Markov Random Field based Methods for Cluttered Scene Stereo

Fahim Mannan
School of Computer Science
McGill University

A thesis submitted for the degree of
M.Sc. in Computer Science
June 2010

Abstract

This thesis studies the performance of different Markov Random Field (MRF) based stereo formulations for cluttered scenes. Cluttered scenes have objects of a specific size distribution placed randomly in 3D space. Real-world examples of such scenes include forest canopy, bushes or foliages in general. One characteristic of such scenes is that they contain a lot of depth discontinuities and partially visible pixels. A natural question which is addressed in this thesis is how well the existing stereo algorithms perform for such scenes. The scenes used in some of the widely used benchmark dataset do not contain stereo pairs with dense clutter. Therefore, we use a cluttered scene model [1] to generate synthetic scenes with different scene parameters such as size and density of objects, and range of depth. In our experiments we apply algorithms with basic and visibility constraints. In the basic category we use: Expansion, Swap, Max Product Belief Propagation (BP-M), Sequential Tree Reweighted Message Passing (TRW-S) and Sequential Belief Propagation (BP-S) with different forms of data and smoothness terms. In the visibility constraint category we use: KZ1 and KZ2 proposed in [2, 3, 4]. The algorithms are applied to the input dataset with different parameter settings. To compare the performance, we consider the percentage of mislabeled pixels, errors in certain regions and the contribution of the errors in those regions to the total error. We also analyze the cause of those errors using the underlying scene statistics.

For the basic formulation, Potts model performs surprisingly well in all the experiments, in the sense that binocularly visible surface points are correctly labeled. In particular, Expansion, TRW-S, and BP-M perform equally well. Algorithms with visibility constraints also perform equally well for binocular pixels and in some cases slightly better than basic formulation.

We did not observe any clear improvement in labeling binocular pixels. However, visibility constraints perform largely better than basic formulation when all the pixels are considered. This is also reflected in the energy measure. Algorithms based on basic formulation shows large gap between the ground truth and output energy. However, formulations with visibility constraints have energy values closer to the ground truth. This is because the visibility constraint restricts the search space to disparity labels that are consistent. We conclude that methods like KZ1 can primarily improve labeling of monocular pixels. For binocular pixels, there is still room for improvement in both formulations, especially in the case of off-by-one errors (i.e. cases where the assigned labels differ from the ground truth by a single disparity).

Résumé

Ce mémoire vise à comparer la performance de différents modèles stéréo par champs aléatoires de Markov sur des scènes encombrées. Ces scènes sont composées d'objets dont les grandeurs suivent une distribution spécifique et dont les positions sont aléatoires dans l'espace 3D. Elles se caractérisent par la présence de plusieurs discontinuités de profondeur et d'occlusions partielles. Des buissons, des feuillages ou une forêt en sont des exemples. Les scènes de référence généralement utilisées ne contiennent pas d'images stéréo de scènes encombrées. Il nous apparaît donc important de vérifier la performance des algorithmes stéréo existants sur ce type de scènes. Par conéquent, nous utilisons un modèle de scènes encombrées pour générer des scènes synthétiques selon différents paramètres tels que la taille et la densité des objets, et l'intervalle des profondeurs. Nos tests appliquent des algorithmes avec contraintes de base et de visibilité. Parmi les algorithmes avec contraintes de base, nous utilisons: Expansion, Swap, Max Product Belief Propagation (BP-M), Sequential Tree Reweighted Message Passing (TRW-S) et Sequential Belief Propagation (BP-S) avec différentes fonctions de coût et de lissage. Parmi les algorithmes avec contraintes de visiblité, nous utilisons: KZ1 et KZ2 proposés dans [2, 3, 4]. Les algorithmes sont appliqués aux images synthétiques avec différentes valeurs de paramètres. Pour comparer la performance, nous considérons dans toute l'image le pourcentage de pixels erronés, c'est-à-dire n'ayant pas la bonne étiquette de profondeur, ainsi que le pourcentage d'erreurs dans certaines régions et sa contribution dans l'erreur totale. Nous analysons la cause de ces erreurs à l'aide des statistiques de la scène.

Pour les algorithmes avec contraintes de base, le modèle de Potts performe bien pour tous les tests, en ce sens que les points visibles dans les deux

images (pixels binoculaires) sont associés à la bonne étiquette. Plus particulièrement, les algorithmes Expansion, TRW-S et BP-M donnent des résultats similaires. Les algorithmes avec contraintes de visibilité donnent aussi de bons résultats pour les pixels binoculaires, mais sans amélioration significative. Par contre, ils performent beaucoup mieux lorsque tous les pixels sont considérés, ce qui se reflète aussi dans la mesure d'énergie. Les algorithmes avec contraintes de base donnent de grandes différences entre les énergies réelle et en sortie. Mais les algorithmes avec contraintes de visibilité donnent une énergie de sortie beaucoup plus similaire à l'énergie réelle, à cause des contraintes de visibilité qui restraignent l'espace de recherche des étiquettes de disparité. Nous concluons que les méthodes, comme KZ1, améliorent l'étiquetage des pixels monoculaires. Pour les pixels binoculaires, les algorithmes des deux catégories peuvent encore être améliorés, plus spécifiquement dans le cas où l'étiquette diffère de la valeur réelle d'une seule disparité.

Acknowledgements

I would like to thank my supervisor Michael Langer for his guidance, and support. He helped me immensely with the thesis by discussing the ideas, patiently reading through all the drafts and giving useful feedback.

Special thanks to Vincent Chapdelaine-Couture for translating the abstract into French. I would also like to thank all my friends in Montreal for their friendship and support.

Last but not least, I would like to thank my parents and my sisters for their constant support and encouragement.

Contents

List of Figures vii

List of Tables xi

1 Introduction 1

1.1 Stereo Reconstruction . . . 1
1.2 Cluttered Scene Stereo Reconstruction . . . 1
1.3 Modeling the Stereo Vision Problem . . . 2
1.4 Objective . . . 2
1.5 Contributions . . . 2
1.6 Motivation and Application Areas . . . 2
1.7 Outline . . . 3

2 Background Review 5

2.1 MAP-MRF formulation for Vision . . . 5
2.2 Graph Cut . . . 8
2.2.1 Graph Construction for Binary Labeling . . . 8
2.2.2 Move-Making Algorithms . . . 10
2.2.2.1 Expansion . . . 10
2.2.2.2 Swap . . . 11
2.3 Belief Propagation and its Variants . . . 12
2.4 MRF based Stereo Algorithms . . . 14
2.4.1 Basic Formulation . . . 14
2.4.2 Additional Constraints . . . 15
2.5 Comparative Study of MRF-based Algorithms . . . 18

3 Modeling Cluttered Scenes **21**
3.1 General Scene Models . . . 22
3.2 Basic Assumptions and Notations . . . 22
3.3 Probability of Disparity . . . 25
3.4 Binocular Visibility . . . 26
3.5 Joint Probability of Disparity . . . 27
3.5.1 Probability of Closer Neighbor $p(z' < z)$ . . . 27
3.5.2 Probability of Equidistant Neighbor $p(z = z')$ . . . 28
3.6 Models Derived from the Joint Probability Model . . . 28
3.6.1 Probability of Discontinuity $p(f_p \neq f_q | f_p)$ . . . 28
3.6.2 Probability of Difference $p(f_p - f_q)$ . . . 29
3.7 Discussion . . . 29

4 Scene Generation **31**
4.1 Choice of Parameters . . . 32
4.2 Synthetic Scene Generation . . . 34
4.2.1 Rendering Stereo Pairs . . . 34
4.2.2 Generating Ground Truth Disparity Map . . . 36
4.2.3 Generating Scale-Invariant Scenes with fixed γ . . . 37
4.3 Scene Generation Experiments . . . 37
4.4 Discussion . . . 42

5 Performance Evaluation **47**
5.1 Algorithms and Choice of Parameters . . . 48
5.2 Performance Criteria . . . 49
5.3 Experimental Results . . . 51
5.3.1 Formulation with Basic Constraints . . . 55
5.3.2 Formulation with Visibility Constraint . . . 59
5.4 Discussion . . . 62

6 Conclusion **79**
6.1 Summary of Our Approach . . . 79
6.2 Summary of Observations and Conclusions . . . 80
6.3 Contributions . . . 81

6.3.1 Comparative Study of Cluttered Scenes 81
6.3.2 Classifying Cluttered Scenes 81
6.3.3 Synthetic Cluttered Stereo Pair Generation 82
6.3.4 Best Performing Parameters 82
6.4 Issues and Open Questions 82
6.5 Future Work 83

A List of Results **85**
A.1 Results for $T_x = 0.1$, $\gamma = 0.1$ and $r = 0.1$ 85
A.1.1 Scene Statistics 85
A.1.2 Error Statistics 85
A.2 Error Contribution Plots 89
A.3 Total Errors (> 1) 95

References **97**

List of Figures

1.1 Sample range image of a forest scene from [5] 3

2.1 s-t min cut on directed graph 9
2.2 a) Initial Configuration b)Swap, and c) Expansion 11
2.3 a) Binocular visibility, b) $< p, l >$ blocks q from seeing the point $< q, l' >$. Therefore $\{< p, l >, < q, l' >\} \in I_{vis}$. 16

3.1 a) Top view of scene showing different scene parameters b) Minkowski sum of a ray and a square of half-width. The square in (a) is shrunk to a point. 24

4.1 a) Scene Generation (Top View) and b) Aligning camera with the pixel grid 35
4.2 Disparity Computation 36
4.3 1) Samples from the generated scenes (left view) and 2) the corresponding disparity map. Red denotes monocular pixels and magenta monocular pixels that are outside the view volume of the other view. 39
4.4 Percentage of binocular pixels for all the scenes. Red center line denotes the median, the lower and upper bounds of the box represents 25 and 75 percentile of the data. The red '+' denotes the outliers and the whiskers denote the extent of the data. 40
4.5 Disparity and Neighbor Statistics for Data and Model with $\gamma \approx 0.54$ with depth range 2 to 8 42
4.6 Disparity and Neighbor Statistics for Data and Model with $\gamma \approx 0.54$ and depth range 8 to 32 43

4.7 Disparity and Neighbor Statistics for Data and Model with $\gamma \approx 0.1$ and z-range between 2 and 8 44

4.8 Disparity and Neighbor Statistics for Data and Model with $\gamma \approx 0.1$ and z-range between 8 and 32 45

5.1 Mean error rate for scenes with only a textured background. Only the methods with non-zero error rate are shown in the figure. In the first row Expansion, Swap, BP-M, TRW-S, BP-S are colored red, green, blue, magenta and cyan respectively. $k_s = 1$ and 2 are represented by solid and dashed lines respectively. V_{max} values 1,2,10 and 100 are represented by O, *, □ and ⋄ respectively. In the second row, Expansion and KZ2 are colored red and blue respectively. The rest of the notations are the same. The λ scales are different between the two sets as was mentioned in Sec. 5.1 53

5.2 Comparison between $k_d = 1$ and 2 for (a) scene 1a and (b) scene 3b. The first row in each case is for the basic formulation and the second one for the additional constraints. 54

5.3 1) $-\log p$ of joint probability for scene 1a. 2) $V_{max} = 1$ or Potts model. 62

5.4 Non-zero error (i.e. error > 0) statistics for methods using the basic formulation. Only errors between 0 – 40% for Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ⋄ respectively, are shown. 65

5.5 Non-zero error (i.e. error > 0) statistics for methods using visibility constraints. Only errors between 0 – 40% are shown for $k_s = 1$, $k_d = 2$ and $V_{max} = 1$. Expansion, KZ1, KZ2 are colored red, green, and blue respectively. 66

5.6 Error statistics of mislabeled pixels that differ by exactly 1 from the ground truth (off-by-one error) for basic formulation. Only errors between 0 – 35% for Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ⋄ respectively, are shown. 67

5.7 Error statistics of mislabeled pixels that differ by exactly 1 from the ground truth for visibility formulation. Only errors between 0 – 35% are shown for $k_s = 1$, $k_d = 2$ and $V_{max} = 1$. Expansion, KZ1, KZ2 are colored red, green, and blue respectively. 68

5.8 Binocular monocular error statistics for basic formulation. Only errors between 0 – 16% for Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ⋄ respectively, are shown. . . 69

5.9 Binocular monocular boundary error statistics for additional constraint. Only errors between 0 – 16% are shown for $k_s = 1$, $k_d = 2$ and $V_{max} = 1$. Expansion, KZ1, KZ2 are colored red, green, and blue respectively. . . . 70

5.10 Binocular discontinuity error statistics for basic formulation. Only errors between 0 – 20% for Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ⋄ respectively, are shown. . . 71

5.11 Binocular discontinuity error statistics for additional constraint. Only errors between 0 – 20% are shown for $k_s = 1$, $k_d = 2$ and $V_{max} = 1$. Expansion, KZ1, KZ2 are colored red, green, and blue respectively. . . . 72

5.12 Binocular continuity error statistics for basic formulation. Only errors between 0 – 35% for Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ⋄ respectively, are shown. . . 73

5.13 Binocular continuity error statistics for additional constraint. Only errors between 0 – 35% are shown for $k_s = 1$, $k_d = 2$ and $V_{max} = 1$. Expansion, KZ1, KZ2 are colored red, green, and blue respectively. . . . 74

5.14 Energy statistics for basic formulation. Only Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ⋄ respectively, is shown. The ground truth is represented by the black curves. 75

5.15 Energy statistics for additional constraint based formulations. The colors used for output energy for Expansion, KZ1, KZ2 are red, green, and blue respectively. For ground truth energy the corresponding colors are cyan, magenta, and black. 76

5.16 All pixel (i.e. both binocular and monocular) non-zero error statistics for methods using the basic formulation. Only errors between 10 – 90% for Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ⋄ respectively, are shown. 77
5.17 All pixel (i.e. both binocular and monocular) non-zero error statistics for methods using additional constraints. Only errors between 5 – 90% are shown for $k_s = 1$, $k_d = 2$ and $V_{max} = 1$. Expansion, KZ1, KZ2 are colored red, green, and blue respectively. 78

A.1 Disparity and Neighbor Statistics for Data and Model with $\gamma \approx 0.1$, z-range between 2 and 8 and baseline 0.1 86
A.2 Error statistics for basic formulation for scenes 5a (left half), and 5b (right half) 87
A.3 Error statistics for formulations with visibility constraints for scenes 5a (left half), and 5b (right half) 88
A.4 Contribution of binocular monocular error for basic formulation. 89
A.5 Contribution of binocular monocular boundary error for formulations with additional constraint 90
A.6 Contribution of binocular discontinuity error for formulations with basic constraints 91
A.7 Contribution of binocular discontinuity error for formulations with additional constraint 92
A.8 Contribution of binocular continuity error for formulations with basic constraints 93
A.9 Contribution of binocular continuity error for formulations with additional constraint 94
A.10 Error statistics (> 1) for methods using the basic formulation 95
A.11 Error statistics (> 1) for methods using additional constraints 96

List of Tables

4.1 Scene Parameters 34

5.1 Algorithms and Parameters for Basic Formulation 48
5.2 Algorithms and Parameters for Visibility Formulation 49
5.3 Summary of Total Error Statistics for Basic Formulation. Grayed out rows represent scenes with depth range 8–32 56
5.4 Summary of Binocular Monocular Error Statistics. Grayed out rows represent scenes with small objects 57
5.5 Summary of Binocular Discontinuity Error Statistics. Grayed out rows represent scenes with small objects. 58
5.6 Summary of Binocular Continuity Error Statistics. Grayed out rows represent scenes with small objects. 59

Chapter 1

Introduction

1.1 Stereo Reconstruction

Stereo vision is a classical problem motivated by how the brain fuses images from two eyes to give perception of depth. In stereo reconstruction, the 3D scene is inferred from a pair of 2D images of a scene taken from two different viewpoints. The problem is solved by first finding the corresponding points between two images and then determining the distance from a predefined coordinate frame using triangulation. The key step in this process is solving the correspondence problem. In this thesis, we specifically address the problem of stereo reconstruction for cluttered scenes.

