

ESTIMATING BORDER OWNERSHIP USING ITERATIVE VECTOR
VOTING AND CONDITIONAL RANDOM FIELDS

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

BY

BUĞRA ÖZKAN

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

FEBRUARY 2014

Approval of the thesis:

**ESTIMATING BORDER OWNERSHIP USING ITERATIVE
VECTOR VOTING AND CONDITIONAL RANDOM FIELDS**

submitted by **BUĞRA ÖZKAN** 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

Prof. Dr. Fatoş T. Yarman Vural
Co-supervisor, **Computer Engineering Department,** _____
METU

Examining Committee Members:

Prof. Dr. Fatoş Tünay 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 _____

Dr. Ayşenur Birtürk
Computer Engineering Department, METU _____

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: BUĞRA ÖZKAN

Signature :

ABSTRACT

ESTIMATING BORDER OWNERSHIP USING ITERATIVE VECTOR VOTING AND CONDITIONAL RANDOM FIELDS

Özkan, Buğra

M.S., Department of Computer Engineering

Supervisor : Assist. Prof. Dr. Sinan Kalkan

Co-Supervisor : Prof. Dr. Fatoş T. Yarman Vural

February 2014, 70 pages

Border ownership is the information that signifies which side of a border owns the border. Estimating this information has recently become very popular for perceptual organization as it allows rectification of ambiguous visual information. It is applied on many computer vision problems such as object detection, depth perception and optical flow. In this thesis, two different approaches are followed to solve the border ownership problem. For the supervised approach, conditional random fields are used as it is the most appropriate method for modelling contextual relations between semantic classes. Tensor voting is the inspire of our second algorithm called Iterative Vector Voting, as it allows modelling different information sources and their interactions. It is an unsupervised voting framework, which is proper for the use of Gestalt visual cues. Experiments show that both two models show significant contribution to the border ownership problem with respect to the successful results gathered on our own large-scale dataset.

Keywords: Border Ownership, Figure-Ground Segregation, Conditional Random Field, Graphical Models, Tensor Voting

ÖZ

YİNELEMELİ VEKTÖR OYLAMA VE KOŞULLU RASTGELE ALAN KULLANARAK SINIR SAHİPLİĞİ BİLGİSİNİN ELDE EDİLMESİ

Özkan, Buğra

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi : Yrd. Doç. Dr. Sinan Kalkan

Ortak Tez Yöneticisi : Prof. Dr. Fatoş T. Yarman Vural

Şubat 2014 , 70 sayfa

Sınır sahipliği, görüntü üzerindeki kenarların sahibi olan alanları belirlemekte kullanılan görsel bir bilgidir. Bu bilginin hesaplanması, muğlak görsel bilginin tamamlanmasında oynadığı rol sebebiyle son zamanlarda büyük önem kazanmıştır. Bu bilgi, obje tanıma, derinlik algılama ve optik akış gibi bir çok önemli bilgisayarla görme probleminde kullanılmaktadır. Bu tez çalışmasında, sınır sahipliği problemi için ideal modeli geliştirmek amacıyla iki ayrı yaklaşım sergilenmiştir. Denetimli yaklaşımda, anlamsal sınıflar arasındaki içeriksel ilişkileri modellemede en uygun aday olması sebebiyle Koşullu Rastgele Alan kullanılmıştır. İkinci yöntemimiz, farklı bilgi kaynaklarını ilişkileriyle modelleme amacıyla kullanılan Tensör Oylama yöntemi ilham alınarak geliştirilen Yinelemeli Vektör Oylama yöntemidir. Bu yöntem, denetimsiz bir oylama yöntemi olup, Gestalt görsel ipuçlarını kullanmaya uygun bir yöntemdir. Yapılan deneyler, geliştirdiğimiz geniş veri seti üzerinde aldığımız sonuçları doğrultusunda, iki modelin de sınır sahipliği problemine önemli katkılarda bulunduğunu göstermektedir.

Anahtar Kelimeler: Sınır Sahipliği, Figür-Zemin Ayrımı, Koşullu Rastgele Alan, Grafiksel Modeller, Tensör Oylama

To my family.

ACKNOWLEDGMENTS

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

First and foremost, I would like to thank my supervisor Assist. Prof. Dr. Sinan Kalkan for his endless guidance, support and patience.

I offer my deepest respect to Prof. Dr. Fatoş T. Yarman Vural for her guidance. I feel very privileged to know such a wise, meritorious person.

I am so grateful to Nazlı Gökçe Terzioğlu for her continued support and encouraging love.

I would like to thank my thesis jury members Assist. Prof. Dr. Ahmet Oğuz Akyüz, Dr. Aysenur Birtürk and Dr. Erkut Erdem for their insightful, guiding comments.

I am also thankful to my project teammates, Mehmet Akif Akkuş and Gaye Topuz, for their help and collaborative works.

Finally, I wish to appreciate the financial support from TÜBİTAK that funded this study (Project Number: 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	xvi
CHAPTERS	
1 INTRODUCTION	1
1.1 Contributions	2
1.2 Outline of the Thesis	3
2 BACKGROUND	5
2.1 Border Ownership Problem	5
2.2 Graphical Models	8
2.2.1 Markov Random Fields	10
2.2.2 Conditional Random Fields	12
2.3 Tensor Algebra	15
2.3.1 Tensor Voting	16

	2.3.1.1	Encoding	19
	2.3.1.2	Voting	21
3		RELATED WORK	25
	3.1	Studies of Border Ownership Problem	25
	3.2	Datasets for Border Ownership Problem	31
4		METHODOLOGY	33
	4.1	Border Ownership Cues	33
		4.1.1 Size	34
		4.1.2 Contrast	35
		4.1.3 Entropy	35
		4.1.4 Convexity	36
		4.1.5 Junctions	37
		4.1.6 Lower Region	38
	4.2	Methods Developed for Border Ownership Problem . . .	38
		4.2.1 Extraction of Border Ownership Information by Conditional Random Fields	39
	4.3	Extraction of Border Ownership Information by Iterative Vector Voting	42
		4.3.1 Iterative Vector Voting	44
		4.3.1.1 The Algorithm for IVV	48
5		EXPERIMENTS	55
	5.1	Dataset and Evaluation	55
	5.2	Evaluation of Visual Cues	56
	5.3	Experiments on the CRF-based model	57

5.4	Experiments on the IVV model	57
6	CONCLUSION AND FUTURE WORK	61
	REFERENCES	63
	APPENDICES	
A	BORDER OWNERSHIP LABELING PROGRAM & DATASET	67

LIST OF TABLES

TABLES

Table 5.1	Region-based cue accuracies on BO	57
Table 5.2	Accuracies of cue combinations on CRF model	57
Table 5.3	Accuracies of separate cue contributions on the IVV model . .	58

LIST OF FIGURES

FIGURES

Figure 1.1	Border ownership example	2
Figure 2.1	Rubin's vase	6
Figure 2.2	A simple directed graph example	9
Figure 2.3	Examples of Gestalt principles	17
Figure 2.4	2-D Tensor Voting methodology schema	18
Figure 2.5	Geometric descriptions of stick and ball tensors	19
Figure 2.6	Tensor decomposition in 2D	21
Figure 2.7	Vote of stick tensor	22
Figure 3.1	BO dataset example - 1	28
Figure 3.2	BO dataset example - 2	28
Figure 3.3	Shapeme set for FG segregation	29
Figure 3.4	Test images for BO experiments	32
Figure 3.5	Example data drawn erroneously from BSDS	32
Figure 4.1	Rectangles - size cue	34
Figure 4.2	Rubin vase - size cue	34

Figure 4.3 Example - contrast cue	35
Figure 4.4 Example - entropy cue	36
Figure 4.5 Example - convexity cue	37
Figure 4.6 Junction types	37
Figure 4.7 (a) Sample image (b) All junctions are colored red (c) BO decisions of T & L junctions. Red arrows show the owner region of shared border. In T-junction, the shared border is the horizontal line, whereas in L-junction it is the whole line (d) Directions of owner regions in the sample image	38
Figure 4.8 Graphical representation of BO	39
Figure 4.9 Example - impact of orientation on border ownership	45
Figure 4.10 Voting schema of stick tensor	46
Figure 4.11 Comparison of voting fields of TV & IVV	47
Figure 4.12 Voting schema of a visual cue vector	47
Figure 4.13 Iterative Vector Voting algorithm schema	48
Figure 4.14 A set of linels extracted from an image. Each linel is shown with a different color.	49
Figure 4.15 Curvature on border	51
Figure 4.16 Pixel-based contrast tensor	51
Figure 4.17 Sample IVV scenario with the results	54
Figure 5.1 Sample visual results of CRF and IVV models	59
Figure A.1 Border ownership labeling program login/register page	68
Figure A.2 Border ownership labeling program register page	68

Figure A.3 Border ownership labeling program take-tour page	69
Figure A.4 Border ownership labeling program tutorial page	69
Figure A.5 Border ownership labeling program labeling page	70

LIST OF ABBREVIATIONS

HVS	Human Visual System
MRF	Markov Random Field
CRF	Conditional Random Field
BO	Border Ownership
FG	Figure-Ground
GUI	Graphical User Interface
GT	Ground-Truth
CV	Computer Vision
BSDS	Berkeley Segmentation Dataset
TP	True Positive
TN	True Positive
FP	False Positive
FN	False Negative
PDF	Probability Density Function
IVV	Iterative Vector Voting

CHAPTER 1

INTRODUCTION

Images are simply two-dimensional projections of the real world. Such a transformation causes loss of visual data, since objects and their backgrounds are represented as regions and their boundaries. Besides, low-textured homogeneous areas are not easy to process in correspondence-based vision tasks (such as optical flow, stereo, structure from motion) as they do not have any distinguishable visual structure. In other words, these types of regions are not able to create any change on the receptive fields. This leads to ambiguous and incomplete visual data. Human vision system (HVS) and artificial vision methodologies have to deal with such deficiency and perceive 3D information from 2D data.

In order to resolve this vision problem, reliable visual information available at the borders are used by "filling-in" mechanisms, by diffusing the information into the regions from borders. Such a mechanism needs the information of owner regions of borders to be utilized, i.e., "Border Ownership" (BO) information.

In the Figure 1.1, a basic example of border ownership is shown: the gray pentagon is the owner of the red border between two pentagons as it is on the front, and thus causes occlusion.

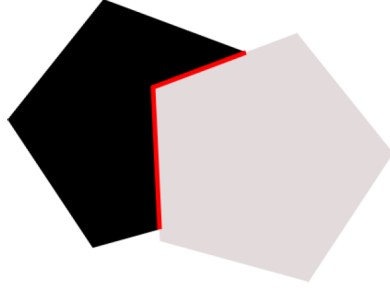


Figure 1.1: Border ownership example
Source: [5]

1.1 Contributions

This study proposes two new methods with promising results to the BO literature: a graphical model using conditional random fields (CRF) and a pixel-based voting algorithm called Iterative Vector Voting (IVV). Both methods utilize Gestalt cues such as size, convexity, entropy, contrast, lower-region and T & L junctions, as suggested by existing psychological, physiological and experimental studies [19, 35, 36, 45].

In the first part of the thesis, a supervised approach has been followed for which training is a must. Border ownership problem is defined as a two-node CRF graphical model, where nodes represent figure and ground. Their contextual integrity takes role on the decision of BO. The edge between these nodes is defined as the border, located between figure and ground. Across the border, cue values are calculated and used as features of CRF model. Utilizing trained graphical model, BO labels are estimated for each border.

The second method for estimating BO is called IVV. It adapts the voting approach of Tensor Voting (TV), and shows the same characteristic with TV. It is a non-iterative, fast unsupervised algorithm. IVV algorithm briefly works as follows: Initial BO labels are extracted for each border pixel using Gestalt rules of same visual cues. These rules decide BO discretely, *e.g.* "the region of higher entropy owns the border". Then, most salient features are extracted through voting and used for final labeling.

TV is adopted while developing IVV for its convenience to the BO problem: It provides different data types the ability to communicate and propagate their information to each other. Besides, its basic formalism has been extended to solve various vision problems such as stereo [22], optical flow [33], motion segmentation [25] and flow visualization [21]. Thus, IVV can be considered as an extended version of TV, applied on the BO problem. Main differences of these two methods are:

- **Representation:** IVV uses vector instead of tensor, as it requires less information to keep.
- **Voting:** In TV, there exist two types of voting, which are sparse and dense. IVV does not need dense voting as the votes located on borders are only needed. Besides, voting procedure is applied multiple times iteratively, with respect to BO propagation. Voting scale of IVV is also smaller than TV to keep votes same after these iterations.

1.2 Outline of the Thesis

The thesis is organized as follows:

- **Background**
All essential background information about BO problem, graphical models, TV, visual cues of Gestalt Psychology and brief information about existing datasets are provided.
- **Related Work**
Current literature about the BO problem is described in details.
- **Methodology**
Two models developed for the BO problem, CRF and IVV models, are explained.

- **Experiments**

Experimental results are presented and compared with the current BO literature.

- **Conclusion and Future Work**

All thesis work is discussed, essential deductions are made for the future work.

CHAPTER 2

BACKGROUND

In this chapter, background information on the BO problem is presented in four sections. First, a formal definition of BO is provided with the purpose of use and its importance for vision. Following this section, Tensor Voting and Conditional Random Field algorithms, which are used to develop a BO model, are presented in second and third sections. Finally in the last section, existing image datasets are discussed briefly with their advantages and disadvantages.

2.1 Border Ownership Problem

Border ownership is a vision problem of identifying which borders belong to which regions in an image. The resulting information, in which border-region relationships are defined, is important for both HVS and computer vision. It is being used in many vision problems such as figure-ground segregation, depth perception, object recognition and optical flow.

The Gestalt cue of "border" is a result of visual occlusion: objects closer mostly occlude the complete view of other objects behind. This situation creates contours, belonging to the closer object as they are responsible for the occlusion. They carry important visual information, thus they can be considered as low-level image features which separate regions in an image.

Vision, due to its nature, needs to overcome the loss of visual information as the images are just two-dimensional (2-D) instances of three-dimensional (3-D) real world. Perception begins at the retina with cluttered 2D visual data.

According to Gestalt psychologists, this data is regulated through the process of figure-ground (FG) organization. FG organization can be examined under two processes: perceptual grouping and FG segregation, which complete each other in a contrary way. Perceptual grouping states that low-level, primitive visual pieces of similar characteristics are grouped together by HVS to extract high-level, semantic structures. On the other hand, FG segregation states that the figure is perceived to stand out from the background, being bounded by a closed contour, behind which the background appears to continue. These two alternative ways in which two abutting regions could be organized into figure and background cannot occur together in our conscious perception. This phenomenon is nicely illustrated as the famous *Rubin's vase* in Figure 2.1. Rubin's vase can be perceived either as two black faces looking at each other, in front of a white background, or as a white vase on a black background. In the case of a FG reversal one line can have two shapes. The shape of the contour formed depends on which side of the line is regarded as part of the figure. This is important, because the visual system represents or encodes objects primarily in terms of their borders. Moreover, elements which are close to one another, or alike, or homogeneous in certain respects tend to be grouped together. This is called *perceptual grouping*. The sudden reversal perceived may be due to subject's shift of attention on the shape of the contour. The observer's perceptual set and individual interests can also bias the situation. Biasing the shapes or contours can make one interpretation stronger than the other one.



Figure 2.1: Rubin's vase
Source: [2]

Some properties of FG relationship:

- Figures hold more memorable association than the ground.

- Figures are located in front of the ground.
- The ground is assumed to be composed of uniform material and seems to extend behind the figure.
- The contour separating the figure from background belongs to the figure.

Both artificial and biological vision systems need to deal with insufficient visual information. As an example to this situation for HVS, it is known that homogeneous or weakly-textured image regions are not able to excite the perception fields of neurons, thus they can not stimulate visual cortex [10]. Although such regions do not consist of any distinguishable structure, human visual system is able to recognize homogeneous image regions thanks to its “filling-in” mechanism.

Optical flow is defined as the pattern of apparent motion of objects and caused by the relative motion of objects and the viewer. It provides worthy information about the spatial arrangement of objects and the rate of changes in their positions. Discontinuities in the optical flow help extracting object regions from the image and learning the shapes, distances and movements of the objects [11].

Binocular disparity, on the other hand, refers to the difference of an image position resulting from two retinal projections, caused by the horizontal separation of eyes [30]. It is used to extract depth information. In computer vision, depth information is measured from the relative positions of similar features extracted from two stereo images of the same scene.

As the definitions and application areas of both optical flow and binocular disparity describe, their roles for perception and vision are significant. However, considering their application areas and related literature in computer vision, they both have a common weakness as they are not capable of extracting information from non-textured areas. Narrowing this gap is only possible by applying “filling-in” mechanisms of HVS to artificial vision systems [28].

Filling-in is a perceptual phenomenon of HVS describing how the inadequate information is completed throughout the physiological blind spot, natural and

artificial scotomata. This mechanism simply diffuses visual information around the borders into the visual region, eventually creates a non-existent visual feature from surrounding area. This phenomenon may cause one to think that some visual cue is perceived inside of a region when it is actually absent there [14]. It combines visual illusion with perceptual completion [28], which has two types: boundary completion, in which illusory contours occur as the continuation of contours in the surrounding area; and feature-based completion occurring due to features such as color, brightness, motion, texture and depth. In computer vision, filling-in mechanism is succeeded in a similar manner with HVS, by delivering the reliable visual information at borders into the region. For such a purpose, border-region assignments should be handled.

To summarize, BO information is very important for both biological and artificial vision as it completes deficient visual information. It is applied on many vision problems such as optical flow, stereo disparity, object detection and depth perception.

2.2 Graphical Models

A probabilistic graphical model is considered as the most associate solution to construct a model corresponding to both probability and graph theories. Thanks to graphical models, it is possible to model a complex system with its contextual relationships. Probabilistic aspect of these models helps us to know how this model is successful at gathering the simpler parts of a graph. The parameters are estimated, the hidden states are inferred with respect to observations through this aspect. On the other hand, graph theoretic aspect enables us to model the whole system as a graph, making it visually solid. It is more feasible to design general-purpose algorithms by using graphical models [24].

Probabilistic graphical models can be represented as $G = (V, E)$ where V stands for the vertices (*i.e.* nodes) and E stands for the edges (*i.e.* arcs) of the graph. Nodes represent random variables as the arcs represent conditional independence assumptions. Arc representation helps to create a compact representation of

probability distributions, which makes inference and learning easier. Here is a simple example of how the conditional independence relationships allow us to specify the joint distribution more compactly, whose graph is shown in Figure 2.2. According to the chain rule of probability, we can define the joint probability of all vertices as:

$$P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|A, B, C). \quad (2.1)$$

B is independent of C considering A as its parent. Besides D is also independent of A as B and C are its parents. In the light of these conditional independence relationships, the joint probability can be rewritten as follows:

$$P(A, B, C, D) = P(A)P(B|A)P(C|A)P(D|B, C). \quad (2.2)$$

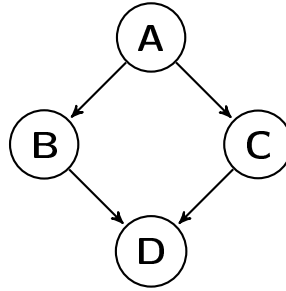


Figure 2.2: A simple directed graph example

There exist two types of graphical models: directed and undirected. Directed graphical models (DGMs) are mostly used in artificial intelligence and machine learning areas. On the other hand, undirected graphical models (UGMs) are widely used in CV. Despite the disadvantages of UGMs such that parameter estimation is computationally more expensive, UGM is a more preferred model for CV as it is symmetric, and thus convenient for spatial data. Besides, each image site can use any feature from the whole image by using CRF, which is not possible for MRF.

An artificial image can be a random group of independent pixels, but it is not valid for any natural image. Pixels, segments or even regions are arranged

spatially in a relationship with each other in a real scene. Contextual interactions can be in various types: An object can be defined as the meaningful combination of different parts or a semantic neighborhood of objects can be defined by the objects included. Thus, image context is assumed to have two main types: *local* and *global*. Local context describes the interaction between the parts of an object whereas the global context shows the interaction of objects (or specific image regions). The problem for generic object detection using natural images is how to model various types of context, considering the relationships between both observations and labels. Undirected graphical models help us to overcome this problem for CV.

2.2.1 Markov Random Fields

Markov Random Field (MRF) is a widely used undirected graphical model in computer vision, whose elements are random variables showing the Markov property. The Markov property simply states that each edge in the graph represents dependency, thus nodes which are not connected with an edge are independent from each other. More clearly, the Markov property consists of three properties of locality, which are:

- **Pairwise Markov Property:** Any two non-adjacent variables are conditionally independent.
- **Local Markov Property:** A variable is conditionally independent of all other variables given its neighbors.
- **Global Markov Property:** Any two subsets of variables are conditionally independent given a separating subset.

MRF is used to model the joint probability of the observations and labels. How an MRF model is utilized in image analysis is as follows:

Let x be the observations from the image (*i.e.* intensity) and y be the corresponding labels of these observations, represented as a vector of random variables. For MRF, the posterior probability over the labels have to be maximized

with respect to the observations, which is $p(x|y)$. Bayes rule defines the conditional probability as $p(y|x) = p(y)p(x|y)$.

Assuming that the observations are conditionally independent given the labels, the posterior distribution over the labels is defined as follows [16]:

$$p(y|x) = \frac{1}{Z} \exp \left(\sum_{i \in S} \log p(f_i(x)|y) + \sum_{i \in S} \sum_{j \in N_i} \beta y_i y_j \right), \quad (2.3)$$

where β is the interaction parameter of the MRF, Z is the normalization parameter, and $f_i(x)$ represents the relevant feature vector of x . Note that f_x is a single-site feature vector, which means it uses data only from a unique site. Conditional independence of data assumption states that the posterior is also MRF besides that the label prior $p(y)$.

MRF has two main drawbacks for image analysis:

- $p(x|y)$ is assumed to have a factorized form as $p(x|y) = \prod p(x_i, y_i)$. The factorization causes the data to be independent of others, but as mentioned before, in real images the data is often dependent on others, especially the neighbors.
- The label interaction is accepted as a prior for MRF. This means that the label interactions do not depend on the observations, but this, also is not valid for real scene images.

In conclusion, for the purpose of using graphical models in computer vision, the posterior distribution of labels $p(y|x)$ has to be estimated, but MRF succeeds this only for the observed sequence, thus it is not able to produce a global solution. Using generative training methods as MRF allows to use features that condition on the observed sequence at no penalty in terms of model complexity. On the other hand, for discriminative methods such as CRF, the features are separate factors for each individual label as far as the model is concerned. In other words, while a generative method models the joint distribution $p(x, y)$ that requires separate models for each observation sequence, a discriminative

framework directly models $p(y|x)$, which reduces computational complexity.

2.2.2 Conditional Random Fields

CRFs are discriminative models which directly model the conditional distribution over labels, *i.e.* $p(y|x)$ as a Markov model. This approach allows to capture arbitrary dependencies between the observations without resorting to any model approximations.

Let observed image data be abbreviated as $x = \{x_i\}_{i \in S}$, whereas the corresponding labels are shown as $y = \{y_i\}_{i \in S}$. Additionally, assume that $G = (S, E)$ is a graph such that y is indexed by the vertex that belongs to G . Due to these assumptions, (x, y) is a CRF if y_i , which are conditioned on x , obey the Markov property. This property can be summarized as:

$$p(y_i|x, y_{S-\{i\}}) = p(y_i|x, y_{N_i}), \quad (2.4)$$

where $S - \{i\}$ is the set of all the nodes in the graph except the node i , N_i is the set of neighbors of the node i in G , and x_w represents the set of labels at the nodes in set w .

With respect to the definition, it can be defined that a CRF is a random field globally conditioned on the observations. According to the Markov-Gibbs equivalence, conditional distribution over all labels y given observations x in a CRF is defined as

$$p(y|x, W) = \frac{1}{Z} \exp \left(\sum_{i \in S} A_i(x_i, y, w_0) + \sum_{i \in S} \sum_{j \in N_i} I_{ij}(x_i, x_j, y, w_1) \right). \quad (2.5)$$

In this equation,

- Z corresponds to the normalizing constant, which is also known as the partition function.
- A_i corresponds to the node/unary/association potential on clique i .

- I_{ij} corresponds to the edge/pairwise/interaction potential on set of cliques i and j .
- W corresponds to the parameter set $W = w_0, w_1$.

Node potential denotes the association of a single graph node to semantic labels. On the other hand, edge potentials denote how much the labels of nodes located on the same edge should interact each other.

Using a logistic function, local class conditional probability function can be written as follows:

$$p'(y_i = 1|f_i(x)) = \frac{1}{1 + e^{-(w_0 + w_1^T f_i(x))}} = \sigma(w_0 + w_1^T f_i(x)), \quad (2.6a)$$

where the node potential is

$$A(y_i, x) = \log p'(y_i|f_i(x)), \quad (2.6b)$$

and the edge potential is

$$I(y_i, y_j, x) = \log p''(y_i, y_j|\phi_i(x), \phi_j(x)). \quad (2.6c)$$

Edge potential can be considered as a function of features (ϕ), that measures the relationship, i.e. defines the semantic relation between two neighbor nodes. It is usually constructed upon the observations, behaviors and relations of data. Thus, for a CV problem, an edge potential can be constructed upon the ratio of color mean values, number of overlapping pixels, distance between two regions/points etc.

Image labeling problem is simply the problem of inferring labels (y), of observations extracted from image (x). Several CV problems such as stereo matching, image segmentation or image restoration can be posed as image labeling problems. To solve this problem, the algorithm should automatically partition the

image into *semantically* meaningful areas, each labeled with a class. With its mathematical definition, the problem is to find the most proper y that maximizes the conditional probability $p(y|x)$:

$$y^* = \arg \max_y P(y|x). \quad (2.7)$$

For labeling, two kinds of information are needed [17]:

- Spectral and spatial features from individual sites (intensity, color, texture, size *etc.*)
- Interactions with neighboring sites (contextual information)

CRFs are suitable to define object interactions and learn contextual relationships in images, if the relations of both labels and features of different sites/pixels/regions are to be fed to the model. It has various real-world applications in image processing: binary CRF for detection of man-made structures and multi-class CRF for purposes such as image classification and contextual object detection.

CRF models perform much better for real-world sequence problems as the results of studies using both CRF & MRF [17, 42] show. The primary advantage of CRFs over hidden Markov models is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference. Additionally, CRFs avoid the label bias problem, a weakness exhibited by maximum entropy Markov models (MEMMs) and other conditional Markov models based on directed graphical models.

To summarize, CRFs are suitable to define object interactions and learn contextual relationships in images. Thanks to CRFs, relations of both labels and features of different sites/pixels/regions are possible to be fed to the model. It has various real-world applications in image processing: binary CRF for detection of man-made structures and multi-class CRF for purposes such as image classification and contextual object detection [16].

2.3 Tensor Algebra

With respect to tensor algebra, generating new vectors is possible only by multiplying vectors by scalars. The resulting vectors have different magnitudes but they represent the same direction. To change both direction and magnitude, neither dot product nor cross product works. Dot product returns a scalar value, on the other hand it is not possible to identify multiple directions separately by cross product.

In physics, it is necessary to represent data having more than one direction. For instance, two basis vectors are needed to represent the forces applied inside a solid object: the first one is for the area vector and the second one is for the force applied on the area. The tensor composed of these vectors is called stress, which means force per unit area. Stress has the units of N/m^2 , and it is represented by tensors.

Tensors hold information about directions and magnitudes of these directions. They are simply mathematical objects used for representing real-world systems. It is proved that tensor algebra is very useful in many engineering contexts such as fluid dynamics, machine learning, and besides, in the analysis of other complex systems such as finance [23].

Vector algebra is a subsection of tensor algebra, as vectors are considered as tensor of rank 1, which has magnitude and one direction. With respect to tensor terminology, tensors of consecutive ranks are named as follows:

- **Scalar:** Tensor of rank 0 (Only magnitude)
- **Vector:** Tensor of rank 1 (Magnitude & 1 direction)
- **Dyad:** Tensor of rank 2 (Magnitude & 2 direction)
- \vdots \vdots

Regarding its practical importance, a 2nd order tensor can be considered as a linear operator that transforms a vector into another vector through a dot product. More generally, tensors are multi-dimensional generalizations of matrices,

which enables the use of multiple directions & magnitudes.

2.3.1 Tensor Voting

Lee & Medioni [18] proposed a unified computational framework for inferring perceptual structures from sparse binary data, which is either noisy, oriented or non-oriented. The perceptual structures they considered are:

- Curves and junctions for 2-D data
- Curves, junctions and surfaces for 3-D data

For data representation, tensor is utilized and all the calculations are handled through linear voting. Lee & Medioni proposed this combinatory method called **Tensor Voting** (TV) for perceptual organization based on the Gestalt principles. The method is initially applied on various computer vision problems such as boundary inference, stereo matching, then it is extended to instance-based learning [23].

Many problems related to computer vision has a common characteristic of being ill-posed, computationally expensive and corrupted by noise. Besides, many of these can be evaluated under the concept of perceptual organization of primitives as their solutions usually have perceptually salient patterns. The idea of TV framework comes from a general, data-driven solution of capturing salient visual structures. It addresses a wide range of problems of perceptual organization, and utilizes Gestalt principles.

Gestalt psychology was developed during 1920s by three German psychologists: Wertheimer, Kafka and Kohler. They found out that HVS subconsciously segregates and groups visual information in order to perceive it as a whole. Gestalt motto states [47]: "The whole is greater than the sum of the parts", *i.e.*, the gatherings of visual pieces provide better information about the perceptual structures than individually. In Figure 2.3, some examples of Gestalt principles are given [23]. In Figure 2.3a, regarding the proximity rule, human perception groups the dots into four by their distances to each other. In Figure 2.3b, dots

are grouped in pairwise horizontally, with respect to their colors. In Figure 2.3c, the curve from A to B is perceived as a whole as it obeys the continuation rule. Finally in Figure 2.3d, the curves are perceived as an ellipse and a square interfered with each other, due to the closure rule.

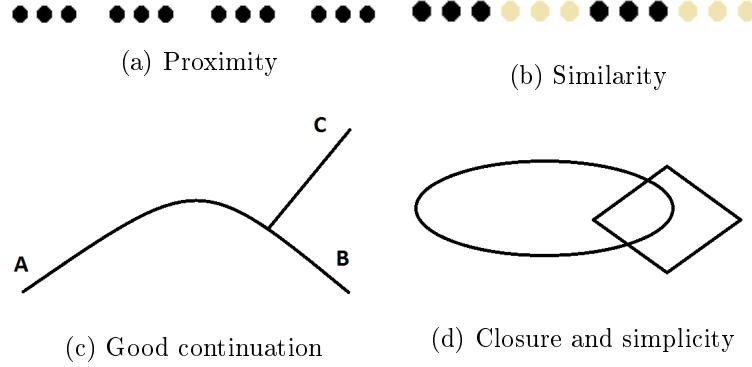


Figure 2.3: Examples of Gestalt principles
Adapted from: [3]

Properties of TV framework are given as follows:

- General
 - provides model-free solutions
- Local
 - local changes affect descriptions locally
- Data-driven
 - needs no training
- Computationally inexpensive
 - able to process large data
- Robust to noise
 - able to tolerate large number of outliers
- Based on Gestalt principles of proximity and good continuation

Data representation of TV is in the form of a 2nd order, symmetric, non-negative definite tensor. The tensor infers both the saliency value and the preferred orientation of perceptual structures including curve, junction and region in 2-D

the token belongs to. How the TV framework works for 2-D is summarized by its main steps in Figure 2.4. The methodology is based on two components [21]:

- Tensor calculus: used for data representation
- Tensor voting: used for data communication & vote propagation

Initially, each input token is encoded into a tensor. If the input is a point, it is encoded as a ball tensor of unit radius. Otherwise if it is a curve, then it is encoded as a stick tensor with direction information.

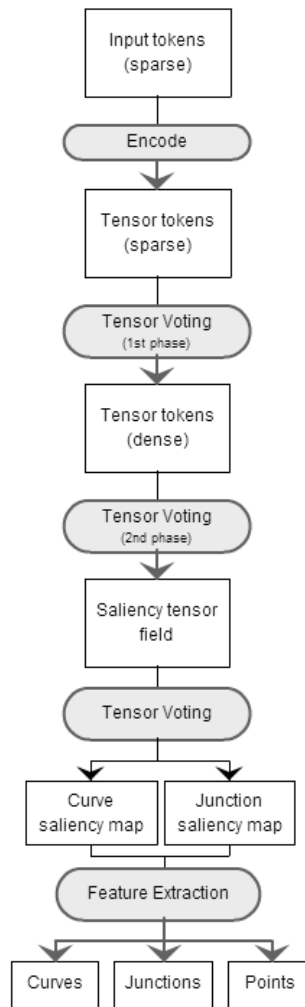


Figure 2.4: 2-D Tensor Voting methodology schema
Adapted from: [23]

After the decoding process, a two-phase tensor voting procedure is applied. The first one is called sparse voting while the second one is called dense. In sparse

voting, all tokens propagate their information to each other in a limited neighborhood, which is called tensor communication. Thanks to this voting, knowledge of curve orientation and the saliency of the knowledge are consolidated, thus tensor tokens become refined. Definition of the neighborhood, which depends on the tensor type, are provided ahead. Later on, in dense voting part, tensors propagate their refined information to every point in their neighborhood. Ball components of the tensors are not incorporated into voting in this phase. Finally, resulting dense tensor map is decomposed into junction and curve maps.

Details of each process in tensor voting framework, which are encoding, voting and decomposing, are mentioned in the following sections.

2.3.1.1 Encoding

A 2nd order symmetric, non-negative definite tensor (briefly, 2-D tensor) can be considered as a 2x2 matrix, or an ellipse in 2D. A visual input for tensor voting framework, which is either a point or a piece of curve, is encoded as a unit **ball** or **stick** tensor respectively. Corresponding tensors, eigenvalues and quadratic forms (2x2 matrices) of oriented (curve) & non-oriented (point) inputs are given in Figure 2.5.

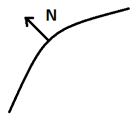



Input	Second order tensor	Eigenvalues	Quadratic form
		$\lambda_1=1$ $\lambda_2=0$	$\begin{bmatrix} n_x^2 & n_x n_y \\ n_x n_y & n_y^2 \end{bmatrix}$
		$\lambda_1=\lambda_2=1$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Figure 2.5: Geometric descriptions of stick and ball tensors
Adapted from: [23]

A 2-D tensor can be decomposed as:

$$T = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T = (\lambda_1 - \lambda_2) e_1 e_1^T + \lambda_2 (e_1 e_1^T + e_2 e_2^T), \quad (2.8)$$

where λ and e are for eigenvalues and eigenvectors, respectively. The first part of the second sum in the equation, which is $(\lambda_1 - \lambda_2)e_1e_1^T$, corresponds to the stick component of the tensor. If T equals to this component, then the tensor is called ***stick tensor***. It has a degenerate elongated ellipsoid structure, as seen in Figure 2.5. Presence of a stick tensor means that a piece of curve exists on that position with e_1 as the curve normal. The size of the stick component, which is $(\lambda_1 - \lambda_2)$ indicates the curve saliency.

On the other hand, the second part of the sum, which is $\lambda_2(e_1e_1^T + e_2e_2^T)$, corresponds to the ball component of the tensor, directly called ***ball tensor***. It represents a perceptual structure having no orientation, or multiple orientations which neutralizes each other at this point. The size of the tensor, again, provides the certainty of information existing on this position. As understood from its name, geometrically ball tensor is in the shape of a circular disk, as seen in Figure 2.5.

The properties of a 2-D tensor are [23]:

- The axes of the ellipse are the eigenvectors of the tensor.
- The aspect ratio of the axes is the ratio of the eigenvalues.
- The major axis is the preferred normal orientation of a potential curve going through the location.
- The shape of the ellipse indicates the certainty of the preferred orientation.
 1. An elongated ellipse represents a token with high certainty of orientation
 2. A degenerate ellipse with only one non-zero eigenvalue represents a perfectly oriented point (*i.e.* a curvel)
 3. An ellipse with two equal eigenvalues represents a token with no preference for any orientation.
- Tensor size encodes the saliency of the information encoded.
 1. Larger tensors convey more salient information than others.

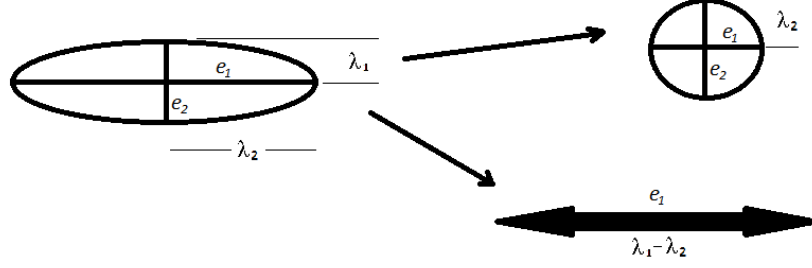


Figure 2.6: Tensor decomposition in 2D
Adapted from: [23]

In Figure 2.6, it is shown that how a generic tensor is decomposed into stick and ball components. As the tensor has parameters $\lambda_1, \lambda_2, e_1, e_2$ as its eigenvalues and eigenvectors, it simply means that:

- The orientation of its normal is at the direction of \hat{e} .
- The saliency of the curve is measured by $\lambda_1 - \lambda_2$

Details of each process in tensor voting framework, which are encoding, voting and decomposing, are mentioned in the following sections.

2.3.1.2 Voting

After the input tensor encoding is completed, two-phase voting is implemented, which are sparse and dense voting. Procedures of these two voting types are same, they both can be considered as a tensor convolution with separate voting kernels. The main difference arises from the voting domain. Besides in dense voting, votes of ball tensors are ignored as they define isolated features.

In sparse voting, tensors vote others located in their neighborhood. A generic tensor is generated at each token location, which equals to the tensor sum of all votes there. Sparse voting deduces the most preferred orientation by refining the initial one for each token. Thus, sparse voting is called as *token refinement*.

On the other hand, tensors vote every point in their neighborhood in dense voting. Initially, each generic tensor is decomposed into its ball and stick com-

ponents. Ball and stick tensors have distinctive voting fields, and these fields define their neighborhoods. They broadcast their information to all discrete cell locations in these neighborhoods. Simply, dense voting extrapolates the information to the whole domain so as to extract features coherently. That's why, it is also called as ***dense extrapolation***.

As described, ball and stick tensors own their voting kernels. Instead, all voting kernels, regardless of the dimension (2-D/3-D), or the type (stick/ball/plate), are derived from the fundamental 2-D stick kernel. Before defining 2-D stick kernel, let us examine the vote of a stick tensor, which is visualized in Figure 2.7.

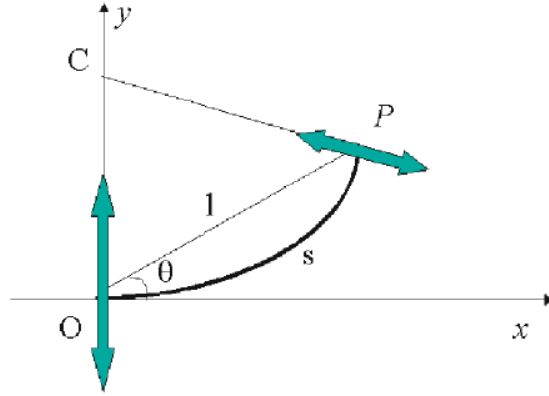


Figure 2.7: Vote of stick tensor
Adapted from: [23]

Mordohai and Medioni [23] claim that the bold curve in Figure 2.7 is the most likely smooth path between P and O , as it is the arc of the osculating circle. Thus, the vote cast at O by a stick tensor at P has the direction of the tangent. Besides, they state that the vote should attenuate with respect to distance and curvature. Under these considerations, the *saliency decay function* is defined, which has the following form:

$$DF(s, \kappa, \sigma) = e^{-\frac{s^2 + c\kappa^2}{\sigma^2}}, \quad (2.9)$$

where s is the arc length, c is the decay control function, σ is the scale of voting, determining the effective neighborhood size and κ is the curvature.

2-D stick and ball votes of a unit stick tensor, of which the parameters are shown in Figure 2.7, are formulated as below in Equations 2.10 & 2.11. Unit vote is used in sparse voting as all tensors have unit magnitude. Otherwise, for a generic stick tensor of arbitrary size, the vote has to be multiplied by the size, *i.e.*, $(\lambda_1 - \lambda_2)$. As easily understood from both the geometric representation and the formulation, ball vote function is simply a fully-rotated version of stick vote function.

$$Vote_{Stick} = DF(s, \kappa, \sigma) [-\sin(2\theta) \cos(2\theta)]^T [-\sin(2\theta) \cos(2\theta)], \quad (2.10)$$

$$Vote_{Ball} = \int_0^{2\theta} R_\theta^{-1} Vote_{Stick}(R_\theta P) R_\theta^{-T} d\theta. \quad (2.11)$$

A voting field can be considered as a map in which directions (orientations) and magnitudes of all votes cast by the voter tensor are shown. Votes are stored in these pre-computed voting fields, thus the expense of computation decreases during voting.

CHAPTER 3

RELATED WORK

In this chapter, current literature on BO problem are explained under two sections. All remarkable researches and studies about BO problem are mentioned in the first section. Later, existing BO datasets are described. Devoting a whole section to datasets is necessary as the data variety is crucial for estimating BO.

3.1 Studies of Border Ownership Problem

BO problem is a quite new research topic and it has received insufficient attention so far. Current studies can be separated into two groups: the ones based on psychological and physiological experiments, tested on artificial images and the ones working with real, complex imagery.

First group of studies basically aim to generate the model based upon the HVS and neural system [8, 13, 31, 32, 39, 50]. They commonly investigate the neural mechanisms for BO determination. Neural networks are among the mostly-used methods for BO problem [8, 13]. Neural networks are the computational models constructed on central nervous systems of animals, which are responsible for handling machine learning and pattern recognition operations. It is suitable for generating realistic border-ownership models, as it simulates human neural system. The methodology followed by this group of studies is mostly in this order:

- Measurement of stimuli changes on psychological and physiological exper-

iments

- Evaluation of the data collected on the first step
- Generating a model based on human neural system

Initial studies focused on finding out which parts of the neural system decide border ownership, and which visual cues have what kind of effects on this decision. Physiological experiments on monkeys have shown that V2 and V4 areas of visual cortex decide the owner of a visual border. They found out that 50% of BO-selective neurons respond to the contrast polarity of the border [50]. A network model of contrast-dependent BO-selective neurons is developed in the light of this knowledge [26]. It consists of three main stages: contrast detection by V1 simple-cell-like units, determination of surrounding contrast configuration, and contrast-dependent BO determination. This study, as other experimental studies, shows the lack of utilizing complex visual information in its model. It states that BO studies using junctions are not physiologically realistic despite the existence of such models, as any neurons selective to complex visual information such as junctions have not been reported yet. There exist experimental facts supporting this assumption: The latency of BO-selective neurons is nearly 10 ms, that's why complex processes like junction detection cannot be involved. Although this fact is still valid, knowledge of T-junctions' being the most reliable cue creates a significant conflict while constructing an ideal model. Zhaoping [49] also uses a model of V2 neurons in order to prove that V2 visual area can produce the ownership signal by itself, without the need of any top-down mechanism or spatial information of T and L junctions.

On the contrary, there exist studies stating spatial properties such as shape or junctions have significant impact on BO decision. Tomasi et al. [44] imitates the initial stage of HVS while extracting shape and motion information from image streams. This initial vision stage is called *early vision*. In the early vision, shape, appearance and motion of objects are detected while semantic interpretation is handled in the latter processes, which belong to higher levels of vision. Zhou et al. [50] found through physiological experiments that there exist cells responsible of encoding which local contours belong to the object.

In the light of this exploration, it can be argued that shape and appearance have impact on border ownership. Kikuchi and Akashi [13] generates a simple neural network model stating that BO is encoded in early vision. Their computer simulation also confirmed that the cells in their model showed similar responses with the cells coding BO information, which Zhou found [50]. This neural network model uses contrast, orientation, curvature and L-junctions as visual cues. The method exhibits a shortfall, which is labeling only one side of boundary by using L-junctions and curvatures as cues, since the neural model forces such an assumption. Additionally, it has two other common deficiencies with the other physiological models: low quality of the data set and generalized assumptions due to the results. In Figure 3.1, simulation data of this study, and their corresponding BO results are shown. Short lines represent the direction of figure to which border belongs at that location. The data is obviously simple and artificial, results of which are quite impossible to be practiced on real imagery. Moreover, as the first two images have reversed contrast polarity and the network showed same response, it is stated that the model has a behavior of contrast independent border-ownership coding cells. Although this deduction is valid for the model, it is certain that an efficient cue like junction is ignored. Albert [6] also compared shape-based cues on BO decision by measuring stimuli changes of perception neurons. Samples from the dataset used is shown in Figure 3.2, which are similarly simple and artificial.

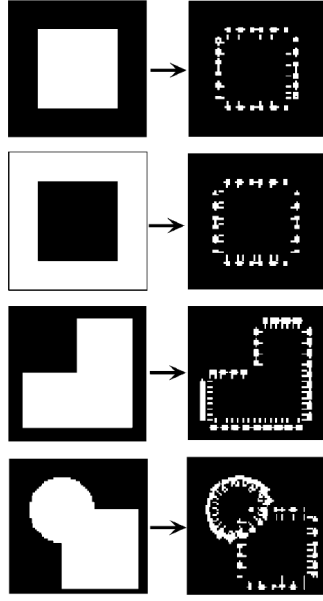


Figure 3.1: BO dataset example - 1
Source: [13]

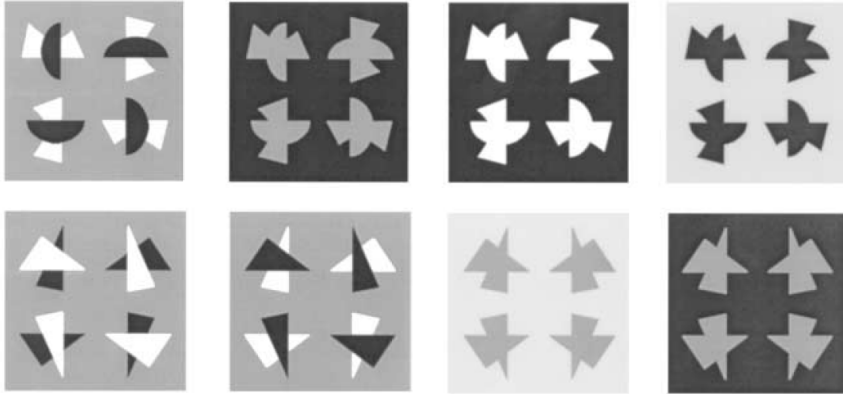


Figure 3.2: BO dataset example - 2
Source: [6]

Second group of studies use natural images, with up-to-date methods such as graphical models. They are mostly prepared under the titles of "figure-ground organization/segregation/separation", but these titles refer the same problem with BO, as the detection of figure-ground reveals BO information. Peterson [29] showed that familiar configurations of meaningful borders, such as boundaries of recognizable objects, provide a powerful cue for BO information. After Peterson's this result, Malik, Fowlkes and Ren [35] tried to generate a decision mechanism for familiar configuration in terms of prototypical local shapes, with-

out the requirement of object detection. Initially, Malik et al. [9] introduced "shape context", which corresponds to a shape descriptor compressing the information of local arrangement of edges in a log-polar structure. It simply works by counting the number of edge points inside each bin, relatively to a center point. Later, they introduced the definition of "shapeme" [36], an orientation-independent generic shape descriptor. The aim of defining shapemes is to learn local FG cues automatically. Shapemes are constructed just by clustering a large set of shape context descriptors. They are used for FG segregation as follows:

- The similarity of the local shape extracted from test image is measured for each shapeme in the shapeme set, which is provided in Figure 3.3.
- A logistic classifier is trained to predict FG label by using this similarity measure set as feature vector.

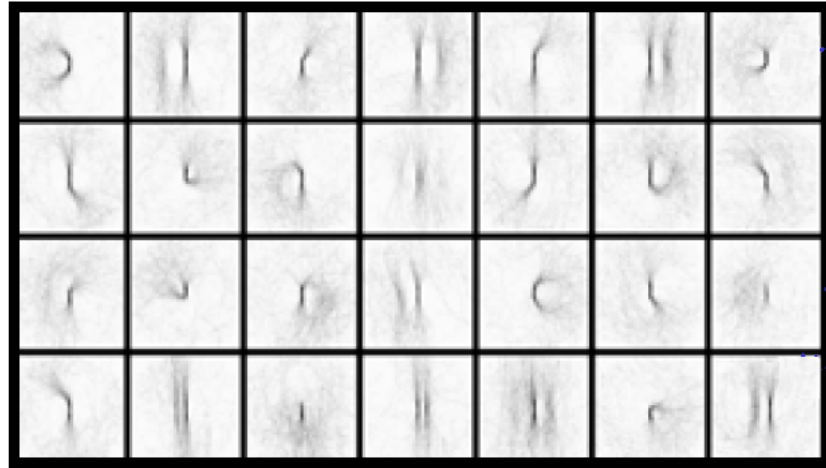


Figure 3.3: Shapeme set for FG segregation
Source: [36]

As stated, shapemes are defined as visual cues without any definition. Results show that they can represent the characteristics of Gestalt cues. In the shapeme table 3.3, cue at the rightmost top corner represents convexity cue while the one at the rightmost bottom corner defines the parallelism cue. Results of this study also prove that shapemes capture other mid-level cues such as texture. Shapeme classifier has shown BO labeling accuracy of 64% on human-marked boundaries

of 200 natural images.

Second model of Malik et al. [36] applies the concept of familiar configuration on the BO problem both through local and global aspects in order. A logistic classifier is developed to locally predict FG labels, based on the shapeme representation. After local findings, a global FG model using conditional random field is used to enforce global consistency by learning T-junction frequency and continuity. Inference on this model is handled by loopy belief propagation. They show that their shapeme-based classifier outperform when compared to a baseline model using cues of size & convexity.

Integration of local and global approaches provides significant advantages to Malik et al. in their study [36]. Firstly, thanks to a local model, Gestalt cues are enabled to use, and this provides an important contribution to results as proven [32]. Secondly, although CRF is used to develop a global model, it enables separating image sites not only to model semantic interactions, but also each site can potentially use features from the whole image, unlike MRF as Kumar [15] states.

Local model provides the estimated probability (p_b) stating that the left side of each border (b) is figure. On the other hand, global model combines the estimated probability with junction information as follows:

$$P(Y|I, \theta) = \frac{1}{Z} \exp \left(\sum \mu(Y_B|I, \theta) + \sum \eta(Y_J|I, \theta) \right), \quad (3.1)$$

where Y_B is associated with labels of BO, where $Y_B = 1$ states that region at the left side of the border is the owner while $Y_B = -1$ states the opposite decision as $Y_B = 0$ means that no border ownership information is gathered. μ is the node potential function on each border, whereas η is a potential function on each junction labeled with Y_J .

Malik used a FG dataset of 200 images with a resolution of 321x481, 100 for testing and 100 for training [36]. For this data set, they obtained an accuracy of 72% by using only the local model of shapemes. With the combination of local and global models, they acquired an increase of accuracy by 7%. Evaluating their

results, they state that human subjects showed a 88% score of labeling, thus they showed an unforeseeable success. As the results are successful, considering such a small data set, a cross validation method could have been used for experiments for a more successful and realistic model.

To sum up, BO studies carried out through the present day reveal two important facts about BO problem: experimental studies are not able to produce reliable results, datasets are inadequate as their contents are either synthetic, simple or their numbers are not sufficient.

3.2 Datasets for Border Ownership Problem

A BO dataset should consist of images with boundary images where segments (regions) and edges (borders) are labeled. Moreover, all borders should be assigned to the regions which they belong to.

There exist a few image databases with their borders extracted. They are mostly generated by tools like LabelMe, which is a free on-line annotation tool to build image databases [38]. These databases do not include BO information either. Besides, those datasets consist of insufficient number of images. Fowkles et al. generated a data set of 200 outdoor images for the purpose of developing a computational model for FG assignment [36].

On the other hand, BO studies have mostly concentrated on visual perception so far. They observe the role of different visual inputs and cues on the stimuli changes in V1 and V2 areas of visual cortex. With respect to eye tracking of test subject (frequently awake monkeys), the neural responses of the visual cues are measured. Thus, the images used in these experiments should be as simple as they only show the characteristics of relevant visual cue. The images that Nishimura & Sakai [26] and Kikuchi & Akashi [13] used in their neural experiments can be seen in Figure 3.4. These gray-scale and binary images, respectively, are quite simple and synthetic images. It is not guaranteed that results would be similar if the same experiments are applied on real images. That's why the ability of generalization and satisfactoriness of these studies are

quite suspicious.

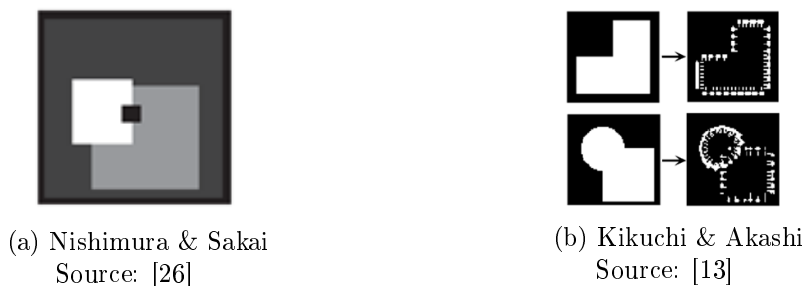


Figure 3.4: Test images for BO experiments

Berkeley Figure/Ground Dataset is used in several BO studies [36]. Each FG labeling in this dataset is associated with its segmentation mask from Berkeley Segmentation Dataset (BSDS). It is completely a real imagery data set. Although the images, segmentations and all relevant code are shared, 400 figure-ground labeling data is available to use among 1636 images. Besides, there exist quite wrong boundaries, which can not be considered to separate figure and ground areas, as in Figure 3.5. In this figure, wrong drawings are colored blue in the second image.



Figure 3.5: Example data drawn erroneously from BSDS
Source: [36]

To summarize, there is a precise need of a new dataset for BO problem, due to reasons such as insufficient number of data and unreliable drawings & labellings.

CHAPTER 4

METHODOLOGY

In order to extract certain border ownership information, two different methods are developed, which are based on supervised and an unsupervised learning, respectively. In the first approach, **CRF** is used for developing a trained graphical BO model. In the second approach, a linear voting method called Tensor Voting is adapted to BO problem, under a brand new method named **Iterative Vector Voting**.

In the following sections, details of the methodology are explained. Initially, visual cues (*i.e.* features) used in the experiments are mentioned. Then, two models developed for BO are introduced, which are as follows:

- CRF-based BO model
- Iterative pixel-based vector voting

4.1 Border Ownership Cues

In both supervised and unsupervised approaches, several cues are tested. Those are size, curvature, junctions, lower region, contrast, convexity, boundary length, contrast and texture. All these cues adapt Gestalt psychology, however some cannot be used in the pixel-based approach due to being region-wide. Although Gestalt principles may sometimes lead human vision perceive erroneously, these cues are among the most accurate ones for figure-ground segregation as previous studies show. In the following sections, they are defined with respect to Gestalt

terminology and illustrated by visual examples.

4.1.1 Size

The smaller of two objects is considered as the figure laid on a larger background. Thus, this rule is also known as "smallness". It is firstly introduced by Rubin [37], and found as the most powerful cue in the study [36], which uses size, lower-region and convexity as Gestalt visual cues. In the following visual example 4.1, it can be observed that black rectangle, which is smaller, is perceived as figure.



Figure 4.1: Rectangles - size cue

Here below in Figure 4.2, two different versions of classical Rubin vase example are visualized, in which the distance between two borders varies. In the first image, the vase is favored while faces are favored as figure in the second image due to the principle of smallness.



Figure 4.2: Rubin vase - size cue
Adapted from: [4]

4.1.2 Contrast

Contrast difference between visual regions is essential to separate objects from background. On the contrary, objects blend into the background. Just as some animals depend upon this principle to camouflage themselves in nature, contrast is a powerful cue for hiding or highlighting objects. As seen in Figure 4.3, as contrast value between words and the background decreases, FG segregation becomes harder.



