

ANALYSIS OF BORDER OWNERSHIP CUES AND IMPROVEMENT OF
DEPTH PREDICTION USING BORDER OWNERSHIP

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

BY

MEHMET AKIF AKKUŞ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2014

Approval of the thesis:

**ANALYSIS OF BORDER OWNERSHIP CUES AND IMPROVEMENT OF
DEPTH PREDICTION USING BORDER OWNERSHIP**

submitted by **MEHMET AKIF AKKUŞ** in partial fulfillment of the requirements for
the degree of **Master of Science in Computer Engineering Department, Middle
East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı
Head of Department, **Computer Engineering**

Assist. Prof. Dr. Sinan Kalkan
Supervisor, **Computer Engineering Department, METU**

Examining Committee Members:

Prof. Dr. Fatoş Yarman Vural
Computer Engineering Department, METU

Assist. Prof. Dr. Sinan Kalkan
Computer Engineering Department, METU

Assist. Prof. Dr. Ahmet Oğuz Akyüz
Computer Engineering Department, METU

Assoc. Prof. Dr. Alptekin Temizel
Information Systems, METU

Assist. Prof. Dr. Erkut Erdem
Computer Engineering Department, Hacettepe University

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: MEHMET AKIF AKKUŞ

Signature :

ABSTRACT

ANALYSIS OF BORDER OWNERSHIP CUES AND IMPROVEMENT OF DEPTH PREDICTION USING BORDER OWNERSHIP

Akkuş, Mehmet Akif

M.S., Department of Computer Engineering

Supervisor : Assist. Prof. Dr. Sinan Kalkan

September 2014, 59 pages

Border Ownership is the problem of identifying which image regions own the image border. This information is essential and important for large variety of high-level vision problems such as object segmentation, object recognition, depth perception, motion perception etc. Current computational approaches to Border Ownership (BO) estimation either use artificial or limited number of real images. In this thesis, we propose a new comprehensive BO database, including 500 indoor and 500 outdoor images whose BO information is labeled by human participants. Using this dataset, BO estimation capability is investigated for several visual cues such as *T-junction*, *L-junction*, *curvature*, *lower region* and *contrast* both individually and combinatorially. Then using these cues, a basic computational model is proposed which estimates BO information based on the majority rule. Moreover, a new method, which merges a feature-based stereo algorithm and BO information, is proposed for more accurate depth prediction on homogeneous areas.

Keywords: border ownership, cue analysis, image database, depth prediction

ÖZ

SINIR SAHİPLİĞİ İPUÇLARININ ANALİZİ VE SINIR SAHİPLİĞİ BİLGİSİ İLE DERİNLİK TAHMİNİNİN İYİLEŞTİRİLMESİ

Akkuş, Mehmet Akif

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi : Yrd. Doç. Dr. Sinan Kalkan

Eylül 2014 , 59 sayfa

Sınır sahipliği resimdeki alanların hangi resim sınırlarını sahiplendiğini belirleme problemidir. Bu bilgi nesne bölütleme, nesne tanıma, derinlik algılama, hareket algılama gibi bir çok üst seviye görme problemi için gerekli ve önemlidir. Sınır Sahipliği (SS) için kullanılan halihazırdaki hesaplama yaklaşımları ya yapay ya da yetersiz sayıda gerçek dünya imgesiyle çalışmaktadır. Bu tezde, 500 iç alan 500 dış alan resmi içeren geniş kapsamlı yeni bir SS veritabanı hazırlandı. Bu veritabanı kullanılarak T-kesişim, L-kesişim, eğrilik, alt-alan ve zıtlık gibi değişik bilgisayarlı görü ipuçlarının bireysel ve katımsal olarak SS tahmin etmedeki yetkinlikleri sınandı. Ve bu ipuçları kullanılarak çoğunluğu esas alan temel bir SS tahmin eden bir hesaplamalı model geliştirildi. Bunlara ek olarak, nitelik tabanlı bir stereo metodu ile SS bilgisi birleştirilerek imgedeki homojen alanlarda daha doğru sonuçlar üreten bir derinlik tahmin metodu önerildi.

Anahtar Kelimeler: sınır sahipliği, ipuçlarının analizi, imge veritabanı, derinlik tahmini

To my family

ACKNOWLEDGMENTS

Firstly, I would like to thank my advisor Prof. Sinan Kalkan for his endless guidance, support and especially patience during my studies. He has become a great role model as an academician for me who is a freshman researcher. I have learnt a lot from him.

Many people contributed to the results presented in this thesis. First, I have to thank Gaye Topuz and Buğra Özkan for their joint-work during the thesis. Then, I have to thank Hande Çelikkanat, with whom I resolved most problems related to programming, Unix and compilation issues. I can not forget to thank Güner Orhan for being a great helper, classmate and project mate. I also thank Fatih Gökçe and Osman Tursun for their companionship and joyful conversations.

I am grateful to my colleagues at the Computer Engineering Department. Alperen Eroğlu and Cağrı Kaya helped me especially in my rush hour for my departmental duties. Without their help, especially final part of my thesis, it would not have been realized in time.

My special thanks goes to Selma Suloğlu for asking me my progress and pushing me to study all the time.

My very special thanks goes to Kadir Fırat Uyanık for being a great social friendship, especially for MTB and swimming.

I would not advance if my parents do not support me all the time, especially last three months of my thesis work. I would like to thank also old members of the KOVAN lab: Fariba Yousefi, Erinç İnci, Sertaç Olgunsoylu, Mustafa Parlaktuna, Mustafa Mızrak, Yiğit Çalışkan, Onur Yürüten and Asil Kaan Bozcuoğlu for making the lab a great lab.

I also thank all the members of the jury whose comments were of great help.

This work was partially funded by the Turkish National Science and Technology Organization (TUBITAK) under the project no 111E155.

TABLE OF CONTENTS

ABSTRACT	v
ÖZ	vi
ACKNOWLEDGMENTS	viii
TABLE OF CONTENTS	ix
LIST OF TABLES	xii
LIST OF FIGURES	xiii
LIST OF ABBREVIATIONS	xvii
CHAPTERS	
1 INTRODUCTION	1
1.1 Motivation	2
1.2 Contributions	2
1.3 Outline of the thesis	5
2 REVIEW OF LITERATURE	7
2.1 Border Ownership in Neuroscience and Psychology	7
2.2 Border Ownership in Computer Vision	8
2.3 Border Ownership Datasets in the literature	10

3	ONLINE LABELING TOOL & BORDER OWNERSHIP DATASET	13
3.1	Online Labeling Tool	13
3.1.1	Registration to the System	14
3.1.2	Explanations for Border Ownership	15
3.1.3	A Tutorial Before Labeling	15
3.1.4	Labeling Page	17
3.2	Border Ownership Dataset	18
3.2.1	Data to Be Labeled	18
3.2.2	Participants	18
3.3	Analysis of the Dataset	20
3.4	Summary	22
4	BORDER OWNERSHIP CUES AND THEIR ANALYSIS	23
4.1	Used Border Ownership Cues	23
4.1.1	Lower Region	23
4.1.2	Curvature	24
4.1.3	T-junction	26
4.1.4	L-junction	27
4.1.5	Contrast	28
4.2	Analysis of the Cues	29
4.2.1	Accuracies of Visual Cues	30
4.2.2	Accuracies of Combined of Visual Cues	31

4.2.3	Analysis of the Conflicting Cues	32
5	MORE ACCURATE DEPTH PREDICTION USING BORDER OWNERSHIP	37
5.1	Depth extracting cues	37
5.2	Depth prediction on homogeneous areas using sparse stereo methods	38
5.3	Formulation of the depth prediction model	39
5.3.1	Bounding edge determination of a mono	41
5.3.2	Voting model for a mono	42
5.4	Adding Border Ownership information to the voting model	43
5.5	Results	45
5.5.1	Results on artificial data	47
5.5.1.1	Evaluation with texture noise	47
5.5.1.2	Evaluation with texture and pepper noise	48
5.5.2	On a real world image	49
5.6	Time issues	52
5.7	Summary	53
6	CONCLUSION	55
6.0.1	Future Work	56
	REFERENCES	57

LIST OF TABLES

TABLES

Table 2.1	Border Ownership Literature on Computer Vision	10
Table 4.1	Occurrences and availability of the cues on all images. Cues are ordered according to number of available borders for a certain cue type. . .	30
Table 4.2	The analysis of different labeling with image type. 2/3 stands for participants are not labeled the same region, 3/3 stands for all three participants labeled the same region	30
Table 4.3	Visual cues and their predictions compared against different consistencies of labels.	32

LIST OF FIGURES

FIGURES

Figure 1.1 The red line is a border between two regions and only one regions owns it. [Best viewed in color]	2
Figure 1.2 Image shows the difference on a artificial image when the BO information is used. Using BO information, extracted surfaces are more accurate as positioned closer to the edges. Left image shows the depth prediction using BO information and right images shows depth prediction with BO information.	3
Figure 2.1 BO cells with the same contrast values in the left and the right halves.	8
Figure 2.2 Basic images used as Border Ownership Datasets in the literature . .	11
Figure 2.3 Sample wrongly-drawn borders in BSDM. [Best viewed in color] .	12
Figure 3.1 Home page of the online labeling tool	14
Figure 3.2 Registration page of the online labeling tool	15
Figure 3.3 Users are trained with real indoor and outdoor images	16
Figure 3.4 Tutorial page on online labeling tool. For a set of simple borders, users are informed whether they labeled right or wrong. [Best viewed in color]	16
Figure 3.5 Labeling page with instructions at the top and a panel at the right shows the progress. <i>Undo</i> button gives a chance to undo the latest click operation, <i>I am not sure</i> button is so as to pass current border if user is not sure enough to label one of the regions, lastly, <i>Drawing On/Off</i> button enables/disables blue and region	17

Figure 3.6	Segment annotation tool is used for creating segments of an image. Left image is the one that will be segmented. Region with red border shows segmented parts till then, Region with yellow points are currently segmented object. Right image shows current segments.	19
Figure 3.7	Various borders in a sample indoor image. [best viewed in color] .	19
Figure 3.8	Various borders in a sample outdoor image. [best viewed in color]	20
Figure 3.9	Consistencies according to age (a) and gender (b). Color legend of two types of images (indoor and outdoor) and for all images in total are listed on the right. Total number of participant is shown in paranthesis [Best viewed in color]	21
Figure 3.10	Consistency ratios of indoor and outdoor images.	22
Figure 4.1	Sample image shows concept of lower region. Image at right corresponding segmented image of the left image. As explained, R1 is at bottom and the closest object, then house, tree and sky accordingly. . . .	24
Figure 4.2	Sample calculation of magnitude of a curvature on a certain point p_1 and p_2 . Note that radius and curvature direction changes along the curvature.	24
Figure 4.3	Sample curvatures on an image. Dashed red lines are the curvatures and arrows show the direction of the ownership.	25
Figure 4.4	Representation of T junction. Note that T-junction has three lines segments.	26
Figure 4.5	Sample T-junctions and ownership directions on an artificial image. Left image: regions are labeled and two T-junctions are shown. Right image: labeled borders with their ids. According to T-junction cue border b_0 and b_1 belong to region R_1 , border b_3 and b_4 belong to region R_3 , border b_5 and b_6 belong to region R_4 etc.	27
Figure 4.6	Sample T-junctions and ownership directions on sample indoor and outdoor image. Dashed rows show the ownership direction [Best viewed in color].	27
Figure 4.7	Representation of L junction. Note that L-junction has two line segments.	28
Figure 4.8	Sample L-junctions and ownership directions on sample indoor and outdoor image. Dashed rows show the ownership direction [Best viewed in color].	28

Figure 4.9 For some borders, it may not be so obvious to determine ownership of the border. This might be due to wrong segmentation by humans (a) or that the borders might seem shared by two objects (b). [Best viewed in color]	31
Figure 4.10 Combination of different cues on indoor and outdoor images. The x-axis shows the number of combined cues and y-axis shows the accuracy by percentage	33
Figure 4.11 Conflicting cues in indoors. (a) A matrix depicting which cues conflict, and the <i>winning cue</i> in case of a conflict. (b) Explanation of each cell in the matrix.	34
Figure 4.12 Conflicting cues in outdoors. (a) A matrix depicting which cues conflict, and the <i>winning cue</i> in case of a conflict. (b) Explanation of each cell in the matrix.	35
Figure 5.1 Extracted primitives (b) for the example image in (a). Magnified edge primitives and edge primitives together with monos are shown in (c) and (d) respectively. Source: [17]	40
Figure 5.2 A set of primitives for illustrating why the relations co-planarity, co-colority and linear dependence are required as restrictions for forming pairs from edges. Source: [17]	41
Figure 5.3 Illustration of how the vote of a pair of edge primitives is computed. The 3D primitives Π_i^e and Π_j^e corresponding to the 2D primitives π_i^e and π_j^e define the plane p . The intersection of p with the ray l that goes through the 2D mono π^m and the camera center C then determines the position of the estimated 3D mono Π^m . The 3D orientation of Π^m is set to be the orientation of the plane p . Source: [17]	42
Figure 5.4 The distribution of the votes for some monos shown in the 3D Euclidean space. Sub-figures (a)-(c) are clear examples where there are two distinct clusters. However, such clear clusters do not always occur (d). Source: [17]	43
Figure 5.5 Stereo image pair from an artificially created scene. [Best viewed in color]	44
Figure 5.6 Regions and corresponding contours (border). Each border and region have unique id. [Best viewed in color]	45

Figure 5.7	Left part is the scene and corresponding depth image. In rectangle, left images shows bounding edges for a certain mono, right images denote voting edges after BO information is applied. Note that some edges primitives are disappeared due to their wrong prediction vote.	46
Figure 5.8	Artificial image with different amount of texture noise. [Best viewed in color]	47
Figure 5.9	Root Mean Squares and Bad Mad Matching Percentage metrics are used with different amount of texture noises.	47
Figure 5.10	Artificial image with different amount of texture and pepper noise. [Best viewed in color]	48
Figure 5.11	Root Mean Squares and Bad Mad Matching Percentage metrics are used with different amount of texture and pepper noises.	48
Figure 5.12	Stereo image pair taken from a Bumblebee stereo vision camera. In the scene, basic colors are used for easing the segmentation process. [Best viewed in color]	49
Figure 5.13	Image shows the difference on a real world image when the BO information is used. Using BO information, extracted surfaces are more accurate as positioned closer to the edges.	50
Figure 5.14	Image shows the difference on a artificial image when the BO information is used. Using BO information, extracted surfaces are more accurate as positioned closer to the edges. Left image shows the depth prediction using BO information and right images shows depth prediction with BO information.	51
Figure 5.15	Elapsed time differs for the depth prediction when the BO information is used. For a mono, elimination of edge primitives, which are in the occluding area or on unrelated edges, shortens elapsed time.	52

LIST OF ABBREVIATIONS

BO	Border Ownership
FG	Figure-Ground
GUI	Graphical User Interface

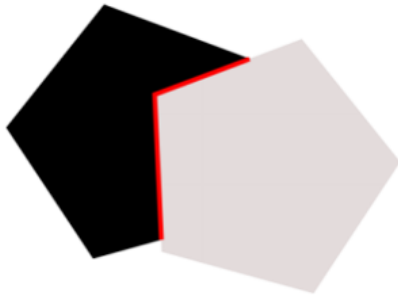
CHAPTER 1

INTRODUCTION

Projection of the three dimensional (3D) world onto electro-optical sensors yields images, and this projection process causes information loss. Most common computer vision problems such as stereo vision and optical flow need non-homogeneous image structures, since they need to find correspondence between different image regions. Similarly, in biological systems, neurons in our visual cortex are not activated on regions with uniform intensities, because they do not cause any change in the receptive fields [25, 15]. This suggests that biological vision system utilizes some supplementary information, because we, humans, can perceive the 3D information available in image areas that have uniform intensities.

It has been shown that, in biological vision systems, incomplete visual information can be completed by reliable visual information available at the borders [4, 5, 8, 11, 34, 22]. For perceiving depth and motion, the valuable information at the borders are diffused/filled in to image areas which mostly have uniform intensities [25, 15]. However, this diffusion process requires determining towards which region the information at a border should be diffused: i.e., the ownership of each border by the regions need to be known in advance. Called Border Ownership (BO), this information links regions and borders in an image.

What has been explained until now can be demonstrated in Figure 1.1, which shows two examples of Border Ownership. In Figure 1.1(a) there are two pentagons and the red line shows the border between them. Here, the border (red line) belongs to the pentagon which is at right. In Figure 1.1(b), the red line belongs to the object which lies at the bottom.



(a) Border belongs to white pentagon



(b) Border belongs to yellow object

Figure 1.1: The red line is a border between two regions and only one regions owns it. [Best viewed in color]

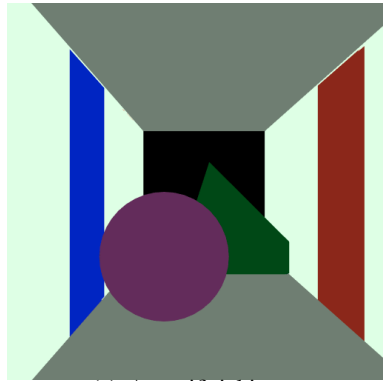
1.1 Motivation

Current datasets for border ownership include a small number of images, based on only one category and one or two participants for each border. For a more coherent analysis, more images with different characteristics should be analyzed. Therefore, in this thesis, we constructed a dataset including 500 indoor and outdoor images with several categories which differ from natural scene to office images. Moreover, it has labels by three participants for each border.

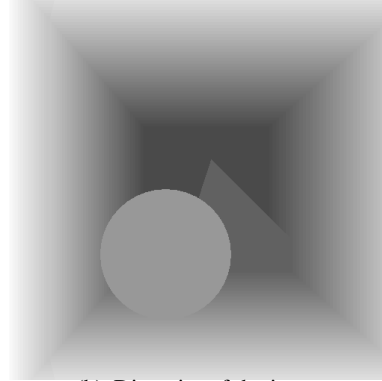
In computer vision, stereo needs to overcome the correspondence problem. While correspondence is easier to find on the edge-like structures, it is almost impossible at weakly-textured or constant intensity image areas. Therefore, disparity information at this kind of homogeneous places is not accurate. At this point, disparity information can be estimated using bounding edges of homogeneous place, but to some extent. The main reason for that is the lack of information of border ownership. For the depth estimation of a point on such an area, entire of the bounding edges can not be used. Only the borders which region owns can be used. Therefore, BO information is essential for such a problem.

1.2 Contributions

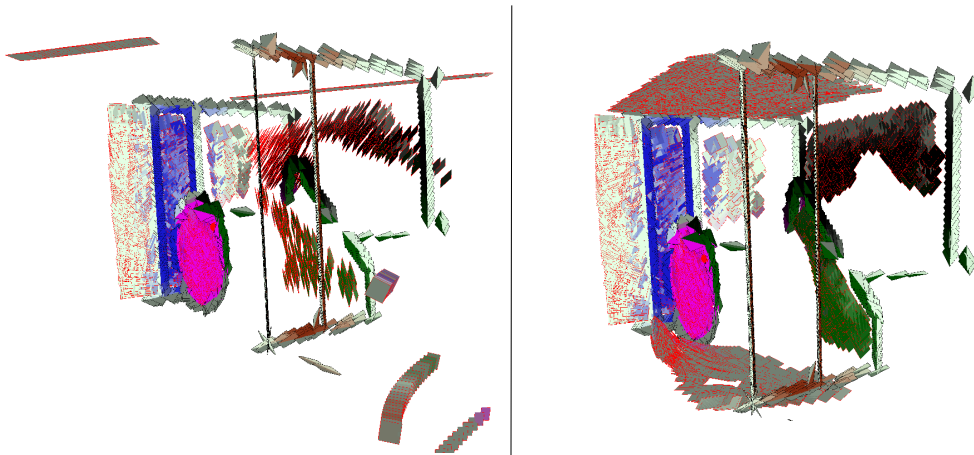
This thesis makes the following contributions:



(a) An artificial image



(b) Disparity of the image



(c) Example scene is shown from the artificial image. (on a 3D visualizer)

Figure 1.2: Image shows the difference on a artificial image when the BO information is used. Using BO information, extracted surfaces are more accurate as positioned closer to the edges. Left image shows the depth prediction using BO information and right images shows depth prediction with BO information.

- A new comprehensive database with 500 indoor and 500 outdoor images. The BO database includes various types of indoor (shopping mall, living room, train station etc.) and outdoor images (animals, historical places, landscapes etc.) with 1000 images in total. Each contour in them are labeled by three different participants to get more accurate human labeled ground-truth database. Furthermore, a user-friendly, well-documented on-line BO labeling tool is created (see Chapter 3.1). Participants used this tool for labeling. On the other hand, another tool named segment annotation tool (see Figure 3.6) and on-line ownership labeling tool give rise to expand dataset with new images.
- Analysis of the local visual cues for border ownership estimation is performed. Those cues are: *lower region, curvature, contrast, T-junction, L-junction*. Each cue is analyzed in terms of their predictive capability.
- Different number and types of cues are combined so as to create more accurate prediction results. Analysis of conflicting cues are also performed. Results are showed in a conflicting matrix.
- A new method merges BO information and a feature based stereo has been developed to obtain more accurate depth perception on homogeneous image areas.

These contributions have been disseminated in the following:

- (Submitted) **Mehmet Akif Akkuş**, Buğra Özkan, Gaye Topuz and Sinan Kalkan, Analysis of Visual Cues and a Computational Model for Border Ownership, *Computer Vision and Image Understanding*.
- **Mehmet Akif Akkuş**, Gaye Topuz and Sinan Kalkan, A Comprehensive Database for Border Ownership, in *IEEE 21th Conference on Signal Processing and Communications Applications (in Turkish)*, 2013

1.3 Outline of the thesis

The outline of the thesis is as follows: The next chapter details background information and literature survey. In chapter 3 , we describe our on-line BO labeling tool and obtained database using this tool. This chapter also makes analysis of the database in terms of consistency. Then chapter 4 introduces visual BO cues and their analysis in terms of prediction capability of the cues. Moreover, it includes analysis of conflicting BO cues. Chapter 5 describes how BO information and feature based stereo algorithm merged to get more accurate depth estimation on homogeneous image areas. Chapter 6 concludes the thesis with a discussion and future work.

CHAPTER 2

REVIEW OF LITERATURE

In this chapter, literature review of Border Ownership is divided into three parts: Psychology and Neuroscience, Computer Vision and the BO database.

2.1 Border Ownership in Neuroscience and Psychology

Border Ownership (BO) selective neurons in human visual system have been studied for the recent decade. Zhou et.al. [36] have found that 18% of the cells in V1 and more than the half of the cells in V2 and V4 respond to the according to direction of the owner of the boundary. Similarly, Qui et.al. [29] have discovered that V2 and V4 brain areas are asserted to be involved in determination of direction of the owner.

There are many studies which claim that visual cues play an important role in the BO estimation process such as contrast [23], depth order [30] and curvature [13]. Another interesting finding [23] is that different types of intensity changes activates the different area of the brain. For instance, the V2 cells activated more when the light gray patch on a dark ground is attached to the left of the figure compared to the dark gray patch on the right (see Figure 2.1).

Qui and Heydt show that selective neurons to the corners in 2D images in V2 regions of macaque monkeys' visual cortex are also selective to the depth order of 3D images and shows larger activation by the near side of the figures.

Among the cues which are mentioned in the previous paragraphs, as also analyzed in this study, curvature is the one of the most informative visual cue. Zhou et.al.



Figure 2.1: BO cells with the same contrast values in the left and the right halves.

[36] have stated that in V2 and V4 regions in the monkey visual cortex BO selective neurons are activated when the borders of figures are convex. As a matter of fact that, particularly V4 neurons whose receptive fields are large and curved are pointed out to be sensitive the concave and convex image areas.

There are some experimental studies which have show that latency of BO selective neurons is 10-25 seconds. This fact indicates that BO information can be determined using local cues rather than junction detection and junction evaluation.

Psychophysical studies have also been involved the BO problem and support what neuroscience put forward considerably. For instance, it has been figured out that human subjects show a strong preference for perceiving convex visual regions as being closer to the viewer compared to concave part [27]. It is also shown that lower-region, T-junctions, L-junctions are related to the BO problem. Vecera et.al. [35] have shown that the relationship between some of the Gestalt rules. They showed that humans are more likely to perceive lower region as figure compared to upper regions. Furthermore, figures are perceived as occluders as being closer to the human subject than the background. T-junctions and L-unctions are important cues as occlusion features.

2.2 Border Ownership in Computer Vision

Border Ownership studies regarding to computer vision are not very common in the literature except the last two decades. First computational studies for modeling BO are neural network based, which use L-junctions and curvatures as BO cues, and used

local propagation and competition mechanism along contours [20, 21]. In a similar manner, Nishimura et. al. [26] proposed another computational model utilizing neural network to show importance of surrounding contrast for BO determination. They achieved this by using artificial images for training and testing.

For the BO problem, Fowlkes et.al [10] utilized size, lower region and convexity cues for the figure/ground segregation. They combined their cues with logistic regression. In their study, combination of the size and lower region were much more powerful than combination of size and convexity cues. Moreover, they stated that size and convexity cues are often related since locally smaller regions are also locally convex. Another finding which they discovered is that junctions are important for occlusion and depth order.

Another model from the Fowlkes et. al. [31] is that based on the shapeme representation. They developed a logistic classifier to locally predict figure/ground labels. Conditional random fields is used to enforce global consistency by learning T-junction frequency and continuity after local findings. They used loopy belied propagation for inference. Their model outperforms when compared to a basic model using size and convexity cues. Using 100 images of Berkeley Image Segmentation dataset, they reported 78.3% accuracy.

Leichter et.al. [24] created a model for Border Ownership prediction using Conditional Random Fields (CRF). In their model they utilized curves and T-junction as a cue and formulated CRF using learnt non-parametric distributions of those cues. Their method significantly improved the currently achieved figure/ground assignment accuracy with 20.7% fewer errors in the Berkeley Image Segmentation Dataset.

In very recent years, Chen et.al. [6] used 5 visual cues:compactness, semantic, position, junction and convexity. They used Adaboost to train their occlusion prediction model. By using the occlusion information, they inferred the layer sequence of the image scene. Moreover they analyzed the occlusion cues and they stated that semantic cue is the best for the rural images followed by position, junction and compactness as well as convexity cues. Furthermore, position cue is the best for artificial images followed by semantic and junction cues. Their another finding is that semantic-position cue pair yielded the best accuracy among the other pairs.

Author(s)	Method	Dataset	Cues
[20], [21]	Neural-netwok	L-junctions, curvatures	Artificial dataset
[10]	Logistic function	size, lower region and convexity	Berkeley SDS
[31]	Conditional Random Fields	shapemes and T-junction	Berkeley SDS
[24]	Conditional Random Fields 20.7% fewer errors in BSDM	curvatures and T-junction	Berkeley SDS
[6]	Adaboost	compactness, semantic, position, junction and convexity	Their own

Table 2.1: Border Ownership Dataset for Computer Vision in the literature

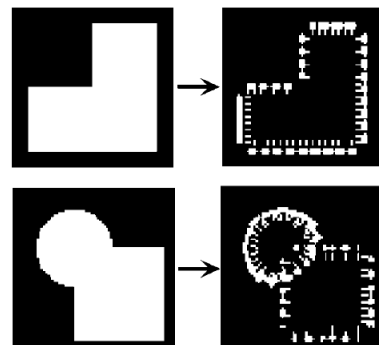
2.3 Border Ownership Datasets in the literature

For a BO dataset, collecting images are not the only ingredient. There should be corresponding regions and borders extracted with unique ids. Moreover, their ownership labels (assignments between borders and regions) have to be done.

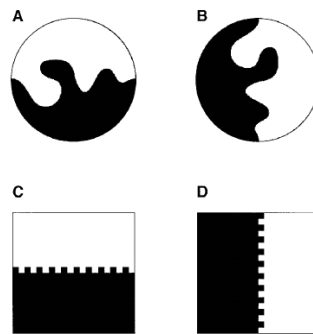
In the literature, there are a few dataset with BO label, but they are not comprehensive with respect to amount of image and number of participants. Current border ownership include a small number of images, some of them are artificial, based on only one category and one or two participants for each border.

In one of the very beginning study on this field, Nishimura et.al [26] used gray-scale or binary images with very low resolution. Kikuchi et.al. [20, 21] also worked with these types of images. Why they used these basic images could be that in their studies they focused on visual perception. They observed how V1 and V2 areas of visual cortex affected by different types of cues and visual stimuli. Therefore, they do not need so complex images. Some example images from their study can be seen on Figure 2.2.

The dataset that Fowlkes et.al [10] created consists of wide variety of indoor and outdoor scenes containing man-made and natural objects, including humans and other animals. These 200 images were chosen at random from the set of 1,000 hand segmented images in the Berkeley Segmentation DataSet (BSDM). Each border in the



(a) From Kikuchi et.al.'s study.
Study:[20]



(b) From Vecera et.al.'s study.
Study:[35]



(c) From Sakai et.al.'s study. Study:[32]



(d) From Nishimura et.al.'s study.
Study:[26]

Figure 2.2: Basic images used as Border Ownership Dateset in the literature



Figure 2.3: Sample wrongly-drawn borders in BSDM. [Best viewed in color]

dataset was labeled by two different human observers with 83.9% consistency on the labellings. In their study, Ren et.al. [31] used the same dataset. This dataset includes wide variety of images, but it has only 200 images. Moreover, some of the images include wrongly-drawn borders that could mislead the system using this dataset. You can see a sample of wrongly-drawn borders with blue color in this Figure 2.3.

In conclusion, It is required to create a new comprehensive dataset which includes various types of indoor (shopping mall, living room, train station etc.) and outdoor images (animals, historical places, landscapes etc.) taken from more participants.

CHAPTER 3

ONLINE LABELING TOOL & BORDER OWNERSHIP DATASET

For explained reasons in preceding chapters, we decided to create a new, comprehensive border ownership dataset incorporates both indoor and outdoor images. To realize this, dataset should include:

- enough indoor and outdoor data
- data accurate enough to be used as groundtruth for computational models that will be developed.

3.1 Online Labeling Tool

As explained earlier, Border Ownership is a problem of identifying which neighbouring region the borders belong to. As a result, our database should include pair of border and region information that border belongs to. An online Border Ownership labeling web page which is with user-friendly and well-documented interface was created (by Selin Akifoğlu and Mehmet Akif AKKUŞ) so as to obtain Border Ownership data simultaneously [3].

Home page of the online labeling tool can be seen in Figure 3.1. This home page explains the problem of Border Ownership and in which purpose this tutorial page is used. Then, it gives a sample image with corresponding sample border and regions. From the home page, users are able to register to the system or log in to continue to labeling process with his username or password.

For a participant to use the system he/she first register to the system, then read the explanations for border ownership, then tries and sees whether learned the concept with an interactive tutorial page, then finally labeling page that includes real indoor and outdoor border ownership data appears. These pages will be explained in the following sections respectively.

