ANALYSIS OF METHODS FOR BORDER OWNERSHIP ESTIMATION

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

BY

SERTAÇ OLGUNSOYLU

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

AUGUST 2015

Approval of the thesis:

ANALYSIS OF METHODS FOR BORDER OWNERSHIP ESTIMATION

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

Prof. Dr. Gülbin Dural Ünver _____
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 _____

Assoc. Prof. Dr. Ahmet Oğuz Akyüz
Computer Engineering Department, METU _____

Assist. Prof. Dr. Yusuf Sahillioğlu
Computer Engineering Department, METU _____

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

Date: **26.08.2015**

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: SERTAÇ OLGUNSOYLU

Signature :

ABSTRACT

ANALYSIS OF METHODS FOR BORDER OWNERSHIP ESTIMATION

Olgunsoylu, Sertaç
M.S., Department of Computer Engineering
Supervisor : Assist. Prof. Dr. Sinan Kalkan

August 2015, 56 pages

Border Ownership (BO) signifies which side of an image border owns the border. This information is very important for many computer vision problems. There are various approaches for estimating BO in the literature. One approach is to use methods that rely on low-level local features; combinations of local visual cues such as convexity or junctions and using spectral features have been proven to be quite effective. Alternatively, global constraints or cues can be imposed on top of local cue predictions to make better BO estimates. There are also physiologically-inspired attempts, which, e.g., model contrast surround modulation in complex cells to estimate BO. In this thesis, we analysed representative BO estimation methods using a comprehensive BO database, which includes 500 indoor and 500 outdoor images. Performance of each method is analysed individually in terms of estimation accuracy. In addition to this analysis, we improve a naive local cue combination method using Conditional Random Fields to impose global consistency on junctions. Moreover, we proposed a new method that uses AdaBoost algorithm to combine visual cues.

Keywords: Border Ownership, Figure-Ground Segregation, Conditional Random Field, Receptive Field, Visual Cues

ÖZ

SINIR SAHİPLİĞİ ELDE ETME YÖNTEMLERİNİN ANALİZİ

Olgunsoylu, Sertaç
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Yrd. Doç. Dr. Sinan Kalkan

Ağustos 2015, 56 sayfa

Sınır sahipliği resimdeki sınırların hangi alanlara ait olduğunu belirleme problemidir. Bu bilgi nesne tanıma, derinlik algılama ve hareket algılama gibi bir çok bilgisayarlı görme problemi için oldukça önemlidir. Literatürde bu probleme dair farklı yaklaşımlar bulunmaktadır. Sadece bölgesel görsel ipuçlarına dayanan yöntemler olduğu gibi resim genelindeki sınırlar arası tutarlılığı ek olarak kullanan yöntemler de bulunmaktadır. Bunlara ek olarak, biyolojik sistemlerdeki yapılardan esinlenerek onlara benzer yöntemler de öne sürülmüştür. Bu tez çalışmasında, literatürdeki farklı yaklaşımlara ait belirgin yöntemler detaylı olarak incelenmiştir. İncelemede 500 iç alan 500 dış alan resmi içeren geniş kapsamlı bir resim veritabanı kullanılmıştır. Her bir yöntem bağımsız olarak detaylı olarak incelenmiş olup tez kapsamındaki yöntemlerin başarımı kıyaslanmıştır. Bu incelemeye ek olarak bölgesel ipuçlarının genel sınır bilgisi ile birleştirilmesi gerçekleştirilmiştir. Ayrıca, "AdaBoost" algoritması bölgesel ipuçlarının bir arada kullanılmasını iyileştirmek üzere kullanılmış ve değerlendirişmiştir.

Anahtar Kelimeler: Sınır Sahipliği, Şekil-Zemin Ayrımı, Koşullu Rastgele Alan

To my family

ACKNOWLEDGMENTS

There are many people to thank for their help but it is possible to name a few.

First of all, I would like to thank to my wife for supporting me. I cannot complete my M.Sc. study without her supports.

I would like to express my respect and thank to my supervisor Sinan Kalkan for his invaluable guidance and support.

I am also thankful to my thesis jury members Prof. Dr. Fatoş Yarman Vural, Assoc. Prof. Dr. Ahmet Oğuz Akyüz, Assist. Prof. Dr. Yusuf Sahillioğlu and Assist. Prof. Dr. Erkut Erdem for their guiding comments.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ALGORITHMS

ALGORITHMS

LIST OF ABBREVIATIONS

BO	Border Ownership
VCC	Visual Cue Combination
CSM	Contrast Surround Modulation
CRF-S	Conditional Random Field on Shapemes
CRF-VC	Conditional Random Field on Visual Cues
SVM-SF	Figure-ground Classification Based on Spectral Features
AB-VC	AdaBoost on Visual Cues
SVM	Support Vector Machine
CRF	Conditional Random Field
GMM	Gaussian Mixture Model
HVS	Human Visual System

CHAPTER 1

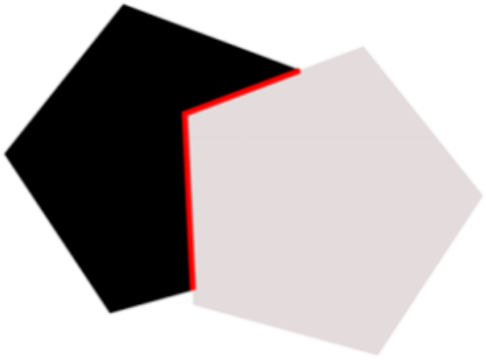
INTRODUCTION

Images are 2D projections of a complex world composed of 3D objects or structures. This projection is fulfilled through electro-optical sensors that transform light and change in light into electronic signals. The transformation process causes loss of visual information such as depth since the projection eliminates one dimension. Moreover, 3D surfaces are projected as regions. Textural details of surfaces are lost up to some extent in this projection. Therefore, the projection yields low-textured homogeneous areas in the image.

Homogeneous areas with low texture pose challenges for artificial vision systems especially in problems such as optical flow, stereo vision, and structure from motion since they require correspondence between different image regions. Biological systems are also affected because neurons in the visual cortex do not get activated by stimuli with uniform intensities because the change in receptive fields caused by such stimuli is insignificant [10, 11]. Biological and artificial vision systems have to deal with such data loss to perceive 3D information from 2D image data.

In order to tackle this problem, reliable information at the boundaries are used in "filling-in" activities to complete the missing information and to rectify ambiguities inside such uniform areas [12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. Such activities diffuse the information into regions from borders. This requires that boundaries of regions should be known in advance. In other words, each border has to be assigned "Border Ownership" (BO) information so that such filling-in activities could utilize the information. Two examples of border ownership (one artificial image and one real world image) is shown in Figure 1.1 where borders are marked with red lines.

(a) White object owns the border



(b) White pentagon owns the border

Figure 1.1: Example of BO information on real world and artificial images. Source: [1].

As shown in Figure 1.2, determining BO information can be quite challenging. There are many factors that affect how HVS assigns ownership to one of the regions that share the boundary. Due to its bistable nature, the drawing could be perceived easily as different groupings: two black face profiles on a white background, a white vase on a black background.

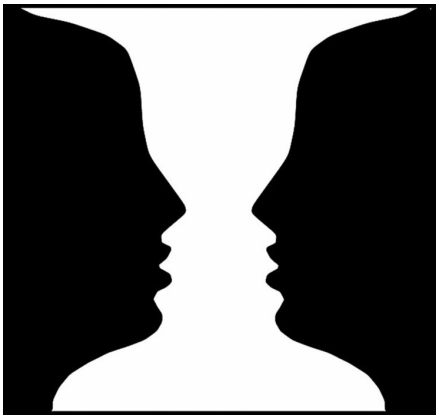


Figure 1.2: Rubin's vase.

1.1 Contributions

The contributions of the thesis are summarized below:

- A detailed analysis of representative BO estimation models on an extensive

human-marked BO dataset is presented. The effect of the crucial parameters of the computational methods have been analysed individually on dataset which consists of 500 indoor and 500 outdoor images. In addition, overall performance of all the methods are compared.

- A new method that applies AdaBoost algorithm on variety of local visual cues existing in the dataset has been developed. Predictive capabilities of the local cues are utilized in AdaBoost as weak classifiers.
- The method using naive combination of local cues is improved by extending it with CRF to impose global consistency on the borders. Instead of considering junctions as local cues, their structural information is used as source of consistency.

These contributions have been disseminated in the following:

- (Submitted) Mehmet Akif Akkuş, **Sertaç Olgunsoylu**, Gaye Topuz and Sinan Kalkan, Analysis of Visual Cues and Computational Models for Border Ownership, *Computer Vision and Image Understanding*.

1.2 Outline of the thesis

The organization of thesis is as follows:

- **Background and Literature Survey**
  Essential information about perceptual organization and BO problem, computational models and algorithms used in this thesis, and current literature and related works on BO problem are provided in Chapter 2.
- **Methods**
  In Chapter 3, computational methods that we analyze in the scope of this thesis are explained in details.
- **Experiments and Results**
  Chapter 4 presents results of analysis work and experiments for each method and overall results.

- **Conclusion**

  Thesis is concluded with a discussion and future work in Chapter 5.

CHAPTER 2

BACKGROUND AND LITERATURE SURVEY

In this chapter, essentials of the BO problem are given as well as its formal definition. Also, related BO studies and computational models used throughout the thesis are explained.

2.1 Border Ownership Problem

Visual systems interpret visual scenes inferring 3D structure of the real world objects from their 2D projections on our retinae. Perception process starts at the retinae on this 2D visual data. The interpretation is achieved in Human Visual System (HVS) employing physiological and cognitive mechanisms. How does HVS accomplish such a significant perceptual task is one of the fundamental questions that multiple disciplines try to find an explanation. The initial attempt to bring an explanation to this problem from cognitive psychological aspect belongs to a group of researchers. These researchers, known as Gestalt psychologist, introduced a formulation to this problem early in the twentieth century [22]. Gestalt means 'shape' or 'form' in German. Therefore, their work is known as Gestalt principles. These principles formulate how the visual input is grouped into unitary forms or shapes in HVS, which kind of perceptual input plays important role in perceptual organization. Although there is no definitive list of Gestalt principles, some of the most commonly discussed are illustrated in Figure 2.1.

This perceptual organization task becomes complicated by occlusions between objects in the scene. The occlusion yields boundaries between neighbouring objects,

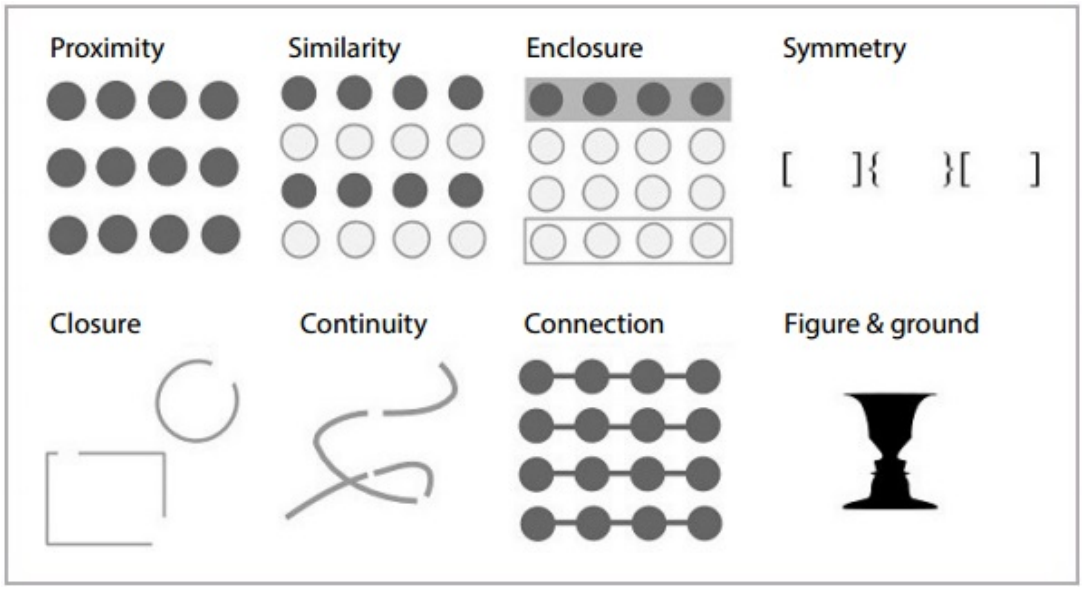


Figure 2.1: Different Gestalt principles cause different perceptual groupings. Source: [2].

which belong to the occluding objects. Gestalt psychologists discovered that BO is an important factor in perception because the perception of shapes depends on the assignment of these borders. The drawing shown in Figure 1.2 illustrates the effect of assignment of BO on perception of shape. The shape perceived heavily depends on the direction of border ownership in this figure. As shown, there are two borders that separate white and black regions. If the owner of these border is considered as the white region, then the shape we perceive become a white vase on a black background. On the other hand, if we take black regions as the owner of these borders, then two black profile faces are seen on a white ground. Since shape perception is affected by BO assignment, perception becomes unstable (or bistable) under such circumstances. It is clearly observed that there is a competition of interpretations between mechanisms in HVS as the alternation leads to some extra delay in perceptual grouping.

Although perceiving which objects own which regions seems obvious, it is not a trivial task. It is heavily unknown that how the visual system achieves this task and there is no computational methods which is completely biologically plausible. Nevertheless, there are two explanations of this process. The first approach relies on the figure-ground organization, which proposes that the brain labels regions as figure and ground. The other implies that border-ownership is assigned to occluding shapes or

regions. There are studies that provide evidence for both approaches. Lamme [23] noticed that certain neurons fire at a higher rate in response to the stimulus originated from figure region compared to the ones from ground region. On the other hand, it is observed that certain neurons fires at a different rates depending on whether BO is assigned to the figure region. However, it is not known that how these two approaches exist together and relate to each other.

2.2 Border Ownership in Neuroscience

BO selective neurons in HVS was discovered early in the recent decade. Zhou et al. [24] found out that 18% of the cells in V1 and more than 50% of the cells in V2 and V4 areas respond at different firing rates depending on the direction of BO.

After their discovery, some insight into how the visual system estimates border ownership information has been attained. It has become apparent that visual cues such as contrast [25], depth order [26], curvature [24, 27] play a key role in the BO estimation process.

Qui and Heydt [26] stated that neurons in V2 regions of macaque monkeys' visual cortex show selectivity to the depth order of 3D images. Those neurons get activated more by the closer parts of the figures.

Layton et al. [25] observed that neurons in V2 area may demonstrate selectivity to contrast changes on boundaries. Light-dark and dark-light contrast changes cause activation in these neurons.

Another important visual cue which found to be used in BO assignment is curvature. Zhou et al. [24] noticed that figures with convex contours activate BO selective neurons in V2 and V4 regions in the monkey visual cortex. It is stated that V4 neurons which has curved and large receptive fields are sensitive to convexity [27].

The BO selective neurons determine the direction of the owner 10-25 ms [24] and the BO sensitivity starts in V1 area. In the light of these findings, it is noticed that local visual cues could be used for BO estimation.

2.3 Border Ownership in Psychology

There are many psychological studies that investigate the BO problem and mechanisms employed in the estimation process. It has been stated [3] that convex visual regions are perceived as figure rather than concave regions by human subjects (see Figure 2.2).

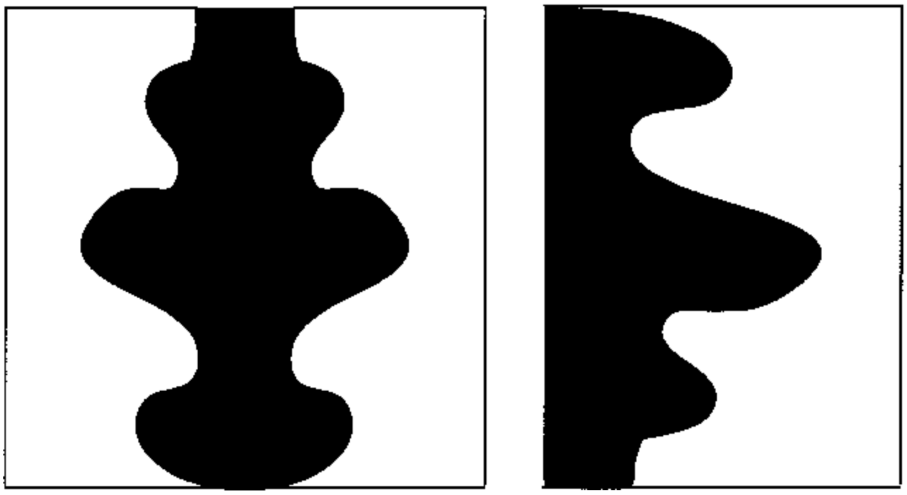


Figure 2.2: Example images in which convex regions correspond to figures. Source: [3].

Lower region, T-junctions, and L-junctions are the other visual cues which are known to be related to the BO problem. Vecera et al. [3] investigated the role of some Gestalt cues in BO estimation. Their experiments showed that lower regions are perceived as figure rather than the upper regions in the image (see Figure 2.3).

Figure 2.3: Example images in which lower regions correspond to figures. Source: [3].

2.4 Border Ownership in Computer Vision

Even though the importance of Border Ownership problem is identified by Koffka [28] in 1935, there are not many computational studies regarding BO until the last two decades. First computational models are neural network based, which make use of local cues such as curvature and L-junctions, and used local propagation and competition mechanisms among the several networks to obtain BO estimation on borders. In a similar way, Sakai et al. [29] developed a neurophysiologically plausible receptive field based model on surrounding contrast. Contrast selective simple cells are used and combined with inhibitory and excitatory units to arrive at a final estimation. In their early attempt, they only used artificial images for testing the model. Later, they extend their model so as to be used on natural images [6].

Fowkles et al. [30] used size, lower region and convexity as local cues for predicting figure-ground in 200 nature images. According to their analysis, the best performing cue was the size with 68% accuracy only by itself. Also, they combine multiple cues fitting a logistic regression to the training data. Pair of lower region and size were more powerful than the combination of size and convexity cues.

Fowkles et al., in another study [9], used Conditional Random Field (CRF) on a local classifier. They developed a logistic classifier to locally predict figure/ground labels. Local shapes are described using local feature categories that they called shapemes. CRF is used to enforce global consistency by learning T-junction frequency. Their model outperforms when compared to a baseline model using size and convexity cues. They analysed their model on a set of 100 labeled Berkeley images, they reported 78.3% accuracy.

Recently, Chen et al. [7] used several visual cues namely, semantic, compactness, position (i.e., lower region), junction as well as convexity cues to predict the occlusion relationships between 200 rural, 250 artificial images as well as 645 outdoor images. In addition to BO estimation, they also inferred layer sequence of the image scene. In order to infer depth order, they developed semantic label map where each region is labeled with human semantic label along with the occlusion relationship between them. They combined the semantic cue with the other local cues and used them as

weak classifiers in the AdaBoost algorithm. According to the their experiment results, semantic cue was the most powerful among those five cues.

2.5 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning model that constructs one or more hyperplanes in a multi-dimensional space [31] based on a set of training data. The hyperplanes divide the space into subspaces. Data points located in a particular subspace are the member of its corresponding cluster. Goodness of the separation depends on both the distance to the nearest data points from the clusters (functional margin) and generalization error of the classifier.

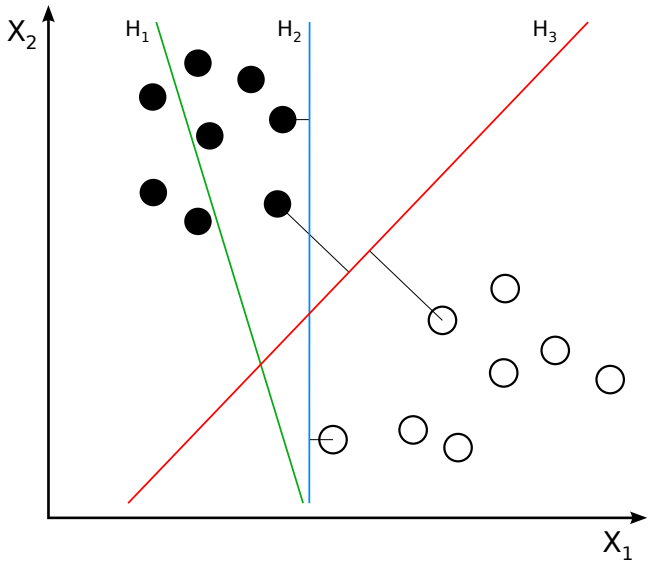


Figure 2.4: Infinitely many hyperplanes can separate two sets of points in a 2D space. SVM finds the one that has maximum functional margin and minimum generalization error. Source: [4].

As seen in Figure 2.4, there are three hyperplanes in a 2D space. Hyperplane H_1 does not separates the classes, on the other hand, H_2 and H_3 are able to separate them correctly. However, H_3 has larger functional margin than H_2, so it yields better classification compared to H_2.

SVM is very well suited to binary classification problems. Since BO problem requires to assign border ownership to one of two regions that share the boundary, SVM is also

suitable for BO estimation.

2.6 AdaBoost

In machine learning, boosting is an approach which aims to create accurate prediction rule by combining a number of relatively weak and less accurate classifiers. AdaBoost, also known as Adaptive Boosting, is machine learning algorithm devised by Freund and Schapire [32]. It is the first boosting algorithm which is widely used and studied. It is known as being adaptive because it tunes the subsequent classifiers for the misclassified instances of the previous classifiers.

Algorithm 1: AdaBoost algorithm

Given: $(x_1, y_1), ..., (x_m, y_m)$ where $x_i \in \mathcal{X}$, $y_i \in \{-1, +1\}$.

1: Initialize $D_1(i) = 1/m$ for $i = 1, ..., m$.

2: Update the weighting.

for $t = 1, ..., T$ **do**

2.1: Train weak learner using distribution D_t.

2.1: Get weak hypothesis $h_t : \mathcal{X} \to \{-1, +1\}$ with error:

$$\epsilon_t = Pr_{i,D_t}[h_t(x_i) \neq y_i]. \tag{2.1}$$

2.2: Choose $\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$.

2.3: Update, for $i = 1, ..., m$:

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}, \tag{2.2}$$

where Z_t is a normalization factor.

end for

3: Output the final hypothesis:

$$H(x) = \text{sign} \left(\sum_{t=1}^{T} \alpha_t h_t(x) \right). \tag{2.3}$$

Algorithm 1 shows the pseudocode for AdaBoost. The algorithm is given a set of training examples $\mathcal{X} = x_1, ..., x_m$ as well as their corresponding labels $y_i \in \{-1, +1\}$. For each iteration $t = 1, ..., T$, weight distribution of the iteration D_t is computed over all the training examples, and a given weak classifier is applied to maintain the weight distribution. All weights are set equally in the beginning; however, algorithm increases the weights of misclassified examples. This weight ma-

nipulation aims to force the weak classifier to focus on the challenging examples. The weak learner should find a weak hypothesis against the distribution D_t. The final hypothesis is computed as a weighted combination of the weak hypotheses.

In this study, we use predictive capabilities of local visual cues as weak classifiers and combine them using AdaBoost algorithm.

2.7 Conditional Random Field (CRF)

Conditional random field (CRF) is probabilistic graphical model to be used for labelling and segmenting structured data. CRF takes the context into consideration unlike standard approaches which ignores relationships in the context while predicting a label for a sample. Using the available global information, CRF encodes relationship between observations which leads to globally consistent interpretations. It is generally used for labelling sequential data in various domains such as natural language processing, bioinformatics, and computer vision.

In computer vision, CRF is suitable for capturing interaction between visual constructs and encode contextual relationships in the visual input. It is proven that CRF could be applied to fundamental vision problems such as object detection. Standard methods utilize the visual information that exists in the local patches, whereas using CRF is helpful for resolving local ambiguities employing global information for solving tasks such as object detection and figure/ground assignment [9, 33].

In this study, we use CRF to enforce global constraints that junctions imply in order to correct predictions which is made using only local cues.

2.8 Summary

There is a progress in studies from Psychology and Neuroscience regarding which visual cues are used for the BO problem and where, in the brain; however, the underlying mechanisms of HVS that estimates BO is largely unknown. Nevertheless, the problem is becoming one of the hot problems in computer vision.

Although there are promising computational studies for estimating BO, there does not exist any study in the literature which analyses existing methods on a comprehensive dataset. In this study, we aim to fill this gap analysing representative method on the METU Border Ownership Dataset [34].

CHAPTER 3

BORDER OWNERSHIP ESTIMATION METHODS

In order to estimate border ownership information, various computational models and methods have been developed as presented in Chapter 2. In this chapter, representative methods analysed in this thesis will be explained in more detail.

3.1 Visual Cue Combination

Psychological studies [3, 35] have shown that visual cues such as contrast, curvature, depth order are important for estimating border ownership. Although how these visual cues are utilized in the human vision system is not completely illuminated, it is known that they play important role in the estimation process. With these findings several computational models using these have been proposed. However, visual cues may not occur on some borders. In the absence of a visual cue, BO estimation fails if it depends on only this particular cue.

Akkus [1] proposed a method based on combining visual cues. The motivation behind this method is to improve estimation of BO by combining visual cues. Collective utilization of visual cues not only atones for unavailability of cues but also improves the confidence by boosting the prediction that majority agrees on.

The visual cues used by Akkus [1] are lower region, curvature, contrast, T-junction, L-junction.

3.1.1 Cues for Border Ownership

3.1.1.1 Lower Region

Regions in the lower positions of a stimulus, i.e. image, are more likely to be perceived as figure than upper regions. Lower region is introduced as a new visual cue, first by Vecera et al. [3]. It is used for estimating border ownership [30], [36].

For a border b separating two regions r_a and r_b, region $\hat{r}_l(b)$ that owns the border is obtained by the lower region cue as follows:

$$\hat{r}_l(b) = \underset{r \in \{r_a, r_b\}}{\arg\min} \underset{(x,y) \in r}{\arg\min}\, y, \tag{3.1}$$

where (x, y) is a 2D coordinate of a point on region r.

As shown in Figure 4.1, hay stack (R1) is in the lowest position in the image, which is also the closest object to the camera. Then, the warehouse (R2) is the second lower region, the trees are the third and so on.

Figure 3.1: Sample image illustrates lower region on a real image. The right hand side, the segmented image shows corresponding regions of the real image. Source: [1].

3.1.1.2 Curvature

Curvature is one of the most informative visual cues. It is stated that figures with convex shapes activate BO selective neurons [24], [37]. The early computational models for estimating BO, neural network based, have used curvature along with the L-junctions as visual cues [38], [39].

A curved border in a 2D image is projection of the edge of a curved surface. Therefore, the border is more likely to be owned by this convex surface, or region in 2D [30]. The higher degree of convexity results in more likelihood of the ownership on regions.