1.2 Cluttered Scene Stereo Reconstruction

We consider cluttered scenes to be any scene, where objects with a certain size distribution are randomly positioned in 3D space. Some examples of such scenes are forest canopy, bushes, hedges, foliages, etc. These scenes pose special challenges because they have large number of depth discontinuities and as a result any method that makes smoothness assumption are likely to have difficulty with reconstructing such scenes. In this thesis, we look at how different stereo formulations perform for cluttered scenes. More specifically, models that formulate the problem in an energy minimization framework are considered.

1.3 Modeling the Stereo Vision Problem

The stereo vision problem has been modeled in a number of ways over the years. Models based on correlation, variational methods, and discrete labeling has been proposed. There also has been a number of work on modeling biological vision and using that to solve the stereo problem. In this thesis, we are only interested in the models based on Markov Random Field. Currently the top performing algorithms are based on some variation of this model. So far there has not been any work on investigating how well these methods perform for cluttered scenes. Most of the performance evaluation work has concentrated only on the standard set of stereo problems which do not exhibit significant amount depth discontinuities. Since cluttered scenes occur in nature very often, it is important to investigate how the MRF-based models perform for these scenes.

1.4 Objective

Our aim is to understand how different optimization techniques with different priors perform for cluttered scenes. We specifically ask, how well do the current methods perform for cluttered scenes? What types of prior perform best for these scenes? Where do the errors occur and by how much? Is there any correlation between the error rate and scene statistics? Is there any room for improvement? By investigating these questions, our goal is to step towards understanding how better priors can be formulated for cluttered scenes.

1.5 Contributions

The main contribution of this thesis is evaluating the performance of an important subset of MRF based stereo algorithms and determining the forms of smoothness and optimizers that work well for different types of cluttered scenes.

1.6 Motivation and Application Areas

General stereo vision has a wide range of applicability in the real world. Cluttered scene reconstruction can have application in forestry where researchers want to make various

measurements (e.g. leaf density, visible amount of light) to determine the growth of a forest or other ecological statistics. An example of a range image for a forest scene is shown in figure 1.1 from [5]. Scene reconstruction from high-resolution satellite stereo images can also benefit from cluttered scene based stereo formulations. One of the challenges in this case is that certain terrains (e.g. forests, or urban regions) can have a lot of depth discontinuities present. In those cases, it is desirable to use algorithms that can robustly handle such scenarios.

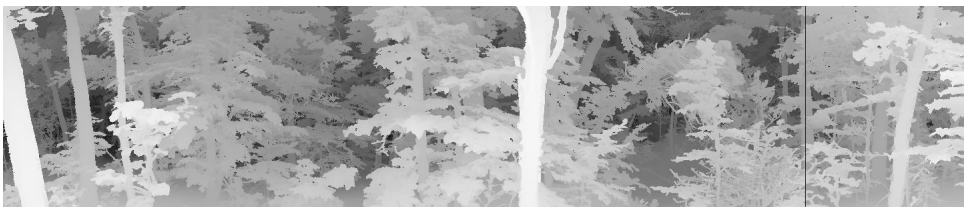


Figure 1.1: Sample range image of a forest scene from [5]

1.7 Outline

The outline of this thesis is as follows: In Chapter 2, we give an overview of the techniques related to MRF based formulation, optimization methods, and compare and contrast some of the relevant previous works. In Chapter 3, we discuss the cluttered scene model that is used in this thesis. Chapter 4 motivates synthetic scene generation process, addresses several issues related to stereo pair generation, specifies the scene parameters for the benchmark dataset and finally experimentally verifies the desired underlying scene statistics. The performance of different stereo algorithms with different parameters settings and different cluttered scenes are presented in Chapter 5. Finally, Chapter 6 concludes the thesis by summarizing the overall approach, addressing the question posed in Section 1.4 and giving future directions.

Chapter 2

Background Review

In this chapter, we discuss some of the existing methods for formulating and solving the stereo reconstruction problem. The scope of this review is restricted to MAP-MRF formulation of stereo vision problems and MAP estimation using optimization techniques such as Graph Cuts and Belief Propagation. An overview of previous works on comparing stereo methods is also presented.

The organization of this chapter is as follows. In section 2.1, we discuss MAP-MRF formulation for general computer vision problems. We review Graph-Cuts in section 2.2 and Belief Propagation with its different variants in section 2.3. In section 2.4, we specifically consider the problem of stereo and review some of the key techniques that have been proposed so far and also considered in this thesis. Section 2.5 presents some of the previous work done on comparative studies with different algorithms and problem formulation and discuss the similarities and differences between those work and ours.

2.1 MAP-MRF formulation for Vision

For many problems in computer vision, we are mainly interested in minimizing an energy function of the form [6]:

$$E(X) = E_{data}(X) + E_{smooth}(X) \tag{2.1}$$

where $E : \mathcal{L} \rightarrow \mathbb{R}$. $\mathcal{L}$ is a set of labels assigned to each pixel, and $X \in \mathcal{L}^{|\mathcal{P}|}$ where $\mathcal{P}$ is the set of pixels.

A set of random variables $X = \{X_i\}$ is a Markov Random Field with respect to some neighborhood $\mathcal{N}$ if it satisfies the following properties:

1. $p(X_i) > 0$
2. $p(X_i \mid X \setminus X_i) = p(X_i \mid X_j \in \mathcal{N}(X_i))$ where $\mathcal{N}(X_i)$ are the neighbors of X_i

The Hammersly-Clifford theorem establishes the equivalence between a MRF and a Gibbs Random Field. A set of random variable, F is a Gibbs Random Field (GRF) with respect to $\mathcal{N}$, if it follows the Gibbs distribution:

$$p(f) = \frac{1}{Z} \exp\{-\frac{1}{T}U(f)\} \tag{2.2}$$

where $f \in F$ and Z is a normalizing constant:

$$Z = \sum_{f \in F} \exp\{-\frac{1}{T}U(f)\} \tag{2.3}$$

Here T is called the temperature and U is the potential function. The potential function is defined as:

$$U(f) = \sum_{c \in \mathcal{C}} V_c(f) \tag{2.4}$$

where V_c is the clique potential and depends on the configuration of the clique. An MRF can be categorized based on the characteristics of the clique potential function and clique size. If the clique size, $|c| = 1$, it is known as a first-order MRF model. If the potential function is independent of position and orientation, it is considered as homogeneous and isotropic respectively. In this thesis and also in most MRF based formulations, only first-order homogeneous and isotropic MRFs are considered. A neighborhood of size c_n, centered at (i, j) is defined as

$$\mathcal{N}(p_{(i,j)}) = \left\{p_{(k,l)} \mid p_{(k,l)} \in \mathcal{P} \wedge (0 < (i-k)^2 + (j-l)^2 \leq c_n)\right\}$$

In the definition of MRF, the probability of a random variable depends on its neighbors. But due to Markov-Gibbs equivalence we do not need to explicitly compute the conditional probability. Rather, we can directly compute the joint probability using the clique potential.

In the MAP-MRF framework we want to find a *configuration*, which is an assignment of values to random variables, that maximizes the posterior probability. This

can be done by taking the negative log probability and minimizing the corresponding function. Since we are only concerned with the optima of the function, we can ignore the normalizing constant.

$$\begin{aligned} p(X|D) &= \frac{p(D|X)p(X)}{\prod_X p(D|X)p(X)} \\ &\propto p(D|X)p(X) \\ &= \exp\{-\frac{1}{T}\sum_{x\in X} E_{data}(D|x)\}\exp\{-\frac{1}{T}\sum_{x\in X} E_{smooth}(x)\} \\ &= \exp\{-\frac{1}{T}(\sum_{x\in X} E_{data}(D|x) + \sum_{x\in X} E_{smooth}(x))\} \\ &= \exp\{-\frac{1}{T}E(X)\} \\ \therefore\ -\log(p(X|D)) &\propto E(X) \end{aligned}$$

In such formulation, MAP estimation is equivalent to finding the minimum of Eq. 2.1. Where E_{data}, which is known as the data term, is defined as $E_{data}(X) = \sum_{x\in X} E_{data}(D|x)$ and E_{smooth}, which is known as the smoothness term, is defined as $E_{smooth}(X) = \sum_{x\in X} E_{smooth}(x)$.

The likelihood can be thought of as modeling sensor noise while the prior models the contextual relationship between pixels. In terms of penalty, the data cost is the penalty for assigning a label to a node. Since in the stereo problem we want to find the corresponding pixels, the data term measures the dissimilarity between corresponding pixels. The smoothness term is the penalty for assigning a pair of neighboring pixels certain labels. It ensures continuity between similar neighboring pixels.

Now that we have seen how to model vision problems using the MAP-MRF framework, we consider the problem of finding the optimal labeling or the labeling that maximizes the posterior probability. That is, we want to solve $X^* = \text{argmin}_X E(X)$.

There are several techniques for doing solving this problem. Early approaches used simulated annealing [7], ICM [8], and other probabilistic techniques. Currently two of the top performing methods are Graph Cut (GC) [9, 10, 11, 12, 13, 14] and Belief Propagation (BP) [15]. Graph Cut works by constructing a graph whose minimum cut corresponds to the minimum energy. Belief Propagation works by passing messages that corresponds to how good a particular label is with respect to all the other nodes in the neighborhood. In the following two sections (2.2, 2.3), we give a brief overview

of these optimization techniques. The details of the algorithms are not used in any of the analysis and are provided for completeness only.

2.2 Graph Cut

Graph cut is a combinatorial optimization technique that works by formulating the problem in terms of a graph problem and then finding a solution to that problem using graph theoretic algorithms. More specifically in Graph Cut, the energy function is represented using a graph and the minimum energy is computed by solving max-flow/min-cut problem. Graph Cut in vision was first used by Grieg et al. in [9], for binary image restoration problem. However, it was not clear at that time how it could be used for more general vision problems. It was in the late 1990s that the approach started to gain popularity [11, 12, 13, 14].

The key step in the Graph Cut formulation is expressing the energy function as a graph. There are several graph-cut algorithms that differ in the way the graph is constructed and how the labeling is performed. In this thesis, we only consider the graph construction proposed by Kolmogorov et al in [16], and a class of move-making optimization algorithms.

2.2.1 Graph Construction for Binary Labeling

We give a brief overview of the graph construction for binary labeling problem from Kolmogorov et al [16]. It is a general construction that does not depend on the specific terms of the energy function or the characteristics of the problem. Before going into the details of the construction process, we first define the class of graph representable energy functions:

Definition A function E is a graph representable function if for a graph $G = (V, E)$ the minimum cut on the graph C equals the minimum energy plus a constant.

The graph $G = (V, E)$ used for representing the energy function is a directed graph whose set of vertices V usually corresponds to the pixels, or set of pixels and the weights of the directed edges E encode the relationship between pairs of pixels. Two additional nodes s and t which corresponds to the binary labels are also added to the graph. These nodes are known as the terminal nodes. Therefore, $V = \{\mathcal{P}\} \cup \{s, t\}$. The

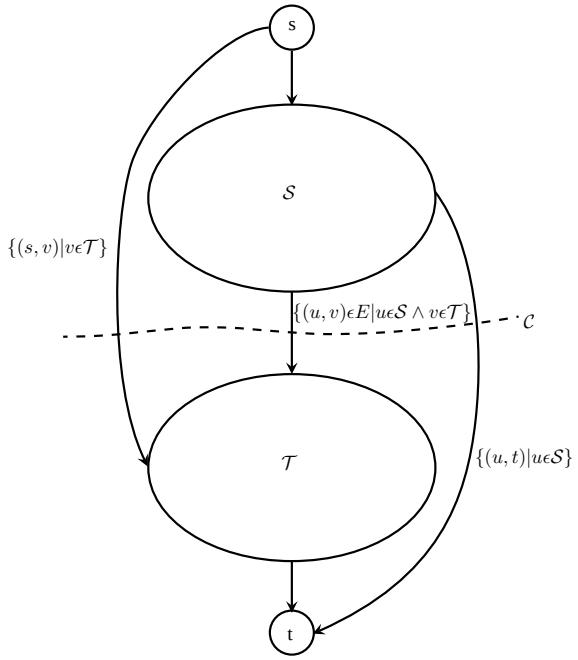


Figure 2.1: s-t min cut on directed graph

edges connecting non-terminal nodes are called the n-links and the ones connecting the non-terminal nodes to the terminal nodes are referred to as the t-links. The weight of the n-links correspond to smoothness penalties and t-links data penalty. In this graph, the minimum cut corresponds to the minimum energy plus a constant. The minimum cut can be computed by solving the max-flow/min-cut problem.

A cut in the graph is the set of edges which when removed creates two disjoint components. Hence, no proper subset of the min-cut $\mathcal{C}$ can be a cut. In a directed graph the set of edges going from set S to T is called the s-t-cut. Therefore, the cut $\mathcal{C}$ (Fig. 2.1) is,

$$\mathcal{C} = \{(u,v) \mid u \in \mathcal{S} \ \wedge \ v \in \mathcal{T}\} \cup \{(u,t) \mid u \in \mathcal{S}\} \cup \{(s,v) \mid v \in \mathcal{T}\}\,.$$

The cost of the cut is the sum of the edges in $\mathcal{C}$. In other words, its the sum of certain t-links and n-links which is equivalent to the sum of data and smoothness costs.

The class of Graph Cut techniques that we consider requires solving binary labeling problem in each intermediate step. So from this point on we only consider the binary labeling problem.

The binary labeling problem can be represented using binary random variables $X = \{x_p \mid p \in \mathcal{P}\}$ where $x_p = 0 \implies x_p \in \mathcal{S}$ and $x_p = 1 \implies x_p \in \mathcal{T}$. Using binary

labels, we can write the energy function as a sum of unary and binary terms as follows:

$$E(X) = \sum_i E_i(x_i) + \sum_{i,j} E_{ij}(x_i, x_j) \tag{2.5}$$

where E_i and E_{ij} denotes a unary and a binary term respectively.

It can be shown that to have a valid graph representation, the energy function must satisfy the following property:

$$\begin{aligned} -E_{ij}(0,0) + E_{ij}(0,1) + E_{ij}(1,0) - E_{ij}(1,1) \geq 0 \\ \therefore\ E_{ij}(0,0) + E_{ij}(1,1) \leq E_{ij}(0,1) + E_{ij}(1,0) \end{aligned} \tag{2.6}$$

This property which is called the *regularity* condition is a binary special case of the *submodularity* condition. Further details on the graph construction can be found in [16]. In the following sections we look at how different algorithms can be used to optimize the energy function using this graph construction.

2.2.2 Move-Making Algorithms

This is a class of algorithm that tries to find the minimum energy by iteratively assigning new labels to existing configurations. They can be categorized based on the size of assignments that are made: 1) Standard move algorithm, and 2) Large scale neighborhood search algorithms. Methods like ICM and Simulated Annealing are all standard move making algorithms since they only modify single pixel at a time. The algorithms that are going to be discussed in this section: Expansion and Swap are of the second category, because they affect the label of a large number of pixels in a single iteration.

The main differences between Expansion and Swap are the set of vertices and interpretation of the terminal nodes. Both the algorithms iteratively decompose the original problem into binary optimization subproblems and stop when no solution is found in any step of the iteration.

2.2.2.1 Expansion

Let the set of configurations that only differ from each other by the label α be:

$$\mathcal{N}_\alpha(f) = \left\{ f' \mid \forall_{p \in \mathcal{P}}\ f_p \neq f'_p \implies f_p \neq \alpha \text{ and } f'_p = \alpha \right\}.$$

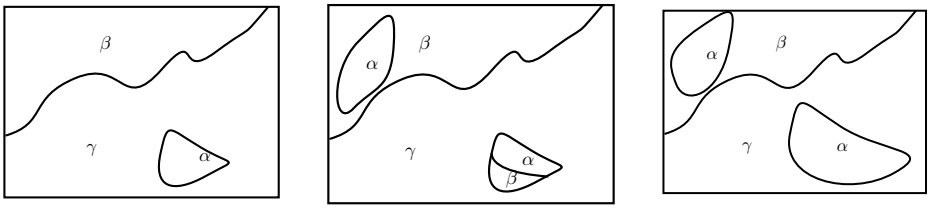


Figure 2.2: a) Initial Configuration b)Swap, and c) Expansion

Such configurations are a single α-expansion move away. In each iteration of Expansion, the algorithm tries to find such a configuration (f^*) that is a single α-expansion move away and minimizes the energy function. The algorithm is given below:

Algorithm: Expansion

success := 1
while *success = 1* **do**
  success := 0
  for *each label* $\alpha \in \mathcal{L}$ **do**
    $\hat{f} := \underset{f \in \mathcal{N}_\alpha(f^*)}{\operatorname{argmin}} E(f)$
    if $E(\hat{f}) < E(f^*)$ **then**
      $f^* := \hat{f}$
      success := 1
    end
  end
end

The minimum cost labeling is found by constructing an st-graph with $V = \{\mathcal{P}, \alpha, \bar{\alpha}\}$. Since this is a binary labeling problem, we have a binary configuration $X = \{x_p | p \in \mathcal{P}\}$ where $x_p = 0 \implies f'_p = f_p$ and $x_p = 1 \implies f'_p = \alpha$. Boykov et al. in [14] showed that $E(f^*) \leq 2cE(f^{opt})$ for optimal configuration f^{opt}. For Potts model $c = 1$.

2.2.2.2 Swap

In the case of the Swap algorithm, binary optimization is done for each pair of labels (α, β). In each iteration, only the pixels, labeled either as α or β are considered. That is, a graph is constructed from the set of pixels $\mathcal{P}$ such that $\forall_{p \in \mathcal{P}} f_p = \alpha$ or $f_p = \beta$.

Let the set of configurations which only differ by swapping labels α and β be:

$$\mathcal{N}_{\alpha\beta}(f) = \left\{ f' | \forall_{p \in \mathcal{P}} f_p \neq f'_p \implies f_p = \alpha \text{ and } f'_p = \beta \text{or} f_p = \beta \text{ and } f'_p = \alpha \right\}$$

Such configurations are a single $\alpha\beta$-swap move away. In each iteration, the Swap algorithm tries to find an assignment of the labels α and β, which is a single $\alpha\beta$-swap move away and minimizes the energy function. The algorithm is as follows: The

Algorithm: Swap

success := 1
while *success = 1* **do**
  success := 0
  for *any pair of labels* $\{\alpha, \beta\} \in \mathcal{L}$ **do**
    $\hat{f} := \underset{f \in \mathcal{N}_{\alpha-\beta}(f^*)}{\text{argmin}} \; E(f)$
    if $E(\hat{f}) < E(f^*)$ **then**
      $f^* := \hat{f}$
      success := 1
    end
  end
end

minimum cost labeling is found by constructing an st-graph with $V = \{\mathcal{P}^{\alpha\beta}, \alpha, \beta\}$ and solving the st min-cut problem.

2.3 Belief Propagation and its Variants

There are two variations of the BP algorithm in terms of how the belief is computed: sum product and max product (or max sum in the case of log probabilities). In BP, an optimal label for each node is determined by passing messages between the nodes. Let m_{pq} be the message passed from node p to node q. Each message represents a score for the receiving node to be assigned some label f_q. This is represented using the notation $m_{pq}(f_q)$. To compute its message, node p tries to find a label f_p that maximizes the compatibility with its neighbors label f_q. This is done by the following message update rule (for an iteration t):

$$m^t_{pq}(f_q) := \max_{f_p} \left\{ \sum E_p(f_p) + \sum E_{pq}(f_p, f_q) + \sum_{r \in N(i), r \neq q} m^{t-1}_{rp} f_p \right\}$$

Once the message propagation stops, the label with the maximum belief is chosen as follows:

$$f_p = \max_{l \in \mathcal{L}} \{b_p(l)\} \quad \text{where } b_p(l) = E_p(l) + \sum_{r \in \mathcal{N}(p)} m_{rp}(l)$$

The algorithm is shown below.

Algorithm: Max Product BP (BP-M)

for *each pixel* $p \in \mathcal{P}$ **do**
  for *each neighbor* $q \in \mathcal{N}(p)$ **do**
    for *each label* $f_q \in \mathcal{L}$ **do**
      $m^t_{pq}(f_q) := 0$
    end
  end
end
for $t = 1$ *to MaxIteration* **do**
  for *each pixel* $p \in \mathcal{P}$ **do**
    for *each neighbor* $q \in \mathcal{N}(p)$ **do**
      for *each label* $f_q \in \mathcal{L}$ **do**
        $$m^t_{pq}(f_q) := \max_{f_p} \left\{ \sum E_p(f_p) + \sum E_{pq}(f_p, f_q) + \sum_{r \in N(i), r \neq q} m^{t-1}_{rp} f_p \right\}$$
      end
    end
  end
end
for *each pixel* $p \in \mathcal{P}$ **do**
  $$f_p = \max_{l \in \mathcal{L}} \{b_p(l)\} \quad \text{where } b_p(l) = E_p(l) + \sum_{r \in \mathcal{N}(p)} m_{rp}(l)$$
end

In [17], Wainright et al establishes a connection between the standard BP algorithm and LP relaxation. Their approach, known as Tree-Reweighted Message Passing (TRW) formulates the MAP estimation problem on a general graph, as a MAP estimation problem on a set of trees. This allows them to prove certain properties of the upper bound of the energy. The update rule is as follows [18]:

$$m^t_{pq}(f_q) := \max_{f_p} \left\{ c_{pq}(E_p(f_p) + \sum_{r \in N(i), r \neq q} m^{t-1}_{rp} f_p) - m^{t-1}_{qp} f_p + E_{pq}(f_p, f_q) \right\}$$

The coefficient c_{pq} in Eq. 2.3 is the main element that makes it different from standard BP-M. The update rule is equivalent to that of BP-M (Eq. 2.3) when $c_{pq} = 1$. The coefficient is computed from the tree-structured distribution of the original problem.

In [19], Kolmogorov proposed a sequential version of the algorithm and showed improved convergence. We use this latter algorithm which is known as TRW-S or sequential Tree-Reweighted Message Passing. In the sequential version the labels are ordered in a particular way and the update rule is applied based on the ordering. The same idea is used by the author in BP-S, which is BP-M with sequential updating rule.

2.4 MRF based Stereo Algorithms

So far we have seen how general computer vision problems are formulated using the MAP-MRF framework and solved using various discrete optimization techniques. In this section, we specifically consider MAP-MRF formulation for stereo problems. In our presentation, a stereo pair is assumed to be rectified and the images in the pair differ only by a horizontal shift. This horizontal displacement or difference between the columns of corresponding pixels is the disparity. In this setting, the optical axes of the cameras are also assumed to be parallel. In the case of stereo, the labels are the disparity values. Most methods solve the optimal assignment problem for one of the images (usually the left), while others solve the problem for all images by either considering a fixed camera or for individual cameras. In the following sections, we discuss two types of formulations that differ in the type of constraints being used by the underlying model.

2.4.1 Basic Formulation

The basic formulation is simply the general form of energy function that was considered at the beginning of this chapter (Eq. 2.1). The equation is rewritten in the context of stereo as follows:

For stereo pairs I_l and I_r with integer disparity values f_p and f_q for pixels p, and q, the energy function that is optimized is of the following form [18]:

$$E = E_p(f_p) + \lambda\, E_{p,q}(f_p, f_q) \tag{2.7}$$

$$E_p(f_p) = d(I_l, I_r, p, q)^{k_d} \tag{2.8}$$

$$E_{p,q}(f_p, f_q) = w_{pq}\, min(|f_p - f_q|^{k_s}, V_{max}) \tag{2.9}$$

As before, equation 2.8 is called the data term and 2.9 the smoothness term. The function $d(.)$ in the data term measures the dissimilarity between corresponding pixels. It can be either a basic dissimilarity measure (e.g. absolute or squared difference between the corresponding pixels) or the more sophisticated Birchfield-Tomasi measure [20]. The data term can be chosen either to be the linear or quadratic (for $k_d = 1$ and 2 respectively) form of the dissimilarity measure.

The smoothness term is a function of label difference. This implies that the smoothness function penalizes discontinuities. Quite often a truncation value like V_{max} in Eq. 2.9 is used. This helps to preserve discontinuities and give better results in practice. In the smoothness term w_{pq} is a function of a pair of pixels which is usually a function of color gradient. It can be defined as follows:

$$w_{pq} = \begin{cases} \lambda_\nabla & \text{if } |I_p - I_q| < I_{threshold} \\ 1 & \text{otherwise} \end{cases} \tag{2.10}$$

. Here, λ_∇ is the gradient penalty and $\lambda_\nabla > 1$. The coefficient λ in Eq. 2.7 specifies the weight that should be given to the smoothness term.

2.4.2 Additional Constraints

The above formulation does not put any restriction on disparity consistency. It is quite possible for multiple pixels in the left image to be mapped onto the same pixel in the right image. Even the corresponding pixels may not map onto each other. Also the formulation does not explicitly consider occlusion and therefore is likely to perform poorly near discontinuities. Such observations have motivated researchers to consider additional constraints.

In the case of visibility reasoning there can be three possible scenarios as shown in Fig. 2.3a. A surface point can either be binocularly visible, semi-occluded (or monocularly visible) or completely occluded. For scenes with fronto-parallel surfaces we can have additional constraints that penalizes many-to-one mapping. There can be

other constraints that allow scenes to have slanted surfaces. However, in this thesis we restrict ourselves to visibility and uniqueness constraints primarily because they are more appropriate for our synthetic scenes and our objective is to analyze the effect of occlusion and discontinuity on a stereo method's performance.

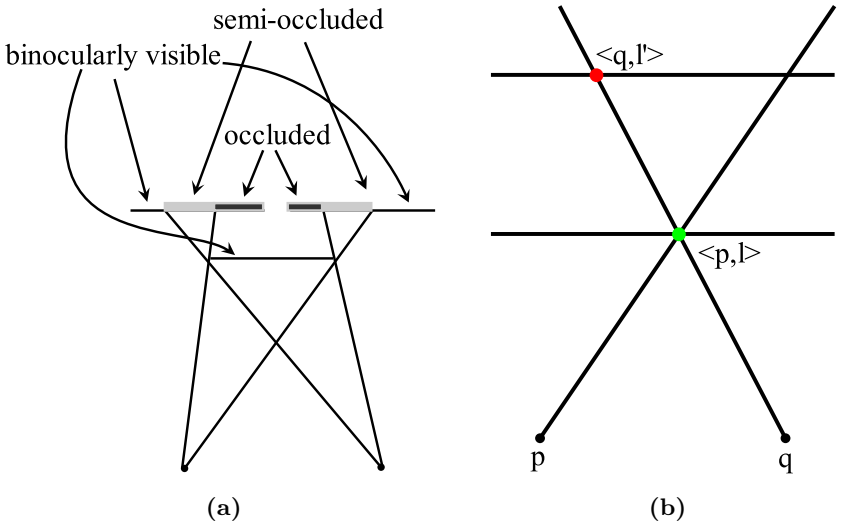


Figure 2.3: a) Binocular visibility, b) $< p, l >$ blocks q from seeing the point $< q, l' >$. Therefore $\{< p, l >, < q, l' >\} \in I_{vis}$.

In the following, we discuss a general formulation for visibility and uniqueness constraints proposed by Kolmogorov and his colleagues in [4]. Specifically we are more interested in the two special cases: KZ1 [3] and KZ2 [2]. The general energy function is of the form [4]:

$$E(f, g) = E_{data}(g) + E^{p}_{smooth}(f) + E^{i}_{smooth}(g) + E_{vis}(f) + E_{consistency}(f, g) \quad (2.11)$$

Two types of configurations, f and g, are used in the formulation. The meaning of the configurations and the terms are explained in our description of the two special cases below.

In the first special case, referred to as KZ1 [3], Eq. 2.11 can be rewritten as (see

[4, 21] for details):

$$E(f) = E_{data}(f) + E_{smooth}(f) + E_{vis}(f) \tag{2.12}$$

$$E_{data}(f) = \sum_{\{<p,f(p)>,<q,f(q)>\}\in I} \min\{0, d(I_l, I_r, p, q)^{k_d} - K\} \tag{2.13}$$

$$E_{smooth}(f) = \sum_{\{p,p'\}\in\mathcal{N}_1} V_{pp'}(f_p, f_{p'}) \tag{2.14}$$

$$\text{where, } V_{pp'}(f_p, f_{p'}) = w_{pp'}\ min(|f_p - f_{p'}|^{k_s}, V_{max})$$

$$\text{and, } \mathcal{N}_1 \subset \{\{p, p'\}|p, p' \in \mathcal{P}\}$$

$$E_{vis}(f) = \sum_{\{<p,f(p)>,<q,f(q)>\}\in I_{vis}} \infty \tag{2.15}$$

Among the two smoothness terms in Eq. 2.11, only E^p_{smooth} is considered. The consistency constraint between f and g is ignored and $E_{data}(g)$ is re-written as $E_{data}(f)$. The configuration f is defined as: $f : \mathcal{P} \rightarrow \mathcal{L}$. The data term shown in Eq. 2.13 is similar to the data term in the basic formulation. One difference is that to satisfy regularity condition the term has to be negative which is why a constant K is subtracted from the dissimilarity measure $d(.)$. The smoothness term is also similar to the basic formulation except that in this formulation the smoothness penalty is calculated for all images rather than just for a single image. The additional term, E_{vis}, enforces visibility constraint. To understand this term, we need to define the notion of "scene point", "interaction" (I), and "visibility interaction" (I_{vis}).

A pair $< p, l >$, where $p \in \mathcal{P}$ and $l \in \mathcal{L}$, defines a 3D-point in the scene. It is the intersection of a ray from pixel p and a plane at distance l from a fixed camera. Let p and q be pixels in two different images i and j respectively. Two points $< p, l >$ and $< q, l' >$ interact (i.e. $\{< p, l >, < q, l' >\} \in I$), if the projection of $< p, l >$ onto j is q. The visibility interaction set I_{vis}, is the set of point-pairs that violate the visibility constraint. It is defined as, $I_{vis} = \{\{< p, l >, < q, l' >\}|\{< p, l >, < q, l' >\} \in I \wedge (l' > l)\}$. Fig. 2.3 (b) shows an example where $< p, l >$ projects onto q and blocks q from seeing $< q, l' >$.

In the second special case, referred to as KZ2 [2], the energy function can be written

as (see [4, 21] for details):

$$E(g) = E_{data}(g) + E_{smooth}(g) + E_{vis}(f(g)) + E_{consistency}(f(g), g) \quad (2.16)$$

$$E_{data}(g) = \sum_{i \in I} g(i) \left(d(I_l, I_r, p, q)^{k_d} - K\right) \quad (2.17)$$

$$E_{smooth}(g) = \sum_{\{i,i'\} \in \mathcal{N}_i} \lambda T[g(i) \neq g(i')] \text{ , where } \mathcal{N}_i = \{\{i, i'\} | i, i' \in \mathcal{I}\} \quad (2.18)$$

$$E_{vis}(f) = \sum_{\{<p,f(p)>,<q,f(q)>\} \in I_{vis}} \infty \quad (2.19)$$

$$E_{consistency}(f, g) = \sum_{<p,q,l> \in I} \infty \, T(g(<p, q, l>) = 1 \wedge (f_p \neq l \vee f_q \neq l)) \quad (2.20)$$

Here, $T(.) = 1$ if its argument is true and 0 otherwise. In the case of rectified stereo pair, g is defined as $g : \{(\mathcal{P}_l, \mathcal{P}_r)\} \rightarrow \{0, 1\}$. In this formulation, pairs of pixels (p, q), where each pixel is from two different images, are considered. This pair is also known as an interaction and can be either "active" (1) or "inactive" (0). If p and q are corresponding pixels then the interaction $g((p, q)) = 1$. The data term only assigns penalty for active interactions. The smoothness term is different from all other formulations in that it is defined on interactions. The term assigns a penalty if two neighboring interactions are not assigned the same label. Two interactions are neighboring if any two pixels in the two interactions are either the same or neighbors of each other.

The last two terms $E_{vis}(f(g)) + E_{consistency}(f(g), g)$ enforces uniqueness constraint. The main idea behind the constraint is that a pixel cannot be in more than one active interaction. The reason why the two terms enforce uniqueness constraint is that when both (p, q) and (p, q') are active then $\{< p, q, l = q - p >, < p, q', l' = q' - p >\} \in I_{vis}$ because $l \neq l'$ and projection of $< p, l >$ is q'. As a result, the visibility term will be infinite.

2.5 Comparative Study of MRF-based Algorithms

There have been other studies that compare MRF-based algorithms for different parameters and scenes. The result of a particular approach can depend on either the form of energy function or the optimization method. Therefore, the key motivation behind these studies has been to understand the effect of different parameters on different scenes.

Scharstein and Szeliski's work [22] was one of the earliest to compare different algorithms. The motivation behind that work was to categorize different dense stereo algorithms and also to characterize the performance of those algorithms for different scenes with different parameters. The authors consider a wide range of algorithms including MRF-based techniques with basic formulation. They identified the common elements between different algorithms and obtained results for 20 algorithms. All the algorithms were implemented under a common framework to make more meaningful comparison. In all cases, the appropriate parameters were modified to evaluate the performance of those algorithms.

To compare the output disparity d_C with the ground-truth d_T, two quality metrics were used:

1. RMS error $E_{RMS} = \sqrt{\frac{1}{N}\sum_{p\in\mathcal{P}}|d_C(p) - d_T(p)|^2}$
2. Percentage of bad matching pixels $E_{avg} = \frac{1}{N}\sum_{p\in\mathcal{P}}(|d_C(p) - d_T(p)| > \tau)$, where τ is the error threshold.

These error statistics are computed in: Textureless, Occluded and Discontinuity regions. The authors found Graph Cut based optimization to be better than other techniques and concluded that using Birchfield-Tomasi and gradient thresholding based smoothness cost gives the best performance. We also find similar conclusion in our analysis. However the authors note that "Choosing the right parameters (threshold and penalty) remains difficult and image-specific".

The motivation behind this thesis is similar. We also consider the amount of error that occurs for different parameter combinations and where they occur (excluding textureless region). However, we focus on a wider range of MAP-MRF based techniques and special type of scenes. Furthermore we show the set of parameters that work best in each scene category, and relate them to the underlying scene statistics.

Tappen et al [23] investigates how Graph Cut and Belief Propagation performs for the same energy formulation. The motivation behind their work was to identify whether the MRF formulation or the inference algorithm causes the difference in performance. To this end, they used the basic MRF stereo formulation and optimized the energy using Swap and BP-M algorithm with the same parameter settings (Birchfield-Tomasi

for data and Potts model for smoothness). The following four statistics were considered for comparing results. These are similar to the ones used by Scharstein et al in [24].

1. Percentage of greater than 1 pixel disparity error in the unoccluded regions.
2. Percentage of greater than 1 error in the textureless regions.
3. Percentage of greater than 1 error near discontinuities.
4. Energy of the solution.

The algorithms were applied on "Map", "Tsukuba", "Sawtooth" and "Venus" dataset. In general they found both the algorithms to have similar performance. Both Swap and BP-M were able to find energy configuration lower than the actual energy. This implies that the data and smoothness terms do not model the problem sufficiently well.

In this thesis, the algorithms are compared in a similar manner. However, a wider range of parameter and algorithm settings are considered.

Szeliski et al, in [18] considers MRF based energy minimization techniques with smoothness-based priors and uses a wide range of optimization algorithms for performance evaluation. They used ICM, Expansion, Swap, BP-M, BP-S, and TRW-S for stereo, photomontage, binary image segmentation and image denoising and inpainting problems with different forms of prior. Their primary objective was to evaluate different energy minimization techniques for different types of priors. The authors used the basic stereo formulation and evaluated the algorithms using "Tsukuba", "Venus" and "Teddy" images from the Middlebury benchmark dataset with V_{max} between 1 and 2, norms L1 and L2, and without gradient threshold. They found Expansion, Swap and TRW-S to be the best performing algorithms in terms of finding the lowest energy.

Compared to [18], we only consider cluttered scenes with different scene parameters for both basic and additional constraints based formulation. We also compare the algorithms in terms of error rates rather than optimal energy and show results for a wide range of parameters. Furthermore we look at the error statistics of the resulting images and relate that to the underlying scene statistics.

Chapter 3

Modeling Cluttered Scenes

This chapter presents the cluttered scene model that is used in this thesis. Cluttered scenes are defined to have objects of certain size and shape distributed randomly in 3D space. There are several real-world examples of these scenes such as foliage, tree canopy, bushes, hedges etc. A model allow us to understand the underlying statistics of cluttered scenes and thus to classify scenes based on those statistics. This is important for us because our goal is to evaluate different methods for different types of cluttered scene and observe the trend in performance due to different underlying statistics. A model lets us choose specific parameters for generating specific scenes, thereby enabling us to methodically evaluate the performance of stereo algorithms.

Ideally a cluttered scene model should mimic the statistics of natural scenes. One such model which is very widely used is the dead leaves model [5, 25]. The objective of the model is to capture the scale-invariant properties of natural scenes. As a result the model only uses the ratio of distance between objects. In our case, we need to use the actual distance of the objects for disparity computation. Therefore, the cluttered scene model from [1], which explicitly derives the probability model with respect to scene point distance, is used.

The organization of the chapter is as follows. In sec. 3.1 we give an overview of previous work related to general scene modeling. The basic assumptions and notations are listed in sec. 3.2. Finally the derivation of the probability models from [1] with some minor modifications is presented in sections 3.3, 3.4, and 3.5.

3.1 General Scene Models

There have been a number of works on modeling the statistics of natural images. A patch of image of size P with K possible pixel values can have P^K possible images. However, not all of those images are meaningful. The study of natural image statistics are motivated by the observation that the small subset of meaningful images should have similar underlying statistics. An important property of natural scenes is that they are translation invariant. This means that for an ensemble of images the statistics at one point is the same as any other point. This allows us to compute the statistics without worrying about the spatial location of pixels. Another important characteristic of natural images is scale invariance. This means that the statistics of the images do not change for different scales. This phenomena has been extensively studied in the literature from the late 1980s [25, 26, 27]. Most of the statistics were studied for intensity images. However, as Huang et al shows in [28], these statistics also hold for range images.

Matheron [29], proposed the "dead leaves model" in the late 1960s for the mathematical morphology community. Since then many researchers used the model to demonstrate scale invariant statistics and also to study the cause of such scale-invariance.

There are two approaches for generating synthetic scenes that conform to the scale-invariance properties of natural scenes [30]. They are: Superposition, and Occlusion Models. In the superposition model, the scene is considered to be a superposition of overlapping objects randomly distributed in the scene, while in the occlusion model the objects are at different depths and the observed statistics are produced due to the occlusion process. In both these models, the object's position and properties (size and texture) are determined using a Poisson process. It can be shown that [5, 25, 31] when the object size follows a $1/r^3$ distribution where r is the size of an object, then the generated scene will have scale-invariant properties. These properties are not investigated in this thesis.

3.2 Basic Assumptions and Notations

[1] addresses the problem of modeling cluttered scenes (as defined in this thesis). In that work, different statistical models are derived assuming constant radius spheres uniformly distributed in the scene. These models include the probability of visibility

and binocular visibility. In a later work [32], the pair-wise depth probabilities for such scenes are modeled. In this thesis, these models are used as the basis for generating synthetic scenes. The objects are assumed to be squares and parallel to the projection plane. This satisfies the uniqueness constraint that is assumed in some of the methods.

Throughout this chapter and the rest of the thesis the following assumptions and notations are used:

Assumptions

1. Objects are parallel to the image plane, and therefore uniqueness constraint holds
2. The horizontal and vertical field-of-view of the cameras are the same
3. Pinhole projection model is assumed and therefore there is no blurring.
4. The size of the projection plane is a function of the focal length and field-of-view. In real cameras this will be the sensor size.
5. The objects are assumed to be within a bounded region.
6. A scene point is assumed to be projected onto a single pixel. Therefore a point in the scene is visible, if a ray from the pixel hits that point. The scene point is occluded if the ray hits some closer object or if the scene point is outside the view-volume. Assuming that all scene points are inside the view-volume, we can simplify derivation by considering the Minkowski Sum of the ray and object (Fig. 3.1). That is the square is shrunk to a single point located at the center of the square and the ray is grown to a cuboid (for a sphere it would be a cylinder). Now a scene point is visible if a cuboid does not contain any square centers (Fig. 3.1).

Common Notations and Conventions

1. Objects are independently and uniformly distributed in the scene.
2. Objects are square with width S and area A.
3. The baseline between the cameras is T_x.

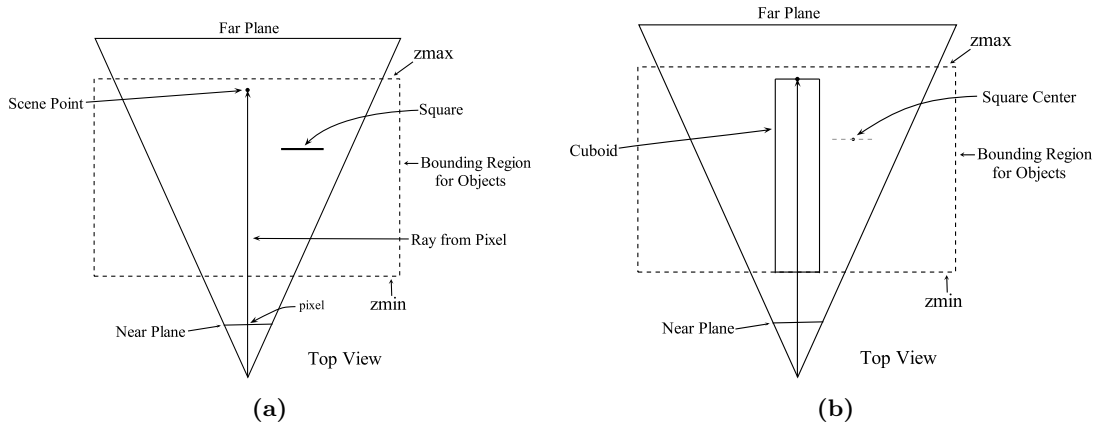


Figure 3.1: a) Top view of scene showing different scene parameters b) Minkowski sum of a ray and a square of half-width. The square in (a) is shrunk to a point.

4. The Focal length of each camera is the distance of the near plane N.

5. Field-of-view is denoted as fov.

6. The extent of the bounding region along the z-axis is z_{min} (or z_0) to z_{max}.

7. The scaling factor from the projection plane to the image coordinate (denoted by the subscript pi) is
$$S_{pi} = \frac{\text{image width in pixels}}{\text{projection plane width}}$$

8. The disparity of a pixel p is denoted as f_p.

9. For transforming distance-based equations to disparity-based, the following substitutions are made:
$$z = \frac{\sigma_s}{f_p} \tag{3.1}$$
$$dz = \frac{\sigma_s}{f_p^2} \tag{3.2}$$
where, $\sigma_s = NT_xS_{pi}$

Based on these assumptions and notations, we look at the different statistics that were presented in [1] with appropriate modifications.

3.3 Probability of Disparity

For a scene point to be visible on the image plane, two conditions must hold:

1. There are no centers inside the cuboid with cross-sectional area A (size equal to that of the object) and extending from z_0 to the position of object (Fig. 3.1b).

2. There is only one square center within a small depth interval $[z, z + dz]$ at the end of the cuboid.

We can model visibility using the Poisson distribution where the probability of having k points is given by the equation $\gamma^k \frac{e^{-\gamma}}{k!}$ where γ is the average number of points within a volume. Let the number of centers within a unit volume be η. For a scene point at distance z to be visible, the first condition can be written as

$$\gamma^0 \frac{e^{-\gamma}}{0!} = e^{-\gamma}$$

where, $\gamma = \eta\, A\, (z - z_0)$ and A is the cross-sectional area of the cuboid.

For there to be a single square center within a very small volume $A\, dz$, the probability is:

$$\gamma \frac{e^{-\gamma}}{1!} = \eta A\, dz \exp\{-\eta\, A\, dz\}$$

Since dz is very small $\exp\{-\eta A\, dz\} \approx 1$. So the total probability of a point at distance z being visible in the image plane is,

$$p(z)dz \approx \eta A \exp\{-\eta A(z - z_0)\}dz \tag{3.3}$$

Using Eq. 3.1 and 3.2, we have:

$$p(f_p) = \frac{\sigma_s}{f_p^2} \eta A \exp\{-\eta A \sigma_s (\frac{1}{f_p} - \frac{1}{f_0})\} \tag{3.4}$$

In the rest of this chapter we do not explicitly perform the substitutions shown in Eqs. 3.1 and 3.2. Rather, the model is presented in terms of distance z and expressed in terms of disparity where appropriate.

3.4 Binocular Visibility

Binocular visibility refers to the case where a 3D point in the scene is visible from both views. This is an important consideration for stereo algorithms because the data term is modeled assuming that the corresponding pixels exist. Modeling the probability of such cases allows us to determine how a stereo algorithm will perform. This is because the techniques we use work best for pixels that are visible to both views. A scene point will be visible from both views if it lies inside the union of the two view volumes and is not occluded by any other scene points. In [1], a probability model for binocular half-occlusion is presented. However, in the model occlusion due to finite view-volume was not considered. In the following derivation the finite view-volume is taken into consideration.

As stated before, a scene point visible from one viewpoint will be invisible from a second viewpoint for one of two reasons:

1. The point is occluded by another point closer to the camera

2. The point is outside the second view volume

Let,
V_i: set of object centers inside view volume i
Γ_i: set of object centers inside cuboid i
$\Phi_i(z)$: width of a view volume i at depth z
$\Delta(x)$: disparity of scene point x
$\beta(x)$: point x is binocular
$\rho(\Gamma_i)$: volume of cuboid Γ_i
A: cross-sectional area of object
Therefore we have,

$$p(\beta(x)) = p(\Gamma_r = \emptyset | \Gamma_l = \emptyset) = p(\Gamma_l \setminus \Gamma_r = \emptyset)\, p(x \in (V_l \cap V_r)) \tag{3.5}$$

The first probability, derived in [1], can be written as,

$$p(\Gamma_l \setminus \Gamma_r = \emptyset) = \frac{\exp\{-\eta\rho(\Gamma_l \cup \Gamma_r)\}}{\exp\{-\eta\rho(\Gamma_l)\}} = \exp\{-\eta\rho(\Gamma_l \setminus \Gamma_r)\} \tag{3.6}$$

where,

$$\rho(\Gamma_l \setminus \Gamma_r) = \begin{cases} A(\frac{2Sz}{T} - z_0) + \frac{2z}{T}S^3 & \text{if } \frac{z-z_0}{z} > \frac{2S}{T} \\ \frac{ST(z-z_0)^2}{z} & \text{otherwise} \end{cases}$$

The second probability is,

$$\begin{aligned} p(x \in (V_l \cap V_r)) &= \frac{\Phi_l(z - T_x)}{A\Phi_l(z - T_x)} \\ &= \frac{w - T_x}{w} \\ &= 1 - \frac{T_x}{w} \quad \text{where, } w = 2z\tan(fov/2) \\ \therefore p(x \in (V_l \cap V_r)) &= 1 - \frac{T_x}{2z\tan(fov/2)} \end{aligned} \tag{3.7}$$

For disparity f_p,

$$p(x \in (V_l \cup V_r)) = 1 - \frac{f_p}{2NS_{pi}\tan(fov/2)}$$

Using the above formulation the probability density of disparity for binocular pixels can be written as:

$$p(\Delta(x) = f_p\,,\, \beta(x)) = p(\beta(x) \mid \Delta(x) = f_p)\, p(\Delta(x) = f_p) \tag{3.8}$$

3.5 Joint Probability of Disparity

MRF based stereo models typically use a smoothness term that tries to capture the underlying pairwise statistics. Therefore, modeling the joint probability for cluttered scenes can be useful for both stereo reconstruction and performance analysis.

We first determine the joint probability of distance and later use that to determine the joint probability of disparity.

The joint probability of neighboring points being at distance z and z' can be determined by considering two separate cases: 1. probability of neighbor being on a closer surface $p(z' < z)$ and 2. probability of both neighbors being on the same surface $p(z = z')$

3.5.1 Probability of Closer Neighbor $p(z' < z)$

Given that a scene point is at distance z, its neighbor will be on a closer surface z' if,

1. For the union of two cuboids contains no square center i.e. $\Gamma_x \cup \Gamma'_x = \emptyset$, the volume is,

$$\rho(\Gamma_x \cup \Gamma_{x'}) = A(z - z_0) + S(z'^2 - z_0^2)(x - x')$$

and the probability is,

$$\exp\{-\eta\rho(\Gamma_x \cup \Gamma'_x)\}$$

2. Both surfaces have a center within a very small distance dz and dz' which is the product of $\eta A dz$ and $\eta 2Sz'(x - x')dz'$

Therefore, the probability

$$p(z' < z) = 2\eta^2 ASz'(x - x') \exp\{-\eta.\rho(\Gamma_x \cup \Gamma'_x)\}dzdz' \tag{3.9}$$

3.5.2 Probability of Equidistant Neighbor $p(z = z')$

Two neighboring pixels will see the same surface at depth z if,

1. The union of two cuboids does not contain any object centers i.e. $\Gamma_x \cup \Gamma'_x = \emptyset$
2. There is an object center at the intersection of $[z, z + dz]$ and $[z', z' + dz']$

$$p(z = z' \mid z) = \eta(A - 2Sz(x - x'))exp\{-\eta.\rho(\Gamma_x \cup \Gamma'_x)\}dz \tag{3.10}$$

Joint probability can be obtained by combining the above two cases. By applying the substitutions in Eq. 3.1 and 3.2, we obtain the joint disparity probability $p(f_p, f_q)$.

3.6 Models Derived from the Joint Probability Model

The joint probability model presented in the previous section can be used to derive the discontinuity and disparity difference models. These models are used later in the thesis because they give better insight into the assumptions made by different stereo algorithms. All the expressions are written in terms of joint probability of disparity.

3.6.1 Probability of Discontinuity $p(f_p \neq f_q | f_p)$

This the probability that, given a scene point at distance z, its neighbor will not be equidistant. It can be easily computed from the joint probability as follows:

$$p(f_p \neq f_q | f_p) = \frac{1}{\sum_{f_q} p(f_p, f_q)} \sum_{f_q < f_p} p(f_p, f_q) + \sum_{f_p < f_q} p(f_p, f_q) \tag{3.11}$$

3.6.2 Probability of Difference $p(f_p - f_q)$

This the probability of having a certain disparity difference. The probability of difference can also be computed from the joint probability as follows:

$$p(f_p - f_q = d) = \frac{1}{\sum_{f_p, f_q} p(f_p, f_q)} \sum_{f_p - f_q = d} p(f_p, f_q) \tag{3.12}$$

3.7 Discussion

In this chapter, we looked at different probability models for cluttered scenes from [1, 32]. We described the model assuming that objects are squares. But the model holds for any shape as long as A is the cross-section area of the object and the objects are fronto-parallel. In the rest of the thesis, we use these models to generate different types of cluttered scenes and evaluate our chosen set of methods. It should be noted that in natural scenes objects are not distributed uniformly in space but can clump together. Also the model does not directly address what the probabilities would be for scale-invariant scenes. However, in the next chapter, we will see that under certain conditions scale-invariant scenes can closely follow the model.

Chapter 4

Scene Generation

In the previous chapter, a cluttered scene model was presented and the motivation behind considering such models was discussed. In this chapter, we concentrate on the scene generation process which includes choosing appropriate parameters for generating scenes with different statistics and rendering stereo pair for those scenes. The generation process for scale-invariant scenes is also described. Finally experimental validations of the generated scenes are presented.

For evaluating different stereo reconstruction techniques (i.e. algorithm and parameter combination), it is important to apply them to a wide variety of scenes. It is also important to categorize scenes based on their underlying statistics so that it is possible to understand the strength and weakness of a technique for a certain type of scene. The scope of this thesis is restricted to cluttered scenes as defined in chapter 3. Real scenes with a desired statistics are hard to obtain and the current widely used dataset does not have sufficient number of cluttered scenes to be applicable for this thesis. This motivates us to generate synthetic scenes which can be used for evaluating different methods.

The model in chapter 3 allows computing the underlying statistics before generating the scenes. It also provides a relationship between the scene parameters and the resulting statistics. This makes it useful for choosing scene parameters that produces the desired statistical properties. The cluttered scene model that was discussed, only considers objects with constant radius. Since our goal is to evaluate different techniques on a wide variety of cluttered scenes, we also consider scenes with non-uniform radius: more specifically scenes with $1/r^3$ size distribution or scale-invariant scenes.

Once a scene is generated, we need to render stereo pairs for that scene. For distance statistics (e.g. in [1]) a single depth map is enough. But for evaluating algorithms we require stereo pairs as input. Besides that, we also need to have a very accurate ground truth disparity map. Although the basic idea behind the rendering process is simple, there are certain subtleties that make the process non-trivial. These issues and their solutions are discussed in detail in this chapter.

The chapter is organized as follows. In Sec. 4.1, the parameter choice and reasoning are discussed. Sec. 4.2 discusses the scene generation process, some of the challenges and their solution. Sec. 4.3 shows the characteristics of the generated scenes and validity of the generation process. Finally the scene generation process and the experimental results are summarized in Sec. 4.4.

4.1 Choice of Parameters

Before choosing scene parameters we first need to categorize cluttered scenes based on some criteria. Currently there are no such classification rules but for our purpose, we can classify the scenes based on the size of objects (r), range of depths, baseline(T_x), and the average number of surface points (γ).

But how should the parameters be chosen? More specifically, what properties of the scene needs to be modified for the experiments and how? To answer these questions, we need to consider how different MRF based stereo formulations work and the underlying assumptions that they make. These were discussed in Sec. 2.4. This thesis mainly considers the underlying single pixel statistics like the binocular visibility, and pairwise statistics like the probability of discontinuity, continuity or disparity difference. In the following we explain how they are computed and why they were chosen. In each case, we compute the statistics for all pixels and for only binocular pixels.

Probability Density of Disparity This is the probability of a pixel having a particular disparity. The theoretical model for this was presented in Sec. 3.3. It is simply the ratio of pixels with certain disparity to the total number of pixels with appropriate normalization using the bin-width. For the binocular case only the binocular pixel's disparity is considered. The motivation behind considering this measure is that most of the models consider the pixels to be binocularly visible. As a result if there are more

binocular pixels in a scene then that scene is likely to have lower total error for all pixels. Furthermore we want to know if the error rate for binocular pixels also change with the probability density in any way. By having different disparity statistics we can understand how this property affects different methods.

Conditional Probability of Discontinuity This is the probability of two neighboring pixels having different disparity. The theoretical model for this was presented in Sec. 3.6.2. It is the ratio of pixels with non equi-disparity neighbors to the total number of pixels with that given disparity. We are interested in this statistic because the smoothness term in general penalizes discontinuity. This penalty depends only on the difference between labels rather than the actual value of the individual labels. This implies that the discontinuity is considered to be independent of depth. Using this statistic we try to understand how performance varies between scenes with different discontinuity statistics.

Probability of Disparity Difference This is the probability of disparity difference between neighbors. Sec 3.6.2 gives the theoretical model for this statistics. It is the ratio of neighbor pairs that have different disparity to the total number of neighbors. This statistic is much closer to the criterion used in almost all of the smoothness terms. Like before this statistic lets us see how different scenes with different statistics affect the performance of different methods.

Now we consider how the parameter values were chosen to make the aforementioned statistics different. From Sec. 3.3, we know that γ and the depth-range directly affect the shape of the disparity curves. The baseline distance primarily affects the range of disparity values. Small baseline causes smaller range of values and the opposite is true for large baseline. Since the overlap between two view depends on the baseline, it also affects the binocular visibility. This parameter is kept fixed for all the main experiments. The size of objects, r, mainly affects the pairwise statistics. If the objects are small then neighboring pixels are more likely to be on different surfaces because the projection of the objects in the image is also small. This results in an increase in the number of discontinuities. The opposite is true for large objects. Also if small objects are closer to the viewer than it is likely to create larger occlusion.

With these different scenarios in mind we consider the parameter values that are shown in table 4.1. Later in this chapter in section 4.3, we compare the similarities and

Table 4.1: Scene Parameters

Scene	baseline	γ	z range	r
1a	0.2	0.54	2-8	0.1
1b				0.025
1c				0.0335 - 0.5359
2a			8-32	0.4
2b				0.1
2c				0.1340 - 2.1436
3a		0.1	2-8	0.1
3b				0.025
3c				0.0335 - 0.5359
4a			8-32	0.4
4b				0.1
4c				0.1340 - 2.1436

differences between the scenes generated using these parameters. But before that the scene generation process is described below.

4.2 Synthetic Scene Generation

The basic idea behind synthetic scene generation is very simple. However in practice there are some subtle issues that need to be considered. Scene generation mainly requires randomly generating the position of objects within a prespecified region and rendering them from two different viewpoints as specified by the baseline parameter. Our approach to scene generation is much closer to the occlusion model [30] approach. In the following subsections we discuss how a stereo pair is rendered and the ground truth disparity map is generated.

4.2.1 Rendering Stereo Pairs

First we discuss the general setup of our synthetic scene generator. For rendering scenes, we consider the image formation model from computer graphics. The image

plane or projection plane is considered to be in front of the center of projection and it is also the near clipping plane. As a result the focal length is equivalent to the distance of the near plane from the center of projection. The size of the projection area on the near plane is determined by the field-of-view and the distance of the near plane. In a real camera it is the sensor size and focal length that determines the field-of-view. Furthermore we consider a pinhole camera model, therefore there is no blurring and all the objects at different depth are in sharp focus. The square normals are parallel to the viewing direction or in other words they are fronto-parallel. The general setup is shown in Fig 4.1a. We also texture map the surface. This is important because all the

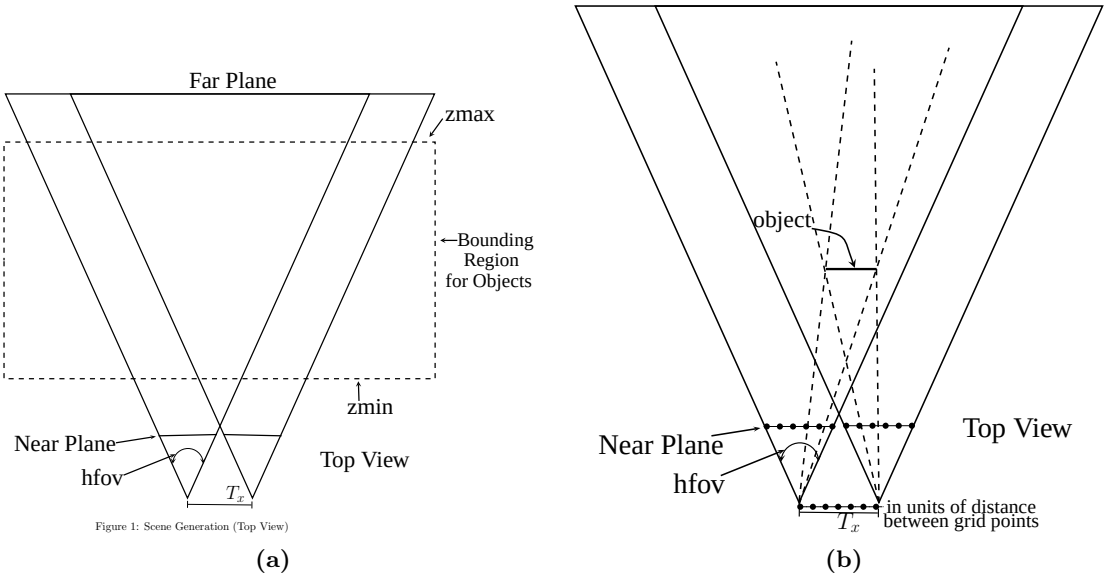


Figure 4.1: a) Scene Generation (Top View) and b) Aligning camera with the pixel grid

algorithms works best for textured surfaces.

Since the same scene is rendered from two different view-points, the size of the projected image in both views should be theoretically the same. However this is not always the case in practice. Fig. 4.1b shows one possible case where the size of the objects will not be the same if the "pixels" on the image plane are not properly aligned. One possible way of overcoming this problem would be to choose the baseline distance in such a way that the projections on the image plane line up with each other. Another possible approach would be to detect such cases and keep regenerating new scenes till

the problem does not occur anymore. One way of detecting such errors would be to backproject an image from one view and reproject it into the other view and then check the length of the diagonal of the square. However, we found both methods to occasionally fail under certain cases.

This problem is mainly caused by the discrete nature of images. Therefore to avoid such problem, we perform the whole rendering process in discrete space. In other words, we pre-compute the disparity and size of the objects in the projected image. During rendering the appropriate transformations are applied and the scene is rendered in the screen space under orthographic projection. It should be noted that since we are rounding to the nearest integer the actual distance or the size of the objects will be slightly different from the specified values. Later in this chapter in Sec. 4.3 we will look closely at these errors closely and see their impact.

4.2.2 Generating Ground Truth Disparity Map

Along with the scenes we also generate the ground truth data. This is an important step since error in the ground truth data will make our evaluation invalid. The process is described below.

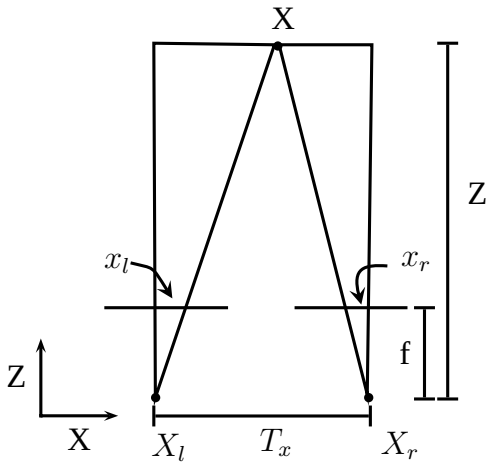


Figure 4.2: Disparity Computation

For each pixel in the left image, the depth value is used to compute the location of the corresponding pixel in the right image. Given the depth of the scene point to be Z, baseline distance T_x, near plane distance N and the scaling factor from image plane

to screen coordinate S_{pi}, disparity d is (Fig. 4.2) :

$$d = S_{pi}\frac{NT_x}{Z} \tag{4.1}$$

Since the camera displacement is horizontal, only the horizontal disparity is required. From the disparity value the corresponding pixel can be computed as $x_r = x_l - d$. The depth value of the corresponding pixel is then compared with the initial one. This value can either be the same or smaller. In the former case, the pixel is binocularly visible and has a valid disparity. In the latter case, the pixel will be occluded. In case if the corresponding pixel is farther away then the scene can be considered to be invalid. This does not happen in our scene generation process.

4.2.3 Generating Scale-Invariant Scenes with fixed γ

To generate the scale-invariant scenes with r^{-3} distribution, we first fix the number of different radii the objects will have and then divide the interval $r_{min} : r_{max}$ uniformly. Next the number of objects N_i of each radius r_i is chosen such that the resulting scene has a fixed γ and follows r^{-3} distribution. For r^{-3} distribution we have:

$$N_i = K\frac{1}{r_i^3} \tag{4.2}$$

For each class of objects of radius r_i, let the cross-sectional area be A_i and density η_i. Then for fixed γ we have:

$$\gamma = \sum_i \eta_i A_i = \frac{1}{V}\sum_i N_i A_i \tag{4.3}$$

From (4.2) and (4.3) we have:

$$K = \frac{\gamma V}{\pi \sum_j \frac{1}{r_j}}$$
$$\therefore\ N_i = K\frac{1}{r_i^3} = \frac{\gamma V}{\pi \sum_j \frac{1}{r_j}}\frac{1}{r_i^3}.$$

4.3 Scene Generation Experiments

In the following, we discuss the statistical similarities and differences between each scene. The statistics are calculated from 100 images of size 256×256. While generating the scenes the size of the objects (last column in Table 4.1) are chosen in such a way

that the smallest object projects to at least 4 pixels (half width is 2 pixels). Therefore in the non scale-invariant case this is determined by the size of object on the farthest plane. In the case of a scale-invariant scenes, this depends on the size of the smallest object on the farthest plane and the largest object on the nearest plane. The range of size in pixels is 2 to 128. In scene coordinates, the width range is 0.034 to 0.536 for a depth range of 2 to 8 and 0.134 to 2.14 for a depth range of 8 to 32. The objects and the background were texture mapped with 2×2 and 16×16 randomly generated textures respectively. The pixels colors were generated from a uniform distribution. In the ground truth disparity map of all the scenes the pixels are marked as either monocular or binocular along with the disparity value. A sample from each scene category and the corresponding disparity map are shown in Figure 4.3.

Figure 4.4 shows the percentage of binocular pixels in all the scenes. For our chosen parameter values the percentage of binocular pixels is always greater than 50%. From the plot it can be seen that smaller objects always decrease the number of binocular pixels. This effect is larger when the objects are closer to the viewer. Objects that are farther away increases the percentage of binocular pixels. The statistics for scale-invariant scene is mostly similar to large objects.

Figures 4.5, 4.6, 4.7, and 4.8 show the different statistics for each of the scenes that were generated. It can be seen that the curves for small objects ($r = 0.025$) at a closer range (2–8) have a sawtooth pattern. This is an artifact due to our scene generation process. Since the size of the projected images are rounded to the nearest integer, the actual width of the squares in the object space changes. For very small objects even the slightest change in square width can have a big impact. As a result the artifact is visible only for very small objects (half-width = 0.025). For larger objects this effect is minimal. If the mean and standard-deviation of the radius is considered for large and small objects we find that for the first scene ($r = 0.1$) they are $\mu \approx 0.1$ and $\sigma \approx 0.0032$, and for the second scene ($r = 0.025$) they are $\mu \approx 0.026$ and $\sigma \approx 0.0033$. The statistical similarities and differences between the scenes are given in the rest of this chapter. First some general observations. In all cases, the shape of the probability curve depends on γ. For a given disparity, the fraction of binocular pixels depends on the size of objects when all other parameters are held fixed. This can be understood from Eq. 3.6 where the probability is dependent on binocular visibility, which in turn has dependency on the size of the object (both directly and indirectly due to cross-sectional area). The

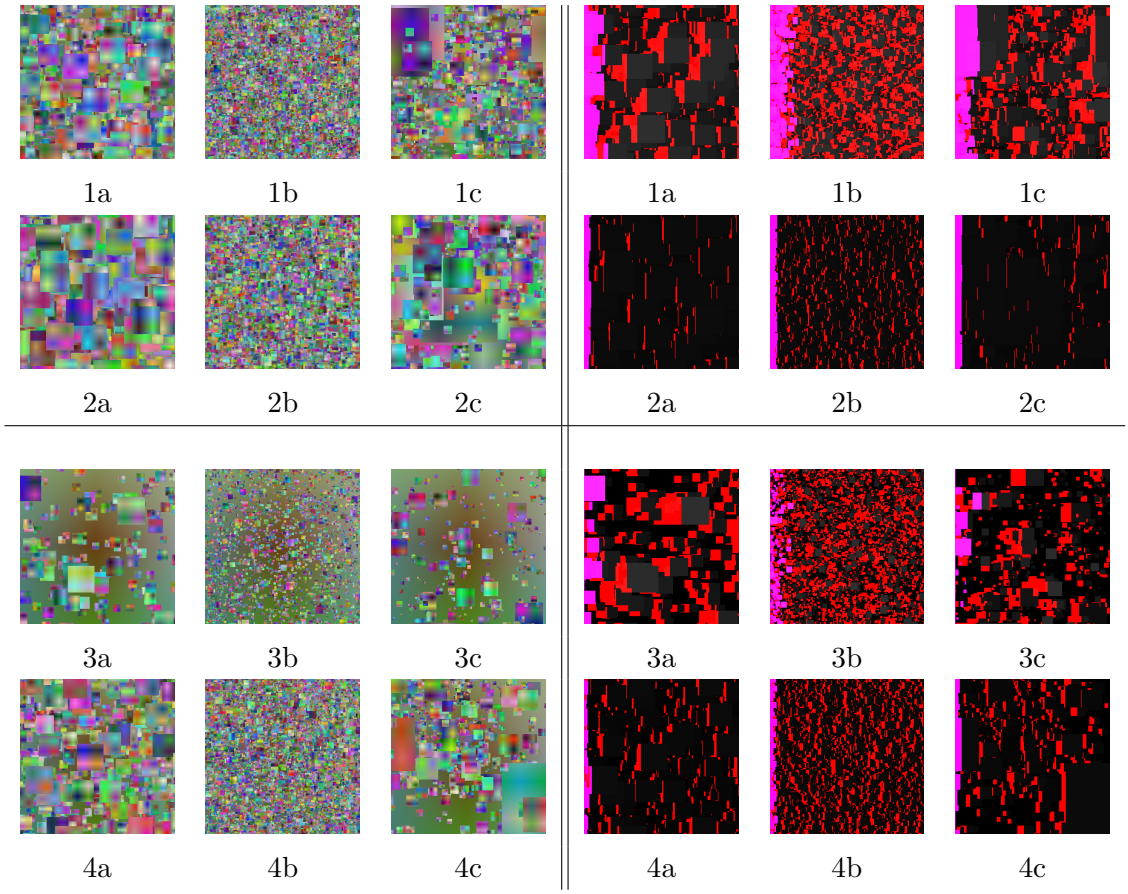


Figure 4.3: 1) Samples from the generated scenes (left view) and 2) the corresponding disparity map. Red denotes monocular pixels and magenta monocular pixels that are outside the view volume of the other view.

range of disparity varies between different scenes based on the depth range. Farther away objects (i.e. 8–32) have smaller range of disparities than closer objects (i.e. 2–8). The discontinuity statistics is largely affected by the size of objects. When all other parameters are held constant, the conditional probability of discontinuity is more for smaller objects than for larger objects. For the disparity difference probability it is easier to compare the negative log probability. This interpretation is closer to the smoothness penalty (Eq. 2.9). From the figures it can be seen that they are almost the same in all cases.

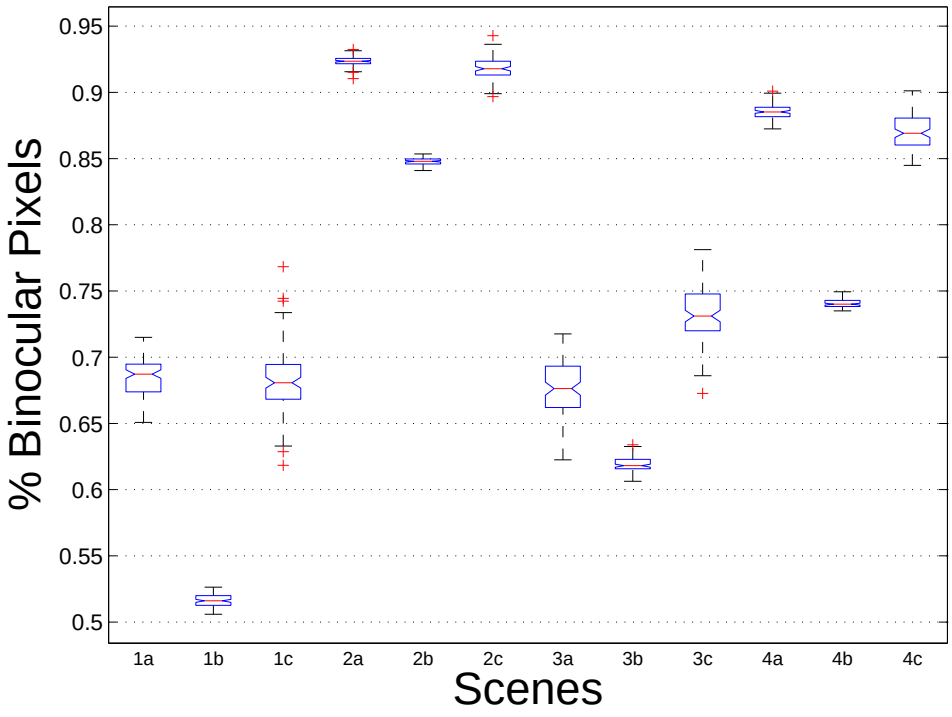


Figure 4.4: Percentage of binocular pixels for all the scenes. Red center line denotes the median, the lower and upper bounds of the box represents 25 and 75 percentile of the data. The red '+' denotes the outliers and the whiskers denote the extent of the data.

Scenes 1a, 1b, and 1c The first three scenes have $\gamma \approx 0.54$ and depth range between 2 to 8. The statistics are shown in Fig. 4.5. The all-pixel disparity probability curves are similar for all three scenes as expected. The binocular disparity for smaller objects shows a larger difference from the all-pixel one. The discontinuity probability for smaller objects is very large and greater than both large and scale-invariant scenes. More importantly the binocularly visible discontinuity is greater for almost the entire disparity range. All three negative log of disparity difference curves are very similar. In all cases, the curve rises linearly within the range $|f_p - f_q| = 0$ and 2. After that the increase is non-linear with the curve rising slowly at first and then rapidly. This indicates that even if the probability of large disparity difference is small, for a certain range, the probability is almost the same.

Scenes 2a, 2b, and 2c Fig. 4.6 shows the statistics of the next three scenes. This set has $\gamma \approx 0.54$ and depth range 8 to 32. The only difference from the first set is the range of distance. The disparity probability density is much larger than the first three scenes. This is because of having a smaller range of disparities. The curves are also steeper because the range of visibility decreases with distance. The discontinuity probability in all three cases is much larger than the previous set. In fact the small objects have the largest probability among all the scenes. This is primarily because the size of the projected image is much smaller than most of the scenes. All three $-\log p$ disparity difference curves look very similar. Like scenes 1a-c the increase is linear for $|f_p - f_q| \lesssim 2$ and after that the curve rises quickly to the maximum value because of the small range of disparities.

Scenes 3a, 3b, and 3c In this set, $\gamma = 0.1$ and depth range is 2–8. The statistics are shown in Fig. 4.7. Unlike all other scenes, the disparity curve is decreasing for this set. This indicates that farther objects (i.e. small disparity) are more visible than closer objects (i.e. large disparity). This implies that, due to the sparse nature of the scene farther away objects are less likely to be occluded by closer objects. Discontinuity probability is in general smaller than the previous two sets and almost flat. In the case of disparity difference the $-\log p$ has two peaks. This is primarily because of the sparsity of the scene and the range of being closer to the viewer. The chosen depth range increases the disparity resolution but because of the sparsity of the scene it is unlikely for two neighboring pixel disparities to differ between $2 - 10$. It should be noted that the model does not fit the data very well in all cases (especially discontinuity and disparity difference) because of the large amount of visible background. This is because the range of distance assumption that is made in Sec. 3.5.1 is not valid for these scenes.

Scenes 4a, 4b, and 4c In this set, $\gamma = 0.1$ and depth range is 8–32. Fig. 4.8 shows the statistics for this set. The disparity probability is larger than the first and third set and is more similar in shape to the first set. The reason is the same as the second set which is the depth range. Since the objects are far away more objects are visible within the view-volume. Furthermore the choice of depth-range reduces visibility as was seen before. This couple with the size of projected objects being closer to the first

set of scene makes the statistics behave similarly to the first set. However because of the smaller range of disparities, the probability values are higher.

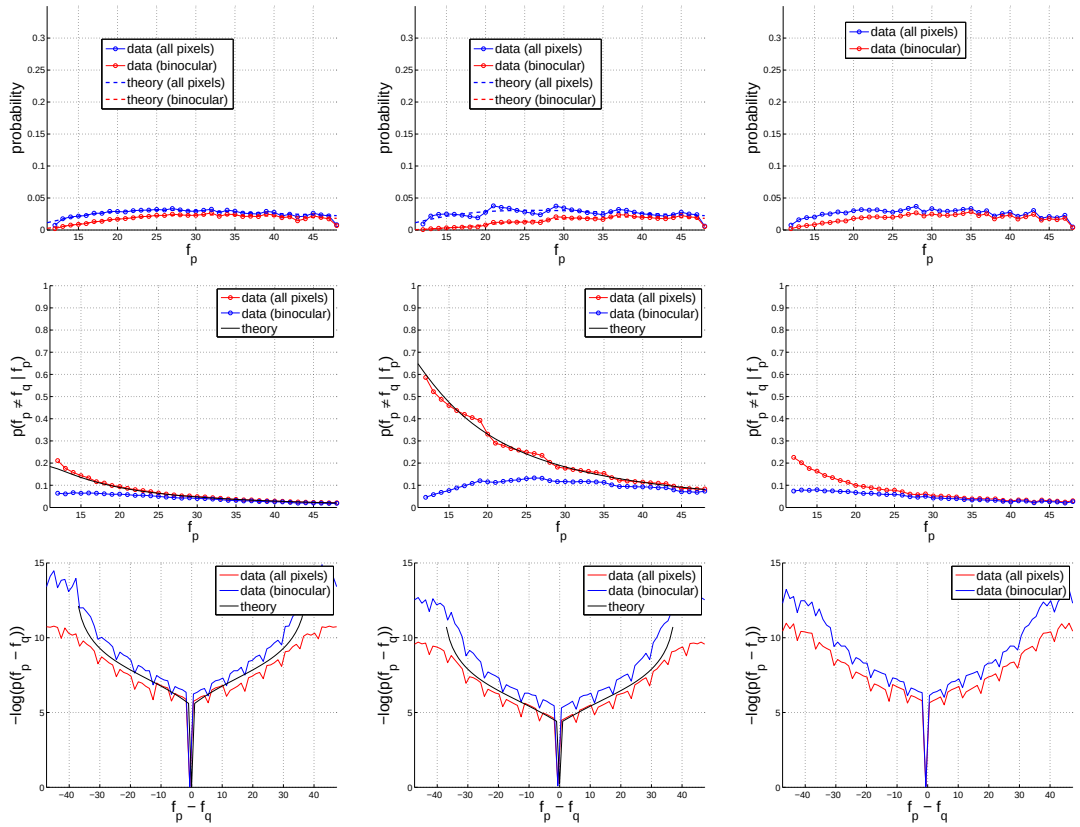


Figure 4.5: Disparity and Neighbor Statistics for Data and Model with $\gamma \approx 0.54$ with depth range 2 to 8

4.4 Discussion

To summarize, our objective was to generate scenes with a wide range of underlying statistics. To this end, we chose a set parameters and their values based on the model presented in Chapter 3. The scenes were generated in such a way that the projections are consistent and the ground truth disparity map has accurate integer disparities.

We also generated statistics from the synthetic scenes. We find that the parameter γ affects the shape of the disparity density curve as expected. Between different scenes

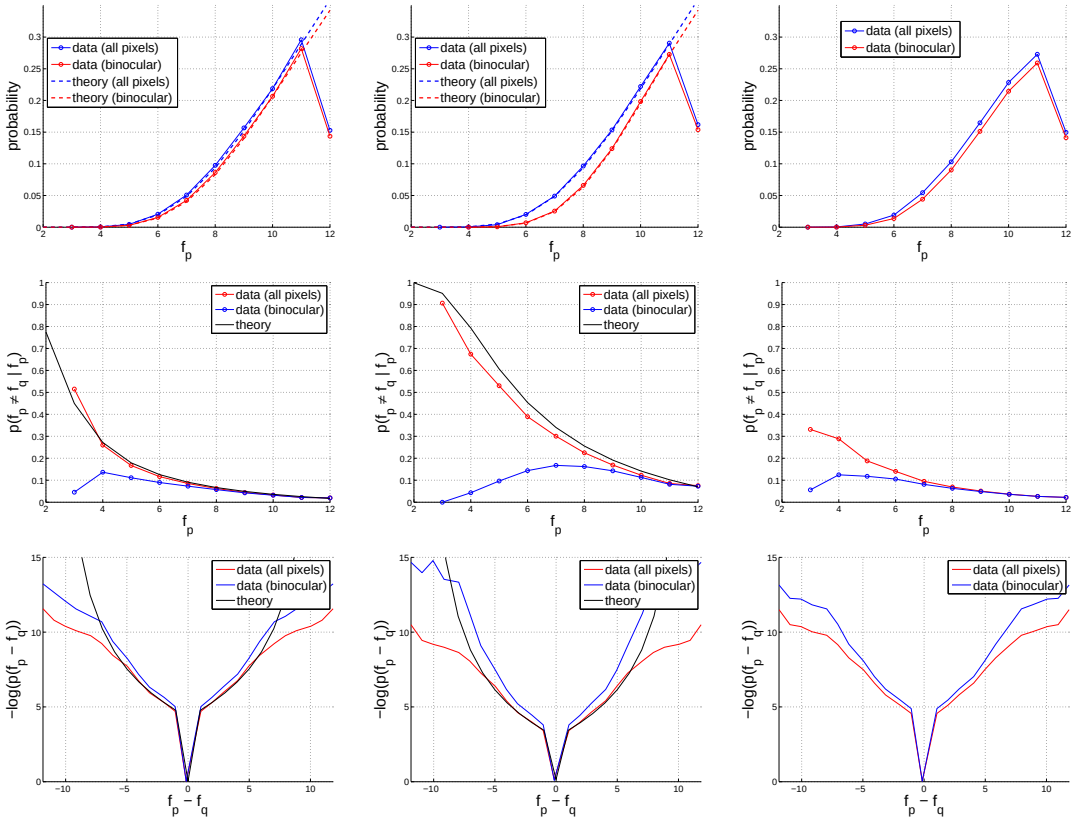
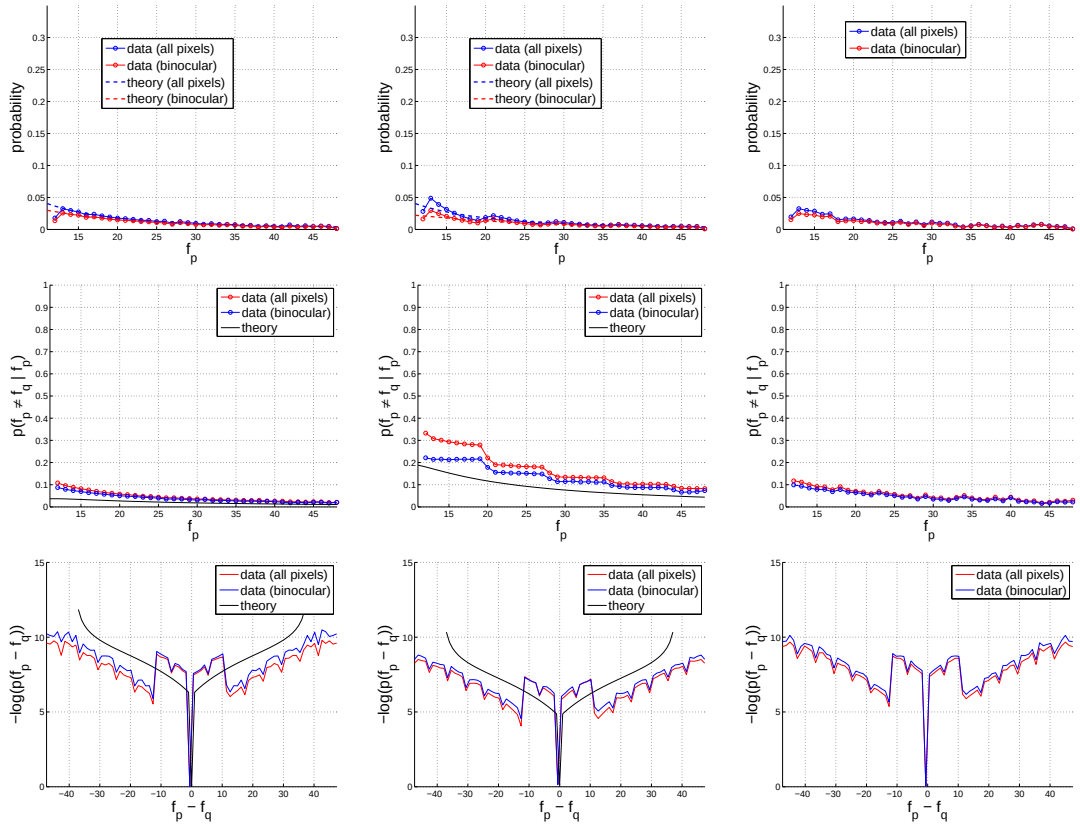


Figure 4.6: Disparity and Neighbor Statistics for Data and Model with $\gamma \approx 0.54$ and depth range 8 to 32

of the same γ, the discontinuity probability depends on the size of the objects. Smaller objects have higher discontinuities. It is interesting to see that the shape of the $-\log p$ of disparity difference curve does not change that much from scene to scene. It rises almost linearly within a small range of disparity difference. Then it stays almost flat within a certain range (for large disparities this range is also larger) and finally increases sharply.

From Fig. 4.4, it can be seen that the percentage of binocular pixels is always lower for smaller objects. This is primarily caused by the increase in discontinuity which happens because of very small projection. For objects that are farther away the percentage of binocular pixels is more. This is mainly because the disparity is small and as a result the amount of occlusion is also small. When small objects are closer to

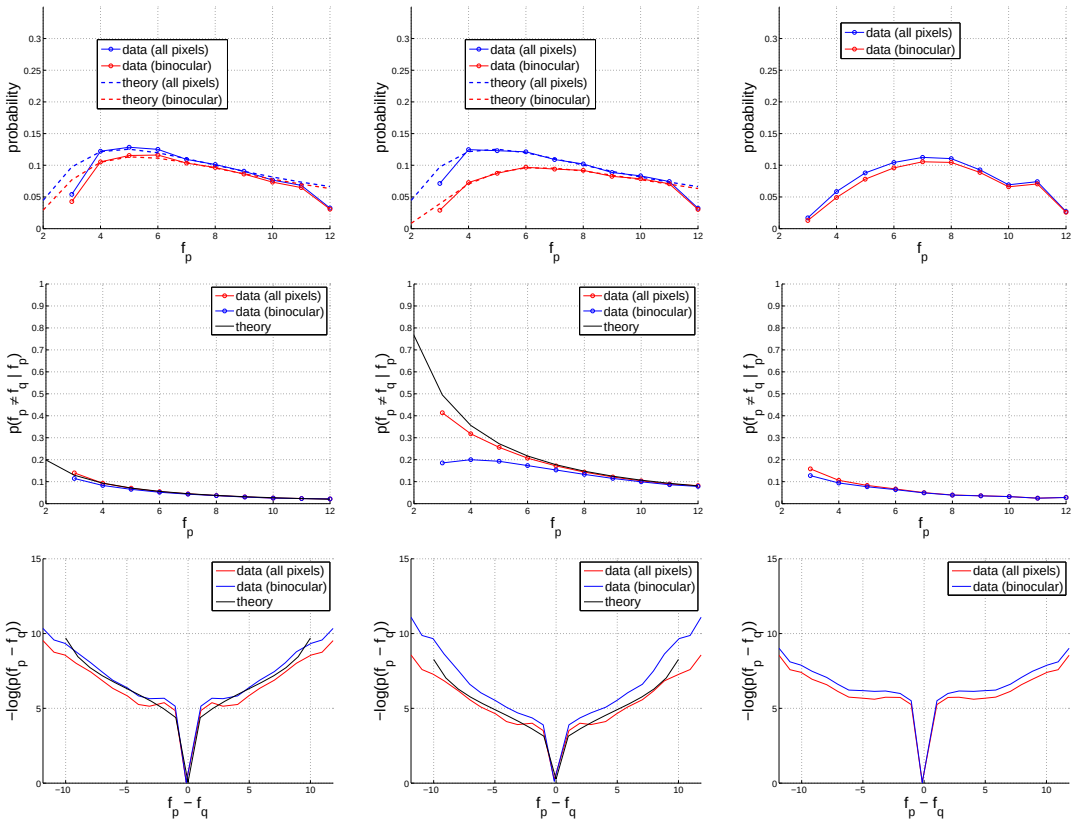


(3a) (3b) (3c)

Figure 4.7: Disparity and Neighbor Statistics for Data and Model with $\gamma \approx 0.1$ and z-range between 2 and 8

the viewer the disparity is large and almost always occludes an area equal to the size of the object.

In the next chapter, we look at the error statistics for these scenes and correlate those observations with underlying scene statistics.



(4a) (4b) (4c)

Figure 4.8: Disparity and Neighbor Statistics for Data and Model with $\gamma \approx 0.1$ and z-range between 8 and 32

Chapter 5

Performance Evaluation

This chapter presents the performance results of different methods for the cluttered scene stereo reconstruction problem. Previous chapters showed how cluttered scenes are modeled and synthetic scenes are generated. The goal of scene generation process was to generate scenes with widely varying statistics. The goal of this chapter is to apply a wide range of methods on those scenes and understand how the performance varies with different methods and input scenes.

Performance measurement is based on the accuracy of the output under different quality metrics. For this we need to decide on a set of quality metrics that are appropriate for cluttered scenes. The primary requirement for these metrics is that they are able to highlight the key characteristics of the scenes as well as the methods.

Statistically different types of cluttered scenes are used for the experiments (e.g. size and density of objects, and range of depth). Different types of parameters (e.g. data and smoothness cost type, smoothness weight, and maximum smoothness penalty) are also chosen for the algorithms to find out which set of parameters perform well in general. Primarily the two types of formulations that were discussed in Sec. 2.4 are evaluated. Their implementation is from [18] and [16].

The chapter is organized as follows: In section 5.1, the parameter choices for the algorithms are discussed. In section 5.2 the performance criteria are explained. The experimental results are presented in section 5.3. Finally this chapter concludes with a discussion of the results in section 5.4.

5.1 Algorithms and Choice of Parameters

In section 2.4 MRF based stereo approaches were divided into two categories: 1) basic formulation and 2) additional constraint based formulations.

The energy that is optimized in the basic formulation was discussed in Sec. 2.4.1. There are two terms in the equation: The *data term* that ensures photo-consistency and the *smoothness term* that ensures continuity between neighbors. The parameters that are of interest are: the norm of the data and smoothness term (k_d and k_s respectively), the maximum smoothness penalty V_{max}, the intensity gradient threshold ($I_{threshold}$), gradient penalty (λ_∇), and the smoothness weighting factor λ. The performance for the basic formulation is evaluated for: Expansion, Swap [14], and variants of Belief Propagation [15] such as BP-M, TRW-S, and BP-S [17, 19, 23, 33]. The different parameter values for the algorithms are given in Table 5.1. The implementation from [18] was used for obtaining the results.

Table 5.1: Algorithms and Parameters for Basic Formulation

Data term (k_d)	Birchfield-Tomasi with $k_d = 1, 2$
Smoothness exponent (k_s)	1, 2
Smoothness Max (V_{max})	1, 2, 10, 100
Gradient Threshold ($I_{threshold}$)	5
Gradient Penalty (λ_∇)	2
λ	1–120 with non-uniform intervals
Algorithms	Expansion, Swap, BP-M, BP-S, TRW-S

There are wide varieties of algorithms in the second category. In this thesis, only KZ1 and KZ2 [2, 3, 4] are evaluated. This is primarily because they have minimal set of constraints and therefore, can be used to easily understand the effect of occlusion and discontinuity. Their formulations were presented in Sec. 2.4.2 (Eq. 2.12 and 2.16).

KZ2 (Sec. 2.4.2, Eq. 2.16) considers pairs of pixels or voxels and assigns either 1 or 0 to each of those voxels. The data term computes the dissimilarity measure for voxels that are set to 1. The smoothness term is defined for interactions (Sec. 2.4.2) and does not depend k_s and V_{max}. Besides the data and smoothness terms, the formulation has additional constraints for ensuring uniqueness. The set of parameters that are of

interest for KZ2 are: k_d, gradient threshold ($I_{threshold}$), gradient penalty (λ_∇), and the smoothness weight factor λ. There is another variable in the data term, K, which contributes to the constant occlusion penalty. But this parameter has a fixed value in our experiments.

In KZ1 (Sec. 2.4.2, Eq. 2.12), all the pixels in the input images are used, and the disparity is computed symmetrically. Like KZ2 it has an additional constraint that ensures consistency between disparity assignments. The parameters that are of interest to us are: k_d, k_s, V_{max}, $I_{threshold}$, λ_∇ and λ.

The implementation of KZ1 and KZ2 is from [4]. Both the algorithms use Expansion for optimizing the energy function. Because of the differences in implementation the error rates for the two formulations cannot be directly compared. This is why, in the second set of experiments, we include the results for basic formulation with Expansion algorithm. Also, because of the differences in implementation, the range of λ values is different from the first set. It should be noted that not all possible combinations of parameter values were used for this set of experiments. The set of values is given in Table 5.2.

Table 5.2: Algorithms and Parameters for Visibility Formulation

Data term (k_d)	Birchfield-Tomasi with $k_d = 1, 2$
Smoothness Exponent (k_s)	1
Smoothness Max (V_{max})	1, 2, 10, 100
Gradient Threshold ($I_{threshold}$)	5
Gradient Penalty (λ_∇)	2(Expansion), 3(KZ)
K	5λ
λ	1/16, 1/8, 1/4, 1/2, 1–10, 15
Algorithms	KZ1, KZ2

5.2 Performance Criteria

A natural question in such a performance evaluation study is how to compare the performance of different algorithms and parameter combinations. The main objective of such comparison is to understand where and how the errors occur and by how much.

5. PERFORMANCE EVALUATION

Sec. 2.5 gave an overview of some of the previously used performance metrics. Besides the commonly used error metric like fraction of mislabeled pixels, we also use some additional metric that measures specific errors. However, these specific error metrics are not mutually exclusive and therefore do not sum up to the total error.

Binocular Error: One widely used performance metric is the percentage of mislabeled pixels. In our experiments we give more emphasis on the percentage of mislabeled *binocular* pixels, which was also used in [23]. This is the ratio of mislabeled binocular pixels to the total number of binocular pixels. The primary reason for using this metric is that the algorithms and energy formulations do not inherently extrapolate disparity labels for monocular pixels. The basic formulations assume that pixels are binocularly visible. Formulations with visibility constraints can identify occluded pixels or inconsistent disparity assignment, and are usually better at labeling monocular pixels. But neither of the formulations explicitly label monocular pixels (however, explicit assignment can be done in the post-processing step in some cases). Therefore, we mostly focus on errors in binocular pixels.

Off-by-one Errors: In many performance studies the error threshold is $\tau = 1$. This is primarily because the input dataset, which contains real-scenes do not have exact integer disparities. However, in our case (recall Sec. 4.2.1), the scenes do not have any fractional disparity. As a result, we set $\tau = 0$, and specifically look for cases where the assigned disparity label differs by 1 from the ground truth.

Binocular Monocular Error (BME): In this metric, the percentage of mislabeled binocular pixels with at least one monocular neighbor (ignoring the labeling of the monocular pixel) is considered. More specifically, it is the ratio of mislabeled binocular pixels with at least one monocular neighbor to the total number of binocular pixels. The reason for using this metric is that, when the smoothness term considers a pair of neighboring pixels, it is implicitly assuming that both the pixels are visible and the neighboring pixel is correctly labeled. Therefore binocular pixels with monocular neighbors are more likely to be mislabeled especially in the absence of any visibility constraints. In our analysis, we find the general trend in this error, and the contribution it has to the total error. However, it is not possible to say anything conclusive just

from this error metric, because there can be other binocular pixels in the neighborhood which can (assuming they are correctly labeled) reduce the influence of the monocular pixel (assuming the monocular pixel is mislabeled).

Binocular Discontinuity Error (BDE): This is the percentage of mislabeled binocular pixels with at least one binocular neighbor with a different disparity out of all the binocular pixels. The smoothness term penalizes these pixels because in this case $f_p - f_q \neq 0$. Such error will give an estimate of how well different methods can handle discontinuities. For scenes with large number of discontinuities (e.g. scenes with small objects) these errors are more likely to have an impact on the overall performance. It should be noted that discontinuity between monocular neighbors are not considered for the same reason why monocular errors are not considered: which is the methods do not explicitly handle monocular pixels.

Binocular Continuity Error (BCE): This is the percentage of mislabeled binocular pixels whose all 4 equi-disparity binocular neighbors out of the total number of binocular pixels. The smoothness term prefers such pixels because here $f_p - f_q = 0$. As a result, there is no penalty. This error measure is useful when there are additional constraints, because it allows us to determine if the additional constraints are affecting the smoothness constraint in any way.

Correlation between Error Rate and Energy: If an energy function accurately models a stereo problem, then the energy of the solution will be close to that of the ground truth. Tappen and Freeman in [23], compared the ground truth and output energy, and showed that the basic stereo formulation does not model the stereo reconstruction problem very accurately, because the ground truth energy is always larger than the output energy. In that work, visibility constraint based formulations were not evaluated. Therefore, we use the same measure in our approach primarily to observe how good the visibility constraint based approaches are.

5.3 Experimental Results

This section presents the results obtained from running experiments with basic (Sec. 5.3.1), and additional (Sec. 5.3.2) constraints.

For each algorithm and parameter combination, the error and energy statistics are averaged over 5 sample image pairs with the same underlying statistics. The reason for using a small number of images is that variations of BP can take a long time to converge. A smaller number of samples make the average and standard deviation of the error rates unreliable. However, the error bars are usually consistent across different parameters which implies that the results are fairly reliable. Furthermore we are only interested in the qualitative trends in the result.

To make the experiments more realistic, Gaussian noise with $\mu = 0$ and $\sigma = 5$ was added with appropriate clamping to the right view. Because of this addition of noise, formulations with gradient threshold (Eq. 2.9) gives better result than those without gradient threshold. Therefore, in the rest of the thesis, only the results for gradient threshold are presented.

To understand the effect of noise and texture pattern on the error statistics, the error rate for the scenes with only one object, namely a textured background with disparity 0 (distance of the background is 2000 units from the camera and stereo baseline $T_x = 0.2$) is examined. Fig. 5.1 shows the non-zero binocular errors.

It can be seen that the error rate goes to zero as λ increases. