SUPERPIXEL BASED EFFICIENT IMAGE REPRESENTATION FOR SEGMENTATION AND CLASSIFICATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

H. EMRAH TAŞLI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

JUNE 2013

ABSTRACT

SUPERPIXEL BASED EFFICIENT IMAGE REPRESENTATION FOR SEGMENTATION AND CLASSIFICATION

Taşlı, H. Emrah

Ph.D., Department of Electrical and Electronics Engineering

Supervisor : Prof. Dr. A. Aydın Alatan

June 2013, 140 pages

The wide availability of visual capture and display devices with increasing resolution and affordable prices, made the visual data an indispensable part of our life. The enormous amount of visual data produced every day is captured, stored and sometimes processed for further analysis. In this era of technological improvement, where an exponential increase in the number and capability of the devices is experienced, researchers have focused on efficient and accurate ways to reach, store, analyse and display the data for various purposes.

At the capture side of the visual content, the number of cameras has rapidly increased in close correlation to the number of mobile phones with built in cameras. As with the quantity increase, the quality of the sensors have also boosted regarding the resolution, color/brightness and noise level performance. On the other side of the pipeline, there has been some major changes at the display side over the last couple of decades. With the introduction of the Plasma and LCD (Liquid-crystal-display) type of displays, sizes have rapidly decreased in the depth dimension. This decrease also made the mobility of the displays possible especially with lower power consumptions. Therefore, mobile equipments with high resolution displays could easily fit in our pockets. Moreover, another major stepping stone towards a richer visual experience is observed with the introduction of 3D capable displays for different sizes and resolutions.

There has been a major increase in the popularity of 3D TVs in the last couple of years. Mobile devices with 3D capability have also been introduced in the market. However, the fast increase in the display side could not be matched as well in the capture and

broadcast side. Therefore, the popularity of the 3D devices have been lower than the expectations. Various factors could be counted as a cause for such a slower reaction. These factors and possible solutions for such problems are presented in this thesis.

This thesis deals with various aspects of the research in visual content analysis and display technologies. The author's previous experience in real time processing of image/video data, human visual perspectives for objective/subjective quality analysis, stereoscopy and 3D perception, image understanding for object recognition, image feature descriptors using low-, mid- and region- level visual cues have been vastly incorporated in this thesis. Applications of the proposed techniques for real world scenarios have been conducted and results are supported with performance evaluations using objective and subjective quality metrics.

Superpixel extraction is proposed as an efficient image representation tool. It has been shown to offer computational efficiency with high segmentation performance. Extraction of the superpixel has been realized using a color and spatial distance metric where the weighting is defined as a trade-off parameter. With extensive comparative tests with the state-of-the-art, the proposed scheme is shown to yield a remarkable alternative in the current superpixel and supervoxel extraction methods with faster execution times and competitive segmentation performances.

The extracted superpixels have been further utilized for user-assisted image segmentation purposes. User assistance is required as drawing lines on the representative parts of the image to define foreground and background regions. An energy minimization technique is then used to define most likely regions to be segmented. The acquired foreground segments could further be used for rendering the stereo pair of an image for 3D visualization purposes. The same energy formulization is also extended on the stereo and video footage for completeness.

The segmented superpixel patches are also presented as mid-level information sources and applied on the image classification task. Pixel-wise image descriptors are studied and extended using the proposed mid-level region descriptor in order to capture the complementary mid-level information present in the image. The experimental results have shown supporting evidence for the proposal where classification scores has considerably increased.

Keywords: Superpixel Extraction, Image Segmentation, Video Segmentation, Graphcuts, 2D/3D Conversion, Stereo Disparity Remapping, Object Recognition, Spatial Pooling, Mid-Level Feature Descripton

ÖZ

BÖLÜTLEME VE SINIFLANDIRMA İÇİN SÜPERPİKSEL TEMELLİ ETKİN İMGE SİMGELEME

Taşlı, H. Emrah

Doktora, Elektrik ve Elektronik Mühendisliği Bölümü

Tez Yöneticisi : Prof. Dr. A. Aydın Alatan

Haziran 2013 , 140 sayfa

Görsel kayıt ve gösterim cihazlarındaki çözünürlük artışı ile birlikte gelen ekonomik satış fiyatları, görsel bilginin hayatın vazgeçilmez bir öğesi olmasına sebep olmuştur. Her gün çok büyük miktarda görsel veri kaydedilip depolandıktan sonra, belki de farklı amaçlar için tekrar işlenerek anlamlandırılmaktadır. Teknolojinin hızla geliştiği ve görsel kayıt cihazların sayısının hızlıca arttığı bu zamanda, araştırmacılar bu büyük veriyi ulaşılır kılmanın ve gerektiğinde farklı amaçlar için işlemenin en verimli yollarını aramaktadır.

Her gün kaydedilen görsel veri miktarındaki artış, sayıları hızla artan taşınabilir cihazların artması ile ilişkilendirilebilir. Bu cihazlardaki sayısal artışın yanı sıra, görsel kalitede de çözünürlük, renk, aydınlık ve gürültü bakımından artış kaydedilmiştir. Diğer tarafta ekran teknolojilerinde de, son çeyrek asırda önemli gelişmeler yaşanmıştır. Plazma ve LCD ekran teknolojilerinin yaygınlaşması ile televizyon ebatlarında derinlik açısından ciddi azalma olmuştur. Bu aynı zamanda, taşınabilir ekranların özellikle düşük enerji tüketimleri ile yaygınlık kazanarak ceplerimize girmelerine sebep olmuştur. Bir başka önemli adım ise üç boyutlu ekranların yaygınlaşarak daha zengin bir görsel deneyim ile tanışmamızı sağlamış olmalarıdır.

Üç boyutlu televizyonlarda son on yıl içinde ciddi bir artış gözlenmiştir. Buna ek olarak üç boyutlu mobil ekranlar da üretilerek tüketiciye sunulmuştur. Fakat, ekran sayısındaki artış içerik üreticileri tarafından aynı oranda karşılık görememiştir. Sonuç olarak, üç boyutlu cihazlar beklenenin altında ilgi görmüştür. Bu durumun altında yatan sebepler ve çözüm önerileri bu tezde sunulmaktadır.

Bu tez, görsel içerik analizinden, görselleştirme teknolojileri konusuna kadar farklı alanlara değinmektedir. Gerçek zamanlı görüntü ve vidyo işleme, insan görsel perspektifi temelli öznel ve nesnel görsel kalite analizi, stereoskopi ve üç boyut algısı, görüntü anlama ve nesne tanıma, alt orta seviye ve bölgesel imge öznitelik tanımlayıcıları gibi konular bu tezde incelenmektedir. Anlatılan yöntemler gerçek hayat senaryolarına uygulanarak sonuçları öznel ve nesnel kalite ölçümleri ile değerlendirilmiştir.

Superpiksel çıkarımı verimli bir imge simgeleme yöntemi olarak sunulmaktadır. Bu şekilde bölütleme performansında artış ve işlem karmaşıklığında ciddi kazanımlar sağlanabilmektedir. Süperpiksel çıkarımında renk ve uzamsal yakınlık kriterlerine dayanan bir metrik kullanılmıştır. Detaylı nesnel karşılaştırmalar ile değerlendirilen yöntem, işlem hızı ve bölütleme performansı ile güncel metotlara ciddi bir alternatif oluşturmaktadır.

Oluşturulan süperpiksel bölgeleri, kullanıcı etkileşimli imge bölütleme yöntemi için kullanılmaktadır. Kullanıcı, imge üzerindeki belirleyici alanları işaret ederek nesne ve arka fon bölütlemesi için sisteme bilgi vermektedir. Bu bilgi ile oluşturulan enerji fonksiyonu en aza indirgenerek, sahnenin bölütlenmesi sağlamaktadır. Elde edilen nesne sınırları, stereo görüntü sentezinde kullanılarak üç boyutlu görselleştirme sağlayabilmektedir. Önerilen yöntem ek olarak stereo ve video içerklere de uygulanarak bütünlük sağlanmıştır.

Süperpiksel bölgeleri ayrıca orta seviye bilgi kaynağı olarak ele alınarak görüntü sınıflandırma probleminde kullanılmışlardır. Güncel piksel temelli öznitelik tanımlıyıcıları örnek alınarak, orta seviye bir imge tanımlama yöntemi önerilmektedir. Bu sayede, alt seviyede yapılan bilgi çıkarımının orta seviyeye de aktarılarak bütünleyici bir yaklaşım sergilenmesi mümkün olabilmektedir. Deneysel çalışmalar ile de destekleyici yönde sonuçlar gözlenmiştir.

Anahtar Kelimeler: Süperpiksel Çıkarımı, İmge Bölütleme, Video Bölütleme, Çizge Kesit, 2 Boyut / 3 Boyut Dönüşümü, Stereo Ayrıklık Tekrar Hedeflenmesi, Nesne Tanıma, Bölgesel Gruplama, Orta Düzey Öznitelik Betimleme

To you, my family..

ACKNOWLEDGMENTS

Looking back five years, the very first time I thought about pursuing a PhD, I would have never imagined such an experience. Hereby, I would like to name a few of the many people who have been with me during this incredible journey.

First of all, I would like to express my sincere gratitude to my advisor, Prof. A. Aydın Alatan for his guidance, support and the friendly research environment he has provided from the bachelor years to my PhD. I am also grateful to him for introducing me to image processing, for his tolerance and support in my atypical and nourishing career path that includes four (and possibly counting) different countries. I would also like to thank my advisor at the University of Amsterdam, Prof. Theo Gevers for his research ideas, his vision and his trust in me.

I would like to acknowledge my thesis progress committee members Prof. Gözde Bozdağı Akar and Assoc. Prof. Uğur Güdükbay, for their valuable feedbacks in our meetings during the last three years. I am also grateful to my thesis committee members Prof. Uğur Halıcı and Assoc. Prof. Çağatay Candan for their remarks and contributions. Lastly, I would like to thank all my professors at METU for their teaching efforts throughout my educational life.

It would not have been possible to write this thesis without the support and joyful company of the many people around me. I will never forget the times we spent at Vestek R&D with Cevahir Çığla and Burak Özkalaycı. I believe, we have been through a very unique experience that I will always gladly remember. Our research with Cevahir has also produced a crucial part of this thesis. For a short time during the qualification exam, Vestek office has hosted Ahmet Saraçoğlu and Serdar Gedik and I am grateful to the nights and pizzas we shared with these four people during the educational and collaborative discussion hours. I would also like to thank my colleagues at Vestek İstanbul for their deep experience and the friendly environment they created. I have always enjoyed the energy, freedom and the creativity during

the unusual working hours there. I also would like to thank the former and current members of the Multimedia Research Group; especially, Yağız Aksoy for his generous help during the administrative processes when I couldn't be there. I would like to acknowledge my colleagues at the Nokia Research Center, especially Kemal Ugur for his hospitality, help and the fruitful discussions we had during my visit. Of course, I like to thank to many other people whom I could not mention for their help, presence and friendship.

Finally, I would like to give my greatest gratitude to my very big family, for their unconditional support and belief. My father Necip Taşlı, for being a good example in every sense; as a father, as an engineer and as a person, I want to thank him for the late car drives in order to get me to sleep when I was a child. My mother Necla Taşlı, for her understanding so that I could freely explore life and for her patience during the hard times I gave her with my stubborn curiosity. My sister Heval, for her unique presence, continuous support and love. My grandparents Makbule, Huseyin and my aunt Emine for their unconditional love and efforts for raising me during my childhood. My parents-in-law Ayşe-Cevat Tığlı, for their motivation and interest in my research. My wife, Özge; you have been always by my side at the ups and downs. Thank you for your imaginative drawings, for our wedding invitation and for your good night fairy tails. I am especially happy that you could take me with you to the wonderland whenever I needed. I hope I can ever return your support and understanding.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

The wide availability of visual capture and display devices with increasing resolution and affordable prices, made the visual data an indispensable part of our life. The enormous amount of visual data that is produced every day is captured, stored and sometimes processed for further analysis. In this era of technological improvement, where an exponential increase in the number and capability of the devices is experienced, researchers have focused on efficient and accurate ways to reach, store, analyse and display the data for various purposes.

At the capture side of the visual content pipeline, the number of cameras has rapidly increased in close correlation to the number of mobile phones with built in camera functionality. As with the quantity increase, the quality of the sensors have also boosted in regards to resolution, color/brightness and noise level performance. On the other side of the pipeline, there has been some major changes at the display side over the last couple of decades. With the introduction of the plasma and LCD (Liquid Crystal Display) type of displays, sizes have rapidly decreased in the depth dimension. This decrease also made the mobility of the displays possible especially with lower power consumptions of LED (Light Emitting Diode) type lightning equipment. Therefore, mobile equipments with high resolution displays could easily fit in our pockets. Moreover, another major stepping stone towards a richer visual experience is observed with the introduction of 3D capable displays for different sizes and resolutions.

There has been a major increase in the popularity of 3D TVs in the last couple of years. Mobile devices with 3D capability have also been introduced in the market. However, the fast increase in the display side could not be matched as well in the capture and broadcast side. Therefore, the popularity of the 3D devices have been lower than the expectations. Various factors could be counted as a cause for such a slower reaction. These factors and possible solutions for such problems are presented in this thesis.

This thesis deals with various aspects of the research in visual content analysis and display technologies. The author's previous experience in real time processing of image/video data, human visual perspectives for objective/subjective quality analysis, stereoscopy and 3D perception, image understanding for object recognition, image

feature descriptors using low-, mid- and region- level visual cues have been vastly incorporated in this thesis. Applications of the proposed techniques for real world scenarios have been conducted and results are supported with performance evaluations using objective and subjective quality metrics.

1.1 Problem Definition

The current capture and display mediums are fixed to the shape of a rectangle. However, this is the result of an evolution starting from a round shape similar to the lens of a camera. One of the many reasons for such an evolution is the ease of image representation using fixed number of elements in the rows and columns of a matrix. Such a representation has supplied various advantages in the data compression and broadcast aspects of the technology. In this matrix representation, the value of each element corresponds to the color and brightness information corresponding to a specific physical location on the image.

There has been previous efforts towards alternative image representations from the perspective that how the matrix values should be defined so that efficient processing might be possible. The contour representation technique presented in [11], computes the boundaries of connected regions of pixels for a given gray level. For each gray level the boundaries of connected regions of pixels are computed. Reconstruction of the original image is possible from the assigned boundary values. Representation of an image as a collection of those boundary lines (contour lines) associated with gray levels is the contour representation of an image.

Pixel representation of an image is often redundant due to the spatial coherence within the image. In order to reduce this redundancy, a preprocessing stage is proposed. The idea of generating a representative region for small image patches is not totally new, however *superpixel* naming convention is introduced by Ren et al. [103]. The method groups pixels into homogeneous image regions, called superpixels (SPs). This preprocessing step has been useful in many image processing applications. Since the superpixel regions on the image, possess similar color and texture characteristics, they provide an efficient representation. This property supports the assumption that pixels in the same superpixel belong to the same semantic object. Inspiring from this idea, all the pixels in a superpixel can be assigned to specific models representing motion, depth or segmentation structures. Such a representation can replace the utilization of pixel primitives in various applications [122, 12]. By the utilization of superpixels as the image representation, the inter-pixel details are captured and preserved in the image. Furthermore, the extracted superpixel structure is crucial for graph-based approaches. When the graph nodes are constructed with superpixels instead of pixels, graph complexity and computation time would substantially reduce.

This thesis proposes a method for extracting superpixels using image color and tex-

tural characteristics and presents the ways to utilize them in the image segmentation and classification framework. Such utilization proposes an efficient representation of the image and hence serves the purpose of reducing the computational complexity of the pixel based methods. Superpixel representation is also proposed as a mid-level information in the object recognition framework.

1.2 Existing Solutions

Efficient representation of an images has been previously addressed in [66]. The epitome of an image is defined as its miniature, condensed version containing the essence of the textural and shape properties. In this sense, Superpixels are also seen as an efficient image representation with reduced resolution (number of graph nodes) and information encapsulation property.

The main purpose of the superpixel extraction is to create an alternative representation of the image. This has been observed to be beneficial for computational and representation purposes. Superpixel extraction is achieved by partitioning the graph where nodes correspond to individual pixels and edge weights are assigned according to a cost function relating inter pixel similarities. In [76], the graph is partitioned recursively in order to minimize a global cost function based on color and texture cues until desired number of superpixels is achieved. This approach satisfies the compactness constraint required for superpixels in order to provide efficient graph representation. However, it suffers from computational complexity. In [47], superpixel extraction is improved in terms of complexity by grouping nodes of the graph via greedy decisions through pairwise region comparisons on edge measures of minimum spanning trees. This method, on the other hand, does not enforce a control on region compactness and number of superpixels. In [88], a lattice structure is enforced by finding horizontal and vertical seams that cut the image optimally via graph cuts. The seams determine the SP boundaries based on region compactness and total SP number. The study in [118] aims to preserve the image topology for SP generation. A recent study [131] proposes a novel method to generate 2D superpixels and 3D supervoxels (SV) in an energy minimization framework utilizing graph cuts. It provides various controls on the SP structure and distribution; however, it suffers from computational complexity during the optimization stage. Another recent study [142] presents a detailed evaluation of various SV methods by extending the common SP methods on temporal volume. The details of how the conventional evaluation metrics should be extended on the spatio-temporal domain are also presented.

Previous literature about extracting superpixels is further detailed in the following section. The main challenges of the superpixel extraction addressed in this thesis can be named as follows. A successful method should preserve local structure by adapting to the local object and region boundaries. Secondly, undersegmentation of the regions

should be avoided for realizing an expressive image representation. Moreover, regular region identification is targeted with quasi-uniform SP regions. Finally, computational complexity should be kept at minimum. The first two challenges are related with the local information encapsulation that enforces adaptation of SP boundaries to the object boundaries. Uniform localization and compactness are required to form regular grid structure among graph models with unbiased neighbor relations. This property has an influence on the precision and accuracy of graph based solutions, especially in image segmentation problem. Computational efficiency is crucial for practical usability of the method.

Pixel based descriptors are widely used in object recognition tasks due to their accepted performance for dense image description [83]. However, the use of middle and higher level descriptors has recently gained attention. Such mid-level cues are claimed to be important for a better scene characterization. In this thesis, the low level image information is explored for extending towards middle level region descriptors. The advantage of the proposed mid-level description is that it does not require a fixed region size or shape to define the support area of the descriptor. Region shape is adaptive depending on the spatial image characteristics. Possible contributions from different levels of information are fused with the proposed hierarchical region adaptive superpixel based descriptor.

The applications of the superpixels on the image segmentation task has been previously addressed. One of the early works in [79], proposes using an initial oversegmentation on the image before computing the energy minimization procedure. Such a preprocessing step would overcome the drawback of the graph based energy optimization methods such that the required amount of time and memory would considerably reduce with the size of the graph. Similar methods have been previously proposed [91, 28] in order to merge regions automatically that are initially segmented by different oversegmentation techniques.

Image representation is of primary importance for various image understanding tasks. A good image representation should capture the distribution of image features faithfully and efficiently. The representation at low levels is a general and basic approach, but its support at the semantic level is poor. The paper in [139] investigates a multi layer neural network for image representation that is inspired by the human visual system. A different type of image representation proposes a transformation in the feature space [62]. The intention is to map the image in the feature space to the target domain in order to generate a robust representation under different visualization conditions. Similar metric learning methods are proposed [141] for obtaining robust transformations in the feature space.

1.3 Highlights and Contributions

This thesis covers a wide range of subjects in the fields of image processing, stereoscopy, image feature descriptors, classification and object recognition. In the first part of the thesis, the superpixel representation of an image is utilized for the image segmentation purposes. The results of the segmentation are used for converting the 2D monoscopic data on stereo for 3D visualization. Moreover, the extension of the technique on the stereo images is also presented. As an application of the stereo segmentation, disparity remapping is proposed as a post-processing step in the stereoscopic content generation pipeline. Finally, the superpixel primitives are used as a mid-level scene descriptor for the object recognition purposes. The contributions of the thesis are many-fold and are explained below in the order of presentation.

Superpixel extraction

The superpixel extraction method proposes contributions to the state-of-the-art in terms of both the computational efficiency and the segmentation performance. In the proposed technique, superpixels and supervoxels are defined depending on the color and spatial similarity. The boundary adaptation idea and the energy function selection are the two main contributions of the proposed method enabling efficient implementation and high segmentation accuracy. Experiments are conducted for different energy function combinations and two of them are selected for the comparisons against the state-of-the-art. The effects of utilizing different color spaces and distance metrics have been examined during the experiments. Utilization of geodesic distance in the energy function has shown improvements in segmentation performance. The proposed convexity constraint is explicitly justified through a graph based interactive segmentation application. According to the extensive comparative tests with the state-of-the-art, the proposed scheme is shown to yield a remarkable alternative in the current superpixel and supervoxel extraction methods with faster execution times and competitive segmentation performances. This has been also supported with the visual results where the generated superpixels and supervoxels are observed to show strong adaptation on the object boundaries.

Mono segmentation & 2D/3D conversion

The extracted superpixels are utilized in the user assisted object segmentation framework with an application on 2D/3D image conversion. The segmentation framework is established using the superpixel primitives. Graph-cut energy minimization technique is iteratively used for multiple object labeling purposes. The efficient iterative implementation of the geodesic distance metric proves to be useful for increasing final segmentation performance especially at the boundary regions of the object. Visual results are presented for a qualitative evaluation.

Stereo segmentation

In order to extend the proposed mono image segmentation on the stereo footage, no extra user assistance is required. The input seeds on the representative locations of just one of the stereo image pairs is used for the stereo segmentation. This way the user is saved from repeating the procedure on the second image. The information propagation is handled via efficient feature point based stereo matching. Hence, the necessity of the computationally demanding dense disparity estimation module is eliminated. The ground truth stereo database is tested for judging objective stereo segmentation performance. With the additional user strokes, the proposed method is shown to generate outstanding results compared to the state-of-the-art methods.

Disparity remapping

The proposed stereo segmentation technique is further used to propose a post-processing step for retargeting stereoscopic footage on different display sizes and resolutions. By the help of the proposed disparity remapping technique, novel disparity adjusted views are synthesized using the produced stereo object segments and background information for the images. To our best knowledge, utilization of segmented stereo objects for virtual depth adjustment purposes has not been addressed before. Subjective evaluations support the usage of such a disparity remapping operation regarding different aspects of visual preferences. It has been observed that the processed images are preferred more frequently for the problematic categories, which in fact, are the target applications for the proposed method.

Video segmentation

Using the segmentation of the initial frame of the video, succeeding frames are automatically segmented using a novel superpixel based feature descriptor. Object and background regions are learned using the proposed superpixel based region descriptors. Support vector machine is used to define the individual likelihood (confidence) of a superpixel to be assigned to the object or background region. Final region segmentation is performed using the graph-cut framework where sink and source energy links are determined by the object and background likelihoods, estimated by the learned region models.

Superpixel based mid-level descriptor for object recognition

Conventional object recognition pipeline is presented in a four step process. The first step, dense feature extraction, utilizes pixel based descriptors for image classification. The proposed hypothesis is that pixel based low-level descriptions are useful but can be further improved with the introduction of mid-level region information. A novel superpixel based region descriptor that encapsulates the mid-level information is proposed in order to explore and evaluate the initial hypothesis. Image regions are described by computing the directional mean differences between a central superpixel and its

various orders of neighborhood. The variance of the neighbors is further included for a better region description. The performance of the proposed descriptor is evaluated on the image classification task. Based on the experimental evaluations, increased average precision scores verify the initial hypothesis that mid-level cues enrich the image description and improve the performance of the low-level cues. Some qualitative results are supplied in order to give a better feeling of the matching performance of the proposed descriptor.

Geometry based region segmentation for spatial pooling
In an attempt to utilize region specific information, the third step, spatial pooling, in the object recognition pipeline is also investigated. The spatial similarities in images for the purpose of object level image classification are explored. This has been achieved by an improvement on the spatial pyramid by adapting spatial regions with respect to the underlying image characteristics. The method has been experimentally evaluated and the results have shown that the region adapted spatial segments improve the accuracy over the baseline. This also supports the intention to encapsulate spatial statistics using the proposed region based segmentation. Increase in the MAP (mean average precision) scores have shown that coherent spatial regions would consistently improve performance for alternative scenarios. Sample qualitative results of the proposed segmentation have also been supplied for illustration of the proposed region adaptation.

1.4 Outline

The outline of the thesis is organized as follows. Following the introduction in Chapter 1, the proposed superpixel and supervoxel extraction method is presented in Chapter 2. Detailed experiments on the segmentation dataset has been provided for performance evaluation.

Chapter 3 presents the user assisted segmentation method. The energy function for optimum image labelling, efficient geodesic distance calculation on the superpixel graph structure are explained in detail. Visual results for segmentation and the proposed 2D/3D conversion technique is provided. Underlying aspects of 3D visualization from the human visual system perspective is also addressed in this chapter.

Chapter 4 is devoted to the extension of the mono image segmentation method on the stereo and video footage. The sparse stereo matching and information propagation on the stereo image are investigated. Subjective tests for visual quality evaluation are conducted and discussed for the performance evaluation of the proposed disparity remapping technique.

Chapter 5 deals with the object recognition task and proposes two main contributions

on the conventional object recognition pipeline. The superpixel primitives are used as a mid-level dense region descriptor. The average precision tests are conducted for evaluating the performance of the proposed superpixel descriptor. Moreover, a geometry based region segmentation is proposed as an improved spatial pyramid in the object recognition pipeline. Similarly, detailed quantitative tests are done in order to verify the initial hypothesis.

Chapter 6 concludes the thesis with final remarks, discussions and a route for the future work.

CHAPTER 2

SUPERPIXEL EXTRACTION

2.1 Introduction

This chapter presents an efficient superpixel (SP) and supervoxel (SV) extraction method that aims improvements to the state-of-the-art in terms of both accuracy and computational complexity. Segmentation accuracy is improved through convexity constrained distance utilization, whereas computational efficiency is achieved by replacing complete region processing by a boundary adaptation technique. Starting from the uniformly distributed, rectangular (cubical) equal-sized superpixels (supervoxels), region boundaries are iteratively adapted towards object edges. Adaptation is performed by assigning the boundary pixels to the most similar neighboring SPs (SVs). At each iteration, SP (SV) regions are updated; hence, progressively converging to compact pixel groups. Detailed experimental comparisons against the state-of-the-art competing methods validate the performance of the proposed technique in terms of both accuracy and speed.

Pixel representation of an image is often redundant due to the spatial coherence within the image. In order to reduce this redundancy, a preprocessing stage is pioneered by Ren and Malik [103]. Their method clusters pixels into homogeneous image regions, called superpixels (SPs). Afterwards, utilization of SPs has become important in many image processing applications. Since the SP regions on the image possess similar color and texture characteristics, they provide an efficient representation. This property supports the assumption that pixels in the same SP belong to the same semantic object. Inspiring from this idea, all the pixels in a SP can be assigned to specific models representing motion, depth or segmentation structures. Such a representation can replace the use of pixels in various applications [122, 12]. By the utilization of SPs as an image representation, the inter-pixel details are captured and preserved in the image. Furthermore, the proposed SP structure is also crucial for graph-based approaches. When the graph nodes are constructed with SPs instead of pixels, graph complexity and computation time would substantially reduce.

SP extraction involves four main challenges: Firstly, a successful method should pre-

serve local structure by adapting to the local object and region boundaries. Secondly, undersegmentation of the regions should be avoided for realizing an expressive image representation. Thirdly, regular region identification is targeted with quasi-uniform SP regions. Finally, computational complexity should be kept at minimum. The first two challenges are related with the local information encapsulation that enforces adaptation of SP boundaries to the object boundaries. Uniform localization and compactness are required to form regular grid structure among graph models with unbiased neighbor relations. This property has an influence on the precision and accuracy of graph based solutions, especially in image segmentation problem. Computational efficiency is crucial for practical usability of the method.

In this chapter, a novel and efficient SP and SV extraction algorithm is presented addressing the four fundamental constraints mentioned above. Local structure is preserved with the selected energy function. Adaptation on the object boundary is satisfied by a color-based similarity measurement and the proposed distance metric takes care of the convexity constraint by penalizing irregularly shaped regions. Computational efficiency is achieved by processing only the pixels at the region boundaries. Following the related work in Section 2.2 for SP extraction, details of the proposed algorithm are presented in Section 2.3. The extension of method on the temporal volume is also explained in Section 2.3. Section 2.4 is devoted to experimental results and the final Section 2.5 concludes the chapter with final remarks and restatement of the contributions. In the rest of the thesis, the word "SP" is used for explaining the algorithmic details for extraction of both SP and SV on the spatio-temporal volume.

2.2 Related Literature

The previous work on SP and SV extraction dates back to less than a decade. We explore the related work in two categories: Graph based and gradient based methods.

Graph Based In graph based approaches, SP extraction is achieved by partitioning the graph where nodes correspond to individual pixels and edge weights are assigned according to a cost function relating inter pixel similarities. In [76], the graph is partitioned recursively, as in Normalized Cuts segmentation [115], in order to minimize a global cost function based on color and texture cues until desired number of SPs is achieved. This approach satisfies the compactness constraint required for SPs in order to provide efficient graph representation. However, it suffers from computational complexity. In [47], SP extraction is improved in terms of complexity by grouping nodes of the graph via greedy decisions through pairwise region comparisons on edge measures of minimum spanning trees. This method, on the other hand, does not enforce a control on region compactness and number of SPs. In [88], a lattice structure is enforced by finding horizontal and vertical seams that cut the image optimally via

graph cuts. The seams determine the SP boundaries based on region compactness and total SP number. The study in [118] aims to preserve the image topology for SP generation. A recent paper [131] proposes a novel method to generate 2D SPs and 3D supervoxels (SV) in an energy minimization framework utilizing graph cuts. It provides various controls on the SP structure and distribution; however, it suffers from computational complexity during the optimization stage. Another recent study [142] presents a detailed evaluation of various SV methods by extending the common SP methods on temporal volume. The details of how the conventional evaluation metrics should be extended on the spatio-temporal domain are also presented.

Gradient Based On the other hand, gradient-based approaches start from initial seeds of rough SPs. Pixel groupings are refined iteratively, depending on the local similarities. Mean-Shift [31], which is one of the well known methods in image segmentation, is adapted for SP extraction by the use of recursive smoothing kernel over pixel feature space. The main weakness of this method is that it does not have a control on the SP properties, such as compactness, distribution and total region number. In [134], an image is considered as a topographic structure and intensity gradient vectors are utilized to form pixel groups. This approach also lacks control on SP properties. TurboPixels concept [76] introduces geometric-flow over initial seeds which are considered as the starting points of the SPs. Level set method is exploited to update and refine SPs based on local image gradients. This approach enables regular distribution of compact SPs with less complexity compared to graph-based approaches. In [146], geodesic distance [19] is exploited to iteratively group neighboring pixels starting from the initial seeds as proposed in TurboPixels [76]. Utilization of geodesic distance enables higher structure sensitivity compared to geometric-flow with almost similar complexity. Initial seed placement in [76, 146] is refined in [28] by rectangular shaped initial SPs. Instead of geometric-flow, boundary pixels are re-assigned to SPs iteratively based on color similarity and spatial distance. At each iteration, SP mean intensity locations and color models are updated and hence enable compact and almost regularly distributed pixels groups. This approach refines SPs through boundary pixels which also significantly decreases the computational complexity. In [8], a similar method is proposed, where all pixels are updated during the refinement rather than only boundary pixels. A recent method [128] proposes a similar boundary update idea for region segmentation, where SPs are not constrained to be convex and have regular distribution. A top-to-bottom partitioning is proposed on the initial large rectangular SPs and an iterative process is exploited to refine SPs, based on color similarity and SP histogram. Besides, no temporal extension capability is presented in this paper.

A broad look at the previous literature has been useful for defining the priorities towards a successful SP extraction scheme. These challenges can be summarized as: 1) Adaptation on the object boundary; 2) Efficient region representation; 3) Quasi-uniform distribution on the image; 4) Fast execution capability. For this purpose,

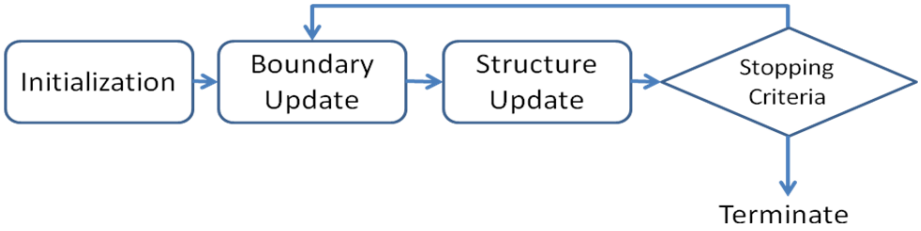


Figure 2.1: Algorithmic flow of the proposed SP generation algorithm

the iterative boundary refinement approach in [28, 121] is improved by constructing a general framework that utilizes color and locational similarity for pixel label assignment. Different parametric evaluations of Euclidean and geodesic distances are achieved in order to create structure sensitive SPs. Hence, through the utilization of alternative energy metrics, a trade-off between compactness and edge adaptation, as well as computational complexity and segmentation accuracy is accomplished.

2.3 Proposed Method

The proposed algorithmic flow of the method can be explored in four main steps: 1) Initialization of the SPs; 2) SP boundary update; 3) SP structure update; 4) Termination. These steps are illustrated in Figure 5.1.

Superpixel Initialization

In the first step, image is divided into equal sized regions according to the desired number of SPs. Each region initially has a rectangular shape and the centers are equally spaced among the image in the region centroids. In the prior methods, regular placement of SPs is a common technique where center pixels are considered as the seeds of pixel groups. Starting from these seeds, SPs are enlarged and the boundaries are constructed. However, the proposed technique approaches the problem from a different perspective. Instead of enlarging from the seed locations, SPs are refined through boundary pixels based on specific energy cost functions. The refinement is achieved iteratively through boundary and structure update steps.

Boundary update

In the boundary update step, a greedy search is conducted on the boundary pixels. During the boundary adaptation, the cost function relating similarity of the pixels to the corresponding SP candidates is minimized. This approach assures that the SPs are composed of connected pixels without any sub-detachment. Computational efficiency is realized by performing a search between boundary pixels and neighboring SP candidates. Label assignment of each boundary pixel is conducted in an eight neighborhood search. Pixel p is assigned to SP Q_i according to the following dissimilarity

cost between the pixel and the neighboring SPs.

$$L(p) = \underset{Q_i}{\operatorname{argmin}} \left(E(p, Q_i) \right), \;\; Q_i \in N_p, \;\; i = 1 : N \tag{2.1}$$

where $L(p)$ is the assigned SP label of the pixel p; $E(p, Q_i)$ is the dissimilarity energy between the corresponding pixel p and SP Q_i. N is the number of neighbor SPs surrounding the boundary pixel p and N_p is the label of these neighboring SPs. Therefore, starting from the initial SP distribution, the boundary pixels are reassigned to the most similar neighboring SPs. When all the boundary pixels are visited, SP centers and mean color values are updated with the current region labels.

Structure Update

During the structure update, the SP model (i.e. mean color values and SP centers) is recalculated based on the removed or merged boundary pixels. This update provides pixel groups to adapt changes along the boundaries and converge to compact SP models in terms of pixel similarity. The boundary and structure update steps are iterated several times until a stopping criteria is met.

Termination

Termination criteria can be set as a fixed number of iteration or it can be computed by the ratio of updated boundary pixels over the unchanged ones.

An example for the proposed evolution of SP boundaries at different stages of the iteration is presented in Figure 2.2. SP boundaries are represented with blue lines, while yellow pixels denote the SP center locations. The initial distribution of the rectangle shaped SPs with a uniform spacing is given in Figure 2.2.a. A greedy search is conducted on the region boundary among the neighboring SPs with respect to the update rule in (5.16). An intermediate SP distribution is shown in Figure 2.2.b. It is observed that the SP boundaries, as well as center locations, show powerful adaptation to local edges without losing their connectedness. At the final step in Figure 2.2.c, the iterative procedure is terminated due to small number of updated boundary pixels, providing edge aware convex SPs with quasi-uniform distribution.

2.3.1 Why Convexity Constrained SPs?

Convexity constrain has been a major criteria for the proposed SP extraction method. The main motivation behind that is to create regular oversegment grids over the entire image. This aim is morphologically meaningful, since objects usually tend to have regular boundaries. Moreover, such a constraint could also be useful for graph based implementations, where individual SPs are assigned as graph nodes.

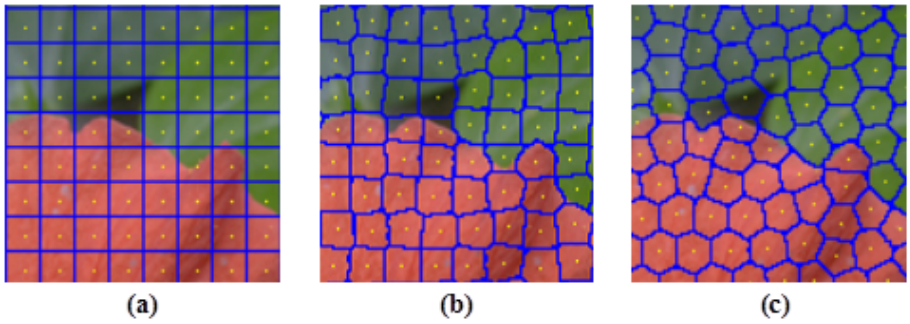


Figure 2.2: Boundary and SP center update at different iterations; a) 0 b) 4 and c) 10

In order to support the motivation towards the proposed convexity constrain on SP extraction, we have incorporated an interactive image segmentation framework with and without convexity bias in the SP structure. Our previous work on interactive object segmentation [122] provides details about the energy function assignment for a binary image classification problem. Human assisted input strokes on representative locations of the image are used to determine the binary segmentation of the image. Graphcut [22] method is utilized for the solution of this combinatorial optimization problem. For the purposes of a true comparison between convexity constraint on the SP structure, the runtime performance of the graphcut optimization method for the image segmentation framework is tested against two different SP extraction methods. In the first scenario, SPs are generated by using only color similarity; hence, no convexity constraint is induced. In the second scenario, a distance function has been incorporated in the SP energy formulation. The details of the energy formulations are explained later in Section 2.3.2. Computation time comparisons for different sized SPs are presented in linear and logarithmic scales in Figure 2.3 for better interpretation. Depending on the naive optimization time results, it can be observed that the computational times differ depending on the structure of the generated SPs. Moreover, as the SP size increases, the difference in computation time also increases. This is an expected result, since the irregularity in the region boundary increases with enlarged region size. In this manner, the number of possible combinations for the graphcut optimization to be tested before

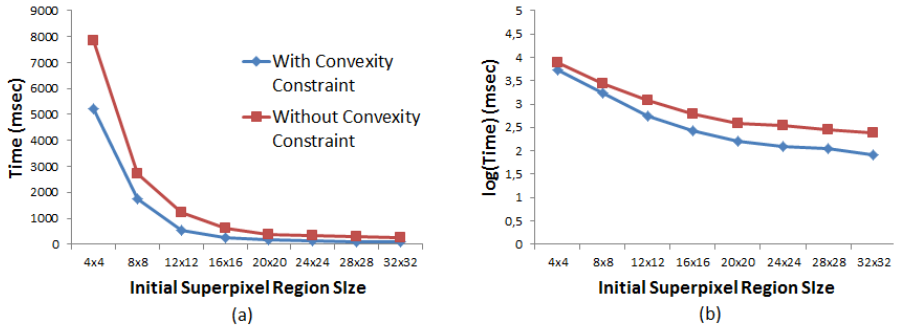


Figure 2.3: Graph cut time computations in (a) linear and (b) logarithmic scale

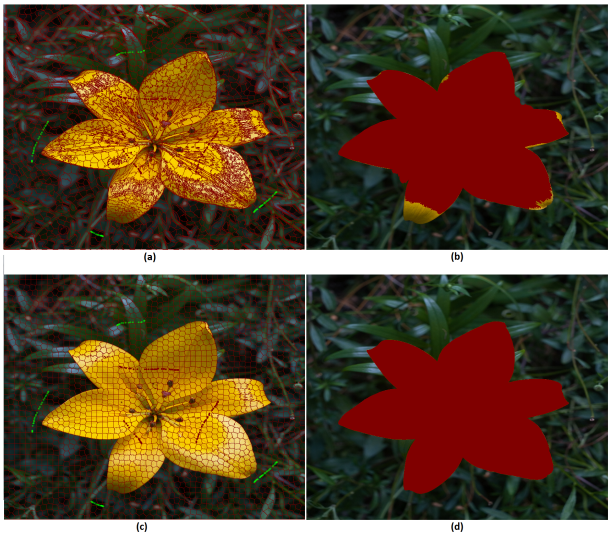


Figure 2.4: (a)(c) : Interactive segmentation inputs on convex and non-convex super pixels, (b)(d) Interactive segmentation results on convex and non-convex super pixels

reaching the optimum energy assignment increases with the increased SP size.

In addition to the computational advantages, accuracy of the segmentation with the two different SP extraction methods (with and without convexity bias) have been evaluated in the same interactive segmentation framework. The SPs on Figure 2.4.a are computed by using only color constraint, and Figure 2.4.c shows the case when distance constraint is induced as well as color. As seen on the figure, introduction of the proposed geometrical constraint generates a very structured and evenly shaped SPs. When the interactive segmentation framework in [122] is used to compute optimum energy labeling for binary segmentation, different results are obtained using the same input scribbles for foreground (red) and background (green). Figure 2.4.b and Figure 2.4.d. show the resulting segmentation when SPs with(&out) convexity bias are used as graph nodes in the energy optimization framework.

The mentioned advantages of utilizing structural SPs, in terms of computational time and segmentation performance, has motivated our study towards developing a method that utilizes a regional structure as a prior in the energy function.

2.3.2 Proposed Energy Formulation

The formulation defined in (5.16) is used on the boundary update of the SP structure. At this step, all the boundary pixels of the SPs are visited and the cost of assigning each boundary pixel to the neighboring SPs is computed. Depending on the computed cost, the boundary pixel is assigned to the region that provides the minimum cost. The cost function used in this study is composed of two main energy terms. The first

term relates the color similarity of the boundary pixel to its neighboring SPs, whereas the second term defines the spatial distance of the pixel from the SP centers. The proposed two term energy function is defined in order to create color wise homogeneous and shape wise structural SP regions. The selection of cost functions and the underlying reasoning is detailed in this section. Varying the cost function parameters yields different segmentation accuracy and computational complexity; hence, it serves as a trade off parameter for the target application.

2.3.2.1 Color Similarity Cost

During the SP boundary update, color similarity of a boundary pixel with its neighboring SPs is of great importance for an accurate label assignment. Stemming from the idea that a SP is composed of homogeneously colored pixels, the boundary pixels are supposed to be assigned to the SPs with maximum similarity. The cost function used in this study for calculating the color similarity is given below.

$$C(p,Q) = \sum_{i=1}^{3} \left(p^i - Q^i\right)^2 \tag{2.2}$$

The color similarity cost $C(p,Q)$ in (2.2) is computed via fusion of 3 channel color information of the boundary pixel p and its neighbor SP Q. p^i represents the mean color of the i^th channel of pixel p and similarly Q^i is the mean color of the i^th channel of SP Q. The experiments are conducted on two different color spaces. RGB is a common format for various image processing applications, whereas LAB has a superior representation performance due to its perceptual uniformity [51]. The cost function given in (2.2) enforces the boundary pixels to be assigned to the color wise most similar SPs and hence strong adaptation at the object boundaries is satisfied.

2.3.2.2 Distance Cost

As mentioned in the previous section, one of the main challenges of SP extraction is to provide quasi-uniform distribution of convex SPs. Utilizing SPs, instead of pixels, provides computational efficiency, however, irregular shaped SPs can degrade performance due to weak neighboring relations [8, 28]. Structural uniformity cannot be met by the use of sole color similarity; therefore, some geometric constraints should be provided. The proposed distance term in the energy function is useful for imposing convexity constraint on the generated SPs. It is realized by penalizing the distant pixels, so that they are re assigned to a different closer SP.

During the selection of the distance metric, Euclidean and geodesic distances [19] are considered. They are selected due to their individual advantages on the final

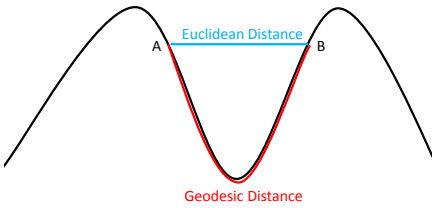


Figure 2.5: Abstract representation of geodesic vs. Euclidean distances

performance. By the selection of the distance function, different trade-off for the convexity constraint has been observed. Euclidean distance $D(p, Q^c)_E$ is computed by comparing the spatial coordinates of the pixel p, (p_x, p_y), and the SP centroid Q^c, (Q^c_x, Q^c_y), as shown below.

$$D(p, Q^c)_E = \sqrt{(p_x - Q^c_x)^2 + (p_y - Q^c_y)^2} \tag{2.3}$$

Geodesic distance has been previously used for image segmentation purposes [100]. The motivation behind utilization of such a metric in the distance computation is due to its region encapsulation property. Euclidean distance can a spatial measure whereas, geodesic distance takes into account the path of the region that brings a pixel to the center of the neighboring SP. Figure 2.5 illustrates the distinction between two distance metrics as an abstract representation between two points.

Geodesic distance, $D(p, Q^c)_G$, between the boundary pixel p and the SP centroid, Q^c is defined as the sum of the cost of the shortest path from p to Q^c [34].

$$D(p, Q^c)_G = min_{P=p_1,p_2,..,p_n} l(P) \tag{2.4}$$

Suppose $P = \{p_1, p_2, ...p_n = Q^c\}$ is a path between the pixels p_1 and $p_n = Q^c$, where p_i and p_{i+1} are connected neighbors. The path length $l(P)$, as defined in (5.20), is the sum of individual neighbor distances $d_N(p_i, p_{i+1})$ (5.21) between adjacent points in the path.

$$l(P) = \sum_{i=1}^{n-1} d_N(p_i, p_{i+1}) \tag{2.5}$$

For the computation of adjacent pixel distance d_N, three color channels (RGB or LAB) can be utilized using the formulations in (5.21). Since no significant performance difference has been observed, the formulation in (5.21) has been selected with k=1 in the final implementation due to its computational efficiency.

$$d_N(p,q) = \sum_{i=1}^{3} (|p_i - q_i|)^k \quad k = 1, 2 \tag{2.6}$$

2.3.2.3 Combining Color and Distance Costs

The individual color and distance terms for SP boundary update are defined previously. This section provides an in depth analysis of color and distance term combinations. Utilization of a combined color and distance formulation is shown to be useful based on the objective results presented in the next section. The proposed dissimilarity energy cost function, defined in (5.16), combines both color and distance terms as shown below.

$$E(p,Q) = \lambda\ C(p,Q) + (1-\lambda)D(p,Q^c) \tag{2.7}$$

where the parameter λ determines the weighting between these two terms.

Figure 2.6 illustrates the update procedure of a boundary pixel "x" on the junction of four different SP neighborhood according to two distance criteria. The neighboring SP centroids are marked with blue dots and the blue line connecting the SP centroid and boundary pixel indicate the shortest path in terms of the distance metric. Computation of the shortest path from the boundary pixel p to the SP centroid Q^c is performed via the formulation provided in [40]. At each iteration, the shortest paths from the neighboring boundary pixels to the SP centroid are computed. Since the termination criteria for path computation is at the boundary, estimation of the shortest paths over the whole image is avoided.

2.3.3 Energy Function for Supervoxel Generation

The extension of the proposed idea towards the spatio-temporal space for SV generation is a necessary and intuitive step for a video representation framework. Similar to SPs in images, SVs has the ability to represent videos in a coherent structure where 3D segmentation of multiple frames become possible. Voxel based segmentation methods can be especially valuable for volumetric region processing. The optimization rule given in (5.16) is revisited for the SV region estimation. In this case, boundary pixels do not define 2D, but 3D volumetric regions. At each iteration of the algorithm, pixels at the volume boundaries are visited and are assigned to the neighboring voxel with the maximum similarity. Section 2.3.2 explains the energy function and the selection procedure. SV generation is realized with the same technique where at this time graph nodes are defined as the pixels of the video frame.

Figure 2.7 shows the initial volumetric cubic region boundaries at the first iteration. Similar to the SP case, SVs adapt to the 3D object boundaries at each iteration. An important parameter to tune for the voxel estimation procedure is the number of the

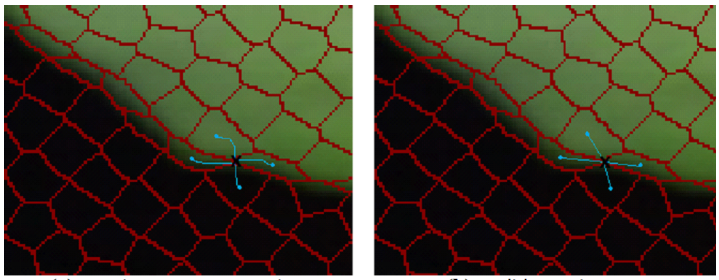


Figure 2.6: Illustration of the shortest (geodesic) path between the SP centroid and the boundary pixel

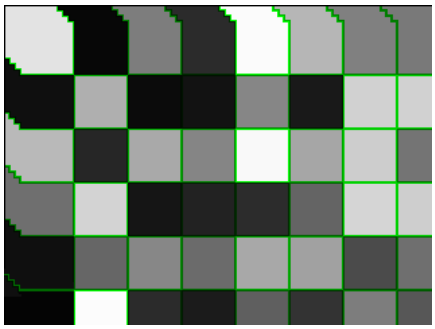


Figure 2.7: Illustration of Supervoxel Boundary

frames to include for each cubic voxel region. Figure 2.7 shows a voxel composed of 5 frames with 20×20 pixels in each frame. Three different temporal dimensions are tested for evaluation; 3,5 and 7. The number of temporal dimension is set as a parameter in the implementation and can be adjusted to any value depending on the application.

2.4 Experimental Results

This section presents quantitative and qualitative results regarding the proposed two SP extraction methods (using Euclidean and geodesic distance) in comparison with the state-of-the-art. The known methods in the literature, Graph-based [47], TurboPixels [76], Structure Sensitive Geo [146] and SLIC [8], are evaluated in terms of accuracy and computation time. The accuracy of the extracted SPs is measured in terms of undersegmentation error, E_{UnSeg} and boundary-recall statistics. Undersegmentation error is calculated by measuring the "bleeding" of the segment boundaries with respect to the ground truth (human) segmentation. Bleeding is measured by the following relation in (2.8)

$$E_{UnSeg} = \frac{1}{N}\left(\sum_{l=1}^{L}\left(\sum_{[S_j|S_j\cap G_l>B]} Area(S_j)\right) - N\right) \tag{2.8}$$

where N corresponds to the number of pixels, L is the number of ground truth segments G_l, and S_j is the extracted SP. In (2.8), pixel area of a SP intersecting with the G_l is computed. B is selected to be equal to 5% throughout the experiments in order to compensate for small errors in ground truth segmentation data.

E_{UnSeg} measures how well the extracted SPs fit the ground truth segment boundaries. The experiments are conducted on the Berkeley segmentation database [103] with their manual segmentation results over 300 different images for a resolution of 481x321. All the presented undersegmentation error values denote the average error over the whole dataset. The second error metric is the boundary recall and it is used to measure the percentage of overlap between the ground truth boundary pixels and the generated SP boundaries within one or two pixel neighborhood. Although this metric is itself inconclusive, it is widely used and clearly gives an idea about the boundary precision of the SP extraction algorithm. Different number of SPs are tested for performance evaluation to observe the performance of the algorithm.

The measurements are performed on 3.06GHz Intel Core i7 CPU with a 6 GB RAM. In the first part of this section, comparative experiments are conducted on different energy functions for the proposed algorithm to optimize accuracy and computational load with respect to various number of SPs and image resolutions. Once the best performances are determined, further comparisons against the state of the art techniques are given in the second part. The source code of the proposed implementation will be made available in the authors' web page [1].

2.4.1 Parameter Optimization

The proposed method has been tested for various energy cost function combinations. The weighting of color and distance cost given in (5.17) has a major impact on SP boundary adaptation. A comparative visual evaluation for different values of λ in (5.17) is presented in Figure 2.8. As the weight of distance term increases, SPs converge to a quasi-uniform distribution with increased convexity, which is desired for graph based approaches. If this ratio is further increased as in Figure 2.8.d, the resulting distribution becomes almost uniform and color homogeneity within SPs is violated. According to the visual interpretation of Figure 2.8, equal color and distance weights are utilized throughout the experiments. However, a different weight selection might also be preferred depending on the application and content.

In Table 2.1, the evaluation of execution times vs measured bleeding ratios are pre-

[1] http://emrahtasli.com/SPExtraction.html

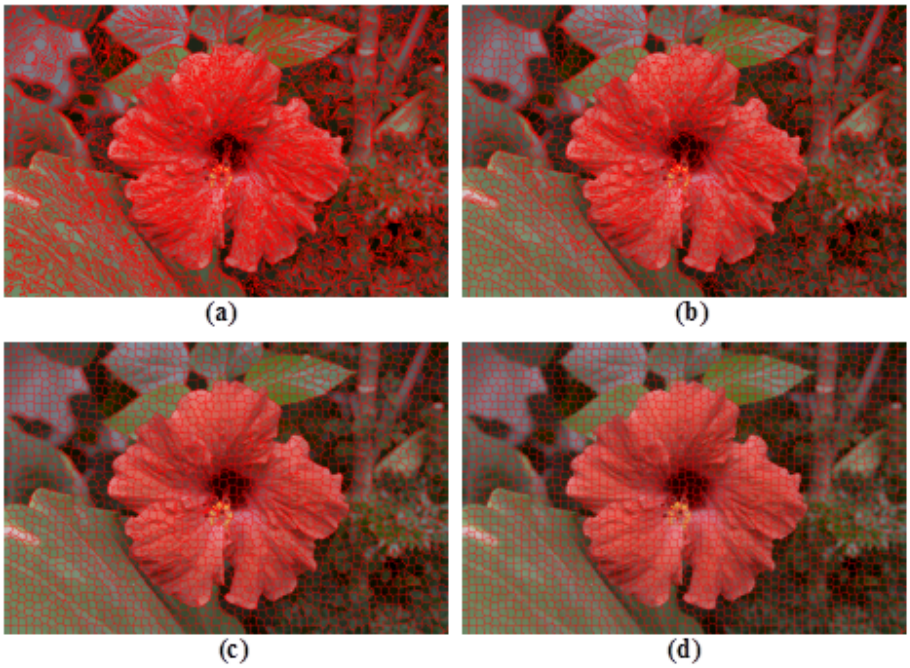


Figure 2.8: SP boundaries under different convexity weights (a) λ=0.9, (b) λ=0.6, (c) λ=0.4 (d) λ=0.1

sented for eight (4×2) different energy function combinations. 1) Only color; 2) only geodesic distance; 3) color and Euclidean distance combination, 4) color and geodesic distance combination. All these combinations are tested both in RGB and LAB color spaces. The measurements are obtained for three different number of SPs as 500, 1000 and 2000 for the sake of completeness. Moreover, the average measurements are illustrated in Figure 2.9 to visualize the performance of different combinations. According to the obtained results, the execution times increase by LAB utilization due to additional overload for RGB to LAB conversion. The fastest execution time is obtained when only the geodesic distance is selected as the energy function; it is followed by the combination of color and Euclidean distance as the energy function. The fastest two selections are further compared with respect to increasing image resolution over original sizes of (481x321) by fixing the number of SPs to 1000. According to the results given in Figure 2.10, geodesic distance enables faster computation for lower resolutions; as the image size increases color and Euclidean combination yields faster computation. The main reason of such a result is that, increasing the image resolution creates larger SPs; hence, more computation time is required during the geodesic distance calculation to reach SP boundary pixels.

Examining Table 2.1, it is interesting to observe that combination of color and Euclidean distance requires less computation compared to using only color term in the energy function. This is a consequence of quasi-uniform distribution of convex SPs provided by the distance term, which minimize the number of boundary pixels. Thus,

Table2.1: Computation Time and Bleeding Ratio Comparison for Different Energy Functions

	Number of Superpixel			
Computation time/ Bleeding (msec/ % ratio)	500	1000	2000	Averages
Color(RGB)	212 / 25.3	262 / 18.7	331 / 13.3	268 / 19.1
Color (LAB)	263 / 22.1	309 / 16.3	372 / 11.6	315 / 16.7
Geo (RGB)	200 / 24.0	191 / 17.4	202 / 12.4	198 / 17.9
Geo (LAB)	266 / 23.6	256 / 17.0	267 / 12.1	263 / 17.6
Color(RGB)+Euclidean	152 / 24.2	209 / 16.5	254 / 12.2	205 / 17.6
Color (RGB)+Geo(RGB)	252 / 22.2	278 / 16.1	318 / 11.8	283 / 16.7
Color (LAB)+Euclidean	209 / 20.8	258 / 15.0	300 / 11.2	256 / 15.7
Color (LAB)+Geo(LAB)	317 / 20.5	345 / 15.0	390 / 10.7	351 / 15.4

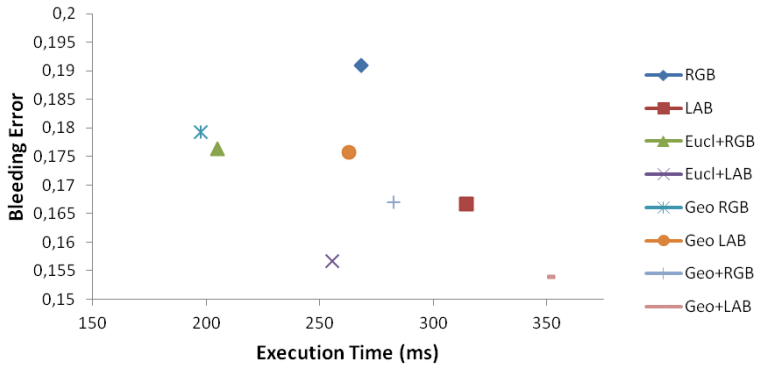


Figure 2.9: Bleeding error vs. execution time for the selected eight energy functions

less number of greedy search among neighboring SPs is performed and this yields faster computation.

LAB color space utilization increases segmentation accuracy for all combinations compared to *RGB*. This is also an expected result, since *LAB* color space is selected due to its perceptual uniformity. The best precision is provided by color and geodesic distance combination in *LAB* domain, which is followed by color and Euclidean distance combination. The same accuracy order is also valid for *RGB* color space as well. This result shows that the bleeding error decreases by the additional distance term, which also enables quasi-uniform distribution. Among the presented different alternatives for the selection of energy function, two of these results turn out to be the optimum solution in terms of execution speed and segmentation accuracy. Hence, these two energy functions, $Euc + RGB$ and $Geo + LAB$, are selected for evaluating the performance of the proposed methodology against the state-of-the-art techniques in the following

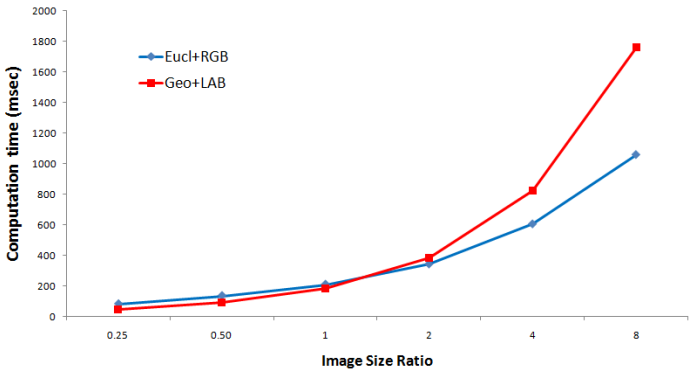


Figure 2.10: Computation time for the selected two energy functions for different image sizes

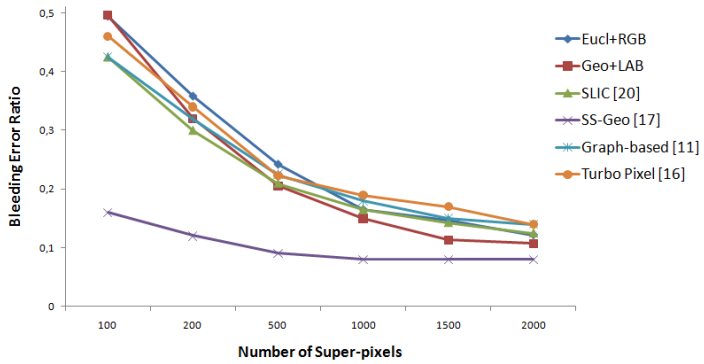


Figure 2.11: Bleeding error comparison for different number of SPs

section 2.4.2. $Geo+LAB$ is selected due to its very low bleeding error and $Euc+RGB$ is selected due to its fast execution time with acceptable bleeding error performance.

2.4.2 Comparison against state-of-the-art Techniques

The quantitative performance evaluation of the proposed method against the state-of-the-art is achieved in terms of the computation time, E_{UnSeg} (2.8) and boundary-recall metrics. Moreover, the resulting boundaries are also presented for a visual evaluation. At this point, it is important to note that segmentation accuracy results of the state-of-the-art techniques are obtained from the corresponding references. The undersegmentation error ratio of the generated SPs provided by Graph-based [47], TurboPixels [76], Structure Sensitive Geo [146] and SLIC [8] and the proposed ($Eucl+RGB$ and $Geo+LAB$) two methods are presented in Figure 2.11. It is observed that Structure Sensitive Geo algorithm [146] has superior performance in terms of undersegmentation error, which is followed by our proposed method, especially when the number of SPs is sufficiently high (≥ 500). SLIC is observed to perform better than $Geo+LAB$ for small number of pixels.

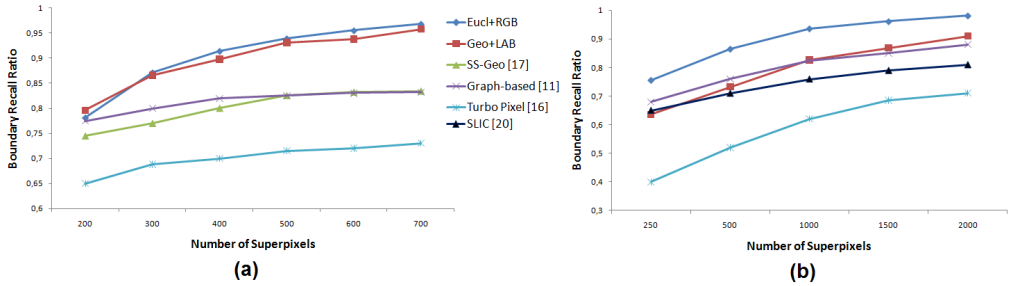


Figure 2.12: Boundary-recall ratio for various number of SPs in (a) 2-pixel neighborhood, (b) 1-pixel neighborhood

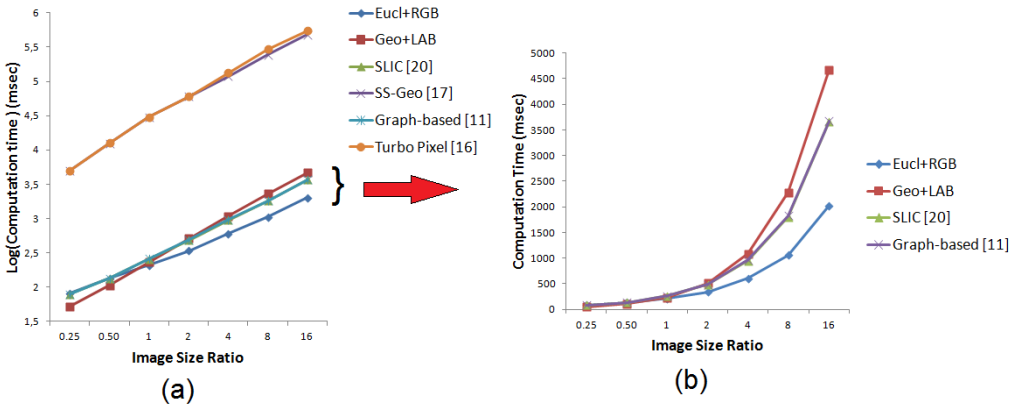


Figure 2.13: Computation time comparison for different image scales a) Logarithmic scale b) Linear Scale

Boundary-recall ratio measures the amount of match between the super pixel boundaries and the ground truth segmentation boundaries. This metric is prone to errors, since it is quite difficult to distinguish the actual boundaries in pixel precision. Hence, two versions of the metric are tested for measuring the ratio of boundary fit. The first version of the metric checks, whether the indicated SP boundary is within two pixel neighborhood of the actual boundary. Similarly one pixel neighborhood test is also conducted. According to the results presented in Figure 2.12.a and Figure 2.12.b the proposed two methods (Eucledian+RGB and Geodesic+LAB) outperforms state-of-the-art techniques.

Final quantitative comparison is conducted in terms of the computational times of the corresponding methods. In this case, the number of SPs is kept constant at 1000 and the images are scaled up and down using bicubic interpolation for different ratios of the original size. Seven different scales of the original image (1/4, 1/2, 1, 2, 4, 8, 16) are used for measuring average running time of the methods. According to the results presented in logarithmic scale in Figure 2.13.a, TurboPixels [76] and Structure

Sensitive Geo [146] require orders of magnitude longer execution times compared to SLIC [8], Graph based [47] and the proposed approach. The results are also shown in logarithmic scales in order to visualize all methods on the same figure. The previous paper [8] presents the efficiency of SLIC compared to the Graph based approach; hence, a final detailed time analysis is conducted in Figure 2.13.b by comparing the proposed methods to SLIC and Graph based method separately.

Apart from the quantitative comparisons, the visual results of the proposed two methods are presented in Figure 2.14 and Figure 2.15 for visual interpretation. It is observed from the visual results that the extracted SP boundaries for the proposed method perform well at the object boundaries. Moreover, region homogeneity and convexity constraint is satisfied with the proposed energy based region assignment.

2.4.3 Supervoxel Extension

As an extension of the SP framework on the temporal dimension, SV extraction has also been implemented and tested. To our best knowledge, there are only limited number of methods presented in the literature for SV extraction. The study in [131] presents an energy based SP extraction and directs to a possible extension on the temporal dimension for a SV generation. A more recent paper in [142] details an evaluation of SV extraction methods. Although this study [142] does not offer a new method to the literature, it provides a framework where previous methods have been analysed in detail. The results in that paper defines the quantitative framework for SV performance evaluation, and it has been also used in our study for evaluating the performance of the proposed SV extension.

The undersegmentation error defined in (2.8) is directly extended to 3D below.

$$E3D_{UnSeg} = \frac{1}{N}\left(\sum_{l=1}^{L}\left(\sum_{[S_j|S_j\cap G_l>B]} Area(S_j)\right) - N\right) \tag{2.9}$$

where N is the number of pixels in the volume, L is the number of ground truth voxels G_l, and V_j is the extracted SV. (2.9) computes the voxel volume of a SV intersecting with the G_l. B is set to 5% throughout this study in order to compensate for small errors in ground truth segmentation data.

2.4.3.1 Evaluation Metric

The segmentation accuracy metric as defined in [142] is used to test the accuracy of the produced voxels with respect to the ground truth volumes. This metric (2.10) measures the fraction of the correctly segmented pixels.

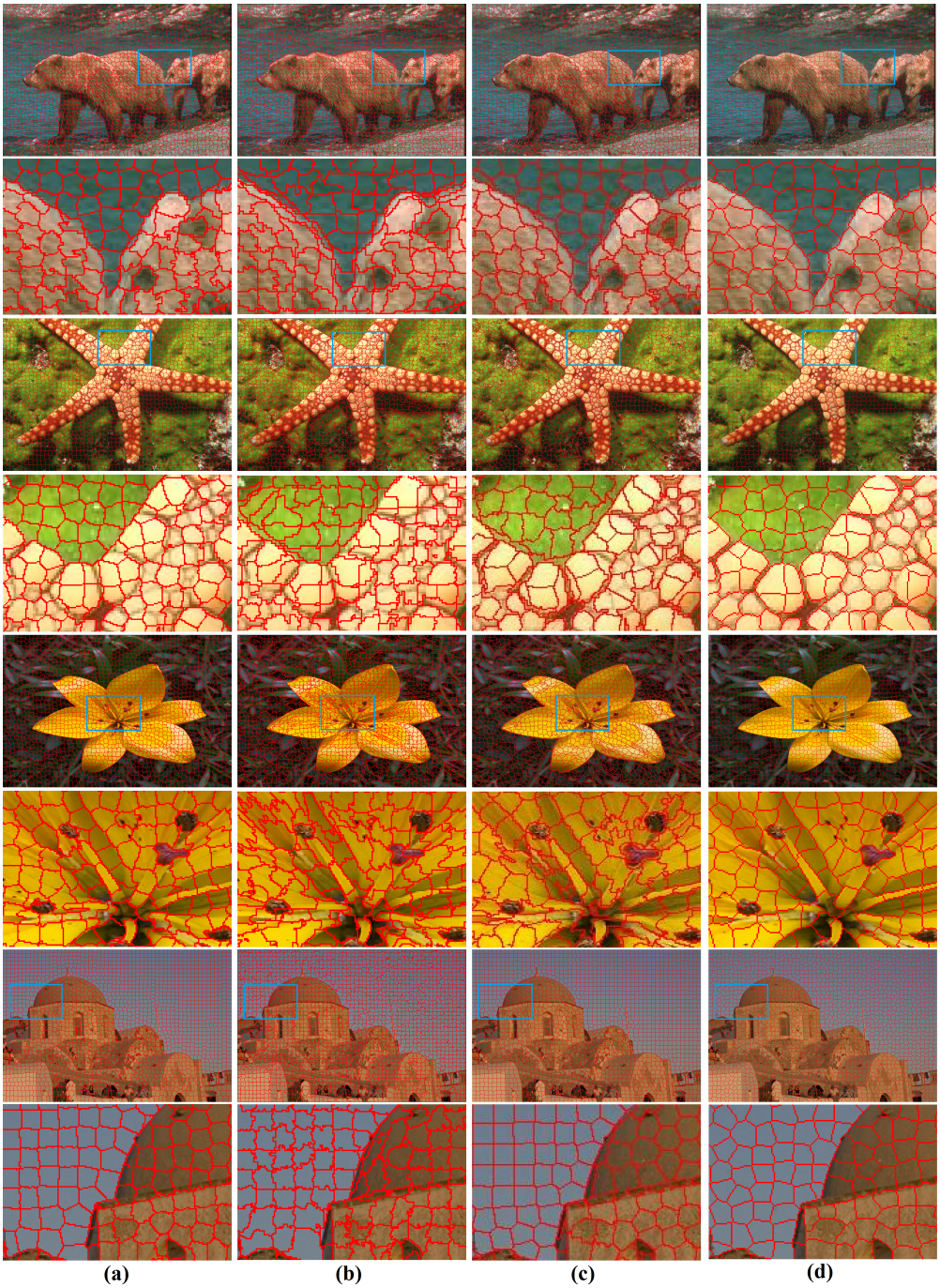


Figure 2.14: SP boundaries (a) $Eucl + RGB$, (b) $Geo + LAB$, (c) SLIC [8], (d) Turbo Pixel [76]

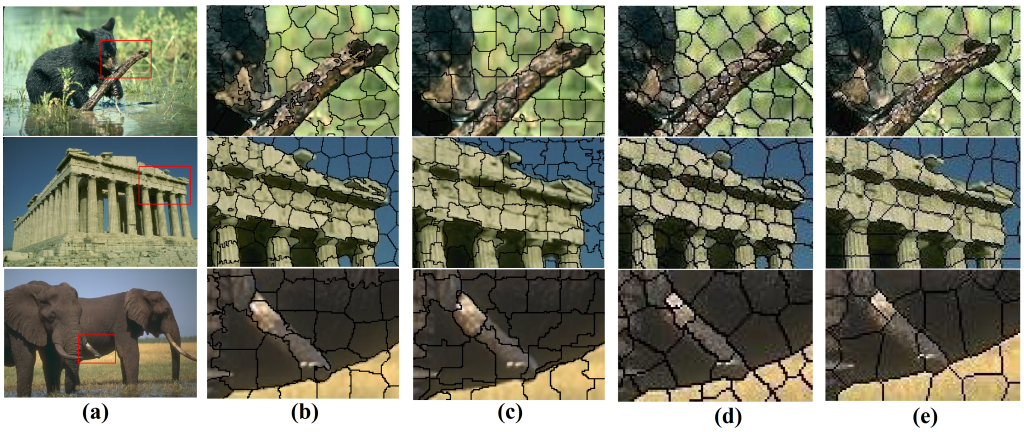


Figure 2.15: (a) Original image, SP boundaries of (b) $Eucl + RGB$, (c)$Geo + LAB$, (d)SS-Geo [146], and (e)Turbo Pixel [76]

$$Acc3D(g_i) = \frac{\sum_{j=1}^{k} Vol(\bar{v_j} \cap g_i)}{Vol(g_i)} \tag{2.10}$$

where g_i represents the ground truth segment and the label of an individual SV, v_j, is defined by the area that it mostly overlaps. The correctly labeled SVs $\bar{v_j}$ define the percentage of overlap with respect to the ground truth volume.

Boundary-recall ratio of the 3D volume is also computed for evaluating the performance of the proposed method. The quantitative experiments are conducted on different number of SP numbers for individual frames and three sets of temporal dimensions (3,5,7 frames) are tested in our experiments. The size of temporal dimension is important on the final performance of the SV. The higher the temporal dimension the harder it gets to isolate a homogeneous voxel from the video. It is mainly because it might get harder to follow a moving object between the frames and hence the voxels might get irregularly shaped; this is penalized by our system. Another possible scenario is that the voxel might just not be available during all the selected frames. It might get occluded or deformed in shape so that it no longer exists. The quantitative results also in parallel with these expectations.

2.4.3.2 Evaluation Dataset

SegTrack dataset [126] is used during the quantitative experiments. It provides a set of videos with manually labeled foreground segments. They are selected from different difficulty levels with respect to color, motion and shape of the foreground segments. There are six videos in the dataset with an average 41 frames per video.

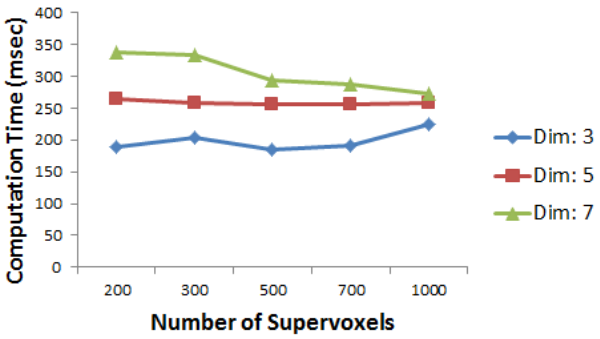


Figure 2.16: SV generation computation time. 3,5,7 dimensional SV for 8 iterations

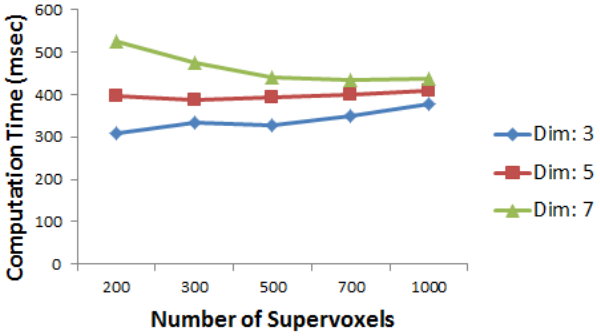


Figure 2.17: SV generation computation time. 3,5,7 dimensional SV for 14 iterations

2.4.3.3 Comparisons

The proposed algorithm is compared against the methods that are implemented in the evaluation paper [142]. The selected methods are namely, SWA [114], GB [47], GBH [57] and mean-shift [31]. The different methods are compared in terms of segmentation accuracy, boundary recall values. Computational time results are also supplied for different SV sizes, temporal dimensions and iteration numbers for the proposed method.

Figure 2.16 and Figure 2.17 show the computational results of the proposed algorithm with SP sizes of 200, 300, 500, 700 and 1000. Temporal dimension of the SV is selected as 3,5 and 7. Figure 2.18 presents the segmentation accuracy of the different methods for varying SV numbers. The proposed method is observed to perform better with respect to this metric. Figure 2.19 shows the boundary-recall measurements under varying SV sizes. It has been observed from these results that the proposed method performs better as the number of SV is increased. The quantitative results are the obtained by averaging on the videos of the dataset.

In addition to the quantitative results, some visuals are also presented in order to show how the SVs evolve during the temporal movement. Figure 2.20 and Figure 2.21 show

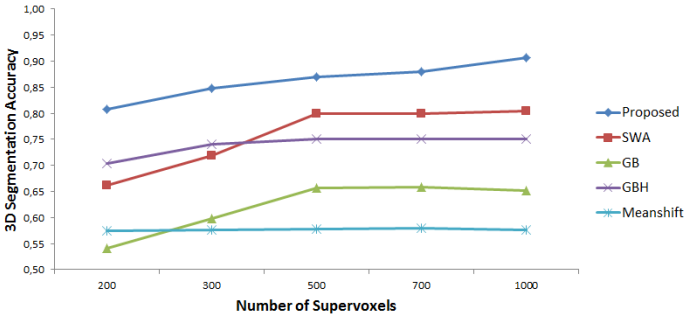


Figure 2.18: 3D Segmentation Accuracy for SP Numbers 200 – 1000 and Frame Number:5

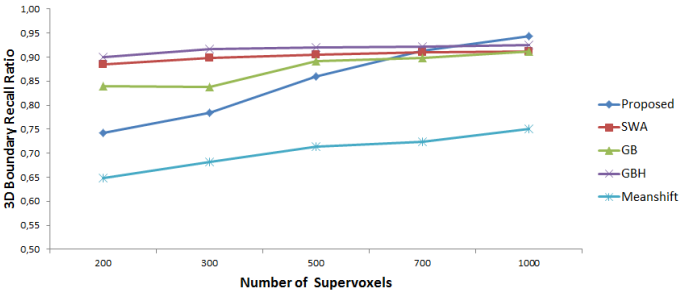


Figure 2.19: 3D Boundary Recall Ratio for SP Numbers 200 – 1000 and Frame Number:5

the changes in video for a selected slice in the image. A horizontal slice is extracted from the image and the evolution of pixels in the same location has been observed for the following frames in the video. At the lower part of Figure 2.20, there is a zoomed version of a part of the temporal voxel boundaries. One can observe the coherency in the voxel boundaries in the succeeding frames. The voxel boundaries can be identified for every 7 frame. Voxels are marked with different colors for visualization. Figure 2.21 also displays similar boundary changes along the temporal dimension. One can observe the difference due to selection of the voxel size and its effect on the generated boundary.

2.5 Conclusion and Discussion

In this chapter a novel superpixel, as well as a supervoxel extraction method is presented with contributions in terms of both computational efficiency and segmentation performance. In the proposed technique, SPs and SVs are updated iteratively through the boundary pixels based on color and spatial similarity. The boundary adaptation idea and energy function selection are the two main contributions of the proposed method enabling efficient implementation and segmentation accuracy. The experiments are conducted for different energy function combinations and two of them are selected for the comparisons against the state-of-the-art. The effects of utilizing different color spaces and distance metrics have been examined during the experiments and it is observed that LAB color space has shown superior performance in terms of boundary adaptation compared to RGB. Similarly, utilization of geodesic distance has shown improvements in segmentation performance compared to the Euclidean metric. Necessity of the proposed convexity constraint is also explicitly justified through a graph based interactive segmentation application. According to the extensive comparative tests with state-of-the-art, it can be concluded that the proposed scheme yields a remarkable alternative for SP and SV extraction methods with faster execution times and competitive segmentation performances. This has been also supported with the visual results where the generated SPs and SVs is observed to show strong adaptation on the object boundaries.

Possible limitation of the proposed method could be observed when the initial SP size is selected too large. In that case, boundary adaptation of the SP might not be possible. A future direction in order to solve such issues might be to adaptively detect optimum SP size or raise a warning in the case that SPs are not well adapted to the object boundaries. This can be done by detecting the edges beforehand and checking for an overlap between the obtained SP boundaries and the computed color/intensity edges. Alternatively, dividing the SPs depending on the existing edges might also be a way to overcome such issues.

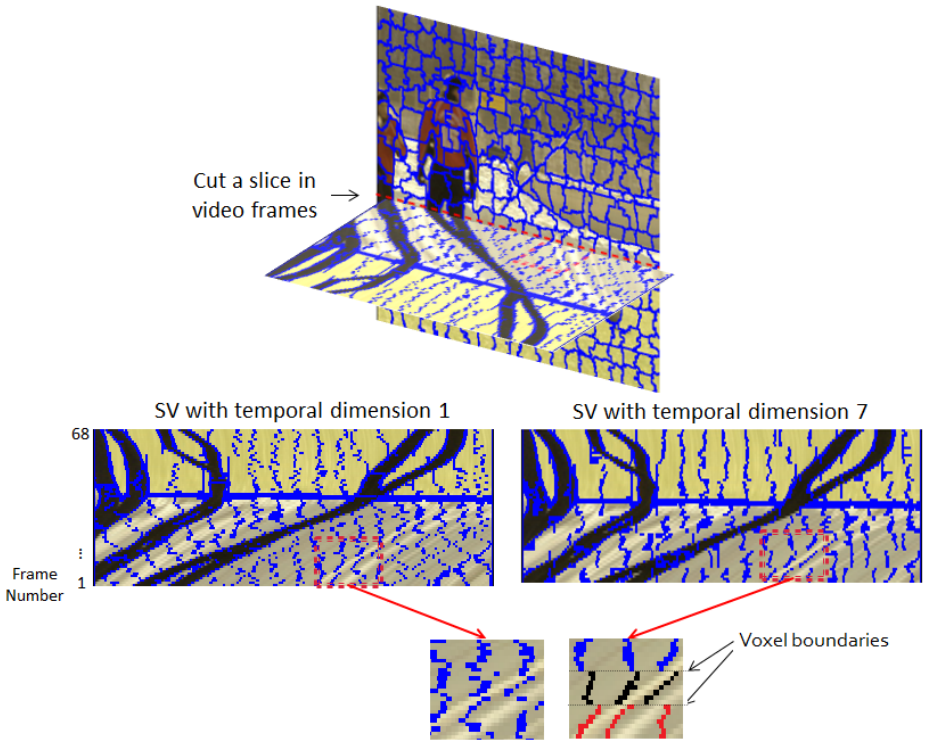


Figure 2.20: SV evolution in video frames for temporal dimension 7

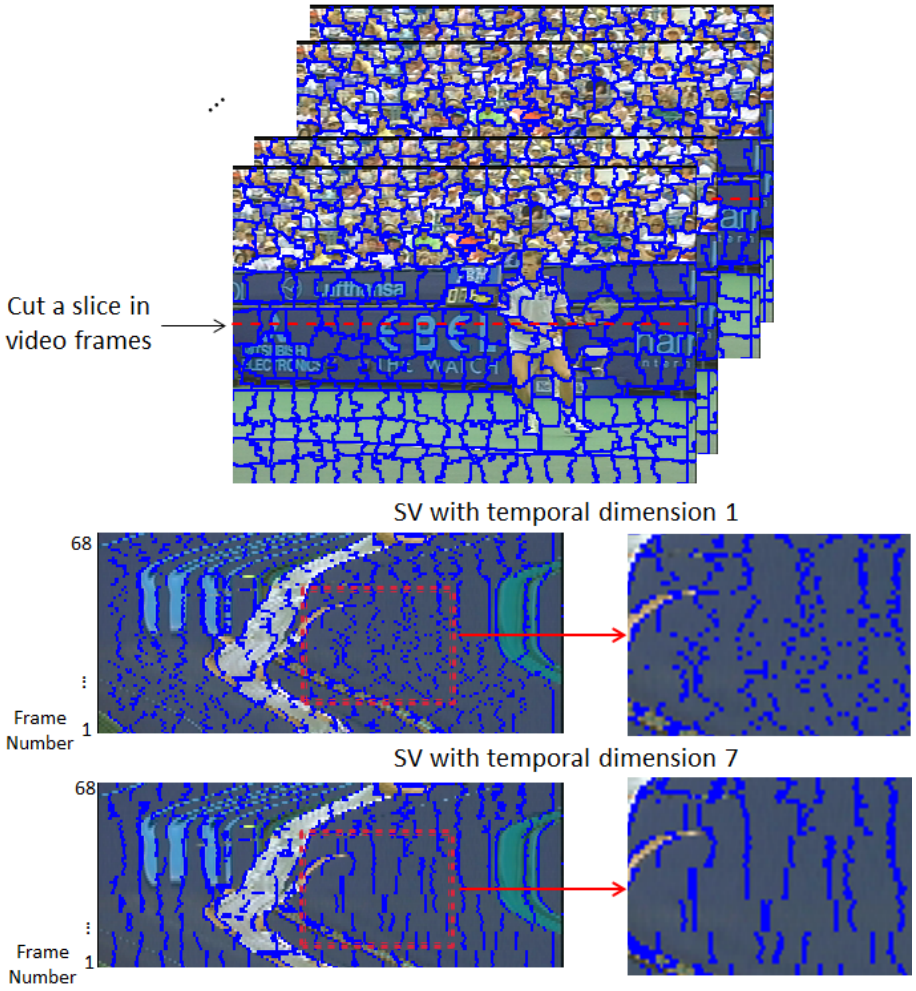


Figure 2.21: SV evolution in video frames for temporal dimension 7

CHAPTER 3

USER-ASSISTED MONO IMAGE SEGMENTATION

3.1 Introduction

Previous chapter presents superpixel based image representation. The applications using this representation is explained in the rest of the thesis. In this chapter, the general framework towards achieving user-assisted (a.k.a. interactive) image segmentation is presented. A multi-label object segmentation method is explained in detail; an application of image segmentation for 2D/3D conversion purposes is discussed in Chapter 3.4. This study approaches image segmentation as an energy minimization problem on a Markov Random Field (MRF). The goal of the proposed energy minimization technique is to achieve minimum energy potential labelling, where pixels labels correspond to image segments. This has been realized on a graphical framework where graph nodes are generated from the superpixel atomic structures as explained previously in Chapter 2. The performance of the proposed technique is evaluated using objective metrics on a ground truth image segmentation dataset.

General pipeline of the proposed method can be summarized as follows; 1) User inputs are used to determine the object locations. 2) Assignment of energy function on the image is done regarding the user inputs and image statistics. 3) Iterative graph cut energy minimization is used to define optimal labeling. The output of the optimum energy labeling will supply the segmented image.

The general outline of the chapter is as follows: Firstly, the related literature about user-assisted image segmentation is summarized. Following that, the proposed image segmentation method is explained in detail in Section 3.3. Following the presentation of the 2D/3D conversion technique in Section 3.4, the chapter is concluded. Extension of the proposed idea on stereo and video data is presented in the next chapter. Experimental results of the proposed method is also supplied in order to have a clear evaluation of the proposed performance.

3.2 Related Work

There has been a considerable amount of work in the field of user-assisted image segmentation for more than a decade now. Drawing lines (a.k.a. scribbles) on the image has been mostly selected as the interaction medium with the image. The human operator first selects the object to be segmented and draws scribbles on the representative areas of the object of interest. Similarly, scribbles on the background region are used to learn which part of the image is intended to be segmented. An object and background model for the given image is generated. Additional user input is also allowed for further processing if the resulting segmentation is not satisfactory. The main constraint on the whole system is that it should be fast enough so that the user can directly interact with the image repeatedly. The time between the user interaction and segment generation should be kept at minimum in order to achieve a natural interface. Moreover, the amount of help from the user should be at minimum in order to propose a system with decent running time. This means that the system should understand quickly and accurately what the user intends to segment. These criteria make the system difficult to be realized. The more the interaction, the more accurate the image segments. However, this requires more of the user which is usually not preferred. The academic literature regarding the user-assisted image interaction is presented in this section in detail. Moreover, some commercial/open source products are also mentioned for completeness.

A simple interaction method to segment an image is to draw bounding lines around the object of interest and select the segment that lies inside the selected region. Such a method does not require any region modelling for the object and/or background. However, such methods are not only time consuming but also prone to errors since pixel wise accurate drawing is difficult for a standard user. The commercial image editing software Adobe Photoshop [1] and the open source version GIMP [3] offer such selection tools. Another basic interaction method is to paint the object with a brush tool and generate the pixel precision segment by selecting the painted pixels. This is also time consuming and erroneous; hence, becomes impractical for a large number of images. A smarter boundary interaction method is to select the rough boundaries of the object and let the system adapt to the boundaries depending on the intensity edges. This method is commercially known as the Magnetic Lasso tool in Adobe Photoshop [1]. The Intelligent Scissors idea [90] is also implemented in the open source GIMP [3] image editing software.

Intelligent Scissors [90] requires user to select object boundary manually. Depending on the given user scribbles, a cost function is minimized for an optimum boundary between the given input points. Figure [90] depicts a typical input point selection for the method. Additional user interaction is possible if the resulting segment is not satisfactory enough. The proposed cost function uses the Laplacian zero-crossings, gradient magnitude and the direction of the gradient.

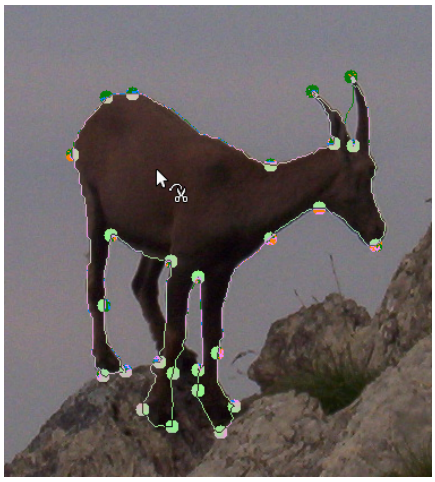


Figure 3.1: Intelligent Scissors method with user interaction points spotted for optimum boundary estimation

The general aim of the Intelligent Scissors method is to find the optimal path between the two user inputs where sharp changes on the path are penalized. Although this method performs well in terms of accuracy where strong gradients exist in the image boundary; it requires too much user interaction and this makes it impractical for large datasets. This method is further developed in [15] and [102] by the utilization of oversegmented region representation.

Another early approach for user interaction is to use initial seed points for region growing towards image segmentation. In this type of approach, the human operator is supposed to initialize a seed point in order to indicate where the object center is located. The object location is then estimated by grouping the pixels around the initial seed point. An early example for such an approach by [149] starts with the initial label and grows the region iteratively depending on an intensity or color based similarity metric. Whenever the similarity between the segment and the candidate pixel goes above a certain value, the pixel is assigned to the same region of the segment. This iterative process depends on a brute-force search on all the pixels for labeling. The major issue in such methods is the unreliability of the final performance since it is mostly depending on the selected threshold and the region statistics. If there is a smooth edge in the object boundary, it might end up generating a totally erroneous segmentation. In order to overcome the ambiguity in the threshold selection, Seeded Region Growing method has been proposed [9]. This method requires user to set the initial region seeds where the region growing should start. The similarity between an individual pixel and the neighbor region is defined as in (3.1).

$$pixSim(p, R) = I(p) - \frac{1}{|R|} \sum_{q \in R} I(q) \tag{3.1}$$

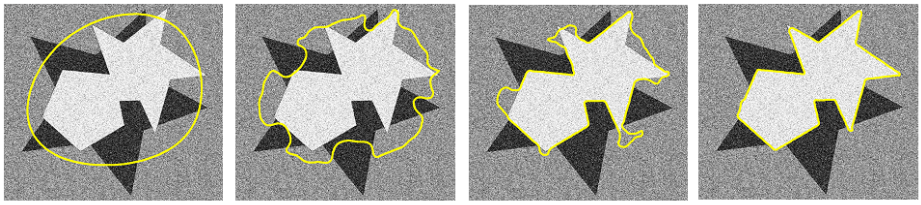


Figure 3.2: Segmentation boundaries for intermediate energy minimization steps (1,3,10,20) for the Active Contour method [68]

$I(p)$ is the intensity value of the pixel p. q is the pixels in region R and $|R|$ is the number of pixels in the region. An improvement of this approach is presented in [117] where users are allowed to draw scribbles on the object regions for a better region identification. Our earlier paper [123] further utilizes and improves this idea by proposing superpixels as the region identifiers for an efficient graph based implementation for a mobile device.

Another important study in the literature is the Active Contours, [68]. This method proposes to use an energy metric for deciding the region boundary between the user inputs. The energy function in this method utilizes two different type of energies; internal and external. Internal energy controls the shape of the contour. The aim is to keep the length of the contour short and changes in the first derivative of the contour at minimal. This is proposed to generate a smooth curve at the object boundary. The external energy forces the contour to adapt the object edges by pulling it towards the sharp gradients in the image.

The user is required to draw a rough contour initially. The energy minimization is conducted via the gradient descent method. It is important to note that the final energy level is not necessarily the global minimum, and hence the initial user input becomes important on the final segmentation accuracy. A visual result on the active contour iteration can be seen in Figure 3.2. This figure is taken from the study [109] and the implementation has been performed using the Active Contour Toolbox by [36]. However, the main problem with this method is that the selection of the initial contour is very crucial in the final output segmentation. This property makes the method prone to errors in the case of wrong user initialization.

The study by Boykov and Jolly in [22] deals with the efficient computation of the global optima under the energy function assignment in (3.2).

$$E(L) = \sum_{p \in P} D_p(L_p) + \sum_{(p,q) \in N_p} V_{(p,q)}(L_p, L_q) \tag{3.2}$$

In this formulation, L is the labeling of image P, D_p is the data penalty function, $V_{p,q}$ is the interaction potential and N_p is the neighborhood of pixel p.

The first term in (3.2) is the unary term and it employs a penalty value for each pixel p by looking at the likelihood of assigning it to foreground or background. In the original study this likelihood is estimated by the histogram of the user given seeds on the image. The second term, the pairwise potential, penalizes similar neighboring pixels when they are assigned to different labels.

The energy minimization framework of graph-cut as defined in [24], satisfies the globally optimum binary solution in terms of the defined energy potentials. The resulting binary segmentation among all possible solutions ensure the energy minimization under the regional (unary) and boundary (binary) energy terms. The additional interactions are inherited in the system by an efficient recomputation method.

The proposed framework in this thesis utilize graph-cut based energy minimization for obtaining final object segmentation. Therefore, further detailed discussion of the graph-cut method is left to the next section 3.3.

An improvement to the seminal work of graph-cut [24] is proposed by [105] where serious effort is invested in the foreground and background region modeling. This is important for determining meaningful energy potentials for the unary term in (3.2). Gaussian mixture model (GMM) of the user specified region is incorporated in the modeling task and more accurate segmentation results have been achieved. In this framework user assistance is defined as drawing a rectangle around the object of interest. The color models of the foreground and background regions are estimated from pixels inside and outside the rectangle and the first step segmentation is performed depending on the estimated models. Based on the intermediate segmentation, the foreground color model is re-estimated for a better region identification and the segmentation procedure is performed again until a stable result is achieved. This method also allows further interaction until the user is satisfied with the final segmentation output.

The GrabCut method [105] has proven success due to its easy interaction, speed, accuracy and wide availability. The Microsoft Office 2010 software is shipped with a foreground extraction functionality based on the GrabCut method.

The main drawback of the graph based energy optimization methods is that they require a considerable amount of time and memory for the computation of global energy minimization. A common way to decrease the amount of computation is to use superpixels in order to reduce the size of the graph. The Lazy Snapping method in [79] proposes using an initial oversegmentation on the image before computing the energy minimization procedure. Since this thesis also utilizes such an oversegmentation approach for obtaining efficient image segmentation, detailed analysis will be provided in the rest of the chapter.

The maximal similarity based region merging (MSRM) [91] method aims at merging regions automatically that are initially segmented by the mean shift technique [31].

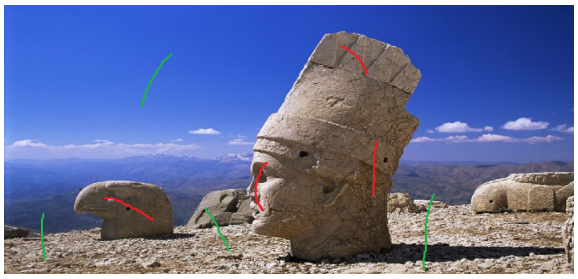


Figure 3.3: Input strokes are shown with 'red' for the foreground region and 'green' for the background region

The object contours are then effectively extracted by labeling all the non-marked regions as either foreground or background. The region merging process is intended to be adaptive to the image content; therefore, no initial threshold setting would be necessary.

3.3 Proposed Image Segmentation

This chapter explains an image segmentation method where previously accepted methodologies are widely incorporated in order to propose an efficient and accurate framework for the user-assisted image segmentation problem. In addition to the single mono image segmentation, the following chapter explains the extension of the method on the stereo and video footage in order to propose a complete pipeline. The proposed method utilizes the energy minimization framework as defined in [24]. This optimization framework satisfies the globally optimum binary solution in terms of the defined energy potentials. The resulting binary segmentation among all possible solutions ensure the energy minimization under the defined regional (unary) and boundary (binary) energy terms. The additional interactions are also inherited in the system using an efficient energy recomputation method in case the resulting segmentation is not visually satisfactory.

3.3.1 User Interaction

This study incorporates a common medium for interacting with the image. User assists the region segmentation process by drawing lines (scribbles) on the image representative regions. Figure 3.3 shows an example interaction where red scribbles correspond to the selection of the foreground object whereas green scribbles correspond to the background region. The input strokes are selected from the locations of diverse color and intensity in an attempt to cover a wider region characteristics for a more accurate result. The final segmentation performance also depends on the superpixel boundary adaptation that will be discussed later in the results section.

The proposed framework also allows further user interaction. At the end of the initial region segmentation, if the result is erroneous or unsatisfactory, user can further interact and correct the possible erroneous regions on the image by supplying additional scribbles.

3.3.2 Energy Function Selection

At this point, it is important to emphasize the underlying idea that produced the seminal work of graph-cut optimization framework. The roots of the analogy between images and statistical physical systems depend on the interpretations by Gemans [52]. Image is basically considered as a statistical mechanical system and the pixel gray levels and the presence and orientation of edges are analysed as states of atoms in a lattice like physical system. The energy function assignment of the mentioned system determines its Gibbs distribution. As the equivalence relation between the Gibbs distribution and Markov Random Field states; the MRF image model is also determined by this energy assignment. The degradation mechanisms in the both image and physical system can be restored using the *maximum aposteriori probability* (MAP) estimate defined by the clique potentials. The MAP estimate is analogous to the isolation of low energy states of the system.

The graphical illustration and neighborhood concept is helpful in defining and understanding the MRF and Gibbs distribution. Let $G = (V, E)$ be an undirected graph with nodes v and edges e. The set of nodes $V = \{v_1, v_2, ..., v_N\}$ and the neighborhood $N = \{N_v, v \in V\}$ are defined in the special context of application. An edge $e_{i,j}$ connects two neighbor nodes v_i and v_j. Let $F = \{f_v, v \in V\}$ denote the family of random variables indexed by the vertices of the graph. Let the possible states (labels) of the vertices are defined as $\Lambda = \{0, 1, 2, ..., L-1\}$ so that $f_v \in \Lambda$ for all v. Let Ω be the set of all possible combinations $\Omega = \{f = (f_{v_1}, f_{v_2}, ..., f_{v_N}) : f_{v_i} \in \Lambda, 1 \leq i \leq N\}$

The event $\{F_{v_1} = f_{v_1}, ..., F_{v_N} = f_{v_N}\}$ is abbreviated as $\{F = f\}$. Under these definitions F is an MRF with respect to the neighborhood N if

$$P(F = f) > 0 \; \forall \; f \subset \Omega; \tag{3.3}$$

$$P(F_v = f_v \mid F_r = f_r, r \neq v) = P(F_v = f_v \mid F_r = f_r, r \in N_v) \tag{3.4}$$

for every $v \in V$ and $(f_{v_1}, f_{v_2}, ..., f_{v_N}) \in \Omega$. For the graph G satisfying (3.3) and (3.4) the joint probability distribution $P(F = f)$ is uniquely determined by the conditional probabilities on the right hand side of (3.4). The property states that a subset $A \subset V$ is said to be complete if each pair of vertices in A defines an edge of the graph. Ordinary 1-D Markov chains are MRF relative to the order of nearest neighborhood system. An r^{th} order Markov process can easily be regarded as MRF by a careful

neighborhood definition where all r^{th} previous states are set as the neighbor nodes. The Gibbs models were first applied to image representation by Hassner and Sklansky [60]. A Gibbs distribution relative to {V,E} is a probability measure π on Ω with the following representation:

$$\pi(f) = \frac{1}{Z} * e^{-U(f)/T} \tag{3.5}$$

Z and T are constants and U is an energy function of the form;

$$U(f) = \sum_{c \in \mathcal{C}} V_C(f) \tag{3.6}$$

Z is the normalizing constant and T stands for temperature.

$$Z = \sum_{f} e^{-U(f)/T} \tag{3.7}$$

C is the set of cliques for N where a subset $C \subset V$ is a clique if every pair of distinct vertices in C are neighbors. Each V_C is a function on Ω with the property that $V_C(f)$ depends only on f_v of f for which $v \in C$. Such a family of V_C is called a potential. Under the definition of the graph G and the neighborhood system N; F is an MRF with respect to N if and only if $\pi(f) = P(F = f)$ is a Gibbs distribution with respect to N. The main benefit of this equivalence property is that the MRF is easily defined using the potentials $V_C(f)$ instead of local characteristics which is almost impossible. A basic example of a potential function can be given as the difference between the degraded and the local mean image, which in fact proves to be useful for image restoration posterior probability assignment. The proof of the equivalence can be found in [17] and [70].

Gemans [52] offer a relaxation technique for the solution of MAP estimate of the MRF. Further effort have been devoted to develop different methods for the approximate or accurate MAP estimation of MRF. Boykov et. al. [23] mainly focus on the global energy minimization technique by solving a minimum binary (or multiway) cut on a graph based representation. The efficient computation of MAP estimate is the main virtue in the proposed study.

The main result of Hammersley-Clifford theorem stating the relation of joint event probability to clique potential in the neighborhood system N is stated as $P(F = f) \propto exp(-\sum_{C} V_C(f))$. The clique potential V_C describes the prior probability of a particular realization of the elements of the clique C. As the MRF restricts cliques in the neighborhood definition the correlation can be stated as;

$$P(F = f)) \propto exp\Big(-\sum_{p \in P} \sum_{q \in N_p} V_{(p,q)}(f_p, f_q)\Big) \tag{3.8}$$

The aim is to estimate f based on the observations O which is related to f with respect to a likelihood function $P(O|f)$. I_p is the observed intensity at pixel p and the event of that realization is $\{I_p = i_p\}$. The stated conditional probability of observing an intensity given the true intensity is related to the noise model as follows.

$$P(O|f)) = \prod_{p \in P} g(i_p, f_p) \tag{3.9}$$

where $g(i_p, f_p) = O(I_p = i_p|F_p = f_p)$ represents the noise model on an individual pixel. Noise sensor is assumed to affect each pixel independently hence the general observed image probability is modeled as the multiplication of probability of individual pixels.

The general aim is to assign labels L where $l \in \Lambda$ to each pixel by maximizing the MAP estimate of the given observations $P(f|O)$. Bayes' rule is used here to convert the problem to previously modelled probability density functions. $P(f|O) \propto P(O|f)P(f)$. Utilization of (3.8) and (3.9) leads to the following result that in order to maximize the posterior probability the following energy function should be minimized.

The minimization necessity declares the energy function as a penalty function. The clique potential that causes an increase in the penalty is defined such that there is absolutely no penalty in the case that the clique members (neighbor graph nodes or pixels in the image restoration case) are of same value.

$$V_{(p,q)}(f_p, f_q) = u_{\{p,q\}}(1 - \delta(f_p - f_q)) \tag{3.10}$$

where $u_{\{p,q\}} \geq 0$ and yields a Potts model if $u_{\{p,q\}}$ is constant for all {p,q}. Hence discontinuities between any pair of labels are penalized equally. The potential is symmetric (independent of orientation) with respect to the set members $u_{\{p,q\}}$ and hence MRF is isotropic. The value of $u_{\{p,q\}}$ can be interpreted as a penalty for the discontinuity between the different labeled cliques. Hence the prior probability $P(f)$ favors continuous labeling against discontinuities. After the elimination of the first summation the prior probability (3.8) becomes:

$$P(F = f)) \propto exp\Big(-\sum_{p,q \in E_N} 2u_{\{p,q\}}(1 - \delta(f_p - f_q))\Big) \tag{3.11}$$

Let the graph $G(V, E)$ with non negative edge weights is constructed with the labeling defined as $\Lambda = \{0, 1, 2, ..., L - 1\}$. The subset of edges $C \in E$ is called a multiway cut if the terminal points (labels) are completely separated after the removal of the cut edges C. The graph obtained after the removal operation is called the induced graph

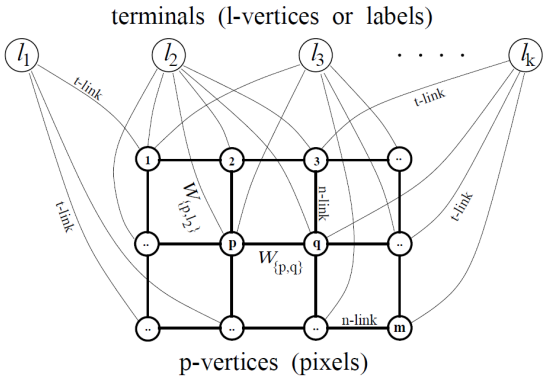


Figure 3.4: Graph $G(V, E)$, terminals $\Lambda = \{0, 1, 2, ..., L-1\}$ and p vertices are $V = \{v_1, v_2, ..., v_N\}$. Each p vertex is connected to at least one terminal vertex [22].

$G(C) = \langle V, E - C \rangle$ The cost of the cut is shown with $|C|$ and is calculated by summing its edge weights.

It is a well known field of study how to efficiently find the minimum cost multiway cut. Moreover, the equivalence of the minimization problem defined in (3.12) to the multiway cut problem makes it more of an interest for the MAP estimation problem. The equivalence is shown in detail in [23] and will be shortly summarized here. The graph G is constructed with two types of vertices p (pixels) and L (labels). Later in this chapter the p vertices will correspond to oversegment patches instead of pixels but not detailed here for the sake of coherency. L vertices are the terminals points to whom p vertices are linked through t-links. p vertices are connected to each other through n-links if they are neighbors referred as E_N. The constructed graph is shown in Figure 3.4. Since the multiway graph cut -by definition- separates all terminals only one t-link is left for each p vertex. A multiway graph is called $feasible$ if each p vertex is left with exactly one t-link. Since the weight assignments are in accordance with the energy minimization formula, it is ensured that the minimum cost multiway cut minimizes $E(f)$ in (3.12). The general multiway minimum cut problem is NP-complete. However, there are approximate solutions with linear running time [35]. The method by Boykov et al [23] which is also a linear time complexity, proposes an iterative solution for the feasible cuts on G. Initially any feasible cut is considered, at each iteration a pair of vertices are reallocated between two terminals, l_i and l_j. Hence a two terminal min cut problem is solved to find whether possible rearrangement might reduce total cost energy. Each iteration considers a new pair of terminals until all distinct pair combinations are visited. The algorithm continues until no possible label switch can further decrease the total energy.

$$E(f) = \sum_{p \in P} \sum_{q \in N_p} V_{(p,q)}(f_p, f_q) - \sum_{p \in P} ln(g(i_p, f_p)) \tag{3.12}$$

Graph construction of a general MRF energy minimization problem is useful in order to convert the problem into a multiway cut minimization problem on which iterative accurate and approximate solutions exit. The iterative method depends on the binary max flow (min cut) solutions. Hence the binary max flow case is of great importance in obtaining a general solution. The general cost energy to be minimized (3.12) is rearranged as follows.

$$E(L) = \sum_{p \in P} D_p(L_p) + \sum_{(p,q) \in N_p} V_{(p,q)}(L_p, L_q) \tag{3.13}$$

where L is the labeling of image P, D_p is the data penalty function, $V_{p,q}$ is the interaction potential and N_p is the neighboring of pixel p. As in connection with (3.12) data penalty is designed so as to increase as the probability of assigned label given the observation decreases. Hence the previous minus sign is eliminated by interaction potential favors coherence by increasing energy under discontinuities. Greig et al. [56] who is first to relate min-cut/max-flow algorithms from combinatorial optimization to energy function minimization has constructed a two terminal graph whose minimum cost cut gives globally optimal binary labeling. The terminal labels in this specific configuration are named as $source, s$ and $sink, t$. Figure 3.5 shows a two terminal labelled graph on a 3×3 image. The vertices were previously considered as image pixels but this work utilizes superpixel primitives in the energy minimization framework.

The energy minimization problem has been successfully converted to the estimation of minimum cut on the constructed graph with carefully assigned neighborhood edge weights (n-links) and data penalties (t-links). Combinatorial optimization fundamentals reveal the equivalence of maximum flow to the minimum cut solution where maximum flow is interpreted as flow from source to sink where edge weights are taken as the pipe capacities. The theorem according to Ford and Fulkerson [50] states the equivalence of max flow to minimum cut solution by showing the maximum flow graph as saturated in the edges are in fact divided into two disjoint regions which actually correspond to the minimum cut of the graph. Moreover, the maximum flow value is equivalent to the minimum cut value since the maximum flow is calculated by the saturated edges, which are actually cuts.

For the solution of min-cut problem, different approaches with polynomial bound are developed. The augmented paths based algorithm by Dinic [41] proposes pushing flow through the non saturated edges until no more flow is possible. Each push in the graph G reduces the residual capacity which is tracked using the residual graph G_f. Residual graph is an update of the original graph with less capacity due to introduced flow in the previous state. At each cycle, the minimum distance path is obtained to push the maximum flow available through the given path. The updated capacities are stored in the residual graph. Maximum capacity is achieved when there is no possible flow in the network or equivalently there is no path from source to sink without crossing a saturated edge (cut).

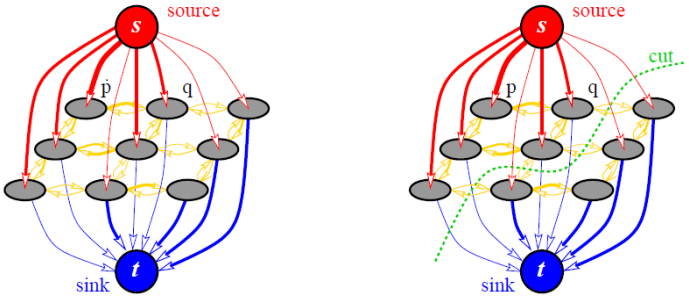


Figure 3.5: Directed graph. Edge costs are reflected by their thickness [22]

There is the other approach named as push & relabel type algorithms [53]. A push operation is sending a part of the excess flow from one node to the other. If there is available capacity between two nodes but the source sink direction is not met (flow from up to down is possible) relabelling operation is realized. After relabelling the current vertex is moved to the front of the list for the next push operation and the traversal push algorithm starts from the beginning of the list.

The method by Boykov et al. [22] presents an improvement to the standard augmented paths type technique. The main handicap with the standard technique is the necessity of repeated calculation of bread first search from sink to source which could be a very expensive operation. The efficiency is satisfied with some modifications on the standard search tree generation. There are two (instead of one) search trees generated one from source and one from the sink. Re-usage of these trees is also presented for higher efficiency. The details of the algorithm is also summarized for completeness.

Two non overlapping search trees are constructed with roots S and T. In S tree edges from each parent to children are non saturated but edges in T are non saturated from child to parent. Nodes outside S and T are free. The nodes on the outer border are active while internal nodes are passive and can no longer grow. An augmenting path is found whenever a contact between two trees is encountered. The general flow of the algorithm is observed in three stages; growth, augmentation and adoption stage.

In the growth stage trees are expanded over the active nodes. Free nodes are included in the tree as children and become active members for next expansion. As soon as all the neighboring vertexes are explored the current node becomes passive. The growth stage is terminated whenever an contact between the active and passive nodes is encountered. The encounter creates a path from source to sink.

The path found in the previous stage is saturated by pushing the maximum flow available from source to sink. This causes some of the edges to be saturated and the nodes to become orphans (no longer connected to the parent).

In the final adoption stage, new parents for the orphans are searched in order to retain

the single S and T tree structure. In no parent is found for the node with positive capacity, the node is isolated and set free, leaving its children as orphans. This stage is repeated until there is no orphan in the tree.

The three stages are repeated until there is no grow possible and trees are separated by saturated edges. This implies that the maximum flow (minimum cut) is reached.

The general energy minimization framework using Gibbs distribution and MRF equivalence is explained above. From this point on, the application of these findings on the segmentation problem will be considered. User assisted segmentation is a popular field of research due to its performance superiority compared to automatic methods. Many successful methods are proposed in the literature [105], [144]. The interactive segmentation problem is converted to the energy minimization problem where optimum segmentation yields minimum energy configuration among all possible segmentation. User input values are used as primary ground truth data for the background and foreground regions. Multi level segmentation is also possible within the same framework. Given the user input scribbles, the rest of the image is automatically segmented depending on a cost function defined as above (3.12).

The well known interactive segmentation work utilizing graph cut energy minimization technique [24] indicates the definition of soft and hard constraints in the energy function. The hard constraints are the user inputs which indicate if the corresponding pixel belongs to the foreground or background. Soft constraints are determined by the similarity of neighboring nodes. Figure 3.5 shows the graph construction of the segmentation problem where S is the source terminal indicating "object" and T is the sink terminal indicating "background".

3.3.3 Superpixel Graph Generation

The explained pixel based energy optimization approach requires generation of pixel size graph. However, this computationally complex approach limits the applicability of the method to low resolution images due to memory and run time issues. Therefore, the extension of the idea on the superpixel domain has been utilized. Superpixel based image representation has gained interest due to increased efficiency by converting pixel based computations to superpixel framework. Widely known mean shift technique [31] can easily be converted to an oversegmentation method with strong region homogeneity. However, the lack of region shape priors limits the usage of mean shift technique on graph based approaches where lattice structure is not guaranteed. Hence, the relatively new proposed turbopixel idea [76] is widely utilized for regular lattice generation. An extension of this idea by imposing a convexity constraint on the oversegment regions has been widely explained in the previous chapter.

The proposed image segmentation technique utilizes the findings of the graph cut en-

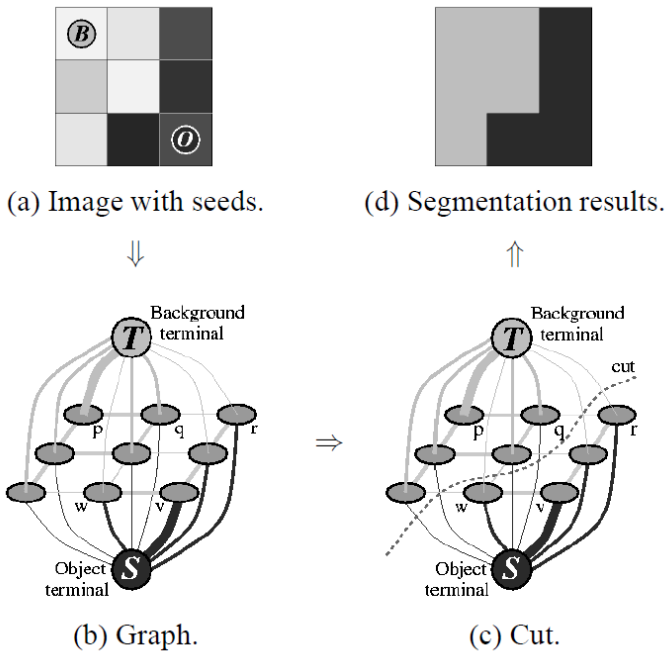


Figure 3.6: 2D segmentation on a 3×3 image. Thickness of edges indicate weights [22].

ergy minimization framework with the superpixel idea in order to obtain an efficient user-assisted image segmentation. It is important to note that segmentation is an intermediate output in obtaining a layer based representation of the segmented objects. Further applications of segmentation (2D-3D conversion, disparity remapping and region based image classification) are further presented in the following chapters of the thesis.

In the selection of oversegment method main concern was to create regular grids on the image region with convexity constraint. Regularity is important for graph generation because it is aimed to keep the number of neighbors close and distance between nodes similar for all nodes. These necessities and straight forward implementation has led to turbopixel idea [76] with the additional convexity constraint as explained in the previous chapter. The emphasized properties has proven advantages over mean-shift [31], watershed [134] and normalized cut type oversegmentation methods. The main advantage of the convexity prior is its compactness constraint while still showing powerful adaptation at the object boundaries. The compactness constraint does not only prevent undersegmentation but also proposes a deterministic run time with a rigid graph structure.

The mentioned advantages can be observed in Figure 3.7. Oversegment boundaries adapts to the local image edges and hence the assumption of assigning same label to the pixels belonging to same superpixel region becomes valid. The validity of the

assumption can also be observed on the same Figure 3.8 where superpixels are painted with the mean RGB values preserving image integrity. The general structure of the superpixel regions resemble a hexagonal honeycomb especially on the smooth areas and hence creating a structured graph with similar neighbor numbers and similar neighborhood distances as intended.

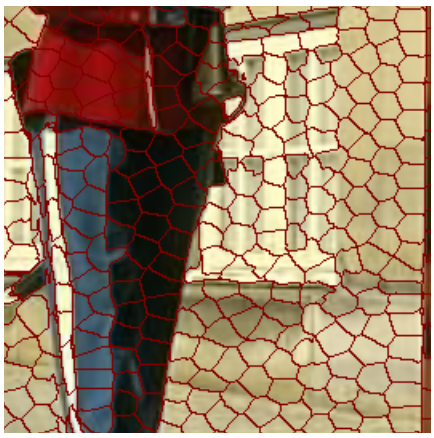


Figure 3.7: Oversegment Boundary Adaptation

After the SP extraction step, an additional region based filtering idea is proposed in order to increase the robustness of final segmentation by boosting inter region similarity. What is aimed at this step is to increase the similarity of neighboring superpixels whenever they belong to the same object. This motivation has been realized by an iterative similarity propagation model as shown in Figure 3.9.