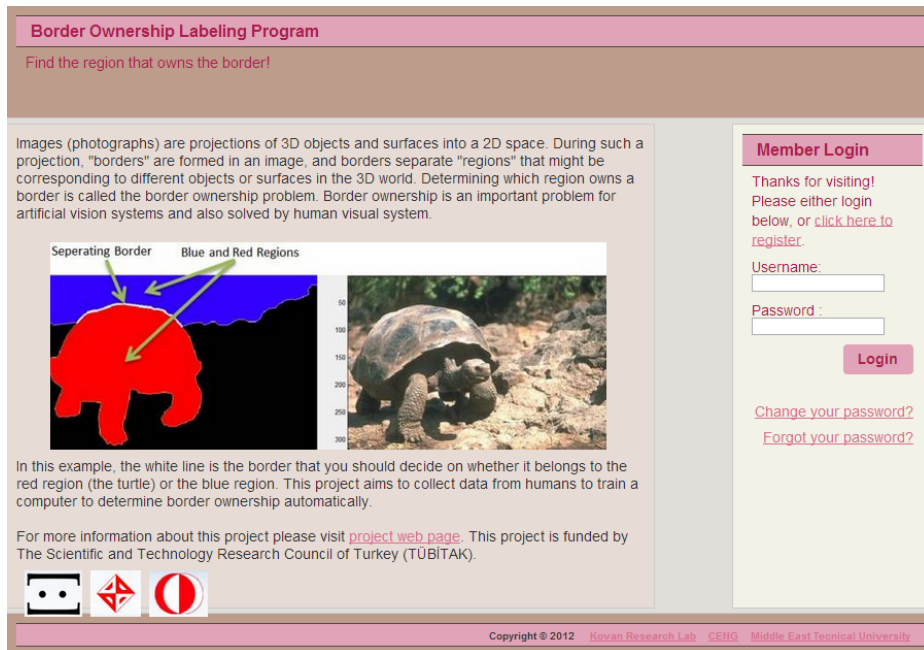


Figure 3.1: Home page of the online labeling tool

3.1.1 Registration to the System

First of all, users are asked to register to the system which is shown in Figure 3.2. They are requested to input their username, password, name, surname. In addition to this information, users are expected to input their age, gender and education level in case they can be used for statistical analysis. This information is used to observe if there is any correlation with consistency of dataset. For example, one of the result what we expect could be like that people who their age more than 20 have more consistent labeling result than the others. Consistency results with respect to age and gender are discussed in the section 4.2.

Figure 3.2: Registration page of the online labeling tool

3.1.2 Explanations for Border Ownership

After login to the system, participants are faced with a page (see Figure 3.3) which consists of several type of sample images. Each image has sample border (which is shown white dashed line) and neighbouring regions (which are shown as transparent blue and red regions). These explanations give an idea how a region can be selected with the given border. As soon as participants feel trained enough for the concept of Border Ownership, the following page is a tutorial page which tests them whether they understand well or not.

3.1.3 A Tutorial Before Labeling

Tutorial page with interactive user interface (see Figure 3.4) gives users a chance of trial before starting with real border ownership data. Participants are asked an artificial image includes 5 very easy borders. For the sake of simplicity, image has very basic shapes such rectangles, circles and pentagons. After each click, they are given a feedback which explains clicked true or not with a reason. Users must click to true region to pass to the next question. In Figure 3.4, you can see the related feedback (a) when user not clicked yet, (b) after wrong click and (c) after correct click.

Border Ownership Labeling Program

Here, you can see the basic explanations about border ownership that you should pay attention while labeling the pictures. The dashed white line is the border that you should decide on whether it belongs to the red region or the blue region.



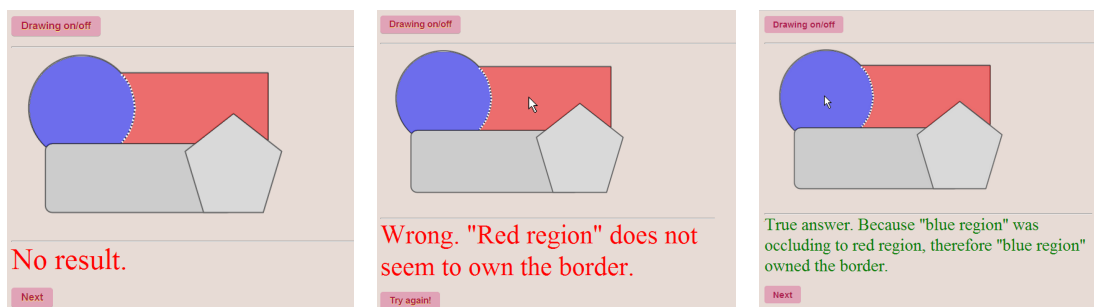
In this example, the dashed border between the red and the blue regions is formed due to the penguin since it is in front of the other objects. Hence, we say that the white border belongs to the red region (i.e. the penguin) because, if there weren't a penguin in the image, the marked border wouldn't exist.

In this example, the white dashed border between the red and the blue regions belongs to the red area (i.e. the goose), which is in front of the background (i.e. the blue region).

Here, the region (i.e. red or blue) to which the white border belongs is not obvious because the regions seem to be at the same depth since they are not in front of each other in any way. Therefore, in this and similar situations, you can choose "I'm not sure" option while labeling regions.

Next

Figure 3.3: Users are trained with real indoor and outdoor images



Drawing on/off

No result.

Next

Drawing on/off

Wrong. "Red region" does not seem to own the border.

Try again!

Drawing on/off

True answer. Because "blue region" was occluding to red region, therefore "blue region" owned the border.

Next

- (a) Participants are expected to click on one of the regions for the border in back given white
- (b) Participant clicked wrong. Feedback given
- (c) Participant clicked correctly. Feedback given

Figure 3.4: Tutorial page on online labeling tool. For a set of simple borders, users are informed whether they labeled right or wrong. [Best viewed in color]

3.1.4 Labeling Page

When the tutorial page is successfully completed, participants are ready for real indoor (including office, shop, living room etc.) and outdoor (including nature, animals) images. In the labeling page (Figure 3.5), users are asked to select one of the red or blue regions. If they think that they did the last operation accidentally, they can undo by clicking *undo* button. Other button *I am not sure* makes it possible to pass the current question if user can not decide exactly which region owns the border. Finally, user can disable the red and blue regions to see the image itself in more detail, and user can enable again by re-clicking to the *Drawing on/off* button.

Each user is asked 30 pictures by default and number of borders varies picture to picture. Progress on images and progress on borders of current image are shown progress bar lies on the right part of screen. There is a *Take a look at explanations* button can be used to see Border Ownership explanations again if user feels unsure what to do.

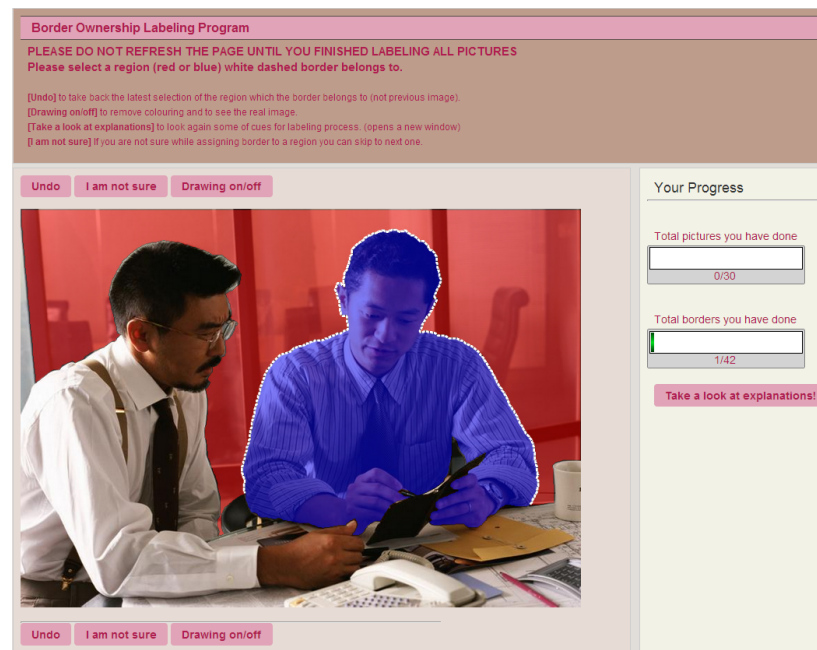


Figure 3.5: Labeling page with instructions at the top and a panel at the right shows the progress. *Undo* button gives a chance to undo the latest click operation, *I am not sure* button is so as to pass current border if user is not sure enough to label one of the regions, lastly, *Drawing On/Off* button enables/disables blue and region

3.2 Border Ownership Dataset

In this section how images and corresponding borders are gathered will be explained. Then, the followed section talks about participants.

3.2.1 Data to Be Labeled

Dataset incorporates 500 indoor and 500 outdoor images in jpeg format, 1000 images in total. Outdoor images are obtained from Berkeley Segmentation Data set [2]. 500 indoor images are obtained from various sources, 219 of which are from LHI [1] dataset, remaining part from diverse copy-free image sharing websites. (see Figure 3.7 for sample indoor images, see Figure 3.8 for outdoor images with sample border and regions)

Regions were available for images taken from LHI and Berkeley Segmentation Data set (they are needed for border to be asked), for other images regions are extracted by hand using a segment annotation tool (see Figure 3.6). It presents ability to select each object (see yellow points) in order and see the resulting annotated segments simultaneously. After obtaining segments for all the indoor and outdoor images, borders and neighbouring regions are easily generated by utilizing these images. In the meantime, borders whose length is smaller than 4% of the the image diagonal are eliminated because of difficulty of seeing and labeling.

To obtain more accurate BO data, each border are asked to at least three different participants. Answers from there different users are analysed under section 4.2.

3.2.2 Participants

Labeling process of Border Ownership was carried out by 151 participants (students and researchers from computer engineering) whose ages varied from 17 to 34. The participants (115 male, 36 female) are told about the purpose of the experiment and joined a tutorial session related to use of online labeling tool.

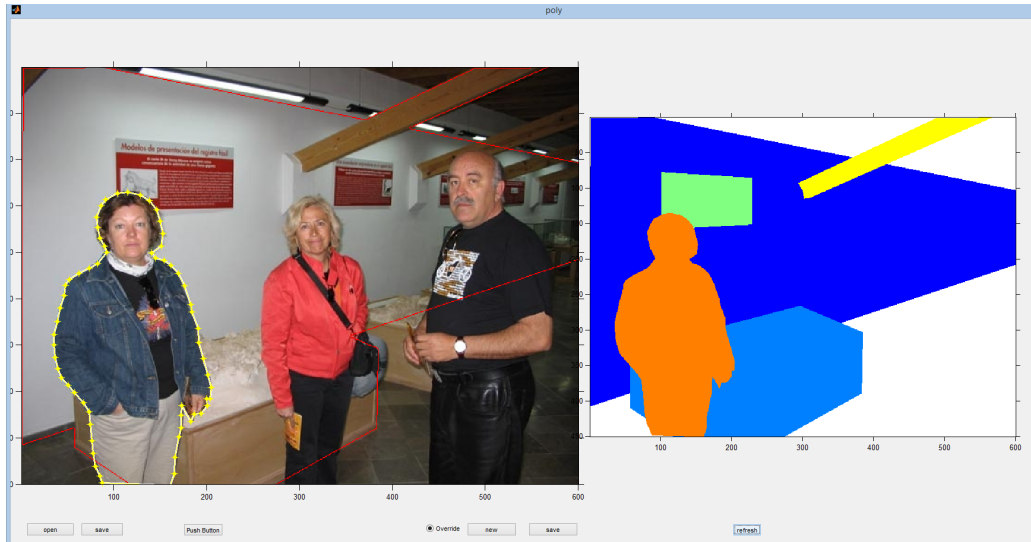


Figure 3.6: Segment annotation tool is used for creating segments of an image. Left image is the one that will be segmented. Region with red border shows segmented parts till then, Region with yellow points are currently segmented object. Right image shows current segments.

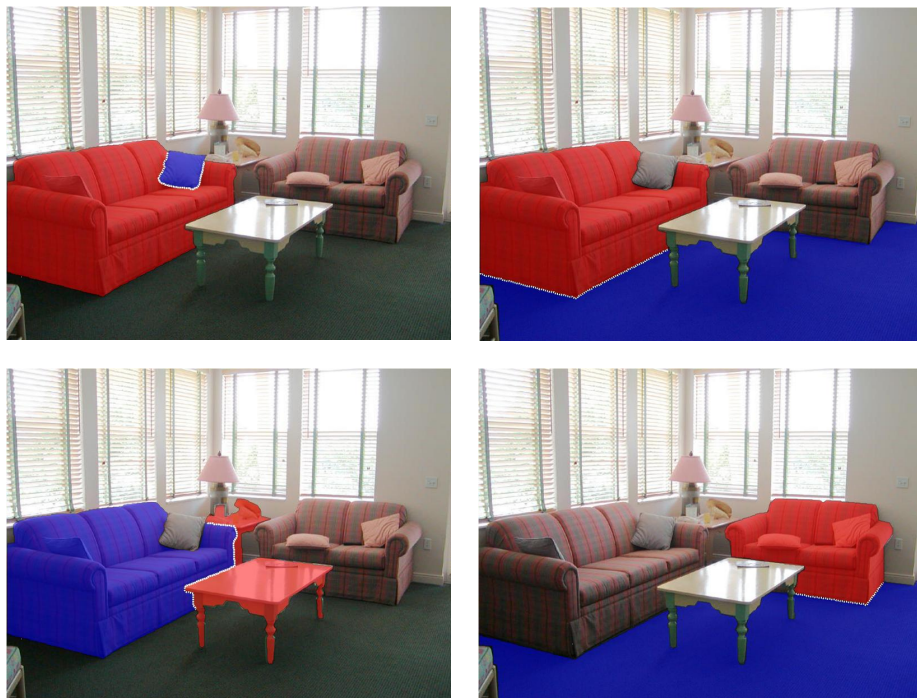


Figure 3.7: Various borders in a sample indoor image. [best viewed in color]

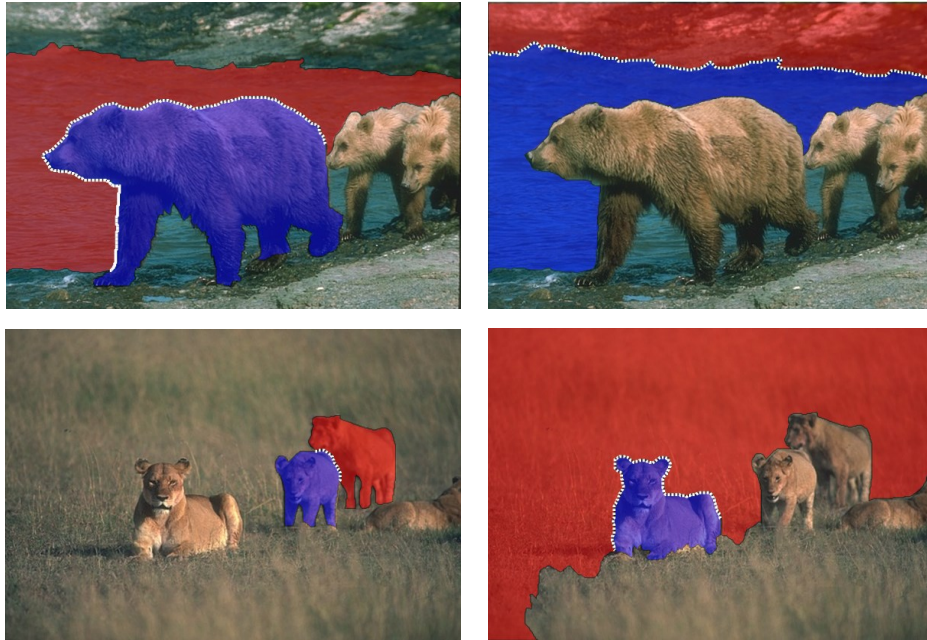


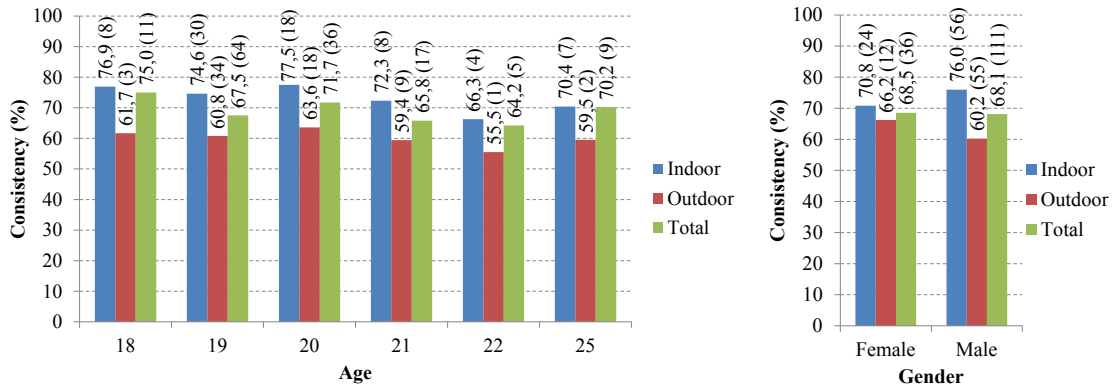
Figure 3.8: Various borders in a sample outdoor image. [best viewed in color]

3.3 Analysis of the Dataset

In this section, an analysis performed on generated Border Ownership dataset. In the following paragraphs, consistency with respect to different type of images (indoor and outdoor), age and gender will be examined.