Figure 4.3: Example - contrast cue
Source: [4]

4.1.3 Entropy

Image entropy is a quantity used to describe the randomness of an image. It is mostly considered to measure the energy, texture & information as it represents the amount of information which must be coded by a compression algorithm. Low entropy image does not have high rate of texture, high number of sudden intensity changes and consequently high energy. Thus it can be compressed to a relatively small size. There exist various ways of entropy calculation, as energy/texture does not have a unique representation but mostly, entropy is calculated as follows:

$$Entropy = - \sum_j P_j \log_2 P_j, \quad (4.1)$$

where P_j is the probability that the difference between two adjacent pixels is

equal to j . Probability P_j is simply calculated by the histogram counts.

Region with more textured structure is more probable owner of the border, compared to the coarse, flat one having lower entropy [10, 36]. In Figure 4.4, a simple real-life example, a scene of moon, is shown supporting entropy cue. Textured surface of the moon shows the characteristics of foreground as the black sky has a flat and non-textured surface, which suits the assumptions of entropy rule for background.



Figure 4.4: Example - entropy cue
Source: [4]

Besides these powerful features, there exist three other visual cues for FG organization, which are surroundedness, symmetry and convexity. These are quiet weak features, compared to previous ones. Also the early studies have provided results supporting this weakness, thus they were not used in the final experiments. In the following two sections, these cues are explained.

4.1.4 Convexity

When all constraints provide no decision of ownership, convex (protruding) rather than concave (indented) patterns tend to be perceived as figures. In Figure 4.5, black convex regions tend to be figures with respect to Gestalt psychology.

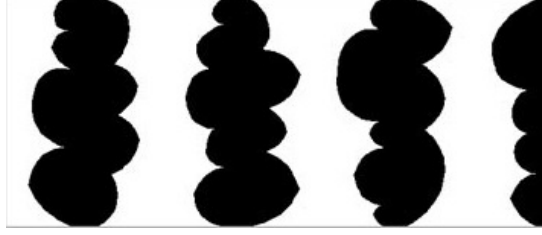


Figure 4.5: Example - convexity cue
Source: [4]

Besides Gestalt visual cues, there exist some other strong visual cues for determining border ownership: *junctions*, *lower-region* and *texture*.

4.1.5 Junctions

Junctions are visual structures that occur at the intersection areas of borders. There exist four kinds of junctions which are named with respect to their shapes: T-junction, L-junction, Y-junction and arrow-junction, shown in Figure 4.6.

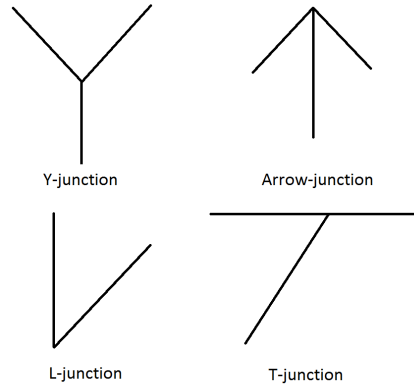


Figure 4.6: Junction types

Among these, T and L junctions provide reliable information for FG organization, if they are located on the boundary [34, 36]. However T-junction is the strongest cue of all, as relevant studies show [10, 36]. Thesis results also support these: T-junctions assign the owner of its widest border very successfully, with a precision of 74.7%. How T and L junctions are evaluated for FG segregation is visualized in Figure 4.7.

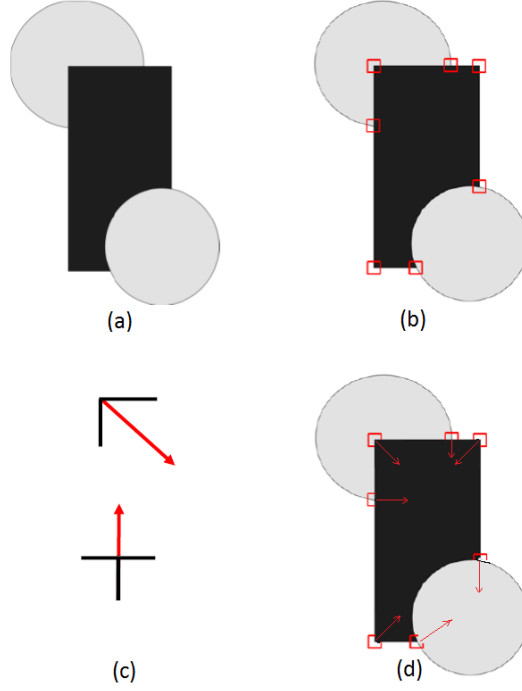


Figure 4.7: (a) Sample image (b) All junctions are colored red (c) BO decisions of T & L junctions. Red arrows show the owner region of shared border. In T-junction, the shared border is the horizontal line, whereas in L-junction it is the whole line (d) Directions of owner regions in the sample image
Source: [27]

4.1.6 Lower Region

Regions in the lower portion of the display are mostly considered as figures. This cue is initially introduced by Vecera et al. [45]. It is also used for the purpose of depth perception [20]. The results show significant success, which is 64.4% accuracy [36].

4.2 Methods Developed for Border Ownership Problem

To estimate the border ownership (BO) information, two methods are developed, which are Conditional Random Fields (CRF) and Iterative Vector Voting (IVV) models. CRF-based BO model is a graphical model utilizing the contextual relationships between neighbor regions, whereas IVV model is an unsupervised, pixel-based voting algorithm based on Tensor Voting (TV).

4.2.1 Extraction of Border Ownership Information by Conditional Random Fields

We define BO as a conditional probability distribution $P(Y|X)$ where X and Y are random variables. X corresponds to pairwise observation sequences of neighbor regions, which provide various visual information about themselves such as contrast or texture, whereas Y are their corresponding labels, conditioned on X .

To acquire a distribution of BO labeling data, generative models must enumerate all possible observation sequences, which is quite impossible. It is also possible to use global features in MRF by connecting a factor to all nodes, but it complicates inference significantly. For a BO model with CRF, it is only needed to do inference and optimization over the label nodes of pairwise model that covers neighbor regions, so it does not cost anything to add more factors.

The algorithm is on the basis of defining contextual relationships between two neighbor regions as a graphical model, training it and refining the initial labels with respect to the model. The model is visualized in Figure 4.8. As mentioned, the nodes represent two neighbor regions R_1 and R_2 whereas the edge stands for the shared border E . Node labels define which region is the owner of the border. Node value 1 is given to the owner region while the other region is labeled with 0.

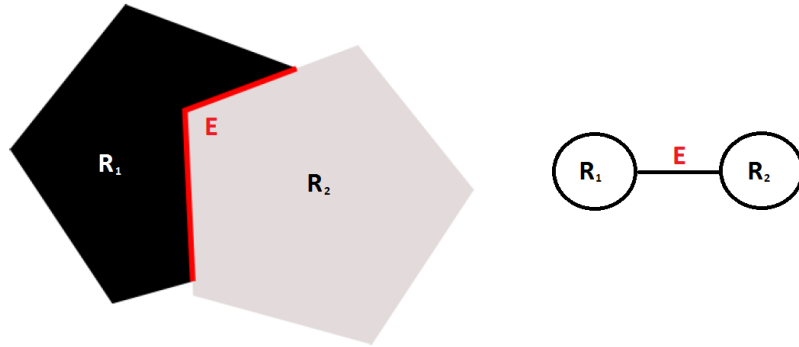


Figure 4.8: Graphical representation of BO

CRF-based graphical model is developed using two spectral and two spatial, region-based features, that are contrast, entropy, T-junctions and size respec-

tively. According to the current literature, these are region-based cues giving the best accuracy as mentioned in Section 4.1. Thesis experiments have also provided such promising results supporting the literature. These features effect both node and edge potentials: each feature builds its own unary potential whereas edge potential is built by concatenating all node features.

The model is very simple: 1 edge connecting 2 nodes, modeled as a Gibbs distribution. The reason of constructing such a model is also simple, as Wallach [46] states: More complex model needs more isolated, characteristic data. As borders align between two neighbor areas, two nodes represent these region while the edge represents the dependency between them.

As a graph of two neighbor regions do not cause any loop and there will not be any intractability problem, exact inference is used to compute the partition function and estimate marginal posterior distributions of labels Y . After estimation, binary label of BO decision is assigned to each border through the simple comparison of $P(Y=1) > P(Y=-1)$: If it is true, the border belongs to the first region, otherwise to the second. All CRF calculations are succeeded with UGM toolbox of Schmidt [40], which is publicly available on web.

Graphical models can be utilized to improve the initial labels of graph sequences with respect to the observations. Initial labels are assigned by decisions of T-junctions due to the following:

- T-junctions have shown an accuracy of 74,7% on BO assignment.
- It is not reasonable to use T-junction as a region-based feature as it does not cover the whole border.

While defining node potentials, entropy, contrast and size cues are used. Initially, lower region and L-junctions are also considered for the model thanks to their success at previous studies, but they are not preferred due to the following:

- Lower region is not sufficient for assigning a probabilistic value for BO: If the region is below, then it is the owner of the border. Besides, lower-region is used as a CRF cue initially despite its discrete value of (0,1), as it

may behave better in combination with other cues. But results have also supported our assumption, thus it is excluded from the finalized version of CRF-based model.

- **L-junctions** do not own a precise definition, the angle between two arms of the junction is a free variable.

Usage and definitions of BO graphical model cues are provided below. How they are calculated and what they represent are explained in details in Section 4.1.

- **Entropy:** Total entropy of region, calculated at gray scale color map.
- **Contrast:** Ratio of gray scale mean values of regions. Actually, it only represents intensity mean values for node potentials, but concatenated feature vector of edge potential implicitly provides contrast information.
- **Size:** Simply pixel numbers of regions.
- **Convexity:** Ratio of region's and its convex hull's areas.

Normalization to the range of $[0,1]$ is applied on all three cues in order to balance their contributions to the CRF model. Region with higher feature value is revalued as 1, whereas the feature value of other region is normalized with respect to its neighbor.

Probability distribution of our two-node BO graphical model is formulated as follows:

$$P(R|x, \mu) \rightarrow - \sum N(R_i, x, \mu) - \sum E(R_i, R_j, x, \mu). \quad (4.2)$$

Posterior probability of BO information is simply calculated as the normalized sum of all mono and pairwise potentials, which determine how labeling is handled with respect to knowledge obtained from both single regions or neighbor pairs of them. Here in the equation above N represents node potential whereas E stands for edge potential and μ denotes the whole parameter set.

Maximum a posteriori (MAP) of BO decisions are constructed upon initial labellings, which are assigned by T-junctions. Node potential function N is defined as the negative log likelihood of region's BO decision probability conditioned on four BO cues:

$$N(R_i) = -\log P(J_T | C_{R_i}). \quad (4.3)$$

J_T returns 1 or -1 with respect to the owner region. If no decision is made, then it returns 0. C_{R_i} defines the feature value of region R_i . Edge potential is constructed for each two neighbor regions by concatenating node potentials as follows:

$$E(R_i, R_j) = [N(R_i)N(R_j)]. \quad (4.4)$$

The whole dataset is separated randomly into two groups of same number: one for training the model and one for testing. For each pair of neighbor regions, a CRF model is constructed. Node and edge features are converted into node and edge potentials using multi-class logistic function. Model parameters are learned by minimizing the energy function which the potentials are fed to. In the test phase, the CRF model for each neighboring pair of regions is inferred by trained parameters. Exact inference is applied to learn BO labels.

4.3 Extraction of Border Ownership Information by Iterative Vector Voting

Physical structure and working mechanism of HVS is very similar to monkey visual system. In their experiments on awake monkeys, Lamme et al. [12] showed that cells of V1 visual cortex produce more response to textured stimuli whether the receptive field is on a visual area that belongs to a figure, rather than ground. This observation posed an important question, whether BO is decided locally on the contrary with previous studies showing that the identification of a visual region is handled through global image processing. All these studies state

that HVS should work on a region at least the same size with figure. But Lamme has revealed that FG segregation is handled by local processing of features, just in a small neighborhood of point in consideration.

Lamme's point of view has significant advantages but limitations, in comparison to the results of other BO studies. Experiments show that HVS is very successful on assigning BO values when a small region, mostly a sub-region, is shown to subjects. Although it is a very controversial issue whether HVS solves this kind of BO problems with a top-down or bottom-up approach, related experiments and results encourage to use local visual structures and features while deciding BO. Besides Lamme's experiments show that constructing a local model makes sense, tensors' being defined as "small meaningful generic tokens" encourages to define local visual structures as tensors.

Defining meaningful visual structures is still an arguable issue as the studies concentrated on measuring visual stimulation mostly focus on pixels, on the other hand there exist studies claiming that the whole boundary takes role on deciding BO information. Besides all local-global or pixel-border discussions, it makes sense to adapt unsupervised approach for BO problem in other ways. Gestalt visual cues are mostly threshold-based features as they provide discrete decisions about BO, simply as follows:

$$label(x) = \begin{cases} 1, & \text{if } x > threshold. \\ -1, & \text{otherwise.} \end{cases} \quad (4.5)$$

With respect to both Lamme's view and Gestalt cues argued above, an unsupervised, pixel-based algorithm seems to be an appropriate solution for estimating BO information. That's why, a pixel-based voting approach is followed in the latter period of the thesis study,. TV is the best choice of voting algorithms, despite a few changes are required to adapt it to the BO problem. Modifications on voting and representation has led to a new voting algorithm called Iterative Vector Voting, which is discussed in the following section.

4.3.1 Iterative Vector Voting

Tensors are utilized for data representation with the purpose of perceptual organization [23]. According to the original 2D tensor framework, tensors can represent two different perceptual structures:

- Non-oriented input: point, which is represented by ball tensor
- Oriented input: curvel, which is represented by stick tensor

Tensors can feed each other by propagating their information, which has two types:

- Orientation information, which is encoded by tensor shape
- Feature saliency, which is encoded by tensor size

While defining BO problem, it is realized that vectors, which are tensors of rank 1, are sufficient for data representation. BO information, whether region-based or pixel-based, should consist of two different types of information described as follows:

- Direction: It defines the direction of the owner region of the boundary.
- Magnitude: It defines how strongly the region owns the border.