The similarity weight between the superpixel nodes p and q_j are calculated individually for each color channel component (RGB color channel is utilized for illustration purposes) (3.14).

$$weightR(p, q_j) = e^{-\frac{-(R_p(i) - R_{q_j}(i))^2}{\sigma^2}} \tag{3.14}$$

p is the center node for which the intensity update is proposed, q_j is the neighbor of node p. $R_p(i)$ and $R_{q_j}(i)$ represent the mean red values of the superpixels at the i^{th} iteration. The exponential function description causes the similarity component to vanish rapidly whenever the mean difference in RGB components increases. σ is computed as the variance of the mean values of SPs in the neighborhood. The similarity weight between nodes are used to update the cumulative red values on the center node p (3.15).

$$meanCumR(p) += \frac{weightR(p, q_j) * R_{q_j}}{neighNum(p)} \tag{3.15}$$

Figure 3.8: Oversegment Regions with Mean Intensity

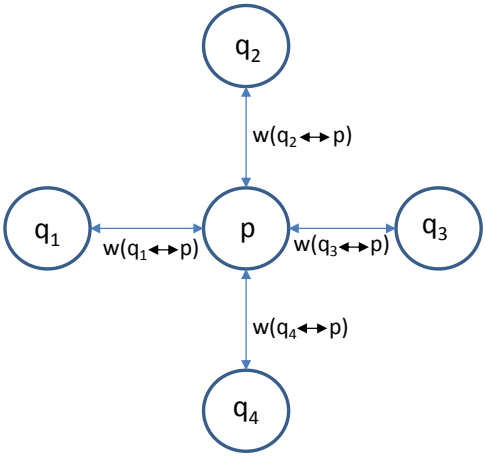


Figure 3.9: Node Weight Propagation

Figure 3.9 shows how node p interacts with the neighbor nodes q_i in a four neighborhood model. Total cumulative weight (3.16) coming from neighbors is also calculated for the final update on the center superpixel (3.17).

$$totWeightR(p) += \frac{weightR(p, q_j)}{neighNum(p)} \tag{3.16}$$

$$R_p(i+1) = \frac{meanCumR(p)}{(1 + totWeightR(p))} \tag{3.17}$$

The mean red value at iteration $i + 1$ is updated using the mean red value in the previous iteration i. As the iteration steps proceed the inter region similarity increases and the amount of updates in terms of total change in RGB values decreases. The proposed update is realized for each color channel independently.

The proposed intensity based weighting model prevents the propagation of intensities through dissimilar regions and hence encouraging the integrity of the object by discriminating it from background and other objects. At each iteration, information from the neighboring superpixels are exchanged. A total number of 10 iterations enables information propagation from 10^{th} order neighbors from each side of the center node. Depending on the superpixel size it covers a considerable amount of the image region successfully. (For an $20x20$ oversegment region 10 iterations enables roughly a total of $400x400$ sized region information to propagate on each center node.)

The proposed information propagation results in a boundary preserving color filtering as shown in Figure 3.10 where oversegment regions are painted with original and updated mean values.

The intensity based similarity between two non neighbor superpixels which is necessary for graph cut energy formulation is normally dependent only on the intensity of two nodes. However, with the utilization of the proposed filter, the intensity information propagates through neighbor nodes. This propagation is powerful along the similar superpixels; in other words, through the nodes which are close in terms of the geodesic distance metric. If the information cumulation on a node from many non neighbor nodes is strong; this proves that these non neighbour nodes are close through a connected geodesic path although they can be further apart in terms of Euclidean distance.

The general segmentation framework is designed such that many (N) object definitions are possible with an individual segment assignment. Each object is assigned a label L where $L = \{1, 2, ..N\}$ and each patch p belongs to either one of the objects O_L or to the background B. User inputs are used to connect the selected patches to the related object or to the background. It is conducted by assigning a comparably higher (infinite) weight to the user selected region. The energy function definition in (3.13)

Figure 3.10: On the left side, mean color values of the SP regions are shown. Region information propagation based filtering shows the changes in mean region intensities on the right side.

includes the cost of assigning a label to the corresponding patch as $D_p(L_p)$. The multi label definition is utilized as combination of succeeding binary segmentations. For each individual object, all the other user selected objects and background is considered as background and a binary segmentation is performed. This procedure is iterated for all the objects individually. The binary segmentation problem is solved by the efficient maximum flow algorithm [22].

The node-node and node-terminal edge weights are assigned according to the energy formula to be minimized. D_p is the weight of the edge from nodes to terminals. This is either set by the user during the initial strokes or automatically determined by the system otherwise. Given inputs for each object and background determine a limited model of the region. Hence, the similarity metric uses this limited model to measure the similarity of a node to the object and background. $D_p(B)$ is the edge weight from node to "Source" (Background) and $D_p(O)$ is the edge weight from node to "Sink" (Object). Similarity measure is related to the input node statistics and the distance of the current patch to the input nodes. Mean and standard deviation of the nodes (oversegment regions) in the selected color domain can be used effectively and efficiently for statistical identification. Weight of the edge between node p and background terminal is assigned as;

$$D_p(B) = max\{similarity(p,q), q \in B \cup O_{\{L-l\}}\} \tag{3.18}$$

weight of edge between node p and Object terminal;

$$D_p(O_l) = max\{similarity(p,q), q \in O_l\}0 \tag{3.19}$$

weight of edge between node p and q where $q \in N_p$;

$$V_{(p,q)}(L_p, L_q) = similarity(p,q), q \in N_P \tag{3.20}$$

and similarity between two nodes of any kind is calculated as shown.

$$similarity(p,q) = e^{-\lambda_1 * intDiff * locDiff} \tag{3.21}$$

Intensity difference ($intDiff$) between the nodes p and q in the above formulation (3.21) can be computed by mean, median or sum of intensity differences in a selected color domain. Histogram-based comparison is also powerful, but computationally complex. In our implementation mean intensity of the superpixel nodes in *Lab* color domain is utilized (3.22).

$$intDiff = \sum_{i=1:3} (abs(Mean_p(i) - Mean_q(i))) \tag{3.22}$$

Lab color space is selected due to its perceptual uniformity. This assures that a measure of distance between two color values is strongly correlated to the visually perceived color difference. The edge weights from nodes to terminals for the scribbled regions are set as follows.

$$D_p(O_l) = max\{similarity(p,q), q \in N_p\} \tag{3.23}$$

Equation (3.23) proposes that the scribbled regions are supposed to be hard wired to the selected terminal. Therefore, assigning maximum of the similarity measure in the neighborhood to the terminal is meaningful instead of assigning a constant high value independent of the region characteristics.

3.3.4 Geodesic Distance Utilization

Input scribbles on the representative regions supply a valuable information regarding the general characteristics of the background and foreground. Gaussian-mixture based region modelling is often used to define region characteristic. Such global methods utilize color intensity relation between an unidentified node and a terminal node for similarity computation. However, we use an additional distance term to assure locality in the similarity equation (3.21). This constraint makes sure that two different regions with similar color distribution can be differentiated on the image.

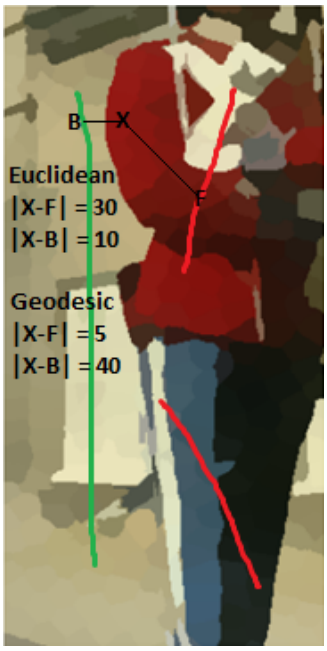


Figure 3.11: Euclidean vs Geodesic Distance: X is closer to B (F) in Euclidean (Geodesic) distance

In the selection of the distance metric, it is previously observed that Euclidean-wise closeness usually does not reveal useful information. The distance of a node to a user scribble might be close especially at the fg/bg boundaries. Point X in the foreground object in Figure 3.11 is close to the background scribble B in Euclidean terms. However, it does not necessarily imply that this point is similar to the region defined by the closest scribble point. Hence, geodesic type distance is utilized for resolving such ambiguities. The geodesic path needed to be traversed in order to reach the target point might be different than Euclidean path. In the case of graph node similarity assignment, Figure 3.11 shows a typical case where object regions close to background scribbles might easily be assigned to background when Euclidean-wise closeness is considered. Euclidean distance between a random point X and background seed B is smaller than the distance between X and foreground seed F. In order to reach point F there is a high contrast object boundary that must be passed. The minimum intensity path between two points is computed for such an energy assignment. The idea of geodesic distance has been previously addressed in a similar study [100].

Figure 3.12 illustrates the geodesic distance idea for the user input scribbles on the object and background regions. The left image shows the minimum geodesic distance of nodes to the regions defined as background by the user. The distances are normalized to $0-255$ scale for visualization, higher intensity implies closer distance. The second image on the right shows the minimum geodesic distance of the nodes to the selected object regions (red lines). Notice that the propagation of node weights decrease as the

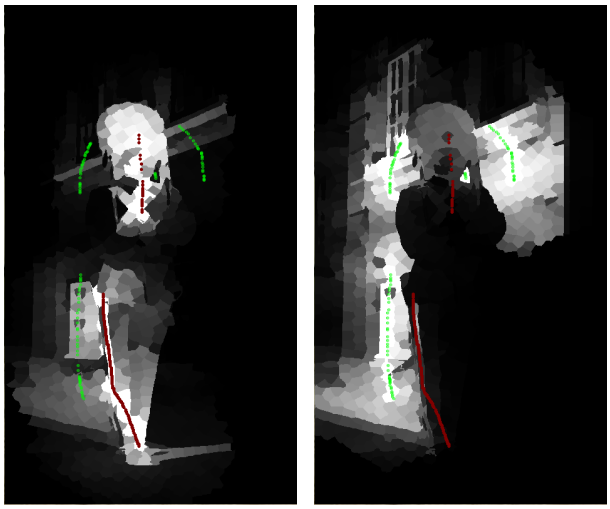


Figure 3.12: Geodesic Distance From the Object and Background Seeds After 5 Iterations

distance of the nodes are further from the input strokes. The utilization of geodesic distance in the graph cut framework provides more realistic energy assignment on the generated graph.

Efficient Distance Computation In this part, an efficient geodesic distance computation technique is presented. An iterative approach is proposed where information is transferred through the graph nodes. Figure 3.13 illustrates the iterative distance computation idea. Hexagonal regions represent the superpixel patches, the nodes in graph. The node weights are assigned as in (3.21). At each iteration, neighbor nodes send out and receive information through their connecting edges. In order to find the geodesic distance between two nodes, sufficient number of iteration has to be completed to transfer information between two nodes. In other words, a path between two nodes has to be established.

Figure 3.13 illustrates how the region coverage expands with increasing iterations. The central node n_1 has the first order neighbors $n_2...n_7$ and the second order neighbors $n_8...n_{19}$. During the first iteration of the algorithm, nodes have only the self-distance information, which is 0 as default. In the following iteration, neighboring 6 (this number is selected for illustration purposes) nodes exchange their edge weights with each other. The distance information of the first order neighbors are stored in the look up table (LUT) of each node. As the iteration number increases, the distance information table surrounding node n_1 expands. At each iteration possible multiple paths between nodes may become available, and only one path has to survive while terminating the other that has a higher cost. Two possible paths are available from n_1 to n_{10} at the end of 2^{nd} iteration; $n_1 \rightarrow n_2 \rightarrow n_{10}$, and $n_1 \rightarrow n_3 \rightarrow n_{10}$. One of them is terminated depending on the calculated geodesic distance. The idea is also

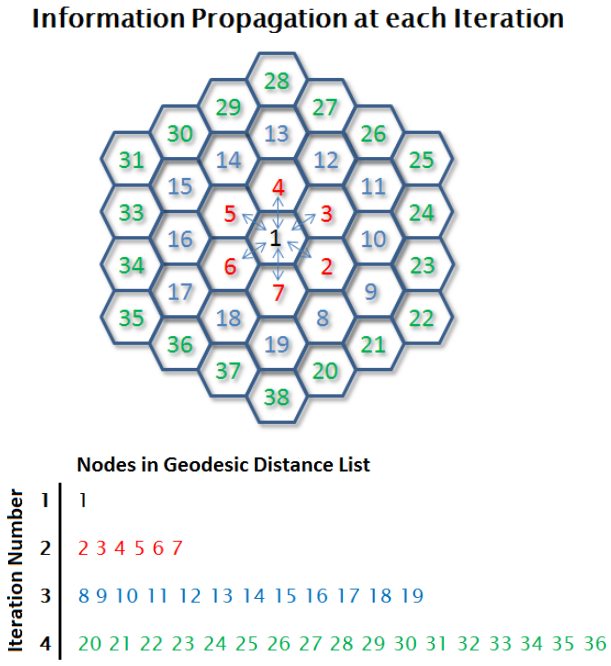
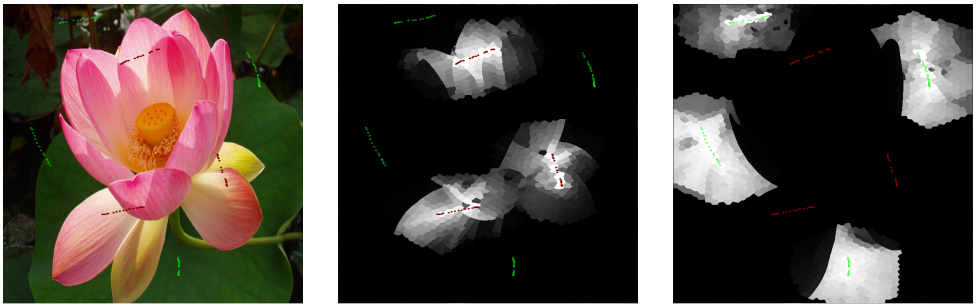


Figure 3.13: Region Coverage Increase per Iteration



(a) Input image with user scribbles

(b) Geodesic distance to object seeds

(c) Geodesic distance to background seeds

Figure 3.14: Geodesic distance to the object and background seeds after 8 iterations are shown. Red scribbles show the user inputs for the object and green scribbles show the inputs for the background.

Algorithm 1 Geodesic Distance Propagation Algorithm Pseudocode

```
1: Initialization:
2: for i = 1 → iteration number do
3:   for j = 1 → total # of nodes in image do
4:     for k = 1 → neighbor # of n_j do
5:       if n_l ∈ LUT of n_j then
6:         if dist(n_j, n_l) > dist(n_j, n_k) + dist(n_k, n_l) then
7:           update path and dist(n_j, n_l)
8:         else
9:           do nothing
10:        end if
11:      else
12:        add n_l in LUT of n_j
13:      end if
14:    end for
15:  end for
16: end for
```

explained in Algorithm Pseudo Code 1.

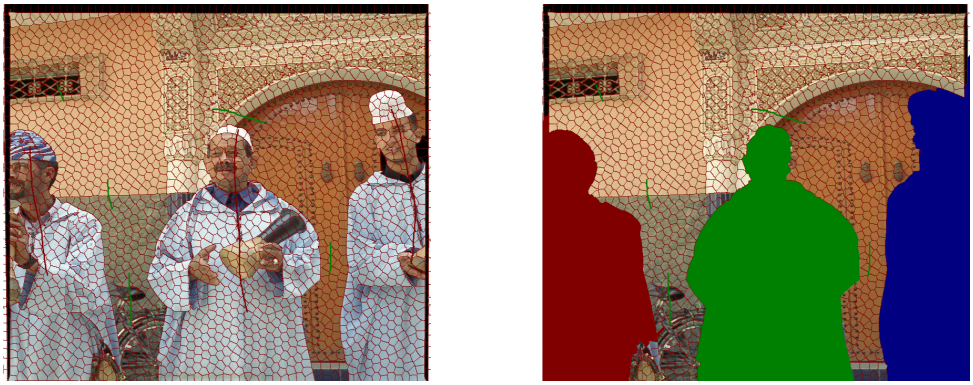
In this illustration an initial superpixel size of $20x20$ has been considered. At the end of 4^{th} iteration, a total number of 38 nodes and an average of $160x160$ pixel region is covered at this hexagonal neighborhood structure. Since the graph nodes in the segmentation framework are constructed on superpixel primitives, this method proves to be computationally very efficient. It is also useful since it offers a deterministic run time depending on the number of iterations.

Figure 3.14 presents the geodesic distances between individual superpixels to the object and background seeds. Distances are computed with the proposed implementation. The computed distances are normalized to 8 bit integer for visualization. The intensity levels indicate how close a node is to a user scribble. Figure 3.14-b shows distance to the foreground scribbles and Figure 3.14-c shows the distance to the background scribbles. This figure clearly illustrates how region information propagates through similar regions and how the object boundaries are preserved. During the simulations in this configuration, the number of iterations is limited to 8.

3.3.5 Experimental Results

Intermediate results of the SP extraction, estimated mean color intensities and geodesic distance calculations are presented in the corresponding sections for keeping the integrity of the explained methods and visual representations. In this part, some qualitative results of the final single image segmentation is presented.

Figure 3.15: Input Seeds and Resulting Segmentation



(a) Input strokes for object and bg regions

(b) Object Segments

Figure 3.16: Multiple object segmentation user inputs and the output image segmentation where each object is painted with different colors.

Figure 3.15 presents both the user selected inputs on the image foreground - background areas and the resulting segmentation. In the user assistance, covering different color regions with the scribbles is intended. This way, a better region modelling could be possible.

Quantitative results regarding the performance of the proposed segmentation method is supplied in the next section together with the stereo extension of the proposed segmentation method.

The proposed binary segmentation is further generalized to multi-label segmentation. This is realized by iteratively segmenting images into multiple regions. Input strokes are defined on the individual object locations as in Figure 3.16. User interaction is performed as follows. The red object scribbles are drawn on the first object. At the end of the interaction a new object is selected and required scribbles on the new object is applied. This is done until all the objects are selected. Finally, the background

Figure 3.17: Performance comparison; Original image, Segmentation using Euclidean distance, Segmentation using Geodesic distance

scribbles are defined on the corresponding locations. The energy assignment of the individual object regions is also performed accordingly. The input scribbles on the selected object are used as foreground, whereas the remaining scribbles are used as the background scribbles and the segmentation is performed accordingly. Visual results for input strokes and output segmentation are presented in Figure 3.16. Each object is shown with different colors in "red", "green" and "blue".

Another important aspect of the proposed framework is the utilization of geodesic distance idea in the graph cut energy optimization framework. In order to emphasize the importance of such a distance utilization, the output segmentation results are shown in Figure 3.17. The energy function assignment using these two metrics produce comparably different results in terms of segmentation performance. It has been observed that the object boundaries are well preserved when the geodesic information is enforced. The over smoothing effect widely encountered in graph cut framework is hence prevented with novel information propagation idea through neighbouring patches. The quantitative results for images in the ground truth dataset reveals the performance increase.

3.4 2D/3D Conversion

With the increased popularity and vast availability of 3D displays, media content has gained great importance, however the lack of sufficient content generation has

been a major handicap towards the popularity of 3D services. 3D movies have been shot to account for the cinema theaters; however, the wider demand for 3D services cannot be addressed due to the limited content. Introduction of the 3D TVs for the end users without a wide broadcast network has led the manufacturers to come up with intermediate solutions; conversion of standard 2D image or video footage into a domain such that at least a pseudo 3D visualization is possible. This section addresses the problem of limited 3D content availability and presents the proposed method for converting monocular image to stereo.

The 3D viewing medium can be of different types; mobile display or a 52" big display, multiview, (auto) stereoscopic (active-passive glasses) or a volumetric display can be the output medium. Each display type has its specific use areas and difficulties. As the market indicators show that the mid-big sized stereoscopic displays are gaining/has gained popularity, small sized mobile displays are predicted to follow the lead.

In the previous section, utilization of superpixel atomic structures are discussed mainly for user-assisted image segmentation purposes. This section focuses on the generation of stereo images from mono views, which basically corresponds to estimation of the depth map of an image followed by a novel view synthesis by the depth image based rendering (DIBR) method [45]. The proposed superpixel based user-assisted image segmentation is used for segmenting the object. Following that, a depth value is assigned to the selected object. Novel stereo view is synthesized considering the human visual perception characteristics. As a further reference in stereoscopic visualization and rendering issues, the studies in [58] and [145] could be visited.

At this point, background information about the human visual system is supplied and the underlying mechanisms of depth perception in a real and artificial scene are explained. This is important for understanding the motivation for the proposed 2D/3D conversion. Later in the following chapter, disparity remapping application for stereoscopic content will also benefit from the background information presented in the following section.

3.4.1 Human Visual System and 3D Perception

Human visual perception depends on various types of information to correctly grasp the geometric properties of a scene. These cues range from intensity based information to motion and maybe more importantly to previously learned structures. The visual cues may be grouped into two main streams in terms of the number of views available. Stereo cues are due to the binocular vision capability of human eyes and are mainly the most important information for 3D perception. The eyes capture two images of the scene from a slightly different viewpoint. This difference in viewpoint causes a horizontal disparity between two images and the amount of disparity is a valuable information about the depth of the scene. Convergence of eyes on the image causes

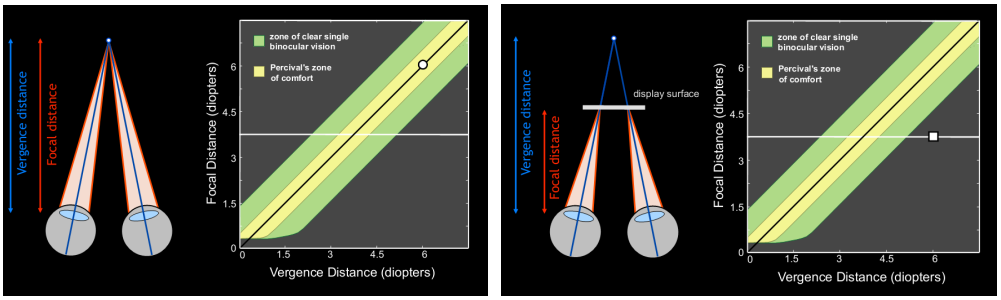
the focus of the lenses to be fixed at the interest point and hence blurring the other regions. The point of convergence and focus are matched to supply visual comfort in 3D perception [14].

On the other hand, monocular information also plays an important role in 3D perception especially when the texture of the view is not rich enough for the eyes to discover the exact disparity information. Experience based learned shapes and illumination rituals are important for easier 3D perception. Sun and home lightning appliance usually shine from above; the relative sizes of known objects (cars, people, buildings) supply a comparable learned depth perception. Linear and aerial perspectives (vanishing lines, haze due to increased distance), motion based occlusion and parallax, shade and focus/defocus are other valuable visual cues in understanding 3D formation of a scene even from a 2D photograph.

As the list of visual cues goes further; it becomes apparent how complex it is to define, model and synthetically create (using only a part of the cues like in 3D displays) the 3D perception. Visual intelligence engages a significant amount of brain activity [61], 70% of all receptors, 40% of the cortex and 4 billion neurons [138] are assigned during the vision process [30].

When a conventional stereoscopic display is considered, the main cue utilized to generate the 3D perception is the disparity between two images. However, there is a main difference between the actual and the virtually created 3D perception. In the real life, the perceived depth and the actual depth of an object is the same. In this virtually created scenario, the depth of the display is always constant, whereas the observed depth might change depending on the imposed disparity in the image. This unnatural situation is the main argument in the discussion of visual discomfort of 3D displays. The so called "accomodation-vergence conflict" is mainly due to the case that accommodative stimulus (focus of the eye) remains fixed on the screen depth, whereas convergence of the eyes adapt to the depth induced by the disparity between the image pairs. Vergence movements are required for adjusting fixation from near to far (divergence) or far to near (convergence) [7]. The stimulus creating the vergence eye movements is primarily the horizontal disparity, and the stimulus for accommodation (focus) is the perceived image blur. In a real world scene; both the retinal disparity and the blur information contribute to a change in fixation. However; in the current stereoscopic displays the disparity information is not accompanied by the required retinal blur. The accommodation of the eyes are fixed at a constant depth no matter how close or far the objects in the scene are. The following figures help visualizing the mentioned conflict.

Figure 3.18-a shows a natural viewing experience where the vergence and focal distance are the same. The yellow region in the figure is named as "Zone of Comfort" where focal and vergence cues agree up to a limit creating comfortable viewing experience. The green layer covering the comfort zone is utilized to define the limits of clear binocular



(a) Natural Viewing (b) Stereoscopic Display

Figure 3.18: Vergence vs Accommodation [61]

vision. An object located in this layer can be perceived at the intended depth but may cause discomfort when exposed continuously. Figure 3.18-b shows the case when the focal and vergence distances are different than each other. This situation generates conflicting depth cues for brain causing visual fatigue. This is the case that happens when a 3D display is viewed.

3.4.2 Related Literature

Depending on the existence of user interaction in the process, 2D/3D conversion methods can be investigated in two main streams; automatic and assisted. Automatic approaches are motivated generally for real time (on the fly) applications, where no time or effort for use interaction is available. Model based, learning based or texture based approaches are proposed for such automatic methods. Some TV producers, e.g. Samsung, Sony, Vestel; have presented 3D equipped products with automatic conversion capabilities. Although the performance of these techniques is questionable, there is a great demand and tendency in the market to offer such products. On the other hand some user-assisted methods are proposed for a better depth image characterization. Such systems can be valuable for visual studios where accurate depth generation is necessary.

Training based automatic conversion methods [110], [67] try to learn image statistics using pixel-level and mid-level cues for predicting the depth of a single image. The paper in [110] proposes MRF based depth estimation by using the ground truth data collected using 3D laser scanner. It uses a hierarchical method of Laws' texture map identification as feature vectors defined in [101]. Some papers offer techniques that totally ignore the actual depth estimation but instead, concentrate on the visual comfort [116]. This method depends on the idea that the object edge discontinues contain the highest information for 3D perception and hence, tries to find the sharpest boundaries in the image to assign higher depth values. A similar idea proposes using visual attention as a depth map. The method in [65] uses the saliency map as the depth map

for stereo image synthesis.

Motion is also another important cue for understanding the depth of a scene. The authors in [85] propose a depth sensing method based on motion parallax. It uses the initial assumption that the scene is stationary and a translational camera motion is present. Under these constraints the depth of the object corresponds to the motion between the successive frames. A more in depth analysis has been pursued in [72]-[71] where the basics of the structure from motion are utilized in order to estimate the camera parameters and sparse 3D structure. The stereo pair for each frame in video is generated by using perspective transformations; hence, estimation of dense depth map (necessary for novel view synthesis) is avoided.

Most of the automatic methods create depth map using only one or a nicely blended combination of many visual cues. However promising they are, it is always a major issue to create a solution that is robust and stable in any general content. This justifies the necessity of human interaction for accurate stereo view generation. User assisted 2D/3D conversion is performed by segmenting an object of interest from the image. Hence, the literature overlaps with the user-assisted image segmentation literature, which has been previously addressed in 3.2.

3.4.3 Proposed Method

The proposed user-assisted 2D/3D conversion method is explained in a three step approach. In the first step, user is required to select the representative locations of the image for segmentation. In the second step, the depth map of the scene is determined depending on the output object segmentation and the selected scene category. The predefined scene categories are shown in Figure 3.19. Finally, the novel stereo pair of the image is synthesized for visualization.

3.4.3.1 User Interaction

The required user assistance has been detailed previously in section 3.3.1. User is expected to interact with the semantically representative parts of the image to locate object and background regions. Figure 3.21-a presents the scribbles on the selected image.

3.4.3.2 Depth Map Generation

The following step in the 2D/3D conversion pipeline is the assignment of the depth map on the object and background segments. This can either be done automatically, or depending on the user selection. This step can be seen as depth layer ordering

Figure 3.19: Predefined depth hypothesis

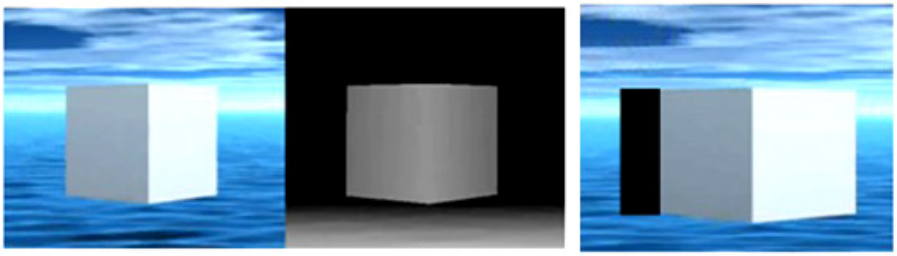


Figure 3.20: A synthetic cube image, its depth map and left image rendered with uncover region painted in black.

where user indicates the relative depths of multiple segmented objects. Background depth is filled using a depth prior. It can be a fixed depth or be assigned to change gradually depending on the scene geometry. The proposed method uses five predefined depth hypothesis as shown in Figure 3.19. In the depth assignment phase, it has to be kept in mind that the ill-posed depth estimation problem does not aim to produce the most accurate depth map, rather a perceptually comfortable, consistent and realistic image rendering is pursued. This common gradient type of depth assignment keeps smooth variations on the selected background region and hence satisfying comfortable 3D visualization. The assignment of relative (not actual) depth values for objects at different depths may be rough; however, the resulting perceptual quality of depth discrimination is quite satisfying when it is combined with the monocular depth cues at the viewer side. Figure 3.21-c shows the resulting depth map generated using the proposed method.

3.4.3.3 Depth Image Based Rendering

Rendering of the virtual view using the depth map of the original view is performed according to the geometrical relations of the scene. The amount of shift in the original RGB values of the pixels is done depending on the disparity value corresponding to the pixel. However it has to be kept in mind that possible multiple assignments of the pixels have to be handled wisely. Similarly, some regions will be left unassigned due to the horizontal shift (disoccluded areas in the virtual frame). The reason for that is the area that is occluded in one view might be visible in the other view. Figure 3.20 shows an image of a cube, its depth map and the rendered left image. The occlusion region

(a) Input scribbles

(b) Output segmented image

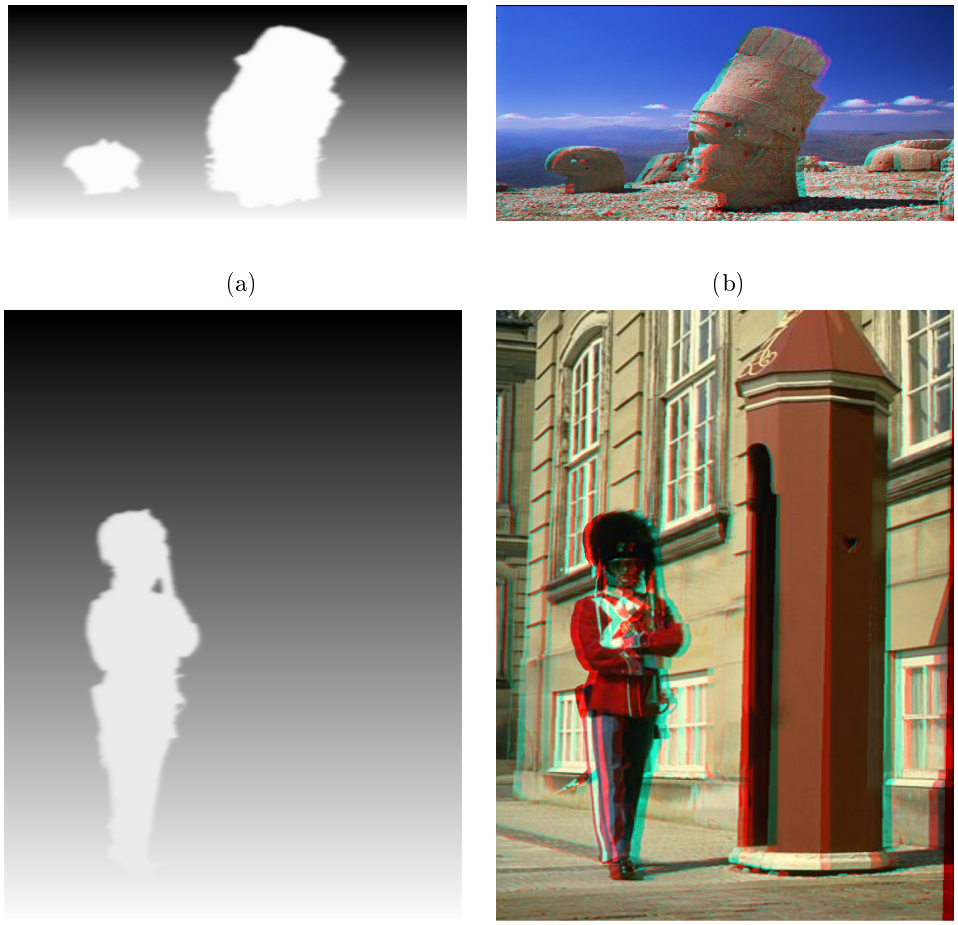
(c) Assigned depth image

(d) Generated anaglyph image

Figure 3.21: 2D/3D conversion pipeline

is shown as black since it is not visible in the original view. These regions can be filled with simple foreground/background color interpolation; moreover, background color mirroring or extrapolation methods are also used. Depending on the region texture even inpainting [33] type of algorithms may result in visible artifacts. Presmoothing of the depth map with a low pass filter has gained a major acceptance for minimization of this type of artifacts.

Figure 3.21 shows the 2D/3D conversion pipeline. User interaction as shown in Figure 3.21-a is followed by image segmentation in Figure 3.21-b. in this example the third depth hypothesis in Figure 3.19 is used for depth generation. The generated depth map in Figure 3.21-c is used to generate the stereo view that is shown in anaglyph format for visualization purposes.



(a) (b)

(c) (d)

Figure 3.22: Generated Depth Map and Stereo Images in Anaglyph Format

3.4.4 Experimental Results and Mobile Device Integration

The results of the proposed 2D/3D conversion technique is qualitatively presented. Figure 3.22 shows the depth assignment and the generated stereo images. The synthesized images are displayed in anaglyph format for visualization.

The proposed 2D/3D conversion technique is also implemented on a mobile device. For this application, the Nokia N900 mobile phone with a special 3D capable autostereoscopic display (parallax barrier type) is used. The "QT" framework is used for cross compiling on the Linux-based "Maemo" operating system. The touch screen of the mobile phone is used for user interaction with the object and the background regions. The GUI is designed such that it allows user to select different depth values (layers) on the multiple objects. Figure 3.23 shows the interface on the phone. The top menu buttons are from left to right are used to; open the image, select the segmentation

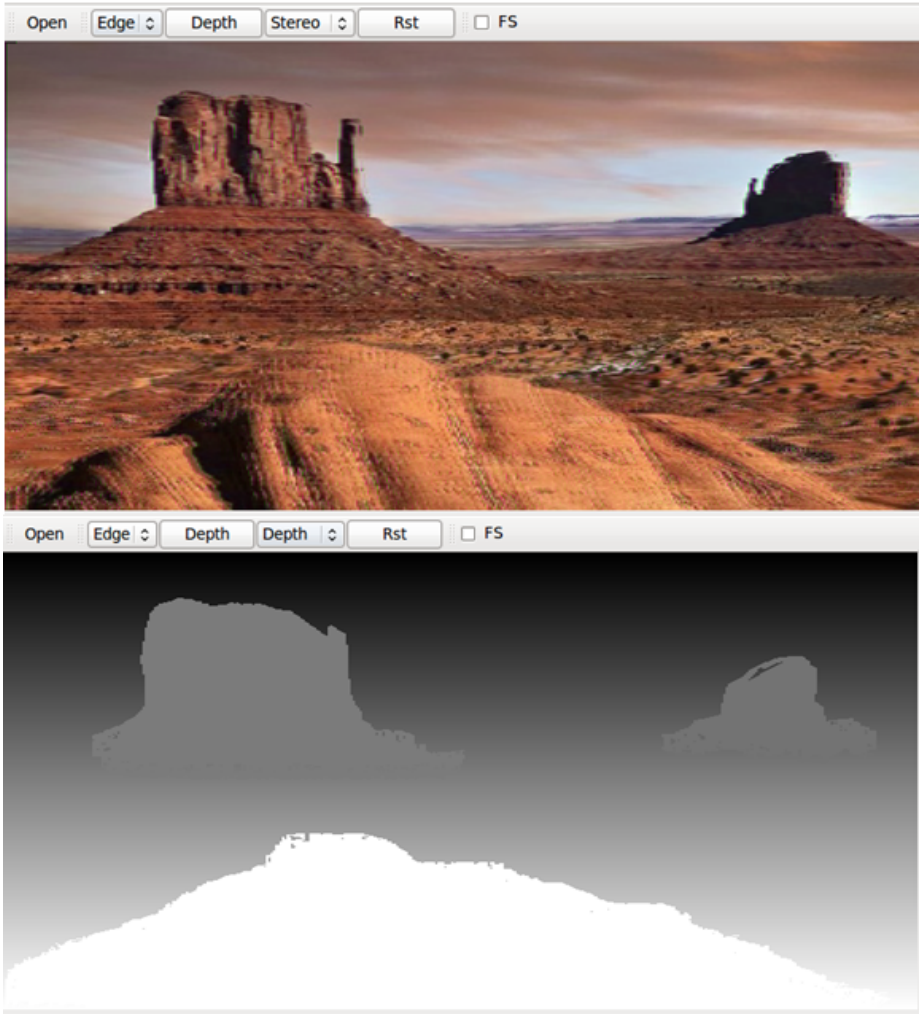


Figure 3.23: The mobile phone interface. The drop down menu on the top enable visualizing both the mono, stereo and depth images.

method, assign depth values for the selected object, select the visualization type, reset image and switch full screen by hiding the top menu.

In the virtual view rendering step, there are a couple options that are explored in the experiments. The first option is to keep the original image unchanged and render the stereo pair with the assigned depth. Another option is to render both pairs with half the original disparity. The advantage of the former is to keep one of the images original, and the advantage of the latter is to render with half the disparity and hence yielding a smaller uncover region. Depending on the user tests, the former is selected in the virtual view rendering phase.

3.5 Conclusion

This chapter presents a general purpose superpixel based user-assisted object segmentation framework with an application on 2D/3D image conversion. The segmentation framework is established using the superpixel primitives and the graph-cut energy minimization technique is used iteratively for multiple object labeling. The oversegmentation step involves a novel weight propagation phase where a similarity based object region integrity is enforced. The efficient iterative implementation of the geodesic distance metric proves to be useful for increasing final segmentation performance especially at the boundary regions of the object. Visual 2D/3D conversion results are presented for a qualitative evaluation.

CHAPTER 4

STEREO IMAGE SEGMENTATION

4.1 Introduction

Previous chapter presents the utilization of superpixels in the user assisted image segmentation framework and application on the mobile device for 2D/3D image conversion. In this chapter, the extension of the mono image segmentation on the stereo footage is further explored. As an application of the stereo image segmentation, a novel disparity remapping technique is proposed .

The energy formulation used for the stereo segmentation is an extension of the method explained in Chapter 3. The stereo extension is realized using a feature based correspondence estimation followed by the energy minimization. Similar to the mono segmentation, superpixel atomic structures are used as the nodes of the graphical framework. The proposed stereo segmentation method is used for retargeting the stereoscopic footage on different display sizes and resolutions. The performance of the proposed technique is evaluated using objective metrics on a ground truth image segmentation dataset. The qualitative user study is also conducted for evaluating the subjective performance of the proposed disparity remapping technique. In order to be complete, the extension of the idea on the video footage is also presented with some qualitative results.

The general outline of the chapter is as follows: Firstly, the related literature about user-assisted stereo image segmentation is explained. Following that, the proposed stereo image segmentation is presented in Section 4.3. The disparity remapping technique is explained later in Section 4.4. Before concluding with final remarks, extension of the proposed idea on the video data is presented in with experimental results in Section 4.5.

4.2 Related Work

This section presents a method to extend previously explained technique on mono image segmentation towards the stereo with minimum extra effort. In that sense, the input strokes on one image are used to produce segments on both image pairs. No additional input is required by the system; hence, the amount of user interaction is kept at minimum.

Previous studies for obtaining stereo image segmentation mostly utilize dense disparity estimation. With the estimated disparity, it is intended to overcome possible perception, disparity, and occlusion issues. However, the methods proposed for estimating per pixel disparities in two view stereo sometimes lack "ground truth" accuracy for arbitrary scenes. The stereoscopic copy-paste idea [82] concentrates on the same problem and offers a method to segment the selected object in stereo image by interactively merging the oversegmented regions. The regions are clustered according to a maximal-similarity merging method, and then refined by graph cut. The propagation of left eye segment on its right eye pair is realized through the disparity information corresponding to the segmented object. A recent study [98] provides details of the joint energy assignment in graph cut method for the stereo image pairs. However, the main constraint on previous methods is the necessity of dense disparity estimation which is a computationally complex step in the whole pipeline.

Since there is only limited literature about stereo image segmentation, mono segmentation techniques are also investigated and utilized in the quantitative comparison. The study in [13] proposes an interactive video segmentation method where structure from motion techniques are utilized for information propagation through the succeeding frames. However, quite long processing and interaction times cause this approach to be highly impractical. The study proposed in [99] utilizes many different cues for obtaining segmentation. Color, gradient, color adjacency, shape, temporal coherence, camera and object motion and easily-trackable points are the cues incorporated in the graph-cut optimization framework. The weighting of the cues are achieved automatically in order to boost performance using the most effective cues for segmentation. However, it also requires long execution and interaction time for the final result. In order to reduce the interaction and execution time, an efficient method is proposed where interaction with only one of the stereo pairs is required. This eliminates the computational burden of dense disparity estimation.

4.3 Proposed Method

The goal of the method is to faithfully segment object and background regions in stereoscopic image pairs. Algorithmic flow is presented in four major steps:

- Superpixel generation for graph generation [121].
- User assistance as scribbles on image representative areas [123].
- Feature matching for information propagation [122].
- Stereo segmentation via graph cut [119].

The generation of superpixel regions and user assistance on the image representation areas are previously explained in Chapters 2 and 3. Therefore, the emphasis is directed to the additional efforts towards achieving stereo segmentation.

4.3.1 Detection and Description of Feature Points

Feature point detection and generating discriminative descriptors have been widely investigated in the computer vision literature. In the proposed technique, feature matching is used for the purpose of estimating the segmented object disparity. These matched keypoints are used to transfer input scribbles supplied by the user on the stereo pair.

Selection of the feature descriptor is an important aspect of the whole process. Most of the widely known methods like SIFT [83] and SURF [16] rely on costly descriptors for detection and matching. Among the state-of-the-art methods, ORB (Oriented FAST and Rotated Brief) feature descriptor [107] steps ahead from its predecessors due to its computational efficiency and acceptable performance [107]. Another point that points out the ORB descriptor is the fact that unlike SIFT it is freely available in OpenCV in the PC and Android environment. ORB feature detector is introduced as a computationally efficient replacement to SIFT [83] that has similar matching performance, and is less affected by image noise. ORB combines the FAST keypoint detector [104] and the BRIEF descriptor [25]. The comparison of the ORB descriptors against previous art can be examined in [107]. Moreover, the study in [42] also compares the detection and tracking performance of different feature detection algorithms.

The matching of the detected keypoints is conducted using a brute-force method by comparing the binary hamming distances of the corresponding descriptors. This has proved satisfactory performance for most of the scenes. Additional epipolar constraint and average disparity thresholding has also been performed for possible outlier elimination.

4.3.1.1 Disparity Estimation

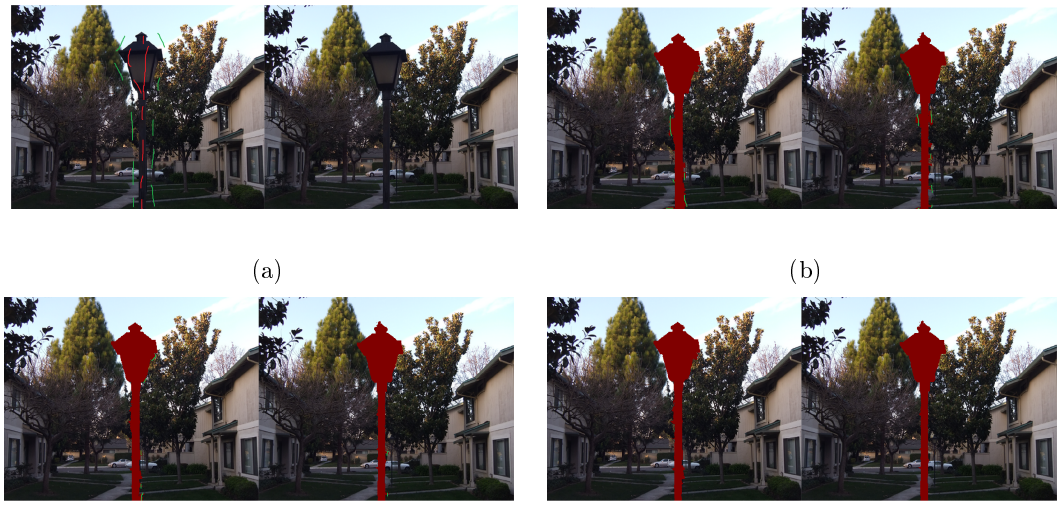
Previous methods on stereo segmentation [98], [82] require a dense disparity map for estimating pixel correspondences. However, this is a computationally complex process

Figure 4.1: ORB feature point match

and causes the application to be non real-time. For an interactive procedure, longer computational times is not tolerable for the human operator. Even with an efficient local implementation [29] the required times for this process is far from real time. The rankings of the dense disparity estimation methods in Middlebury [111] web site clearly state this fact with the quantitative experiments. Moreover, the estimated disparity values are also prone to errors even with the performance oriented global implementations [86]. Our proposed method eliminates this procedure by applying an efficient sparse feature matching idea. Stereo feature matches as shown in Figure 4.1 are used to find the average disparity of the segmented object. The estimated object disparity is used to transfer the scribbles supplied from one image to the other. In this scribble transfer, epipolar consistency is also considered. Hence, the feature matches are eliminated for possible outliers using an adaptive thresholding method. For a keypoint K_x the outlier control is done by finding the average disparity in the neighborhood N_x. It is computed by finding all the keypoints K_y and averaging the disparity sum by the number of keypoints M in the neighborhood N_x (4.1).

$$AvrDisp(N_x) = \frac{1}{M} * \sum_{K_y \in N_x} (disp(K_y)), K_y \in N_x \tag{4.1}$$

If the matched disparity of the keypoint K_x is not compatible with the computed average disparity, it is discarded. The matched image feature points are visualized in a top-bottom format in Figure 4.1 where lines connecting the feature points indicate the matching performance.



(a) (b)

(c) (d)

Figure 4.2: Proposed method enables repeatable interaction for obtaining satisfying segmentation results

4.3.1.2 Energy Minimization on the Stereo Image

At the final step of the stereo segmentation algorithm, the input seeds and their stereo correspondences are integrated in the graph-cut energy formulation as explained in Section 3.3.2. The calculated average disparity in the segmented object is used for transferring seed location information on the stereo pair. The procedure uses the idea that the segmented object area has similar disparity. Therefore, the average disparity level estimated through the sparse correspondences can be used to transfer object and background seeds on the stereo pair. The relocated input seeds are used in the binary segmentation framework to create final stereo segmentation output as shown in Figure 4.3.

4.3.1.3 Additional User Input

In the proposed framework, the user is allowed to add additional inputs to the system for a better performance. When the resulting segmentation requires additional adjustments, the user might further indicate erroneously segmented regions on the image as in the initial phase. Left click of the mouse is used to indicate foreground object and right click to indicate the background region that is erroneously segmented at the initial stage. This step clearly enhances the final segmentation performance, but causes an additional user interaction effort and time. In the results section, we have indicated the resulting increase in segmentation performance as well as the time required for the additional input step.

Figure 4.2 presents a scenario where initial inputs do not produce a satisfactory segmentation output. Figure 4.2-a shows the initial seed assignment. The output segmentation by the proposed graph cut method is shown in Figure 4.2-b. On the same figure, one can observe the additional user strokes especially at the low gradient object boundary. This procedure can be repeated until user is satisfied by the result.

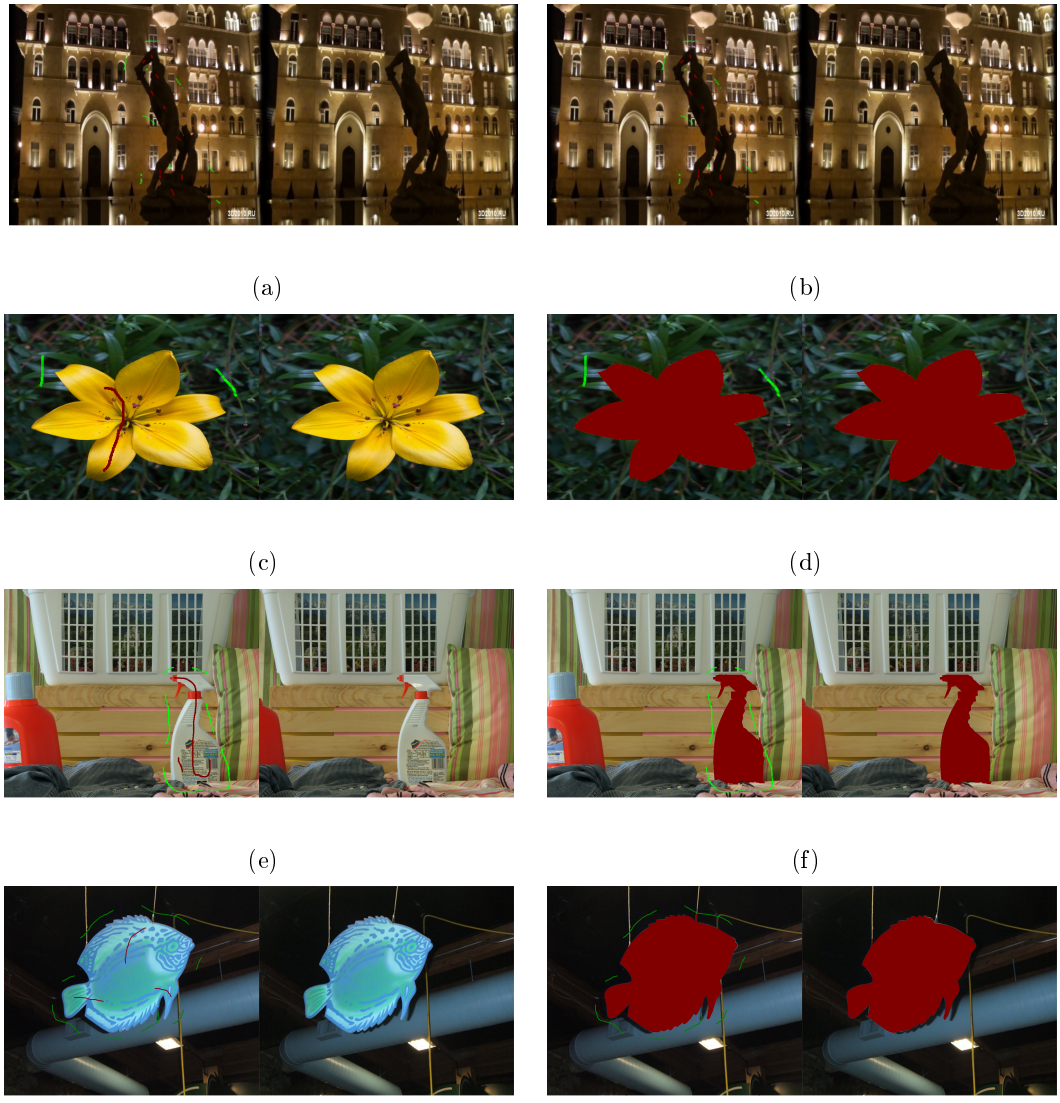
4.3.2 Experimental Results

The performance of the proposed stereo segmentation technique has been quantitatively evaluated. A ground truth dataset [98] containing binary stereo segmentation results has been used for the tests. The dataset contains 30 stereo images and some of them are from Middlebury dataset [111]. The images are labeled as foreground, background and unknown regions. Unknown regions correspond to object boundaries, where it is hard to accurately decide between foreground and background; i.e. hair region around the head of a person. Stereo segmentation results have been tested towards segmentation accuracy which measures the ratio of the correctly labeled pixels over the total number of pixels.

The proposed segmentation algorithm is compared with the state-of-the-art methods Livecut [99], SnapCut [13] and StereoCut [98]. These methods are selected due to the utilization of similar scribble based user assistance for segmentation. Average segmentation error of the proposed method with limited and extensive user interaction for Euclidean and geodesic distance metric have been computed.

Any interactive system can be repeatedly tuned by supplying more user inputs in order to produce satisfying results. However, the aim is to keep the time of user assistance at minimum. We firmly believe that the true performance of a user assisted segmentation technique cannot be evaluated by pure segmentation results; amount of interactions should also be considered. Hence, the time required for obtaining proposed segmentation results is recorded. On the average, the proposed system requires less than one minute per image including user interaction and CPU processing time. Approximately 3 seconds for preprocessing (including superpixel generation and sparse feature matching) and 50 ms for graph cut optimization is recorded for a very high resolution ($1920x1080$) stereo image on a 3.06 GHz PC. Moreover, it should also be noted that the method in [98] strictly requires a dense disparity estimation in order to obtain a stereo segmentation. Time required to estimate the disparity takes more than one minute with their utilized method [48] at this resolution. Therefore, the proposed solution with sparse feature matches is computationally much feasible compared to such a dense disparity estimation method.

Average segmentation error ratios for the images in dataset are shown in Table 4.1. The table also presents the required time for human interaction and computer processing. There are multiple conclusions that can be derived from the quantitative



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 4.3: Input scribbles and proposed stereo segmentation

analysis. Firstly, utilization of geodesic distance in the similarity cost enhances the segmentation performance. It has been previously presented in Section 3, Figure 3.17 for a visual comparison. Secondly, proposed method with sparse feature matches is computationally much feasible compared to a dense disparity estimation approach. With the standard or additional user assistance, satisfying results are obtained in minimal processing and interaction time. Considering the objective metrics only, it can be concluded that the proposed technique provides competitive performance with respect to the state-of-the-art methods. Visual stereo segmentation results are presented in Figure 4.3.

Table4.1: Average segmentation error

Methods	Label Error %	Process Time	Interaction Time
Livecut [99]	1.07	$> 5sec$	∼1 min
Snapcut [13]	0.37	$> 5sec$	∼1 min
StereoCut [98]	0.31	$> 60sec$	∼1 min
Proposed Euclidean	0.39	∼3 sec	∼30 sec
Proposed Geodesic	0.32	∼3 sec	∼30 sec
With additional input	0.26	∼3 sec	∼1 min

4.4 Disparity Remapping

This section explains how the proposed stereo image segmentation technique is applied as a post processing step in a stereo content generation pipeline, also presented in our paper in [120]. The basic layout and formation of the stereoscopic content production setup is seen as easy as placing two cameras side by side; however, the virtue of creating natural and comfortable viewing experience lies in correctly assembling the required settings for the geometry of the scene, specific display medium and viewer placement during the show. The ambiguity in the display side; namely the fact that the output display size and resolution cannot be predicted during the shooting; is the main challenge in the production step. The produced content can tolerate only a limited amount of deterioration from the actual target size and resolution. This difficulty in the production step motivated us to attack the problem of adjusting stereoscopic content for retargeting purposes. This makes visually appealing 3D perception possible in multiple output medium types other than the targeted size and resolution.

Human visual system depends on various cues while perceiving the depth of a scene. Horizontal disparity information between the two retinal images is processed by the brain to produce a single vision and stereoscopic depth [94]. In order to achieve this, eye movements are required to position the image onto the fovea. Vergence movements are required for adjusting fixation from near to far (divergence) or far to near (convergence) [7].

When an object is very close to the camera during shooting, it is displayed with great negative disparity ($> 10\%$). This may result in an uncomfortable viewing experience and possibly a temporary diplopia, a problem that happens when two stereoscopic pairs cannot be combined by the brain due to the excessive disparity. Hence, the amount of disparity should be well optimized for the display medium [61]. It is not convenient to apply a stereoscopic content optimized for a specific type of display (e.g. large theater screen) on a different target medium (e.g. a standard 42" TV or a 4" mobile display). The reason is due to the fact that it might create either unimpressive (cardboard effect) or uncomfortable (diplopia) viewing experience due to the mismatched disparity values. However, directors or visual artists (namely, stereographers) might willingly impose excessive negative disparity on the objects mostly for a very short time to create an impressive impact. On the contrary, it might be the case that an object that deserves a specific attention is shot with a configuration that no impressive depth differentiation is present. The director in such a case might prefer to emphasize the object which is originally located on a similar depth with the background. The proposed method can be used in such a scenario to impose a synthetic depth discontinuity between the object and the background in order to create an artistic impact.

The mentioned various possible issues and scenarios cannot be handled simultaneously during the production phase. In other words, it is not possible to create a general solution that gives optimum quality in different viewing conditions. Hence, post processing techniques, as proposed in this section, are developed to recreate (synthesize) stereoscopic scenes for a better visual experience.

4.4.1 Related Work

Image retargeting topic has been previously addressed in the fields other than stereoscopy. The study in [113] proposes a method for warping mono images adaptively on different output formats. A comparative study on monoscopic image retargeting applications can be examined in [106]. In the case of stereoscopic content, image retargeting aims at virtually adjusting the perceived depth of the image on the screen so that no visual fatigue or visual conflict is observed.

The study in [73] proposes a cost based automatic disparity remapping idea utilizing the perceptual saliency cues extracted from the scene. This method approaches the problem as a warping issue and is also of practical importance since it does not require a conventional disparity estimation but a sparse correspondence analysis. The usefulness of the results are validated with objective and subjective tests. Another warping based method is also presented in [143]. This paper proposes a linear mapping method to adjust the depth range of a stereoscopic video according to the viewing configuration. Display size, pixel density and viewer distance are considered for the final warped image. The main limitation is stated as possible warping issues in case of small objects in the scene.

Another direction of disparity mapping methods require estimation of dense disparity map. The study in [140] applies a non linear mapping operation on the estimated disparity values. The mapping formulation is motivated by the basic gamma curve mapping idea for visual representation of images on displays. This study also refers to subjective tests for validating increase in visual comfort after the remapping operation. However, the methods proposed for estimating per pixel disparities in two view stereo sometimes lack "ground truth" accuracy for arbitrary scenes. The steroscopic copy-paste idea [82] concentrates on the same problem and offers a method to segment the selected object in stereo image by interactively merging the oversegmented regions. The regions are clustered according to a maximal-similarity merging method, and then refined by graph cut. The propagation of left eye segment on its right eye pair is realized through the disparity information corresponding to the segmented object. The geometric perspectives of remapping operation has been detailed by Devernay and Duchene [38], where baseline and viewpoint modification aspects are covered and a hybrid method is proposed. It also supplies background information about the reasons of observed visual deteriorations which mainly result from erroneous acquisition of the scene.

Since the problem is well suited for post processing applications, commercial products are also available in the market. The product released by Foundry [2] requires geometric calibrations, such as camera parameters, followed by novel view synthesis and occlusion handling. Another method utilizing user interaction for the modification of the 3D morphology is presented in [135].

4.4.2 Proposed Technique

The proposed method allows modification of the depth of an individual object by moving it virtually closer to or further away from the camera. This is done using the previously explained stereo segmentation output. The segmented object on the stereo image is horizontally moved in order to achieve the intended virtual depth adjustment. The direction of movement is defined by the user depending on the intended visual effect.

Novel View Synthesis The motivation behind moving an object closer to the camera is to create a more appealing effect on the viewers by differentiating the selected object from its background. This effect can be preferred by the post processing artists, when a special emphasis on the object is required. The second case, where selected object is pushed further away from the camera, might be useful when the stereoscopic setup is not properly configured for the target viewing conditions. Excessive negative disparity might prevent the scene from being comfortably viewed. Proposed virtual depth adjustment might heal possible misconfiguration and enhance the perceptual quality of 3D vision on the target medium.

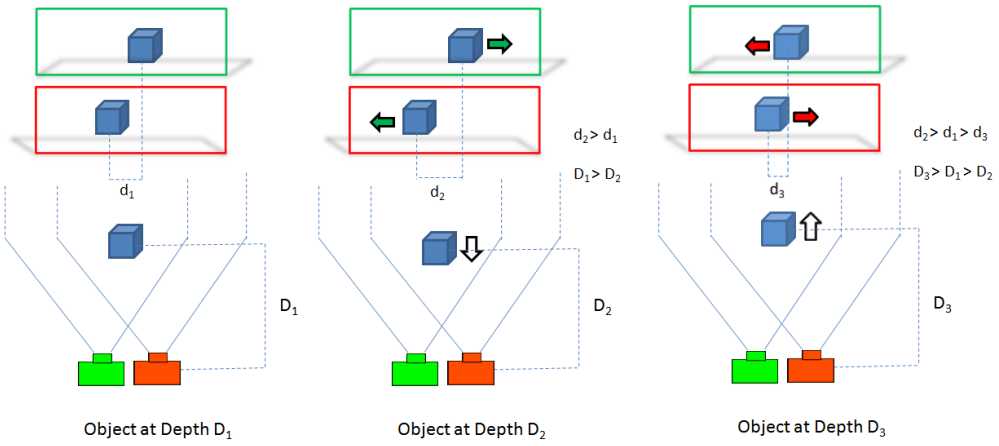


Figure 4.4: When the object is moved closer to (further away from) the camera, ($D_3 > D_1 > D_2$), disparity of the object increases (decreases) ($d_2 > d_1 > d_3$).

In order to synthesize the disparity altered views, it is necessary to generate horizontally shifted versions of the segmented stereo objects. Virtual move of the segmented objects in camera direction, forward or backward, is only possible by moving the segmented objects horizontally in reverse directions. For a side-by-side image, where left eye image is located on the left side as in Figure 4.5, moving the stereo objects towards each other will create a perception of moving the object closer to the camera, and similarly it is enough to move objects apart from each other to push objects towards the background. During the horizontal move, one of the images can be kept as it is, and the stereo pair is moved with the required disparity. However, we prefer to modify both of the stereo pairs. The reason is due to the fact that, with this setup the disparity shift applied per image will be half, compared to the former case. This minimizes the deformation and visual artifacts on both images as also proposed in [73].

The movement of the segmented objects and the corresponding disparity and depth changes are illustrated by a parallel camera setup in Figure 4.4. The inverse relation between disparity and depth is presented geometrically. (d_i, D_i) pair corresponds to the disparity and depth values of the segmented image where $i \in 1, 2, 3$ successively corresponds to the original object location, object moved closer to the camera and object moved further away from the camera. Horizontally shifted views are shown in the results section in Figures 4.5-c and 4.5-d.

Due to the horizontal movement of segmented objects, some parts of the original image are covered by the moved object and similarly some uncover regions arise at the opposite side of the moved object boundary. Missing information of the uncover regions are filled from the background with conventional inpainting type methods [33]. Object borders are usually difficult to define with the pixel precision. Therefore, image matting methods have been proposed to overcome such image synthesis tasks [59]. In

(a) (b)

(c) (d)

(e) (f)

Figure 4.5: The corresponding outputs at the intermediate steps of the algorithm

the proposed technique, this problem is addressed by smoothing the disparity at the object boundaries. This approach would cause the boundary pixels to be shifted with a slightly different disparity. With this configuration, a smooth boundary inpainting is performed. Furthermore, the recently introduced stereo inpainting method could also be utilized for a seamless hole filling operation [89]. Uncover regions and resulting region filling operation after disparity remapping operation are presented in Figure 4.5-e and Figure 4.5-f.

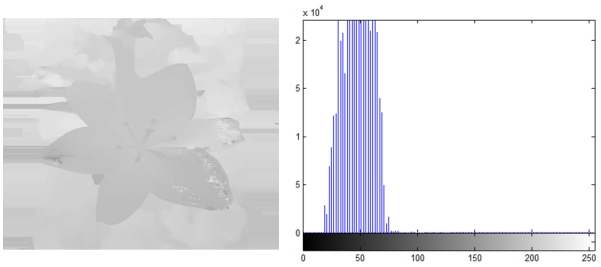
In order to present the effect of the proposed disparity remapping technique, the resulting disparity map of the generated stereo images are shown in Figure 4.6. The method in [28] is used for estimating the disparity. The disparity histograms of the remapped images are also illustrated for tracking the changes between the disparity altered images.

4.4.3 Experimental Results

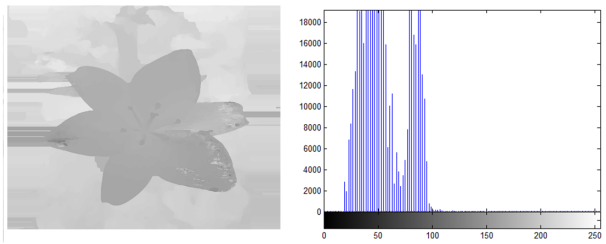
The performance of the proposed disparity remapping technique has been qualitatively evaluated. 12 set of images from are selected for subjective evaluation of the disparity remapping operation. Each set is composed of one original and two disparity altered versions. Images used in subjective tests are selected depending on the initial disparity configuration from the stereo segmentation dataset [98], Middlebury and web.

Subjective tests are conducted on a 42" 3D LED TV with 1080x1920 resolution with actual height and width of 53x94 cm. The participants are seated at a 3.2m distance which is approximately 6 times the picture height. Input stereo images are of the same resolution with the display; hence, no resizing operation is performed at the TV scaler. TV has a stereoscopic pattern retarder type display which uses circularly polarized passive glasses. Each odd and even pixel lines correspond to left and right views of the image and they are paired with the polarization direction of the glasses. Detailed analysis on different types of stereoscopic displays can be obtained in the following study [63].

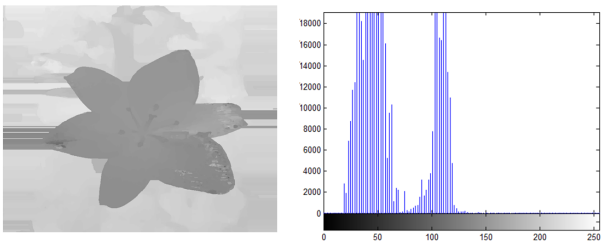
The human based experiments have been performed in accordance with the ITU-R BT.500-11 [4]. The recommendations for the subjective assessment of stereoscopic television pictures has also been considered [6]. 18 non-expert subjects participated during the qualitative evaluation tests. The ages of the participants ranged from 22 to 35 years. Those who normally required optical correction kept their glasses in addition to the circularly polarized stereoscopic glasses. No information regarding the experimental hypothesis have been shared with the participants. They have only been asked to evaluate the randomly ordered stereoscopic content according the given criteria. The test images are grouped under three categories; 1) Control group, 2) Object very close to camera, 3) Image with limited disparity range. The images in the 'control group' are selected such that no visual difficulty is present and a clear 3D



(a) Object moved backward



(b) Object at original location



(c) Object moved forward

Figure 4.6: Disparity estimate and disparity histogram of original and remapped stereo images

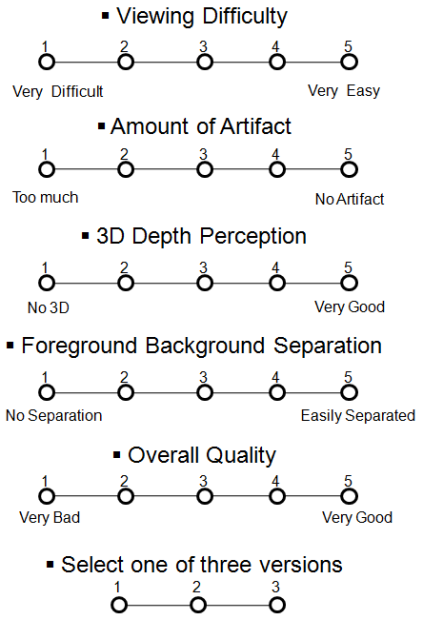


Figure 4.7: Likert scale and evaluation criteria

perception is satisfied. The disparity range of images in this category is limited to +5% of the image width where no visual fatigue is expected [61]. The second category contains the images where the object of interest is quite close to the camera. In this category, negative disparity of the selected object is close to or larger than 10% of the image width where convergence problem is expected. The final category consisted of images with quite limited disparity range (less than 2% of image width) where salient object is weakly discriminated from the rest of the background.

Each set of 12 images contained one original and two disparity adjusted stereo versions. Disparity remapping is performed over the geodesic distance based segmentation results. In the first category, selected object is moved closer and further away from the camera (3% increase and decrease in the object disparity compared to image width) within the visual comfort zone as shown in Figure 4.8-a. The second category images are processed so that the object is moved two steps further away from the camera (3% and 6% decrease in the object disparity) inside the comfort zone as shown in Figure 4.8-b. The selected objects in the final category which are tagged as having limited disparity range are moved two steps forward (3% and 6% increase in the object disparity) in the comfort zone as shown in Figure 4.8-c. The red circle in the figures correspond to the selected object. Notice that remaining scene is kept unchanged in terms of the perceived depth. These three versions of 12 image sets constitute the 36 image dataset for subjective evaluation. Each category have equal number of 4 examples in the dataset.

Table4.2: Mean opinion scores of the subjective evaluation

First Category: Normal Viewing Case						
	Q1	Q2	Q3	Q4	Q5	Q6
Original	3.93	3.96	3.92	3.9	3.8	48%
Forw. 3%	3.17	3.38	3.38	3.68	3.15	13%
Backw. 3%	3.83	3.71	3.58	3.7	3.68	39%

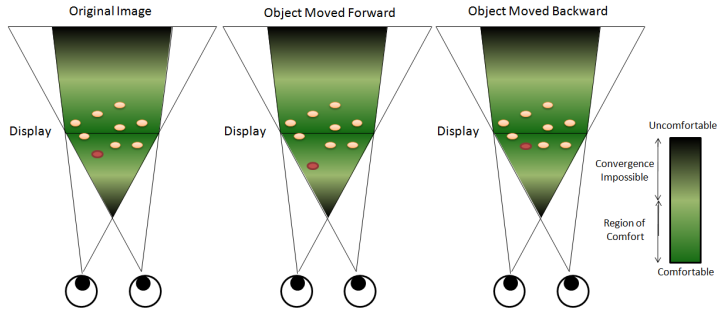
Second Category: Object quite Close to Camera						
	Q1	Q2	Q3	Q4	Q5	Q6
Original	2.62	3.32	2.69	2.82	2.42	22%
Backw. 3%	3.02	3.25	2.86	2.92	2.86	46%
Backw. 6%	2.86	3.1	2.52	2.42	2.55	32%

Third Category: Object with Limited Disparity						
	Q1	Q2	Q3	Q4	Q5	Q6
Original	3.62	3.76	3.66	3.76	3.33	33%
Forw. 3%	3.68	3.42	3.74	3.80	3.42	27%
Forw. 6%	3.46	3.33	3.77	3.94	3.39	40%

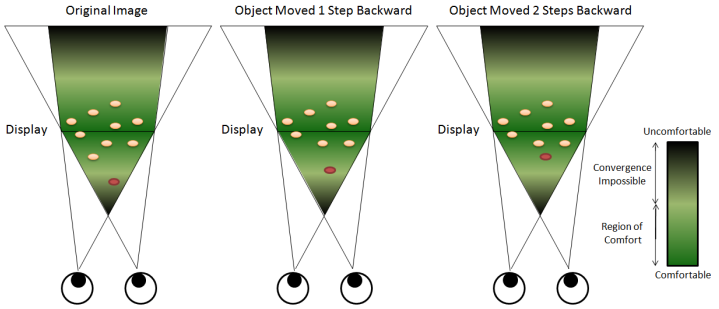
The test runs around 15-20 minutes depending on the participant. Participant is provided with the control of the slide show. He/She could go back and forth in the dataset as much as he/she wants and no time limitation has been pressured. For the given five criteria, the subjects rated the images in a 5 point Likert scale [80], where 1 indicates worst and 5, the best grade. Figure 4.7 shows the Likert evaluation criteria used in the test. A final question indicates a general selection between the three versions of each set of images. The first five questions are answered for each image and the final question is answered for each set of three images.

The question regarding viewing difficulty measures the observed eye-strain while visualizing the content. The second criteria about observed amount of artifact is important for evaluating the rendering performance especially at the occlusion boundaries. Third question regarding depth perception for various cases, is a key element for assessment of the proposed method hypothesis. The fourth question, Fg/Bg separation, measures the amount of separation between the object and background, some specific cases are covered with this question in which object of interest with clear depth perception is hardly discriminated from the background. The overall quality is also rated in question 5 and that depends totally on the participants own subjective criteria. Finally, a selection between the three versions of the image set is requested.

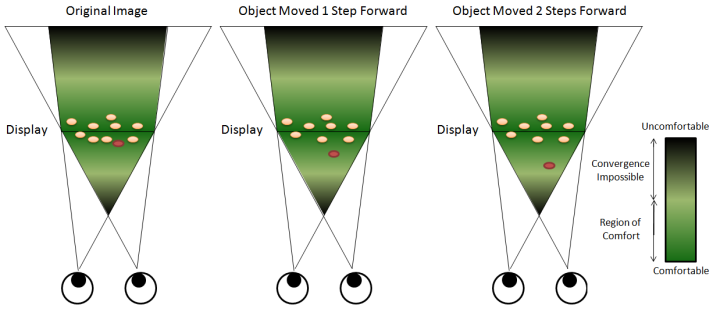
The results for each category are evaluated using the Mean Opinion Scores (MOS). Computed MOS for each category are supplied in Table 4.2. The first five questions record the ratings from a 1 to 5 scale as shown in Figure 4.7. The final question (Q6) corresponds to the selection percentage. It indicates the preference of the image among three alternatives. MOS for the first category indicate that the original images for an appropriate disparity range is chosen with higher percentage (47%) compared to the



(a) Normal viewing case



(b) Object quite close to camera



(c) Object with limited disparity

Figure 4.8: Different viewing scenarios that have been utilized during the subjective tests

processed images. This result is expected since the images in the control group are classified as having decent 3D perception and post processing of the object by moving virtually in forward or backward direction creates visual artifact (Q2) and considerable viewing difficulty, especially in forward move case (Q1). Having a control group in such a visual experiment is highly crucial in order to judge the validity of the results. MOS for the control group images has defined confidence on the subjective evaluation of the participants.

The second category consists of images that are shot with inappropriate negative disparity, and hence, the scene creates an undesired convergence causing visual discomfort as it can be observed from the results of Q1. The visual discomfort considerably decreases as the object is virtually moved back on to the display depth. On the other hand, perceived visual artifact increases with increased horizontal movement of the segmented object. MOS for the third question regarding depth perception reveals useful information. The score of the original image is quite low due to the wrong camera-display configuration. Perception score increases as the object is moved 3% backwards. However, the score decreases again when the object is moved 6% backwards. Ideally the perception scores are expected to rise as the object is moved further away from the camera with the decreased visual fatigue. One possible reasoning behind such a result is the increased visual artifact with increased occlusion which diminishes the quality of 3D perception. Similar characteristics have been observed for Q4 rating foreground and background separation. The overall quality and selection rates indicate a serious tendency towards the images with 3% backward move.

The final category containing limited disparity images are processed such that the selected object is moved closer to the camera. The scenario is shown in Figure 4.8-c. The results indicate that the images with 3% forward move are ranked first according to the viewing difficulty metric. The observed visual artifact tends to increase as the amount of object movement increases. MOS of the third question is a good indication of the success of the proposed method. This conclusion is also supported by the rankings of foreground - background separation. Overall quality rankings (Q5) indicates the superiority of the 3% forward move. It can be argued that 3% forward move is an optimum compromise considering visual artifact and depth perception. Although the resulting selection of images seems to contrast the overall quality ratings, the participants seem to be inclined to the scenes where objects are closer to the camera despite the observed visual artifact.

Figure 4.9 shows the disparity altered images in anaglyph format for a visual understanding. All the original and disparity remapped images used in the tests can be reached in the authors' web site [1] in the side-by-side format.

[1] http://emrahtasli.com/DispRemap.html

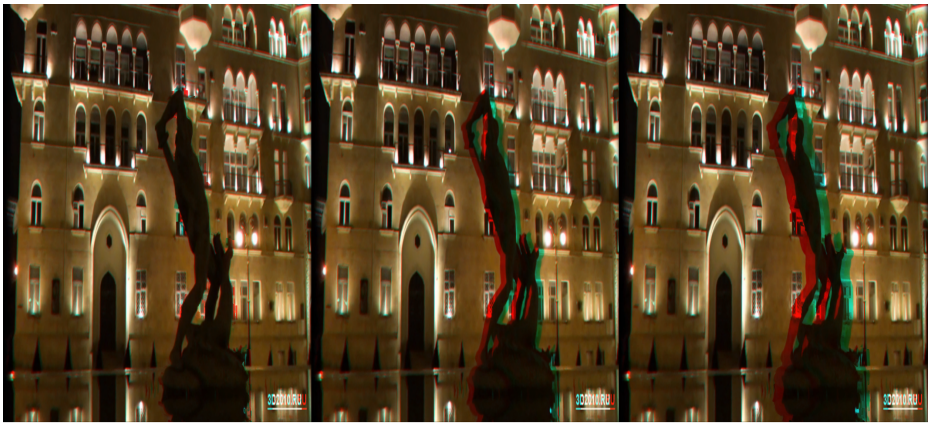


Figure 4.9: Disparity altered images shown in anaglyph format. Left: Selected object moved backward. Middle: Original stereo image. Right: Selected object moved Forward.

4.4.4 Discussions

Looking back at Table 4.1 that presents the segmentation accuracy and required time for human interaction and computer processing, it is observed that the average segmentation error decreases from 0.39 % to 0.32% as a result of the utilization of geodesic metric. In addition to the proposed geodesic metric, an additional user input scenario is also incorporated for evaluating segmentation accuracy as explained in Section 4.3.1.3. When the user is allowed to further interact with the system and correct possible erroneously segmented regions as in Figure 4.2, one can obtain superior segmentation accuracy as shown in Table 4.1. The second important conclusion that can be derived from Table 4.1 is regarding the computational complexity of the proposed method. By avoiding the dense disparity estimation, our proposed method outperforms the compared state-of-the-art in terms of computation time.

Possible limitations of the proposed method in different stages of the pipeline can be stated as follows. In the monocular segmentation step, low gradient region boundaries are prone to be erroneously segmented by the superpixels. Low intensity changes on the object boundaries can be inaccurately segmented by the generated superpixels. Dividing superpixels into smaller regions to compensate for such inconsistencies can be addressed as a future direction. Possible limitation regarding the stereo extension might be inaccurate matching of feature points on the stereo correspondences. However, this has not been observed as a real concern, since stereo matching is performed in a confined neighborhood with the epipolar constraint. View synthesis has been performed by using single image inpainting methods. This can be further improved by the recently introduced stereo inpainting method [89].

User study has shown a clear preference towards the proposed remapped images. More-

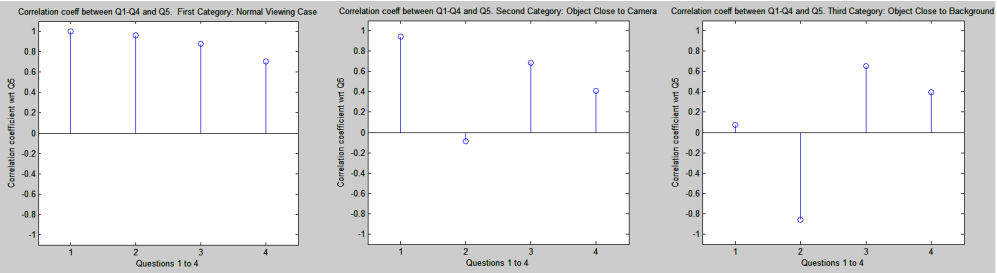


Figure 4.10: Correlation coefficient values between question 5 and questions 1-4 reveal further information regarding the selection criteria during subjective evaluations.

over, a detailed analysis about correlation between each question and observed overall quality reveals further information. Q3 and Q4 shows similar correlation in all of the cases. This proposes that depth perception and foreground-background separation is a major criteria in evaluating overall quality of the image. Q2 (Viewing difficulty) shows strong correlation in the first and second case; however, correlation decreases in the third case. This indicates that in the third category where object is too close to the background, the overall quality is almost independent of the observed viewing difficulty. Q2 (Amount of Artifact) shows strong positive correlation in the first case, very low correlation in the second case and strong negative correlation in the third case. These results indicate that people evaluate quality depending on more Q3 and Q4, despite there is great amount of observed artifact in the image.