In this dataset each border asked three different participants. For some borders, participants are agreed on the same region, for some they are not. This type of labeling is called as *consistent* if they are agreed upon. Labeled borders have the high ratios of consistency (69.4%) for indoor images and (62.2%) for outdoor images. Also, it should be stated that since participants had a chance to pass the border, without selecting any of the region, if they were not sure to select it. Thus, these labellings are not taken into consideration for consistency.

The dataset was labeled by 36 female and 117 male participants with different ages varying from 17 to 34. Consistencies are filtered according to age and gender. Results respect to age and gender are show in Figure 3.9(a) and Figure 3.9(b) respectively. There is a slight decrease on the consistency ratios as the age increases and there is



(a) Consistencies according to ages. Some of the ages are not shown due to low participations (b) Consistency in terms of gender participations

Figure 3.9: Consistencies according to age (a) and gender (b). Color legend of two types of images (indoor and outdoor) and for all images in total are listed on the right. Total number of participant is shown in paranthesis [Best viewed in color]

a minor difference for males and females with respect to indoor and outdoor images. Since number of participants in terms of age and gender are not equal nor nearly equal it is not so much possible to conclude the result as in the Figure 3.9. As a result, consistency of border ownership does not seem to be correlated with gender and age.

Other analysis is with respect to image type: indoor and outdoor (see Figure 3.10). It can be easily shown that indoor images are more consistent than outdoor images. The reason for this difference could be that indoor images have more structured, regular borders than outdoor images, and figure-ground segregation is more clear in general compared to outdoor images.

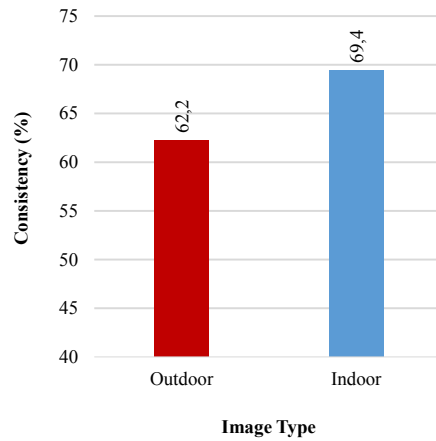


Figure 3.10: Consistency ratios of indoor and outdoor images.

3.4 Summary

To summarize, new dataset for Border Ownership includes various types of indoor (shopping mall, living room, train station etc.) and outdoor images (animals, historical places, landscapes etc.) with 1000 images. Dataset presents not just one labeling for borders but also it has at least three labeling for each borders. On the other hand, segment annotation tool and online ownership labeling tool give rise to expand dataset with new images.

It is believed that this dataset can be used as a benchmark for computational models as well as for analyzing the underlying mechanisms of different vision problems, especially that of Border Ownership.

CHAPTER 4

BORDER OWNERSHIP CUES AND THEIR ANALYSIS

In this chapter, visual cues used for Border Ownership determination will be explained. Cues below are mostly used for previous studies: *lower region, curvature, contrast, T-junction, L-junction*.

4.1 Used Border Ownership Cues

4.1.1 Lower Region

Particularly in outdoor images, object with the lower positions in the image is more likely to be close to the camera. In other words, they are more likely to be perceived as figures. This cue is used for image scene layering [10, 35, 6] as well as border ownership [24].

As shown in Figure 4.1, straw stack (R1) is the lower than the others and the closest object at them same time. Then, the house (R2) is the second lower region and goes on like this. On the other hand, according to formulation of this cue given below (see (4.1)), border between R1 and R2 belongs R1 since R1 is the lower. Thus, lower region in our model estimates correctly for this border.

For a border b with two neighboring regions r_a and r_b , the lower region cue predicts the owning border $\hat{r}(b)$ of b as follows:

$$\hat{r}_l(b) = \arg \min_{r \in \{r_a, r_b\}} \arg \min_{(x,y) \in r} y, \quad (4.1)$$



Figure 4.1: Sample image shows concept of lower region. Image at right corresponding segmented image of the left image. As explained, R1 is at bottom and the closest object, then house, tree and sky accordingly.

where x and y are the points on 2D coordinate system.

4.1.2 Curvature

Curvature is the another cue which is mostly used for border ownership estimation [10, 6]. Curvature in 2D image tends to correspond to 3D object surface and convex side of the border is likely to own the border. Also, there is a correlation between degree of a convexity and ownership likelihood. In other words, the more border is curved, the more likely the convex region owns the border. (see Figure 4.2)

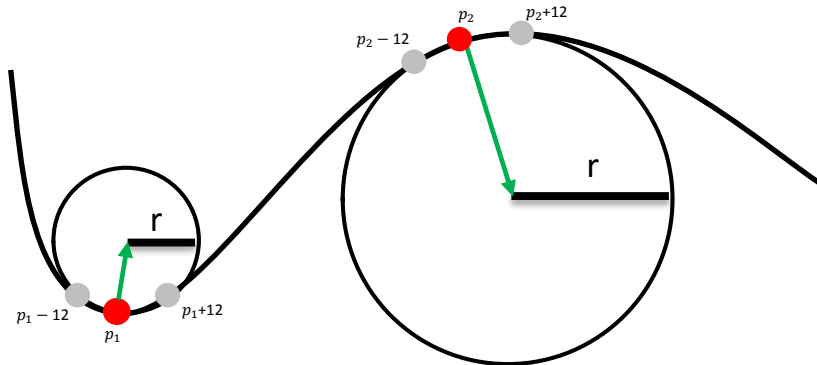


Figure 4.2: Sample calculation of magnitude of a curvature on a certain point p_1 and p_2 . Note that radius and curvature direction changes along the curvature.

In our curvature model, curvature at a point has *magnitude* and *direction*. *Magnitude*

of a curvature is inverse of the radius of a circle fitted at point. To fit a circle, we select two more points before and after at that point (see Figure 4.2) since there can be drawn unique circle passing from these three consecutive points $(p_{i-12}, p_i, p_{i+12})$. 12 is selected empirically.

For calculation of magnitude $m(p_i)$ at a certain point p_i on the border b , first circle c is fitted with the radius r formulated as:

$$m(p_i) = 1/r. \quad (4.2)$$

The direction $d(p_i)$ at a point p_i is the vector connecting the point p_i and the center of the circle $ce(c)$ formulated as:

$$d(p_i) = ce(c) - p_i. \quad (4.3)$$

To determine the region owning the border, all points are calculated along the boundary as follows:

$$\hat{r}_c(b) = \arg \max_{r \in \{r_1, r_2\}} \sum_{p \in b} (d(p) \rightarrow r) * m(p), \quad (4.4)$$

where $d(p) \rightarrow r$ is 1 if the direction of curvature coincides with the region r , and 0 otherwise. Then, the point contributes as magnitude along the direction.

Sample curvatures on an image can be show on Figure 4.3.



Figure 4.3: Sample curvatures on an image. Dashed red lines are the curvatures and arrows show the direction of the ownership.

4.1.3 T-junction

Junctions are another type of cues which give highly strong information about 3D structure where they occur [16]. T-junction is found out best among junctions since it gives accurate results on depth ordering and figure-ground segregation and occlusion detection [14, 31].

In this study, determination of T-junction points are based on corner points. First of all, corner points are detected using curvature-based corner detector, proposed by He & Yung [12]. Then we extract the line segments $l_{s_1} \dots l_{s_L}$ meeting that corner point and the angles between them $\theta_1 \dots \theta_L$. Line segments (the red lines) and angles can be seen in Figure 4.4.

Using the information of line segment and angles, first type of junction is determined then, the ownership direction of the border. For a junction to be T-junction, it has to have three line segments. Next, for ownership direction, θ_s with the biggest angle is selected and lines, namely borders, belongs to the region including θ_s . Figure 4.5 shows the assigning borders to a region using T-junction. Note that it is special to T-junction that each junction assigns two borders, while other cues assign only one border at a time.

These processes are implemented for all the junctions in the image and sample T-junctions and their ownership are available on Figure 4.6.

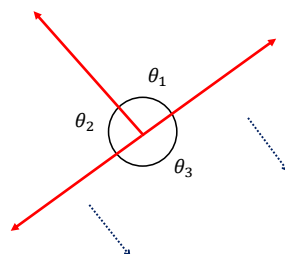


Figure 4.4: Representation of T junction. Note that T-junction has three lines segments.

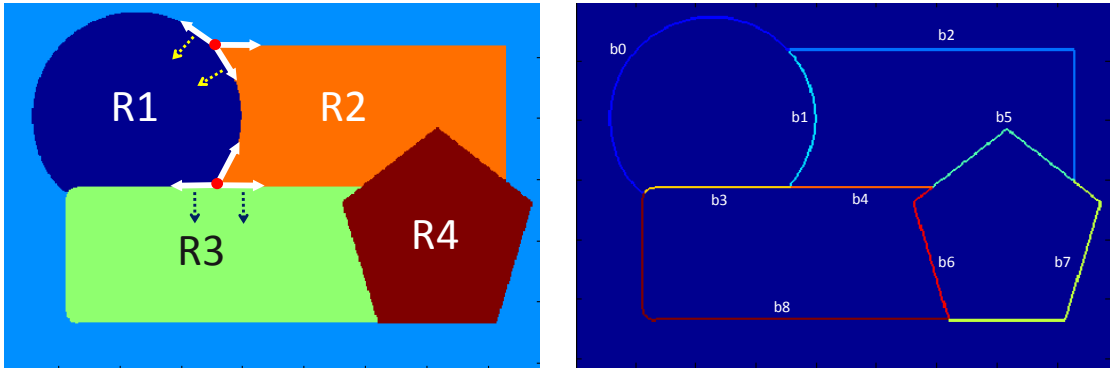


Figure 4.5: Sample T-junctions and ownership directions on an artificial image. Left image: regions are labeled and two T-junctions are shown. Right image: labeled borders with their ids. According to T-junction cue border b_0 and b_1 belong to region R_1 , border b_3 and b_4 belong to region R_3 , border b_5 and b_6 belong to region R_4 etc.

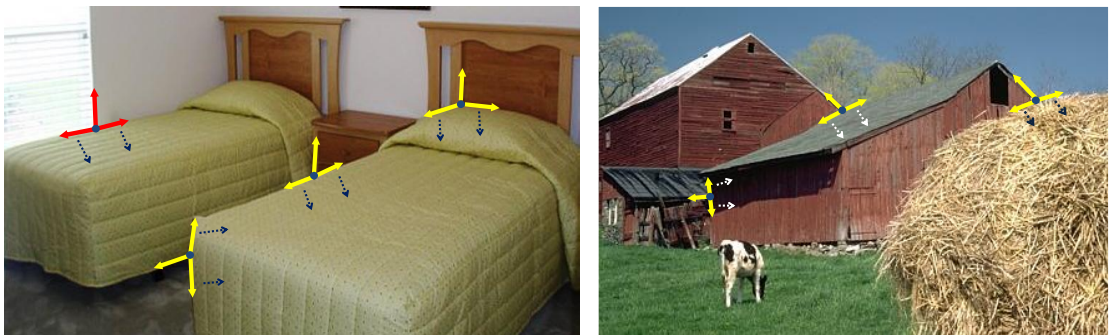


Figure 4.6: Sample T-junctions and ownership directions on sample indoor and outdoor image. Dashed rows show the ownership direction [Best viewed in color].

4.1.4 L-junction

Similarly to T-junction, L-junction is another corner based cue for Border ownership. But in this case, there are only two line segments and two angles (see Figure 4.7).

According to this cue, inner part of the junction is more likely to own the borders. In other words, region at the same side with small angle owns the borders. For instance, as can be seen on Figure 4.5 border b_2 at top right belongs to region R_2 and at right bottom b_7 belongs to region R_3 using L-junction. Moreover, Sample L-junctions on a sample indoor and outdoor images can be seen on Figure 4.8.

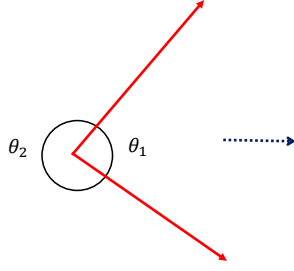


Figure 4.7: Representation of L junction. Note that L-junction has two line segments.



Figure 4.8: Sample L-junctions and ownership directions on sample indoor and outdoor image. Dashed rows show the ownership direction [Best viewed in color].

4.1.5 Contrast

Last cue utilized in this study is contrast. According to this cue, brighter objects are more close to the camera, so it is more likely to own the border.

The region \hat{r}_{co} owning a border b is as follows:

$$\hat{r}_{co}(b) = \arg \max_{r \in \{r_1, r_2\}} \sum_{p \in r} \frac{1}{N} I(p), \quad (4.5)$$

where $I(p)$ is the intensity of pixel P and N is the total number of pixel on the region.

In this study, these five cue have been utilized. In the following section, accuracy of each cue for Border Ownership estimation will be explained.

4.2 Analysis of the Cues

In this section, we analyse each cue in terms of their predictive capability. Then, we combine different number and types of cues so as to create more accurate prediction results.

First of all, number of available borders changes significantly between different types of cues. Comparative results can be seen on Table 4.1. This table illustrates that number of occurrences and availability is the same for the cue Lower region and contrast. As we did not ask the some curvatures due to their small length, number of available border decreases considerably. Since t-junction can vote for more than one border its availability is bigger than its occurrences (see Section 4.1.3). As you can see, L-junction votes only for one border and its occurrences and availability is the same. This table also shows that L-junction and T-junction are more seldom than the others.

Table 4.2 shows that distribution of different labeling types within different image type. In general, 65.4% of the dataset has consistent border labeling (3/3), 30.8% of the dataset includes different labels (2/3), while 3.8% does not include any label (“I am not sure”). Such an analysis allows understanding of where participants disagree more about BO and investigation of underlying reasons for that. Figure 4.9 shows sample images which difficult to select one of the regions. It is very clear that labeling for indoor dataset is more consistent and participants are more unsure about labellings.

Table 4.3 shows prediction capability of each cue against two kind of dataset, the first of which is “single dataset” that includes one labeling for each border, the second of which “is triple dataset” that includes three labeling for each border. Single dataset was obtained by researchers in this study, triple dataset was obtained by over 100 participant from students studying computer engineering. Triple dataset is divided into two categories according to labeling type that explained on the Table 4.3. One of which includes only consistent borders and other includes both consistent borders and non-consistent borders. For non-consistent borders the region which was labeled by two participants selected. Note that in the existence of three participants, there must

Table 4.1: Occurrences and availability of the cues on all images. Cues are ordered according to number of available borders for a certain cue type.

Type of Cue	# of occurrences	# of available border
Lower region	32837	32837
Contrast	32837	32837
Curvature	32837	26434
T-junction	12961	14655
L-junction	12791	12791

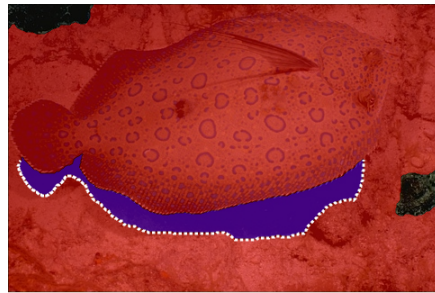
Table 4.2: The analysis of different labeling with image type. 2/3 stands for participants are not labeled the same region, 3/3 stands for all three participants labeled the same region

Labeling Type	Image Type		
	Indoor	Outdoor	Total
(3/3) consistency	69.4%	62.2%	65.4%
(2/3) consistency	22.3%	37.7%	30.8%
Participant labeled “I am not sure”	8.3%	0.1%	3.8%

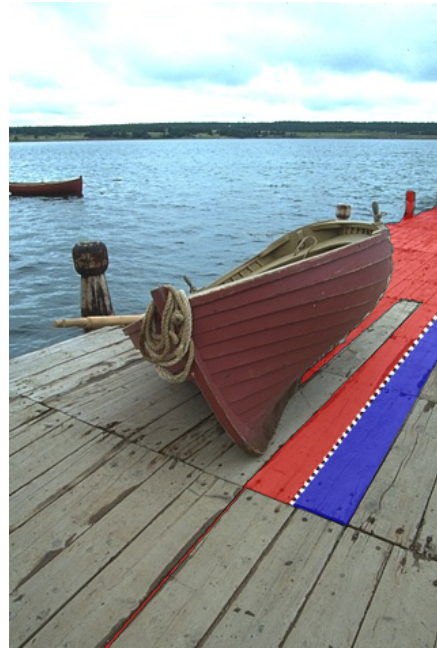
be at least two participants selecting one of the regions. Although “3/3 consistency” is the best accuracy among them, “2/3 + 3/3 consistency” was used in the later part of the analysis for the sake of more available data.

4.2.1 Accuracies of Visual Cues

From Table 4.3 following results can be deduced: (i) Accuracy with respect to cues changes significantly. This is presumed since each cue works well under different circumstances. For instance, while T-junction and curvature are the better under indoor, lower region is better on outdoor images. (ii) T-junction and curvature are strong cues, which is in agreement with the literature [37]. L-junction is the third best and Lower-region is the fourth best cue. Contrast is the worst cue with lower accuracy than the random choice. (iii) Generally, indoor images have more accuracy than the indoor images. Reason for this possibly is that indoor images are more likely to include regular structures than outdoor images. Hence, it might be reason why junction based cues are generally better indoor images.



(a) Ambiguous border 1



(b) Ambiguous border 2

Figure 4.9: For some borders, it may not be so obvious to determine ownership of the border. This might be due to wrong segmentation by humans (a) or that the borders might seem shared by two objects (b). [Best viewed in color]

4.2.2 Accuracies of Combined of Visual Cues

Another experiment is combining different number of cues so that better Border Ownership prediction can be obtained. Relative prediction accuracy can be shown on Figure 4.10, which includes indoor (4.10(a)) and outdoor results (4.10(b)). While combining multiple cues the majority rule was used. That's to say, the region was selected on which more cues are agreed. If the majority does not occur, which the half of the cues predict different regions, the region with more accuracy was selected. Another issue should be explained here is that availability of cues. Some cues are not available on a certain border. In other words, it is not possible for every cue to occur on a border. For instance, while cues such as lower region, curvature are available for every border, cues such as T-junction and L-junction are available on the border only where junction occurs.

Closer look to 4.10 explains the following results. As can be seen from the figure, for

Table 4.3: Visual cues and their predictions compared against different consistencies of labels.

	“single dataset”		“triple dataset”			
	Indoor(%)	Outdoor(%)	“3/3 + 2/3 consistency”		“3/3 consistency”	
			Indoor(%)	Outdoor(%)	Indoor(%)	Outdoor(%)
Curvature	65.6	61.0	65.6	59.2	72.0	64.9
T-junction	77.0	66.0	79.0	63.3	82.0	66.9
L-junction	50.0	55.0	58.0	50.6	59.6	54.1
Lower Region	46.0	50.0	52.0	50.9	54.7	52.0
Contrast	44.0	45.8	44.7	47.4	44.1	47.4

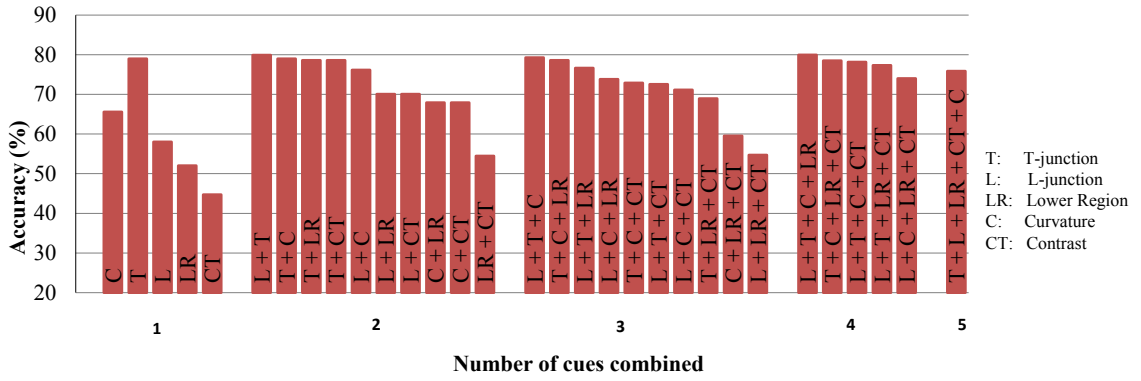
indoor images combination of for cues namely T-junction, L-junction, Curvature and Lower-region are the best with 80% accuracy among 26 combinations. L-junction and T-junction combination (79.9% accuracy) and L-junction, T-junction and curvature combination (79.3% accuracy) have nearly the same accuracy as the best cue combination. Moreover, T-junction is best among the other cues and have the highest contribution in combinations.

Regarding to outdoor images, combination of T-junction, curvature and lower region is the best with the accuracy of 68.6% among 26 combination. Then combination of L-junction and curvature (68.1%) follows. Interestingly, although curvature and T-junction are individually best cues, combination of them is ranked fifth in two cue combination. Reason for this might be that T-junction is so rare in outdoor images due to lack of regular structures.

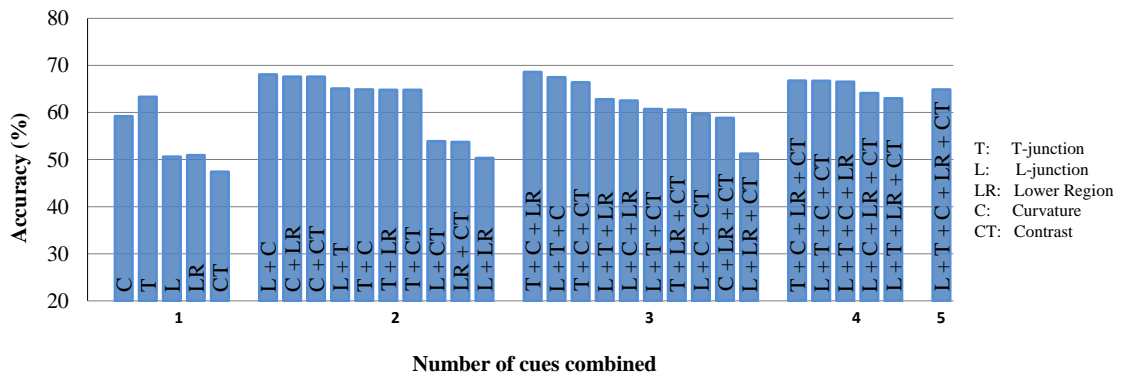
Among a total of 26 combinations from our 5 cues, while combination of 4 cues are best for indoor images and combination of 3 cues are best for outdoor images. What is interesting is that combination of more cues is not meant to perform better accuracies for both indoor and outdoor images.

4.2.3 Analysis of the Conflicting Cues

Previous section covered the analysis how accuracy changes when they are combined. Now, this section covers conflicting cues. Figure 4.11(a) and 4.12(a) show the matrix of conflicting cues for indoor and outdoor images of BO prediction. Rows and



(a) Combined cues on indoor images



(b) Combined cues on outdoor images

Figure 4.10: Combination of different cues on indoor and outdoor images. The x-axis shows the number of combined cues and y-axis shows the accuracy by percentage

columns of 5 by 5 matrix stands for each cue. Each cell stores number of conflicting values about two different cues. The numbers in the middle depicts number of cases that cues conflict divided by total number of cases that cues co/occur. In addition to number of cases, percentage of cases are shown on the upper right and down left corners of each cell. These numbers stand for correct prediction of each cue in case of conflict.

For indoor images, ratio of conflict is quite low between curvature and L-junction and between T-junction and L-junction, while number of conflict between lower-region and L-junction and between lower-region and curvature is quite high (see Figure 4.11(a)). On the other hand, for outdoor images, curvature and L-junction are quite in accordance with each other while lower region and curvature conflicts most (see Fig-

ure 4.12(a)). To sum up, L-junction, T-junction and curvature are in agreement with each other for BO prediction while lower region and contrast are generally conflicting for indoor images. However, in outdoor images, T-junction, L-junction and curvature have more conflict compared to indoor images. These result is in agreement with what we observe that, indoor images have more corners creating T and L junctions.

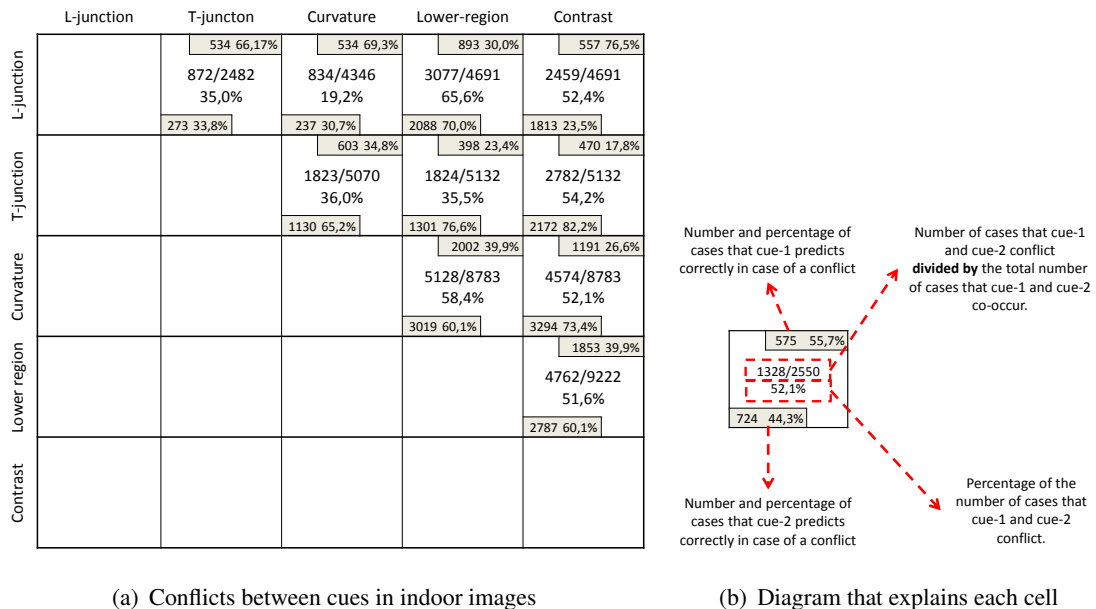
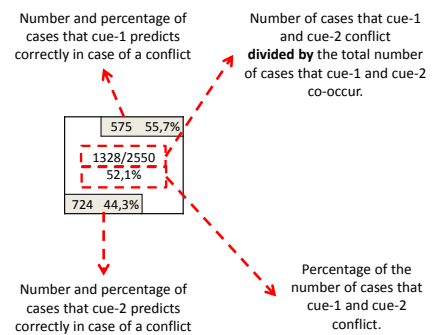


Figure 4.11: Conflicting cues in indoors. (a) A matrix depicting which cues conflict, and the *winning cue* in case of a conflict. (b) Explanation of each cell in the matrix.

	L-junction	T-junction	Curvature	Lower-region	Contrast
L-junction		319 60,3% 562/1148 49,0% 210 39,7%	567 73,5% 789/2541 31,1% 204 26,5%	749 47,1% 1615/2550 63,3% 840 52,9%	575 44,3% 1328/2550 52,1% 724 55,7%
T-junction			420 55,6% 790/1837 43,0% 335 44,4%	243 39,8% 688/1843 37,3% 368 60,2%	305 33,4% 969/1843 52,6% 607 66,6%
Curvature				951 38,7% 2466/4023 61,3% 1508 61,3%	628 30,6% 2060/4023 51,2% 1425 69,4%
Lower region					912 44,2% 2070/4040 51,2% 1153 55,8%
Contrast					

(a) Conflicts between cues in outdoor images



(b) Diagram that explains each cell

Figure 4.12: Conflicting cues in outdoors. (a) A matrix depicting which cues conflict, and the *winning cue* in case of a conflict. (b) Explanation of each cell in the matrix.

CHAPTER 5

MORE ACCURATE DEPTH PREDICTION USING BORDER OWNERSHIP

[9] 3D interpretation from their 2D projections, namely *depth prediction*, is one of the field of the computer vision that has been studied for a long time. To realize this, there are basically two types of cues: monocular cues that 3D information can be obtained using only one view of the scene and multi-view cues that at least two views of the scene are available for 3D information.

This chapter addresses that for depth prediction on homogeneous areas, Border Ownership (BO) information increases the accuracy when it is used with sparse stereo vision algorithm.

5.1 Depth extracting cues

Texture gradient, perspective distortion, occlusion, shading, atmospheric effect are examples of monocular cues. Texture gradient and perspective distortion are similar cues implying that closer objects to the camera are bigger and far objects are smaller. Occlusion is the another frequently used cue making statement about image scene layering, where occluding object is in front of the occluded object. Shading of a surface gives idea about how the surface is formed such as flat or curved surface etc. Last but not least, another cue is atmospheric effect, where closer objects seem sharper and less blurred, whereas far objects seem more blurred and less sharp.

It is important to realize that monocular cues are not capable to provide exact depth in-

formation, they can only provide relative depth information or scene layering, namely depth ordering, between different surfaces. Creating a computational model using monocular cues is quite limited due to difficulty of the formulation of the different cues. In their study, Clerk et. al. have shown that recovering 3D surface information from texture gradient needs that some certain assumptions has to be made [7].

Different from monocular cues, multi-view depth cues such as stereo can provide absolute depth information rather than relative depth information without prior information of the scene. These cues are used to determine absolute depth information by triangulation of two or more views of the scene. Stereo vision requires correspondance of globally-distinguishable image points between two images. Finding correspondence is not always trivial especially on homogeneous or weakly-textured image areas due to lack of distinguishable structures. Even in the existence of textured image areas, finding correspondence can be difficult. Using camera geometry search area for correspondences can be limited to line-like area, namely epipolar line, which makes easier the matching process.

There exists two main computational methods for stereo vision: sparse and dense approaches. Whereas sparse methods utilize image features for correspondence, and consequently, produce sparse depth information, dense methods deal with correspondence problem at lower level and compute stereo information for every pixel. Moreover, sparse methods are cost effective because of the sparsity and they can work with images which are big and have big disparities. On the other hand, dense methods need textured image areas and they are computationally expensive.

5.2 Depth prediction on homogeneous areas using sparse stereo methods

Existing studies have shown that 3D information at the edges is able to recover the missing depth information at weakly textured, homogeneous image areas [18]. To realize this, sparse (feature-based) stereo algorithm is utilized to require 3D information at the edges. In their study, model of Kalkan et.al. starts with creating local features corresponding to edge-like structures and monos (structures on homogeneous image area) using stereo image pair. Then, depth information at the edges is obtained using

a sparse stereo method [28]. Therefore, it is available to predict depth at the homogeneous image patch using the depth at edges with a voting model. This voting model only use the 3D line orientation at the 3D local edge features.

The missing part in the mentioned study is how to cope with the *occlusions*. Occluding edges should vote only occluding image areas during prediction. Otherwise, occluded surface tends to be closer to the occluded surface rather than being at the its exact position. This occlusion problem can be solved using *Border Ownership (BO)* information. BO information is able to give ownership of an image area in terms of borders, or edges.

This chapter details that BO information can be used to enhance the accuracy of a depth prediction at the homogeneous image areas.

5.3 Formulation of the depth prediction model

In their study in 2007 Kalkan et. al. used different representation for homogeneous and edge-like image structures, namely primitives [18, 19]. Primitives can be edge-like and homogeneous or either 2D or 3D. The corresponding 3D primitives of 2D edge-like primitives is extracted using stereo vision. For homogeneous primitives, 3D primitive is predicted from the 3D edge-like primitives. Figure 5.1. shows extracted primitives for a sample image. They developed a novel voting-based method for predicting depth at monos. They used a voting model since it is appropriate for producing result from data which consist of outliers. In this model, there is a set of voters stating their opinion about a certain mono. To make decision about the mono, the voting model combines these votes in a reasonable way.

For a depth prediction problem, the mono π^m to be voted for its depth and 3D orientation needs voters which are bounding edge primitives π_k^e (for $k = 1, \dots, n$). A pair P_j of two edge primitives π_i^e and π_k^e are required from the set of bounding edges.

The edge pair should assure some certain conditions:

- Edge primitives π_i^e and π_k^e should share the same color with the mono.

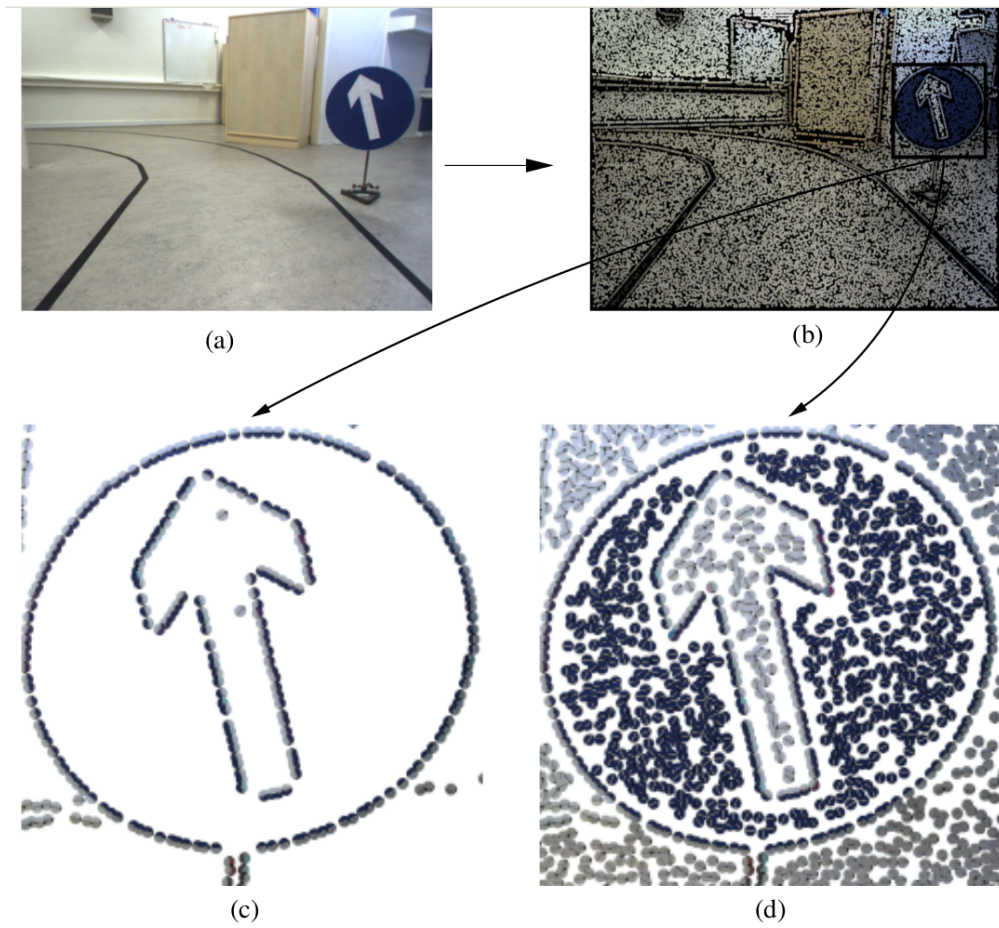


Figure 5.1: Extracted primitives (b) for the example image in (a). Magnified edge primitives and edge primitives together with monos are shown in (c) and (d) respectively. Source: [17]

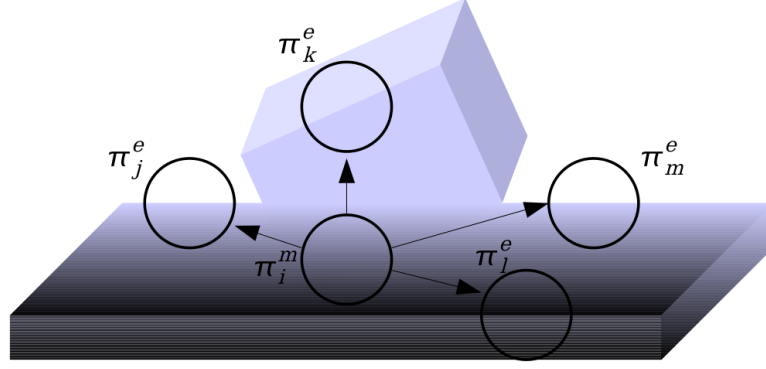


Figure 5.2: A set of primitives for illustrating why the relations co-planarity, co-colority and linear dependence are required as restrictions for forming pairs from edges. Source: [17]

- The 3D primitives Π_i^e and Π_k^e of the edge primitives π_i^e and π_k^e should be on the same plane.
- Edge primitives π_i^e and π_k^e should be linearly independent so that they define a plane.

Figure 5.2. depicts the listed restrictions on a sample mono and set of bounding edge primitives. The edge primitives π_j^e and π_l^e are co-color with the mono (1). They are on the same plane (2) and they are not on the same line (3). Therefore edge primitives π_j^e and π_l^e are the possible candidates for voting the mono π_i^m . On the other hand, edge primitives π_k^e and π_j^e are neither co-color nor do they define a plane. Hence, they are not the possible candidate for voting.

5.3.1 Bounding edge determination of a mono

Bounding edge determination is crucial for a mono to estimate its depth and 3D orientation. For this, search is made in a set of directions d_t , $t = 1, \dots, N_d$. Using two distance thresholds R_{min} and R_{max} , only edge primitives between these distances are searched. Found edge primitives are added to a list of bounding edges. In addition, there can be some missing edge primitives due to discrete search operation. These edge primitives can be added by grouping.

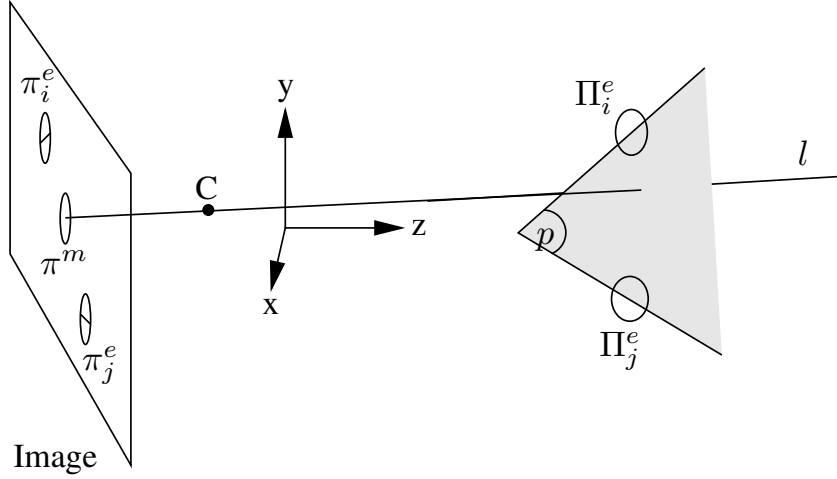


Figure 5.3: Illustration of how the vote of a pair of edge primitives is computed. The 3D primitives Π_i^e and Π_j^e corresponding to the 2D primitives π_i^e and π_j^e define the plane p . The intersection of p with the ray l that goes through the 2D mono π^m and the camera center C then determines the position of the estimated 3D mono Π^m . The 3D orientation of Π^m is set to be the orientation of the plane p . Source: [17]

5.3.2 Voting model for a mono

A pair P_i of two edge primitives π_j^e and π_k^e and corresponding 3D edge primitives Π_j^e and Π_k^e which assure the mentioned conditions can define a plane p with 3D normal n and position P .

As it can be seen from the Figure 5.3, the intersection of the plane p with the ray l going through 2D mono π^m and camera center C determines position of the 3D mono Π^m . The orientation of the 3D mono Π^m is the same as the orientation of the plane.

For a certain 2D mono π^m , there exists good number of votes from bounding edge pairs to determine 3D representation Π^m of the mono. Taking weighted average of the votes is the one alternative way. but it is not a good idea due to the outliers. Clustering the votes and taking the best cluster among them is the better alternative. Vote clusters can be seen in 3D Euclidean in figure 5.4. The cluster that includes the most votes can be labeled as the cluster that has the highest reliability. Equation 5.1. defines the best cluster as the most crowded one.

$$\Pi^m = \arg \max_{c_k} \#c_k, \quad (5.1)$$

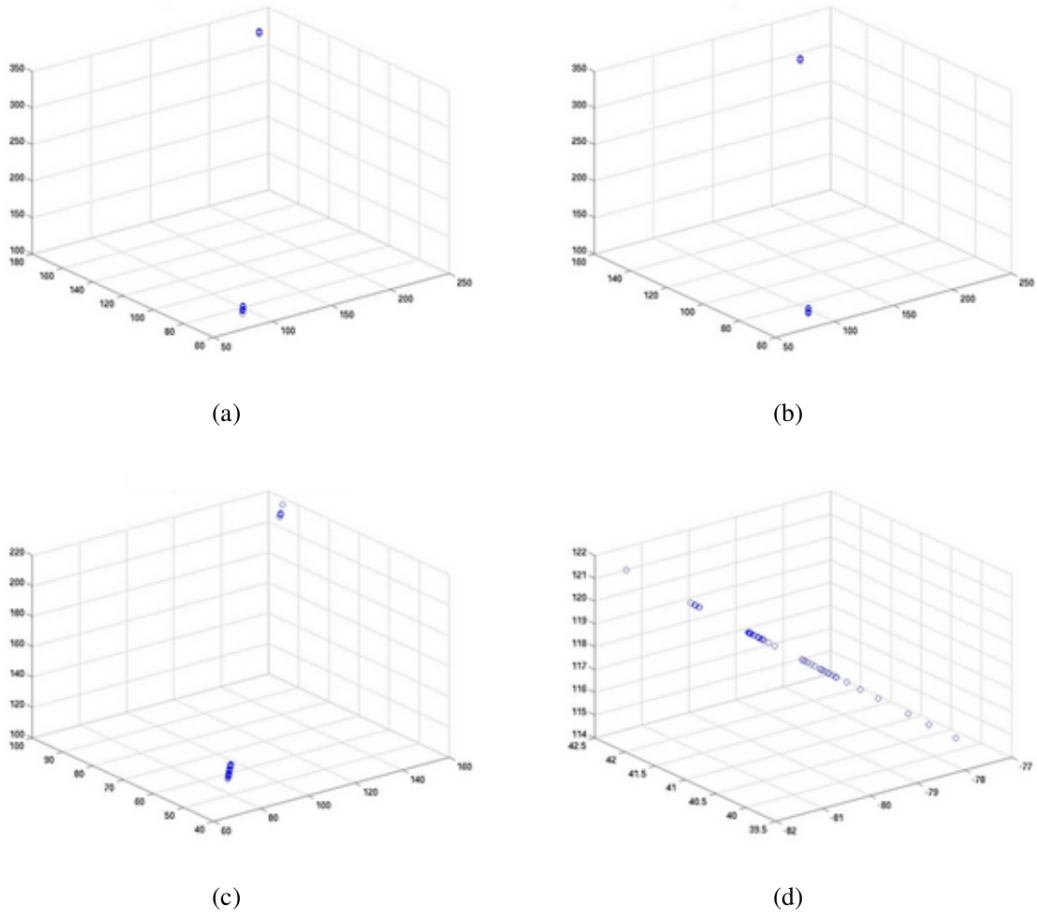


Figure 5.4: The distribution of the votes for some monos shown in the 3D Euclidean space. Sub-figures (a)-(c) are clear examples where there are two distinct clusters. However, such clear clusters do not always occur (d). Source: [17]

where c_k denotes clusters for $k = 1, \dots, N_k$ and $\#$ is the cardinality of a cluster. Clustering is practical for eliminating the outlying votes. As shown in the Figure 5.4. clustering is not always trivial for the nature of the distribution.

5.4 Adding Border Ownership information to the voting model

As stated in previous sections, edge primitives on the occlusion edges can mislead the voting model while determining the position of a mono on the occluded image areas.

First of all, image segments and borders (see Figure 5.6) are generated using the RGB image (see Figure 5.5 for stereo image pair). This generation process can be either

with by hand using an image annotation tool or using one of the image segmentation algorithm. For the sake of simplicity and more accurate image segments, image annotation tool is used (see section 3.2.1. for more information). Then, using basic BO computational model BO information can be obtained for each border.

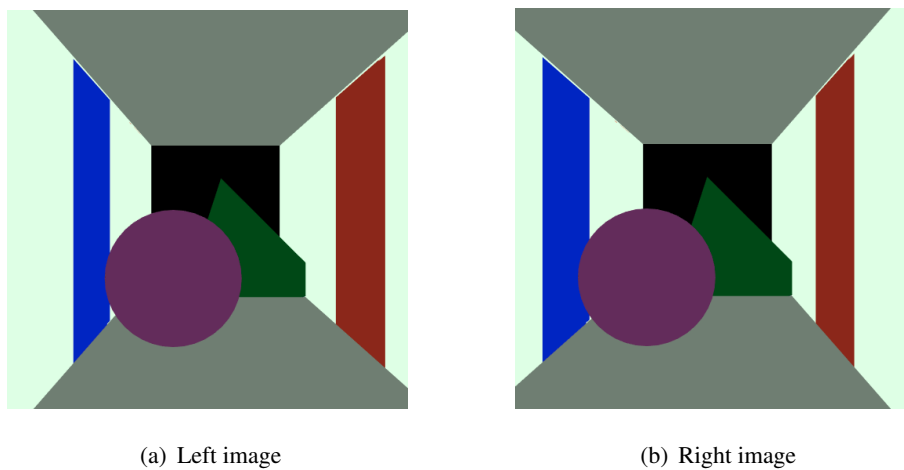


Figure 5.5: Stereo image pair from an artificially created scene. [Best viewed in color]

BO information prevent voting model from choosing edge primitives which normally should not vote for a certain mono. You can see the difference in Figure 5.7. to see difference bounding edge primitives to be used for a certain mono (labeled as black point on the left image) before using BO information (left image) and after BO information (right image). As you realized that for monos which are on the occluded image area, edge primitives lying on the occluding edge can not vote and these primitives are eliminated from the list of voters. Moreover, irrelevant image primitives due to missing edges points also can be eliminated (see the last row on the Figure 5.7).

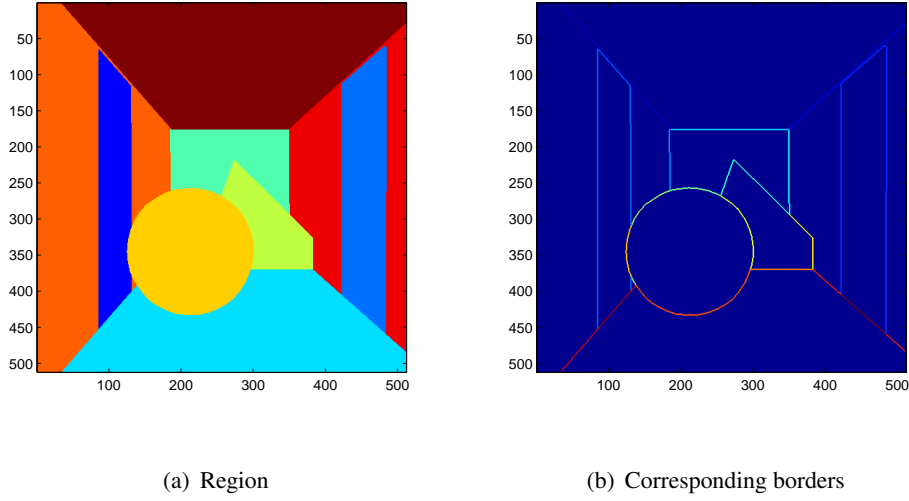


Figure 5.6: Regions and corresponding contours (border). Each border and region have unique id. [Best viewed in color]

5.5 Results

This section evaluates the performance of depth prediction with BO information and compares its performance to standard depth prediction method. The object is to show depth prediction using BO information helps resulting with less erroneous depth estimation. To show this, comparisons are performed on an artificial scene where noise of the scene can be controlled to see behavior of the two approaches. What we expect is that depth prediction using BO is also helpful in different types of noises.

Evaluation is performed by comparing with groundtruth. For this, two metrics which have been frequently used in the literature are utilized: Root-Mean-Squares (RMS) and Bad-Matching-Percentage (BMP). These measures are taken from the study of Schartein et. al. [33]. RMS (eq. 5.2) and BMP (eq. 5.3) are defined as follows respectively:

$$\text{RMS}(S) = \frac{1}{\#S} \sum_{p \in S} (|d_c(p) - d_G(p)|^2)^{1/2}, \quad (5.2)$$

$$\text{BMP}(S) = \frac{1}{\#S} \sum_{p \in S} (|d_c(p) - d_G(p)| > 1). \quad (5.3)$$

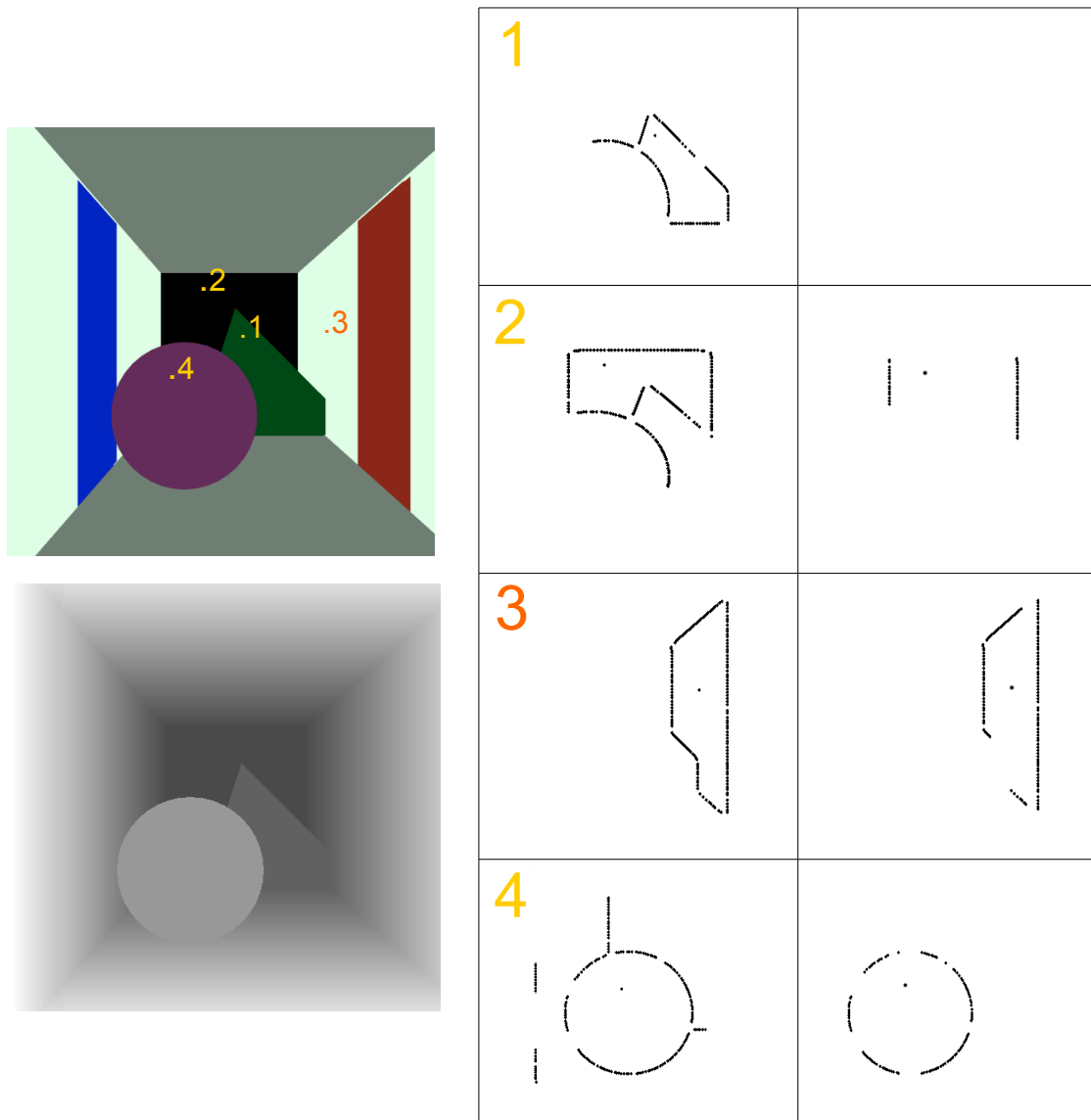


Figure 5.7: Left part is the scene and corresponding depth image. In rectangle, left images shows bounding edges for a certain mono, right images denote voting edges after BO information is applied. Note that some edges primitives are disappeared due to their wrong prediction vote.

where S denotes the set of the points with disparity information, $d_c(p)$ and $d_G(p)$ are respectively the computed and the ground truth disparity information at point p .

5.5.1 Results on artificial data

5.5.1.1 Evaluation with texture noise

On artificial data first texture noise (white noise) is added with different amount of texture frequency ($n \in \{0, 0.01, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2\}$). A subset of input images can be seen in Figure 5.8.

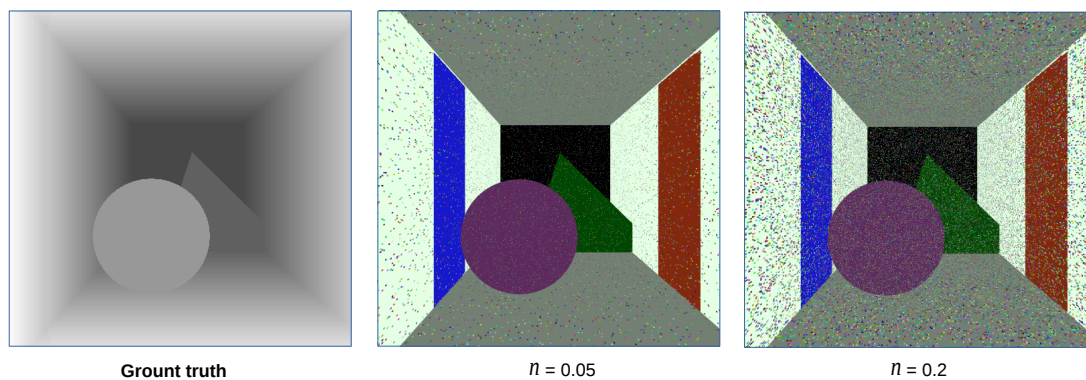


Figure 5.8: Artificial image with different amount of texture noise. [Best viewed in color]

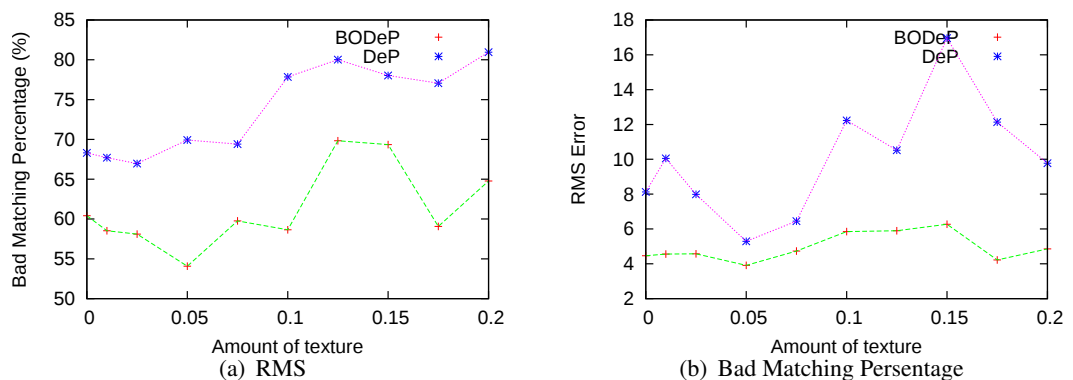


Figure 5.9: Root Mean Squares and Bad Mad Matching Percentage metrics are used with different amount of texture noises.

5.5.1.2 Evaluation with texture and pepper noise

Different from the noisy set of images in previous section, in this section pepper noise added with several frequencies ($u_n \in \{0, 0.01, 0.05, 0.1, 0.15, 0.2\}$). A subset of input images can be seen in Figure 5.10.

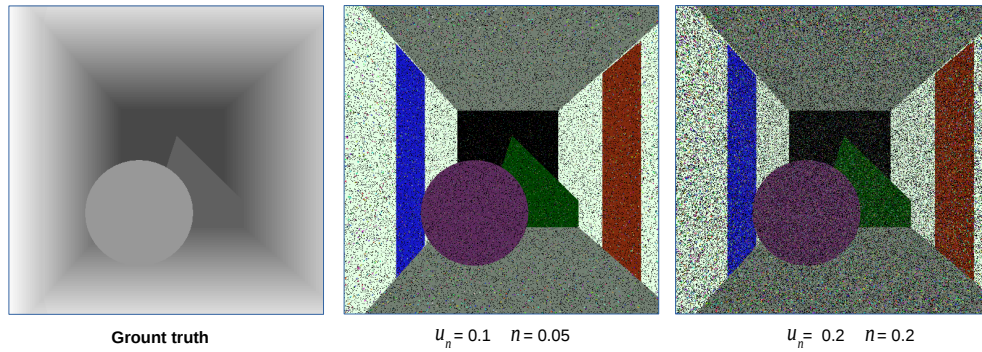


Figure 5.10: Artificial image with different amount of texture and pepper noise. [Best viewed in color]

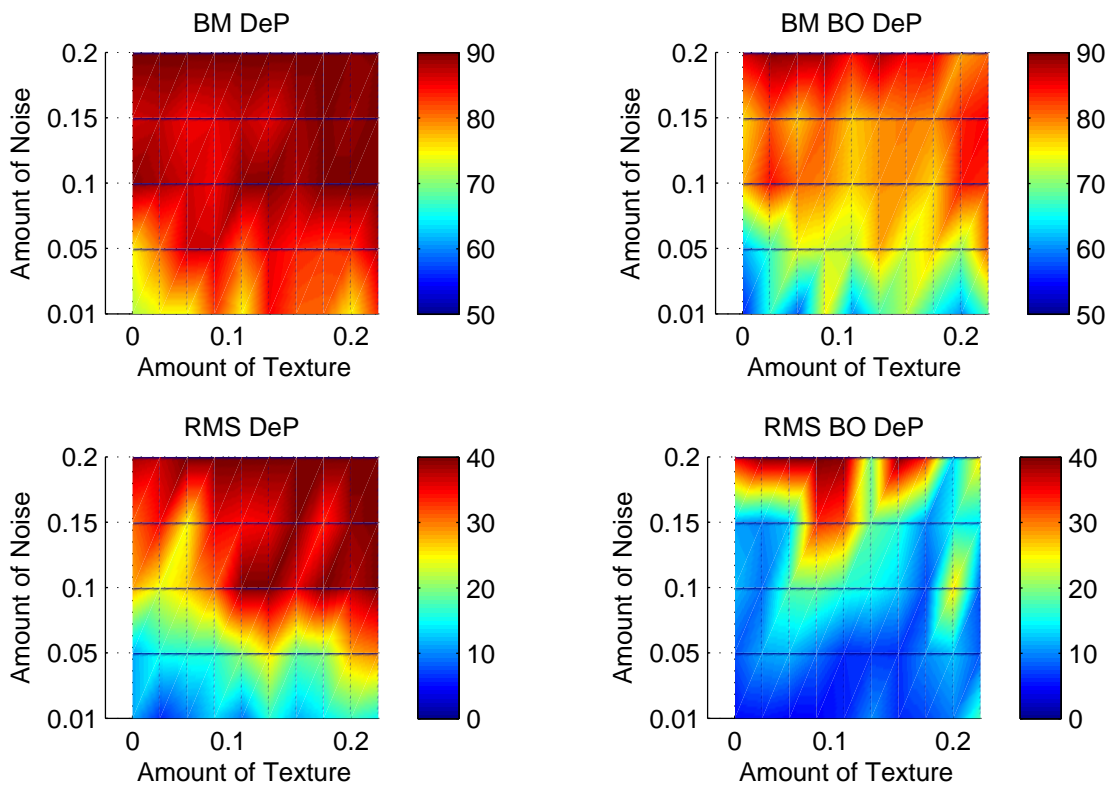


Figure 5.11: Root Mean Squares and Bad Mad Matching Percentage metrics are used with different amount of texture and pepper noises.

5.5.2 On a real world image

Depth prediction algorithm is also performed on a real world image. Stereo images are taken from a Bumblebee Stereo Vision camera. This camera has two calibrated camera lenses which offer taking images with different resolutions. In this study 1024x768 image resolution has been used.

For simplicity, scene consists of one object in the middle and a few surfaces with different colors. Surfaces with different colors are required for acquiring edges in the image. Furthermore, in the scene there must be edges whose orientation is vertical or almost vertical to find good correspondences since edges whose orientation is parallel to epipolar line have big uncertainties. Taken stereo image pair can be seen on Figure 5.12

Depth prediction results on the real world image can be seen on Figure 5.13. In the 5.14(a), extracted edge primitives are shown, then image pair located below shows the difference when the BO information is used. Using BO information, extracted surfaces are more accurate as located closer to the edges.

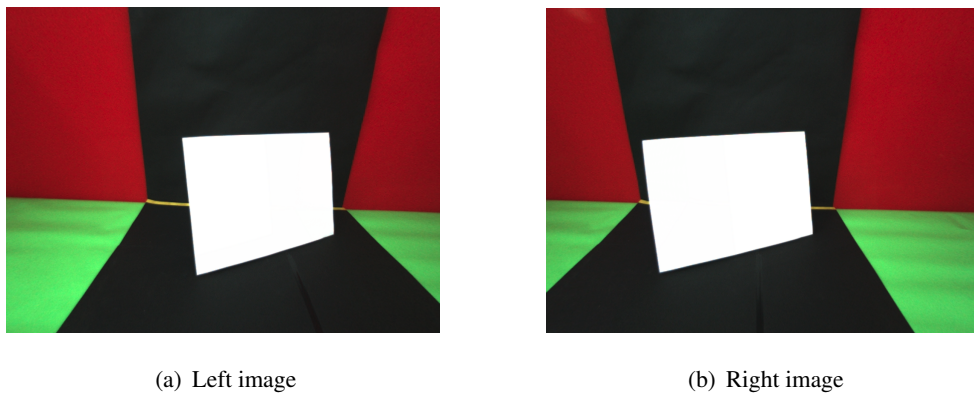
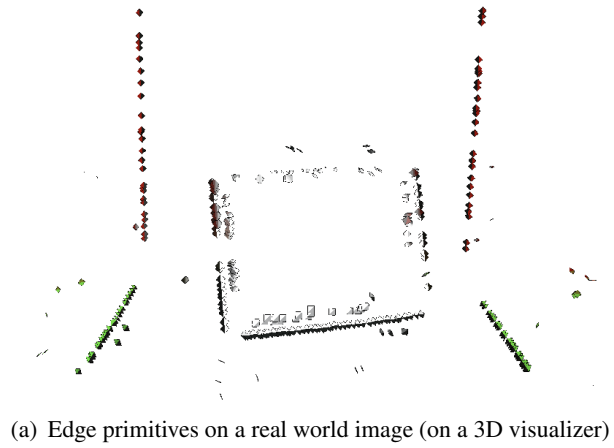
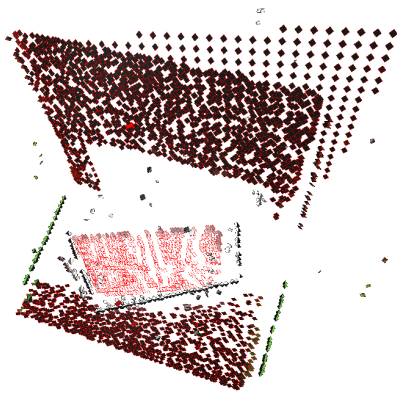


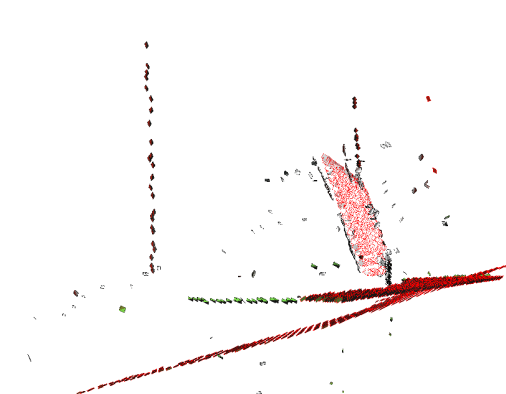
Figure 5.12: Stereo image pair taken from a Bumblebee stereo vision camera. In the scene, basic colors are used for easing the segmentation process. [Best viewed in color]



(a) Edge primitives on a real world image (on a 3D visualizer)

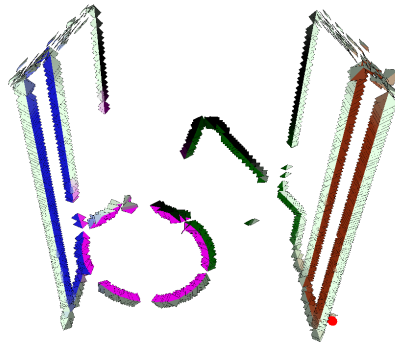


(b) Example scene from a real world image (on a 3D visualizer)

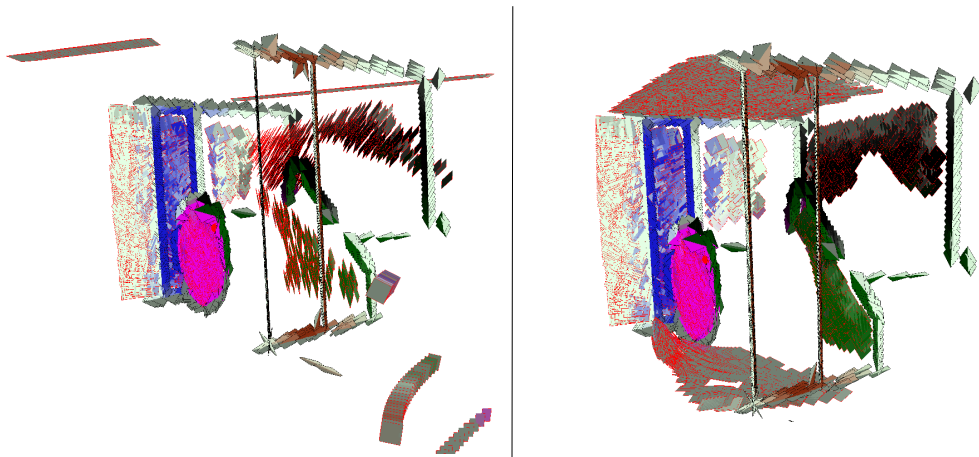


(c) Example Scene from a real world image (on a 3D visualizer)

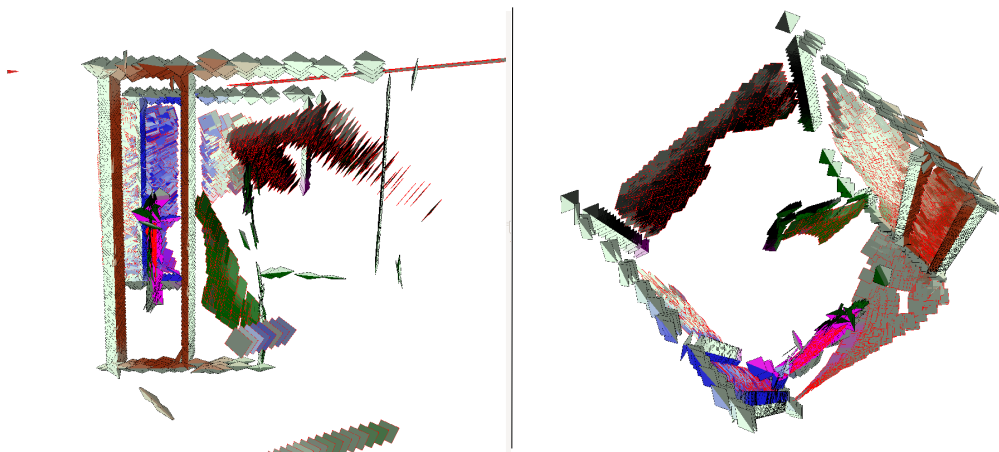
Figure 5.13: Image shows the difference on a real world image when the BO information is used. Using BO information, extracted surfaces are more accurate as positioned closer to the edges.



(a) Edge primitives on the artificial image (on a 3D visualizer)



(b) Example scene is shown from the artificial image. (on a 3D visualizer)



(c) Example scene is shown from the artificial image. (on a 3D visualizer)

Figure 5.14: Image shows the difference on a artificial image when the BO information is used. Using BO information, extracted surfaces are more accurate as positioned closer to the edges. Left image shows the depth prediction using BO information and right images shows depth prediction with BO information.

5.6 Time issues

The amount of time that depth prediction algorithm requires is related to amount of homogeneous image areas and size. For a 512x512 pixel² image, it needs 80-85 seconds on average without BO information and 50-55 seconds when the BO is used¹.

Utilization of BO information changes elapsed time dramatically with different amount of texture noise for artificial image. Results of elapsed time for a certain amount of texture noise can be seen on Figure 5.15.

BO information eliminates some edge primitives so the votes. Therefore, number of edge primitives to be clustered will be decreased (see section 5.3.2). Since clustering process is one the most time consuming part of the depth prediction model, elapsed time drops considerably. Interesting point in the figure is that for the image which does not include noise, depth prediction requires much more time than the others since it has more clear edge primitives and segments increases the time.

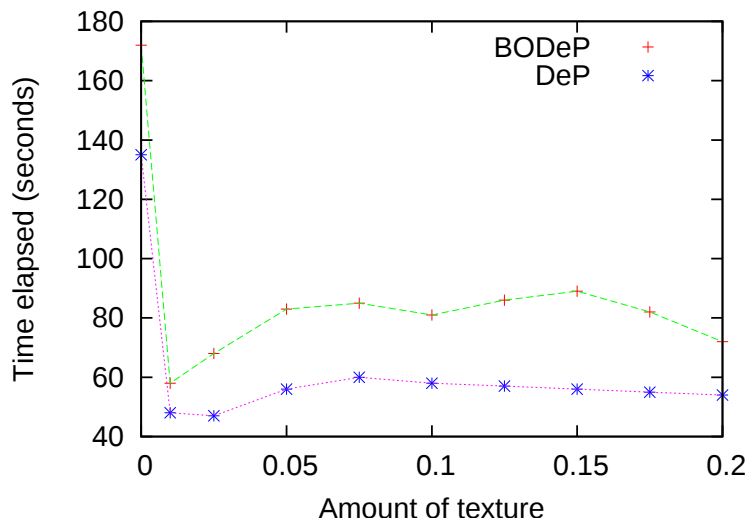


Figure 5.15: Elapsed time differs for the depth prediction when the BO information is used. For a mono, elimination of edge primitives, which are in the occluding area or on unrelated edges, shortens elapsed time.

¹ These results are gathered on a computer includes dual core processors with 4 GB ram

5.7 Summary

In this chapter, first types of depth prediction cues are detailed. Since monocular cues produces relative depth information, multi-view depth cues are used for absolute depth information. Then, under multi-view cues stereo is a useful method with two alternative types. Dense methods are those which performs well on textured image areas. On the other hand, sparse methods are better on low-textured image areas.

This study mainly use Kalkan et.al.'s 20007 study [18], which uses sparse stereo method for depth prediction on homogeneous image areas, and adapts BO information to obtain more accurate depth information on weakly textured or homogeneous image areas. The study presents promising results in terms of accuracy and time consumption. What is missing in the study is that BO information is not changed with the image when the applied noise changes, in other words static BO information is used for the simplicity. This part is kept as future work.

CHAPTER 6

CONCLUSION

In this thesis, a new comprehensive database is created with 500 indoor and 500 outdoor images. The BO database includes various types of indoor (shopping mall, living room, train station etc.) and outdoor images (animals, historical places, landscapes etc.) with 1000 images in total. Each contour in these images are labeled by three different participants to get more accurate human labeled groundtruth database. Furthermore, a user-friendly, well-documented online BO labeling tool is developed. Participants used this tool for labeling. Besides, another tool named segment annotation tool and online ownership labeling tool give rise to expand dataset with new images.

We performed analysis of the local visual cues for border ownership estimation. They are: *lower region, curvature, contrast, T-junction, L-junction*. Each cue is analyzed in terms of their predictive capability. Different number and different types of cues are combined in order to create more accurate prediction results. Analysis of conflicting cues are also performed and showed in a confliction matrix.

A new method for a feature based stereo has been developed to obtain more accurate depth perception on homogeneous image areas. This study mainly use Kalkan et.al.'s 20007 study [18], which uses sparse stereo method for depth prediction on homogeneous image areas, and adapts BO information to obtain more accurate depth information on weakly textured or homogeneous image areas. The study presents promising results in terms of accuracy and time consumption.

6.0.1 Future Work

There are several part that this study can be improved in the future. First of all, we combined visual cues using a basic majority voting. Our visual cues only give the BO direction, not with a confidence value. By getting the confidence value from each visual cue, combinatorial analysis can be realized more complex methods such as using logistic regression. We believe that other approaches can produce more realistic and accurate results.

In depth prediction part, only artificial and very basic real world scene are utilized. More complex real world and artificial stereo images can be added. Furthermore, depth prediction method with BO can be improved so as to predict areas not only flat surfaces, but also spherical surfaces.

REFERENCES

- [1] Lhi dataset, <http://www.imageparsing.com/>, February 2013.
- [2] Berkeley segmentation dataset, <http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/>, June 2014.
- [3] Border ownership labelling programme, <http://www.kovan.ceng.metu.edu.tr/bo/>, November 2014.
- [4] B.L. Anderson, M. Singh, and R.W. Fleming. The interpolation of object and surface structure. *Cognitive Psychology*, 44(2):148–190, 2002.
- [5] H. G. Barrow and J. M. Tenenbaum. Interpreting line drawings as three-dimensional surfaces. *Artificial Intelligence*, 17(1):75–116, 1981.
- [6] Xiaowu Chen, Qing Li, Dongyue Zhao, and Qinqing Zhao. Occlusion cues for image scene layering. *Computer Vision and Image Understanding*, 117(1):42 – 55, 2013.
- [7] Maureen Clerc and Stephane Mallat. The texture gradient equation for recovering shape from texture. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):536–549, 2002.
- [8] T. S. Collett. Extrapolating and Interpolating Surfaces in Depth. *Royal Society of London Proceedings Series B*, 224:43–56, 1985.
- [9] Ringo Doe. This is a test entry of type @ONLINE, June 2009.
- [10] Charless C Fowlkes, David R Martin, and Jitendra Malik. Local figure-ground cues are valid for natural images. *Journal of Vision*, 7(8):1–9, 2007.
- [11] W. E. L. Grimson. A Computational Theory of Visual Surface Interpolation. *Royal Society of London Philosophical Transactions Series B*, 298:395–427, September 1982.
- [12] Xiao Chen He and Nelson HC Yung. Corner detector based on global and local curvature properties. *Optical Engineering*, 47(5):057008–057008, 2008.
- [13] Jay Hegdé and David C Van Essen. A comparative study of shape representation in macaque visual areas v2 and v4. *Cerebral Cortex*, 17(5):1100–1116, 2007.

- [14] Derek Hoiem, Andrew N Stein, Alexei A Efros, and Martial Hebert. Recovering occlusion boundaries from a single image. In *IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.
- [15] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.
- [16] Allan Jepson and Whitman Richards. What makes a good feature. *Spatial vision in humans and robots*, pages 89–126, 1993.
- [17] Sinan Kalkan. *Multi-modal statistics of local image structures and its applications for depth prediction*. PhD thesis, Niedersächsische Staats- und Universitätsbibliothek Göttingen, 2008.
- [18] Sinan Kalkan, Florentin Wörgötter, and Norbert Krüger. Statistical analysis of second-order relations of 3d structures. In *VISAPP (2)*, pages 13–20. Citeseer, 2007.
- [19] Sinan Kalkan, Florentin Wörgötter, and Norbert Krüger. Depth prediction at homogeneous image structures. In *VISAPP (2)*, pages 520–527, 2008.
- [20] M. Kikuchi and Y. Akashi. A model of border-ownership coding in early vision. In *International Conference on Artificial Neural Networks*, pages 1069–1074, 2001.
- [21] M. Kikuchi and K. Fukushima. Assignment of figural side to contours based on symmetry, parallelism, and convexity. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 123–130, 2003.
- [22] H. Komatsu. The neural mechanisms of perceptual filling-in. *Nature Reviews Neuroscience*, 7(3):220–231, 2006.
- [23] Oliver W Layton, Ennio Mingolla, and Arash Yazdanbakhsh. Dynamic coding of border-ownership in visual cortex. *Journal of Vision*, 12(13):1–21, 2012.
- [24] Ido Leichter and Michael Lindenbaum. Boundary ownership by lifting to 2.1d. In *IEEE 12th International Conference on Computer Vision*, pages 9–16. IEEE, 2009.
- [25] H. Neumann, A. Yazdanbakhsh, and E. Mingolla. Seeing surfaces: The brain’s vision of the world. *Physics of Life Reviews*, 4(3):189–222, 2007.
- [26] H. Nishimura and K. Sakai. Determination of border ownership based on the surround context of contrast. *Neurocomputing*, 58-60:843 – 848, 2004.
- [27] Mary A Peterson and Elizabeth Salvagio. Inhibitory competition in figure-ground perception: Context and convexity. *Journal of Vision*, 8(16):1–13, 2008.

- [28] Nicolas Pugeault and Norbert Krüger. Multi-modal matching applied to stereo. In *BMVC*, pages 1–10. Citeseer, 2003.
- [29] Fangtu T Qiu, Tadashi Sugihara, and Rudiger von der Heydt. Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience*, 10(11):1492–1499, 2007.
- [30] Fangtu T Qiu and Rudiger Von Der Heydt. Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules. *Neuron*, 47(1):155–166, 2005.
- [31] Xiaofeng Ren, Charless C Fowlkes, and Jitendra Malik. Figure/Ground assignment in natural images. In *European Conference on Computer Vision*, pages 614–627, 2006.
- [32] Ko Sakai, Haruka Nishimura, Ryohei Shimizu, and Keiichi Kondo. Consistent and robust determination of border ownership based on asymmetric surrounding contrast. *Neural Networks*, 33:257–274, 2012.
- [33] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.
- [34] S. Treue, R. A. Andersen, H. Ando, and E. C. Hildreth. Structure-from-motion: perceptual evidence for surface interpolation. *Vision Research*, 35(1):139–48, 1995.
- [35] Shaun P Vecera, Edward K Vogel, and Geoffrey F Woodman. Lower region: A new cue for figure-ground assignment. *Journal of Experimental Psychology-General*, 131(2):194–205, 2002.
- [36] H. Zhou, H. S. Friedman, and R. von der Heydt. Coding of Border Ownership in Monkey Visual Cortex. *J. Neurosci.*, 20(17):6594–6611, 2000.
- [37] H. Zhou, H.S. Friedman, and R. Von Der Heydt. Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience*, 20(17):6594–6611, 2000.