segmentation error.

grouping leads to a significant speedup and memory reduction of the labeling problem whereas the segmentation is comparable. Thus large-scale labeling problems can be solved in an efficient manner with a small approximation loss.

Contribution

The number of variables of the energy function is reduced by an algorithm that merges variables to small sets of non overlapping groups, so that each group can be represented by one single variable. The merging follows the idea of [FH04] and [KNKY11]. In [FH04], Felzenszwalb and Huttenlocher proposed a very well-known superpixel algorithm, where the grouping is based on appearance. Kim et al. [KNKY11] proposed to group pixels based on the terms of the energy function. In contrast to the work of Kim et al. [KNKY11], the task-specific similarity and the appearance of neighboring pixels / groups is combined using Dempster's theory of evidence. Using this theory allows to compute the basic belief that two neighboring variables should be merged. Furthermore the size of a group is not directly penalized by proposing new merging constraints (MAXEDGE and COMPACTEDGE), that follow the idea to allow large groups of variables in homogeneous regions. Thus, the number of variables can be reduced drastically. Instead of an accurate maximum a posteriori solution (MAP), the goal of the proposed algorithm is to reduce the segmentation error. Therefore Dempster's theory of evidence, that is complementary to the terms of the energy function, is used for the feature fusion. The proposed method is eval-



uated on standard benchmark images to show that the grouping achieves a better performance than the methods of [FH04] and [KNKY11]. Furthermore the algorithm is quantified on video sequences and high-resolution images to show that the segmentation, performed on top of the grouping, results in a similar segmentation with a dramatic reduction in computational costs and memory requirements.

5.2.1 Variable Grouping based on Dempster's Theory of Evidence

This section describes the details of the proposed variable grouping and shows the similarities and differences to existing approaches. For the grouping of the variables, the definitions given in [KNKY11] and the notation in [SSR12] is used. A variable grouping of graph G is a graph $G' = (\mathcal{V}'_G, \mathcal{E}'_G)$ with energy function E' produced by a surjective map $m_G : \mathcal{V}_G \to \mathcal{V}'_G$ and the edge set $\mathcal{E}'_G = \{(s,t) \in \mathcal{V}'_G \times \mathcal{V}'_G \mid \exists (i,j) \in \mathcal{E}_G : m_G(i) = s \text{ and } m_G(j) = t\}$. Thus, the energy function for a variable grouping G' reads:

$$E'(\mathcal{L}) = \sum_{i \in \mathcal{V}} \tau_i(\hat{\mathcal{L}}_{m_G(i)}) + \sum_{(i,j) \in \mathcal{E}} \tau_{i,j}(\hat{\mathcal{L}}_{m_G(i)}, \hat{\mathcal{L}}_{m_G(j)}) \,, \tag{5.6}$$

where $\hat{\mathcal{L}}$ is the labeling of the variable grouping. Solving this energy function on top of the grouping can be seen to correspond to the existing practice of using superpixels as a preprocessing step and defining the energy minimization problem on superpixels instead of pixels. Since most superpixels are directly derived from image properties, they perform poorly because the properties of the energy function, e.g. the unary term, are ignored. Figure 5.11 shows an example of a variable grouping and the corresponding graph based on the new energy function.

The idea of grouping nodes is as follows: A score function w_{ij}, measuring how similar two connected nodes i and j are, is assumed. That function is chosen so that small values indicate a strong similarity and large values dissimilarity.

(i) the first step is to sort all edges of the graph in ascending order so that edges with a small weight come first,

(ii) for each edge in the list nodes that fulfill a given constraint are merged until the problem is sufficiently reduced.

The efficient graph-based segmentation method, proposed by Felzenszwalb and Huttenlocher [FH04], is an example for such a grouping algorithm. To balance the size of a group and its internal coherence, a global criterion is used to decide if two nodes or groups of nodes can be merged. Algorithm 5.1 is identical to [FH04] and [KNKY11] using a slightly different notation. The merging constraint used in [FH04] and [KNKY11] is based on the so-called internal difference

$$\mathrm{Int}(C) = \max_{(i,j) \in \mathrm{MST}(C,\mathcal{E})} w_{ij} \,,$$

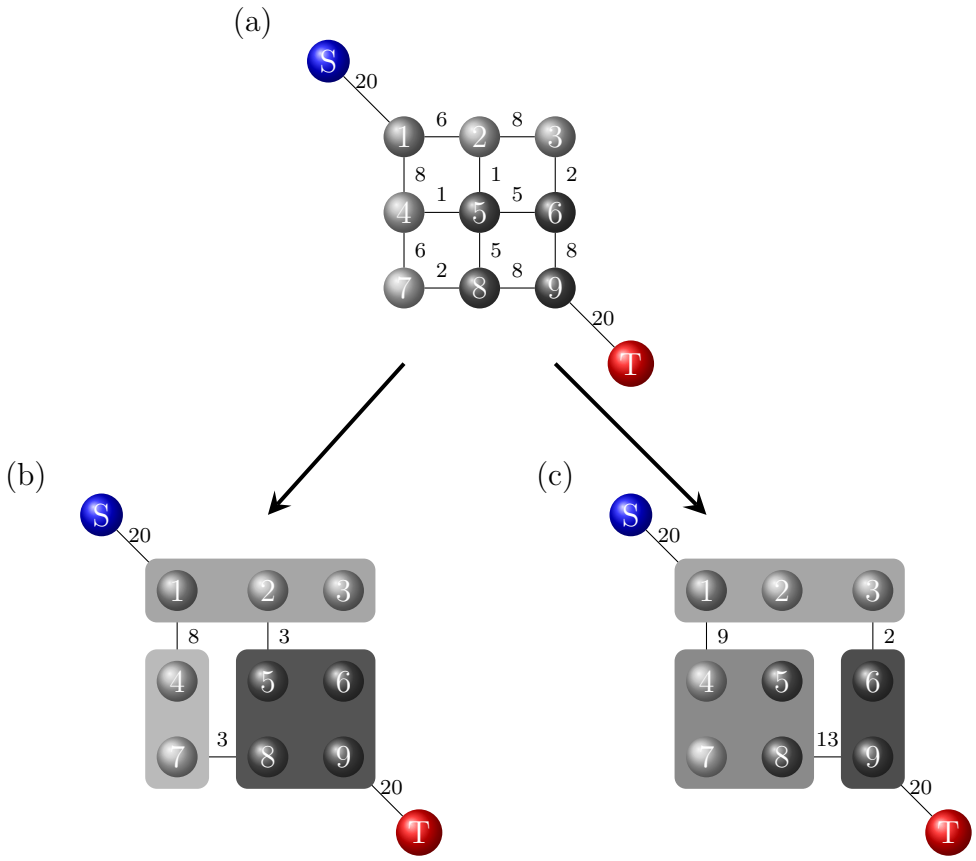
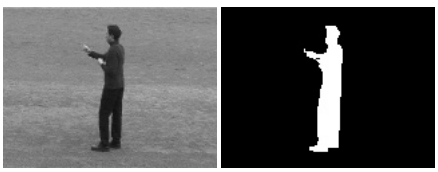


Figure 5.11: Example variable groupings. The nodes from the original graph (a) are merged into three different groups of variables (b) and (c), respectively. The weights of the new graph are changed according to the new energy function. A good grouping (b) does not change the MAP solution of the original graph.

where MST(C,$\mathcal{E}$) is the minimum-weight spanning tree within the group of nodes C with a set of edges $\mathcal{E}$. Int(C) is small if the nodes in group C are similar according to the defined edge weights. To decide whether two groups are merged, the algorithm compares the weight of the connecting edge between the two groups C_1 and C_2 and compares it with the internal difference Int(C_i) of both groups. For the goal of grouping variables for energy minimization, this criterion makes sense since groups of variables should be similar and agree about their labeling. For the decision, [FH04, KNKY11] use the function MInt(C_1,C_2) defined as

$$\text{MInt}(C_1,C_2) = \min\{\text{Int}(C_1) + \tau(C_1), \text{Int}(C_2) + \tau(C_2)\}\,,$$

where $\tau(C) = \frac{k}{|C|}$ penalizes the size of a group based on a free parameter k. Accord-



Algorithm 5.1: Dempster-Shafer based Variable Grouping

1: $(\mathcal{V}_\mathcal{G}', m)$ = DempsterShaferGrouping(G,φ, w)
2: **Input:**
3: $G = (\mathcal{V}_\mathcal{G}, \mathcal{E}_\mathcal{G})$ // an instance of the graph
4: $\varphi_i, \varphi_{i,j}$ // node and edge energies
5: $w : \mathcal{E}_\mathcal{G} \to \mathbb{R}$ // dissimilarity weights
6: **Output:**
7: $\mathcal{V}_\mathcal{G}'$ // set of grouped variables
8: m // surjective map
9: **Algorithm:**
10: $\mathcal{V}_\mathcal{G}' \leftarrow \mathcal{V}_\mathcal{G}, \mathcal{E}_\mathcal{G}' \leftarrow \mathcal{E}_\mathcal{G}$
11: $m \leftarrow \{(i,i) \mid i \in \mathcal{V}_\mathcal{G}\}$
12: $\pi \leftarrow \text{sort}(\mathcal{E}_\mathcal{G}, w)$ // sort weights in ascending order
13: **for** $e = 1, \ldots, |\pi|$ **do**
14: $(i,j) \leftarrow \pi_e$
15: **if** $m(i) = m(j)$ **then**
16: **continue** // nodes already merged
17: **end if**
18: **if** w_{ij} fulfills given constraint **then**
19: merge C_i and C_j in $m, \mathcal{V}_\mathcal{G}'$
20: **end if**
21: **end for**

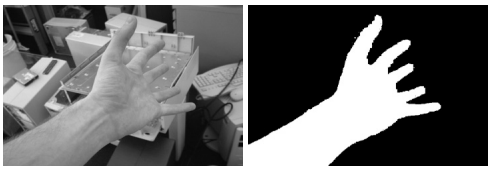
ing to Algorithm 5.1, when edge $w_{ij} \in \mathcal{E}_\mathcal{G}$ fulfills the equation

$$w_{ij} \leq \text{MInt}(C_i, C_j) \tag{5.7}$$

C_i and C_j are merged. As mentioned in [FH04], this graph based method is very efficient and easy to implement in $O(|\mathcal{E}_\mathcal{G}| \log |\mathcal{E}_\mathcal{G}|)$ time and memory.

Merging Function

The grouping resulting from the algorithms in [FH04] and [KNKY11] can be described as compact since the free parameter k in $\tau(C)$ penalizes the size of a group. In [KNKY11] the goal was to produce compact groups of variables that will have the same label according to the minimum energy state. Therefore the weight functions are based on the unary or pairwise potentials of the energy function. In contrast, the proposed algorithm should group as many variables as possible that are likely to have the same label according to the minimum energy state and to the ground truth labeling.



To allow big groups of variables, e.g. in homogeneous regions, new merging constraints based on the maximum weight among outgoing edges are proposed. Instead of using a global criterion, balancing the size and the internal coherence of a group all nodes that are connected by a sufficiently small edge are merged. E.g. one could use the function $w_{ij} \leq W$ to merge all nodes connected by an edge smaller than the parameter W. However, it is easy to see that such a simple constraint does not produce groups of pixels that agree with either the minimum energy state or the ground truth segmentation. To produce groups of homogeneous variables, two new merging constraints based on the local edge weights of two nodes are proposed. The first constraint takes into account the maximum value of any edge connected to one of the two nodes. Therefore two components connected by the edge w_{ij} are grouped if

$$w_{max}(i,j) := \max\{w_{ik}, w_{lj} \mid (i,k),(l,j) \in \mathcal{E}_{\mathcal{G}}\} \leq W_1 \qquad (\text{MAXEDGE}). \tag{5.8}$$

This means that two nodes are merged if the weights of all edges adjacent to the edge (i,j), including the edge w_{ij}, are smaller than the parameter W_1, which indicates that these nodes are somewhat similar. The threshold W_1 is computed according to the distribution of the edge weights (e.g. so that 66% of the edge weights are smaller than W_1). Thus, the amount of reduction can be controlled implicitly by the parameter W_1. The idea of the proposed constraint is to have large groups of variables in all images regions except the borders of the objects. If a node (pixel) is near the border of an object there should be one edge with a high weight. With (5.8) this edge guarantees that the node is not merged with any neighbor.

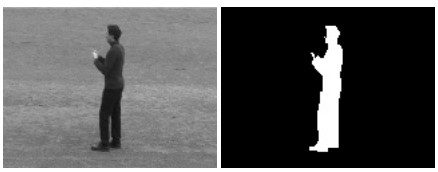
To allow somehow small compact groups of variables in regions that do not fulfill the MAXEDGE constrained, e.g. at the borders of an object or in noisy image regions, a second merging constraint is proposed. This constraint combines the MAXEDGE property with the global constraint based the minimum-weight spanning tree and the size dependent function τ, to balance the size of a group and its internal coherence. Thus, the decision is made according to

$$\text{MAXEDGE} \;\text{ or }\; w_{ij} \leq \text{MInt}(C_i,C_j) \qquad (\text{COMPACTEDGE}), \tag{5.9}$$

The differences of the proposed merging functions are discussed in the experiment section.

Weight Functions

Besides the merging constraint, the weight function w_{ij}, is the most important measure to decide which nodes are merged. Three classes of weight functions w_{ij} are considered. The first two are well known weight functions that shall serve as comparison with the proposed one.



Felzenszwalb and Huttenlocher: In [FH04] Felzenszwalb and Huttenlocher take the pixel difference as the grouping weight. If I_i and I_j are the feature vectors of pixels i and j in the image, the weight is set to the norm of the difference:

$$w_{ij}^{FH} = ||I_i - I_j|| \,. \tag{5.10}$$

The experiments on image segmentation will show that this measure is not performing comparably, since the properties from the energy minimization problem are ignored.

Kim et al.: An approach very similar to [FH04] and the proposed one, was proposed by Kim et al. in [KNKY11]. For comparison with the proposed method the defined UNARYDIFF weight function is used. In the experiments on standard benchmark images this weight function outperformed the others for the problem of binary image segmentation. The UNARYDIFF weight is defined as

$$w_{ij}^{ud} = ||\varphi_i - \varphi_j|| \,, \tag{5.11}$$

using the unary terms of the defined energy function. The weight describes the disagreement of the states between two variables and measures the task-specific similarity of two neighboring nodes.

Dempster-Shafer weighting function: The proposed weight function includes the unary potentials τ_i and τ_j and the pairwise potential τ_{ij} of nodes i and j. Thereby, the image information that is included in the pairwise potential and the information included in the unary potential, typically derived from a discriminative classifier is taken into account. Hence the proposed weight function can be seen as a combination of the two earlier presented ones which combines the image features with the task specific similarity. To combine both information Dempster's theory of evidence is used, which is also different from the aforementioned approaches. Therefore weights based on the unary and pairwise potentials are defined as

$$w_{ij}^{pairwise} = \tau_{ij}(x_i, x_j) \quad \text{and} \quad w_{ij}^{unary} = ||\tau_i - \tau_j|| \,. \tag{5.12}$$

Since the co-domains of the weights are different, they are normalized individually to [0,1]. That means for two variables with a similar feature vector $w_{ij}^{pairwise} \approx 1$ and $w_{ij}^{pairwise} \approx 0$ if the feature vectors are different. For w_{ij}^{unary} it means $w_{ij}^{unary} \approx 0$ if the negative log likelihood for two variables is similar for both states. Based on these weight functions, two mass functions over the hypothesis set $\Omega = \{\Omega_1, \Omega_2\}$ are defined. In this context, the hypothesis Ω_1 means that the two variables are similar and Ω_2 that they are dissimilar:

$$\begin{aligned}
&m_1(\Omega_1) = b_1 \cdot w_{ij}^{pairwise} \quad , m_1(\Omega_2) = b_1 \cdot (1 - w_{ij}^{pairwise}) \,, \\
&m_1(\emptyset) = 0 \quad , m_1(\Omega) = b_1 \,, \\
&m_2(\Omega_1) = b_2 \cdot (1 - w_{ij}^{unary}) \quad , m_2(\Omega_2) = b_2 \cdot w_{ij}^{unary} \,, \\
&m_2(\emptyset) = 0 \quad , m_2(\Omega) = b_2 \,,
\end{aligned} \tag{5.13}$$

where b_i describes the belief of the different information sources. In all the experiments the believe is weighted equally with $b_1 = b_2 = 0.5$. Now the two mass functions are combined with Dempster's rule of combination (3.47) and define the weights:

$$\begin{aligned} w_{ij}^{DS} &= 1 - Bel(\Omega_1) = 1 - m(\Omega_1) = 1 - m_1(\Omega_1) \otimes m_2(\Omega_1) \\ &= 1 - \left(\frac{m_1(\Omega_1) \cdot m_2(\Omega_1) + m_1(\Omega_1) \cdot m_2(\Omega) + m_1(\Omega) \cdot m_2(\Omega_1)}{1 - (m_1(\Omega_1) \cdot m_2(\Omega_2) + m_1(\Omega_2) \cdot m_2(\Omega_1))} \right) \end{aligned} \tag{5.14}$$

In contrast to [KNKY11], the proposed weight function allows for the combination with other information sources, such as the user initialization, the optical flow in video sequences, depth images or appearance information of an object.

5.2.2 Experiments

The proposed grouping allows to compute an approximate segmentation result. Since the resulting graph used for the energy minimization process is much smaller, the segmentation result differs from the original MAP-solution. Since the goal of a segmentation scheme is an accurate segmentation, that corresponds to a low segmentation error, and not an accurate MAP-solution, the proposed grouping is quantified using three performance measures:

(i) the segmentation quality with respect to the ground truth solution,

(ii) the minimum possible segmentation error of the grouping with respect to the ground truth segmentation,

(iii) and the ratio of runtimes solving the MAP-problems (including the time for the grouping).