4.5 Video Extension

The user assisted image segmentation has been explored previously for mono and stereo footage. In order to complete the analysis, the extension of the method on video is investigated in this section. Proposed video segmentation method utilizes the segmentation of the object in the first frame with the help of user assistance. As soon as the first frame is segmented with a satisfactory accuracy, the object and background region characteristics of the image are estimated with analysis on the segmented regions. The estimated object and background model is further used for processing the succeeding frames. Superpixel primitives that are used in the initial step are also generated on each succeeding frame and a similar superpixel based image segmentation is performed on the rest of the video. In this section, a superpixel based region modelling has been investigated for the target foreground - background region identification. For this purpose, a novel a superpixel based feature descriptor is proposed, which in fact is one of the first feature descriptors defined on the irregular superpixel lattice.

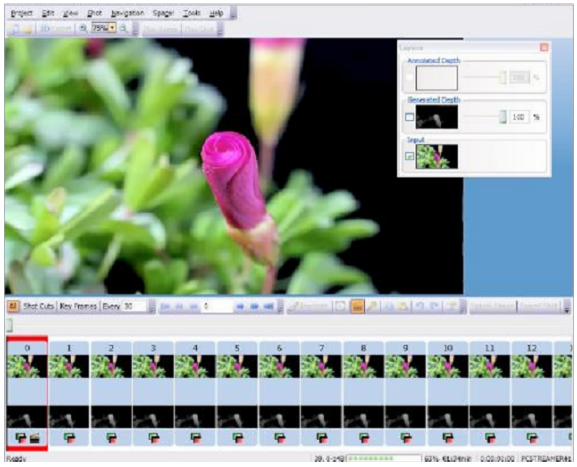


Figure 4.11: Philips BlueBox video segmentation & depth generation tool graphical user interface

4.5.1 Related Work

There are numerous methods addressing the video segmentation problem utilizing user assistance. Most of the pixel based methods still suffer from disturbingly long computation times before obtaining the final segmentation result. In most of the methods, it takes up to hours to process an average resolution one minute video. Therefore, this service is provided by the professional art studios at reasonably high prices. Philips Electronics has offered a commercial product under the name *BlueBox* [5] that enables interactive object segmentation on video footage. The primary aim of this product is to generate depth information from mono and stereo video for Philips autostereoscopic displays. *BlueBox* is offered as a service, coming with client tools and a dedicated hardware. An image from the user interface of the offered tool is shown in Figure 4.11.

The paper by Agarwala et al. [10] proposes a boundary tracking based video segmentation using splines that follow object boundaries between keyframes. The method uses both boundary color and shape cues for region identification. This method fails especially in the case that single type of cue is insufficient to select the object. On the other hand, the study proposed in [99] utilizes many different cues for obtaining interactive video segmentation. Color, gradient, color adjacency, shape, temporal coherence, camera and object motion and easily detectable points are the cues incorporated in the graph-cut optimization framework. The cue weighting is done automatically in order to boost performance using the most effective cues for segmentation. This study also suffers from very long pre and post processing times for the final segmentation. Up to 30 minutes are required to process a 100 frame footage. This makes this system totally impractical when a longer footage is to be segmented. Figure 4.12 shows the different cues used in the system.

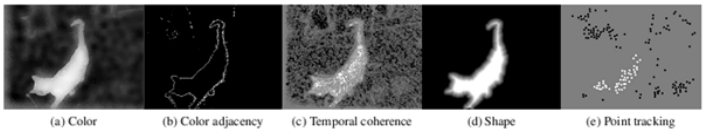


Figure 4.12: Different cues inherited in the system [99]

The paper proposed by Li et al.[78], applies a 3D segmentation approach on spatio-temporal volume by a tracking-based local refinement. Users are required to segment every tenth frame, and graph cut computes the selection between the frames using global color models from the key-frames. Watershed based oversegment regions are partitioned as foreground and background. This method requires a great deal of manual assistance on many frames in addition to corrections. Proposed 3D segmentation graph energy assignment is shown in Figure 4.13.

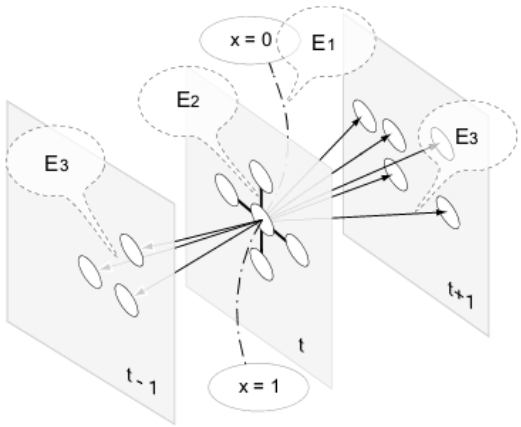


Figure 4.13: 3D graph cut energy assignment [78]

Interactive video cut-out, [136], presents a system where user draws scribbles in 3D space. A hierarchical mean-shift preprocessing is employed to cluster pixels into super-nodes, which greatly reduces the computation of the min-cut problem. Three step approach is proposed; 1) hierarchical mean-shift as preprocessing, 2) interactive user interaction and a global min-cut optimization, 3) a local min-cut optimization is done to refine the final foreground boundary. This method again suffers from high amount of time required to complete the proposed segmentation. Only the preprocessing time for a 720x480 resolution image of 200 frames takes up to 30 minutes. An example fo 3D user interaction is shown in Figure 4.14.

A learning based technique proposed in [13] utilizes collaboration of a set of local classifiers, each adaptively integrating multiple local image features for interactive segmentation. Local classifiers on object boundary are propagated onto the next frame by motion estimation. Local classification results are used to generate a foreground probability map. Video matting technique is used for final segmentation. No detailed

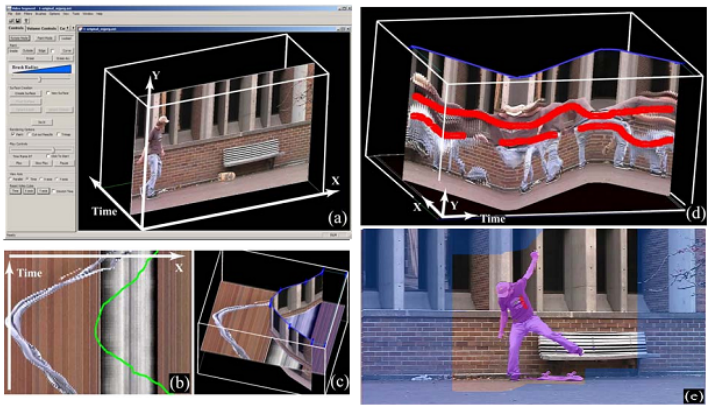


Figure 4.14: 3D User interaction for video segmentation [136]

analysis is supplied for the general time complexity, however less than 10 seconds has been proposed per medium resolution frame. Motion blur and defocus are named as the limitations of the method for defining accurate object boundaries.

A 2D-3D conversion method using video segmentation is also addressed in this section. The study in [137] uses structure from motion techniques for propagating initial frame segmentation on the succeeding frames. As an application, 2D-3D conversion has been proposed where predefined depth templates are inserted on the segmented objects. However, very long processing and interaction times (on the average more than one minute per frame) cause this approach to be impractical.

A recent automatic video segmentation method proposed by [75] starts with detection of object like regions (key-segments) on the unlabeled video. These regions are scored according to shape and motion cues for clustering purposes. Final segmentation is done according to the model learned from the key frames. This method can perform well only if the required cues are available in the given video.

4.5.2 Proposed Method

The proposed method utilizes a superpixel based video segmentation. In this technique, first frame of the video footage is interactively segmented using the previously explained graph cut framework. Geodesic distance information is used in the local energy assignment. Additional interaction is allowed until the user is satisfied with the final segmentation. Since there is considerable coherency between two succeeding frames of a video footage, the proposed system utilizes this coherency by defining a superpixel based region classification. The proposed method can be explained in four steps.

• Superpixel based feature descriptor extraction.

- Classification model training for fg/bg regions.
- Superpixel region confidence estimation.
- Global optimization using region confidences.

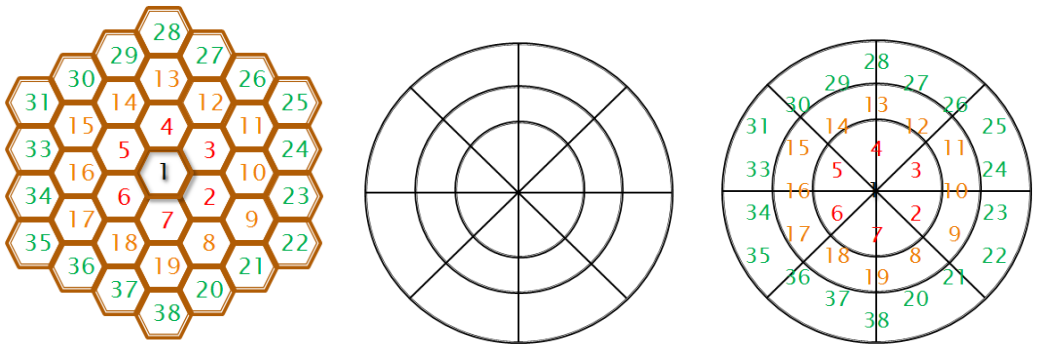
4.5.2.1 Superpixel Feature Descriptor Extraction

Superpixel (SP) based feature generation has been previously addressed for image annotation [124] and image retrieval [133] purposes. The paper in [124] proposes a non-parametric image parsing method where label queries are made using SPs in order to reduce the complexity. 1708 dimensional SP features are defined where shape, location, texture, color and appearance information of the SPs are stored. Another study [133] uses local SP histograms and local binary patterns [92] as the SP descriptors for content based image retrieval purposes. These methods utilize general feature descriptor ideas on the SPs without any special emphasis on its nature. Moreover, high dimensional features overrule the computational efficiency gained by using SPs.

The proposed SP feature descriptor aims to infer as much information as possible from the local color, location and neighborhood state of the SP nodes. Moreover, the computational efficiency is still highly valued. The initial segmented foreground and background regions are used to extract information using the proposed features. The proposed SP features are presented in the Figure 4.15.

Location	Color	1st order Neighborhood	2st order Neighborhood	3st order Neighborhood
Centroid X	Mean Ch1	Mean Diff in 0–pi/4	Mean Diff in 0–pi/4	Mean Diff in 0–pi/4
Centroid Y	Mean Ch2	Mean Diff in pi/4–pi/2	Mean Diff in pi/4–pi/2	Mean Diff in pi/4–pi/2
	Mean Ch3	Mean Diff in pi/2–3*pi/4	Mean Diff in pi/2–3*pi/4	Mean Diff in pi/2–3*pi/4
	Variance Ch1	Mean Diff in 3*pi/4–pi	Mean Diff in 3*pi/4–pi	Mean Diff in 3*pi/4–pi
	Variance Ch2	Mean Diff in pi–5*pi/4	Mean Diff in pi–5*pi/4	Mean Diff in pi–5*pi/4
	Variance Ch3	Mean Diff in 5*pi/4–3*pi/2	Mean Diff in 5*pi/4–3*pi/2	Mean Diff in 5*pi/4–3*pi/2
		Mean Diff in 3*pi/2–7*pi/4	Mean Diff in 3*pi/2–7*pi/4	Mean Diff in 3*pi/2–7*pi/4
		Mean Diff in 7*pi/4–2*pi	Mean Diff in 7*pi/4–2*pi	Mean Diff in 7*pi/4–2*pi

Figure 4.15: Superpixel features used in the proposed SP feature descriptors



(a) Superpixel up to 3^{rd} neighborhood

(b) Diagram of neighborhood regions

(c) Diagram overlayed with superpixel regions

Figure 4.16: SP neighborhood and directional bins used in the proposed feature descriptor

The SP based feature descriptor proposed in this thesis accumulate three types of information. The location information is utilized by computing the X and Y locations of the SP centroid. This is a valuable information since two succeeding frames can only have a limited displacement in both X and Y directions. A motion based model can also be incorporated in the descriptor for a superior SP centroid localization. Color based features include the three channel mean color values of the SP. *RGB* and *LAB* color spaces has been tested for performance evaluation. *LAB* color space has been selected due to its superior performance in region identification. The third category features are composed of SP neighborhood relations. We believe that the neighborhood relations are valuable for defining a region. This is basically motivated from the pixel based widely accepted descriptors such as SIFT and HOG where oriented gradient information is utilized. Our proposed system utilizes up to 3^{rd} level neighborhood and stores the sum of mean color difference in three channels in an 8 bin angular representation as shown in Figure 4.16. This can be compared to SIFT but differs mainly due to the fact that there is no regular lattice in the SP framework.

4.5.2.2 Classification Model Training

Once the SP features are obtained for the object and background regions of the initial frame, the region model can be constructed using a classification framework. In this thesis, Support Vector Machine (SVM) maximum margin classifier [32] is used for region modelling. The stable C implementation of the SVM library supplied by [26] is utilized in the classification framework.

SVM is a maximum margin classifier where the goal is to separate the bi-class annotated data by a hyperplane. The quadratic solution of a convex function is obtained using the stochastic gradient decent in order to achieve the maximum margin. Super-

vised learning is used to find the optimal hyperplane for separating the data. The set of input training data (SP features) and the class labels are used to create the model of the data for prediction purposes. The test data (SP features in succeeding frames) is assigned on one of the two categories using the trained model.

SVM classifier models the training data in the feature space. The hyperplane is located so that the features of the separate categories are divided by a clear gap that is as wide as possible. Test features are then mapped into that same feature space and binary prediction or likelihoods are obtained depending on the region the features are assigned to in the generated model.

The goal of the classification in this stage is to answer the ultimate binary output question: "Does the current SP belong to background or foreground?" The linear discriminant function used to answer this question is shown below:

$$f(x) = w^T x + w_0 \tag{4.2}$$

The output of the above equation is used as the prediction if the SP belongs to foreground or not. In this equation x is feature vector and w is the weight term followed by the bias term w_0 where $x \in R^N$, $w \in R^N$ and $b \in R^1$. For a given set of training vectors $x_1, x_2..x_n$ and training labels $y_1, y_2, ...y_n$ where $y_i \in \{+1, -1\}$, the aim is to find the optimum weighting vector w that best separates the training data. Among the all possible hyperplanes, SVM select the one where the distance (margin) of the hyperplane from the closest data points is as large as possible.

The distance d_{x_i} of a feature point x_i to the hyperplane can be found as:

$$d_{x_i} = \frac{w^T x + w_0}{||w||} \tag{4.3}$$

If the training samples are linearly separable, the optimal hyperplane can be found by maximizing the distance of the training vectors closest to the hyperplane.

$$\text{maximize} \quad d_{x_i} \quad \text{subject to;} \quad y_i(w^T x + w_0) \geqslant d \quad i = 1, 2, ...n \tag{4.4}$$

This problem is solved using the *Lagrange* multipliers. If the training samples are not linearly separable, the maximization is done by using a trade off parameter to compensate for the misclassified samples.

A single SP feature as defined in the previous subsection, corresponds to a point in 32 dimensional space. This might seem like a limited representation of the image with such a low dimensional feature selection, especially when compared with the 128 dimensional SIFT descriptor. However, for the purposes of the study this suffices to

be descriptive enough considering the coherency in the successive video frames. The training phase is realized after the first frame of the video is segmented using the still image segmentation principles. Segmented foreground and background regions of the SP in the first image are used as binary labels for training the SVM model. Support Vectors for the given data is calculated depending on the SVM parameters. Online training using the segmentation results of the succeeding frames can also be considered for continuous model update and adaptation to changing illumination and object deformation issues. However, online training carries the risk of unsupervised data assignment which might cause deviation from the actual model in the succeeding frames. Hence, online training should be used considering its advantages and disadvantages.

4.5.2.3 Region Confidence Estimation

The SP features are assigned a likelihood (unary potential $D_p(L_p)$ in Section 3.13) depending on the distance d_{x_i} of the feature point to the hyperplane 4.3. Estimated likelihoods of the query SPs are used as a confidence term of the SP belonging to the assigned region. Figure 4.17-b and Figure 4.18-b show the region confidence estimates in the 2^{nd} and 10^{th} frames of the video. The confidence levels decrease as the frame number increases. Hence some online information propagation could be introduced in the classification model even with some user assistance at specific frames of the video.

4.5.2.4 Global Optimization Using Region Confidences

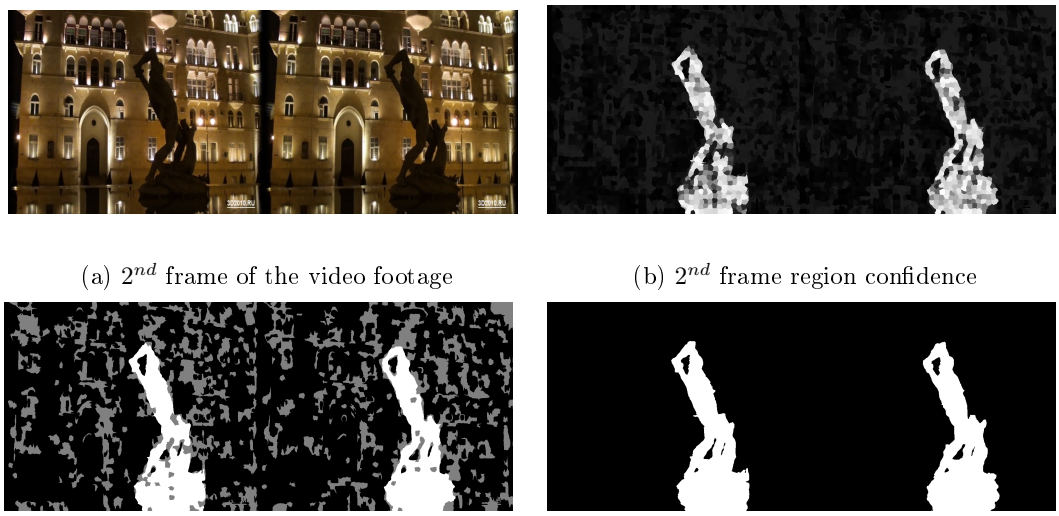
The confidence of the individual SP regions are used in the binary combinatorial optimization framework for the final segmentation decision. The region confidence values are used in graph cut formulation as the unary potentials ("T" linkes). The binary potential weights are assigned as explained in Section 3.3.2 depending on the similarity of the SP mean values. Figures 4.17-c and 4.18-c show the graph cut energy assignment in the 2^{nd} and 10^{th} frames of the video. The confidence levels become less informative as the frame distance to the reference frame increases.

The final segmentation output is computed after the assignment of the confidence levels. Figures 4.17-d and 4.18-d show 2^{nd} and 10^{th} frames without any user assistance.

4.5.3 Experimental Results

The visual results of the explained region confidences, graph cut energy assignments and automatic segmentation outputs are shown in Figure 4.17 and Figure 4.17.

The proposed system is also tested on the SegTrack dataset [126]. The dataset provides a set of 6 videos with an average 41 frames per video. The videos are selected from



(a) 2^{nd} frame of the video footage

(b) 2^{nd} frame region confidence

(c) 2^{nd} frame graph cut energy assignment

(d) 2^{nd} frame output segment

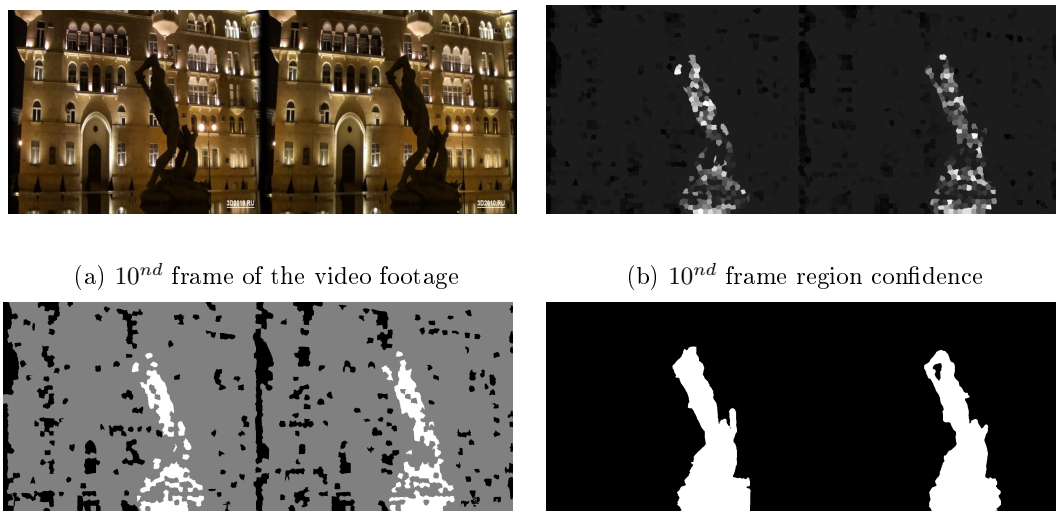
Figure 4.17: Proposed video segmentation pipeline. Output segment is automatically generated using the first frame user inputs

different difficulty levels with respect to color, motion and shape of the foreground segments. In the tests, user assistance is provided every 10 frames to achieve accurate results. The 2D/3D conversion of the videos using the proposed technique is provided for visual analysis.

4.6 Conclusion

The highlights of the proposed stereo segmentation method can be listed as follows: The interactive segmentation framework utilizes MRF based energy minimization. Superpixel primitives are used in the graph generation phase for efficient maximum flow calculation. User assistance is required as input seeds on the representative locations of just one of the stereo image pairs to save user from repeating the procedure for the second image. The information propagation is handled via efficient feature point based stereo matching. Hence, the necessity for the computationally demanding dense disparity estimation module is eliminated. The ground truth stereo database is tested for judging objective stereo segmentation performance. With additional user strokes, the proposed method is shown to generate outstanding results compared to state-of-the-art methods.

The proposed stereo segmentation technique is also presented as a post processing step for retargeting stereoscopic footage on different display sizes and resolutions. By the help of the proposed technique, novel disparity adjusted views are synthesized using the produced stereo object segments and background information for the images. To our



(a) 10^{nd} frame of the video footage

(b) 10^{nd} frame region confidence

(c) 10^{nd} frame graph cut energy assignment

(d) 10^{nd} frame output segment

Figure 4.18: Proposed video segmentation pipeline. Output segments are automatically generated using the first frame user inputs

best knowledge, utilization of segmented stereo objects for virtual depth adjustment purposes has not been addressed before. Subjective evaluations support the usage of such a disparity remapping operation regarding different aspects of visual preferences. Processed images are preferred more frequently for the problematic categories, which in fact, are the target applications for the proposed method.

Using the user-assisted image segmentation, the succeeding video frames are automatically segmented using a novel superpixel based feature descriptor. Object and background regions are learned using the proposed superpixel based region descriptors. Support vector machine is used to define the individual likelihood (confidence) of a superpixel to be assigned to the object or background region. Final region segmentation is performed using the graph cut framework where sink and source energy links are determined by the object and background likelihoods, estimated in the previous step.

CHAPTER 5

SUPERVISED IMAGE CLASSIFICATION

In the previous chapters, superpixel generation, utilization of these atomic structures in mono/stereo image/video segmentation and applications on 2D/3D conversion and disparity remapping scenarios are widely discussed. In this chapter, attention is directed to the object recognition task and ways of utilizing superpixels in various steps of the image classification pipeline is investigated.

5.1 Introduction

Recognizing and localizing semantic objects in a complex scene is a challenging problem that is solved efficiently and successfully by the human visual and cognitive system. The field has been vastly studied in the literature and various methods for object detection, recognition and segmentation have been proposed. Since these processes are connected to each other, it is usually difficult to isolate one from other [84], [127]. The information from each process contributes to the others and this creates a multi dimensional problem with strong feedback between the dimensions.

Object recognition is usually defined as the ability to assign labels to objects at multiple conceptual levels, from specific identification to coarse categorization. Possible identity preserving transformations like scaling, rotation, occlusion, changes in intensity, size and pose might be present during the assignment procedure. Ideally, a classification system should provide accurate performance in the presence of such transformations. However, no method has offered a human-like performance so far yet. This naturally leads to the following question: Where is the "gap" in the image understanding pipeline?

The aim in this chapter is to explore various steps in the object recognition process by incorporation of spatial information using the superpixel (SP) structure. Therefore, SP 2 based region segmentation and region description has been primarily investigated and the effects on the object recognition task is experimented. The contributions of this chapter are two-fold: Firstly, SP based mid-level region cues are incorporated in the feature description phase. Secondly, SP based region segmentation is proposed

as a spatial pooling method where pooled regions are defined in accordance with the underlying image characteristics.

This section is organized as follows. Related work and motivation of the various aspects in object recognition is explained in Section 5.2. Section 5.3 supplies a background information on the utilized object recognition pipeline in accordance with the state-of-the-art. Section 5.4 provides details on the construction of the superpixel descriptor and the experimental evaluation of the proposed technique is presented in Section 5.4.2. The following section 5.5 on region segmentation illustrates the adaptive region boundary extraction and utilization on the spatial pyramid pooling technique. The experimental results are discussed before concluding with final remarks and future directions.

5.2 Related Work

Object Recognition

Object recognition tasks have been vastly studied in the literature [43, 74, 125]. As a general consensus, the typical object recognition pipeline is usually studied in four major steps: 1) extraction of local image features, 2) encoding of local image descriptors, 3) pooling of encoded descriptors into a global image descriptor, 4) training and classification of pooled image descriptors for the purpose of object recognition. In this chapter, the first and third step will be widely discussed in which the local image features are extracted and spatially pooled for incorporating spatial image region statistics.

In the literature, several studies focus on evaluating the performance of the first step in the pipeline in terms of the classification and matching accuracy. Pixel based shape, color, and texture descriptors are proposed for such purposes [97]. Biological insight is also incorporated to obtain invariance under various viewing conditions [112]. Other studies propose combining different levels (low - mid - high) of information [148]. The second step of the object recognition pipeline has also been widely addressed in the literature. For encoding a set of local descriptors into a single high dimensional feature vector, the Fisher Vector method in [96], achieves state-of-the-art performance. The (third) pooling step is also shown to provide improvements. Especially spatial and feature space pooling techniques have been widely investigated [21, 54, 74]. Concerning the final step of the pipeline, discriminative classifiers like SVM are widely accepted as efficient and accurate in terms of classification performance. Judging from the non optimal final performance of the-state-of-the-art [44], one can claim that there is still room for improvement in the pipeline.

The use of scene geometry for image classification is firstly proposed by the work of Lazebnik et al. with the "spatial pyramid" idea [74]. In that paper, hard segment boundaries are utilized for creating hierarchical rectangular windows as region seg-

ments. As a further step, this chapter argues that this relation is useful but can be better exploited where hard region assignment is improved using a SP based region segmentation. The motivation behind such partitioning is to utilize the locality with the combination of the global information in frames, ie. an image on the highway is more likely to contain a car than a toaster [74]. With such geometric partitions, one can better utilize the statistical information related to the location of individual segments, ie. "Sky" is usually in the top part of the image.

A previous work on generic pooling for image classification, proposes utilization of similarities between image categories [129]. This has been shown to improve the classification scores with the introduced correspondence between the equivalence classes. This study emphasizes the idea that geometric properties of the regions have a statistical relevance to image categories.

Partitioning of images into semantic regions by using the top-down knowledge with bottom-up grouping approach has been previously discussed in the literature [18], [64]. Such methods aim to generate semantic segments that correspond to single objects or regions. This has shown improvements in classification scores when accurate segmentation results are achieved.

In this chapter, the motivation of the proposed adaptive region segmentation is to generate coarse boundaries on the spatial regions. This is performed according to a given geometric prior where a convexity constrained energy term preserves the shape of the initial geometry. This would yield adaptation of the region boundary while preserving the region geometry.

Neuroscience Perspective

The goal of the studies regarding the semantic gap in the image understanding procedure is to determine where the machines lack accuracy compared to humans. In order to address this issue, the way the brain solves visual object recognition task has been investigated. The fact that half of the primate neocortex is engaged during the visual processing, proves the complexity of the whole recognition process [46]. Moreover, recent studies propose strong evidence that a cascade of computations in the Inferior Temporal (IT) Cortex (it is the cerebral cortex on the inferior convexity of the temporal lobe in primates including humans) are engaged in the visual object recognition process [39]. However, the underlying algorithm that produces this result stays mostly undiscovered.

Neuroscientists deal with the underlying aspects of brain functionality during the recognition process. The focus in this chapter is not to investigate the neural implications of visual understanding. However, it is important to emphasize the results of the recent studies. These results can be valuable to better understand the object recognition process. It has been observed that the IT pattern of activity can be very informative

for achieving robust and real time visual object categorization [77, 108]. Even a simple weighted summation of IT spike counts (without any advanced classification method) can lead to high rates of validated performance for a wide range of object recognition tasks. On the other hand, there is still very limited information regarding the encoding of individual IT responses. What is known is that the IT neurons are activated by at least moderately complex combinations of visual features [108].

To summarize the discussion from the neuroscience perspective; IT neuron outputs are very explanatory and valuable for visual understanding. Hence, efforts towards the clustering methods might be directed towards obtaining "good" feature descriptions and encodings. The goal of the study explained in this chapter is to develop and analyze extensions in feature description and encoding schemes with exploration of hierarchically classified pixel (low level), region (mid level) and scene based (high level) feature descriptors.

Mid-Level Cues

Finding the correct type of features is very important in the image classification task. Local features are generally used to extract pixel level information from the interest points or in a dense grid. On the other hand, global features are used to define the whole image using a descriptor. In this thesis, mid-level descriptors are proposed in order to incorporate mid-level region information of the image. For that purpose, superpixel atomic structures are utilized. Superpixels can seen as an efficient image representation with reduced resolution and information encapsulation property. SP extraction is previously explained in Chapter 2.

The study in [20] proposes mid-level features for object recognition and presents a detailed analysis on different levels of pooling strategies. They define macro-feature vectors as jointly encoded small neighborhoods of SIFT descriptors. The neighborhoods are defined by a fixed size of squares that encode multiple SIFT descriptor into one as the macro-feature vector. This method pursues a similar spatial information utilization as proposed in this thesis. However, they use only fixed sized (multiple) square regions independent of the region properties. The proposed technique in this thesis on the other hand, aims at combining spatial characteristics of the region and encoding it into a descriptor that has flexible and adaptive coverage depending on the spatial region properties. A recent work [95] that investigates the role of local and global information in image classification also focuses on exploring the performance limitations of current techniques. Another study that aims at labeling image regions depending on the similarity of the SP features in the training set is presented in [124]. In that study, scene-level matching with global image descriptors is followed by SP level matching of mid-level features. The study in [148] addresses the low-, mid-, and high-level cues. Individual classifiers are trained on different levels of descriptors and

classification outputs are combined for the final decision. Descriptor level grouping has also been addressed in a more recent study [49] where local histograms from larger neighboring regions have shown to improve classification performance. This method uses a fixed neighborhood definition to aggregate the local histograms; whereas, in this thesis a flexible and more natural region description is proposed.

5.3 Image Classification Pipeline

As previously stated, the typical object recognition pipeline consists of four major steps also presented in Figure 5.1; 1) extraction of local image features, 2) encoding of local image descriptors, 3) pooling of encoded descriptors into a global image descriptor, 4) training and classification of pooled image descriptors for the purpose of object recognition. Prior to the detailed analysis of the proposed feature descriptor and spatial pooling enhancement, general image classification pipeline will be explained.

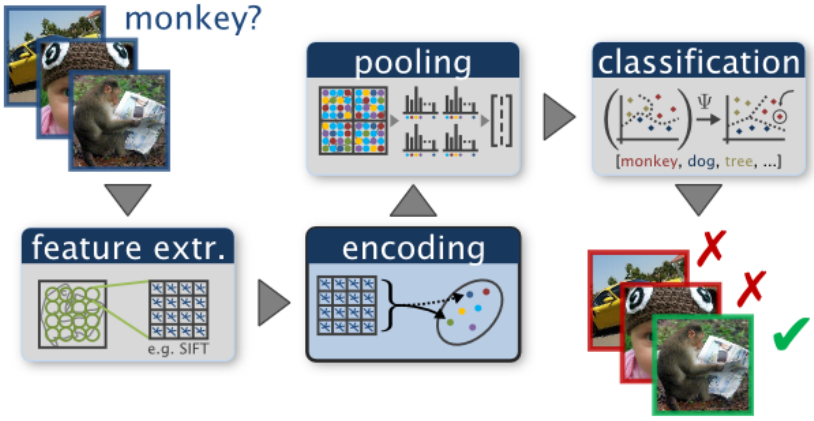


Figure 5.1: Algorithmic flow of a general image classification pipeline [27]

5.3.1 Local Image Features

The literature on producing robust local image descriptors dates back to a couple of decades. However, the focus in this study is not to investigate the performances of different local image descriptors in the image classification problem. Therefore, the well known image descriptor SIFT [83] has been selected due to its widely accepted descriptive performance. The publicly available VLFeat toolbox [132] has been used in the descriptor computation. This is a much faster and a very close approximation to the original implementation in [83]. The increase in the speed is important for dealing with large datasets and dense sampling of the image points. Dense sampling of the image is important for understanding the global statistics and the "gist" of the image [125]. However, pixel resolution sampling might still be redundant. In this study every

one out of 4 pixels is regularly sampled and the local descriptor has been computed accordingly. It has been observed that reducing the dimension of the SIFT features by using principal componenet analysis (PCA) [69] might provide not only efficiency but also an increase in the final classification performance.

5.3.2 Feature Encoding

Encoding of the local image descriptors is based on the idea of partitioning the feature space into structural regions in the sense that these can be defined by a grouping rule. These regions are named as *visual words*, and by the combination of all the visual words, the *visual vocabulary* is generated. In order to define the regions of the visual words in the d dimensional feature space, K-means clustering approach is utilized. The set of n local image descriptors $x_1, ..., x_n$ are used to define the k visual regions. The total number of training image descriptors n is defined by the dense number of image samples multiplied by the total number of images in the training dataset. K-means clustering aims to partition the n observations into k sets so as to minimize the within-cluster sum of squares.

$$\underset{s}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2, \tag{5.1}$$

where μ is the mean of points in S_i.

The standard Lloyds algorithm [81] iterates between the search of best mean values μ_i from the samples $x_j \in S_i$ and best grouping from the estimated means. For a given of k means $\mu_1^1, \mu_2^1, ...\mu_k^1$, the algorithm iterates by alternating between the two "assignment" and "update" steps. In the assignment step, each observation is assigned to the cluster whose mean is closest to it (i.e. partition the observations according to the Voronoi diagram generated by the means).

$$S_i^t = \{x_p : ||x_p - \mu_i^t|| \leq ||x_j - \mu_j^t||, \qquad \forall 1 \leq j \leq k\}, \tag{5.2}$$

where x_p is assigned to one S_t. In the update step, the means to the centroids of the observations in the new clusters are calculated.

$$m_i^{t+1} = \frac{1}{|S_i^t|} \sum_{x_j \in S_i^t} x_j. \tag{5.3}$$

The algorithm converges when there is no longer change in the assignment phase and hence no update is necessary.

Bag-of-words model has been effectively used for object classification and video retrieval purposes [147]. The success of this orderless method lies in the categorization

of the whole scene without requiring any spatial coherency. For example, if many number of words that define a face is found in a scene, there is a high probability that a face exists in that scene. This would be independent of the position of these words and their spatial relation. The likelihood of a category existing in a scene is usually computed by the histogram of the visual words. The vocabulary of size k with the estimated means $\mu_1, ..., \mu_k$. Assignment S_i of the individual descriptors x_i is done as shown in (5.2). The histogram of the assigned values produce the general description of the scene. The success of *bag-of-words* model has attracted attention and hence improvements on the discrete representation of the feature space has been proposed. A weak point on the bag-of-words idea is that it only allows a discrete vocabulary assignment; however, a feature point can be similar to different words in the vocabulary at the same time. It has been shown that codeword uncertainty can be improved with the soft assignment idea [130]. This method has produced superior categorization performance for state-of-the-art datasets. Recent studies proposed similar encoding techniques in order to better utilize the uncertainty. Fisher Kernel [96] method improved the soft assignment idea by introducing the first and second order statistics of the Gaussian mixture model (GMM) for the feature encoding procedure. GMM model is derived from the K-means vocabulary centers as initialization points.

GMM clustering aims to model the parameters of a probability distribution $p(x|\theta)$ (5.4) and (5.5). The set of individual samples are used to estimate the D dimensional mean (μ), covariance (Σ) and the weight π of the mixture of Gaussian.

$$p(x|\theta) = \sum_{k=1}^{K} p(x|\mu_k, \Sigma_k)\pi_k, \tag{5.4}$$

$$p(x|\mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^D det\Sigma_k}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}, \tag{5.5}$$

where $\theta = (\pi_1, \mu_1, \Sigma_1, ...\pi_K, \mu_K, \Sigma_K)$ is the vector of estimated parameters of the model. π_i is the weight; $\mu_i \in R^D$ is the mean and $\Sigma_i \in R^{DxD}$ is the covariance matrices of each Gaussian mixture. Parameters are estimated by the Expectation Maximization method [37] from the available sample points $x_1, ...x_N$. The likelihood of individual samples belonging to one of the K regions are estimated as in (5.6)

$$q_{ki} = \frac{p(x_i|\mu_k, \Sigma_k)\pi_k}{\sum_{j=1}^{K} p(x_i|\mu_j, \Sigma_j)\pi_j}, k = 1, ...K \tag{5.6}$$

Fisher encoding [96] is an alternative way to *bag-of-visual-words* method where non-discrete label assignment is encouraged. Fisher vectors are composed of two gradient calculations where the average of first and second order distance between the feature points and GMM means are computed. u_k captures the distance between the image descriptors and the mean of the corresponding Gaussian. The second order distance v_k similarly computes the variance, see equation 5.7.

$$u_k = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{K} q_{ik} \frac{x_i - \mu_k}{\sigma_k}, \tag{5.7}$$

$$v_k = \frac{1}{N\sqrt{2\pi_k}} \sum_{i=1}^{K} q_{ik} [\frac{(x_i - \mu_k)^2}{\sigma_k^2} - 1], \tag{5.8}$$

where σ^2 is the variance term in the diagonal covariance matrix Σ, and q_{ij} represent the soft assignment term defined in (5.6). The final encoding of the vector is done by direct concatenation of u_i and v_i for all N components. This structure produces a vector of size $2DK$. as shown in (5.9).

$$f_{fisher} = [{u_1}^T, {v_1}^T, ..., {u_K}^T, {v_K}^T]^T. \tag{5.9}$$

5.3.3 Pooling

The state-of-the-art object recognition pipeline includes a pooling step where the responses of encoded descriptors are combined either spatially or in the feature space. The aim of combination of features is to transform the joint feature representations such that the local information is preserved while irrelevant details are eliminated. The success of the spatial pyramid idea [55, 74] illustrates an empirical motivation towards spatial grouping for classification purposes.

Spatial pooling is introduced as a weak geometry constraint on the image classification challenge. It was initially introduced as an extension to the *bag-of-words* representation but could be easily applied to any other encoding scheme. In the previous seminal work of [74], image is partitioned into sub-regions and the feature encoding is realized individually on these sub regions. Imposing locality constraints on the encoding step has shown increase in the final classification accuracy. The individual local region descriptors are concatenated to produce the final global scene descriptor. In the original work, the base line is selected by setting the spatial pyramid regions as 1×1, 3×1 and 2×2 on the image geometry. When the fisher vectors corresponding to all the regions are concatenated, a descriptor of 8 times the original size is constructed.

5.3.4 Classification

In the classification phases, support vector machines (SVM) is usually utilized with the publicly available library [26]. SVM is a method of binary classification (1-1 or sequential 1-all for multi class case) using supervised learning for data analysis and pattern recognition. The set of input training data and their labels are used to create

the model. The test data is supplied to the trained system and assignment is done on one of the two categories. An SVM model is a representation of the features in the feature space, mapped so that the features of the separate categories are divided by the largest margin possible. Test features are then mapped into the same space and prediction is done by computing the likelihood of the region that the test feature is closest to. SVM performs classification between two classes by finding a decision surface that is based on the most informative points of the training set.

For training the encodings in the object recognition task, human annotations are used for training with a linear kernel. Even the non-linear kernels tend to yield better classification accuracy, linear kernels are usually selected for computational efficiency. It has also been observed that SVM performs better if the data is normalized. parameter C of the SVM (regularization-loss trade off) is determined on a validation set (on the provided train and validation split in the dataset).

5.4 Mid-Level Cues from Superpixels

The individual steps in the object recognition pipeline is presented in the previous section. This section focuses on the first step in the pipeline where feature extraction is performed. Conventional feature extraction techniques are explored from the perspective that mid-level information could be incorporated in order to obtain a superior scene description. It is hypothesized that pixel based low-level descriptions are useful but can be improved with the introduction of mid-level region information. Hence, the methods to acquire such mid-level information from the image regions are investigated in order to improve the classification and retrieval accuracy. Detailed experimental evaluations on classification and retrieval tasks are performed in order to validate the proposed hypothesis.

Pixel based descriptors are widely used in object recognition tasks due to their accepted performance for image description [83]. However, the use of middle and higher level descriptors is important for a better scene characterization. In the proposed method, the aim is to extend the low level descriptors towards middle level region descriptors. The advantage of the proposed mid level description is that it does not require a fixed region size or shape to define the support area of the descriptor. Region shape is adaptive depending on the spatial image characteristics. Therefore, the proposed descriptor is based on the superpixel mean color and variance information in a spatial neighborhood. Different region and superpixel sizes as shown in Figure 5.2 and Figure 5.3 are used to explore possible contributions by fusing spatially different levels of information.

The proposed Superpixel based Angular Differences (SPAD) method uses the intensity difference between the SPs in a local neighborhood. The angular intensity differences in the SP neighborhoods are accumulated in order to define the region covered by

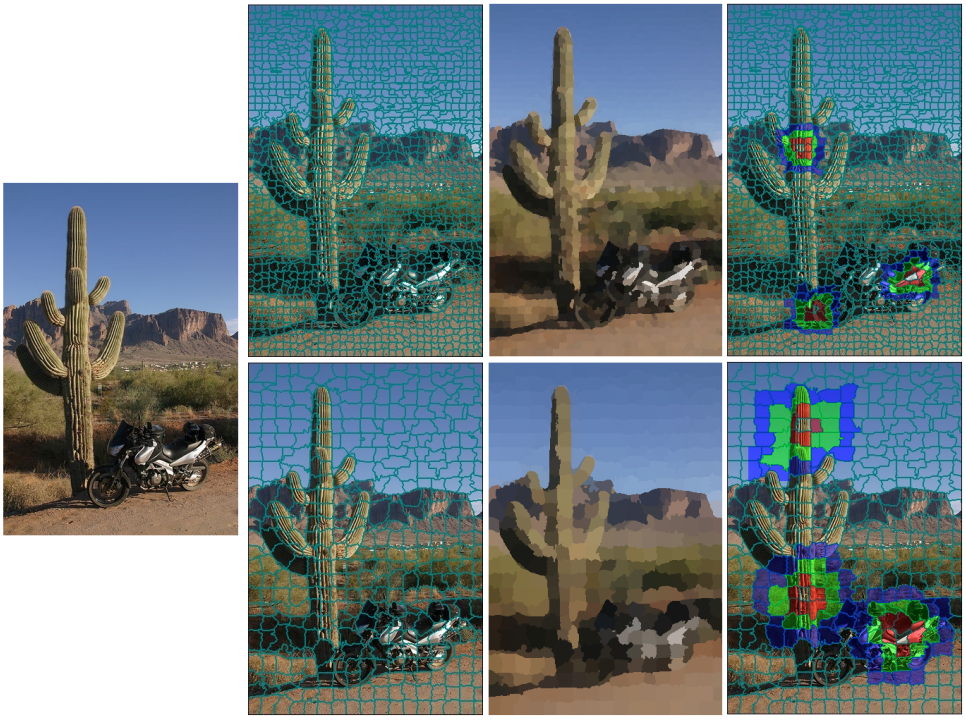


Figure 5.2: Describing an image with superpixels. Top: SPs with size 10x10 SP. Bottom: 20x20 SP. From left to right: Original image; SP boundaries on the image; Mean RGB values for each SP region; first (red), second (green) and third (blue) order neigborhoods of randomly selected 3 SP regions.

the irregular shaped SPs. Figure 5.4 presents the proposed idea where central and neighboring SPs are generated in a realistic configuration for illustration purposes. The coverage of the neighborhood depends on the size of the extracted SP and the number of neighbor levels. Local SP neighborhood in Figure 5.2 and Figure 5.3 shows the extracted SP boundaries on the original image. On the colored area, the different orders of neighborhoods of the central SP are emphasized with "red", "green" and "blue" colors.

5.4.1 Superpixel based Angular Differences (SPAD)

The proposed descriptor extraction method is explained in four steps as follows: 1) Extraction of SPs for different sizes in Section 5.4.1.1 (3×3, 5×5, 10×10, 20×20). 2) Generation of SP neighborhood structure in Section 5.4.1.2. 3) Computation of the angular intensity differences and variances for different (1^{st},2^{nd}, and 3^{rd}) levels of neighborhood in Section 5.4.1.3. 4) Fusion of the computed angular differences for different sizes of SPs in Section 5.4.1.4.

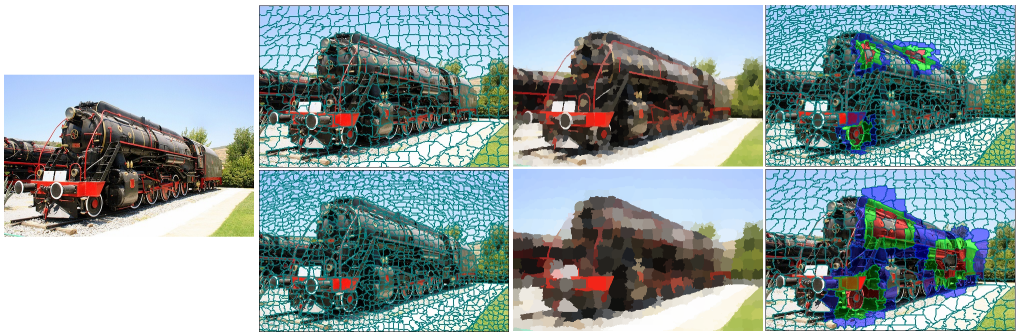


Figure 5.3: Describing an image with superpixels. Left: SPs with size 10×10 SP. Right: 20×20 SP. From top to bottom: Original image; Mean RGB values for each SP region; first (red), second (green) and third (blue) order neigborhoods of randomly selected 3 SP regions.

5.4.1.1 Superpixel Extraction

For the purpose of the proposed mid-level descriptor, extracted SP patches should possess several structural properties. The SP extraction method as explained in Chapter 2 preserves local structure by adapting to the local object and region boundaries. Moreover, undersegmentation of the regions is avoided to yield an expressive image representation. Uniform localization and compactness are also satisfied to form regular grid structure among the graph models with unbiased neighbor relations. In order to generate a scalable descriptor, different sizes of SPs are hierarchically extracted based on the initial grid structure (3×3, 5×5, 10×10, 20×20).

5.4.1.2 Superpixel Neigborhood Structure

Each SP patch p corresponds to a node $v \in V$ of an undirected graph $G = (V, E)$. Each edge $e \in E$ of the graph is assigned a weight depending on the similarity of the nodes that it connects. For each SP, the neighborhood of p is defined as N_p^n where n corresponds to the order of the neighborhood with $n \in \{1, 2, 3\}$ in our implementation. For the given parameter settings, rough calculation of the region coverage with 3 levels of neighborhood for 20×20 SP size resutls in $(2n + 1) \times 20 \rightarrow 140 \times 140$ pixels for $n = 3$. This coverage can be adjusted with different sized SPs or neighborhood levels. In the proposed implementation, up to the 3^{rd} level of neighborhood with the following SP sizes: 3×3, 5×5, 10×10, 20×20 are utilized.

While generating the neighborhood structure, an iteration on the individual nodes is conducted in order to define the neighborhood relations. To obtain a color wise distance d_{p,q_i} between the adjacent nodes p and q_i ($q_i \in {N_p}^n$), the distance metric is computed over three color channels:

$$d^c_{p,q_i} = e^{\frac{-(\mu_p{}^c - \mu^c_{q_i})^k}{\sigma^c_p}} sign(\mu^c_p - \mu^c_{q_i})^{k-1}, k = 1, 2 \tag{5.10}$$

, where μ^c is the mean color of the c^{th} index of the color channel and σ_p is the variance of the mean color values in the n^{th} neighborhood:

$$\sigma_p^{c\,2} = \frac{1}{\|N_p^n\|} \sum_{i=1:\|N_p^n\|} (\mu_p{}^c - \mu^c_{q_i})^2, \tag{5.11}$$

where $\|N_p^n\|$ is the total number of neighbors of the SP p within the n^{th} neighborhood.

In order to compute the angular difference, the angular orientation of each SP with respect to the central SP is required. The angular orientation $arg(p, q_i)$ (argument of the vector $(\vec{p} - \vec{q_i})$ in R^2) between the adjacent nodes p and q_i ($q_i \in N_p^n$) is computed as:

$$arg(p, q_i) = \begin{cases} \arctan(\frac{(p^y - {q_i}^y)}{(p^x - {q_i}^x)}) & \text{if} \quad x > 0 \\ \arctan(\frac{(p^y - {q_i}^y)}{(p^x - {q_i}^x)}) + \pi & \text{if} \quad x < 0 \;\; y \geq 0 \\ \arctan(\frac{(p^y - {q_i}^y)}{(p^x - {q_i}^x)}) - \pi & \text{if} \quad x < 0 \;\; y < 0 \end{cases} \tag{5.12}$$

where p^x, p^y correspond to the x and y pixel coordinates of the SP p.

The calculated distance and angular orientations are used in the next step to compute the angular intensity differences.

5.4.1.3 Angular Difference Computation

We divide the angular space in 8 equal bins to compute the intensity differences of superpixels for different orders of neighborhood. Figure 5.4 illustrates the proposed idea where different colored centers contribute to the intensity difference term in the 8 bin angular orientations.

D^c_θ is the angular intensity difference between the center SP p and its neighbors at the selected angle θ and color channel c. In our implementation, we use 8 bin orientations where $\theta \in \{0,\ \pi/4,\ \pi/2,\ 3\pi/4,\ \pi,\ 5\pi/4,\ 3\pi/2,\ 7\pi/4\}$. The angular difference D^c_θ is computed as the summation of the projection of 3 closest (in terms of angular orientation) SPs in the selected neighborhood order (5.13). Figure 5.4 shows the projected points for $\theta = 0$ for the 1^{st} neighborhood and $\theta = 3\pi/2$ for the 2^{nd} neighborhood. The dashed lines show the projection of SP centers on the corresponding orientations and intensity differences are accumulated on each orientation as follows:

$$D^c_\theta = \sum_{q_i, i=1:3} d^c_{p,q_i} \cos(arg(p, q_i) - \theta), \tag{5.13}$$

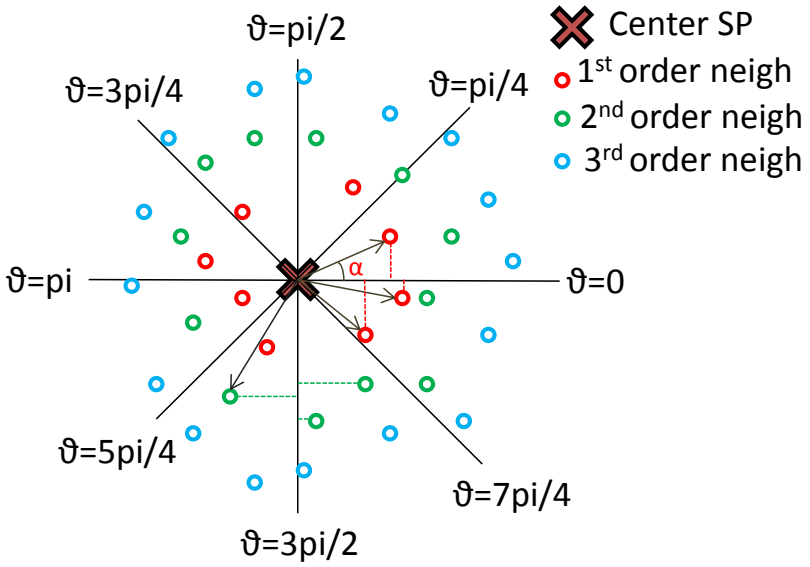


Figure 5.4: Computation of angular differences on the superpixel grid. Projection of the closest superpixels are accumulated on the final intensity difference. X represents the central SP and circles in different colors ("red", "green", and "blue") represent the 1^{st},2^{nd}, and 3^{rd} order neighbor superpixel centers

where the 3 closest neighbors are selected as:

$$i = \underset{j}{\operatorname{argmin}}(|\theta - arg(p, q_j)|), \quad j \in N_p \tag{5.14}$$

Incorporating Second Order Statistics In addition to the angular intensity difference, we also incorporate the angular distribution of second order statistics of the SP patches. As in (5.13), we compute the angular variances in the SP patches as shown below in (5.15).

$$V_\theta^c = \sum_{q_i, i=1:3} {\sigma_{q_i}^c}^2 \cos(arg(p, q_i) - \theta) \tag{5.15}$$

where ${\sigma_{q_i}^c}^2$ is the variance of the c^{th} color channel in SP q_i.

5.4.1.4 Descriptor Fusion

The computation of angular difference D_θ and angular variance V_θ for 8 orientations produce a 8×1 length vector each. In the proposed method, up to 3 levels of neighborhood information are used to generate a 48×1 sized vector for D_θ and V_θ together.

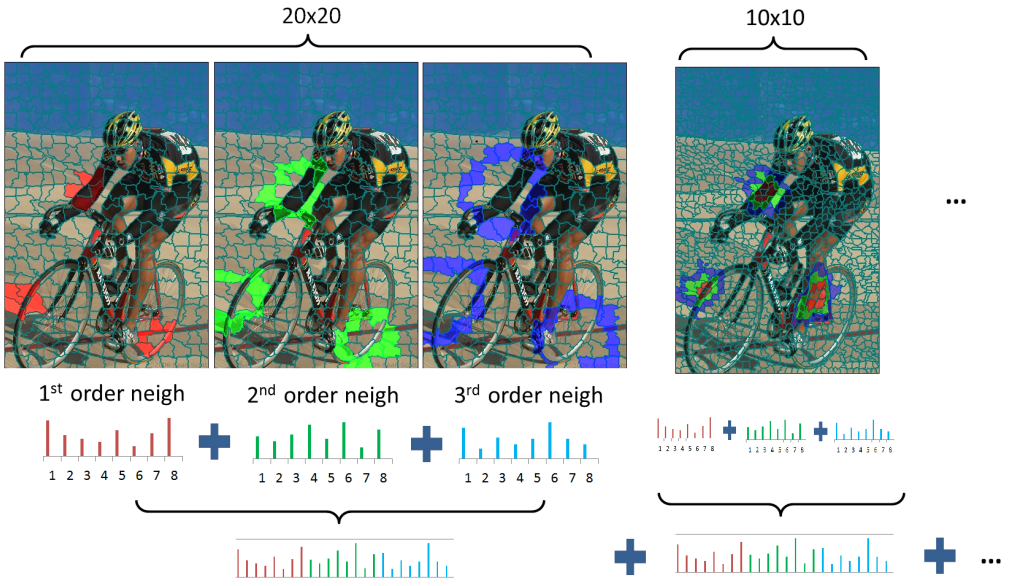


Figure 5.5: Angular difference computation. Red, green, and blue colored regions correspond to the 1^{st},2^{nd}, and 3^{rd} order neighborhood of the central SP. Angular differences are combined for different neighborhood and SP sizes.

This vector constitutes the final region descriptor for the given hierarchy as illustrated in Figure 5.5 for different orders of neighborhoods and SP sizes.

Different sizes of SPs are used to obtain scale invariance and cover distinct mid-level region cues that we aimed for. The final structure of the descriptor when the angular difference and variance are combined is shown below.

$$v = [D^n_{\theta_1}\ D^n_{\theta_2}\ \dots\ D^n_{\theta_8}\ V^n_{\theta_1}\ V^n_{\theta_2}\ \dots\ V^n_{\theta_8}]^3_{n=1}$$

As a final step, the two descriptors D^n_θ and V^n_θ are independently ℓ_2 normalized over all neighborhoods. The normalization step has provided with an increase in the final classification accuracy.

5.4.2 Experimental Results

Image Classification The descriptive performance of SPAD is evaluated on image classification. This task aims at detecting the predefined class of each image in a test set based on training samples. For this purpose, the Pascal VOC 2007 Dataset [43] is used, which consist of a total number of 9,963 images (5,011 for training and 4,952 for testing). Some examples of the 20 classes in the dataset are: person, motorbike, air plane, cat, cow, bottle, sofa, etc. The measure used to evaluate the performance of a given system is the Average Precision (AP) metric. The Mean Average Precision

(MAP) is the averaged AP over all the classes tested.

The conventional image classification pipeline used in the experiments is presented in previous section 5.3. In the feature extraction state, VLFeat toolbox [132] is used to compute the SIFT descriptors. Following the ℓ^2 normalization of the SIFT descriptors, principal component analysis (PCA) is performed and the dimensionality of the SIFT features are reduced to 64. Encoding the local image descriptors is achieved using the Fisher Vectors (FV), since it has been reported to outperform other encoding methods on the classification task [27].

Test scores are ranked depending on the output likelihood of each image to belong to the classes in the training set. With the proposed modification on the pipeline, SPADs are computed on each image instead of dense SIFT descriptors. Fisher vectors and SVm are used in accordance with the conventional pipeline.

An evaluation of the proposed system using various scales of SPAD is performed and the results are shown in Table 5.1. SPs are extracted from different grid sizes: 3×3, 5×5, 10×10, and 20×20, see Figure 5.5. SPADs are computed hierarchically on different grids (SPAD3, SPAD5, SPAD10, SPAD20) for every image. The MAP scores are calculated for each size and the combination of different sized descriptors are investigated for improving the accuracy. The combinations are performed at different steps of the pipeline, as called *early-fusion* and *mid-fusion*. *Early-fusion* encodes all scales of SPAD together at the feature extraction level and builds a standard sized fisher descriptor for the image. on the other hand, *mid-fusion* concatenates the fisher vectors corresponding to individual scales that are separately encoded. This would yield a larger image descriptor size compared to *early-fusion*.

Table 5.1 shows the MAP scores of different scales of the proposed descriptor. As the initial size of the SP goes smaller, the precision increases. This is expected since larger sized SPs will group bigger image regions and hence making the image understanding difficult. Moreover, combinations of different scales produce more accurate results since several levels of region information is incorporated in the combined features. As shown in the final result, *mid-fusion* is observed to outperform *early-fusion* in terms of final MAP.

Table 5.2 evaluates the baseline method, using dense SIFT, and its combinations with the proposed approach. The image descriptors of each method are combined using *mid-fusion*. The combination of dense SIFT descriptors with SPAD offers better performances and a 2.7% improvement in terms of MAP over the baseline is observed.

A class level detailed look at the AP scores is presented in Figure 5.6. The increase in the AP score per class varies between 0.1% and 5.6%. On all of the classes there is an increase in AP; and on the large majority of the classes, a very stable improvement, between 2% and 4% is observed. Increase is obtained regardless of the nature of the data, due to the adaptivity of the proposed descriptor. This observation supports the

Table5.1: SPAD classification MAP scores for Pascal VOC 2007, using Fisher vectors with k=256 Gaussians. Descriptors are combined using *early-fusion* and *mid-fusion*.

Method	Fisher 256	Dimensions
SPAD 3	0.381	$48 \times 2 \times k$
SPAD 5	0.356	$48 \times 2 \times k$
SPAD 10	0.300	$48 \times 2 \times k$
SPAD 20	0.252	$48 \times 2 \times k$
SPAD 3,5 Mid	0.406	$2 \times 48 \times 2 \times k$
SPAD 3,5,10 Mid	0.417	$3 \times 48 \times 2 \times k$
SPAD 3,5,10,20 Mid	**0.421**	$4 \times 48 \times 2 \times k$
SPAD 3,5,10,20 Early	**0.410**	$48 \times 2 \times k$

Table5.2: MAP scores for the standard pipeline and combination with SPAD for Pascal VOC 2007, using Fisher vectors with k=256 Gaussians

Method	Fisher 16	Fisher 64	Fisher 256
SIFT standard	0.440	0.491	0.549
SIFT & SPAD-Early	0.457	0.514	0.563
SIFT & SPAD-Mid	**0.468**	**0.527**	**0.576**

initial hypothesis concerning the information gained by utilizing the mid-level cues.

Image Matching In this part of the thesis, an image descriptor specific to image classification task is proposed. The aim is to accumulate mid-level information from the scene which is hypothesized to be complementary to the low-level information such as SIFT. Dense SIFT descriptors are used in the image classification pipeline without rotation invariance property and are accepted as the conventional way to describe a scene [27]. Therefore, rotation variant SP descriptor is used in the classification experiments. To complete the evaluation, the proposed method is extended to have rotation invariance property and is evaluated on the image matching task. For rotation invariance, the orientation component of SIFT is computed and it is used to align the SP descriptors along the main orientation. Hereby, some visual results in Figure 5.7 from the dataset [87] are shown; as an illustration of the rotation invariance property.

5.4.3 Discussion

Object recognition can be improved by exploring the limitations at the feature extraction step. Current low-level image descriptors are widely explored for such purposes; however, utilization of mid-level cues can capture additional spatial information. Previous mid-level techniques mostly define a fixed image region and accumulate the

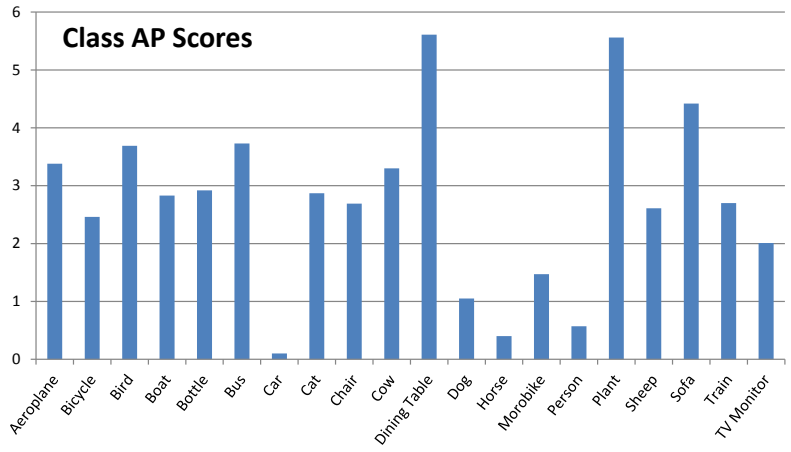


Figure 5.6: AP score increase with the proposed SPAD-Mid combination compared to the standard SIFT, for individual classes of Pascal VOC.

low-level information in this predefined window. Different scales of the SIFT descriptor can also collect information from a larger but fixed sized area on the image. On the contrary, in this section a descriptor in the SP domain is proposed and the regions are defined according to the spatial characteristics of the image. The advantage of such an approach is to incorporate region specific information in the descriptor. One can also argue about the similarities of the proposed method with the LBP descriptor [93], especially in the hierarchical neighborhood idea. However, the two techniques differ. The LBP method stores the sign of the differences in the predefined locations of the image. The binary vectors of the sign differences are then accumulated in a histogram on a predefined window. The proposed method on the other hand, stores not only the sign but also the magnitude of the difference and covers a region that is adaptive in terms of shape and size.

The experimental results show supporting evidence that the proposed method is useful for improving the accuracy of the object recognition task. Visual results on image matching performance is also promising and can be further studied.

5.5 Geometry Based Region Segmentation for Spatial Pooling

In the previous section, a method for region descriptor using SPs is presented. This technique intends to utilize mid-level information in the image that are hypothesized to be complementary to the low level pixel features. In this section similar region description idea is generalized on the pooling step of the image classification pipeline. In that aspect, a geometry-constrained region segmentation approach for image classification is proposed. Scene geometry is supplied as an input parameter for region segmen-

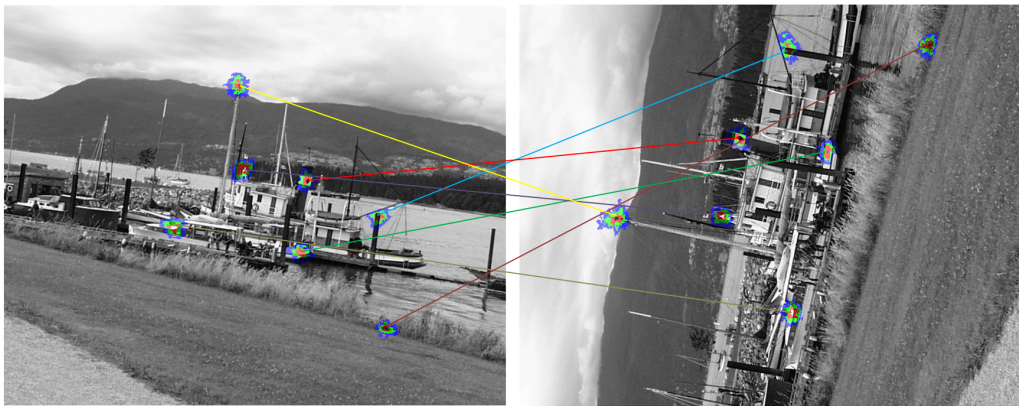


Figure 5.7: Matched points on the rotated and scaled image pairs. Different colors represent the image neighborhoods.

tation. The resulting segmentation is performed in accordance with the predefined geometric guidelines. Using an approximate global geometric correspondence exploits the idea that images of the same category share a spatial similarity. This assumption is evaluated and justified in an object classification framework, in which generated region segments are used as an enhancement to the widely utilized "spatial pyramid" method. Fixed region pyramids are replaced by the proposed locally coherent geometrically consistent region segments. Performance of the proposed method on object classification framework is evaluated on the 20 class Pascal VOC 2007 dataset and a consistent increase in the MAP score for different experimental scenarios is observed.

In this section, region segmentation in order to obtain natural image layout for the purpose of image classification is explored. In order to do that, a method that segments image regions automatically for a given input geometry is proposed. Region segmentation assigns pixels in one-to-one correspondence to the connected image regions. The geometry utilization for image classification is inspired by the work of Lazebnik et al. with the "spatial pyramid" idea [74]. Hard image segment boundaries are defined for creating hierarchical rectangular windows as region segments. The motivation behind such partitioning is to utilize the locality with combination of the global information in frames.

Proposed method offers an improved spatial pooling idea where an extension to the standard spatial pyramid by enforcing a similarity measure on the region segmentation is proposed. Some image categories occur more likely in the specific regions of the image, ie. sky is on the top part, a car is in the middle or lower part. The formation of image descriptors on the manually defined spatial regions might succeed in providing information about the layout of image features. However, one can argue about the optimality, since the manually defined grid structure may not adapt to fit the spatial statistics of natural images. With the proposed geometry based segmentation, the

statistical characteristics is further enforced by allowing boundary adaptation on the object/region boundaries.

The proposed region segmentation is done on a graphical framework. The image is first segmented in superpixels as explained in Chapter 2. Each small SP region corresponds to the nodes of an undirected graph $G = (V, E)$ where edges e connecting the SPs are assigned a weight w_e depending on the similarity of the SP nodes v.

In the proposed region segmentation, manual image geometries are used as initial region segments. Region boundaries are iteratively adapted depending on the color-wise and spatial distance between the individual SP and the candidate region. During the region boundary update, the cost function relating similarity of the SP to the corresponding image region candidate is minimized. This approach aims to keep image regions connected without any sub-detachment.

The proposed method initially segments the image into large number of (~ 600) color wise similar SP regions. The final goal is to partition the image into small number of (3 or 4) spatially coherent regions by dynamically moving superpixel patches using the initial geometry. With an iterative update procedure, an energy objective is pursued, where superpixels are assigned to the region that satisfies minimum energy cost. Region updates are terminated either after a fixed amount of iteration or if the energy reduction after the update is smaller than a threshold.

5.5.1 Proposed Region Segmentation

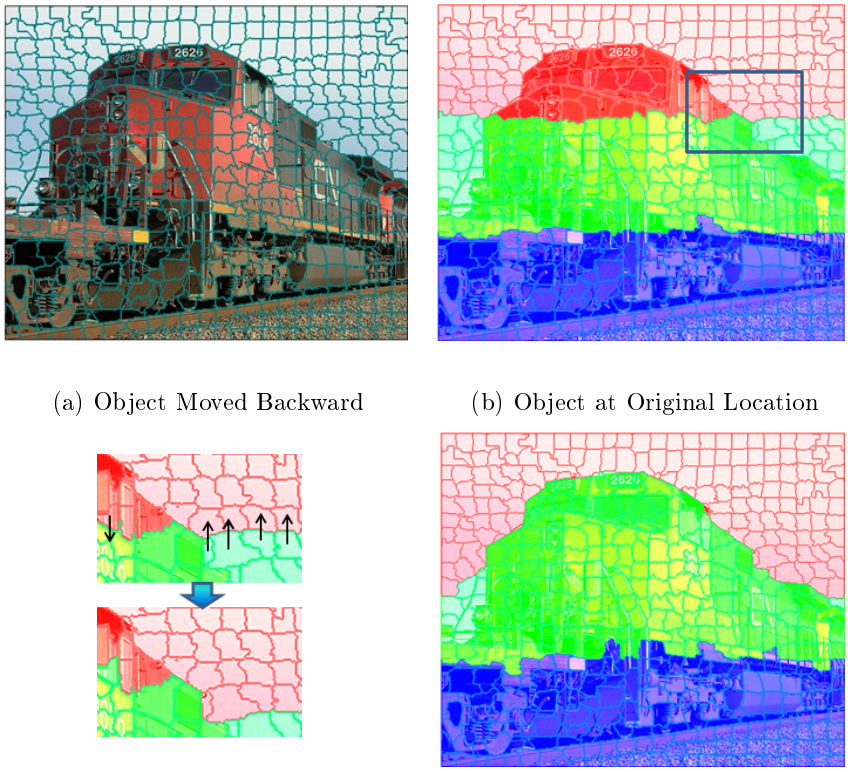
The proposed region segmentation method includes three main algorithmic steps: 1) Initialization of the regions with the input geometry. 2) Region boundary update. 3) Region structure update.

1) In the first step, generated SPs are assigned to the regions according to the initial geometry. The SP boundaries as shown in Figure 5.8-a are initialized to the input 3×1 geometry in Figure 5.8-b.

2) The second boundary update step performs a greedy search on the boundary SPs. Figure 5.8-c shows the SP updates on the regions boundaries. During the boundary adaptation, the cost function that relates the similarity of the SP to the corresponding region candidates is tried to be minimized. This approach assures that the final regions are composed of connected SPs without any sub-detachment. SP to region assignment is performed according to the formula given in (5.16),

$$L(p) = argmax_{i=1:N} \left(Si(p, Q_i) \right), \tag{5.16}$$

where $L(p)$ is the region label of the SP p; $S(p, Q_i)$ is the similarity cost between the corresponding SP p and region Q_i. N is the number of neighboring region candidates.



(a) Object Moved Backward

(b) Object at Original Location

(c) Object Moved Forward

(d) Object Moved Forward

Figure 5.8: Region segment generation for 3×1 geometry. a) Initial SPs are generated. b) 3×1 geometry is imposed on the SP structure. c) Region adaptation is performed on the boundary SPs d) After a number of iterations final spatial pyramid regions are obtained

Therefore, starting from the initial region geometry, boundary SPs are reassigned to the most similar neighboring regions.

3) During the structure update, the region statistics are recalculated based on the removed or merged boundary SPs. This update provides SP groups to adapt changes along the region boundaries and converge to compact and coherent region model. The boundary and structure update steps are iterated until the stopping criteria is met. Termination criteria can be set as a fixed number of iteration or it can be computed depending on the decrease in energy cost during the update step. Figure 5.8-d presents the generated region segments after the termination condition is met.

The optimization rule given in (5.16) updates the region boundary. Each boundary SP is visited and assigned to the region that provides maximum similarity. The proposed cost function used is composed of two main energy terms as denoted in (5.17). The first term relates the color similarity of the boundary SP to its neighboring regions. The second term defines the spatial distance of the SP to the region centers.

$$E(p,Q) = \lambda\ C(p,Q) + (1-\lambda)D(p,Q^c) \tag{5.17}$$

The term λ in (5.17) is a trade off parameter to be tuned depending on the content. Selection of λ imposes the geometry constraint on the generated region segments. As λ is increased, the input geometry constraint will be relaxed in favor of color similarity in the region. The value of 0.5 is used in all the experiments. This value is selected as a mid point between the imposed geometry constraint and region color similarity. *Lab* color space is utilized in the experiments due to its perceptual uniformity. Color distance is computed over the individual color channels i, see eqn (5.18).

$$C(p,Q) = \sum_{i=1}^{3} |\ p_i - Q_i|^2 \tag{5.18}$$

The spatial distance between the boundary SP p and the region centroid Q^c is computed using the geodesic distance. It is defined as the length of the shortest path from p to Q^c, as given in (5.19) [34].

$$D(p,Q^c)_G = min_{P=p_1,p_2,..,p_n} l(P) \tag{5.19}$$

Suppose $P = p_1, p_2, ...p_n = Q^c$ is a path between the SPs p_1 and $p_n = Q^c$ where p_i and p_{i+1} are connected neighbors. The path length $l(P)$, as defined in (5.20), is the sum of individual neighbor distances $d_N(p_i, p_{i+1})$ between adjacent SPs in the path.

$$l(P) = \sum_{i=1}^{n-1} d_N(p_i, p_{i+1}) \tag{5.20}$$

For the computation of adjacent SP distance d_N, three color channel (Lab) distance is utilized (5.21).

$$d_N(p, q) = \sum_{i=1}^{3} (p_i - q_i)^k \quad k = 1, 2 \tag{5.21}$$

No significant performance difference has been observed in the selection of k, hence, it is selected as 1 in all the experiments due to its computational efficiency.

Computation of the shortest path from the boundary SP to the region centroid is performed via the shortest path algorithm in [40]. At each iteration, shortest paths from the neighboring boundary SPs to the region centroid are computed. Since the termination criteria for path computation is at the boundary, calculation of the shortest paths over the whole image is avoided.

5.5.2 Experimental Results

The benefit of the proposed region segmentation method is evaluated as an improved pooling idea for image classification. The utilized training based classification method aims to assign the test samples in the dataset to one of the predefined classes. Similar to the previous section, PASCAL VOC 2007 [43] image classification dataset is used for the evaluation.

The details of the pipeline is previously explained in Section 5.3. Proposed region segmentation explores the pooling step of the pipeline in order to combine the responses of the encoded descriptors in a spatially coherent region structure. The common spatial pyramid pooling introduces a weak geometry in the encoding phase. The conventional image regions used in the pyramid are as follows: 1×1, 1×3 (three horizontal stripes), and 2×2 (four quadrants) grids. Concatenation of the fisher vectors from the eight spatial regions as in Figure 5.10, produces an image descriptor that is eight times the initial encoded size. The contribution of the proposed region segmentation is a replacement of this fixed image partitions. Instead, it is hypothesized that the spatial information in the pyramid segments can be better exploited by the proposed geometrically constrained region segmentation.

Mean average precision (MAP) scores of the 20 class dataset is presented in Table 5.3. The experiments are conducted for two different GMMs (128 and 256). Comparative tests for different geometry assignments and possible pyramid combinations are performed. (1×1, 3×1, 2×2 and possible combinations of these individual geometries).

	Number of GMMs			
	128		256	
Spatial Pyramid Type	Conventional	Proposed	Conventional	Proposed
1×1	52.90	52.90	54.87	54.87
3×1	55.40	55.71	57.09	57.55
2×2	53.71	54.57	55.60	56.98
$1 \times 1 + 3 \times 1$	56.40	56.88	58.17	58.28
$1 \times 1 + 2 \times 2$	55.28	56.12	57.02	58.05
$1 \times 1 + 3 \times 1 + 2 \times 2$	56.99	57.61	58.71	59.36

Table5.3: Pascal VOC classification results (MAP) with different spatial pyramid combinations

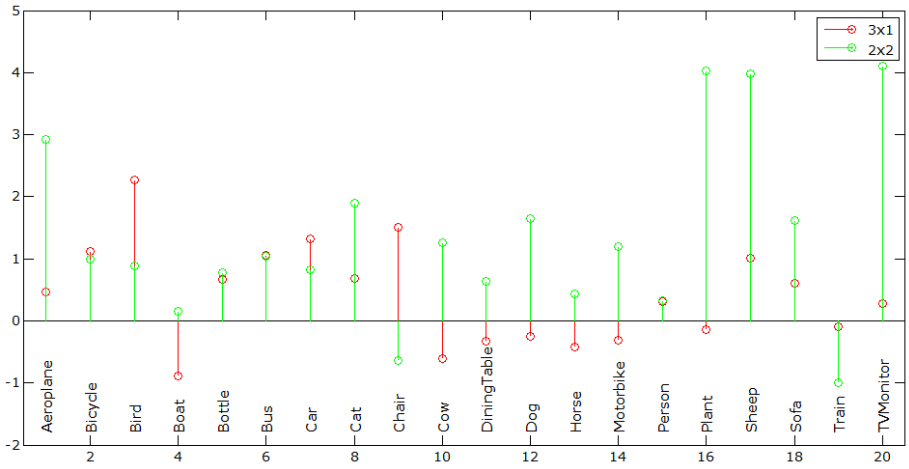


Figure 5.9: Change in AP scores for individual classes with the 3×1 and 2×2 geometry

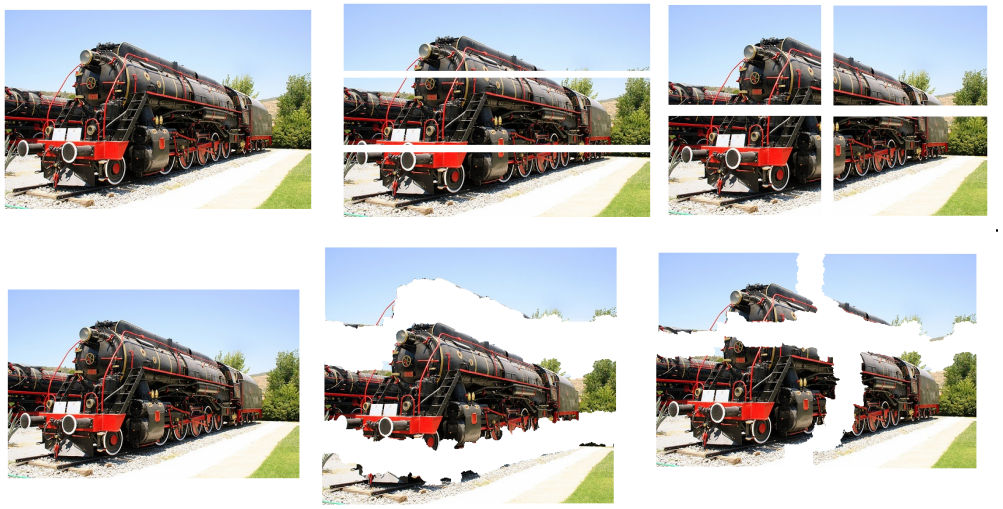


Figure 5.10: Spatial pyramid regions with conventional and proposed segmentation for 1×1, 3×1, 2×2 configurations

Standard deviation in MAP scores is observed to be less than 0.2%.

A detailed look at the change in the AP scores of the individual classes for the proposed segmentation might supply more information. Figure 5.9 shows the difference of the class specific average precision scores for the 20 Classes. Red corresponds to the change in AP for the 3×1 and blue for 2×2 spatial regions. One can observe that in the 3×1 geometry, 12 out of 20 classes have benefited form the proposed segmentation. On the 2×2 case, 18 out of 20 classes have increased accuracy. This result would show the advantage of proposed method especially for 2×2 geometry. One can also claim that the 2×2 geometry is inadequate for encapsulating region properties for the used Pascal VOC Dataset hence greater improvement is observed with the proposed region segmentation.

5.5.3 Discussion

A general observation of the MAP scores in Table 5.3 indicates that spatial pooling with the proposed region segmentation introduces a stable increase in the classification accuracy for all of the scenarios. The MAP results are observed to be consistent with the initial hypothesis that proposed region segmentation can better encapsulate local coherency compared to the conventional fixed region assignment. However, one can still argue that there is still room for improvement with different geometry selection.

Figure 5.11: Spatial pyramid regions with conventional and proposed segmentation for 1×1, 3×1, 2×2 configurations

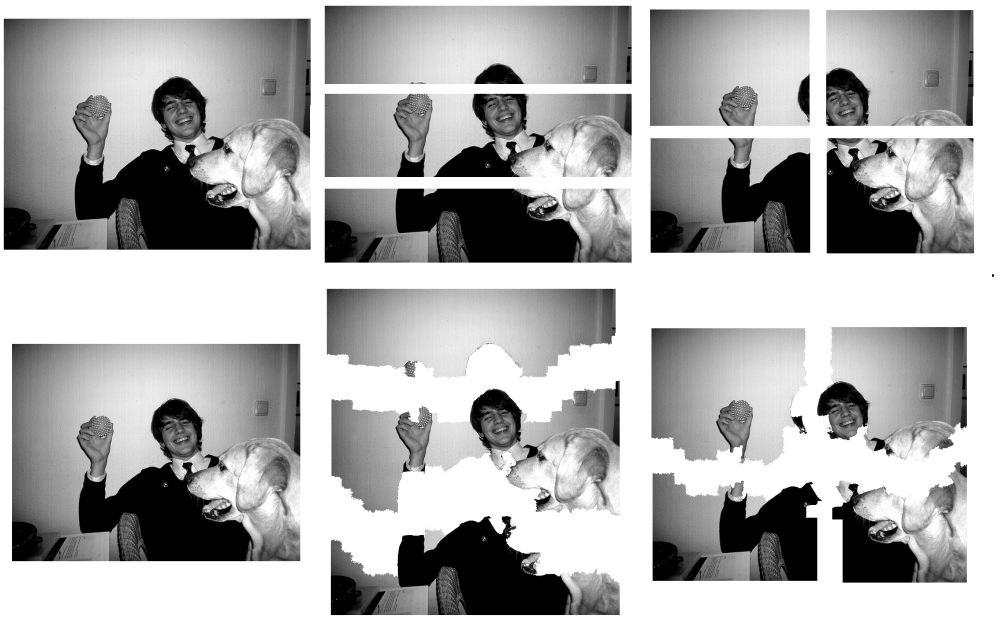


Figure 5.12: Spatial pyramid regions with conventional and proposed segmentation for 1×1, 3×1, 2×2 configurations

5.6 Conclusion and Future Work

In this section, the state-of-the-art object classification pipeline is presented. The application of the superpixel image segmentation on the different steps of the pipeline is explained. Primary focus is given on the feature extraction (first) and the spatial pooling (third) steps of the pipeline.

In the conventional dense feature extraction step, the mid-level region information is incorporated in order to obtain a superior scene description. The hypothesis that pixel based low-level descriptions are useful but can be further improved with the introduction of mid-level region information is evaluated. A novel SP based region descriptor that encapsulates the mid-level information is proposed. Image regions are described by computing the oriented mean differences between a central superpixel and its various orders of neighborhood. The variance of the neighbors is further included for a better description. The performance of the proposed descriptor is evaluated on the image classification task. For the experimental evaluations, baseline score is achieved using dense SIFT descriptors and 2.7% MAP improvement over the baseline is achieved. Based on the experimental evaluations, the initial hypothesis that mid-level cues enrich the image description and improve the performance of low-level cues has been verified.

As a future direction, the proposed method can be extended on different color spaces and channels. Utilization of different color spaces such as LAB, can improve the description power with its perceptual uniformity property. Moreover, in order to evaluate the matching performance of the proposed descriptor, quantitative results could provide an objective comparison with pixel based methods.

In an attempt to utilize region specific information, the spatial pooling step is also investigated. The spatial similarities in images for the purpose of object level image classification are explored. This has been achieved by an improvement on the spatial pyramid by adapting spatial regions with image characteristics. The method has been experimentally evaluated and the results show that the region adapted spatial segments improve over the image classification baseline. This also supports the intention to encapsulate spatial statistics using the proposed region based segmentation. Increase in the mean average precision scores have shown that coherent spatial regions would consistently improve performance for alternative scenarios. Sample visual results of the proposed segmentation have also been supplied for illustration of the region adaptation. Looking at the MAP scores one can still argue that there is still room for improvement with different geometry selection.

Individual class precision scores have shown that 2×2 geometry benefits more from the proposed segmentation. This can be related to inadequate region coherency in the 2×2 geometry. This indicates a future perspective towards exploration of scene geometry for region segmentation. Instead of the conventional fixed geometry, a preprocessing

stage for estimating the scene geometry can be performed. This might help obtaining flexibility both on the region borders and scene geometry.

CHAPTER 6

CONCLUSION

This thesis covers a wide range of fields and applications where the superpixel representation of an image is utilized. Superpixels are used in the graph based mono/stereo image/video segmentation framework. 3D related applications, e.g., 2D/3D conversion and disparity remapping, are explored. The results are evaluated both qualitatively and quantitatively with subjective and objective tests. Moreover, the superpixel primitives are presented as mid-level cues in the object recognition framework and a mid-level region descriptor is proposed. Moreover, adaptive spatial pooling using geometry based region segmentation has shown superior results with respect to mean average precision performance (MAP). [7]

6.1 Summary

After the introduction and outline of the thesis in Chapter 1, 2^{nd} chapter presents in detail an efficient superpixel (SP) and supervoxel (SV) extraction method. Improvements to the state-of-the-art in terms of both accuracy and computational complexity are observed. Segmentation accuracy is improved through convexity constrained distance utilization, whereas computational efficiency is achieved by replacing complete region processing by a boundary adaptation technique. Starting from the uniformly distributed, rectangular (cubical) shaped equal-sized superpixels (supervoxels), region boundaries are iteratively adapted towards the object edges. Region adaptation is performed by assigning the boundary pixels to the most similar neighboring superpixels. At each iteration, superpixel regions are updated; hence, progressively converging to compact pixel groups.

In Chapter 3, the general framework towards achieving user assisted image segmentation is presented. A multi-label object segmentation method is explained in detail; an application of the proposed image segmentation for 2D/3D purposes is discussed. The user assisted image segmentation is handled in a graphical domain where nodes are constructed by the superpixels. This yields a considerable amount of computational efficiency compared to pixel based approaches. Segmentation performance is satisfied

by the boundary adaptation power of the superpixels that are extracted using color and distance metrics. Image segmentation is approached as an energy minimization problem on a Markov Random Field (MRF). The goal of the proposed energy minimization technique is to achieve minimum energy potential labelling. The graph-cut method is used for this purpose on the superpixel primitives. The performance of the proposed technique is evaluated using objective metrics on a ground truth image segmentation dataset.

Chapter 4 presents the extension of the mono image segmentation on the stereo footage. As an application of the stereo image segmentation, a novel disparity remapping technique is proposed. The energy formulation as in mono segmentation is extended on the stereo data. This is realized using a feature based correspondence estimation followed by the energy minimization. The proposed technique is used as a post processing step for retargeting stereoscopic footage on different display sizes and resolutions. The performance of the proposed segmentation is evaluated using objective metrics on a ground truth image segmentation dataset. Moreover, the subjective user study is also conducted for evaluating the performance of the proposed disparity remapping technique.

Chapter 5 directs the attention to the image classification task. Different ways of utilizing superpixels in various steps of the image classification pipeline is investigated for this purpose. Initially, superpixel based mid-level region cues are incorporated in the feature description phase. Secondly, superpixel based region segmentation is proposed as a spatial pooling method where pooled regions are defined in accordance with the underlying image characteristics. Detailed quantitative results support the initial hypothesis that utilization of superpixels as mid-level information increases the recognition accuracy compared to only pixel based representations. Moreover, the adaptive spatial pooling method has also shown improvement over the conventional spatial pyramid technique.

6.2 Conclusion

The superpixel extraction method proposed in chapter 2 sets the backbone of the presented applications for image segmentation and classification. There has been major quantitative improvements with the proposed superpixel extraction method in terms of computational efficiency and segmentation performance. The iterative boundary adaptation idea and energy function selection are the two main contributions presented in this thesis. The success of the technique is verified with detailed quantitative experimental evaluations. Different energy metrics are utilized for evaluating the performance. The geodesic distance with LAB color space has shown the most accurate performance. On the other hand, Euclidean distance with RGB color space has also shown good performance and with a very high computation efficiency. With such ob-

servations, the proposed method proves to be a remarkable alternative for the current superpixel extraction techniques in the state-of-the-art.

The user assisted image segmentation framework is established using the superpixel primitives and graph-cut combinatorial optimization. The utilization of geodesic distance in the region similarity assignment yields advantages compared to Gaussian mixtude model (GMM) based region modelling techniques. Moreover, the proposed geodesic distance computation method provides an efficient information propagation technique on the superpixel lattice. The automatic extension of the user inputs on the stereo pair is useful with minimal user assistance. With additional user strokes, the proposed method is shown to generate outstanding results with efficient computation times.

The stereo segmentation technique is presented as a post processing step for retargeting stereoscopic footage on different display sizes and resolutions. By the help of the proposed disparity remapping technique, alternative stereo images with remapped disparity values are synthesized. To the best of our knowledge, utilization of stereo object segmentation for virtual depth adjustment purposes, has not been addressed before. With user experiments, the initial hypothesis is supported, i.e., the disparity adjusted images could provide superior visual experience. By the help of this technique, it becomes possible to change the object disparity such that accommodation-convergence conflict is less distracting.

Single image segmentation technique is further extended on the video footage where after the segmentation of the initial frame, succeeding video frames are automatically segmented. A learning based framework is utilized for modelling the segmented foreground and background regions. The large margin classifier support vector machine (SVM) is used to define the individual likelihood of the superpixels to the object or background region.

The utilization of superpixels as mid-level cues has been evaluated as a dense region descriptor for the purpose of object recognition task. This novel feature descriptor has been observed to increase recognition accuracy for alternating scenarios. As a conclusion of the experimented parameter settings, hierarchical fusion of different sized superpixels provide up to 3% increase as a mean average precision in the tested 20 classes of the Pascal VOC dataset. The results support the initial motivation that acquiring mid-level information from the superpixel primitives could carry a complementary information with respect to the low-level cues.

The advantage of proposed region adaptive spatial pooling step is experimentally validated using quantitative and visual results. The results support the main intention of encapsulating spatial statistics using the proposed geometry based region segmentation. Increase in the mean average precision scores have shown that coherent spatial regions would consistently improve performance for alternative scenarios. Sample vi-

sual results of the proposed segmentation have been supplied for illustration of the region adaptation.

6.3 Discussions and future directions

The proposed techniques in this thesis have been both qualitatively and quantitatively evaluated. Minor and major increases in the performance have been observed for individual tasks. In the final discussion section of this thesis, possible limitations of the individual methods and proposals for the future direction are presented.

Possible limitation of the proposed superpixel extraction technique could be observed when the initial superpixel size is too large. In that case, superpixel boundary adaptation might not be accurate as expected. A future direction in order to solve such issues might be to adaptively detect optimum superpixel size or raise a warning in the case that superpixels are not well adapted to the object boundaries. This can be done by detecting the edges beforehand and checking for an overlap between the obtained superpixel boundaries and the computed color/intensity edges. Alternatively, dividing the superpixels depending on the existing edges might also be a way to automatically overcome such issues. However, this might loosen the control on the total region size as an input parameter.

Extension of the image segmentation on the video footage is proposed as a proof of concept idea. Therefore, minimal evaluation has been addressed to that direction. Detailed investigation in this aspect can produce more accurate results. The proposed method could be objectively evaluated for alternating scenarios and video types. Moreover, a user study on the proposed 2D/3D conversion on the video footage can supply deeper understanding.

Stereo extension part of the method is quantitatively evaluated using the stereo segmentation dataset. However, a possible limitation could be observed in the information propagation part where user interacts only with one of the stereo images. In the proposed information propagation part, a feature matching based disparity estimation is realized. However, if the feature detection does not generate reliable results, the estimated average disparity could also produce erroneous values. This could easily be prevented by asking the user to supply interaction on the other stereo pair. This would decrease the proposed efficiency but could be easily tolerated for some erroneous scenarios.

As a future direction for the proposed mid-level feature descriptor, it can be further extended on different color spaces and color channels. Utilization of different color spaces such as LAB, can improve the description power with its perceptual uniformity property. Moreover, in order to evaluate the matching performance of the proposed superpixel descriptor, quantitative results could provide an objective comparison with

respect to the pixel based methods.

The results on the region adaptive spatial pooling indicate the fact that 2×2 geometry benefits more from the proposed segmentation. This can be explained by the weakest region similarity in the 2×2 geometry compared to other testes scenarios. This indicates a future perspective towards exploration of the scene geometry for region segmentation. Instead of the conventional fixed geometry, a preprocessing stage for estimating the scene geometry can be performed. This could help obtaining flexibility both on the region borders and scene geometry.

REFERENCES

[1] Adobe photoshop image editting software http://www.adobe.com/products/photoshop.html.

[2] The Foundry, http://www.thefoundry.co.uk/ (last accessed, june 2013).

[3] The GNU Image Manipulation Program (GIMP), http://www.gimp.org/.

[4] Methodology for the subjective assessment of the quality of television pictures. *ITU, Recommendation BT.500-11, 2002.*

[5] Philips bluebox 3d content creation service. *http://www.business-sites.philips.com/3dsolutions/3dcontentcreationproducts/bluebox.*

[6] Subjective assessment of stereoscopic television pictures. *ITU, Recommendation BT.1438, 2000.*

[7] ATSC Planning Team 1 Interim Report. *Advanced Television Systems Committee, Inc*, 2011.

[8] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012*, 2012.

[9] R. Adams and L. Bischof. Seeded region growing. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1994.

[10] A. Agarwala, A. Hertzmann, D. H. Salesin, and S. M. Seitz. Keyframe-based tracking for rotoscoping and animation. In *ACM SIGGRAPH 2004 Papers.*

[11] T. Asano and S. Kimura. Contour representation of an image with applications. *Proc. SPIE Vision Geometry*, 1995.

[12] S. S. Ayvaci, A. Motion segmentation with occlusions on the superpixel graph. *Proc. of the Workshop on Dynamical Vision, Kyoto, Japan, 2009.*

[13] X. Bai, J. Wang, D. Simons, and G. Sapiro. Video snapcut : Robust video object cutout using localized classifiers. *ACM SIGGRAPH*, 28, 2009.

[14] M. Banks, J. Read, A. R.S., and S. Watt. Stereoscopy and the human visual system. *SMPTE Motion Imaging Journal.*

[15] W. A. Barrett and A. S. Cheney. Object-based image editing. In *ACM SIGGRAPH*, 2002.

[16] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *ECCV 2006.*

[17] J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B*, 1974.

[18] E. Borenstein and S. Ullman. Combined top-down/bottom-up segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[19] G. Borgefors. Distance transformations in digital images. *Computer Vision, Graphics and Image Processing, 1986.*

[20] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *Computer Vision Pattern Recognition*, 2010.

[21] Y. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun. Ask the locals: multi-way local pooling for image recognition. In *International Conference on Computer Vision*, 2011.

[22] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:359–374, 2001.

[23] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–655, 1998.

[24] Y. Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images.

[25] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. *European Conference on Computer Vision*, 2010.

[26] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011.

[27] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. *British Machine Vision Conference*, 2011.

[28] C. Cigla and A. A. Alatan. Efficient graph-based image segmentation via speeded-up turbo pixels. In *International Conference on Image Processing, 2010.*

[29] C. Cigla and A. A. Alatan. Efficient edge-preserving stereo matching. *ICCV Workshop on Live Dense Reconstruction from Moving Cameras*, 2011.

[30] A. Coltekin. Foveation for 3d visualization and stereo imaging. In *PhD Dissertation, Helsinki University of Technology*, 2006.

[31] D. Comaniciu, P. Meer, and S. Member. Mean shift: A robust approach toward feature space analysis. *IEEE Pattern Analysis and Machine Intelligence*, 2002.

[32] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, 1995.

[33] A. Criminisi, , P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting, 2004.

[34] A. Criminisi, T. Sharp, and A. Blake. Geos: Geodesic image segmentation. In *European Conference on Computer Vision*, 2008.

[35] E. Dahlhaus, D. S. Johnson, C. H. Papadimitriou, P. D. Seymour, and M. Yannakakis. The complexity of multiterminal cuts. *SIAM Journal on Computing*, 23:864–894, 1994.

[36] É. Debreuve, M. Barlaud, J.-P. Marmorat, and G. Aubert. Active contour segmentation with a parametric shape prior: Link with the shape gradient. In *IEEE International Conference on Image Processing*, 2006.

[37] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. In *Journal of the Royal Statistical Society*, 1977.

[38] F. Devernay and S. Duchene. New view synthesis for stereo cinema by hybrid disparity remapping. *IEEE International Conference on Image Processing*, 2010.

[39] J. DiCarlo, D. Zoccolan, and N. Rust. How does the brain solve visual object recognition? *Neuron*, 73, 2012.

[40] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1959.

[41] E. A. Dinic. Algorithm for Solution of a Problem of Maximum Flow in a Network with Power Estimation. *Soviet Math Doklady*, 11:1277–1280, 1970.

[42] K. Eugene. Comparison of the OPENCV feature detection algorithms, http://computer-vision-talks.com, 2011.

[43] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 Results, 2010.

[44] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[45] C. Fehn. Depth-Image-Based Rendering, Compression and Transmission for a New Approach on 3D-TV. In *SPIE Stereoscopic Displays and Virtual Reality Systems XI*, 2004.

[46] D. Felleman and V. Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1991.

[47] P. Felzenswalb and D. Huttenlocher. Efficient graph based image segmentation. *International Journal on Computer Vision, 2004.*

[48] P. Felzenszwalb and D. Huttenlocher. Effcient belief propagation for early vision. *International Journal on Computer Vision, 2006.*

[49] B. Fernando, E. Fromont, and T. Tuytelaars. Effective use of frequent itemset mining for image classification. In *European Conference on Computer Vision*, 2012.

[50] L. Ford and D. Fulkerson. *Flows in Networks, Princeton Univ. Press, 1962.* Princeton Univ. Press, 1962.

[51] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach.* Prentice Hall Professional Technical Reference.

[52] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Pattern Analysis and Machine Intelligence*, 1984.

[53] A. V. Goldberg and R. E. Tarjan. A new approach to the maximum-flow problem. *Journal of the ACM*, 35(4):921–940, Oct. 1988.

[54] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *International Conference on Computer Vision*, 2005.

[55] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *International Conference on Computer Vision*, 2005.

[56] D. M. Greig, B. T. Porteous, and A. H. Seheult. Exact Maximum A Posteriori Estimation for Binary Images. *Journal of the Royal Statistical Society Series B*, 1989.

[57] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *International Conference on Computer Vision Pattern Recognition*, 2010.

[58] U. Gudukbay and T. Yilmaz. Stereoscopic view-dependent visualization of terrain height fields. In *IEEE Transactions on Visualization and Computer Graphics*, 2002.

[59] S. W. Hasinoff, S. B. Kang, and R. Szeliski. Boundary matting for view synthesis. *Computer Vision and Image Understanding, 2006.*

[60] M. Hassner and J. Sklansky. The use of markov random fields as models of texture. *Computer Graphics and Image Processing*, 12(4):357 – 370, 1980.

[61] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks. Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision*, 8, 2008.

[62] J. Hoffman, E. Rodner, J. Donahue, T. Darrell, and K. Saenko. Efficient learning of domain-invariant image representations. *International Conference on Learning Representations*, 2013.

[63] N. S. Holliman, G. Dodgson, N. Favalora, and L. Pockett. Three-dimensional displays: A review and applications analysis. *IEEE Transactions on Broadcasting, 2011.*

[64] A. Ion, J. Carreira, and C. Sminchisescu. Image segmentation by figure-ground composition into maximal cliques. In *International Conference on Computer Vision*, 2011.

[65] K. J., B. A., J. Y.J., and P. D. 2d-to-3d conversion by using visual attention analysis. In *SPIE Stereoscopic Displays and Applications*, 2010.

[66] N. Jojic, B. J. Frey, and A. Kannan. Epitomic analysis of appearance and shape. In *International Conference on Computer Vision*, 2003.

[67] J.-I. Jung and Y.-S. Ho. Depth map estimation from single-view image using object classification based on bayesian learning. In *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2010*, 2010.

[68] W. A. Kass, M. and D. Terzopoulos. Snakes: Active contour models. In *International Journal of Computer Vision*, 1988.

[69] Y. Ke and R. Sukthankar. PCA-SIFT a more distinctive representation for local image descriptors. In *Proceedings of International Conference on Computer Vision Pattern Recognition*, 2004.

[70] Kemeny, Knapp, and Snell. *Introduction to Markov Random Fields*. Springer, 1976.

[71] S. Knorr and T. Sikora. An ibr approach for realistic stereo view synthesis of tv broadcast based on structure from motion. In *IEEE International Conference on Image Processing (ICIP)*, 2007.

[72] M. Kunter, S. Knorr, A. Krutz, and T. Sikora. Unsupervised object segmentation for 2D to 3D conversion. SPIE, 2009.

[73] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross. Nonlinear disparity mapping for stereoscopic 3D. *ACM Transactions on Graphics*, 29(3):10, 2010.

[74] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition*, 2006.

[75] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. *International Conference on Computer Vision*, 2011.

[76] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi. TurboPixels: Fast Superpixels Using Geometric Flows. *Trans. Pattern Analysis and Machine Intelligence*, 2009.

[77] N. Li, D. D. Cox, D. Zoccolan, and J. DiCarlo. What response properties do individual neurons need to underlie position and clutter invariant object recognition. *Journal of Neurophysiology*, 2009.

[78] Y. Li, J. Sun, and H.-Y. Shum. Video object cut and paste. In *ACM SIGGRAPH*, 2005.

[79] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum. Lazy snapping. *ACM Transactions on Graphics*, 23:303–308, August 2004.

[80] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology, 1932.*

[81] S. Lloyd. Least square quantization in pcm. In *Proceedings of Transactions on Information Theory*, 1982.

[82] W. Lo, J. van Baar, C. Knaus, M. Zwicker, and M. H. Gross. Stereoscopic 3d copy & paste. *ACM Trans. Graph.*, 2010.

[83] D. Lowe. Object recognition from local scale-invariant features. *International Conference on Computer Vision*, 1999.

[84] D. Marr. Vision: A computational investigation into the human representation and processing of visual information. In *W.H. Freeman and Company, NY*, 1982.

[85] Y. Matsumoto, H. Terasaki, K. Sugimoto, and T. Arakawa. Conversion system of monocular image sequence to stereo using motion parallax. 1997.

[86] X. Mei, X. Sun, M. Zhou, S. Jiao, S. Wang, and X. Zhang. On building an accurate stereo matching system on graphics hardware. *ICCV Workshop on GPU for Computer Vision Applications, 2011.*

[87] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, T. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. In *International Journal of Computer Vision*, 2005.

[88] A. Moore, S. Prince, J. Warrell, U. Mohammed, and G. Jones. Efficient graph based image segmentation. *International Conference on Computer Vision and Pattern Recognition, 2008.*

[89] B. Morse, B. Howard, S. Cohen, and P. Price. Patchmatch-based content completion of stereo image pairs. *3DimPVT 2012.*

[90] E. N. Mortensen and W. A. Barrett. Intelligent scissors for image composition. ACM SIGGRAPH, 1995.

[91] J. Ning, L. Zhang, D. Zhang, and C. Wu. Interactive image segmentation by maximal similarity based region merging. *Pattern Recognition*, 2010. Interactive Imaging and Vision.

[92] P. M. Ojala, T. and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2002.

[93] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Trans. Pattern Analysis and Machine Intelligence*, 2002.

[94] R. Ono, H. Angus and P. Gregor. Binocular single vision achieved by fusion and suppression. *Perception and Psychophysics*, 1977.

[95] D. Parikh. Recognizing Jumbled Images: The Role of Local and Global Information in Image Classification. In *International Conference on Computer Vision*, 2011.

[96] F. Perronnin, J. Sanchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, 2010.

[97] N. Pinto, Y. Barhomi, D. Cox, and J. DiCarlo. Comparing state-of-the-art visual features on invariant object recognition tasks. In *IEEE Workshop on Applications of Computer Vision*, 2011.

[98] B. Price and S. Cohen. Stereocut: Consistent interactive object selection in stereo image pairs. *IEEE International Conference on Computer Vision.*

[99] B. Price, B. Morse, and S. Cohen. Livecut : Learning-based interactive video segmentation by evaluation of multiple propagated cues. *International Conference on Computer Vision*, 2009.

[100] B. L. Price, B. Morse, and S. Cohen. Geodesic Graph Cut for Interactive Image Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[101] D. R. Laws' texture energy in texture. In *Machine Vision: Theory, Algorithms, Practicalities, Academic Press*, 1997.

[102] L. Reese and W. A. Barrett. Image editing with intelligent paint. In *Proceedings of Euro-graphics*, 2002.

[103] X. Ren and J. Malik. Learning a classification model for segmentation. In *International Conference on Computer Vision*, 2003.

[104] E. Rosten and T. Drummond. Machine learning for high speed corner detection. *ECCV*, 2006.

[105] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH 2004*.

[106] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir. A comparative study of image retargeting. *Biometrika*, 29:160, 2010.

[107] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB : an efficient alternative to SIFT or SURF. *International Conference on Computer Vision*, 2011.

[108] N. Rust and J. DiCarlo. Ambiguity and invariance: Two fundamental challenges for visual processing. *Journal of Neuroscience*, 2010.

[109] J. Santner. Interactive multi-label segmentation. In *PhD Dissertation, Graz University of Technology*, 2010.

[110] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[111] D. Scharstein and R. Szeliski. Middlebury stereo repository. In *http://vision.middlebury.edu/stereo/*, 2010.

[112] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[113] A. Shamir and O. Sorkine. Visual media retargeting. *ACM SIGGRAPH ASIA 2009*, pages 1–13, 2009.

[114] E. Sharon, A. Brandt, and R. Basri. Fast multiscale image segmentation. In *Computer Vision Pattern Recognition*, 2000.

[115] J. Shi and J. Malik. Normalized cuts and image segmentation. In *IEEE Computer Vision and Pattern Recognition*, 1997.

[116] W. J. Tam, C. Vázquez, and F. Speranza. Surrogate depth maps for stereoscopic imaging: different edge types. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 2007.

[117] K.-H. Tan and N. Ahuja. Selecting objects with freehand sketches. In *International Conference on Computer Vision*, 2001.

[118] D. Tang, H. Fu, and X. Cao. Topology preserved regular superpixel. *IEEE International Conference on Multimedia and Expo*, 2012.

[119] E. Tasli and A. Alatan. User assisted stereo image segmentation. *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video*, 2012.

[120] E. Tasli and A. Alatan. Interactive disparity remapping for stereo images. *Accepted for Publication Signal Processing: Image Communication*, 2013.

[121] E. Tasli, C. Cigla, T. Gevers, and A. Alatan. Super pixel extraction via convexity induced boundary adaptation. *IEEE International Conference on Multimedia and Expo (ICME)*, 2013.

[122] H. E. Tasli and A. A. Alatan. Interactive object segmentation for mono and stereo applications : Geodesic prior induced graph cut energy minimization. *ICCV Workshop on Human Interaction on Computer Vision*, 2011.

[123] H. E. Tasli and K. Ugur. Interactive 2D 3D image conversion method for mobile devices. In *3DTV-Conference*, 2011.

[124] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *European Conference on Computer Vision*, 2010.

[125] A. Torralba, W. Murhpy, K. P. amd Freeman, and M. A. Rubin. Context based vision system for place and object recognition. In *International Conference on Computer Vision*, 2003.

[126] D. Tsai, M. Flagg, and J. Rehg. Motion coherent tracking with multi-label mrf optimization. In *British Machine Vision Conference*, 2010.

[127] S. Ullman. High-level vision : object recognition and visual cognition. In *The MIT Press, Cambridge, MA*, 1996.

[128] M. van den Bergh, X. Boix, G. Roig, B. de Capitani, and L. van Gool. Seeds: Superpixels extracted via energy-driven sampling. In *European Conference on Computer Vision*, 2012.

[129] J. C. van Gemert. Exploiting photographic style for category-level image classification by generalizing the spatial pyramid. In *ACM International Conference on Multimedia Retrieval*, 2011.

[130] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. In *European Conference on Computer Vision*, 2008.

[131] O. Veksler, Y. Boykov, and P. Mehrani. Superpixels and supervoxels in an energy optimization framework. In *Perspectives in neural computing*, 2010.

[132] A. Vevaldi and B. Fulkerson. Vlfeat an open and portable library of computer vision algorithms. In *Proceedings of ACM International Conference on Multimedia*, 2010.

[133] R. Vieux, J. Benois-Pineau, and J.-P. Domenger. Content based image retrieval using bag-of-regions. In *International conference on Advances in Multimedia Modeling*, 2012.

[134] L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Pattern Analysis and Machine Intelligence*, 1991.

[135] C. Wang and A. A. Sawchuk. Disparity manipulation for stereo images and video. *Proceedings of SPIE*, 2008.

[136] J. Wang, P. Bhat, R. A. Colburn, M. Agrawala, and M. F. Cohen. Interactive video cutout. *ACM SIGGRAPH*, 2005.

[137] B. Ward, S. B. Kang, and E. Bennett. Depth director: A system for adding depth to movies. *Computer Graphics and Applications, IEEE*, 2011.

[138] C. Ware. Information visualization - perception for design. In *Elsevier Inc. ,Morgan Kaufmann Publishers*, 2004.

[139] H. Wei, Q. Zuo, and B. Lang. A bio-inspired model for image representation and image analysis. *IEEE International Conference on Tools with Artificial Intelligence*, 2011.

[140] C. Wu, C. Li, Y. C. Lai, and L. Chen. Disparity remapping by nonlinear perceptual discrimination. *International Conference on 3D Systems and Applications*, 2011.

[141] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, 2002.

[142] C. Xu and J. Corso. Evaluation of super-voxel methods for early video processing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[143] T. Yan, R. W. H. Lau, and Y. Xu. Depth mapping for stereoscopic videos. *International Journal of Computer Vision, 2012.*

[144] W. Yang, J. Cai, J. Zheng, and J. Luo. User-Friendly Interactive Image Segmentation Through Unified Combinatorial User Inputs. *IEEE Transactions on Image Processing*, 2010.

[145] T. Yilmaz and U. Gudukbay. Stereoscopic urban visualization based on graphics processor unit. In *Optical Engineering*, 2008.

[146] G. Zeng, P. Wang, J. Wang, R. Gan, and H. Zha. Structure-sensitive superpixels via geodesic distance. *International Conference on Computer Vision, 2011.*

[147] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. In *International Journal of Computer Vision*, 2001.

[148] S. Zheng, Z. Tu, and A. Yuille. Detecting object boundaries using low-, mid-, and high-level information. In *Computer Vision Pattern Recognition*, 2007.

[149] S. W. Zucker. Region growing: Childhood and adolescence. In *Computer Graphics and Image Processing*, 1976.

Curriculum Vitae

H. EMRAH TAŞLI

emrah.tasli@gmail.com

PERSONAL INFORMATION

Date of Birth	1982
Place of Birth	İzmir - Turkey
Gender	Male

EDUCATION

2008 – 2013	Ph.D. Signal Processing Inst. Middle East Technical University, Ankara
2005 – 2007	MS, Communications Engineering, Munich Technical University, Munich
2000 – 2005	BS, Middle East Technical University, Ankara
	Major Program in Department of Electrical-Electronics Engineering
2003 – 2005	Minor Program in Mathematics,

LANGUAGES

Turkish	Native
English	Fluent
German	Good

SKILLS

Personal

- Strong analytical and problem-solving skills
- The big-picture & goal-orientation mindset
- Effective communication skills
- Very good documentation and technical writing skills

Programming

- Programming languages: C, C++, Matlab, Python, HTML
- OOP and Parallel computing in GPU, CUDA
- IDEs: MS Visual Studio, QT Creator, Eclipse, Matlab
- Algorithm parallelization for high performance computing
- Source Control: CVS, Mercurial, SVN, Rational ClearCase

Conceptual and Mathematical Field

- Computer Vision, Pattern Recognition, Artificial Intelligence
- Machine Learning, Statistical Signal Processing
- 2D/3D Image/Video Processing
- Classification and Clustering Methods
- Lossy/Lossless Data Compression
- Stereoscopy and Human Visual Perception
- Algorithm development/verification to product line support
- Real time constrained algorithm optimization
- Market experience with academic perspective

WORK EXPERIENCE & PREVIOUS PROJECTS

11/2012 – Current: **Post-Doc Researcher at University of Amsterdam,** Amsterdam, Netherlands

- Research & Development for Image/Video Classification
  - C/C++ and Matlab implementation
- Pattern Recognition, Computer Vision and Machine Learning on large scale data
- Supervision and teaching activities
  - Assisting Intellectual Multimedia Systems lecture
  - Supervision of three student profile projects
- Engaged in project proposals for research funds

01/2008 – 08/2012: **Design Architect at Vestel Technologies R&D,** Ankara, Turkey

- Algorithm Design for video post-processing engine (Pixellence)
  - Algorithm development with C/C++ Matlab
  - 2D/3D Image Enhancement (Contrast & Sharpness, Active Noise Reduction)
  - Frame Rate Conversion for Video Judder Removal
  - Local dimming and crosstalk reduction on LED 3D displays
  - Human Pose Estimation with Kinect
- Design to production line support
- Parallel computing on GPUs with Nvidia CUDA
- Project management and documentation

09/2010 – 12/2010: **Strategic Intern at Nokia Research Center,** Tampere, Finland

- Research on visual aspects of 3D visualization
- Development of user assisted 2D/3D image conversion method on mobile devices
- Algorithm implementation on the mobile platform with C++ using QT

01/2007 – 09/2007 **Master Thesis at Rohde & Schwarz,** Munich, Germany

- Software Solution for Video over IP transmission
- C++ Implementation on PowerPC Processor
- Video Data transfer through the IP on a RTOS
- Experience in network architectures; UDP, RTP, MPEG TS, FEC, Multicast, IGMP, ASM, SSM, RTOS, DVB

11/ 2006 – 01/2007 **Project "WCDMA uplink receiver implementation"**

- Design of a high SNR WCDMA uplink receiver
- Implementation in Matlab

08/2006 – 12/2006 **Imaging Scientist Infineon Technologies**, Munich, Germany

- Image processing on AIMS images
- Optimal proximity correction (OPC) in the optical lithography process
- Automatic reporting with MATLAB Report Generator

03/2006 – 07/2006 **Image Video Compression Laboratory**

- Lossy & Lossless image and video compression
- Matlab implementation of variable bit-rate video encoding and decoding
- Experience in DCT, DWT, Motion Compensation, MpegX, H.26X.

02/2006 – 03/2006 **Internship in Siemens AG**, Munich CT IC3, Germany

- Research on TPM (Trusted Platform Module) for mobile equipment integration
- Reporting and use scenario implementation

01/2005 –05/2005 **Bachelor Project**

- Audio/Video enabled human mimicking robot

- A software & hardware design to recognize & repeat human gestures

07/2004 – 08/2004 Internship, TUBITAK (Information Technologies and Electronics Research Institute)

- Automatic Video Text Detection for huge image databases.

Honors & Awards

DAAD (German Academic Exchange Service) Master Scholarship
METU honor student
Runner-up of the national mathematics Olympiads

Selected Publications

Journal Papers

[2013] H. Emrah Tasli and A. Aydin Alatan "Interactive Disparity Remapping for Stereo Images"; Elsevier Signal Processing: Image Communication; accepted.

[2013] H. Emrah Tasli, Cevahir Cigla, A. Aydin Alatan "Efficient Superpixel Extraction via Convexity Induced Boundary Adaptation"; submitted to Elsevier Image Vision Computing.

Conference Papers

[2013] H. Emrah Tasli, Ronan Sicre, Theo Gevers and A. Aydin Alatan, "Geometry Constrained Region Segmentation" Submitted to ACM International Conference on Multimedia 2013

[2013] H. Emrah Tasli, Cevahir Cigla, Theo Gevers and A. Aydin Alatan, "Superpixel Extraction via Convexity Induced Boundary Adaptation" International Conference on Multimedia and Expo 2013

[2012] H. Emrah Tasli and A. Aydin Alatan, "User Assisted Stereo Image Segmentation" 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 3DTV-CON 2012

[2011] H. Emrah Tasli and A. Aydin Alatan, "Interactive object segmentation for mono and stereo applications: Geodesic prior induced graph cut energy minimization." ICCV 2011, Workshop on Human Interaction on Computer Vision.

[2011] H. Emrah Tasli.; Murat Sayinta ; A. Aydin Alatan, " Pixel-Wise Intensity Compensation for Locally Dimmed Backlight Displays Based on an Objective Metric" Society of Information Display Symposium, SID 2011

[2011] H. Emrah Tasli and Kemal Ugur, "Interactive 2D 3D image conversion method for mobile devices," 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 3DTV-CON 2011

Local Conference Papers

[2013] H. Emrah Tasli, Theo Gevers and A. Aydin Alatan, "User Assisted Stereo Image Retargetting" IEEE Conference on Signal Processing and Communications Applications, SIU 2013

[2012] H. Emrah Tasli and A. Aydin Alatan, "User Interactive Segmentation Method on Stereo Images" IEEE Conference on Signal Processing and Communications Applications, SIU 2012

[2012] H. Emrah Tasli, and A. Aydin Alatan, "Crosstalk Reduction for Pattern Retarder Stereoscopic Displays" IEEE Conference on Signal Processing and Communications Applications, SIU 2012

[2012] Cevahir Cigla, H. Emrah Tasli, A. Aydin Alatan, "Efficient Superpixel Extraction for Image Segmentation" IEEE Conference on Signal Processing and Communications Applications, SIU 2012.

[2011] H. Emrah Tasli and A. Aydin Alatan, "Pixel compesation for locally dimmed backlight displays," IEEE Conference on Signal Processing and Communications Applications, SIU 2011

Patents

[2012] H. Emrah Tasli "A Method for Reducing Crosstalk in Stereoscopic Displays and Display Systems" Pending

[2011] Burak Ozkalayci, H. Emrah Tasli "Superresolution enhancement method for N-view and N-depth multiview video" EP2369850A2

[2011] Burak Ozkalayci, H. Emrah Tasli "Superresolution based N-view N-depth multiview video coding" EP2373046A2

[2011] H. Emrah Tasli "A method for local dimming boost up using depth map", TR201100485

[2011] H. Emrah Tasli "Edge led local dimming" TR20110038

[2011] H. Emrah Tasli, Cevahir Cigla "A method for local dimming boost using salient features" EP2372686A1

[2010] H. Emrah Tasli " Brightness correction for LCD displays with backlight modulation" EP2312567A1

PhD Thesis

[2013] "Superpixel-based Efficient Image Representation for Segmentation and Classification" Advisor: Prof. A. Aydin Alatan, Middle East Technical University, Ankara, Turkey.

Master Thesis

[2007] "A software solution for MPEG2 Datastream delivery over IP through Gigabit Ethernet (ASI over IP)", Advisor: Prof. Eckehard Steinbach, Munich Technical University, Germany