Basic data representation and low computation cost of voting are two significant features of TV. Besides, the other very important advantage of TV is the way of creating voting fields. Saliency decay function, which assigns the votes by decreasing by distance and orientation, is the associate function used both by ball and stick tensors. On the other hand, border ownership information does not diminish on smooth borders unless sudden orientation changes occur. An example for this situation is shown in Figure 4.9.



Figure 4.9: Example - impact of orientation on border ownership

There exist two yellow-labeled borders in Figure 4.9, which are sequentially located at the bottom-left and bottom-right corners. The first one, corresponding to the front side of the house, is a smooth border where orientation is stable. There exists a T-junction on this border, which is colored as black and located between wood floor, blue floor and the house. The information is delivered to the whole boundary without any loss as there exist neither a change on orientation nor another strong information such as a junction. On the other hand, another T-junction exists on the border of the second object, which is located on the bottom-right side and colored black. BO information of T-junction cannot be propagated to the whole border, since the border has many recesses and ledges, *i.e.*, there exist many orientation changes.

As the example above shows, BO information depends on orientation together with distance. If there exists sudden orientation changes on the border, the information decreases by distance. Thus, voting algorithm for BO must be different from the original tensor voting framework. In the original algorithm, votes are aggregated by both distance and orientation, which is formulated as follows:

$$DF_{TV}(s, \kappa, \sigma) = e^{-\frac{s^2 + c\kappa^2}{\sigma^2}}, \quad (4.6)$$

where s is the arc length between two tensors, κ is the curvature, c is a variable to control the degree of decay with curvature and σ is the scale of voting, which are denoted in Figure 4.10. In this figure, the vote of a stick tensor located on T1 to the location T2 is visualized. The vote is also a stick tensor with different orientation and magnitude. θ is the angle between the tangent of the osculating circle at the voter tensor, as l defines the distance between T1 and T2.

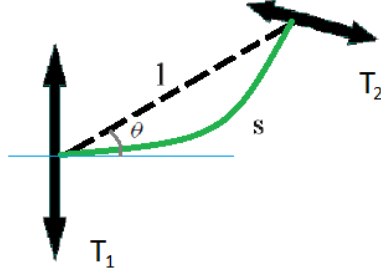


Figure 4.10: Voting schema of stick tensor

To attenuate votes by orientation and distance together for IVV, the solution is simple: the distance parameter is removed, the scale of voting is decreased and voting is iterated. Thanks to such a solution, it is succeeded that distance cannot show any effect on border ownership information without any orientation change. Thus, decay function of IVV becomes:

$$DF_{IVV}(\kappa, \sigma) = e^{-\frac{c\kappa^2}{\sigma^2}}. \quad (4.7)$$

We adapt the voting algorithm from stick voting algorithm of TV completely. The reason for such an adaptation is that stick voting keeps the information of curve continuation. It emits the maximum vote to the boundary curve which is predicted to continue from these locations. Thus, our voting function, which depends on both orientation and saliency decay function, is defined as follows:

$$Vote(\theta, \sigma) = DF_{IVV}[-\sin(2\theta)\cos(2\theta)]^T[-\sin(2\theta)\cos(2\theta)]. \quad (4.8)$$

Due to the change in voting functions, voting fields of TV and IVV differ. Distance is not considered in IVV, besides voting scale is smaller (Figure 4.11).

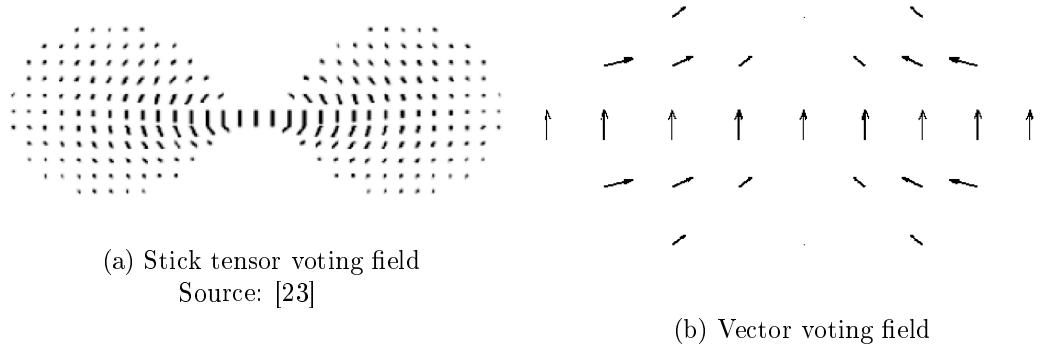


Figure 4.11: Comparison of voting fields of TV & IVV

Vectors are more compact versions of tensors, carrying less information. Thus, their geometric representations are not different than tensors, shown in Figure 4.12. In this figure, vector V_1 and its vote at V_2 , which is a vector too, are visualized. Here orientations of both V_1 and V_2 show the same region, that represents BO, but their saliences are different due to vote attenuation.

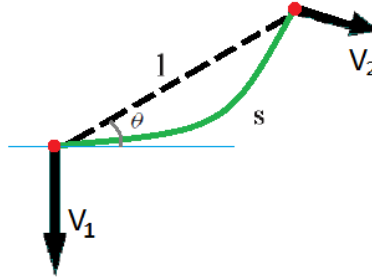


Figure 4.12: Voting schema of a visual cue vector

Defining meaningful visual data structures is an arguable issue. But in this study, pixel-based approach is preferred with respect to two reasonable assumptions:

- As the scale of voting in Iterative Tensor Voting has to be smaller compared to Tensor Voting due to its iterative nature, pixel is a better choice to prevent vote loss.
- BO studies, mostly working on visual areas V1 and V2 [12, 13, 26, 41] measure stimuli changes point-by-point on image, which are quite similar

to the pixel-based approach of IVV.

Additionally, in IVV, there is no need for dense voting as the votes on the borders are enough to assign the BO information. Under these assumptions and rules, IVV algorithm is explained in details in the following section.

4.3.1.1 The Algorithm for IVV

The IVV algorithm consists of three main phases, which are visualized in the sequence diagram of Figure 4.13.

- Cue Extraction
- Curvature Extraction
- Voting

The inputs of the algorithm, as usual, are the original image and relevant boundary image. Initially, both visual cues and curvature map are extracted from these. Visual cues are represented by vectors: their magnitudes show the saliency of cue, *i.e.*, the magnitude of border ownership information while their directions show the owner region of the border pixel.

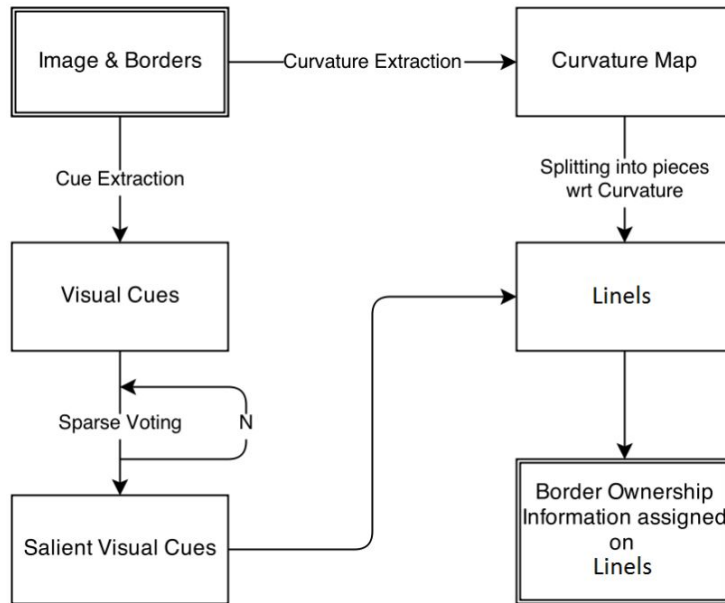


Figure 4.13: Iterative Vector Voting algorithm schema

Simultaneously with cue extraction, curvature map is used for determining *linels*. *Linell* is not a word in English, it is just inspired by *curvel* term. Although *curvel* word does not have a formal definition either, it is firstly declared by Medioni [23] as "perfectly oriented point". Thus, curvel can be considered as the smallest, meaningful piece of curve. In the light of this definition, linel is defined as "largest smooth piece of line". Linels are simply extracted by cutting the border from points having curvature of local maxima. An example image of linels, which are colored differently, is shown in Figure 4.14.

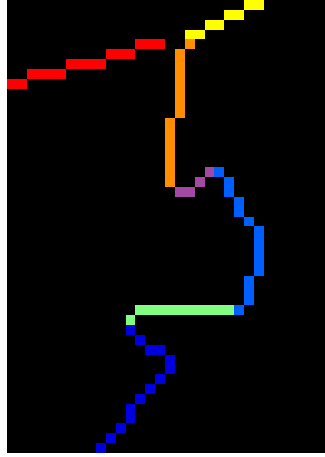


Figure 4.14: A set of linels extracted from an image. Each linel is shown with a different color.

After visual cue vectors and curvels are extracted, the voting process comes next. Visual cue vectors propagate their border ownership information iteratively to each other, in the limits of pre-defined saliency decay function.

After the voting procedure, all border pixels have more reliable, enriched BO information. However, as in the original TV framework, insignificant votes are to be eliminated. For such purposes, thresholding via local maxima points are used. These points are found as usual as follows:

x is local maxima for the function f if it succeeds:

- $f'(x) = 0$
- $f''(x) < 0$

Thresholding by local maxima means just holding the vectors of local maxima and eliminating all others. Reason for such an elimination is to choose pixels of salient features, to create an HVS-like model. It is argued that early vision of HVS initially uses salient, attention-taking points instead of examining the whole boundary [43].

Finally, after thresholding insignificant votes, the BO information is assigned to each line simply by addition of all cue vectors belonging to the line. Resulting vector of addition provides us the joint decision of salient cue vectors about border ownership.

Details of each phase of the algorithm are provided below under the titles of Curvature, Visual Cues and Voting.

- **Curvature**

The method for calculating curvature is simple: It first fits a circle to the origin point and its neighbors, then calculates the analytical curvature of the origin this circle.

The distance of neighborhood, which determines the right and left neighbors of the origin point, is learned through the observations. It is chosen as 12 with respect to these observations, which means that there should be 12 pixels on the border, between the origin and the neighbors on which the circle is placed. We define the curvature as $\frac{1}{r}$ where r is the circle radius. In other words, the curvature on point_{*i*} equals to:

$$C_i = \frac{1}{r},$$

where r is the radius of circle of points $(p_{i-12}, p_i, p_{i+12})$, visualized in Figure 4.15

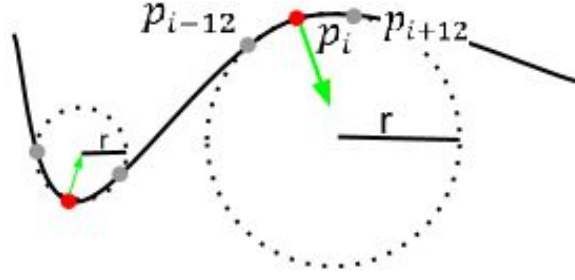


Figure 4.15: Curvature on border

p_{i-12} and p_{i+12} are right and left neighbors of the origin point (p_i). The green-colored arrow shows the direction of the curvature at the same point.

- **Visual Cues**

The method consists of four pixel-based visual cues, one of which are spatial and other three are spectral, respectively as follows:

- T-junctions
- Curvature
- Entropy
- Contrast

These visual cues are represented by vectors for each border pixel. In Figure 4.16, how a visual cue is converted to a vector is visualized.

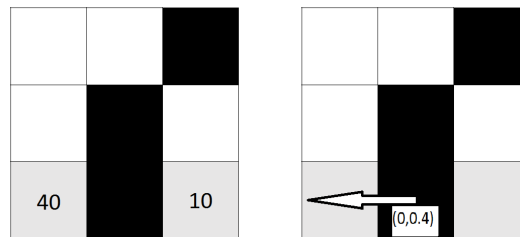


Figure 4.16: Pixel-based contrast tensor

Each visual cue has its own map, where pixels are assessed with their cue values. The ratio of cue values on both sides of the border defines the magnitude of cue vector, as the vector perpendicular to the border yields the direction.

Assume the map of contrast cue vectors are extracted. Contrast is an important distinguishing visual cue, and the rule of BO related to contrast can be simply defined as follows: Regions having stronger color, *i.e.* of higher color intensity, are mostly the foreground, as Hoiem et al. [10] state.

In Figure 4.16, pixels lying on both sides of the border, which are colored gray, have gray-scale colors of 40 and 10, respectively. These pixels are briefly called as neighbor pixels. With respect to pixel-based BO approach, one of these neighbor pixels is the owner of the border pixel between them. This border pixel, which is colored black, is called the *seed pixel*.

With respect to the contrast rule, the neighbor pixel with value 40 is the owner of the seed pixel. Thus, the direction of the visual cue vector is perpendicular to the border, towards the region of this neighbor pixel, as shown in Figure 4.16. The magnitude on the other hand, which represents the information saliency, is measured by the ratio of gray scale values. Greater difference on cue values means higher saliency. The magnitude is scaled down to the range [0-1]. It means that a ratio higher than 10 is considered a vector with magnitude of 1, corresponding to the most reliable information.

Entropy cue follows the same way with the contrast cue. A region with more textured structure is more probable owner of the border, compared to the coarse, plain one [10, 36]. More textured structure means higher energy, *i.e.*, higher entropy, thus entropy of 3x3 neighborhood is calculated for each neighbor pixel, as defined in Section 4.1. As T-junctions indicate very salient BO information, they are represented by vectors of magnitude 1, which is the maximum. Thus, they transmit very reliable BO decision to their neighbors.

As the curvature map is calculated earlier in the line extraction procedure (Figure 4.15), curvature values and directions are directly used from this map when constructing curvature cue vectors. Lower-region cue can also be considered among successful pixel-based cues. However, as it provides discrete BO decision, lower region cue can not be utilized for voting.