Following, the three measures are described in detail:

Segmentation error: The segmentation error is defined analogously to [BRB+04] as the ratio between the number of misclassified pixels and the number of pixels in unclassified regions:

$$R_{se}(\mathcal{L}) = \frac{\sum_{i \in \mathcal{V}_\mathcal{G}} [\mathcal{L}_i \neq \mathcal{L}_i^{gt}]}{\text{no. pixels in unclassified regions}}, \tag{5.15}$$

where $\mathcal{L}^{gt}$ is the ground truth labeling. Thus, the segmentation error is low, if most pixels are assigned the right label. The normalization using unclassified image regions takes into account the prior information (user labels), that are assumed to be correct. This is the so-called Hamming distance of the segmentation result and the ground truth segmentation.

Minimum segmentation error: Another measure to quantify the quality of a grouping is given by the minimum segmentation error, that counts the minimum

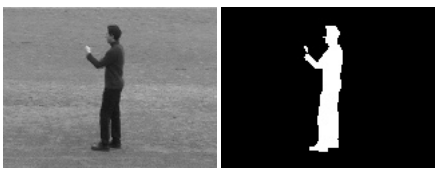
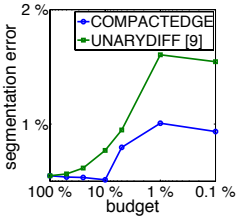


Table 5.1: Comparison of the proposed algorithm and two similar methods proposed in [FH04] and [KNKY11]. All values are averaged over 50 benchmark images using stroke (lasso) initializations. As can be seen our proposed method COMPACTEDGE performed best in terms of quality with a smaller budget. The proposed MAXEDGE has the lowest minimum segmentation error with the drawback of a bigger budget.The graph visualizes $R_{se}(\mathcal{L})$ for one image and different budgets.

Method	Avg. budget	Avg. $R_{mse}(\mathcal{L})$	Avg. $R_{se}(\mathcal{L})$
full MAP (reference)	100 (100)	0 (0)	0.075 (0.058)
FH-alg [FH04]	10.22 (10.22)	209.74 (209.74)	0.074 (0.063)
UNARYDIFF [KNKY11]	10.72 (10.84)	255.1 (219.08)	0.073 (0.065)
MAXEDGE	47.72 (15.21)	58.42 (4.21)	0.069 (0.058)
COMPACTEDGE	6.25 (5.00)	321.5 (63.52)	0.061 (0.058)



number of misclassified pixels by an optimal segmentation.

$$R_{mse}(\mathcal{L}) = \sum_{i \in \mathcal{V}'_{\mathcal{G}}} \min \left(\sum_{j \in m^{-1}_{\mathcal{V}'_{\mathcal{G}}}(i)} [\mathcal{L}^{gt}_j = FG], \sum_{j \in m^{-1}_{\mathcal{V}'_{\mathcal{G}}}(i)} [\mathcal{L}^{gt}_j = BG] \right) . \tag{5.16}$$

This measure looks for groups of pixels that contain foreground and background pixels. Normalized by the number of pixels in unclassified image regions, see Equation (5.15), this is a lower bound for the segmentation error.

Ratio of runtimes: To compute the ratio of runtimes, the time to compute the grouping and solve the reduced problem is compared with the time solving the original sized problem.

The evaluation of the proposed method is divided in three different experiments.

(i) standard small-scale images,

(ii) high-resolution images and

(iii) video sequences.

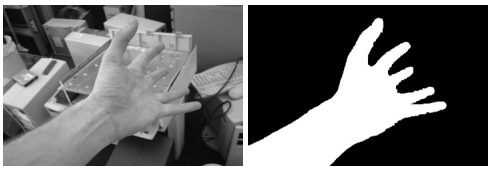
To evaluate the grouping on small-scale images, the Microsoft segmentation benchmark proposed by Blake et al. [BRB+04][3,4] is used. Since there exist no segmentation benchmark with large-scale problems, high-resolution images with up to 26 million pixels found on the web are used for the evaluation. For the problem of binary video segmentation video sequences from the KTH action dataset [SLC04][5] and videos provided by Sand and Teller [ST06][6] are used. To compare the segmentation results produced on top of the grouping with the segmentation results of the

[3] http://research.microsoft.com/en-us/um/cambridge/projects/visionimagevideoediting/segmentation/grabcut.htm

[4] http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/

[5] http://www.nada.kth.se/cvap/actions/

[6] http://rvsn.csail.mit.edu/pv/



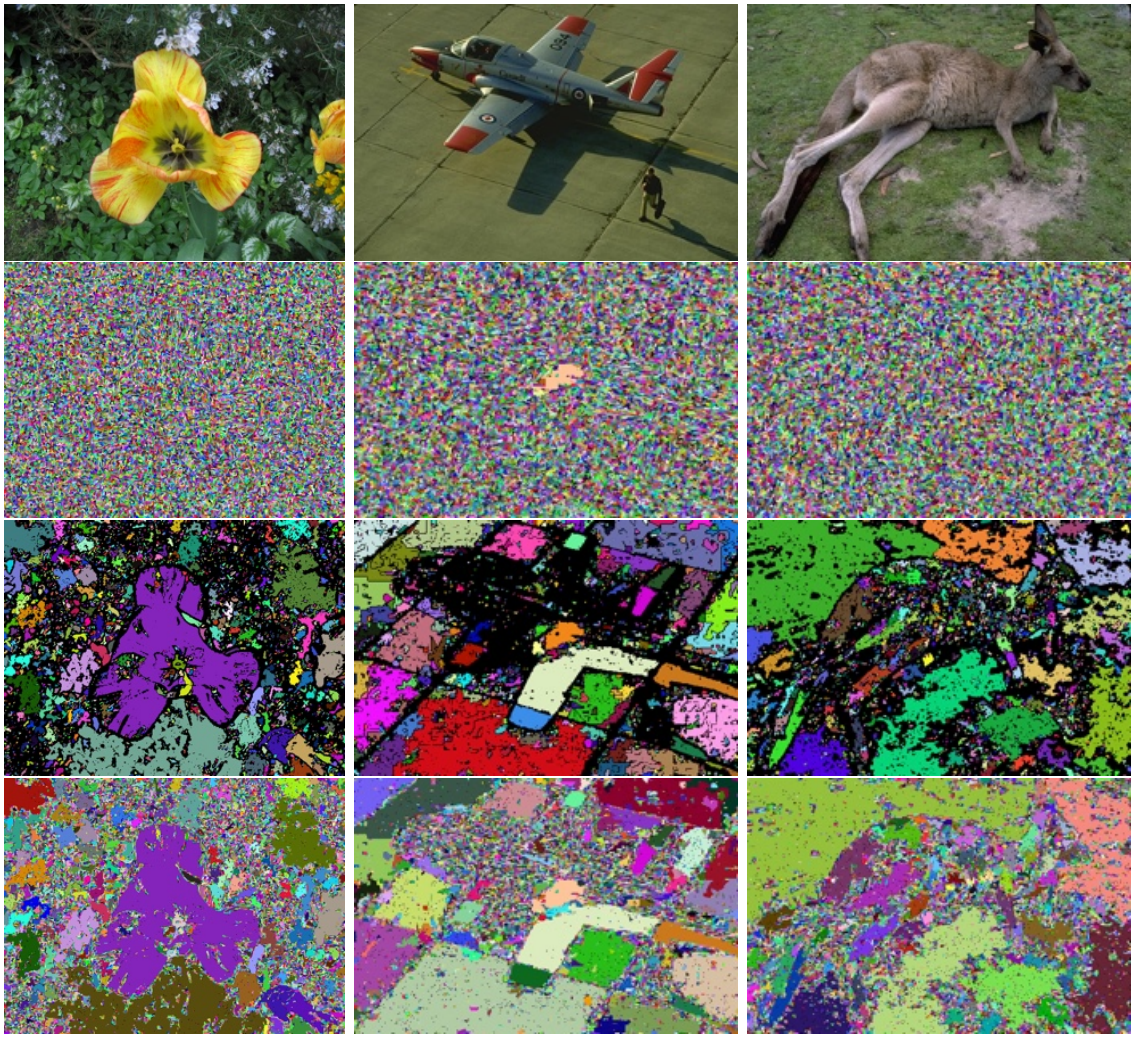
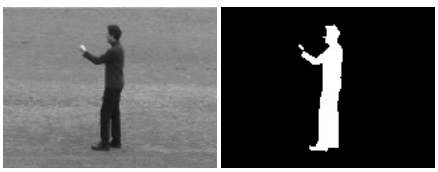


Figure 5.12: Example for the different approaches for variable grouping. Rows: (i) original image; (ii) variable grouping using [KNKY11]; (ii) proposed method using MAXEDGE; (iv) proposed method using COMPACTEDGE; In contrast to [KNKY11] where the grouping produces superpixels that are comparable in size our proposed methods group large homogeneous regions to single variables.



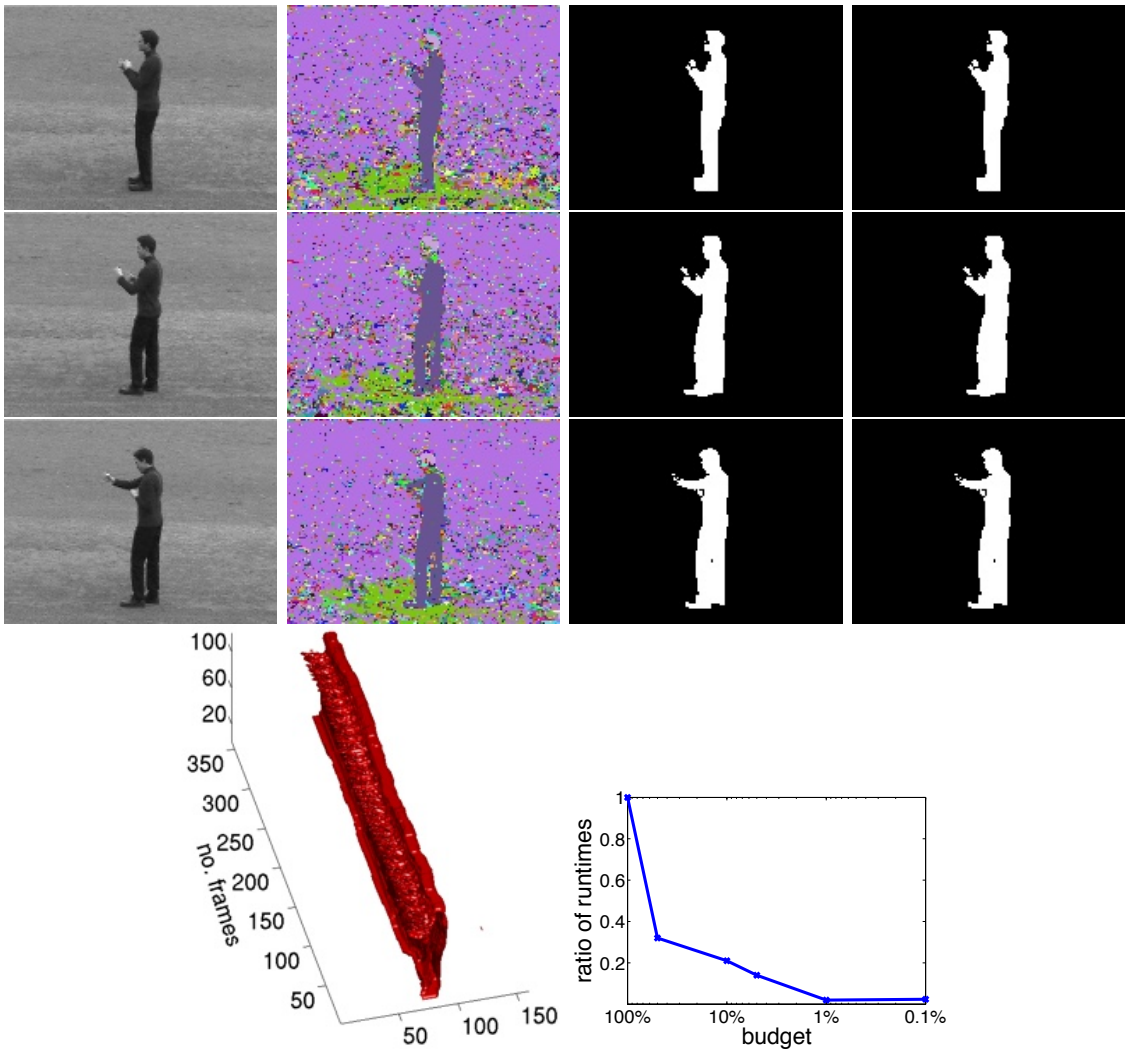


Figure 5.13: Variable grouping for video segmentation. The first rows correspond to the frames 20, 220 and 350 of the Boxing sequence from [SLC04]. The last row visualizes the isosurface of the segmentation result and the ratio of runtime for different budgets. Columns: (i) original frame; (ii) variable grouping with the proposed algorithm; (iii) segmentation result solving the full MAP; (iv) segmentation result solving the approximated MAP. The segmentation results are almost identical even if the approximated solution used a budget of 10%. The ratio of runtime for this example is ≈ 0.21.

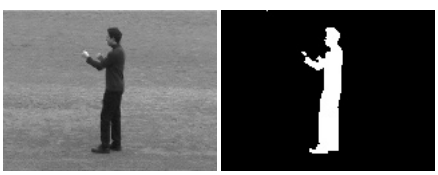
MAP-problem, the same energy function proposed by Blake et al. [BRB+04] and the same set of parameters are used. The experiments were run on a MacBook Pro (Mid 2010) with 2.4 GHz Intel Core i5 processor and 4GB Ram. For all experiments the proposed algorithm is compared with the approaches of Felzenszwalb and Huttenlocher [FH04] and Kim et al. [KNKY11].

Small-scale images: Table 5.1 shows the evaluation of the proposed algorithm on the Microsoft segmentation benchmark in comparison to the works of Felzenszwalb and Huttenlocher and Kim et al.. Since independently the benefit of the proposed weight w_{ij}^{DS} and the merging constraints is rather small, only the combination that outperformed existing approaches is evaluated.

It is observable, that the combination of Dempster's theory of evidence and the proposed constraint has a smaller average segmentation error with an even smaller budget. The small minimum segmentation error using the MAXEDGE constraint highlights that the idea to group large homogeneous regions to one single variable makes sense and the proposed weights based on Dempster's theory of evidence reliably find those regions. In combination with small groups at the objects boundaries the proposed COMPACTEDGE constraint clearly outperforms the existing approaches. Figure 5.12 presents a visual comparison of the different approaches.

High-resolution images: To evaluate the segmentation quality and the possible speedup of the proposed method high-resolution images with more than 20 MP are used and down sampled to several image sizes. Similar to the experiments on small-scale images and video sequences the difference in segmentation quality is small and the reduction of runtime is dramatic for those high-resolution images. As already shown by Delong and Boykov [DB08] the BK-algorithm is inefficient and unusable if the graph does not fit into the physical memory. For those large MAP inference problems the ratio of runtime was approximately 0.08 using a budget of 5%. Due to the limitations of the BK-algorithm the proposed method greatly extends its applications.

Video sequences: The proposed algorithm can also be applied to group variables for the problem of video segmentation using a three dimensional pixel neighborhood. To evaluate the performance of the proposed method different video sequences are segmented. It can be seen from Figures 5.13 and 5.14 that the proposed algorithm achieves a similar segmentation like the full MAP solution with a much smaller budget and a dramatic reduction of runtime. E.g. for the hand video in Figure 5.14 (200 frames) the number of variables is reduced from 69.1 million to 3.5 million. For a visual comparison of the results, only the first 40 frames of the hand sequence are used. This is due to the fact, that the full MAP solution was only computed for 40 frames since solving the full MAP problem for all 200 frames was not possible due to memory reasons. The full MAP problem for the KTH-sequence shown in Figure 5.13 has seven million variables and the results shown use a budget of approximately 10% resulting in 0.7 million variables with a comparable segmentation result. In all



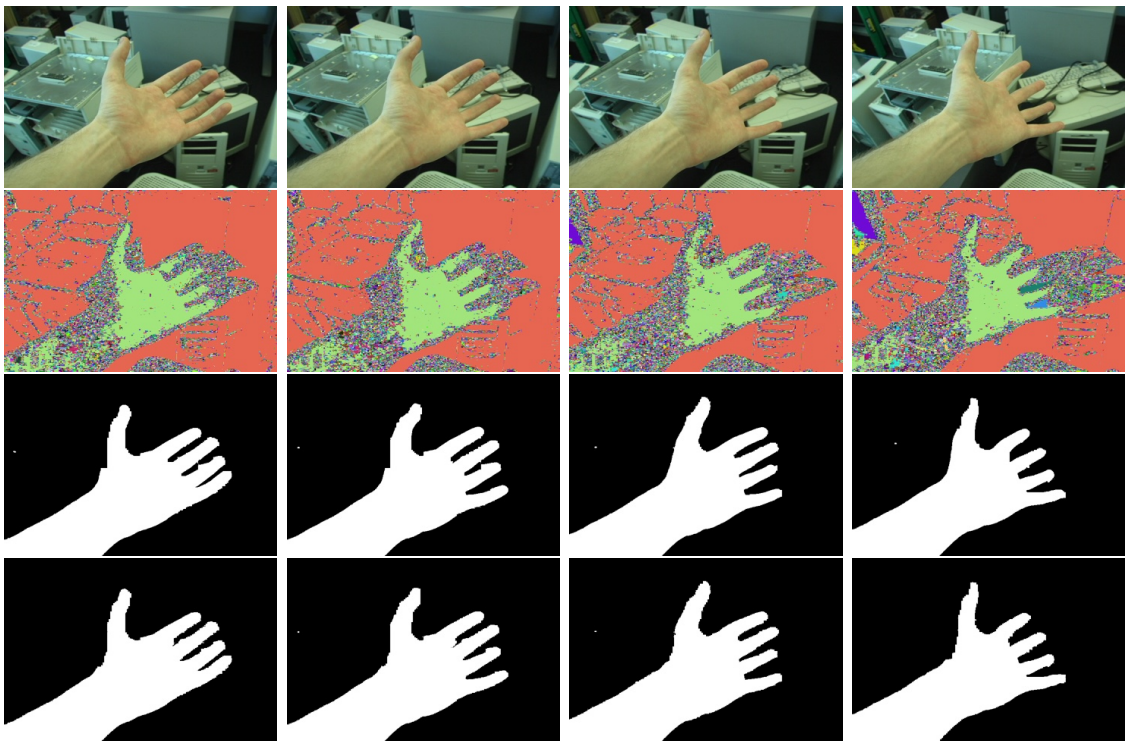


Figure 5.14: Variable grouping for video segmentation. The columns correspond to the frames 5, 15, 25 and 38 of the hand sequence [ST06]. Rows: (i) original frame; (ii) variable grouping with the proposed algorithm; (iii) segmentation result solving the full MAP; (iv) segmentation result solving the approximated MAP. The segmentation results are almost identical even if the approximated solution used a Budget of 5%. The ratio of runtime for this example is ≈ 0.1.

examples the segmentation algorithm is initialized using a few strokes in the first frame[7].