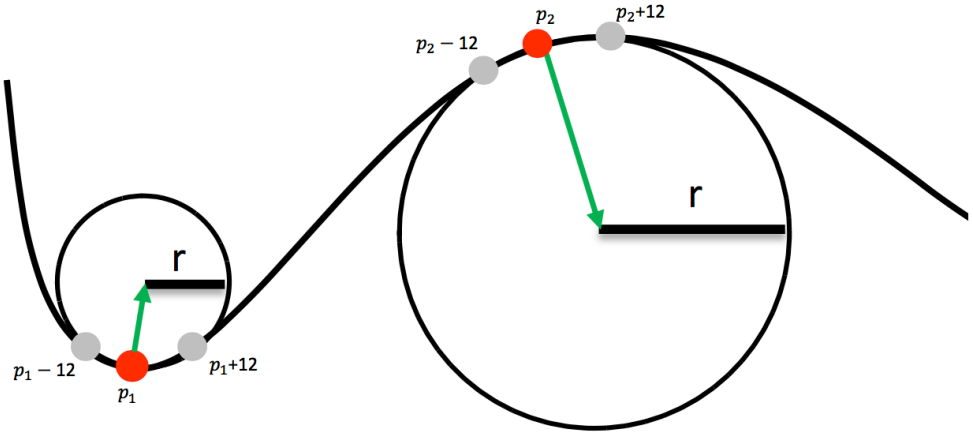


Figure 3.2: Illustration of circle fitting on a curve. Source: [1].

The curvature of a point is represented as a vector whose magnitude, $m(p_i)$, is the inverse of the radius of a circle fitted at that point. As shown in Figure 3.2, circle o_{p_i} fits at a point p_i along a border b such that two more points which are on the border and equidistant to point p_i are selected. These three points define an arc on circle o_{p_i}.

$$m(p_i) = 1/r, \tag{3.2}$$

where r is the radius. The direction of the curvature, $s(p_i)$, is the direction of the vector that connects p_i to the center of the circle o_{p_i}.

$$s(p_i) = c(o_{p_i}) - p_i, \tag{3.3}$$

where $c(o_{p_i})$ is the center of the circle.

For a border b separating two regions r_a and r_b, region $\hat{r}_c(b)$ that owns the border is obtained by the curvature cue as follows:

$$\hat{r}_c(b) = \underset{r \in \{r_a, r_b\}}{\arg\max} \sum_{p \in b} (s(p) \rightarrow r) \cdot m(p), \tag{3.4}$$

where $s(p) \rightarrow r$ is 1 if the direction of the curvature points to region r, otherwise

0. As seen in Equation 3.4, all the points along a border contribute to curvature confidence. Region that yields highest confidence score is determined as the owner.

Figure 3.3: Sample image shows curvatures on an image. Convex regions are more likely to own the borders as shown. Source: [1].

3.1.1.3 Contrast

Contrast[1] is one of the visual cues utilized for estimating border ownership [25]. The brighter objects are supposed to be closer to the camera. Being closer implies that these brighter regions are more figurelike than darker regions.

For a border b separating two regions r_a and r_b, region $\hat{r}_{co}(b)$ which is owner of the border is determined as follows:

$$\hat{r}_{co}(b) = \underset{r \in \{r_a, r_b\}}{\arg\max} \sum_{p \in r} \frac{1}{N} I(p), \tag{3.5}$$

where $I(p)$ intensity value of pixel p and N is the total number of pixels within the region.

3.1.1.4 T-junction

Junctions are significant features that keep information about 3D structure of objects. T-junction accurately determines surface that occludes others. Hence, it has been used for detecting occlusions [40] and figure-ground segregation [9]. Given a junction j, it is a T-junction if there are three line segments and one of the angles is close to $180°$(see Figure 3.4).

[1] Although the right term would be "intensity", "contrast" is used to follow naming convention around the concept in the literature.

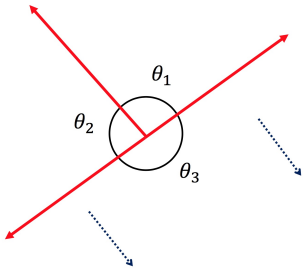


Figure 3.4: Visual representation of T-junction. Source: [1].

The region $\hat{r}_T(b)$ that owns the border b is determined to be the region facing the largest angle. This is illustrated in Figure 3.5.

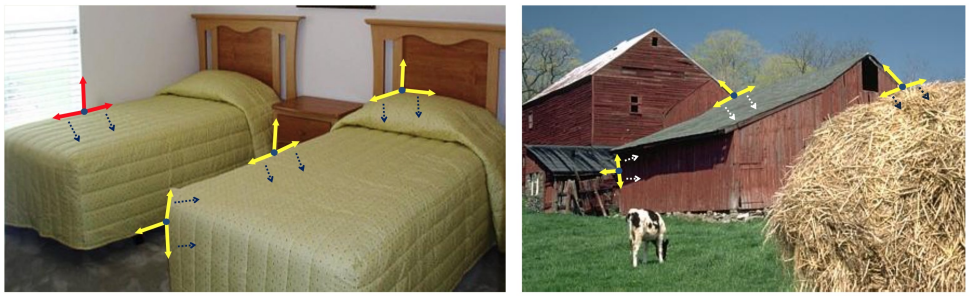


Figure 3.5: Sample indoor and outdoor images where T-junctions determine border ownership. Dashed arrows point out the owning region. Source: [1].

3.1.1.5 L-junction

L-junction is also an informative cue for border ownership [38], [39]. Unlike T-junction, L-junction consists of two line segments and two angles.

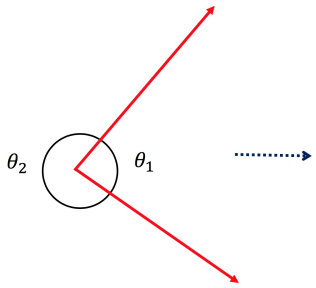


Figure 3.6: Visual representation of L-junction. Source: [1].

The region $\hat{r}_L(b)$ that owns the border b is determined to be the region facing the

smallest angle (see Figure 3.6).

Figure 3.7: Sample indoor and outdoor images where L-junctions determine border ownership. Dashed arrows point out the owning region. Source: [1].

3.1.2 Combination of Visual Cues

The visual cues explained in Section 3.1.1 can be utilized as any possible combinations of them so as to improve prediction accuracy. The approach chosen to combine the cues is the majority rule. In other words, the region $\hat{r}_{comb}(b)$ owning the border b is determined as more cues are agreed on (see Equation 3.6).

$$\hat{r}_{comb}(b) = \underset{r \in \{r_a, r_b\}}{\arg\max} \sum_{c \in \mathcal{C}} \mathbf{1}_{\{\hat{r}_c(b)=r\}}, \tag{3.6}$$

where $\mathcal{C}$ is cues used in combination and $\mathbf{1}_{\{\hat{r}_c(b)=r\}}$ is 1 if prediction of cue c on border b is region r, otherwise 0.

3.2 Figure-ground Classification Based on Spectral Features

Deciding which side of a border is the figure can be done utilizing global cues as well as local ones. Local cues depend on only small image patches; hence, they can be extracted and processed using less computational resources unlike global cues.

Ramenahalli et al. [5] proposed a method accompanied with a novel type of local feature that describes spectral properties of boundary image patches. This feature vector encodes spectral anisotropies via notion of oriented energy spectrum for both directions namely, parallel and orthogonal to the border. This method uses the feature vector to train a SVM classifier so as to classify borders on unknown patches.

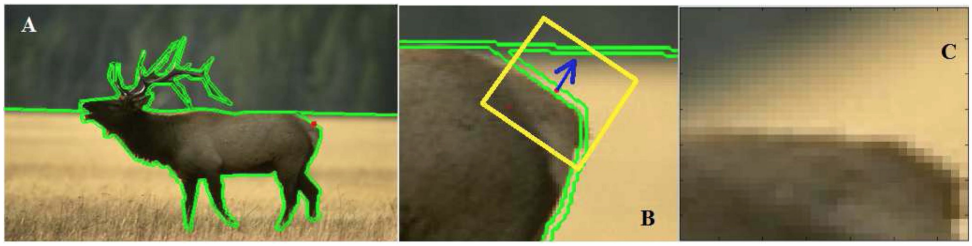


Figure 3.8: Patch extraction. (A) An outdoor segmented image where red dot shows the location of the patch to be extracted. (B) An image patch in its original orientation. Yellow box shows the window that bounds the patch area. (C) Final image patch after rotation applied. Source: [5].

In this method, an image patch $\boldsymbol{p}(x, y)$ of size N x $N = 36$ x 36 that strides the border such that occluding regions cover the patch in half shares is selected, as Ramenahalli et al. did in [5]. Then, the patch is rotated such that figure is the lower half and the ground is the upper half as described in Figure 3.8. The oriented energy spectrum of a region (figure or ground), parallel and orthogonal to the figure-ground boundary is defined in Equation 3.7 and Equation 3.8, respectively.

$$E_r^{\parallel}(u, y) = |\boldsymbol{P_r^x}(u, y)|^2, \tag{3.7}$$

where y coordinate varies orthogonally to the boundary and $\boldsymbol{P_r^x}(u, y)$ is the discrete one-dimensional Fourier transform with respect to x axis, parallel to the boundary.

$$E_r^{\perp}(x, u) = |\boldsymbol{P_r^y}(x, u)|^2, \tag{3.8}$$

where x coordinate varies parallelly to the boundary and $\boldsymbol{P_r^y}(x, u)$ is the discrete one-dimensional Fourier transform with respect to y axis, orthogonal to the boundary. The average oriented energy of a region is computed as follows:

$$\bar{E}_r^{dir} = \frac{1}{K} \sum_c E_r^{dir}(c), \tag{3.9}$$

where E_r^{dir} can be either $E_r^{\parallel}(u, y)$ or $E_r^{\perp}(x, v)$, c is x or y coordinate and K is the number of points c depending on the desired direction of the spectrum.

The total oriented energy of a region r is defined as,

$$T_{dir}(r) = \int \bar{E}_r^{dir}(u) du, \tag{3.10}$$

where dir indicates in which direction the total energy to be computed.

The 4-dimensional feature vector is defined such that total energy parallel and orthogonal to the figure-ground border for both upper region r_u and lower region r_l in the patch as shown in Equation 3.11.

$$\boldsymbol{f} = [T_{\perp}(r_l), \qquad T_{\parallel}(r_l), \qquad T_{\perp}(r_u), \qquad T_{\parallel}(r_u)]. \tag{3.11}$$

Feature vector, obtained from a selected image patch, is a positive sample (tagged with $+1$). Then, elements of the same feature vector is flipped horizontally to obtain negative samples (tagged with -1). The SVM is trained using these positive and negative samples.

3.3 Contrast Surround Modulation

Sakai et al. proposed localized, asymmetric surround modulation as a mechanism for figure-ground segregation [6]. Neurophysiological studes have reported that the surround modulation plays an important role in this process [41], [23], [42], [43], [44]. The model proposed consists of three stages: contrast detection using Gabor filters, surround modulation, and border ownership determination (see Figure 3.9). In the model, entire image region I is spanned with a matrix of model cells such that one unit length, i.e. a pixel, corresponds to visual angle of $0.125°$.

3.3.1 Stages of the model

The response of a model cell in stage i for orientation ori is represented by $O^i_{ori}(x, y)$, where (x, y) is 2D cartesian coordinate of the model cell.

Contrast detection

In the first stage, the image contrast is detected and it is normalized afterwards. The contrast detection is done via convolution (*) of the pixels with Gabor masks in a particular orientation, and the output is filtered using half-wave rectification (see 3.12).

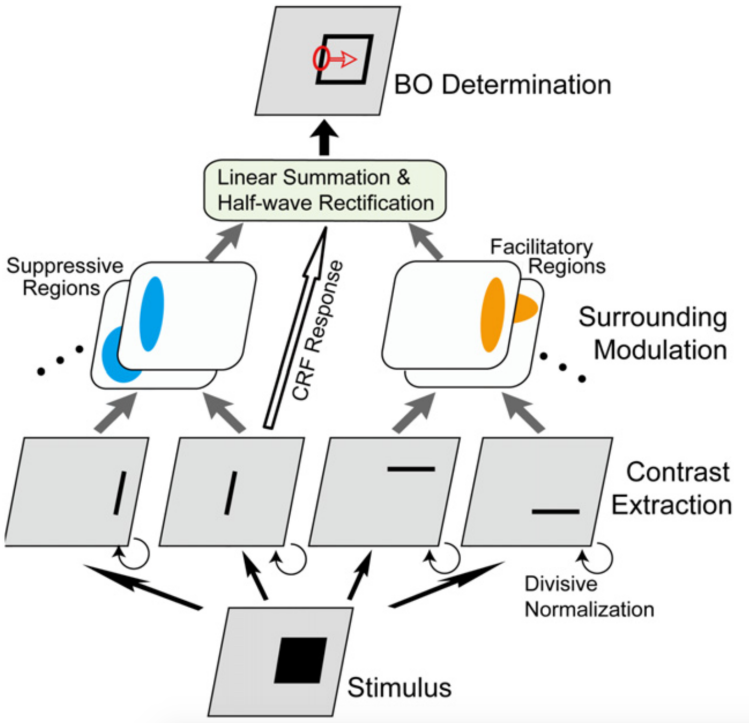


Figure 3.9: Stages of surround modulation. Source: [6].

$$T_{ori}(x_0, y_0) = \begin{cases} I_0 * G_{ori}, & \text{if } T_{ori} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.12}$$

where I_0 is a subregion of an input image whose center is (x_0, y_0). The gabor mask has a Gaussian distribution in orientation ori such that the standard deviation of the Gaussian, the wavelength and phase of the cosine are 1.25°, 1.25°, and 90°, respectively. Eight types of preferred orientation is employed so as to obtain more precise determination of contrast. The orientations are 0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315°. After contrast is detected, it is normalized as described in Equation 3.13.

$$D_{ori}(x_0, y_0) = \frac{c_1 T_{ori}(x_0, y_0)}{1 + c_2 T_{all}(x_0, y_0)}, \tag{3.13}$$

where

$$T_{all}(x_0, y_0) = \sum_{ori; x, y \in I_1} I_1(x, y) T_{ori}(x, y). \tag{3.14}$$

In Equation 3.14, c_1 and c_2 are equal to 10.0 and 0.01, respectively. I_1 is a circular region, centered at x_0, y_0, with a Gaussian distribution (standard devaiation = 0.75°). The response of the cell is normalized by the total amount of contrast within

its neighborhood for eight orientations. Then, the response passes through compressive nonlinearity function $s(x)$:

$$O^1_{ori}(x_0, y_0) = s(D_{ori}(x_0, y_0))), \tag{3.15}$$

where

$$s(x) = \frac{1}{1 + \exp(c_{slope}(c_{thres} - x))}, \tag{3.16}$$

c_{thresh} and c_{slope} are set to 9.0 and 2.0 [29] respectively. $s(x)$ resembles the Naka–Rushton function [45].

Surround Modulation

In the second stage, iso-orientation signals are used for suppression and cross-orientation signals for facilitation in the subregion:

$$O^2_{ori,i}(x_0, y_0) = (F_i * O^1_{cross(ori)})(x, y) - (S_i * O^1_{iso(ori)})(x, y), \tag{3.17}$$

where

$$O^1_{cross(ori)}(x, y) = \sum_{o \notin \{ori, ori+180\}} O^1_o(x, y), \tag{3.18}$$

$$O^1_{iso(ori)}(x, y) = \sum_{o \in \{ori, ori+180\}} O^1_o(x, y). \tag{3.19}$$

F_i and S_i are facilitatory and suppressive regions, respectively. Each region is a 2D Gaussian whose center and standard deviations are selected randomly within physiological range as did in [6]. There are 40 suppressive and facilitatory regions and i indicates the index of the region. For eight orientations and existing 40 regions for each, there are 320 types of cell response for a single location in the second stage.

Border ownership determination

After suppression and facilitation, done in the second stage, direction of border ownership is determined in the third stage. In this stage, cell responses are computed only for the location in which human-marked contours (HMCs) exist. Response of the cells which are not located on HMC is zero. Determination is not done for those

locations for the sake of performance. Response of a cell whose location is (x_0, y_0) is described in Equation 3.20.

$$O^3_{ori,i}(x_0, y_0) = O^1_{HMC,ori}(x_0, y_0)(O^1_{HMC,ori}(x_0, y_0) + k_{nat} O^2_{ori,i}(x_0, y_0)), \quad (3.20)$$

where k_{nat} quantifies contribution of surround modulation, which is originally set to 1.0. If $O^3_{ori,i}(x_0, y_0) < 0$, then $O^3_{ori,i}(x_0, y_0) = 0$. Strength of border ownership is computed for eight directions for each F, S region pairs at each location. Border ownership strength for direction dir at location (x_0, y_0) is obtained as follows:

$$BO_{dir,i}(x_0, y_0) = O^3_{dir,i}(x_0, y_0) - O^3_{dir+180,i}(x_0, y_0). \quad (3.21)$$

A vector summation is applied to the responses of 320 types of model cell at a location to obtain final border ownership direction at that location. If the difference between the predicted direction and the original direction is less than $90°$, it is considered as correct determination.

3.4 AdaBoost on Visual Cues

Chen et al. [7] proposed a method to determine border ownership using AdaBoost algorithm on visual cues. Their method utilizes similar low-level visual cues that we explained in Section 3.1. They also have an additional cue called "Semantic cue". Semantic cue is simply a mapping between image pixels and semantic label from a predefined label set. As shown in Figure 3.10, each semantic region is assigned with a label. They extract occlusion relationship of semantic regions and make use of it in their classifier. The other cues are lower region, curvature, junctions, and compactness.

Figure 3.10: Example of a semantic cue. Source: [7].

All these cues except semantic cue exist in our dataset. On the other hand, they do not make use of contrast cue. In our method we combine our visual cues with theirs so as to define a feature vector for our classifier.

3.4.1 Visual cues

In this section, all the cues will be explained with their mathematical definitions. The cues that we already mentioned in Section 3.1 will not be explained in detail.

3.4.1.1 Lower region

Given two regions r_i and r_j, lower region is defined as follows:

$$LR(r_i, r_j) = \frac{1}{1 + \exp(\frac{\hat{y}_j - \hat{y}_i}{H})}, \tag{3.22}$$

where $\hat{y}_i$ and $\hat{y}_j$ are average height of regions r_i and r_j, respectively. H is the height of the image.

3.4.1.2 Compactness

Compactness is the only new cue that we do not use in the cue combination method. It is known that the more regular regions have bigger compactness and compact regions are more likely to own the border (see Figure 3.11). Compactness cue of a region r is formulated as follows:

$$Comp(r) = \exp\left(\frac{L^2}{A}\right), \tag{3.23}$$

where L is the contour length of region r and A is the area of the region.

3.4.1.3 Curvature

Curvature cue between two regions r_i and r_j is defined as follows[2]

$$Curv(r_i, r_j) = \frac{1}{1 + \exp(-C(b)/l)}, \tag{3.24}$$

[2] Although there exists several definitions to the cue definitions, all of them are conceptually equivalent to each other.

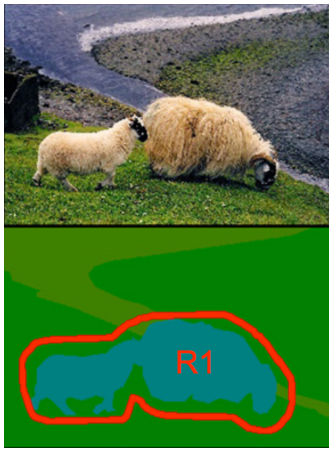


Figure 3.11: Example of a region which is more compact. Source: [7].

where

$$C(b) = \sum_{p \in b} (s(p) \rightarrow r) \cdot m(p), \tag{3.25}$$

where $C(b)$ is curvature of the border b that r_i and r_j share. l is the number of pixels which are located on b. $s(p) \rightarrow r$ and $m(p)$ are explained in detail in Equation 3.4.

3.4.1.4 Junctions

There are two types of junctions used namely, T-junction and L-junction.

As mentioned before, 3 regions and 3 angles forms T-junction. For two regions r_i and r_j, T-junction is formulated as follows:

$$TJ(r_i, r_j) = \frac{\theta_{r_i}}{\theta_{r_i} + \theta_{r_j} + \theta_{r_k}}, \tag{3.26}$$

where r_k is the third region that forms T-junction with other regions. θ_{r_i} is the angle facing region r_i.

Also, L-junction is formulated similarly in Equation 3.27:

$$LJ(r_i, r_j) = \frac{\theta_{r_i}}{\theta_{r_i} + \theta_{r_j}}. \tag{3.27}$$

3.4.1.5 Contrast

Contrast cue for a region r is defined as follows:

$$Cont(r) = \sum_{p \in r} \frac{1}{N} I(p), \tag{3.28}$$

where $I(p)$ is intensity value of pixel p and N is the number of pixels within r.

3.4.2 AdaBoost Classifier

Visual cues explained in the previous section are combined to create a feature vector. Those feature vectors are used to train an AdaBoost classifer, the algorithm explained in Section 2.6. Definition of feature vector for regions r_i and r_j that share a border is as follows:

$$FV(r_i, r_j) = \begin{bmatrix} LR(r_i, r_j) \\ Comp(r_i) \\ Comp(r_j) \\ Cont(r_i) \\ Cont(r_j) \\ Curv(r_i, r_j) \\ TJ(r_i, r_j) \\ LJ(r_i, r_j) \end{bmatrix}^T . \tag{3.29}$$

If r_i occludes r_j; in other words, r_i owns the border, $FV(r_i, r_j)$ is considered as positive sample. Otherwise, it is a negative sample. Unlike Chen et al. [7], we use contrast and L-junctions in our feature vector additionally. Also, our feature vector lacks of semantic cue which does not exist in our dataset.

After training the classifier with these positive and negative samples, feature vectors that we extract from unknown samples are classified in order to determine BO.

3.5 Conditional Random Field on Shapemes

Psychophysical studies have shown that familiar object boundary configurations are most likely recognized prior to figure/ground assignment and it is a powerful cue for

figure/ground segregation [46]. These findings are also consistent with the traditional Gestalt emphasis on being a global phenomenon of this process.

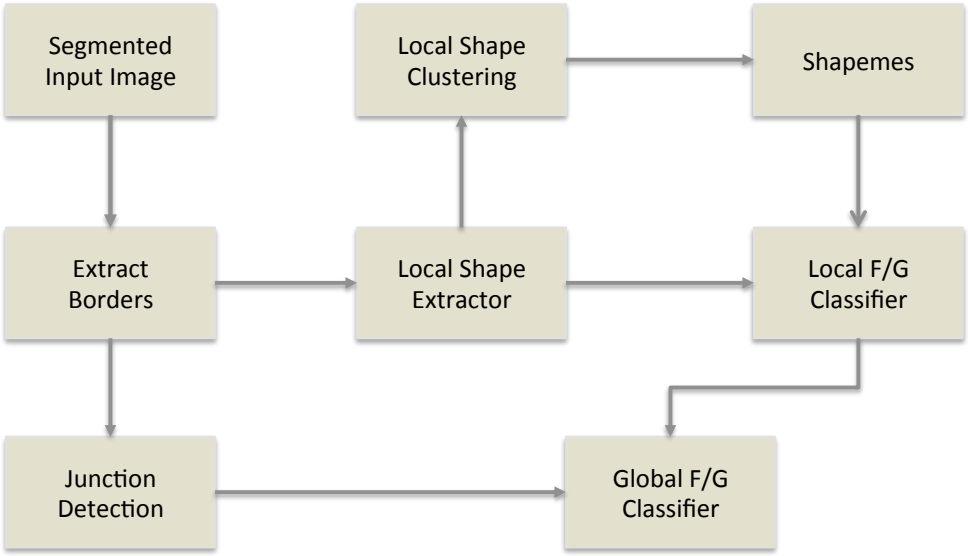


Figure 3.12: Overall architecture of CRF-S. Given an segmented input image, first border image is obtained. Then, local shapes are extracted and passed to the local classifier. On the other hand, after detecting junctions in the same border image, global classifier refines predictions of local classifier using junction consistencies.

Ren et al. [9] proposed an approach to figure-ground assignment problem which depends on not only local visual cues but also global configuration of the image constructs. This method consists of local classifier and a global model which enforce junction consistency on top of local predictions. As shown in Figure 3.12, given an input image with existing segmentation information and contour structure, the local model predicts figure/ground label at each image location, after which the global model that uses Conditional Random Field (see Section 2.7) corrects ambiguities using junction consistencies. In this section, both local and global models are explained.

3.5.1 Local Figure/Ground Model

Most of the border ownership studies employs mid-level visual cues as classical figure/ground cues. These mid-level cues have precise mathematical/geometric definitions; however, it is not trivial to extract them such that they strongly conform to their definitions. Instead of coming up with a mathematical definiton for every possible

local cues, local shapes are encoded using generic shape descriptor to achieve prototypical shapes, each of which encodes a local visual cue in empirical way. These prototypical shapes are also called Shapemes. Shapemes was first used in [47] to create an index of object specific shapes. In this model, it is used to capture mid-level cues with no need to define what these cues are.

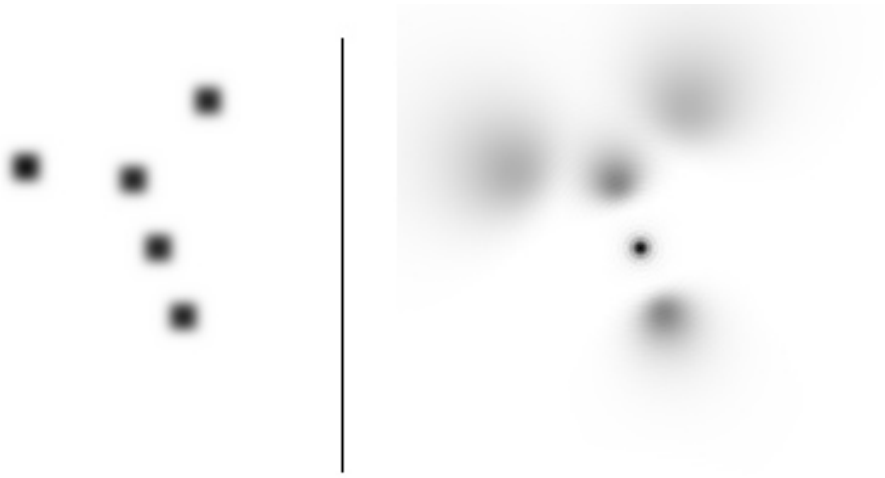


Figure 3.13: Image to the left is an example of Gaussian blur. Blur is applied uniformly to all pixels in the image. On the right hand side, geometric blur is applied to the same image. In geometric blur, pixels further away from the center become more blurred. Source: [8].

The generic shape descriptor used is Geometric Blur operator [48]. Given an input image I, and E its edge map. $GB_x(y)$ is inner product of E with a spatially varying Gaussian where x is the center of geometric blur and y is the location that geometric blur applies to. The standard deviation of the Gaussian is $\alpha|y-x|$, where $\alpha = 0.5$. In this study, blurred image GB_x is rotated such that contour orientation at x is always vertical. Then, the blurred and rotated image is sampled at 4 different radii (increasing with a factor of $\sqrt{2}$) and 12 orientations to form a feature vector of dimension 48 (see Figure 3.13).

After each local cue is represented using a Geometric blur descriptor, these vectors are clustered to find prototypical shapes, i.e. shapemes. Shapemes in this study are orientation independent since they are aligned to border orientations. Visualization of a sample set of prototypical shapes is shown in Figure 3.14.

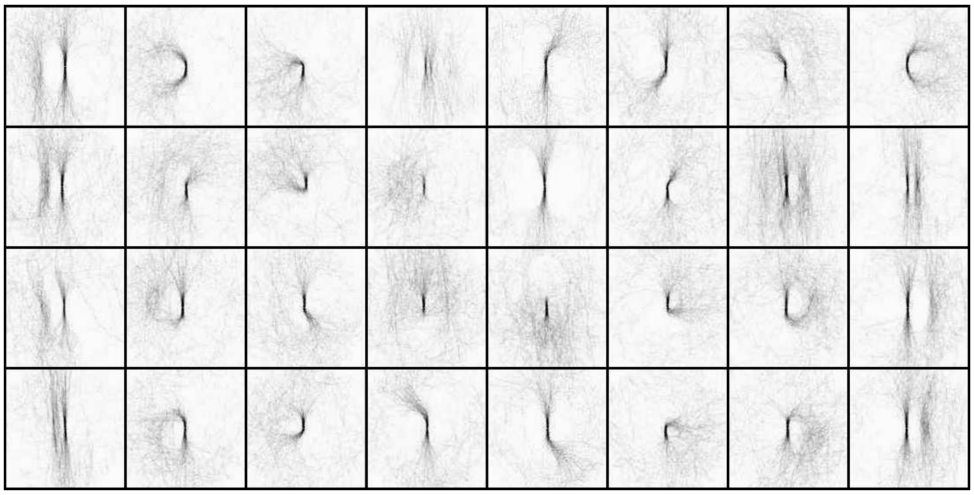


Figure 3.14: Shapemes from a set of boundary images. Each image in this figure shows the average shape in each cluster which encode a particular type of mid-level visual cue. Source: [9].

In this study, we use 64 shapemes constructed from human-labeled contour data, as Ren et al. [9] did in the original study. To represent local shapes Gaussian mixture model is used. A local shape is represented by a feature vector f of dimension 64, where each element in the vector is the log posterior probability of the corresponding component mixture. These features are used to classify which side of the border is figure. The classification is done via a logistic classifier which is fit to human-labeled ground-truth data using standard iteratively re-weighted least squares.

3.5.2 Global Figure/Ground Model

Border ownership studies have shown that local visual cues are highly informative for figure/ground assignment. However, global structure of the borders carries important information about the scene and the objects included in it. Especially, junctions are probably the most important and informative global information about contours in an image. This model combines local cues with the global information so as to obtain more consistent and accurate border ownership estimation. In this model, conditional random field is used to enforce global consistency on the local model's predictions.

For every border b in the image, the local model's estimation about the ownership is termed as X_b. $X_b = 1$ if the "left" side of b is figure, otherwise $X_b = -1$. Also, X_V

is the collection of variables, sorted in a clock-wise way, for all borders that join at a junction V. Probability distribution of the BO model is an exponential distribution as follows:

$$P(X|I,\Theta) = \frac{1}{Z(I,\Theta)}\exp\left\{\sum_{b}\Phi(X_b|I,\Theta) + \sum_{V}\Psi(X_V|I,\Theta)\right\} \tag{3.30}$$

where Φ is unary potential function on each border b, which is local model's contribution, and Ψ is smoothing function that corrects local estimate using global consistency at each junction V.

The border potential Φ using local model's estimate is defined as follows:

$$\Phi(X_b) = \beta X_b \log\left(\frac{p_b}{1-p_b}\right), \tag{3.31}$$

where p_b is the probability that the "left" side of b is figure. This is the average of the probabilities of all pixels on this border.

The smoothing term, i.e. junction potential, assigns a weight to each junction type. Let V is a junction which consists of k borders $b_1, b_2, ..., b_k$. $T(X_V)$, type of junction V, is determined as $T(X_V) = \{X_{b_1}, X_{b_2}, ..., X_{b_k}\}$. The junction potential Ψ is defined as follows:

$$\Psi(X_V) = \sum_{t\in T_a}\alpha_t \cdot \mathbf{1}_{\{T(X_V)=t\}} + \sum_{t\in T_c}\gamma\cdot\theta(X_V)\cdot\mathbf{1}_{\{T(X_V)=t\}}, \tag{3.32}$$

where T_a set of all possible junction types, and T_c is set of junctions where a continuity term θ is defined. In this study, continuity term is defined for T-junction and L-junction, and it is the angle between two borders facing the ground. In this study, all CRF computations are succeeded using UGM toolbox [49] that enable us training and using CRF model as described in this section.

3.6 Conditional Random Field on Visual Cues

In this study, we propose a method which combines predictive capabilities of visual cues analyzed in [1] with global configurations of junctions. As seen in Figure 3.15 this method is quite similar to the model explained in previous section; however, we replaced shapeme-based local classifier with combination of visual cues explained in

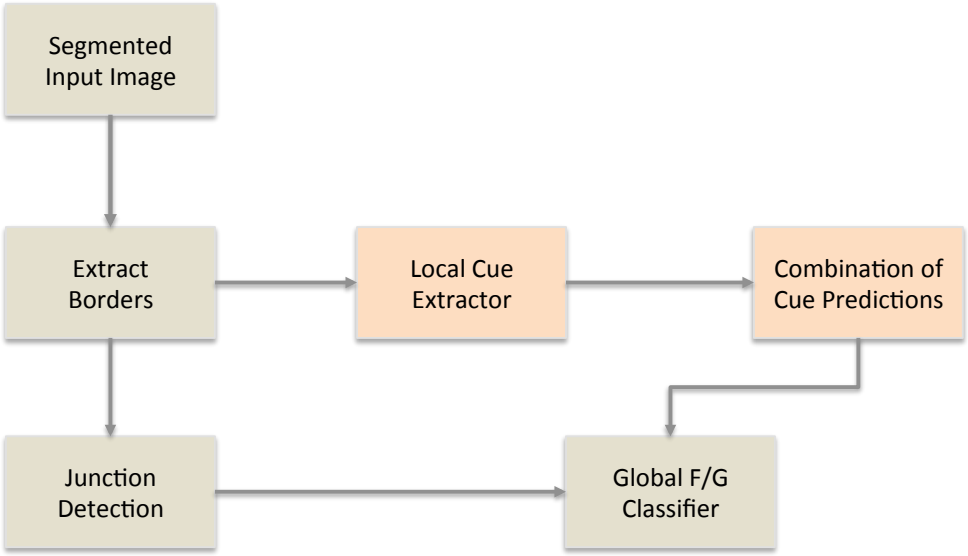


Figure 3.15: Overall architecture of CRF-VC. Given an segmented input image, first border image is obtained. Then, local cues are extracted and combined according to majority voting. On the other hand, after detecting junctions in the same border image, global classifier refines predictions of local classifier using junction consistencies.

Section 3.1 is used in junction potential function. The only changing part is probability function of the local classifier. The smoothing function is same as what is used in CRF-S. Therefore, probability function p_b becomes as follows:

$$p_b = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbf{1}_{\{\hat{r}_c(b) = r_{left}\}}. \tag{3.33}$$

Since junctions contribute to prediction as part of smoothing term, they are not included in visual cues used in this method. Therefore, the cues are as follows:

$$\mathcal{C} = \{contrast, curvature, lower\_region\}. \tag{3.34}$$

CHAPTER 4

EXPERIMENTS AND RESULTS

In this chapter the methods explained in the previous chapter are analyzed individually in depth. Moreover, they are compared to each other in terms of their accuracies and running times.

4.1 Dataset and Evaluation

For the experiments, we used the METU Border Ownership Dataset [34]. In the dataset, there are 1,000 images, 500 of which are indoor and 500 of which are outdoor images. The dataset is divided into training set and test set, each of which consists of 250 indoor and 250 outdoor images. Figure 4.1 shows example images from the dataset. There are 15,082 borders in total in the dataset, which are human labeled. In order to asses overall performance of each method, we repeat the experiment for 10 iterations. In each iteration, dataset is split into training and test set randomly. For the methods that require training, the training set is used. All the methods are tested using 500 images in the test set.

Performance of the methods are evaluated in terms of accuracy, which is formally defined as follow:

$$\text{Accuracy} = \frac{\text{Number of TP + TN}}{\text{Number of TP + TN + FP + FN}}, \tag{4.1}$$

where TP, TN, FP, FN are true positive, true negative, false positive, and false negative, respectively.

(a) An indoor image from the dataset

(b) Segmented version of the original image.

Figure 4.1: An example image from the METU Border Ownership Dataset.

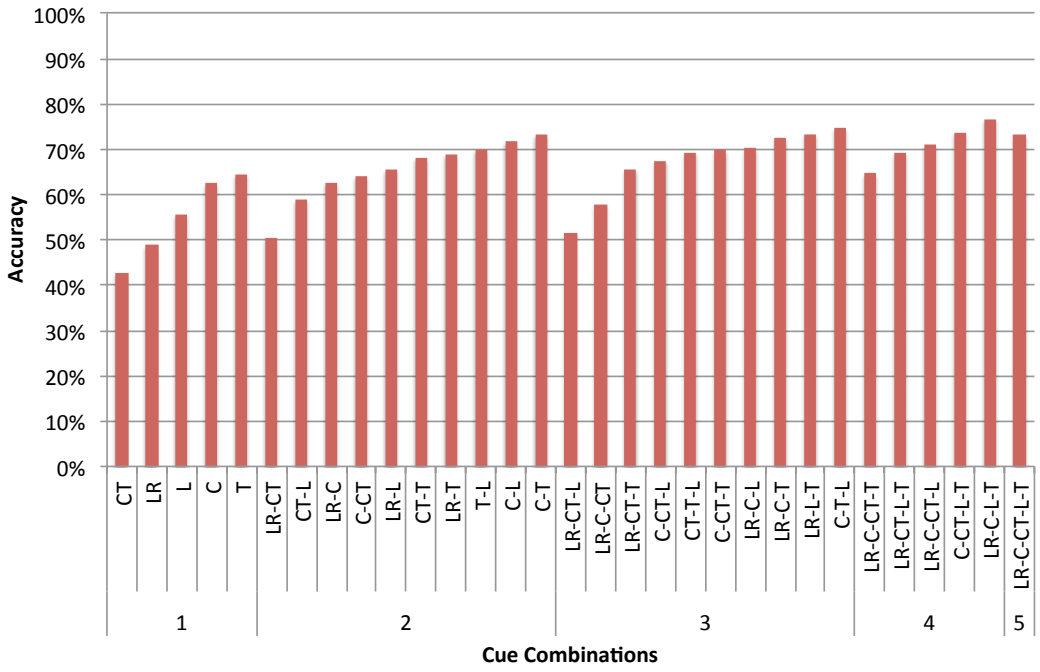
4.2 Individual Analysis of Border Ownership Methods

In this section, we individually analyzed all the methods explained in Chapter 3.

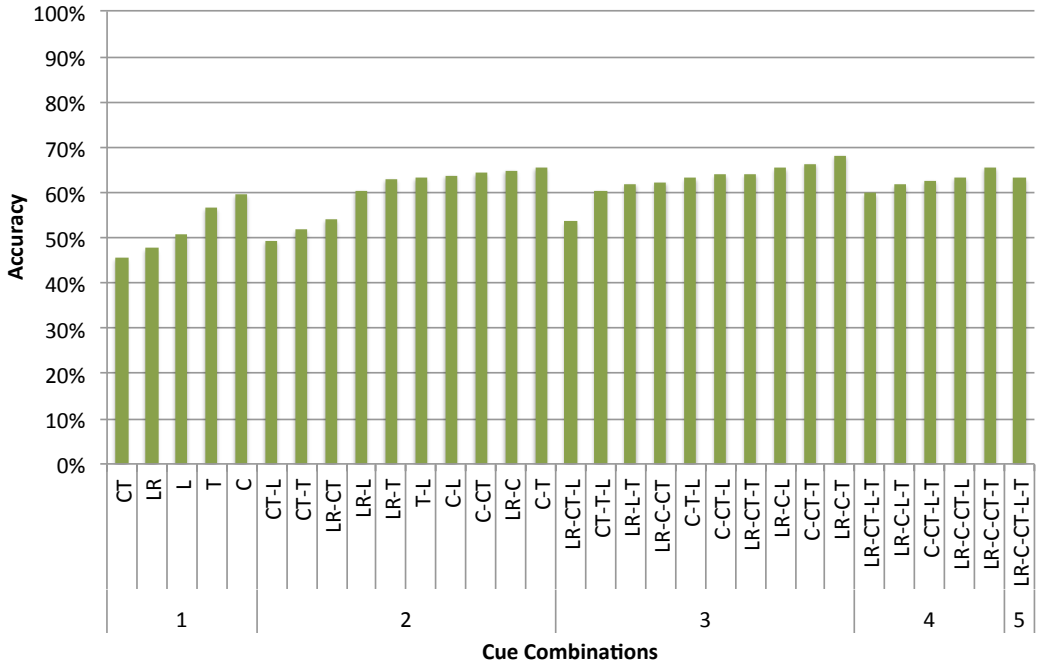
4.2.1 Analysis of Visual Cue Combination

In this experiment, we analyzed predictive capabilities of individual cues as well as whether it yields better predictions if we combine them on indoor and outdoor images. As mentioned in Section 3.1, there are five cues used; hence, there exists 31 combinations of these visual cues. Figure 4.2 shows prediction accuracy of these combinations on indoor (a) and outdoor (b) images.

As shown in Figure 4.2a, the best combination for indoor images consists of four cues namely, lower-region, curvature, T-junction, and L-junction with 76.9% accuracy. The best single cue is T-junction with 64.6% accuracy. Indoor images are composed of well-structured objects which form junctions, especially T-junctions. Therefore, they are quite available in indoor images and yield good accuracy. The best two-cue combination is T-junction and curvature with 73.1%. T-junctions perform better when it is combined with curvature which is also a powerful visual cue for border ownership (62.6% single cue accuracy). The winner three-cue combination is curvature, T-junctions, and L-junctions. It has almost the same accuracy with the best combination (74.6%). On the other hand, when all cues are combined together it does not have the best performance. This is discussed later in this section.



(a) Combined cue accuracy on indoor images



(b) Combined cue accuracy on outdoor images

Figure 4.2: BO determination accuracy with respect to combination of different cues on indoor and outdoor images.

Outdoor performance of cue combinations is shown in Figure 4.2b. Accuracy of border ownership estimation on outdoor images are significantly lower than the accuracy on indoor images. Outdoor scenes do not contain regular structure as much as indoor environments do; therefore, determination of BO is harder for outdoor images using visual cues. The best is combination of lower-region, curvature, and T-junction with 68.2% accuracy. Curvature is the best single cue for outdoor images. When we combine curvature with T-junction, this combination has the best accuracy among the two-cue combinations. The best four-cue combination has 65.5% accuracy. This combination is lower-region, curvature, contrast, and T-junction. Although it is superset of the best combination its accuracy is worse. Besides, combination of all cues is even worse. This is observed in indoor images as well.

There are 31 possible combinations of these five cues. A four-cue combination works best for indoor images; however, three-cue combination is the winner for outdoor images. Overall, the best is the combination of lower-region, curvature, and T-junction with 70.6% accuracy in overall (72.5% indoor and 68.2%). Adding more cues does not mean they are supposed to perform better. Some cues in the combination conflict with each other; for that reason, they decrease the accuracy.

4.2.2 Analysis of Figure-ground Classification Based on Spectral Properties

This method extracts feature vectors from the input image, which describes spectral properties of boundary image patches. The key parameter of this method is the size of the boundary patches. In the original study, they have chosen patch size as 36 x 36. In this study, we varied the window size in order to analyze its impact on overall performance of the model.

Figure 4.3 shows model accuracy with respect to window size for both indoor and outdoor images. This method performs better with indoor images as visual cue combination does. The best accuracy is 73.2% for indoor and 66.8% for outdoor with window size 36. For larger window sizes performance does not change drastically, it performs with almost same accuracy. However, when we shrink the size of boundary patches, accuracy decreases. It drops to 54.7% for indoor and 52.1% for outdoor when window size is 12 pixel. If window size is small, the boundary patch mostly

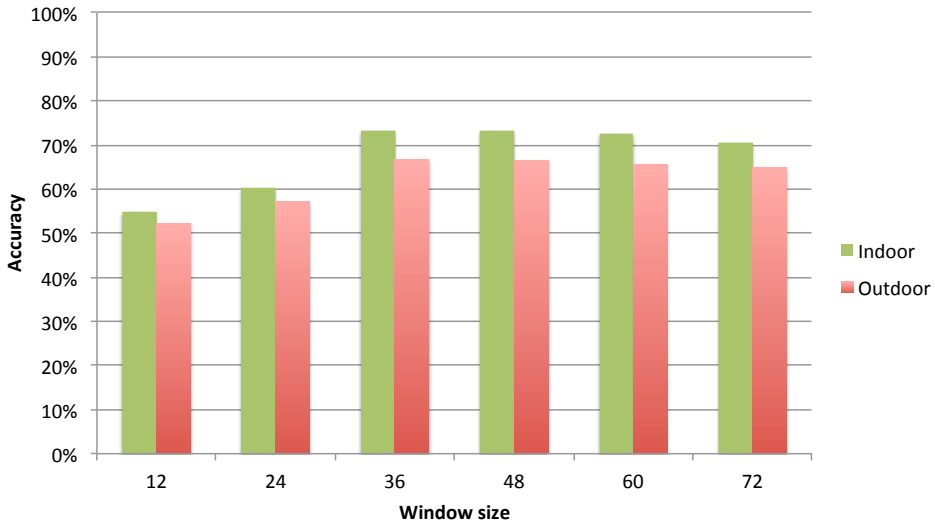


Figure 4.3: Accuuracy of SVM-SF with respect to varying window size. The method performs best when the window size is 36-pixel.

consists of pixels that belongs to border. Therefore, oriented energy cannot be computed from pixels of regions that share the boundary. For those patches, the feature vectors extracted are not conclusive for determining the owner region. Therefore, image patches should be big enough to cover boundary regions' pixel so that it can determine which region owns the border.

The SVM model is trained on the training which consists of 500 images. The 10-fold cross validation score on the training set is 88.3%.

4.2.3 Analysis of Contrast Surround Modulation

CSM method, described in Section 3.3, uses Gabor filters in multiple directions to detect contrast change in each direction. This method uses receptive field that spans whole image with a matrix of model-cells. Response of each model cell at a certain location is obtained through multi-stage calculations from an image patch whose center is the location of the cell.

Window size selection and orientations used in Gabor filters are main factors that could affect performance of the model. Therefore, we analyzed the impact of these factors on accuracy of the model. In addition, the strength of surround modulation

is weighted using a coefficient called k_{nat} (see Equation 3.20). We analyzed how surround modulation affects the model performance in our experiments.

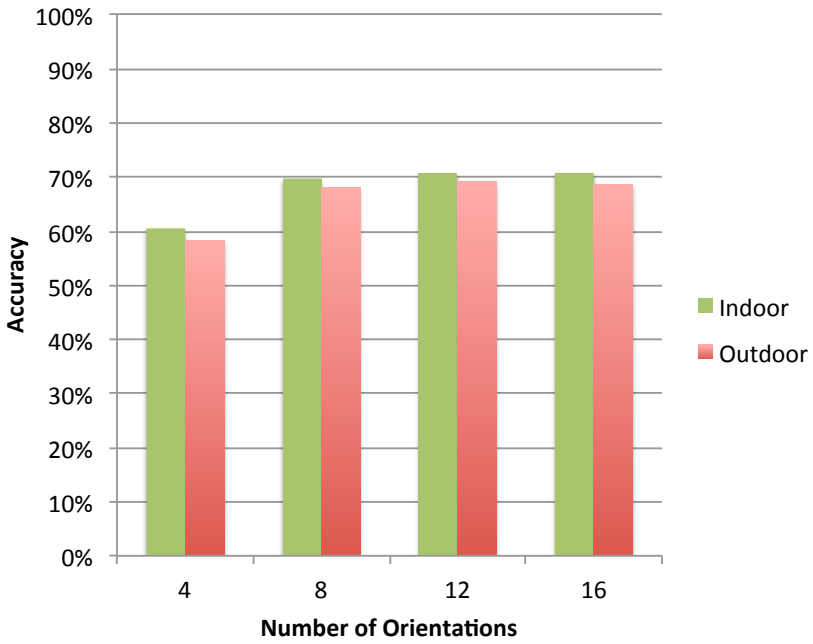


Figure 4.4: Impact of number of Gabor filter orientations on accuracy of CSM method.

In the original study, Sakai et al. [6] extended their previous model increasing the number of orientations so as to capture contrast changes in natural images. In our experiments, we analyzed their alteration on Gabor filters. We implemented both approach, Gabor filters with four orientations and eight orientations, then tested these two implementations with the same set of images. Moreover, we analyzed the performance of the extended model for increasing number of orientation. For this analysis, we select window size and k_{nat} are selected as 24 X 24 and 1.0, respectively. Figure 4.4 shows BO determination accuracy with respect the number of orientations used in Gabor filters. Gabor filters with eight orientations have 69.7% accuracy for indoor and 68.1% accuracy for outdoor whereas it is 60.4% and 58.2% with four orientations for indoor and outdoor, respectively. Increasing number of orientations to 12 and 16 produces slightly better results (12 orientation: 70.6% indoor, 69.1% outdoor and 16 orientation: 70.7% indoor, 68.6% outdoor). As natural images are composed of complex geometric structures contrast changes might occur in any directions. For that reason, Gabor filter with four different orientation cannot detect the contrast transi-

tion in some cases. Extending model with Gabor filters using eight or more directions increases accuracy of the model for natural images.

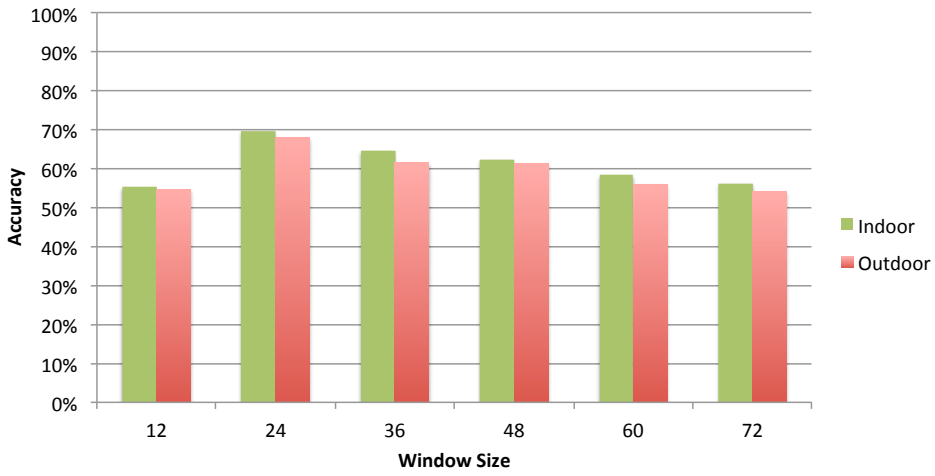


Figure 4.5: The size of the subregion that model cell compute its response within affects the performance CSM.

As mentioned, response of each cell is computed from pixels that a window covers. We analyzed performance with respect to window size in our experiments. For this anaylsis, we use Gabor filtering in 12 orientations. Also, k_{nat}, strength of surround modulation, is selected as 1.0. As shown in Figure 4.5, accuracy varies as window size changes. The model performs best with 69.7% accuracy for indoor and 68.1% accuracy for outdoor when the window size is 24. For window sizes larger than 24 pixel, there is a decrease in performance for larger sizes. Larger subregion or window could include pixels from some regions which do not share the border in question. Therefore, pixels residing in these regions would be irrelevant and mislead the determination.

In addition to window size and Gabor orientations, we also analyzed how surround modulation affects the BO determination. In the original study, strength of surround modulations is controlled through a coefficient named k_{nat}. It is chosen as 0.0125 for artificial images and 1.0 for natural images in [5]. We varied k_{nat} from 0.0 to 1.2. Figure 4.6 shows the change in accuracy with respect to the strength of surround modulation. It is obvious that surround modulation has important contribution to determination of BO as figure shows. When there is no surround modulation, i.e., $k_{nat} = 0$, the accuracy is 48.7% in overall. Accuracy increased as surround modula-

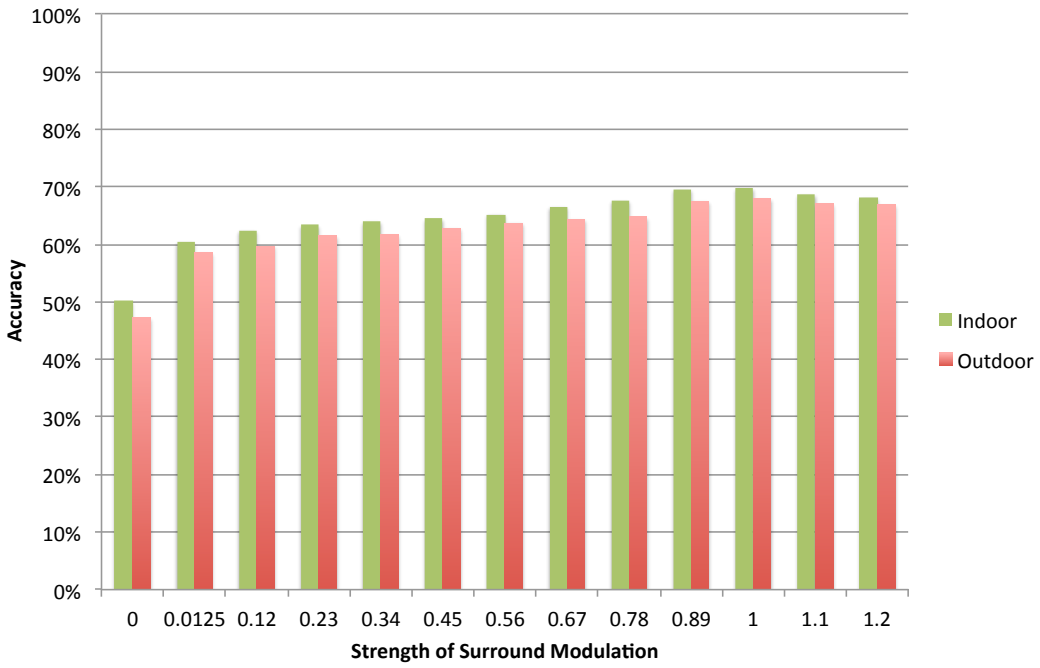


Figure 4.6: Surround modulations has weighted contribution to response of the model cell. CSM method has the best performance with full contribution of surround modulation.

tion is taken into account. However, for k_{nat} values larger than 1.0, the accuracy starts decreasing. It has the best performance when k_{nat} is equal to 1.0. Without surround modulation performance is not satisfactory since it seems that BO is determined by chance.

Overall, the model performs best with Gabor filter using 12 orientations, 24-pixel window, and full contribution of surround modulation (70.6% indoor and 69.1% outdoor).

4.2.4 Analysis of AdaBoost on Visual Cues

AdaBoost on Visual Cues (AB-VC) is a simple method that combines weak classifiers, visual cue predictors. It finds the best weight combination that maximizes accuracy for weak classifiers. We tested the AdaBoost model on both indoor and outdoor images. As shown in Table 4.1, the model performs with 70.8% accuracy for indoor and 68.3% for outdoor images. The overall accuracy of the model is 69.6%.

Table 4.1: Accuracy of AB-VC method.

	Accuracy (%)
Indoor	70.8
Outdoor	68.3
Overall	69.6

4.2.5 Analysis of Conditional Random Field on Shapemes

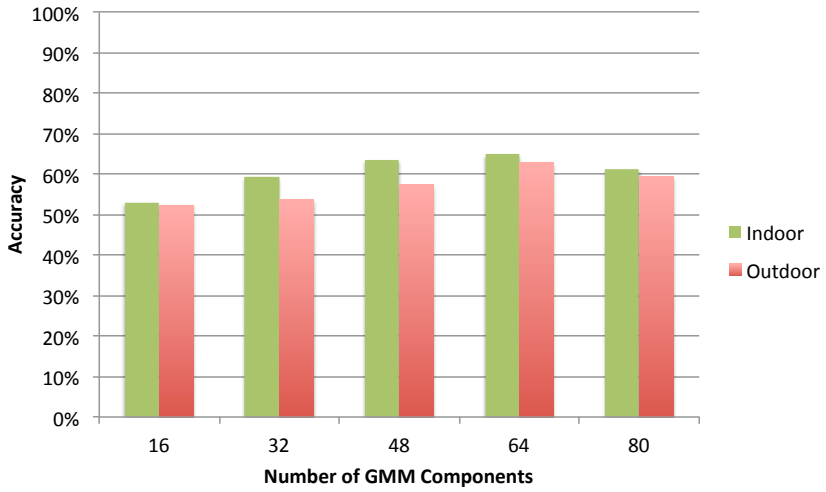


Figure 4.7: Performance of the local classifier with respect to number of GMM mixtures used in the model.

We explained the method that use CRF on top of Shapeme based local classifier in Section 3.5. This model consists of local classifer and global classifier. Local classifier determines which region owns the border from shape descriptor that encodes geometric characteristic of the boundary neighborhood. Local classifier uses set of prototypical shapes to identify frequently observed geometric properties. Number of prototypical shapes corresponds to number of components that local classifier uses in GMM. In our experiments, we analyzed performance of the local classifier with respect to number of components used in GMM. Changing the number of components affects descriptive capability of the local classifier; in other words, how many different local shapes can be described through GMM components. Figure 4.7 shows accuracy of the local classifier with respect the number of GMM components. Increasing number of components results in better accuracy for both indoor and outdoor images. However, increase in accuracy is steeper for outdoor images than it is

for indoor images. Since outdoor images lack of regular structures and usually include complex shapes, descriptive capability of the local classifier has important in figure-ground classification. It performs best with 63.8% accuracy (62.8% indoor and 64.9% outdoor) when the number of GMM components is selected as 64.

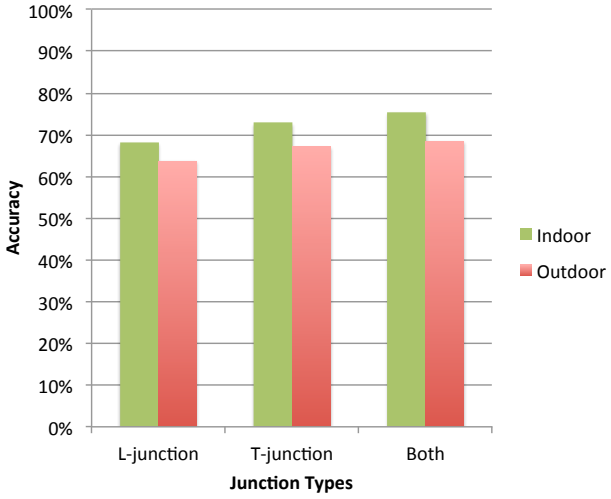


Figure 4.8: Contribution of the junction types to accuracy of the global model. CRF-S performance increases when both junctions are used to enforce global consistency.

On the other hand, the global CRF model uses junction to enforce global consistency in figure-ground assignment. It uses the global information to correct predictions of the local model when there is a conflict. We examined contribution of junction types to prediction accuracy in our experiments. As it is shown in Figure 4.8, when both T-junction and L-junction are used to control consistency, the model performs the best with 71.9% accuracy (75.4% indoor and 68.4% outdoor).

4.2.6 Analysis of Conditional Random Field on Visual Cues

Another method is to combine global power of CRF with predictive capabilities of local visual cues. In this method, we use the global information that T-junction and L-junction provide on CRF model instead of combining junction information as another local cue. This method does local classification based on combination of visual cues namely, lower-region, curvature, and contrast. Then, CRF model corrects their prediction according to global consistency that junctions provide. Figure 4.9 shows

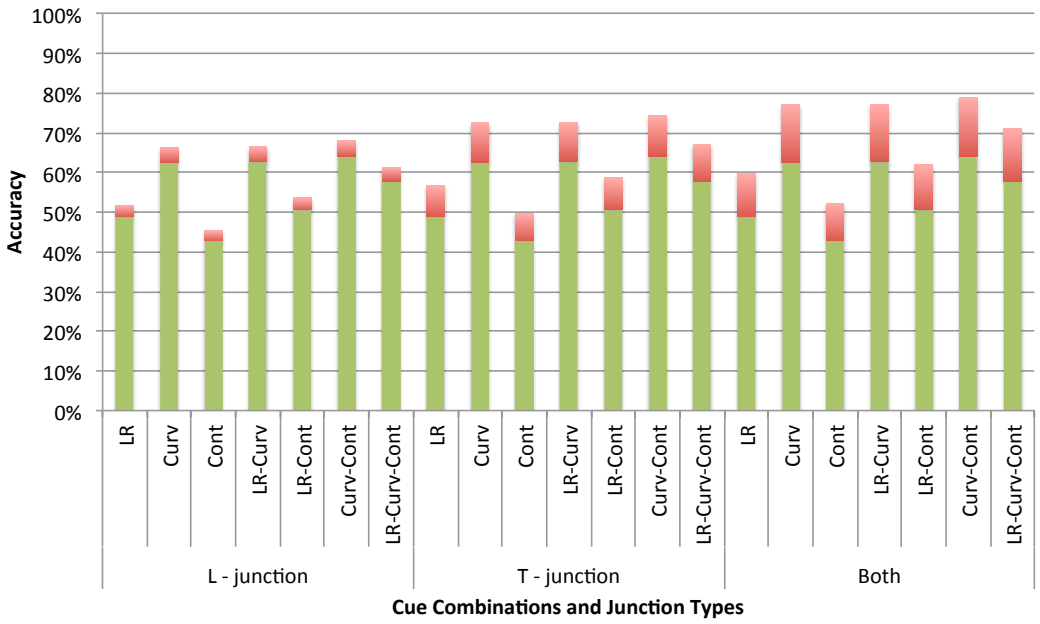
accuracy of cue combination with CRF for both indoor (a) and outdoor (b) images. In these figures, accuracy of the model is presented for seven possible combination of three cues for three cases where which junction(s) are used in CRF. These are only L-junction, only T-junction and both T-junction and L-junction.

For indoor images (see Figure 4.9a), best combination is curvature and contrast when CRF uses both of the junctions to correct their predictions. Accuracy of the model for this combination is observed up to 78.9%. For outdoor images (see Figure 4.9b), lower-region and curvature are the best combination with both junctions. Best accuracy achieved for outdoor images is 70.7%. Regarding the junction types, CRF smoothing works best when T-junction and L-junction are used together. Improvement done by global cues are higher in indoor images since junctions are more frequently exist in indoor images than outdoor images.

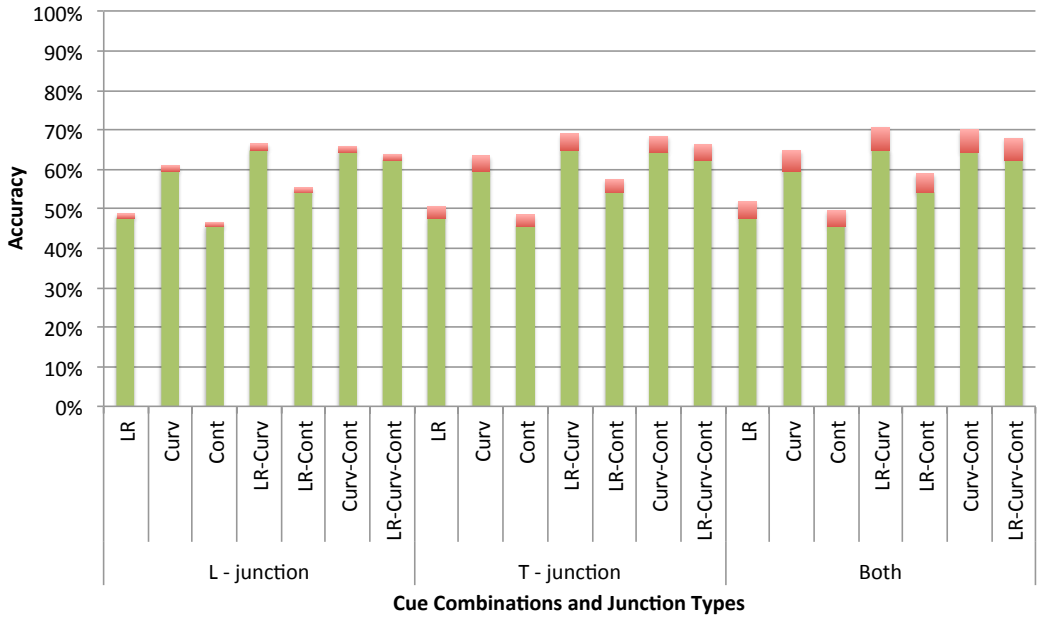
4.3 Overall Comparison

In this section the methods are compared to each other and their performances are analyzed for indoor and outdoor images. As mentioned in Section 4.1, we shuffle and split dataset into two subsets (training set and test set) for 10 iterations. The results reported in this section are averages of those 10 iterations.

Table 4.2 shows average accuracy and standard deviation of each method. Also accuracies for indoor and outdoor images are illustrated in Figure 4.10. The best method is CRF-VC. It has 74.2% accuracy (76.6% indoor, 71.7% outdoor). The second method is CRF-S with 72.1% accuracy (74.4% indoor, 69.8% outdoor). SVM-SF has accuracy 68.2% accuracy in overall, which is 71.3% for indoor and 65.2% for outdoor images. Although its accuracy is not the best, it is a simple and effective method for determining BO. Although VCC method has good prediction accuracy, it is interesting that it has the best accuracy when we combine some of the visual cues, not all of them. Having more cue does not mean it performs better if the majority voting is used. On the other hand, when we combine 6 different visual cues through AdaBoost algorithm, we reach to 69.6% accuracy (70.8% indoor, 68.3% outdoor). AB-VC performs better than VCC if all local cues are combined; however, VCC reaches to better



(a) CRF-VC on indoor images



(b) CRF-VC on outdoor images

Figure 4.9: CRF-VC performance on indoor and outdoor images. Figures show accuracy of model for all combinations of visual cues for each junction type used in global model. Green bars show the accuracy of cue combination, red bars stacked on greens show improvement that CRF provides.

accuracy (70.2% in overall) compared to AB-VC if fewer number of cues are combined.

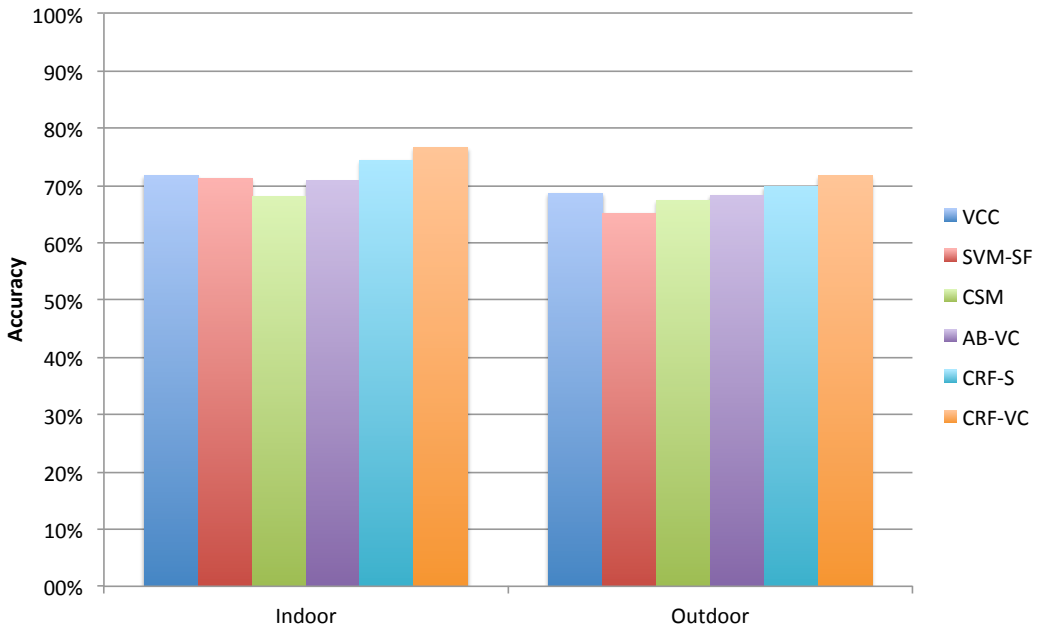


Figure 4.10: Comparison of all methods for indoor and outdoor images as well as average performance of the methods.

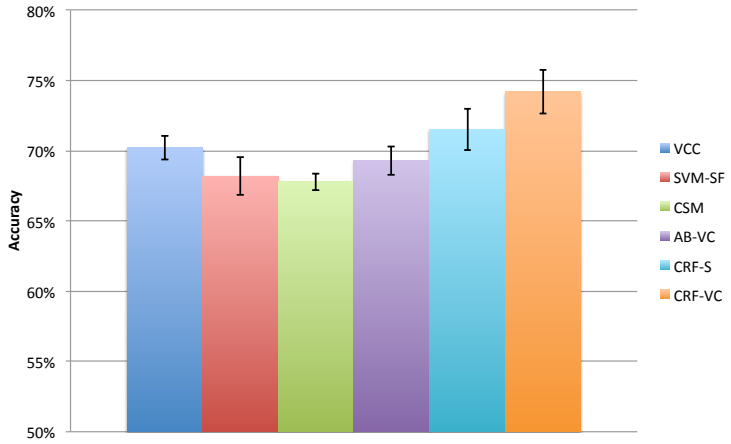


Figure 4.11: Overall accuracies of the methods and their standard deviations.

As shown in Figure 4.10, accuracies of all methods change depending on whether the image is outdoor or indoor. These methods are heavily incorporate visual cues and their organization in the image; therefore, their performance decrease if there are complex objects or regions exist in the image.

Table 4.2: Average accuracies of all methods and their standard deviations.

	Accuracy (%)	Standard Deviation
VCC	70.2	1.2
SVM-SF	68.2	2.0
CSM	67.8	0.7
AB-VC	69.6	1.6
CRF-S	72.1	2.1
CRF-VC	**74.2**	**2.2**

These findings show that enforcing global consistency may be helpful as the methods that use CRF have better accuracies compared to others. Moreover, accuracy increases when we take junctions out of cue combinations and used them to correct local predictions through CRF. CRF-VC method has better accuracy than the pure cue combination. In VCC method, junctions are just a visual cue. The method does not make benefit of their global characteristics. Therefore, even the junctions could have corrected the prediction if majority votes against them they are just discarded. However, CRF handles this global information gracefully and corrects the local predictions accordingly.

Table 4.3 compares the accuracies of the methods with their reported accuracies in the literature. The reported accuracy for VCC is 74.0% [1]. In our experiments, it performed with 70.2% accuracy. The accuracy of SVM-SF is almost equal to what Ramenahalli et al. [5] have reported. CSM has shown better performance in our experiments compared to the results of the original study [6]. Chen et al. [7] achieved 92.0% accuracy with their AdaBoost method; however, their "semantic cue" has significant role in the performance (Semantic cue feature has 91.0% accuracy). Our AB-VC method has 69.6% accuracy without semantic cue. Ren et al. [9] reported that their CRF based classifier performed with 78.0% accuracy. On the other hand, our shapeme based CRF classifier has 74.2% accuracy in overall.

4.4 Running Time Analysis

In this section, the methods are compared in terms of their running time performance. All methods are run on the same machine. 100 images, 50 indoor and 50 outdoor, are

Table 4.3: Accuracies of all methods in our study along with the reported accuracies in the literature. Note that the original studies were tested on different images.

	Our Study (%)	Original Study (%)
VCC	70.2	74.0
SVM-SF	68.2	68.1
CSM	67.8	64.7
AB-VC	69.6	92.0
CRF-S	72.1	78.0
CRF-VC	74.2	-

used to measure running time.

Table 4.4: Running times of all methods per border.

Method	Time (sec)
VCC	**2.2**
SVM-SF	4.6
CSM	6.1
AB-VC	2.4
CRF-S	4.5
CRF-VC	3.3

Table 4.4 shows the average running time of BO determination per border for each method. VCC and AB-VC methods are the fastest among the methods since they utilize only low-level simple visual cues whose extraction process does not require intensive computational power. The second is CRF-VC with 3.3 seconds running time. It is followed by CRF-S and SVM-SF with 4.5 seconds and 4.6 seconds, respectively. CSM is the slowest method with 6.1 seconds. The reason for CSM's slowness is due to the fact that its determination process consists of multiple stages which performs more complex operations than the other methods.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this thesis, methods widely used for border ownership estimation are analyzed on a comprehensive dataset which consists of indoor and outdoor images. The evaluated methods are naive visual cue combination, SVM classification based on spectral properties, contrast surround modulation, visual cue combination through adaptive boosting, CRF on shapeme based local classifier, and CRF on naive visual cue combination.

From the experiments and the results, the following conclusions have been drawn:

- It is observed that BO estimation becomes more challenging when regions or objects have complex structures. Accuracies of the methods are lower for outdoor images compared to indoor images.

- Local visual cues bare important information regarding to BO; however, they could conflict with each other for some cases. Adaptive boosting has a strategy to resolve these conflicts. It tweaks weight of each predictive cue according to the training data. Therefore, weighted combination of local cues yields better results compared to naive cue combination.

- CRF based methods take context into account by employing consistency that junctions imply. The accuracy of local classifier is improved by correcting local predictions enforcing global consistency on junctions. The results show that global information has an important role in the estimation process.

5.1 Future Work

Although the dataset used in this study is more comprehensive than its alternatives, lack of semantic labelling information prevents us from analyzing the role of such top-down information in the BO estimation process. Therefore, this study could be proceeded further with the future work:

- The dataset could be extended with semantic information that human subjects could provide.
- BO estimation methods that utilize top-down information such as semantic label could be analyzed and compared with the methods in this study.
- Although AdaBoost is a useful ensemble learning algorithm for BO determination, other classifier ensemble methods such as Random Forests [50] or Rotation Forest [51] could be employed in classification.

REFERENCES

[1] Mehmet Akif Akkus. Analysis of border ownership cues and improvement of depth prediction using border ownership. Master's thesis, Middle East Technical University, 2014.

[2] Twain Taylor. How to use the gestalt principles for visual storytelling. http://blog.fusioncharts.com/2014/03/how-to-use-the-gestalt-principles-for-visual-storytelling-podv/, 2015. Last visited date: August 15, 2015.

[3] Shaun P. Vecera, Edward K. Vogel, and Geoffrey F. Woodman. Lower region: A new cue for figure-ground assignment. *Journal of Experimental Psychology: General*, 131:194–205, 2002.

[4] Wikipedia. Support vector machine. http://en.wikipedia.org/wiki/Support_vector_machine, 2015. Last visited date: August 7, 2015.

[5] Sudarshan Ramenahalli, Stefan Mihalas, and Ernst Niebur. Figure-ground classification based on spectral properties of boundary image patches. In *46th Annual Conference on Information Sciences and Systems, CISS 2012, Princeton, NJ, USA, March 21-23, 2012*, pages 1–4, 2012.

[6] 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.

[7] Xiaowu Chen, Qing Li, Dongyue Zhao, and Qinping Zhao. Occlusion cues for image scene layering. *Computer Vision and Image Understanding*, 117(1):42 – 55, 2013.

[8] Alex Berg. Geometric blur. http://www.acberg.com/gb.html, 2015. Last visited date: August 7, 2015.

[9] Xiaofeng Ren, Charless C. Fowlkes, and Jitendra Malik. Figure/ground assignment in natural images. In *ECCV*, pages II: 614–627, 2006.

[10] Heiko Neumann, Arash Yazdanbakhsh, and Ennio Mingolla. Seeing surfaces: The brain's vision of the world. *Physics of Life Reviews*, 4(3):189 – 222, 2007.

[11] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. (Lond.)*, 195(1):215–243, Mar 1968.

[12] Barton L. Anderson, Manish Singh, and Roland W. Fleming. The interpolation of object and surface structure. *Cognitive Psychology*, 44(2):148 – 190, 2002.

[13] H.G. Barrow and J.M. Tenenbaum. Interpreting line drawings as three-dimensional surfaces. *Artificial Intelligence*, 17(1–3):75 – 116, 1981.

[14] T. S. Collett. Extrapolating and interpolating surfaces in depth. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 224(1234):pp. 43–56, 1985.

[15] W. E. L. Grimson. A computational theory of visual surface interpolation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 298(1092):pp. 395–427, 1982.

[16] B. Julesz. *Foundations of Cyclopean Perception*. University of Chicago Press, 1971.

[17] H. Komatsu. The neural mechanisms of perceptual filling-in. *Nat. Rev. Neurosci.*, 7(3):220–231, Mar 2006.

[18] L. Pessoa, E. Thompson, and A. Noe. Finding out about filling-in: a guide to perceptual completion for visual science and the philosophy of perception. *Behav Brain Sci*, 21(6):723–748, Dec 1998.

[19] D. Terzopoulos. The computation of visible-surface representations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 10(4):417–438, Jul 1988.

[20] Stefan Treue, Richard A. Andersen, Hiroshi Ando, and Ellen C. Hildreth. Structure-from-motion: Perceptual evidence for surface interpolation. *Vision Research*, 35(1):139 – 148, 1995.

[21] K. Nakayama, Z. J. He, and S. Shimojo. *Visual surface representation: a critical link between lower-level and higher level vision*, pages 1–70. M.I.T. Press, 1995.

[22] Max Wertheimer. Untersuchungen zur lehre von der gestalt. ii. *Psychologische Forschung*, 4(1):301–350, 1923.

[23] V.A.F. Lamme. The neurophysiology of figure-ground segregation in primary visual cortex. *Journal of Neuroscience*, 15(2):1605–1615, 1995. cited By (since 1996) 391.

[24] R. Von Der Heydt H. Zhou, H. Friedman. Coding of border ownership in monkey visual cortex. *Image and Vision Computing*, 20(17):6594–6611, 2000.

[25] Oliver W. Layton, Ennio Mingolla, and Arash Yazdanbakhsh. Dynamic coding of border-ownership in visual cortex. *Journal of Vision*, 12(13):8, 2012.

[26] Fangtu T. Qiu and Rüdiger Von Der Heydt. Figure and ground in the visual cortex: v2 combines stereoscopic cues with gestalt rules. *Neuron*, pages 155–166, 2005.

[27] J. Hegde and D. C. Van Essen. A comparative study of shape representation in macaque visual areas v2 and v4. *Cereb. Cortex*, 17(5):1100–1116, May 2007.

[28] K. Koffka. *Principles of Gestalt psychology*. Harcourt, New York, 1935.

[29] Ko Sakai and Haruka Nishimura. Surrounding suppression and facilitation in the determination of border ownership. *Journal of Cognitive Neuroscience*, 18(4):562–579, 2015/06/17 2006.

[30] 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.

[31] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[32] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997.

[33] Sanjiv Kumar and Martial Hebert. Discriminative fields for modeling spatial dependencies in natural images. In *In NIPS*. MIT Press, 2003.

[34] Metu border ownership dataset. http://www.kovan.ceng.metu.edu.tr/bo/, 2015. Last visited date: August 7, 2015.

[35] M. A. Peterson and E. Salvagio. Inhibitory competition in figure-ground perception: context and convexity. *J Vis*, 8(16):1–13, 2008.

[36] Ido Leichter and Michael Lindenbaum. Boundary ownership by lifting to 2.1d. In *ICCV*, pages 9–16. IEEE, 2009.

[37] Jay Hegde and David C. Van Essen. A comparative study of shape representation in macaque visual areas v2 and v4. *Cereb. Cortex*, 17(5):1100–1116, 2007.

[38] Masayuki Kikuchi and Kunihiko Fukushima. Assignment of figural side to contours based on symmetry, parallelism, and convexity. In Vasile Palade, Robert J. Howlett, and Lakhmi C. Jain, editors, *KES*, volume 2774 of *Lecture Notes in Computer Science*, pages 123–130. Springer, 2003.

[39] Masayuki Kikuchi and Youhei Akashi. A model of border-ownership coding in early vision. In Georg Dorffner, Horst Bischof, and Kurt Hornik, editors, *ICANN*, volume 2130 of *Lecture Notes in Computer Science*, pages 1069–1074. Springer, 2001.

[40] Derek Hoiem, Andrew N. Stein, Alexei A. Efros, and Martial Hebert. Recovering occlusion boundaries from a single image. In *ICCV*, pages 1–8. IEEE, 2007.

[41] David Fitzpatrick. Seeing beyond the receptive field in primary visual cortex. *Current Opinion in Neurobiology*, 10(4):438 – 443, 2000.

[42] H. Supèr, H. Spekreijse, and V. A. F. Lamme. Two distinct modes of sensory processing observed in monkey primary visual cortex (V1). *Nature neuroscience*, 4(3):304–310, March 2001.

[43] D. K. Xiao, S. Raiguel, V. Marcar, J. Koenderink, and G. A. Orban. Spatial heterogeneity of inhibitory surrounds in the middle temporal visual area. *Proceedings of the National Academy of Sciences of the United States of America*, 92(24):11303–11306, November 1995.

[44] Karl Zipser, Victor A. F. Lamme, and Peter H. Schiller. Contextual modulation in primary visual cortex. *The Journal of Neuroscience*, 16(22):7376–7389, 1996.

[45] K. I. Naka and W. A. H. Rushton. S-potentials from colour units in the retina of fish (cyprinidae). *The Journal of Physiology*, 185(3):536–555, 1966.

[46] M.A. Peterson and B.S. Gibson. Must figure-ground organization precede object recognition? an assumption in peril. *Psychological Science*, 5:253–259, 1994. cited By (since 1996) 102.

[47] G. Mori, S. Belongie, and J. Malik. Shape contexts enable efficient retrieval of similar shapes. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–723–I–730 vol.1, 2001.

[48] Alexander C. Berg and Jitendra Malik. Geometric blur for template matching, 2001.

[49] Mark Schmidt. Ugm: A matlab toolbox for probabilistic undirected graphical models. http://www.cs.ubc.ca/ schmidtm/Software/UGM.html, 2007. Last visited date: August 7, 2015.

[50] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[51] J.J. Rodriguez, L.I. Kuncheva, and C.J. Alonso. Rotation forest: A new classifier ensemble method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1619–1630, Oct 2006.