- **Voting**

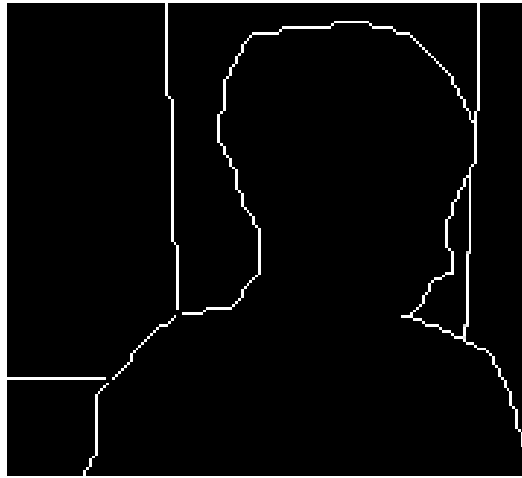
Four separate cue vector maps are constructed by assigning four different cue vectors to every border pixels. After these maps are obtained, five step voting algorithm are applied sequentially as follows:

1. Apply iterative voting algorithm for each cue map separately, each with a specific number of iterations
2. Apply local maxima thresholding to these votes, to acquire the most reliable updated votes
3. Calculate the vector sum on each linel for each cue
4. Apply majority voting on each linel to gather an associate decision
5. Apply second majority voting on BO decisions of linels to assign a BO label to the whole boundary

All steps of IVV are visualized in Figure 4.17. In Figures 4.17a and 4.17b, a sample image and its border mask are shown. In Figures 4.17c and 4.17d, a map of contrast vectors and image of voted cue vectors are given respectively. It is observed that the number of salient vectors arise around right shoulder of the child, as there exist two T-junctions and many contrast & entropy vectors exist and show the same region as the owner. Thresholded, associate BO vectors are shown in Figure 4.17e whereas finalized BO decisions are provided with different colors in Figure 4.17f. Correctly assigned borders are colored green while wrongly labeled ones are shown as red. Besides, the border with white color represents that no BO value is assigned here in the ground truth, due to a visual conflict.



(a) A sample image



(b) Border image



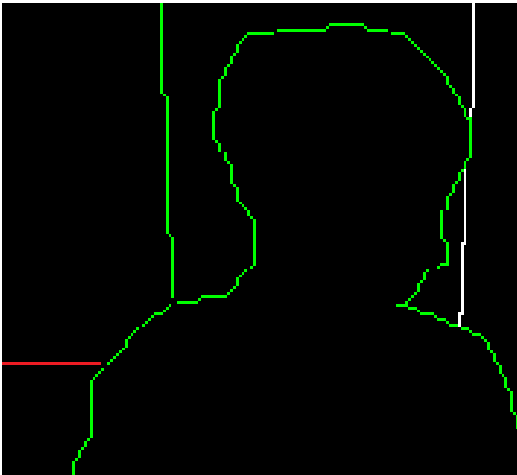
(c) Contrast cue map



(d) Voted cue Map



(e) Associate revised cue map



(f) Final BO labeling

Figure 4.17: Sample IVV scenario with the results

CHAPTER 5

EXPERIMENTS

In this chapter, three separate experiments have been conducted:

- **Evaluation of Visual Cues:** In this phase, BO decision accuracies of each visual cue are measured. Besides, their contributions to both supervised and unsupervised models are examined.
- **Experiments of CRF-based model:** In this phase, success rates of CRF cues are provided and evaluated, both separately and totally.
- **Experiments of IVV model:** Results of four pixel-based IVV cues are examined, both in different combinations and as a whole.

In Appendix A, the BO dataset and web page are described in details. At last, all results of each phase are provided in the following sections.

5.1 Dataset and Evaluation

The dataset used for the experiments is the largest of all [1]. Although the whole dataset consists of 1003 images actually, 884 of these are utilized due to their reliable drawings & labellings.

Other properties of the dataset:

- 440 indoor images

- 404 outdoor images
- 18647 boundaries

Since the users could not have given BO decision of 544 borders, these borders are excluded from the experiments.

Evaluation of results are handled through accuracy calculation except for T-junctions. Accuracy (ACC) is the proportion of true results in the whole population. Performance of T-junctions is measured as precision value, *i.e.* positive predictive value (PPV), to provide a reasonable comparison with other cues. It is because a border may not own a T-junction. ACC and PPV are computed as follows:

$$ACC = \frac{\# \text{ of TP} + \text{TN}}{\# \text{ of TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (5.1)$$

$$PPV = \frac{\# \text{ of TP}}{\# \text{ of TP} + \text{FP}}, \quad (5.2)$$

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

5.2 Evaluation of Visual Cues

In Table 5.1, average accuracies of region-based features estimating BO are provided.

Size and lower-region cues provide accuracies close to random (50%). But their contributions to the region-based model are much more than these values. Besides, the most successful cue is T-junction as expected.

Table 5.1: Region-based cue accuracies on BO

Visual Cue	Performance Value	Performance Type
T-junction	74,7 %	PPV
Convexity	66,0 %	ACC
Entropy	63,6 %	ACC
Contrast	60,7 %	ACC
Size	54,4 %	ACC
Lower Region	52,0 %	ACC

5.3 Experiments on the CRF-based model

In Table 5.2, the results of the CRF-based model are provided in different combinations of cues. T-junctions are used with the purpose of prior labeling as this cue is quite successful at assigning BO information alone and it is required that a prior label is assigned before CRF update.

Table 5.2: Accuracies of cue combinations on CRF model

Visual Cue	Accuracy
T-junction	36,7 %
T-junction + (Convexity-Entropy)	41,3 %
T-junction + (Contrast-Entropy)	57,1 %
T-junction + (Convexity-Entropy-Contrast)	59,4 %
T-junction + (Convexity-Entropy-Contrast-Size-Lower Region)	68 %

These results reveal two things: Size and lower-region cues are more successful when utilized in a model with multiple cues rather than alone, as the performance improvement of 9% shows. Secondly, how to combine cues is important as different combinations of convexity-entropy-contrast cues reveal. In this table, success ratio of T-junction is given as accuracy, in contradiction to previous results, to see the contributions of other cues clearly.

5.4 Experiments on the IVV model

In the IVV model, initially each cue gives a BO decision itself, then the decisions of all cues are combined to reach a consensus for each linel via weighted majority

voting. Thus it is easier to get separate accuracies for each cue of IVV.

IVV needs the parameter of "number of iterations" to be pre-defined, as a discrete value. Through several experiments of different iteration numbers with each cue separately, results of which are given in Table 5.3), the ideal parameters are found.

Table 5.3: Accuracies of separate cue contributions on the IVV model

Visual Cue / No of Iter.	2	3	4	5
T-junction	72,2%	73,1%	73%	71,8%
Curvature	57,4%	58%	56,7%	59,4%
Entropy	66,7%	71%	65,4%	58,3%
Contrast	64%	68%	64,2%	60,7%

Even a small loss is encountered after the 4th iteration for T-junction cue, iteration number for this cue does not effect the performance. Iteration numbers of contrast and entropy cues are chosen as 3, whereas it is 5 for the curvature cue, as their accuracies are maximum at these numbers of iteration. Another deduction from Table 5.3 is that cue accuracies of IVV are much higher than region-based performances of same cues. Reason of such an improvement is that salient features can be suppressed on region-based approach due to averaging on whole border. Why it gives a better result when the number of iterations is higher can be explained with the same situation while calculating curvature as best results of curvature values are gathered within a neighborhood of 9. With respect to the cue combination of these parameters, the IVV model has provided a performance with a total accuracy of **77%**.

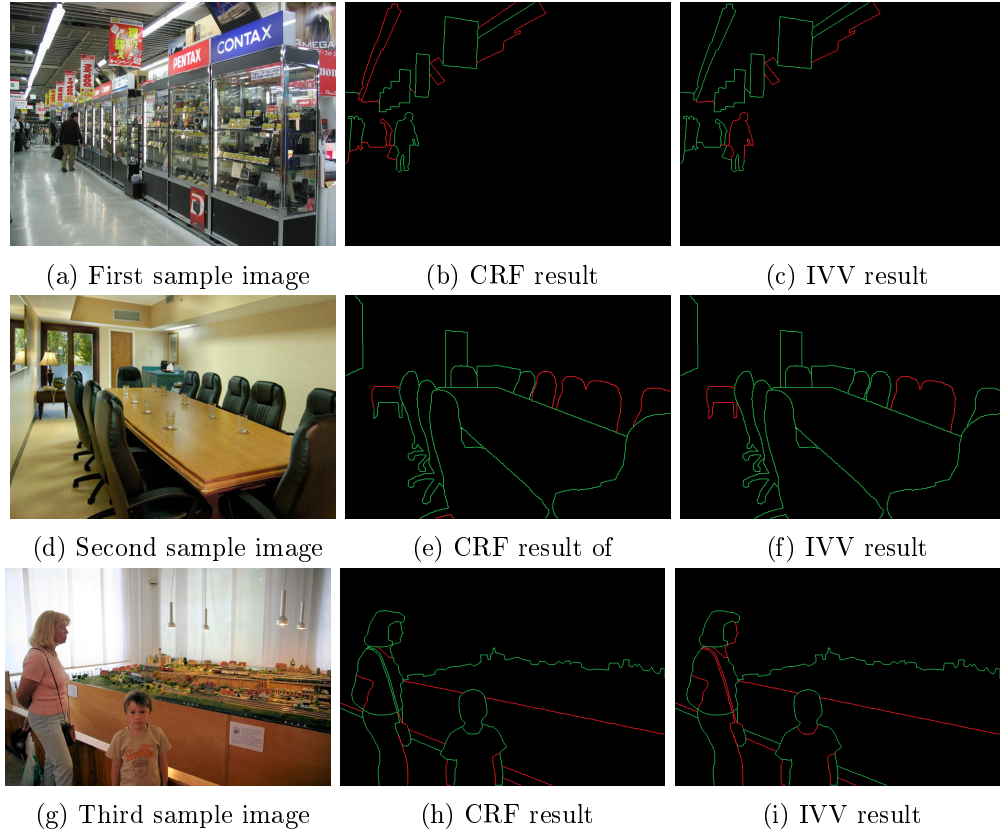


Figure 5.1: Sample visual results of CRF and IVV models

Sample visual results for both CRF and TV models are provided in Figure 5.1. In the first column, original images are given. In the second and third rows, CRF and IVV model results of these images are provided respectively. Correctly labeled borders are colored green, while the wrong ones are colored red in these result images. For the first image, IVV algorithm has provided a better result while in the second image one, CRF model is more successful. Such a discrepancy of success arises from the approach. CRF-model becomes more successful at the images showing more tendency to region-based cues such as area. Lastly in the third image, the results of both algorithms are quite successful due to high number of T-junctions.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This thesis propose two new models for the border ownership (BO) problem by utilizing a supervised and an unsupervised model on a new, comprehensive BO dataset using Gestalt cues such as size, convexity, entropy, contrast, lower-region and T-junctions.

There are just two studies of BO to compare results against, which show results above the average and work on real images. The first study is the study of Ren, Fowlkes and Malik, in which shapemes are defined as image-specific cues [35]. Results of the study is 72% accuracy for human-marked segmentation set while 64% accuracy is obtained for the dataset of which segmentations are automatically generated curves. The results of the thesis are much better, although a precise comparison is impossible to make as the datasets are different.

The other one is the study of Leichter et al. [19], in which a 2.1D model is applied using T & L junctions, convexity, lower region, fold/cut and parallelism as cues. 2.1D model provides ordinal depths of regions, which is a great prior for BO assignment. This study briefly models the PDF ordinal depth information and maximizes it by a CRF model utilizing six different cues. Its performance is measured as 82,8% accuracy, which is the highest score so far. But a disadvantage of this study is that the whole dataset is outdoor images. 2.1D model suits better to outdoor data due to its characteristics, thus same success is not expected for indoor imagery.

Although both CRF and IVV models have produced promising and successful

results, there are more things to consider as the future work:

- All visual cues have the same amount of impact on the BO decision. It seems logical that an auto-weighting algorithm should work; however, there does not exist any study in the literature which defends or opposes this aspect.
- The scale of voting and iteration numbers are not assigned automatically. They are chosen by assumptions and observations.
- IVV is a pixel-based approach, thus even very small inaccurate drawings of borders cause great differences in results. There exist some solutions such as selecting further neighbors instead of neighbors with distance 1, but this is not a reliable solution.
- Cue combinations are handled with respect to the results gathered. Here in this procedure, physiological studies and experiments can be utilized by measuring their impacts on HVS and constructing a reasonable heuristic.

REFERENCES

- [1] Border ownership labelling program. <http://www.kovan.ceng.metu.edu.tr/bo/>. Accessed: 2013-08-06.
- [2] Rubin's vase. http://en.wikipedia.org/wiki/Rubin_vase. Accessed: 2014-02-02.
- [3] Visual examples of gestalt principles (1). <http://www.psypress.co.uk/mather/resources/topic.asp?topic=ch09-tp-01>. Accessed: 2014-02-02.
- [4] Visual examples of gestalt principles (2). <http://www.slideshare.net/GlennaShaw/the-gestalt-of-slides>. Accessed: 2014-02-02.
- [5] M. Akkus, G. Topuz, B. Ozkan, and S. Kalkan. A comprehensive database for border ownership. In *Signal Processing and Communications Applications Conference (SIU), IEEE 21st*, pages 1–4, April 2013.
- [6] M. K. Albert. Cue interactions, border ownership and illusory contours. *Vision Research*, 41(22):2827–2834, 2001.
- [7] P. Arbelaez, C. Fowlkes, and D. Martin. The berkeley segmentation dataset and benchmark. see <http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds>, 2007.
- [8] J. S. Bakin, K. Nakayama, and C. D. Gilbert. Visual responses in monkey areas v1 and v2 to three-dimensional surface configurations. *The Journal of Neuroscience*, 20(21):8188–8198, 2000.
- [9] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. In *NIPS*, volume 2, page 3, 2000.
- [10] D. Hoiem, A. Efros, and M. Hebert. Recovering occlusion boundaries from an image. *International Journal of Computer Vision*, 91(3):328–346, 2011.
- [11] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1):185–203, 1981.
- [12] J. F. Jehee, V. A. Lamme, and P. R. Roelfsema. Boundary assignment in a recurrent network architecture. *Vision Research*, 47(9):1153–1165, 2007.
- [13] M. Kikuchi and Y. Akashi. A model of border-ownership coding in early vision. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Artificial Neural Networks — ICANN 2001*, volume 2130 of *Lecture Notes in Computer Science*, pages 1069–1074. Springer Berlin Heidelberg, 2001.
- [14] H. Komatsu. The neural mechanisms of perceptual filling-in. *Nature Reviews Neuroscience*, 7(3):220–231, 2006.

- [15] S. Kumar. *Models for learning spatial interactions in natural images for context-based classification*. PhD thesis, Carnegie Mellon University, 2005.
- [16] S. Kumar. Discriminative graphical models for context-based classification. In R. Cipolla, S. Battiato, and G. M. Farinella, editors, *Computer Vision*, volume 285 of *Studies in Computational Intelligence*, pages 109–134. Springer Berlin Heidelberg, 2010.
- [17] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML, pages 282–289, San Francisco, CA, USA, 2001.
- [18] M. Lee and G. Medioni. Inferred descriptions in terms of curves, regions and junctions from sparse, noisy binary data. In *Proc. IEEE Int. Symp. Computer Vision*, pages 73–78, 1995.
- [19] I. Leichter and M. Lindenbaum. Boundary ownership by lifting to 2.1d. In *International Conference on Computer Vision*, pages 9–16. IEEE, 2009.
- [20] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1253–1260. IEEE, 2010.
- [21] G. Medioni, C.-K. Tang, and M.-S. Lee. Tensor Voting: Theory and Applications. In *Proceedings of RFIA*, 2000.
- [22] P. Mordohai and G. Medioni. Stereo using monocular cues within the tensor voting framework. In T. Pajdla and J. Matas, editors, *Computer Vision - ECCV 2004*, volume 3024 of *Lecture Notes in Computer Science*, pages 588–601. Springer Berlin Heidelberg, 2004.
- [23] P. Mordohai and G. Medioni. Tensor voting: a perceptual organization approach to computer vision and machine learning. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2(1):1–136, 2006.
- [24] K. Murphy. An introduction to graphical models. *A Brief Introduction to Graphical Models and Bayesian Networks*, 10, 2001.
- [25] M. Nicolescu and G. Medioni. Motion segmentation with accurate boundaries: A tensor voting approach. In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR’03, pages 382–389, Washington, DC, USA, 2003. IEEE Computer Society.
- [26] H. Nishimura and K. Sakai. Determination of border ownership based on the surround context of contrast. *Neurocomputing*, 58–60(0):843–848, 2004.
- [27] B. Ozkan and S. Kalkan. Extraction of border ownership information by conditional random field model. In *IEEE 21st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4, April 2013.
- [28] L. Pessoa, E. Thompson, and A. Noë. Finding out about filling-in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences*, 21(6):723–748, 1998.

- [29] M. A. Peterson. Object recognition processes can and do operate before figure-ground organization. *Current Directions in Psychological Science*, 3(4):105–111, 1994.
- [30] N. Qian. Binocular disparity and the perception of depth. *Neuron*, 18(3):359–368, 1997.
- [31] F. T. Qiu, T. Sugihara, and R. von der Heydt. Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience*, 10(11):1492–1499, 2007.
- [32] F. T. Qiu and R. Von Der Heydt. Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules. *Neuron*, 47(1):155–166, 2005.
- [33] H. Rashwan, M. Garcia, and D. Puig. Variational optical flow estimation based on stick tensor voting. *Image Processing, IEEE Transactions on*, 22(7):2589–2599, July 2013.
- [34] X. Ren, C. Fowlkes, and J. Malik. Cue integration for figure/ground labeling. In *Neural Information Processing Systems Conference*, 2005.
- [35] X. Ren, C. Fowlkes, and J. Malik. Familiar configuration enables figure/ground assignment in natural scenes. *Journal of Vision*, 5(8):344–344, 2005.
- [36] X. Ren, C. C. Fowlkes, and J. Malik. Figure/ground assignment in natural images. In *European Conference on Computer Vision-ECCV*, pages 614–627. Springer, 2006.
- [37] E. Rubin. *Visuell wahrgenommene figuren*. PhD thesis, Gyldendalske Boghandel, 1915.
- [38] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008.
- [39] K. Sakai and Y. Hirai. Neural grouping and geometric effect in the determination of apparent orientation. *Journal of the Optical Society of America A (JOSA A)*, 19(6):1049–1062, 2002.
- [40] M. Schmidt. Ugm: Matlab code for undirected graphical models. <http://www.di.ens.fr/~mschmidt/Software/UGM.html>. Accessed: 2014-02-02.
- [41] E. Seidemann and W. T. Newsome. Effect of spatial attention on the responses of area mt neurons. *Journal of Neurophysiology*, 81(4):1783–1794, 1999.
- [42] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics, 2003.
- [43] J. B. Subirana-Vilanova and W. Richards. Perceptual organization, figure-ground, attention and saliency. 1991.

- [44] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.
- [45] S. P. Vecera, E. K. Vogel, and G. F. Woodman. Lower region: a new cue for figure-ground assignment. *Journal of Experimental Psychology: General*, 131(2):194, 2002.
- [46] H. M. Wallach. Conditional random fields: An introduction. *Technical Reports (CIS)*, page 22, 2004.
- [47] M. Wertheimer. Laws of organization in perceptual forms. In *A source book of Gestalt psychology*, pages 71–88. Routledge & Keegan Paul, 1938.
- [48] B. Yao, X. Yang, and S.-C. Zhu. Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 169–183. Springer, 2007.
- [49] L. Zhaoping. Border ownership from intracortical interactions in visual area v2. *Neuron*, 47(1):143–153, 2005.
- [50] 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.

APPENDIX A

BORDER OWNERSHIP LABELING PROGRAM & DATASET

Due to the insufficient number or low quality of data, a new BO dataset is required. The project team (Asst. Prof. Sinan Kalkan, Mehmet Akif Akkuş, Gaye Topuz and Buğra Özkan) initially created an on-line annotation tool, with a user-friendly GUI and clear instruction set [1]. Via the annotation tool, a database has been generated which consists of 1003 images with its corresponding labeled borders.

The database consists of 503 indoor and 500 outdoor images in JPEG format. All of the outdoor images are taken from BSDS [7]. They all have resolution of 321x481. For the indoor dataset, 219 images are obtained from LHI database [48] and the other 284 ones are gathered from various image-sharing websites. The width for indoor images is same for all, which equals to 800 pixels. The height varies from 443 to 1343 pixels.

Annotated GT data was available with the LHI and BSDS datasets, as they are necessary for gathering the regions and their borders. For the other images, GT data is extracted through segmentation by hand. In continuation of the segmentation process, image borders are extracted, and the ones smaller than 4% of the length of image diagonal are eliminated. The simple goal for the attendant is to click either blue or red area on the GUI considering it as the owner of the white border between these two.

Usage scenario of the on-line annotation tool is simple and stepwise as the necessary information is provided before labeling. After entering the web site, initially

a register/login page appears as in Figure A.1. While registering, the attendant must also share the age, education level and gender information (Figure A.2) as any dependency analysis may be held with respect to these parameters.

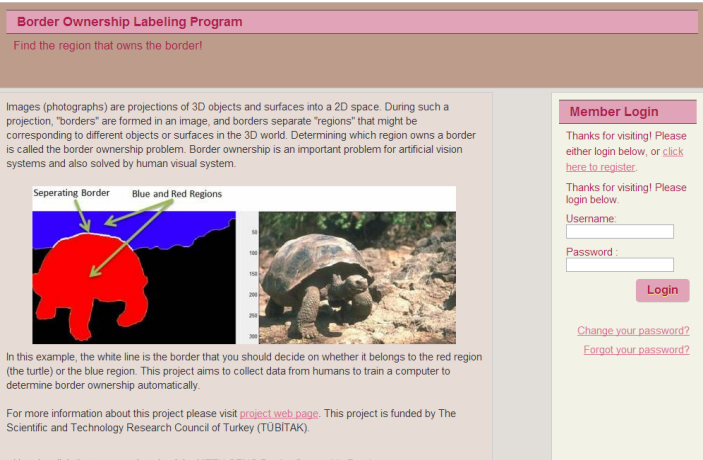


Figure A.1: Border ownership labeling program login/register page

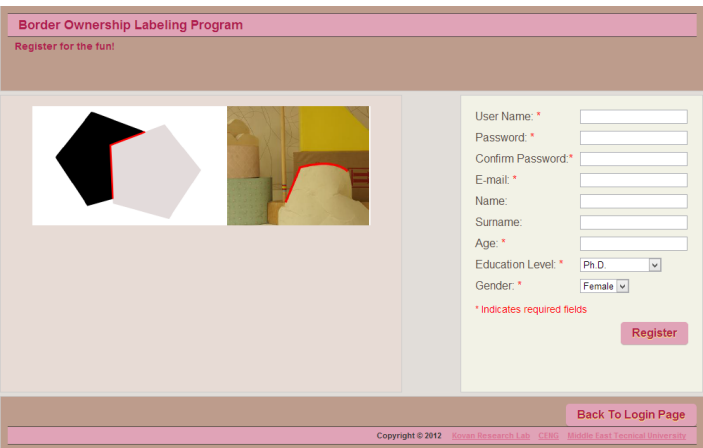


Figure A.2: Border ownership labeling program register page

After login procedure, the take-tour page is the next, where the test participant is informed by animations about BO problem and the usage of website (Figure A.3). Later on, before the labeling process, the user is tested to learn whether he/she understood BO problem by using a simple artificial data on the tutorial page (Figure A.4). After each correct labeling, the user is informed about why the border belongs to the area selected. The tutorial is a must before labeling starts. Unless the participant labels all borders correctly, he/she is not able to pass the tutorial and start labeling.

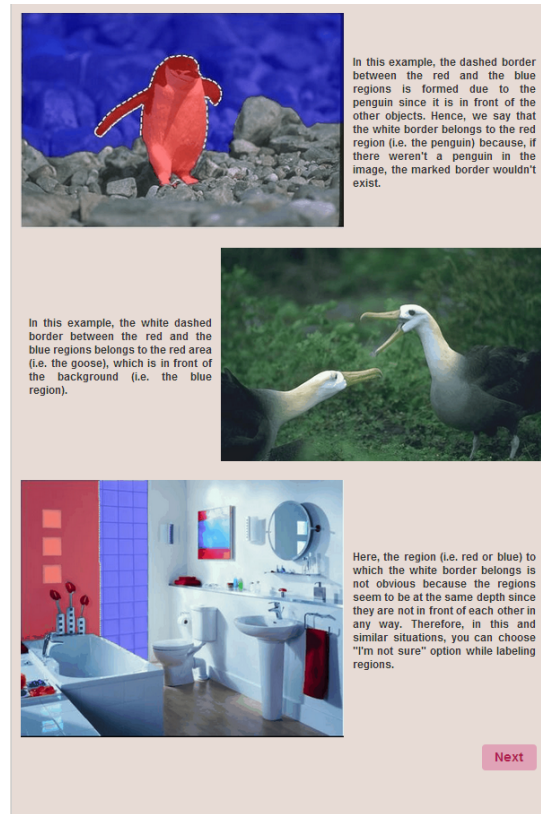


Figure A.3: Border ownership labeling program take-tour page

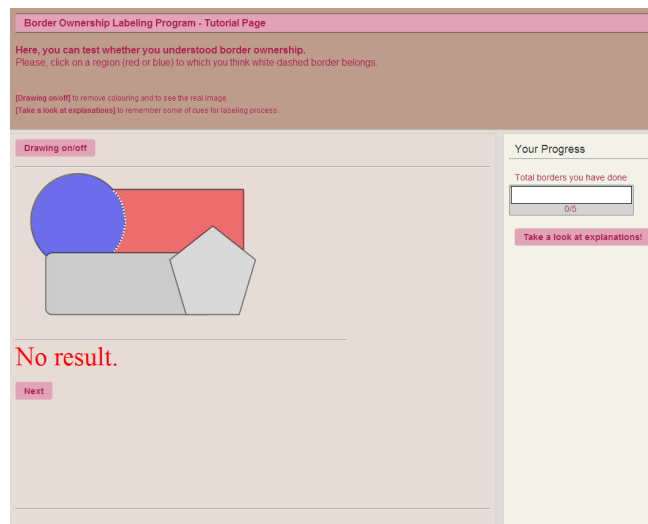


Figure A.4: Border ownership labeling program tutorial page

After tutorial part is completed successfully, labeling part starts. Two visual areas that causes occlusion are colored transparently as red and blue, respectively (Figure A.5). In case the test participant is not able to guess which region is

the owner of the border, he/she can pass the image by clicking "I am not sure" button. Besides, the participant can change his BO decision by clicking "Undo" button and disable the colored border layer by clicking "Drawing on/off" button.

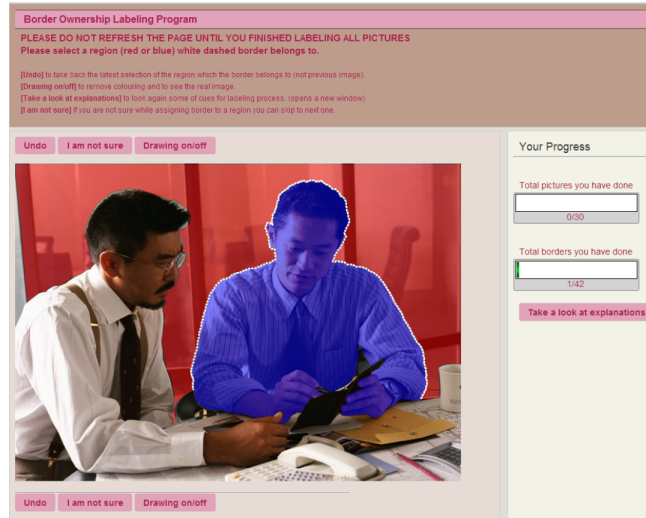


Figure A.5: Border ownership labeling program labeling page

For the data representation of borders & regions, JSON data type is used. In addition to the images, mat files, an SQL file and a read-me file are provided. Each mat file obtains the information of related image with its regions. SQL file consists of all labeling database. The read-me file provides all information to utilize the database.

In conclusion, the dataset consists of various kinds of real, indoor & outdoor images, with their correctly labeled BO GT masks, thanks to our on-line annotation tool. It is expected that both the tool and dataset will contribute to the literature more as the number of studies about BO problem increases.