5.3 Discrete Energy Function including Dempster's Theory of Evidence

In the previous chapters, the focus was how to speed up discrete segmentation methods using SlimCuts or variable grouping. This chapter now shows how to integrate Dempster's theory of evidence for the feature fusion into a discrete energy minimizing framework.

[7]Parts of the video segmentation results are included in this thesis as a flip-book.

Segmentation of foreground objects in video sequences is a fundamental step in many computer vision applications and has been widely studied in the last years. A popular application in movie post production is the integration of virtual objects into a sequence [CSRO12]. Because of many aspects in real-world scenarios video segmentation is a very challenging task. Illumination changes or background appearance changes, caused by people walking around, are typical problems that need to be treated.

Time-of-Flight (ToF) cameras are perfect candidates to simplify the problem of binary video segmentation. ToF cameras use active sensors to measure the time taken by infrared light to travel to the object and back to the camera. The travel time corresponds to a certain depth value. Thus, ToF cameras are able to determine the depth value for the pixels in an image, which can be seen as additional information for each pixel. Typically, the depth information is less sensitive to environment changes. Combined with appearance, this yields a more robust segmentation method. Motivated by the fact that a simple combination of two information sources might not be the best solution, a novel scheme based on Dempster's theory of evidence is proposed. In contrast to existing methods, the use of Dempster's theory of evidence allows to model inaccuracy and uncertainty. The inaccuracy of the information is influenced by an adaptive weight, that provides a measurement of how reliable a certain information might be.

The proposed algorithm is related to many recent works on binary video segmentation [HGW01, KCB+05, CCBK06, WZYZ10]. In [HGW01] and in [KCB+05], stereo images where used to estimate the scene depth. They showed that the combination of estimated depth and color improves the segmentation result. However, the estimation of the scene depth is a non trivial problem that is prone to errors in real-world scenarios.

The most related method is the so-called ToFCut method proposed by Wang et al. [WZYZ10]. They combine depth and color cues in a discrete energy function and weight the information adaptively.

Contribution

In contrast to ToFCut, a novel method to fuse color and depth information in a discrete energy function is proposed. Therefore, Dempster's theory of evidence is integrated in the discrete energy function. Using the proposed feature fusion within Dempster's framework allows to explicitly model inaccuracy and uncertainty, see Figure 5.15. This modeling provides an elegant way to incorporate the reliability of a feature channel. The information about how reliable a feature channel might be, can be either defined manually, based on prior information, or using our proposed adaptive weighting function. The adaptive weighting uses the symmetric Kullback-Leibler divergence as a measure of reliability. Therefore, distances of foreground

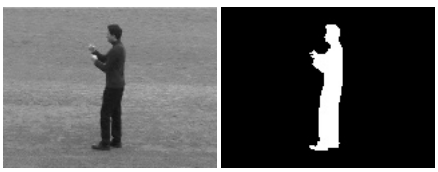


Figure 5.15: Example segmentation result by fusing color and depth information using Dempster's theory of evidence. The explicit modeling of uncertainty forces the algorithm to segment the person in the foreground even if the depth information of the person in the background is similar. Input data taken from [WZYZ10].

and background histograms based on the segmentation result of the previous frame are computed. In comparison with the adaptive weighting of ToFCut, the proposed weighting is more stable. The experimental validation on the data set used in [WZYZ10] shows that the proposed method outperforms ToFCut.

5.3.1 Feature Fusion using Dempster's Theory of Evidence

This section shows how to integrate the feature fusion, based on Dempster's theory of evidence, in a discrete energy function.

In [WZYZ10], the discrete energy function, see Equation (3.26), is extended by means of additional depth information. Therefore, the unary potential has the form:

$$\tau_i(\mathcal{L}_i) = -\gamma_c \cdot \log p_c(I(i) \mid \mathcal{L}_i = S) - \gamma_d \cdot \log p_d(D(i) \mid \mathcal{L}_i = S)\,, \tag{5.17}$$

where $D(i)$ is the depth of pixel i and S is either foreground or background. The likelihood p_c is a Gaussian Mixture Model learned using 3D histograms with 8^3 bins in the RGB color space and the likelihood for depth p_d is modeled by two Gaussian distributions. The parameters γ_c and γ_d are used to adaptively weight the impact of both cues. They are based on the discriminative capabilities of the two likelihoods. The color confidence is computed using the Kullback-Leibler divergence (KL) between the gray-scale histograms of frames I^{t-1} and I^t (denoted by δ^{KL}_{lum}) and the KL divergence between foreground and background color histograms of frame I^{t-1} (δ^{KL}_{rgb}) . This yields the confidence of the color term

$$\mathcal{R}_c = \exp\left(-\frac{\delta^{KL}_{lum}}{\eta_{lum}}\right) \cdot \left(1 - \exp\left(-\frac{\delta^{KL}_{rgb}}{\eta_{rgb}}\right)\right)\,, \tag{5.18}$$

with parameters η_{lum} and η_{rgb}. The depth confidence $\mathcal{R}_d$ is computed using the distance between the average depth values for foreground and background in frame

I^{t-1} ($\Delta\mu = |(\mu^f + \mu'^f) - (\mu^b + \mu'^b)|/2$). Here, μ^f, μ'^f, μ^b and μ'^b are the mean values of the Gaussian distributions p_d. This yields

$$\mathcal{R}_d = 1 - \exp\left(-\frac{\Delta\mu}{\eta_d}\right), \tag{5.19}$$

with the additional parameter η_d. Finally, the adaptive weights are defined by normalizing the confidences: $\gamma_c = \mathcal{R}_c/(\mathcal{R}_c + \mathcal{R}_d)$ and $\gamma_d = \mathcal{R}_d/(\mathcal{R}_c + \mathcal{R}_d)$. For more details on the likelihood terms and the adaptive weighting the reader is referred to [WZYZ10].

In contrast to the ToFCut approach, the symmetric Kullback-Leibler divergence is used, since the symmetric distance does not depend on the order of the feature channels. The symmetric KL divergence is also used to measure the distance between foreground and background depth histograms in frame I^{t-1}, since the given definition using $\Delta\mu$ lacks in precision.

The symmetric Kullback-Leibler divergence for two normalized histograms H_1 and H_2 is given by:

$$\delta^{sym.KL}(H_1,H_2) = \delta^{KL}(H_1,H_2) + \delta^{KL}(H_2,H_1), \tag{5.20}$$

where $\delta KL(H_1,H_2)$ is the standard Kullback-Leibler divergence defined by:

$$\delta^{KL}(H_1,H_2) = \sum_i H_1(i) \log\left(\frac{H_1(i)}{H_2(i)}\right). \tag{5.21}$$

The unary potential used by ToFCut is defined as a weighted sum of negative log likelihoods, see Equation (5.17), and can be reformulated as:

$$\tau_i(\mathcal{L}_i) = -\log\left[p_c(I(i)|\mathcal{L}_i = S)^{\gamma_c} \cdot p_d(D(i)|\mathcal{L}_i = S)^{\gamma_d}\right], \tag{5.22}$$

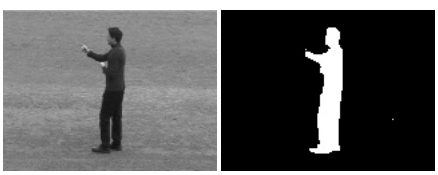
which can be interpreted as follows: if the confidence for a channel is near zero, the likelihood is near one, see Figure 5.16 That means, to ignore a channel the corresponding likelihoods are pushed near one. This is a neither intuitive nor elegant solution. Furthermore, this non-linear solution heavily depends on a good adaptive weighting function.

In contrast to ToFCut, the proposed unary potential is defined using Dempster's basic probability assignment:

$$\tau_i^{DS}(\mathcal{L}_i) = -\log m(\mathcal{L}_i = L), \tag{5.23}$$

where the mass function $m = m_c \otimes m_d$ fuses the information of color and depth according to Dempster's rule of combination. Thus the complete energy function reads:

$$E(\mathcal{L}) = \sum_{i\in\mathcal{V}} \tau_i^{DS}(\mathcal{L}_i) + \sum_{(i,j)\in\mathcal{E}} \tau_{i,j}(\mathcal{L}_i,\mathcal{L}_j), \tag{5.24}$$



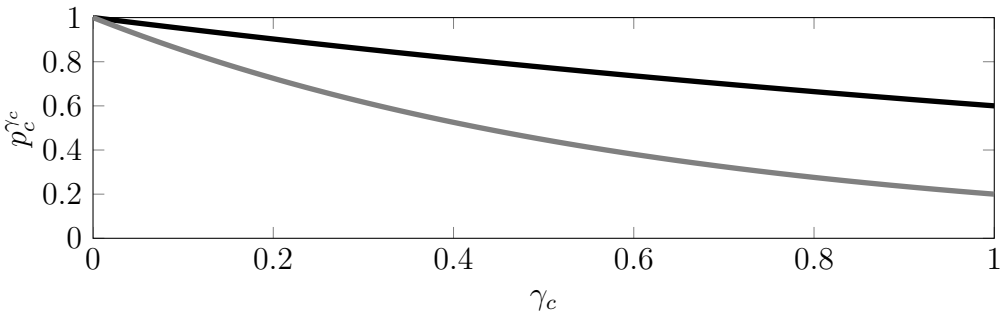


Figure 5.16: Unary color potential of ToFCut with varying confidence. If the color confidence γ_c decreases, the likelihood increases.

The proposed unary potential τ_i^{DS}, elegantly models the uncertainty of a channel by defining the corresponding mass functions appropriately. Since Dempster's rule of combination, that is associative, is used, additional information e.g. motion can also be included straight forward.

Mass Functions

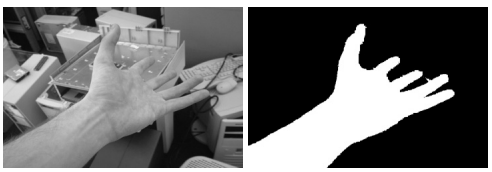
The most important difference between the proposed method and ToFCut is the feature fusion using Dempster's theory of evidence instead of summing up weighted negative log-likelihoods. Therefore the main contribution is the definition of appropriate mass functions, that model inaccuracy and uncertainty in an elegant way. The mass functions modeling color and depth information are defined by:

$$\begin{aligned} m_c(\Psi) &= \frac{(1-\gamma_c)(1-(p_c(I(i)|\mathcal{L}_i=\mathrm{FG})+p_c(I(i)|\mathcal{L}_i=\mathrm{BG})))}{K}, \\ m_c(S) &= (1-m_c(\Psi))\frac{p_c(I(i)|\mathcal{L}_i=S)}{p_c(I(i)|\mathcal{L}_i=\mathrm{FG})+p_c(I(i)|\mathcal{L}_i=\mathrm{BG})} \end{aligned} \tag{5.25}$$

for the color term and

$$\begin{aligned} m_d(\Psi) &= \frac{(1-\gamma_d)(1-(p_d(D(i)|\mathcal{L}_i=\mathrm{FG})+p_d(D(i)|\mathcal{L}_i=\mathrm{BG})))}{K}, \\ m_d(S) &= (1-m_d(\Psi))\frac{p_d(D(i)|\mathcal{L}_i=S)}{p_d(D(i)|\mathcal{L}_i=\mathrm{FG})+p_d(D(i)|\mathcal{L}_i=\mathrm{BG})} \end{aligned} \tag{5.26}$$

for the depth term, where S is either FG or BG. The uncertainties $m_c(\Psi)$ and $m_d(\Psi)$ of the models are defined by summing up the conditional probabilities. This means that the uncertainty of a model is high, if foreground and background likelihoods are small. The normalization factor K is chosen so that $m_c(\Psi)+m_d(\Psi)=1$, which means that the sum of modeled uncertainty is one. This is exactly the definition



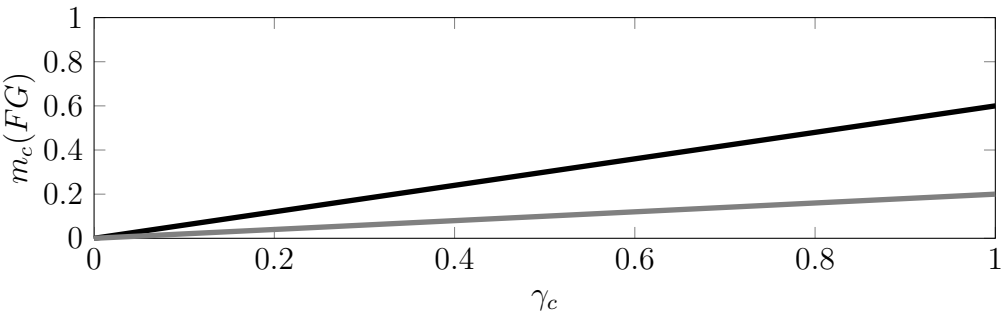


Figure 5.17: Unary color potential of the proposed method with varying confidence. If the color confidence decreases, the mass decreases as well.

used in the proposed variational segmentation scheme, see Chapter 4.1 and Equation (4.13). The parameters γ_c and γ_d are the adaptive weights coming from the histogram analysis. They can be used to further increase or decrease the importance of a feature channel. In contrast to the ToFCut model, the mass of a pixel decreases if the corresponding weight decreases, see Figure 5.17 This weighting is much more intuitive and easier to control.

Color and Depth Likelihoods

Additionally, an improved color model is used, since the one proposed in [WZYZ10] is sensitive to small bins and lacks in precision, leading to suboptimal segmentation results. Similarly to [WZYZ10], two 3D histogram with $H = 8^3$ bins in the RGB color space are used for foreground and background, respectively. For each bin a 3D-Gaussian, with mean μ_k^S, covariance matrix Σ_k^S and weight w_k^S, for $k \in 1 \dots H$ and $S \in \{\text{FG, BG}\}$, is learned. The conditional probability is than given by:

$$p_c(I(i) \mid \mathcal{L}_i = S) = \sum_{r \in \mathcal{N}} w_r^S G(I(i) | \mu_r^S, \Sigma_r^S)\,, \tag{5.27}$$

where $\mathcal{N}$ is the index set of neighboring bins of $I(i)$ in 3D. In contrast to ToFCut, the normalization term is omitted to make the model more robust. Normalizing the given sum with the sum of weights (see [WZYZ10]) makes the color model sensitive to small bins.

To model the depth likelihoods, the conditional probability proposed by Wang et al. [WZYZ10] are used. Therefore, pixels are classified as dark or bright based on the threshold $T_1 = 60$. Foreground and background likelihoods are than each

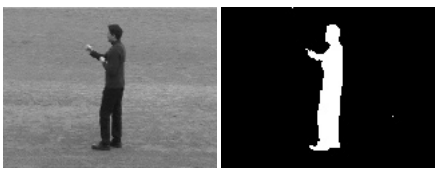


Table 5.2: Comparison between the proposed method DS (highlighted in blue) and ToFCut obtained on four video sequences. The mean percentage error, computed across the whole sequence, is provided for equal weight fusion (EW) and adaptive weight fusion (AW). The results obtained by ToFCut are taken from [WZYZ10]. Furthermore the results are compared to standard graph cut using only color (color) and only depth information (depth).

Seq. ID	WL		MS		MC		CW	
No. Frames	200		400		300		300	
Alg.	ToFCut	DS	ToFCut	DS	ToFCut	DS	ToFCut	DS
Error (EW)	1.37	0.54	0.51	0.23	0.16	0.06	11.68	2.21
Error (AW)	1.35	0.51	0.51	0.23	0.15	0.06	0.38	0.26
Error (color)	9.91		6.88		0.59		1.83	
Error (depth)	1.68		0.92		0.26		4.62	

modeled by two Gaussian distributions using dark and bright pixels, respectively:

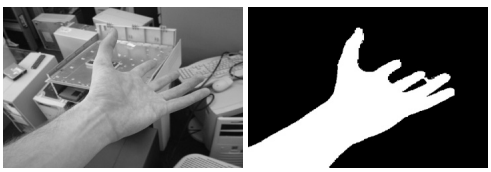
$$p_d(D(i) \mid \mathcal{L}_i = FG) = \begin{cases} G(D(i) \mid \mu^{FG}_{dark}, \sigma^{FG}_{dark}) & I(i) < T_1 \text{ and } D(i) > T_2 \\ G(D(i) \mid \mu^{FG}_{bright}, \sigma^{FG}_{bright}) & I(i) \geq T_1 \text{ and } D(i) > T_2 \\ 0 & D(i) \leq T_2 \end{cases},$$
$$p_d(D(i) \mid \mathcal{L}_i = BG) = \begin{cases} G(D(i) \mid \mu^{BG}_{dark}, \sigma^{BG}_{dark}) & I(i) < T_1 \text{ and } D(i) > T_2 \\ G(D(i) \mid \mu^{BG}_{bright}, \sigma^{BG}_{bright}) & I(i) \geq T_1 \text{ and } D(i) > T_2 \\ 1 & D(i) \leq T_2 \end{cases}. \tag{5.28}$$

Furthermore, a threshold T_2 on the depth map is defined, to exclude pixels from the training of the Gaussians. This threshold forces pixels with a depth value smaller than T_2 to be segmented as background and improves the two models models. Thus, the single parameter T_2 is intuitive, easy to adjust and can be computed automatically by a histogram analysis.

5.3.2 Experimental Results

In this Section, the evaluation of the proposed method is presented. For qualitative and quantitative analysis the ToFCut data set with the corresponding ground truth data[8] is used. The data set consists of four video sequences, each with 200 - 400 frames. The sequences simulate different scenarios, e.g. changing lightning

[8] http://vis.uky.edu/%7Egravity/Research/ToFMatting/ToFMatting.htm

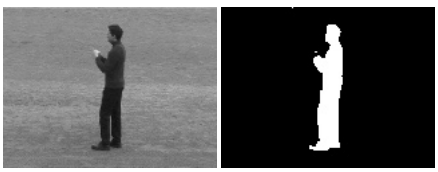


Sequence: WL MS MC CW

Figure 5.18: Example video segmentation results on four sample frames from each of the video sequences.

conditions, changing depth and objects with similar depth. Table 5.2 presents the obtained results and compare them to ToFCut and graph cut by means of mean percentage error of misclassified pixels [BRB+04, KCB+05, WZYZ10], that is the Hamming distance of the segmentation result and the ground truth segmentation, see Equation (2.2). In the experiments an equal weight fusion of color and depth information is used by setting $\gamma_c = \gamma_d = 0.5$. Furthermore, an adaptive weight fusion based on the proposed histogram analysis is evaluated. The quantitative results show that for both systems, equal weight fusion and adaptive weight fusion, the proposed fusion with Dempster's theory outperforms ToFCut. Important to notice is, that the proposed algorithm only needs to adjust two intuitive parameters:

- γ, the weighting of neighboring discontinuities
- T_2, the threshold of the depth map.



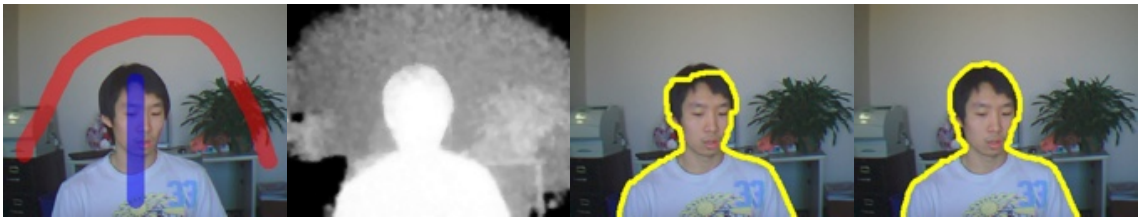


Figure 5.19: Example interactive segmentation result. From left to right: Color image with initialization (FG in blue/BG in red), depth image, segmentation result using ToFCut with equal weights, proposed DS fusion with equal weights.

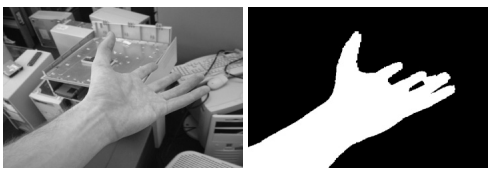
The parameters η_{lum},η_{rgb} and η_d, controlling the adaptive weighting, remain constant in all the experiments, while in [WZYZ10] they had to be adjusted for each sequence manually. Furthermore, the results show that the proposed fusion works well on many sequences without an adaptive weighting. Qualitative results for all sequences using the adaptive weighting are presented in Figure 5.18. They show that the small segmentation error corresponds to a high-quality segmentation. The corresponding adaptive weights are visualized as functions over time in Figure 5.21. It can be seen, that the weights automatically adapt to changes in the scene or environment. E.g. for the sequence CW (Figure 5.21 lower right), the depth reliability decreases when the second person enters the scene. This due to the fact, that this person has a very similar depth when compared to the foreground person. When the second person is partially occluded or leaves the scene, the corresponding depth reliability increases.

Besides video segmentation, interactive image segmentation is a challenging task. Since there exists no benchmark including depth images, the same data set is used. Qualitative results are presented in Figure 5.19. Since color and depth models are learned from rough user strokes, the models are likely to be incomplete. By using the proposed fusion based on Dempster's theory of evidence, this is elegantly modeled by the proposed mass functions and the segmentation clearly outperforms ToFCut. In Figure 5.20, both methods are quantified using a fixed initialization and different reliability weights γ_d. It can be seen, that the proposed method is less sensitive to the weight and the results are more intuitive. Thus, it is much easier to manually adjust the parameter or automatically measure it using histograms.

5.4 Discussion

In Chapter 5 three main contributions are proposed:

- *SlimCut*: An efficient method for graph simplification of maximum a posteriori problems.



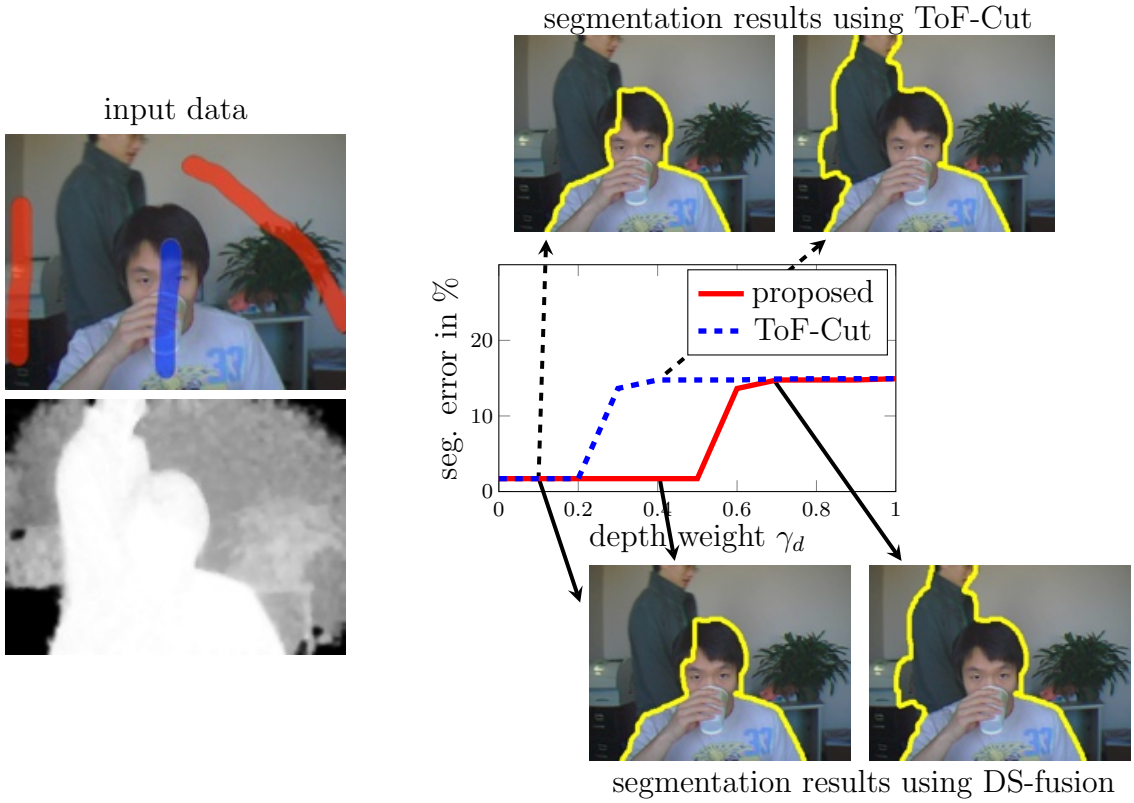
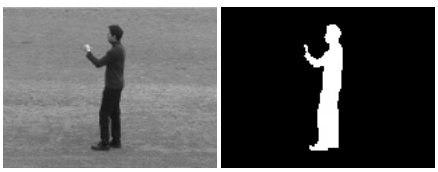


Figure 5.20: Dependence on the weight γ_d. For the given input data (color and depth image and rough user initialization), the dependence of the segmentation result on the reliability weight is visualized. For the given example the ToF-Cut method only produces reasonable results for very small weights ($\gamma_d \in [0,0.2]$). The proposed method, using Dempster's theory for feature fusion, is less sensitive to γ_d and produces reasonable results for weights up to 0.5 ($\gamma_d \in [0,0.5]$).

- An efficient algorithm to group variables using Dempster's theory of evidence.
- A novel video segmentation scheme for RGB-D image sequences.

First, an efficient method for graph simplification of maximum a posteriori problems, the so-called *SlimCut*, is presented. It constructs a *Slim Graph* by merging nodes that are connected by *simple edges*. A proof that the maximum flow of the constructed, much smaller graph remains identical is given. Hence it can be applied to any maximum flow algorithm. The experiments demonstrated that the speedup is between 14 and 70 percent on small-scale problems compared to the BK-algorithm. On high-resolution images with up to 26 MP, the proposed method was up to 877 times faster. It is shown that the proposed method requires much less memory allowing segmentation of images of reasonable sizes even on mobile devices. A further



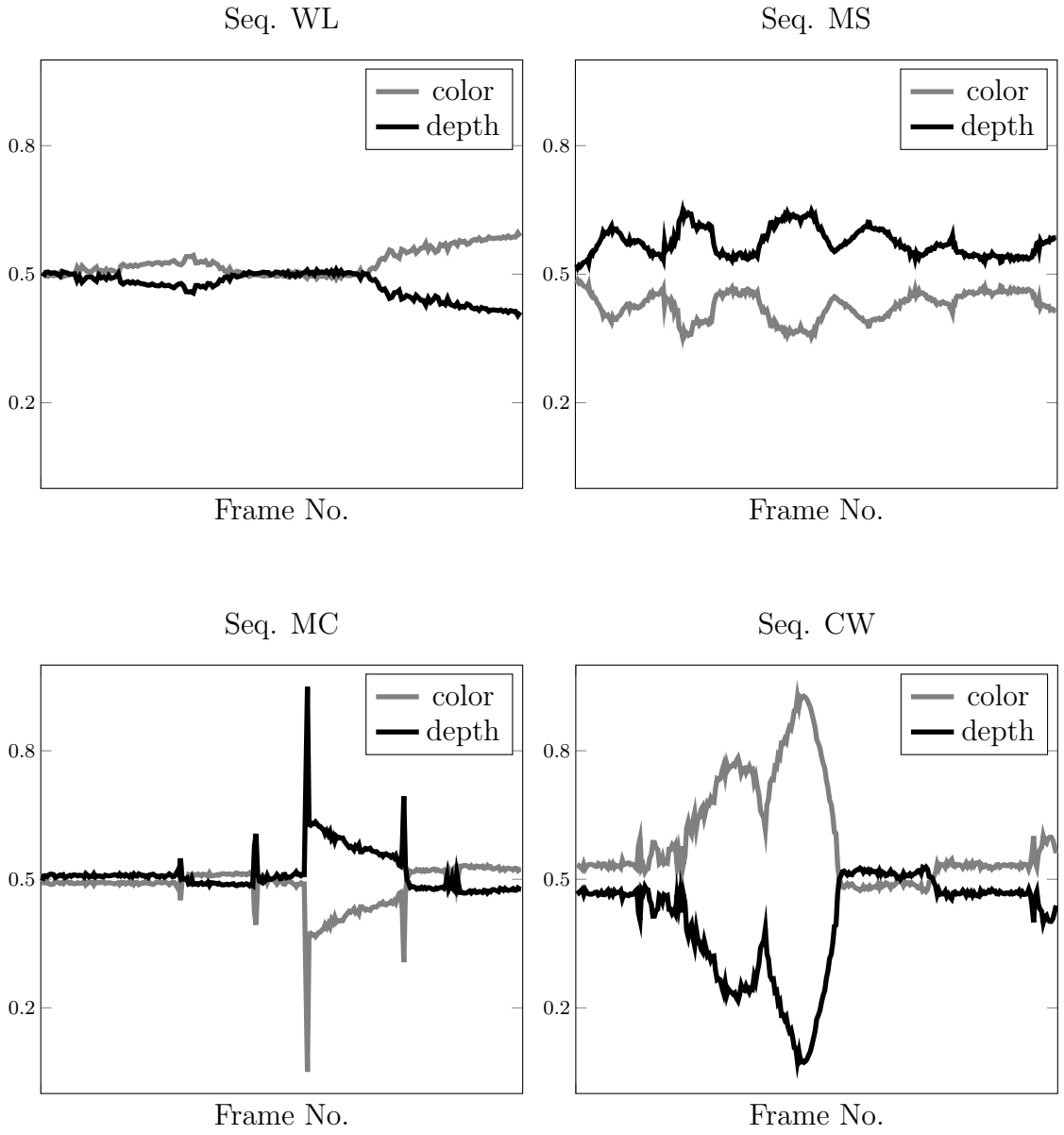
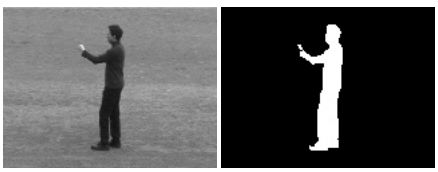


Figure 5.21: Adaptive Weights as functions over time for the four video sequences based on the proposed symmetric Kullback-Leibler divergence. It can be seen that the weights automatically adapt to the environment.

reduction of computation time can be achieved by using parallel hardware architecture. In addition the visualization of the *Slim Graph* can be utilized to guide the user during the segmentation process resulting in less user interaction. In comparison to other works, the proposed simplification does not use special hardware like multiple processors or GPU. Thus the algorithm can be applied to resource-limited systems, as demonstrated on Apple's iPhone 4.

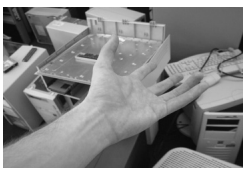
To overcome the dependence on the number of such *simple edges*, a second efficient algorithm for graph simplification is proposed. It uses Dempster's theory of evidence and new constraints for the graph based grouping to group large homogeneous regions to one single variable of the problem. The experiments on segmentation demonstrated that the segmentation error using the proposed method is smaller or comparable to the full MAP solution. Several experiments on large-scale problems with millions of variables have demonstrated that the reduction in runtime is dramatic while the segmentation quality stays comparable. Both proposed algorithms are widely applicable to MAP inference problems in computer vision.

Furthermore, a novel RGB-D video segmentation scheme is proposed. It uses Dempster's theory of evidence to fuse color and depth information. With Dempster's theory of evidence it is possible to define the uncertainty of a feature in an elegant way using prior information or an adaptive weight. The adaptive weight is computed using the symmetric Kullback-Leibler divergence to make it more robust. Additionally, adjusted color and depth models are presented to improve the segmentation results. The quantitative evaluation shows that the proposed method outperforms the reference ToFCut method. In comparison, the proposed method has less parameters that are more intuitive and easy to adjust. Since Dempster's theory of evidence naturally models the uncertainty, the proposed method is also applicable for interactive segmentation. An additional property of the proposed fusion scheme is the naturally given possibility to include further information like motion or user priors.



Chapter 6

Conclusion

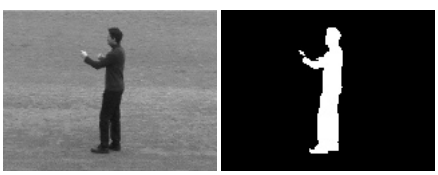


6.1 Summary

This dissertation addresses the problem of interactive binary segmentation using energy minimizing frameworks, namely level sets and graph cuts. In the last few decades, interactive binary segmentation frameworks have been widely studied and used in Computer Vision. This is due to the fact, that the problem of binary image segmentation is one of the fundamental problems in low level Computer Vision. This thesis addresses two problems raised in the last years. First, new sensors and features have become popular and the need for a fusion framework became necessary. Second, due to the scale of current Computer Vision problems, computationally more efficient algorithms are essential.

The first part of this dissertation targets the problem of an elegant and intuitive framework for feature fusion. Dempster's theory of evidence, as a generalization of Bayes' theory, is used to fuse information arising from different sensors. It is included in the energy functions (variational or discrete ones) and extends current state-of-the-art segmentation frameworks. The results show, that the use of Dempster's theory of evidence, instead of Bayes' theory, for feature fusion improves the segmentation in terms of segmentation quality. Utilizing Dempster's theory of evidence, it is possible to extend the well known level set approach by means of user interaction. In a small user study it was shown that the proposed interactive framework is able to compete with other interactive segmentation frameworks. By using a novel confidence measure, estimating the accuracy of a feature, and Dempster's theory of evidence for feature fusion, a RGB-D video segmentation scheme based on graph cuts is introduced that outperforms other methods.

The problem of computationally more efficient algorithms solving large scale Computer Vision problems is addressed in the second part of this dissertation. Two algorithms, reducing the size/scale of the problem, are presented. The first one, called *SlimCuts*, reduces the underlying graph of a discrete energy minimization problem while maintaining the maximum flow property. A proof is given, that building the *Slim Graph* does not change the maximum flow. Thus the graph and the problem is reduced without changing the solution. Obviously, this kind of graph reduction is limited by the graph itself. Therefore, a second algorithm for grouping similar variables of the energy function is proposed. The terms of the energy functions, interpreted as information of similarity, are fused within the framework of Dempster's theory of evidence to define the joint similarity. Using this similarity and an extended graph based grouping algorithm, the problem size is reduced. The experiments have shown, that the proposed algorithms drastically reduce the graph. In case of the variable grouping, the experiments also show that the change in the segmentation result is negligible.



6.2 Contributions

Listed below are the major contributions discussed in this dissertation.

Integration of Dempster's theory of evidence: Appropriate mass functions, modeling the likelihoods within Dempster's theory of evidence, are proposed for the problem of energy minimizing image segmentation. A novel continuous energy function is proposed that integrates Dempster's theory of evidence and extends the classical level set framework. Thus, the features arising from different sensors are fused more intuitive and more elegantly. Several experiments demonstrate the properties of the proposed feature fusion, that is able to directly model inaccuracy and uncertainty and is therefore able to resolve conflicts.

Interactive level sets: The proposed continuous energy function including Dempster's theory of evidence for feature fusion is furthermore extended by means of user interaction. Since user information can be sparse, modeling inaccuracy and uncertainty is essential. The proposed energy function models these information using a user-defined image model, that can be sparse, and a user-defined shape prior. Thus, the user interactions have local influence (caused by the shape prior) and global influence (caused by the image model). In comparison to other methods, the proposed interactive level sets needs significantly less user interactions for adequate segmentation results on real images.

SlimCuts: *SlimCuts* are proposed to simplify the underlying graph of a maximum a posteriori problem. *Simple edges* are characterized and the graph is reduced by contracting nodes connected by a *simple edge*. It is proven that those edges do not contribute to the minimum cut. By contracting this edges, the minimum cut and thus the segmentation result does not change. Hence it can be applied to reduce the problem size of any maximum a posteriori problem that is solved with maximum flow algorithms. The experiments on image segmentation demonstrate a drastic reduction in runtime for large-scale problems. In contrast to other works, the proposed method does not use any special hardware and can therefore be used on resource-limited systems like mobile phones. Furthermore, the workflow of an interactive segmentation framework can be optimized by visualizing the *Slim Graph*.

Variable grouping: Grouping variables of the MAP-problem based on their similarity is another promising direction to reduce the problem size. The similarity in the proposed method is defined by the terms of the energy function and this information is fused using Dempster's theory of evidence. This algorithm allows to group large homogeneous regions to single variables of the problem. Thus, the problem size is reduced drastically while the segmentation results stays comparable in terms of segmentation quality.

Multi-sensor fusion for video segmentation: A novel video segmentation framework based on graph cuts for RGB-D image pairs is proposed. Dempster's theory of

evidence is integrated in the discrete energy function to fuse color and depth information and to model uncertainty. The uncertainty of a feature is measured using an adaptive weight or defined by some prior information on the sequence. The adaptive weight is computed by an extended function using the symmetric Kullback-Leibler divergence. Experiments on benchmark sequences show that the proposed video segmentation framework outperforms others in terms of quality with less parameters to control. Due to the intuitive definition of mass functions used by Dempster's theory of evidence for feature fusion, the proposed method is also suited for the problem of interactive image / video segmentation.

The contributions have been published at ACCV, EMMCVPR, CIARP, WACV, SCIA, DAGM and ISVC.

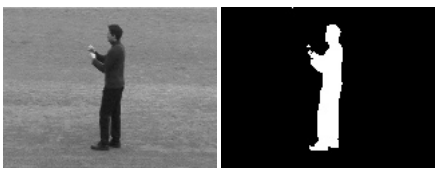
6.3 Possible Directions for Future Work

This thesis is concluded with some ideas and directions for future research. It was shown in this thesis how Dempster's theory of evidence can be integrated into energy minimizing frameworks for binary image segmentation. There is clear need for extending these methods into multi-label segmentation systems. Some progress in this field has been made by the work of Arndt et al. [ASR13]. Large regions of an image are pre-segmented utilizing Gaussian mixture models and spherical coordinates. The classical GrowCut algorithm [VK05] is extended by a novel weight function and the pre-segmented image is used as initialization.

The impact of using Dempster's theory of evidence for feature fusion in multi-label frameworks can be huge. Since uncertainty, inaccuracy and conflicts are modeled directly, this fusion should outperform the classical Bayesian feature fusion. But an appropriate definition of mass functions for multi-label problems is still an open problem. The proposed mass functions can be used, but the whole power of Dempster's theory is not exploited since the conflict and the mass of joint events for more than two classes are not modeled until now.

Dempster's theory of evidence is a generalization of the Bayesian theory, thus the integration into other energy functions for problems like object reconstruction, image restoration or disparity estimation can be advantageous. Besides Dempster's rule of combination, there exist many other fusion rules within the framework of Dempster's theory of evidence. For the problem of image segmentation, Dempster's rule of combination outperformes most of those alternatives. However, for other problems other fusion rules might be advantageous.

Another promising direction of research is the combination of graph reduction and variable grouping algorithms to further decrease the problem size of large-scale problems. E.g. combining the results of the proposed *SlimCut* algorithm with the



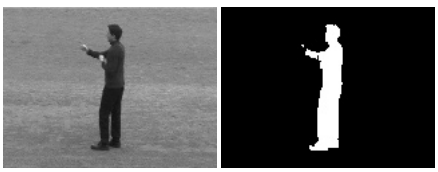
proposed variable grouping and the pre-segmentation proposed in [ASR13] can possibly lead to much smaller problem sizes.

Something that has not received much attention till now, is the use of additional features for the pairwise term $\tau_{i,j}$, penalizing strong gradients. In the experiments it was observed that, e.g. depth disparities are a strong indicator for object boundaries. Even though this information is spatially not that accurate, it should help to find the correct boundary of an object. Because of its spatial inaccuracy, the information can not be utilized similar to the color penalty used so far. In fact, a large gradient in the depth image can be utilized as an indicator for all the pixels in some defined neighborhood. Thus, aggregating the gradient information in a circular or rectangular neighborhood and combine this information with the color gradient can improve the model. For the combination of both information Dempster's theory of evidence can be used.

The thesis is concluded by the following observation: Many problems in Computer Vision are solved using multiple sensors or features. Dempster's theory of evidence provides a general framework and generalizes Baye's framework with the possibility to directly model inaccuracy and uncertainty. It has been shown that integrating or extending current frameworks by means of Dempster's theory of evidence offers the possibility to improve the results and allows for a more intuitive understanding. Thus, this theory is of high interest for other applications in the fields of Computer Vision or Machine Learning as well.

Appendix

Appendix



A.1 Building the Euler Lagrange Equation

Given the classical Chan-Vese energy function for image segmentation [CV01]:

$$\begin{aligned} E(\mu_{FG},\mu_{BG},\varphi) = \lambda_1 \int_\Omega |I(x) - \mu_{FG}|^2 H(\varphi)\, dx \\ + \lambda_2 \int_\Omega |I(x) - \mu_{BG}|^2 (1 - H(\varphi))\, dx \\ + \nu_1 \cdot \int_\Omega |\nabla H(\varphi)|\, dx + \nu_2 \cdot \int_\Omega H(\varphi)\, dx\,. \end{aligned} \tag{A.1}$$

Minimizing the energy with respect to φ while keeping μ_{FG} and μ_{BG} fixed can be performed by solving the the corresponding Euler-Lagrange equation, that is a necessary condition for a minimum. The energy function has the form:

$$E(\varphi) = \int_\Omega \mathcal{L}(\varphi,\nabla\varphi)\, dx \tag{A.2}$$

The corresponding Euler-Lagrange equation is given by:

$$\frac{dE}{d\varphi} = \frac{\partial \mathcal{L}}{\partial \varphi} - \frac{\partial}{\partial x_1}\frac{\partial \mathcal{L}}{\partial \varphi_{x_1}} - \frac{\partial}{\partial x_2}\frac{\partial \mathcal{L}}{\partial \varphi_{x_2}} = 0\,. \tag{A.3}$$

Inserting the energy function from Equation (A.1) or building the derivatives of the corresponding terms of the energy function, respectively, leads to:

$$\frac{\partial \mathcal{L}}{\partial \varphi} = \lambda_1 |I(x) - \mu_{FG}|^2 \delta(\varphi) - \lambda_2 |I(x) - \mu_{BG}|^2 \delta(\varphi) + \nu_1 \delta'(\varphi)|\nabla\varphi| + \nu_2 \delta(\varphi)\,, \tag{A.4}$$

plus

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \varphi_{x_i}} &= \nu_1 \delta(\varphi) \frac{\partial \varphi}{\partial x_i}\frac{1}{|\nabla\varphi|} \quad \text{with } i \in \{1,2\} \\ \Rightarrow \frac{\partial}{\partial x_i}\frac{\partial \mathcal{L}}{\partial \varphi_{x_i}} &= \nu_1 \frac{\partial \varphi}{\partial x_i}\frac{1}{|\nabla\varphi|}\left(\frac{\partial}{\partial x_i}\delta(\varphi)\right) + \nu_1 \delta(\varphi)\left(\frac{\partial}{\partial x_i}\left(\frac{\partial \varphi}{\partial x_i}\frac{1}{|\nabla\varphi|}\right)\right) \quad \text{with } i \in \{1,2\}\,. \end{aligned} \tag{A.5}$$

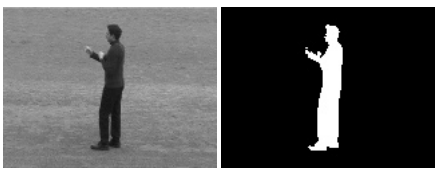
Overall it leads to:

$$
\begin{aligned}
\frac{dE}{d\varphi} &= \lambda_1 |I(x) - \mu_{FG}|^2 \delta(\varphi) - \lambda_2 |I(x) - \mu_{BG}|^2 \delta(\varphi) + \nu_1 \delta'(\varphi) |\nabla\varphi| + \nu_2 \delta(\varphi) \\
&\quad - \sum_{i=1}^{2} \left[\nu_1 \frac{\partial\varphi}{\partial x_i} \frac{1}{|\nabla\varphi|} \left(\frac{\partial}{\partial x_i} \delta(\varphi) \right) + \nu_1 \delta(\varphi) \left(\frac{\partial}{\partial x_i} \left(\frac{\partial\varphi}{\partial x_i} \frac{1}{|\nabla\varphi|} \right) \right) \right] \\
&= \lambda_1 |I(x) - \mu_{FG}|^2 \delta(\varphi) - \lambda_2 |I(x) - \mu_{BG}|^2 \delta(\varphi) + \nu_1 \delta'(\varphi) |\nabla\varphi| + \nu_2 \delta(\varphi) \\
&\quad - \nu_1 \sum_{i=1}^{2} \left[\delta'(\varphi) \left(\frac{\partial\varphi}{\partial x_i} \right)^2 \frac{1}{|\nabla\varphi|} + \delta(\varphi) \left(\frac{\partial}{\partial x_i} \left(\frac{\frac{\partial\varphi}{\partial x_i}}{|\nabla\varphi|} \right) \right) \right] \\
&= \lambda_1 |I(x) - \mu_{FG}|^2 \delta(\varphi) - \lambda_2 |I(x) - \mu_{BG}|^2 \delta(\varphi) + \nu_1 \delta'(\varphi) |\nabla\varphi| + \nu_2 \delta(\varphi) \\
&\quad - \nu_1 \delta'(\varphi) \frac{\sum_{i=1}^{2} \left(\frac{\partial\varphi}{\partial x_i} \right)^2}{|\nabla\varphi|} - \nu_1 \delta(\varphi) \sum_{i=1}^{2} \left[\frac{\partial}{\partial x_i} \left(\frac{\frac{\partial\varphi}{\partial x_i}}{|\nabla\varphi|} \right) \right] \\
&= \lambda_1 |I(x) - \mu_{FG}|^2 \delta(\varphi) - \lambda_2 |I(x) - \mu_{BG}|^2 \delta(\varphi) + \nu_1 \delta'(\varphi) |\nabla\varphi| + \nu_2 \delta(\varphi) \\
&\quad - \nu_1 \delta'(\varphi) \frac{|\nabla\varphi|^2}{|\nabla\varphi|} - \nu_1 \delta(\varphi) \cdot \operatorname{div} \left(\frac{\nabla\varphi}{|\nabla\varphi|} \right) \\
&= \delta(\varphi) \left[\lambda_1 |I(x) - \mu_{FG}|^2 - \lambda_2 |I(x) - \mu_{BG}|^2 + \nu_2 - \nu_1 \cdot \operatorname{div} \left(\frac{\nabla\varphi}{|\nabla\varphi|} \right) \right]
\end{aligned}
\tag{A.6}
$$

This partial differential equation can be solved using standard numerical techniques [SR09].

A.2 Application - N-View Human Silhouette Segmentation in Cluttered, Partially Changing Environments

This application will show how to fuse the proposed variational framework (see Chapter 4) with a different segmentation scheme for a fully automatic segmentation of multi-view human silhouettes. The segmentation of foreground silhouettes of humans in cameras is a fundamental step in many computer vision and pattern recognition tasks. An approach is presented which, based on color distributions, automatically estimates the foreground by integrating data driven 3D scene knowledge from multiple static views. These estimates are integrated into the variational level set framework based on Dempster's theory of evidence to provide a final segmentation. The advantage of this approach is that ambiguities on color information, used by the level set approach, can be resolved in many cases utilizing 3D scene knowledge and 2D boundary constraints. The application is directly based on a joined publication with Feldmann et al. [FSRW10].



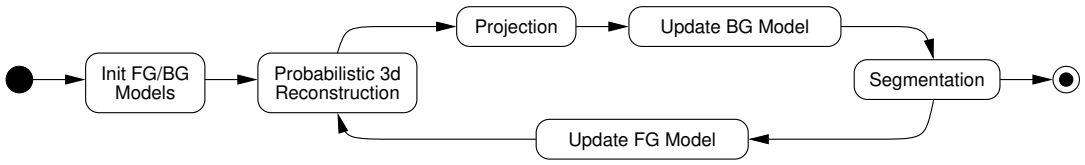


Figure A.1: Segmentation loop utilizing probabilistic 3D fusion as data driven feedback mechanism to enhance the segmentation by automatically adapted color distributions.

Segmentation by Probabilistic 3D Fusion

The segmentation via probabilistic 3D fusion proposed by Feldmann et al. [FDW09] is based on two ideas: First, a probabilistic 2D segmentation of fore- and background in all camera images of a static, calibrated multi camera setup is performed based on color distribution models. To make this segmentation more robust and adaptive, the second part integrates 3D scene information reconstructed from all cameras. The 3D information is used as a feedback mechanism to the segmentation task. Hereby the color distributions are adapted automatically to achieve better segmentation results. The basic assumption is that observed objects are surrounded by multiple cameras to obtain complete 3D reconstructions of the foreground.

The steps of the approach are depicted in Fig. A.1. First, coarse fore- and background models are generated. They are used with the current camera images to create a probabilistic 3D voxel reconstruction of the scene. Probabilistic in this context means that each reconstructed voxel has a specific occupation probability derived from the probabilities of the corresponding pixels in all views to be foreground. The 3D reconstruction is projected into the camera images, thresholded and in this way provides a masked area of foreground in the images. Image areas which are not covered by this mask are used to update the background model. By utilizing this updated model a segmentation is performed to precisely determine the foreground silhouettes. The silhouettes are used to update the foreground model accurately in a succeeding step. The fore- and background models are then used to create a probabilistic 3D reconstruction of the foreground by using the next camera frame and the loop restarts.

Fore- and Background Model

To model fore- and background, the random variable $\mathcal{L}_x \in \{0,1\}$ decides whether a pixel x at a given time t is fore- or background ($\mathcal{L}_x = 1$ respectively $\mathcal{L}_x = 0$). Based on a given feature vector I_x the color distribution $p(I_x \mid \mathcal{L}_x = 1)$ models the foreground and is used to infer the conditional probability $P(\mathcal{L}_x = 1 \mid I_x)$. The foreground model is generated based on the foreground segment for each frame

separately and consists of two parts:

$$p(I_x \mid \mathcal{L}_x = 1) = (1 - P_{\mathrm{NF}}) \sum_{k=1}^{K_{\mathrm{fg}}} \omega^k \eta(I_x, \mu^k, \Sigma^k) + P_{\mathrm{NF}} \mathcal{U}(I_x) \ . \tag{A.7}$$

The first part models known foreground in terms of a Gaussian Mixture Model (GMM) with the density function $\eta(I_x, \mu, \Sigma)$ where μ^k and Σ^k are mean and variance of the k^{th} of K_{fg} components of the mixture and ω^k is the component's weight. B models a uniform color distribution which is necessary to integrate suddenly arising new foreground. Both parts are coupled by the probability $P_{\mathrm{NF}} = \frac{1}{2}$ of new foreground. The model is generated continuously by utilizing k-means clustering of the colors of the foreground silhouette during consecutive frames. The background model consists of two parts as well:

$$p(I_{x,t} \mid \mathcal{L}_{x,t} = 0) = (1 - P_{\mathrm{S}}) \sum_{k=1}^{K_{\mathrm{bg}}} \omega_t^k \eta(I_{x,t}, \mu_t^k, \Sigma_t^k) + P_{\mathrm{S}} \sum_{k=1}^{K_{\mathrm{bg}}} \omega_t^k p(I_{x,t} \mid \mathcal{S}_{x,t}^k = 1) \ . \tag{A.8}$$

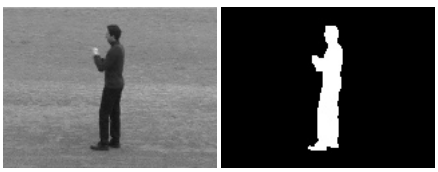
The first part models the color distribution of the background similar to the model in Equation (A.7) with K_{bg} components. In contrast to Equation (A.7), the model is updated over the whole observation time t. The second part models the occurrence of shadows and highlights. Both parts are again coupled with an additionally probability of shadows $P_{\mathrm{S}} = \frac{1}{2}$. The shadow and highlight model is modeled in analogy to the background color model, i.e. the weights are reused. To determine shaded areas or areas of highlights, the colors are examined in the YUV color space. A luminance ratio λ is calculated in the Y channel: $\lambda = \frac{Y_t}{Y_B} = \frac{I_{x,t}^1}{\mu_t^{k,1}}$. Two thresholds are introduced to detect shadows, if $\tau_S < 1$, and highlights, if $\tau_H > 1$. The resulting shadow model is:

$$p(I_{x,t} \mid \mathcal{S}_{x,t}^k = 1) = \begin{cases} \frac{1}{(\tau_H - \tau_S)\mu_t^{k,1}} \prod\limits_{d=2,3} \eta(I_{x,t}^d, \mu_t'^{k,d}, \Sigma_t^{k,d}) & \text{if } \tau_S \leq \lambda_t^k \leq \tau_H \\ 0 & \text{else} \end{cases} \ . \tag{A.9}$$

The scale factor $\frac{1}{(\tau_H - \tau_S)\mu_t^{k,1}}$ is needed to achieve the density's integration to result in 1. The background model in Equation (A.8) is updated continuously by integration of all previous frames over time by utilizing an online Expectation Maximization (EM) approach as presented in [FDW09].

Probabilistic 3D Fusion

To update fore- and background models, a method is needed to reliably identify foreground regions in the images. In case of multi camera setups it is feasible to



exploit the strong prior of geometric coherence of the scene observed from multiple views by using the approach of a Bayesian probabilistic 3D reconstruction [FB05]. The volume seen by the cameras is discretized into voxels $\mathcal{V} \in \{0,1\}$. For each voxel the probability of being foreground is derived from the foreground probabilities of the corresponding pixels in all cameras according to the model definition in [FDW09]. Four a-priori probabilities are introduced into the reconstruction model. First, the probability of voxel occupation: $P(\mathcal{V}) = \frac{1}{2}$. Additionally, three error probabilities P_{DF}, P_{FA} and P_{O}. P_{DF} means a *detection failure*, i.e. a voxel should be occupied but is not due to e.g. camera noise. P_{FA} means a *false alarm*, i.e. a voxel should not be occupied but erroneously is, e.g. due to shadows. Finally, P_{O} means an *obstruction*, i.e. a voxel should not be occupied but is on the same line of sight as another voxel which is occupied and, hence, classified incorrectly. The conditional probability of foreground of an unoccupied voxel is, thus, $\mathcal{V}$: $P(\mathcal{L}_n = 1 \mid \mathcal{V} = 0) = P_{\mathrm{O}}(1 - P_{\mathrm{DF}}) + (1 - P_{\mathrm{O}})P_{\mathrm{FA}}$. The conditional probability of background of an unoccupied voxel is $\mathcal{V}$: $P(\mathcal{L}_n = 0 \mid \mathcal{V} = 0) = 1 - [P_{\mathrm{O}}(1 - P_{\mathrm{DF}}) + (1 - P_{\mathrm{O}})P_{\mathrm{FA}}]$. Values of 5% for P_{DF}, P_{FA} and P_{O} provide reasonable results. The joint probability distribution defined in [FDW09] is used, and marginalize over the unknown variables $\mathcal{L}_n$ by observing the features (colors) $I_1, \ldots, I_N$ at the corresponding pixels in the images of the cameras $1, \ldots, N$ by:

$$P(\mathcal{V} = 1 \mid I_1, \ldots, I_N) = \frac{\prod_{n=1}^{N} \sum_{f \in \{0,1\}} P(\mathcal{L}_n = f \mid \mathcal{V} = 1) p(I_n \mid \mathcal{L}_n = f)}{\sum_{v \in \{0,1\}} \prod_{n=1}^{N} \sum_{l \in \{0,1\}} P(\mathcal{L}_n = f \mid \mathcal{V} = v) p(I_n \mid \mathcal{L}_n = l)} \,. \quad \text{(A.10)}$$

The resulting probabilistic 3D reconstruction is back-projected into the camera images and then used to identify fore- and background segments.

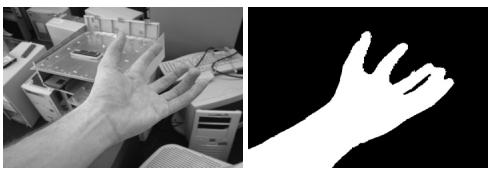
Probabilistic Foreground Detection

By using the probability densities $p(I_x \mid \mathcal{L}_x = 1)$ and $p(I_x \mid \mathcal{L}_x = 0)$ the conditional probability $P(\mathcal{L}_x = 1 \mid I_x)$ that a pixel belongs to the foreground based on an observed color I_x can be calculated using Bayes' rule

$$P(\mathcal{L}_x = 1 \mid I_x) = \frac{P(\mathcal{L}_x = 1) p(I_x \mid \mathcal{L}_x = 1)}{p(I_x)} = \frac{P(\mathcal{L}_x = 1) p(I_x \mid \mathcal{L}_x = 1)}{\sum_{l \in \{0,1\}} P(\mathcal{L}_x = l) p(I_x \mid \mathcal{L}_x = f)} \quad \text{(A.11)}$$

which under assumption of no a-priori knowledge about the unconditional probabilities $P(\mathcal{L}_x = f)$ and a resulting uniform distribution cancels out to:

$$P(\mathcal{L}_x = 1 \mid I_x) = \frac{p(I_x \mid \mathcal{L}_x = 1)}{\sum_{l \in \{0,1\}} p(I_x \mid \mathcal{L}_x = l)} \,. \quad \text{(A.12)}$$



Integrating Probabilistic 3D Fusion into Variational Segmentation

Given the probabilities $P(\mathcal{L}_x = 1 \mid I_x)$ for each feature vector I_x arising from the probabilistic foreground detection, see Equation (A.12), the following mass function is defined:

$$\begin{aligned} m_{\text{fg}}(\emptyset) &= 0\,, & m_{\text{fg}}(\Omega) &= 1 - \nu_2\,, \\ m_{\text{fg}}(FG) &= \nu_2 \cdot P(\mathcal{L}_x = 1 \mid I_x), & m_{\text{fg}}(BG) &= \nu_2 \cdot (1 - P(\mathcal{L}_x = 1 \mid I_x))\,, \end{aligned} \tag{A.13}$$

with a weighting parameter $\nu_2 \in [0,1]$. This parameter can be interpreted as the belief in the probabilistic foreground detection. With a parameter $\nu_2 < 1$ inaccuracy of the foreground detection is integrated. As a consequence, the evolving boundary is directly driven by the intensity information of the image and the result of the probabilistic 3D fusion.

The mass function m_{fg} is now integrated into the variational approach for image segmentation , see Equation (4.8), using Dempster's rule of combination:

$$m_{\text{new}} = m \otimes m_{\text{fg}} = m_1 \otimes m_2 \otimes \ldots \otimes m_k \otimes m_{\text{fg}}\,. \tag{A.14}$$

The energy functional for segmentation fusing image features and probabilistic foreground detection can be written as:

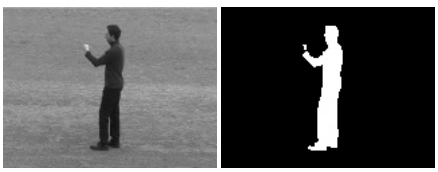
$$\begin{aligned} E(\varphi) = & \underbrace{-\int_\Omega H(\varphi) \log m_{\text{new}}(\Omega_1)\, d\Omega - \int_\Omega (1 - H(\varphi)) \log m_{\text{new}}(\Omega_2)\, d\Omega}_{\text{fusion of image features and probabilistic foreground detection}} \\ & + \nu_1 \int_\Omega |\nabla H(\varphi)|\, d\Omega\,. \end{aligned} \tag{A.15}$$

Compared to the Bayesian approach the proposed framework is able to correct wrong classifications coming from the probabilistic foreground detection and vice versa, because channels with a strong support are favored.

Experimental Results

A qualitative and a quantitative analysis of the algorithm is presented based on the images of the *Dancer Sequence* in [FDW09] and recordings of gymnasts with seven Prosilica GE680C cameras in a circular setup.

In a qualitative analysis the results of the approach of [FDW09] are compared to the results of a variational segmentation, with GrabCut [RKB04] and the results of the proposed combined approach. The probabilistic segmentation of [FDW09] is initialized with a-priori recorded background images. These images varied in lighting and details which was automatically compensated by the presented approach. In case of the variational segmentation and GrabCut, the result of the probabilistic



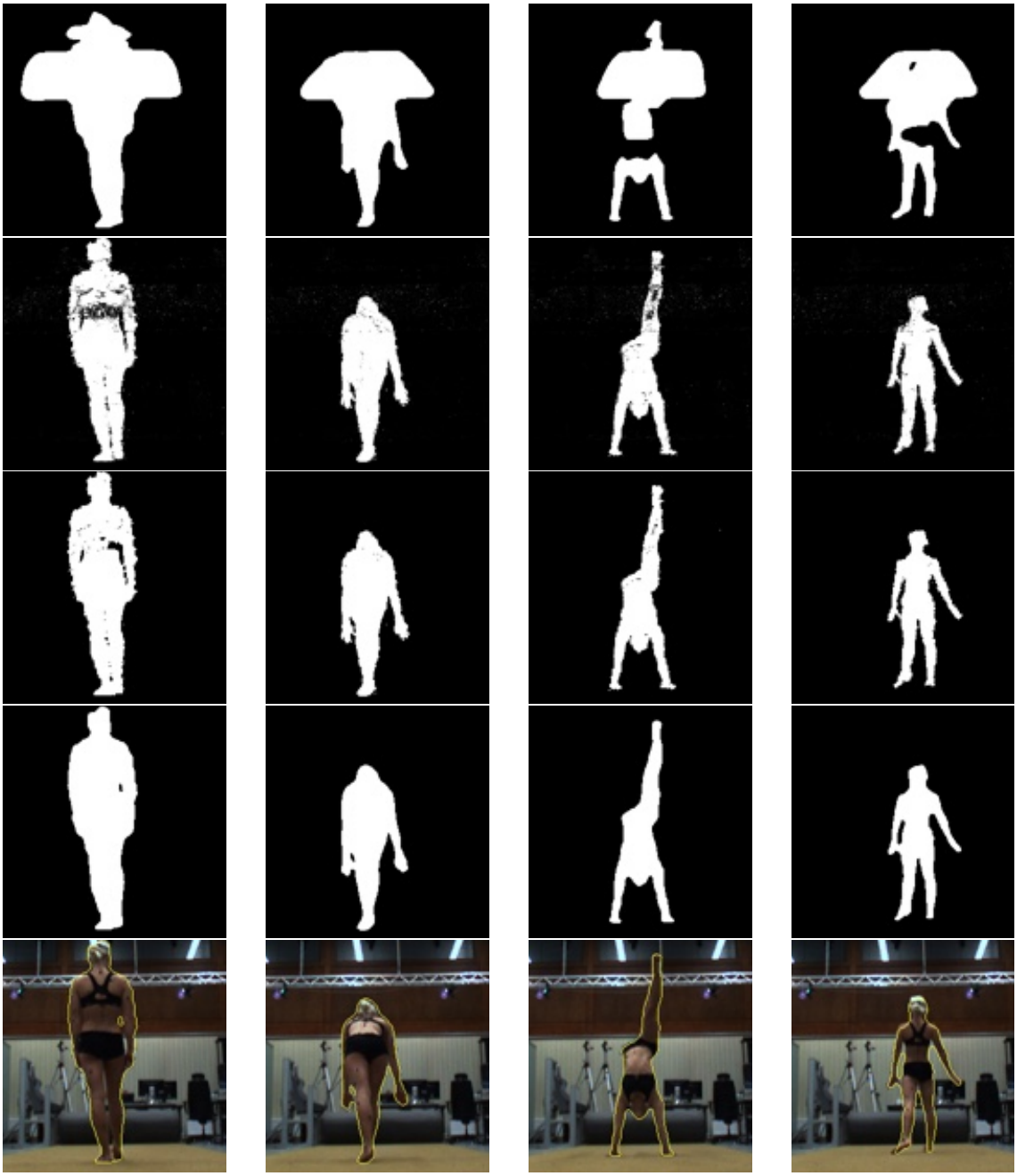


Figure A.2: Frames 100, 300, 500 and 700 of camera 7. First row: Variational segmentation only. Second row: Segmentation by probabilistic fusion only without post processing. Third row: Combined approach with GrabCut segmentation. Fourth row: Proposed combined approach with variational segmentation. Fifth row: Input image and detected contour of combined approach. All single approaches have difficulties in areas with nearly identical color distributions of fore-/background. Only the proposed combined approach is able to cope with these kinds of ambiguities.

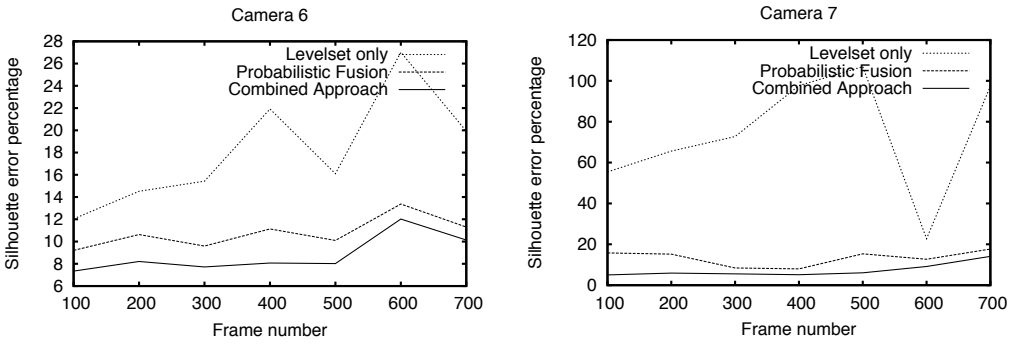


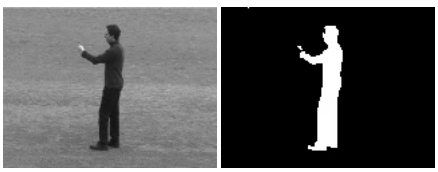
Figure A.3: Silhouette error percentage in steps of 100 frames in cameras 6 and 7 of the gymnast sequence. The proposed approach generates the best results with the fewest errors compared to the variational approach and the approach of [FDW09].

3D fusion is used to initialize the segmentation. In the combined approach the information from the probabilistic 3D fusion is used as the initial boundary and integrated into the variational segmentation framework as proposed in Equation (A.15).

In Figure A.2 exemplary results are presented of all four approaches performed on a difficult scene with very similar color distributions of fore- and background. It is clearly observable that neither the variational approach nor the segmentation by probabilistic fusion are able to fully cope with that kinds of ambiguities. The variational approach integrates large parts of the wooden background into the foreground silhouette while the approach of [FDW09] leads to very low probabilities of foreground in the ambiguous areas. Solely, the proposed approach leads to satisfying results in such difficult scenarios. As an alternative to variational segmentation, the results of the probabilistic segmentation could also be used as initialization for GrabCut. But only the combination of initialization by probabilistic segmentation and fusing this information utilizing the Dempster-Shafer approach can close erroneous holes and, thus, recover from false classifications in a meaningful manner (see Figure A.5).

Due to the convincing results of Figure A.2 a quantitative analysis of the three approaches is performed where the error compared to hand labeled data is measured. Exemplary results of the cameras 6 and 7 are presented in Figure A.3. Camera 6 has been chosen because this view contains background motion and this demonstrates that the adaptivity of [FDW09] is not compromised by the presented approach. The results of Camera 7 are selected to link the qualitative results in Figure A.2 with quantitative results to clarify the benefits of the presented approach. In all cases the proposed approach provides better results over the full sequence.

Finally, a qualitative analysis of the proposed approach on the dancer from [FDW09] is performed. The qualitative results show, that again, the proposed approach gains



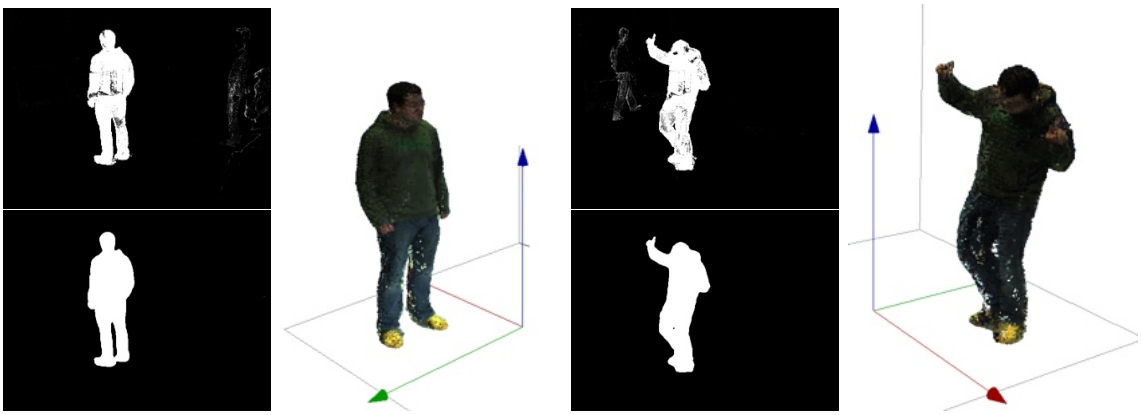


Figure A.4: Column 1, top: Probabilistic segmentation of the second frame (after first model update) of camera 6 of the dancer sequence; Bottom: Segmentation of proposed approach. Column 2: Resulting 3d reconstruction of proposed approach. Column 3, top: Probabilistic segmentation of frame 645; Bottom: Segmentation of the proposed approach. Column 4: Resulting 3d reconstruction.

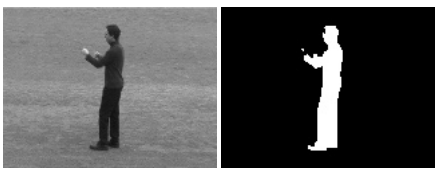
Figure A.5: Random images of situations where GrabCut (10 iterations) fails while the proposed approach gains meaningful improvements. Left to right: Alternating the results of GrabCut and the Dempster-Shafer level set approach. GrabCut misses to close holes in the silhouettes in difficult situations due to similar color distributions.

better segmentation results (see Figure A.4) than the probabilistic segmentation. Additionally the experiments on the dancer sequence demonstrate that the proposed approach is applicable in these kinds of difficult scenarios with occluding noise and, thus, unites the benefits of robust segmentation and robust dense 3D reconstruction results.

Appendix

List of Publications

B



B.1 Refereed Publications

Multi-Sensor Fusion using Dempster's Theory of Evidence for Video Segmentation

Björn Scheuermann, Sotirios Gkoutelitsas and Bodo Rosenhahn
In: Proc. of the 18th Iberoamerican Congress on Pattern Recognition (CIARP), November 2013, Havanna, Cuba (Accepted for oral presentation)

Cleaning Up Multiple Detections Caused by Sliding Window Based Object Detectors

Arne Ehlers, Björn Scheuermann, Florian Baumann and Bodo Rosenhahn
In: Proc. of the 18th Iberoamerican Congress on Pattern Recognition (CIARP), November 2013, Havanna, Cuba

Foreground Segmentation from Occlusions using Structure and Motion Recovery

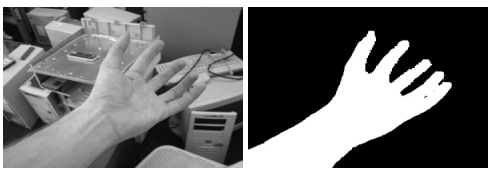
Kai Cordes, Björn Scheuermann, Bodo Rosenhahn and Jörn Ostermann
In: Computer Vision, Imaging and Computer Graphics. Theory and Applications, Communications in Computer and Information Science (CCIS), May 2013

"Region Cut" - Interactive Multi-Label Segmentation Utilizing Cellular Automaton

Oliver Jakob Arndt, Björn Scheuermann and Bodo Rosenhahn
In: Proc. of the IEEE Workshop on Applications of Computer Vision (WACV), January 2013, Clearwater Beach, Florida, USA

Efficient Pixel-Grouping based on Dempster's Theory of Evidence for Image Segmentation

Björn Scheuermann, Markus Schlosser and Bodo Rosenhahn
In: Proc. of the 11th Asian Conference on Computer Vision (ACCV), November 2012, Daejeon, Korea



Learning Object Appearance from Occlusions using Structure and Motion Recovery

Kai Cordes, Björn Scheuermann, Bodo Rosenhahn and Jörn Ostermann
In: Proc. of the 11th Asian Conference on Computer Vision (ACCV), November 2012, Daejeon, Korea

Occlusion Handling for the Integration of Virtual Objects into Video

Kai Cordes, Björn Scheuermann, Bodo Rosenhahn and Jörn Ostermann
In: Proc. of the Seventhth International Conference on Computer Vision Theory and Applications (VISAPP), February 2012, Rome, Italy

SlimCuts: GraphCuts for High Resolution Images Using Graph Reduction

Björn Scheuermann and Bodo Rosenhahn
In: Proc of the Eighth International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR), July 2011, Saint Petersburg, Russia

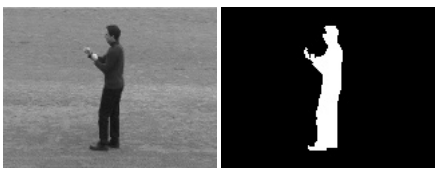
Ego-Motion Compensated Face Detection on a Mobile Device

Björn Scheuermann, Arne Ehlers, Hamon Riazy, Florian Baumann and Bodo Rosenhahn
In: Proc. of the Seventh IEEE Workshop on Embedded Computer Vision (ECVW2011, CVPR Workshop), June 2011, Colorado Springs, USA

Interactive Image Segmentation Using Level Sets and Dempster-Shafer Theory of Evidence

Björn Scheuermann and Bodo Rosenhahn
In: Proc. of the 17th Scandinavian Conference on Image Analysis (SCIA), May 2011, Ystad Saltsjöbad, Sweden

Feature quarrels: The Dempster-Shafer Evidence Theory for Image Seg-



mentation Using a Variational Framework

Björn Scheuermann and Bodo Rosenhahn
In: Proc. of the Tenth Asian Conference on Computer Vision (ACCV), November 2010, Queenstown, New Zealand

N-View Human Silhouette Segmentation in Cluttered, Partially Changing Environments

Tobias Feldmann, Björn Scheuermann, Bodo Rosenhahn and Annika Wörner
In: Proc. of the 32nd Annual Symposium of the German Association for Pattern Recognition (DAGM), September 2010, Darmstadt, Germany

Automated Extraction of Plantations from Ikonos Satellite Imagery using a Level Set Based Segmentation Method

Karsten Vogt, Björn Scheuermann, Christian Becker, Torsten Büschenfeld, Bodo Rosenhahn and Jörn Ostermann
In: Proc. of the SPRS Technical Commission VII Symposium, July 2010, Vienna, Austria

Analysis of Numerical Methods for Level Set Based Image Segmentation

Björn Scheuermann and Bodo Rosenhahn
In: Proc. of the Fifth International Symposium on Visual Computing (ISVC), November 2009, Las Vegas, USA (Accepted for oral presentation)

B.2 Patents

Markus Schlosser, Björn Scheuermann, Bodo Rosenhahn. **"Method and apparatus for multi-label segmentation"**. European Patent App. # 13155916 (Publication date: September 11, 2013)

Björn Scheuermann, Oliver Jakob Arndt, Bodo Rosenhahn. **"Method and apparatus for image segmentation"**. European Patent App. # 13162354 (Publication date: October 23, 2013)

Bibliography

[AB94] Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 16(6):641–647, 1994.

[AMFM11] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(5):898–916, 2011.

[AO07] Tomasz Adamek and Noel E. O'Connor. Using dempster-shafer theory to fuse multiple information sources in region-based segmentation. In *International Conference on Image Processing (ICIP)*, pages 269 – 272. IEEE, 2007.

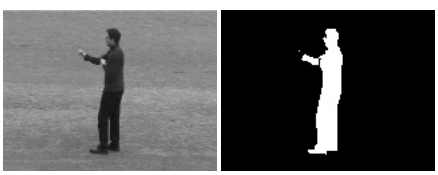
[AS95] David Adalsteinsson and James A. Sethian. A fast level set method for propagating interfaces. *Journal of Computational Physics*, 118(2):269 – 277, 1995.

[ASR13] Oliver Jakob Arndt, Björn Scheuermann, and Bodo Rosenhahn. "Region Cut" - interactive multi-label segmentation utilizing cellular automaton. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 309–316. IEEE, 2013.

[BC07] Thomas Brox and Daniel Cremers. On the statistical interpretation of the piecewise smooth mumford-shah functional. In *Scale Space and Variational Methods in Computer Vision*, pages 203–213. Springer, 2007.

[BElV+07] Xavier Bresson, Selim Esedo lu, Pierre Vandergheynst, Jean-Philippe Thiran, and Stanley Osher. Fast global minimization of the active contour/snake model. *Journal of Mathematical Imaging and vision*, 28(2):151–167, 2007.

[Beu91] Serge Beucher. The watershed transformation applied to image segmentation. In *Conference on Signal and Image Processing in Microscopy and Microanalysis*, pages 299–314, 1991.



[BFL06] Yuri Boykov and Gareth Funka-Lea. Graph cuts and efficient nd image segmentation. *International Journal of Computer Vision (IJCV)*, 70(2):109–131, 2006.

[BGW91] Josef Bigün, Goesta H. Granlund, and Johan Wiklund. Multidimensional orientation estimation with applications to texture analysis and optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 13(8):775–790, 1991.

[BJ01] Yuri Boykov and Marie-Pierre Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *IEEE 8th International Conference on Computer Vision (ICCV)*, volume 1, pages 105–112. IEEE, 2001.

[BK96] András A. Benczúr and David R. Karger. Approximating s-t minimum cuts in $\tilde{O}(n^2)$ time. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, STOC '96, pages 47–55, 1996.

[BK03] Yuri Boykov and Vladimir Kolmogorov. Computing geodesics and minimal surfaces via graph cuts. In *IEEE 9th International Conference on Computer Vision (ICCV)*, pages 26–33. IEEE, 2003.

[BK04] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(9):1124–1137, 2004.

[BKR11] Andrew Blake, Pushmeet Kohli, and Carsten Rother. *Markov random fields for vision and image processing*. The MIT Press, 2011.

[BRB+04] Andrew Blake, Carsten Rother, Matthew Brown, Patrick Perez, and Philip Torr. Interactive image segmentation using an adaptive gmmrf model. In *European Conference Computer Vision (ECCV)*, volume 3021 of *Lecture Notes in Computer Science*, pages 428–441. Springer, 2004.

[Bri03] Robert Edward Bridson. *Computational aspects of dynamic surfaces*. PhD thesis, Stanford University, 2003.

[BS07] Xue Bai and Guillermo Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *IEEE 11th International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007.

[BT08] Arvind Bhusnurmath and Camillo J. Taylor. Graph cuts via l_1 norm minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(10):1866–1871, 2008.

[BVZ98] Yuri Boykov, Olga Veksler, and Ramin Zabih. Markov random fields with efficient approximations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 648–655. IEEE, 1998.

[BVZ01] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(11):1222–1239, 2001.

[BW06] Thomas Brox and Joachim Weickert. A TV flow based local scale estimate and its application to texture discrimination. *Journal of Visual Communication and Image Representation*, 17(5):1053–1073, 2006.

[Can86] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 8(6):679–698, 1986.

[CCBK06] Antonio Criminisi, Geoffrey Cross, Andrew Blake, and Vladimir Kolmogorov. Bilayer segmentation of live video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 53–60. IEEE, 2006.

[CFRA07] Daniel Cremers, Oliver Fluck, Mikael Rousson, and Shmuel Aharon. A probabilistic level set formulation for interactive organ segmentation. In *Proceedings of the SPIE Medical Imaging*, volume 6512, 2007.

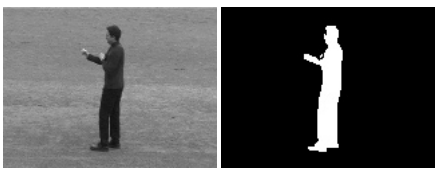
[CGK+97] Chandra S. Chekuri, Andrew V. Goldberg, David R. Karger, Matthew S. Levine, and Cliff Stein. Experimental study of minimum cut algorithms. In *Proceedings of the 8th annual ACM-SIAM symposium on Discrete algorithms*, SODA '97, pages 324–333. Society for Industrial and Applied Mathematics, 1997.

[CKS97] Vicent Casseles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *International Journal of Computer Vision (IJCV)*, 22(1):61–97, 1997.

[CLW04] Joshua E Cates, Aaron E Lefohn, and Ross T Whitaker. Gist: an interactive, gpu-based level set segmentation tool for 3d medical images. *Medical Image Analysis*, 8(3):217–231, 2004.

[CM02] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(5):603–619, 2002.

[CRD07] Daniel Cremers, Mikael Rousson, and Rachid Deriche. A review of statistical approaches to level set segmentation: integrating color, texture,



motion and shape. *International Journal of Computer Vision (IJCV)*, 72(2):195–215, 2007.

[CS05] Daniel Cremers and Stefano Soatto. Motion competition: A variational approach to piecewise parametric motion segmentation. *International Journal of Computer Vision (IJCV)*, 62(3):249–265, 2005.

[CSB08] Daniel Cremers, Frank R. Schmidt, and Frank Barthel. Shape priors in variational image segmentation: Convexity, lipschitz continuity and globally optimal solutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6. IEEE, 2008.

[CSFB08] Salim Ben Chaabane, Mounir Sayadi, Farhat Fnaiech, and Eric Brassart. Color image segmentation based on dempster-shafer evidence theory. In *IEEE 14th Mediterranean Electrotechnical Conference*, pages 862–866. IEEE, 2008.

[CSFB09] Salim Ben Chaabane, Mounir Sayadi, Farhat Fnaiech, and Eric Brassart. Dempster-shafer evidence theory for image segmentation: application in cells images. *International Journal of Signal Processing*, 2009.

[CSRO12] Kai Cordes, Björn Scheuermann, Bodo Rosenhahn, and Jörn Ostermann. Learning object appearance from occlusions using structure and motion recovery. In *The 11th Asian Conference on Computer Vision (ACCV)*, volume 7726 of *Lecture Notes in Computer Science*, pages 611–623. Springer, 2012.

[CSS03] Daniel Cremers, Nir Sochen, and Christoph Schnörr. Towards recognition-based variational segmentation using shape priors and dynamic labeling. In *Scale Space Methods in Computer Vision*, volume 2695 of *Lecture Notes in Computer Science*, pages 388–400. Springer, 2003.

[CSV00] Tony F. Chan, B. Yezrielev Sandberg, and Luminita A. Vese. Active contours without edges for vector-valued images. *Journal of Visual Communication and Image Representation*, 11(2):130–141, 2000.

[CV99] Tony F. Chan and Luminita A. Vese. An active contour model without edges. In *Scale-Space Theories in Computer Vision*, volume 1682 of *Lecture Notes in Computer Science*, pages 141–151. Springer, 1999.

[CV01] Tony F. Chan and Luminita A. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001.

[CZ05] Tony F. Chan and Wei Zhu. Level set based shape prior segmentation. In *IEEE 10th International Conference on Computer Vision (ICCV)*, volume 2, pages 1164–1170. IEEE, 2005.

[DB08] Andrew Delong and Yuri Boykov. A scalable graph-cut algorithm for nd grids. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[DB09] Andrew Delong and Yuri Boykov. Globally optimal segmentation of multi-region objects. In *IEEE 12th International Conference on Computer Vision (ICCV)*, pages 285–292. IEEE, 2009.

[Dem68] Arthur P. Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):205–247, 1968.

[DPTA+13] Foued Derraz, Laurent Peyrodie, Abdelmalik Taleb-Ahmed, Miloud Boussahla, and Gerard Forzy. Fast unsupervised segmentation using active contours and belief functions. In *Computer Analysis of Images and Patterns*, volume 8047 of *Lecture Notes in Computer Science*, pages 278–285. Springer, 2013.

[DZ86] Silvano Di Zenzo. A note on the gradient of a multi-image. *Computer Vision, Graphics, and Image Processing*, 33(1):116–125, 1986.

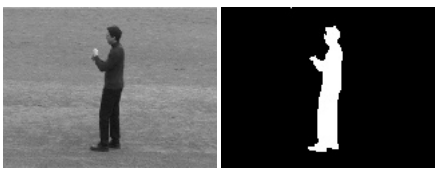
[EFS56] Peter Elias, Amiel Feinstein, and Claude Shannon. A note on the maximum flow through a network. *IRE Transactions on Information Theory*, 2(4):117–119, 1956.

[FB05] Jean-Sébastien Franco and Edmond Boyer. Fusion of multi-view silhouette cues using a space occupancy grid. Technical Report 5551, INRIA, April 2005.

[FDW09] Tobias Feldmann, Lars Dießelberg, and Annika Wörner. Adaptive foreground/background segmentation using multiview silhouette fusion. In *31st DAGM Symposium*, volume 5748 of *Lecture Notes in Computer Science*, pages 522–531. Springer, 2009.

[FF56] Lester R. Ford and Delbert R. Fulkerson. Maximum flow through a network. *Canadian Journal of Mathematics (CJM)*, 8:299–404, 1956.

[FH04] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision (IJCV)*, 59(2):167–181, 2004.



[FSRW10] Tobias Feldmann, Björn Scheuermann, Bodo Rosenhahn, and Annika Wörner. N-view human silhouette segmentation in cluttered, partially changing environments. In *32nd DAGM Symposium*, volume 6376 of *Lecture Notes in Computer Science*, pages 363–372. Springer, 2010.

[GBO10] Tom Goldstein, Xavier Bresson, and Stanley Osher. Geometric applications of the split bregman method: segmentation and surface reconstruction. *Journal of Scientific Computing*, 45(1-3):272–293, 2010.

[GG84] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 6(6):721–741, 1984.

[GPS89] D. M. Greig, B. T. Porteous, and Allan H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 271–279, 1989.

[HC71] John M. Hammersley and Peter Clifford. Markov fields on finite graphs and lattices. *Unpublished manuscript*, 1971.

[HGW01] Michael Harville, Gaile Gordon, and John Woodfill. Foreground segmentation using adaptive mixture models in color and depth. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 3–11. IEEE, 2001.

[HKR+01] Karin Hogstedt, Doug Kimelman, V. T. Rajan, Tova Roth, and Mark Wegman. Graph cutting algorithms for distributed applications partitioning. *ACM SIGMETRICS Performance Evaluation Review*, 28(4):27–29, 2001.

[HM08] Michal Haindl and Stanislav Mikeš. Texture segmentation benchmark. In *19th International Conference on Pattern Recognition (ICPR)*, pages 1–4. IEEE, 2008.

[HNB+06] Ben Houston, Michael B. Nielsen, Christopher Batty, Ola Nilsson, and Ken Museth. Hierarchical rle level set: A compact and versatile deformable surface representation. *ACM Transactions on Graphics (TOG)*, 25(1):151–175, 2006.

[HP60] Paul V. C. Hough and Brian W. Powell. A method for faster analysis of bubble chamber photographs. *Il Nuovo Cimento*, 18(6):1184–1191, 1960.

[HP74] Steven L. Horowitz and Theodosios Pavlidis. Picture segmentation by a directed split-and-merge procedure. In *Proceedings of the second international joint conference on pattern recognition*, volume 424, page 433, 1974.

[HS05] Matthias Heiler and Christoph Schnörr. Natural image statistics for natural image segmentation. *International Journal of Computer Vision (IJCV)*, 63(1):5–19, 2005.

[IK88] John Illingworth and Josef Kittler. A survey of the hough transform. *Computer Vision, Graphics, and Image Processing*, 44(1):87–116, 1988.

[Ina91] Toshiyuki Inagaki. Interdependence between safety-control policy and multiple-sensor schemes via dempster-shafer theory. *IEEE Transactions on Reliability*, 40(2):182–188, 1991.

[Inc02] Adobe Systems Incorporated. *Adobe Photoshop User Guide*. Adobe Systems Incorporated, 2002.

[KB03] Timor Kadir and Michael Brady. Unsupervised non-parametric region segmentation using level sets. In *IEEE 9th International Conference on Computer Vision (ICCV)*, pages 1267–1274. IEEE, 2003.

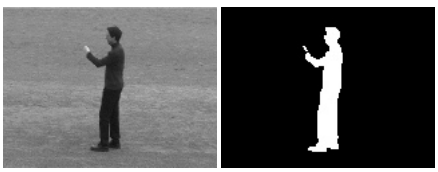
[KCB+05] Vladimir Kolmogorov, Antonio Criminisi, Andrew Blake, Geoff Cross, and Carsten Rother. Bi-layer segmentation of binocular stereo video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 407–414. IEEE, 2005.

[KLR10] Pushmeet Kohli, Victor Lempitsky, and Carsten Rother. Uncertainty driven multi-scale optimization. In *32nd DAGM Symposium*, volume 6376 of *Lecture Notes in Computer Science*, pages 242–251. Springer, 2010.

[KNKY11] Taesup Kim, Sebastian Nowozin, Pushmeet Kohli, and Chang D Yoo. Variable grouping for energy minimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1913–1920. IEEE, 2011.

[Kom10] Nikos Komodakis. Towards more efficient and effective LP-based algorithms for MRF optimization. In *European Conference Computer Vision (ECCV)*, volume 6312 of *Lecture Notes in Computer Science*, pages 520–534. Springer, 2010.

[KS96] David R. Karger and Clifford Stein. A new approach to the minimum cut problem. *Journal of the ACM (JACM)*, 43(4):601–640, 1996.



[KSK+08] Maria Klodt, Thomas Schoenemann, Kalin Kolev, Marek Schikora, and Daniel Cremers. An experimental comparison of discrete and continuous shape optimization methods. In *European Conference Computer Vision (ECCV)*, volume 5302 of *Lecture Notes in Computer Science*, pages 332–345. Springer, 2008.

[KT05] Pushmeet Kohli and Philip H. S. Torr. Efficiently solving dynamic markov random fields using graph cuts. In *IEEE 10th International Conference on Computer Vision (ICCV)*, volume 2, pages 922–929. IEEE, 2005.

[KWT88] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision (IJCV)*, 1(4):321–331, 1988.

[KZ04] Vladimir Kolmogorov and Ramin Zabin. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(2):147–159, 2004.

[LB07] Victor Lempitsky and Yuri Boykov. Global optimization for shape fitting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007.

[LCW03] Aaron E. Lefohn, Joshua E. Cates, and Ross T. Whitaker. Interactive, gpu-based level sets for 3d segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 2878 of *Lecture Notes in Computer Science*, pages 564–572. Springer, 2003.

[LGF04] Frank Losasso, Frédéric Gibou, and Ron Fedkiw. Simulating water and smoke with an octree data structure. *ACM Transactions on Graphics (TOG)*, 23(3):457–462, 2004.

[Li09] Stan Z. Li. *Markov random field modeling in image analysis*. Springer, 2009.

[LLM11] Nicolas Lermé, Létocart Létocart, and François Malgouyres. Reduced graphs for min-cut/max-flow approaches in image segmentation. *Electronic Notes in Discrete Mathematics*, 37(1):63–68, 2011.

[LMP01] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, 2001.

[LSGX05] Herve Lombaert, Yiyong Sun, Leo Grady, and Chenyang Xu. A multilevel banded graph cuts method for fast image segmentation. In *IEEE 10th International Conference on Computer Vision (ICCV)*, volume 1, pages 259–265. IEEE, 2005.

[LSK+09] Alex Levinshtein, Adrian Stere, Kiriakos N. Kutulakos, David J. Fleet, Sven J. Dickinson, and Kaleem Siddiqi. TurboPixels: fast superpixels using geometric flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31(12):2290–2297, 2009.

[LSS09] Jiangyu Liu, Jian Sun, and Heung-Yeung Shum. Paint selection. *ACM Transactions on Graphics (TOG)*, 28(3):69, 2009.

[LSTS04] Yin Li, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Lazy snapping. *ACM Transactions on Graphics (TOG)*, 23(3):303–308, 2004.

[MB98] Eric N. Mortensen and William A. Barrett. Interactive segmentation with intelligent scissors. *Graphical models and image processing*, 60(5):349–384, 1998.

[MFTM01] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE 8th International Conference on Computer Vision (ICCV)*, pages 416–423. IEEE, July 2001.

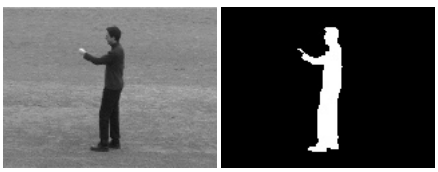
[MM03] Juan B. Mena and José A. Malpica. Color image segmentation using the dempster-shafer theory of evidence for the fusion. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (34) 3/W8*, pages 139–144, 2003.

[MO10] Kevin McGuinness and Noel E. O'Connor. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444, February 2010.

[MS85] David Mumford and Jayant Shah. Boundary detection by minimizing functionals. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22–26. IEEE, 1985.

[MS89] David Mumford and Jayant Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42(5):577–685, 1989.

[MSV95] Ravi Malladi, James A. Sethian, and Baba C. Vemuri. Shape modeling with front propagation: A level set approach. *IEEE Transactions on*



Pattern Analysis and Machine Intelligence (TPAMI), 17(2):158–175, 1995.

[OP03] Stanley Osher and Nikos Paragios. *Geometric Level Set Methods in Imaging, Vision, and Graphics*. Springer, 2003.

[OS88] Stanley Osher and James A. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. *Journal of Computational Physics*, 79(1):12–49, 1988.

[PB99] Jan Puzicha and Joachim M. Buhmann. Multiscale annealing for grouping and unsupervised texture segmentation. *Computer Vision and Image Understanding (IJCVIU)*, 76(3):213–230, 1999.

[PFK+05] Kilian M. Pohl, John Fisher, Ron Kikinis, W. Eric L. Grimson, and William M. Wells. Shape based segmentation of anatomical structures in magnetic resonance images. In *Computer Vision for Biomedical Image Applications*, pages 489–498. Springer, 2005.

[Pot52] Renfrey Burnard Potts. Some generalized order-disorder transformations. *Proceedings of the Cambridge Philosophical Society*, 48(2):106–109, 1952.

[PR90] Manfred Padberg and Giovanni Rinaldi. An efficient algorithm for the minimum capacity cut problem. *Mathematical Programming*, 47(1):19–36, 1990.

[PRR02] Nikos Paragios, Mikael Rousson, and Visvanathan Ramesh. Matching distance functions: A shape-to-area variational approach for global-to-local registration. In *European Conference Computer Vision (ECCV)*, volume 2351 of *Lecture Notes in Computer Science*, pages 775–789. Springer, 2002.

[PS07] Alexis Protiere and Guillermo Sapiro. Interactive image segmentation via adaptive weighted distances. *IEEE Transactions on Image Processing*, 16(4):1046–1057, 2007.

[RBD03] Mikaël Rousson, Thomas Brox, and Rachid Deriche. Active unsupervised texture segmentation on a diffusion based feature space. In *IEEE 9th Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 699–704. IEEE, 2003.

[RBW07] Bodo Rosenhahn, Thomas Brox, and Joachim Weickert. Three-dimensional shape knowledge for joint image segmentation and pose tracking. *International Journal of Computer Vision (IJCV)*, 73(3):243–262, 2007.

[RKB04] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In 3, editor, *ACM Transactions on Graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[RP02] Mikael Rousson and Nikos Paragios. Shape priors for level set representations. In *European Conference Computer Vision (ECCV)*, volume 2351 of *Lecture Notes in Computer Science*, pages 78–92. Springer, 2002.

[RPSM10] Mike Roberts, Jeff Packer, Mario Costa Sousa, and Joseph Ross Mitchell. A work-efficient gpu algorithm for level set segmentation. In *Conference on High Performance Graphics*, pages 123–132. Eurographics Association, 2010.

[RZ02] Michèle Rombaut and Yue Min Zhu. Study of dempster–shafer theory for image segmentation applications. *Image and Vision Computing*, 20(1):15–23, Januar 2002.

[Set96] James A. Sethian. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences*, 93(4):1591–1595, 1996.

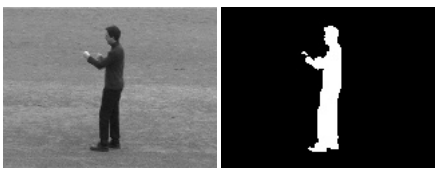
[Set99] James Albert Sethian. *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*, volume 3. Cambridge university press, 1999.

[SF02] Kari Sentz and Scott Ferson. Combination of evidence in Dempster-Shafer theory, sandia national laboratories. Technical report, No. SAND2002-0835, 2002.

[SG06] Ali K. Sinop and Leo Grady. Accurate banded graph cut segmentation of thin structures using laplacian pyramids. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 4191 of *Lecture Notes in Computer Science*, pages 896–903. Springer, 2006.

[SGR13] Björn Scheuermann, Sotirios Gkoutelitsas, and Bodo Rosenhahn. Multi-sensor fusion using dempster's theory of evidence for video segmentation. In *The 18th Iberoamerican Congress on Pattern Recognition (CIARP)*, volume 8259 of *Lecture Notes in Computer Science*, pages 431–438. Springer, 2013.

[Sha76] Glenn Shafer. *A mathematical theory of evidence*. Princeton university press, 1976.



[SHB07] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image processing, analysis, and machine vision.* Thomson-Engineering, 2007.

[SK10] Petter Strandmark and Fredrik Kahl. Parallel and distributed graph cuts by dual decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2085–2092. IEEE, 2010.

[SLC04] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local SVM approach. In *17th International Conference on Pattern Recognition (ICPR)*, pages 32–36, 2004.

[SM00] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(8):888–905, 2000.

[SR09] Björn Scheuermann and Bodo Rosenhahn. Analysis of numerical methods for level set based image segmentation. In *5th International Symposium on Advances in Visual Computing (ISVC)*, Lecture Notes in Computer Science, pages 196–207. Springer, 2009.

[SR10] Björn Scheuermann and Bodo Rosenhahn. Feature quarrels: The dempster-shafer evidence theory for image segmentation using a variational framework. In *The 10th Asian Conference on Computer Vision (ACCV)*, volume 6493 of *Lecture Notes in Computer Science*, pages 426–439. Springer, 2010.

[SR11a] Björn Scheuermann and Bodo Rosenhahn. Interactive image segmentation using level sets and dempster-shafer theory of evidence. In *Scandinavian Conference on Image Analysis (SCIA)*, volume 6688 of *Lecture Notes in Computer Science*, pages 656–665. Springer, 2011.

[SR11b] Björn Scheuermann and Bodo Rosenhahn. Slimcuts: Graphcuts for high resolution images using graph reduction. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 6819 of *Lecture Notes in Computer Science*, pages 219–232. Springer, 2011.

[SS04] Memet Sezgin and Bülent Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–168, 2004.

[SSR12] Björn Scheuermann, Markus Schlosser, and Bodo Rosenhahn. Efficient pixel-grouping based on dempster's theory of evidence for image segmentation. In *The 11th Asian Conference on Computer Vision (ACCV)*, volume 7724 of *Lecture Notes in Computer Science*, pages 745–759. Springer, 2012.

[ST04] Daniel A. Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *36th symposium on Theory of computing (STOC)*, pages 81–90. ACM, 2004.

[ST06] Peter Sand and Seth J. Teller. Particle video: long-range motion estimation using point trajectories. In *IEEE 12th Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2195–2202. IEEE, 2006.

[Str99] John Strain. Tree methods for moving interfaces. *Journal of Computational Physics*, 151(2):616–648, 1999.

[SV05] Glenn Shafer and Vladimir Vovk. The origins and legacy of Kolmogorvs Grundbegriffe. *The game-theoretic probability and finance project working paper*, 4:8, 2005.

[Sze10] Richard Szeliski. *Computer vision: algorithms and applications*. Springer, 2010.

[Toi96] Pekka J. Toivanen. New geodosic distance transforms for gray-scale images. *Pattern Recognition Letters*, 17(5):437–450, 1996.

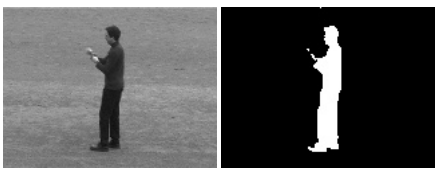
[UPT+08] Markus Unger, Thomas Pock, Werner Trobin, Daniel Cremers, and Horst Bischof. Tvseg-interactive total variation based image segmentation. In *British Machine Vision Conference (BMVC)*, pages 1–10. Citeseer, 2008.

[VBM10] Olga Veksler, Yuri Boykov, and Paria Mehrani. Superpixels and supervoxels in an energy optimization framework. In *European Conference Computer Vision (ECCV)*, volume 6315 of *Lecture Notes in Computer Science*, pages 211–224. Springer, 2010.

[VK05] V. Vezhnevets and V. Konouchine. Growcut: Interactive multi-label nd image segmentation by cellular automata. In *Proceedings of Graphicon*, pages 150–156, 2005.

[VN08] Vibhav Vineet and PJ Narayanan. Cuda cuts: Fast graph cuts on the gpu. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–8. IEEE, 2008.

[VS91] Luc Vincent and Pierre Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 13(6):583–598, 1991.



[WBC+05] Jue Wang, Pravin Bhat, R. Alex Colburn, Maneesh Agrawala, and Michael F. Cohen. Interactive video cutout. *ACM Transactions on Graphics (TOG)*, 24(3):585–594, 2005.

[Whi98] Ross T. Whitaker. A level-set approach to 3d reconstruction from range data. *International Journal of Computer Vision (IJCV)*, 29(3):203–231, 1998.

[WZYZ10] Liang Wang, Chenxi Zhang, Ruigang Yang, and Cha Zhang. Tofcut: Towards robust real-time foreground extraction using a time-of-flight camera. In *Fifth International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 2010.

[Yag87] Ronald R. Yager. On the dempster-shafer framework and new combination rules. *Information sciences*, 41(2):93–137, 1987.

[Zuc76] Steven W. Zucker. Region growing: Childhood and adolescence. *Computer graphics and image processing*, 5(3):382–399, 1976.

[ZY96] Song Chun Zhu and Alan Yuille. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 18(9):884–900, 1996.

Curriculum Vitae

Björn Scheuermann

16.02.1983	geboren in Gehrden, Germany

Beruf

seit 10/2014	Entwicklungsingenieur bei der *Robert Bosch GmbH*, Hildesheim
10/2008 - 09/2014	wissenschaftlicher Mitarbeiter am *Institut für Informationsverarbeitung* an der *Leibniz Universität Hannover*

Studium

10/2003 - 09/2008	Studium der Mathematik mit Studienrichtung Informatik und Anwendungsfach Wirtschaftsinformatik an der *Leibniz Universität Hannover* Abschluss mit Diplom (Dipl.-Math.)

Bundeswehr

10/2002 - 06/2003	Grundwehrdienst

Schulbildung

08/1999 - 06/2002	Gymnasiale Oberstufe der KGS Ronnenberg Abschluss Abitur (Durchschnittsnote: 2.8)
08/1995 - 07/1999	Gymansialzweig der KGS Ronnenberg
08/1993 - 07/1995	Orientierungsstufe Ronnenberg
08/1989 - 07/1993	Theodor-Heuss-Grundschule Empelde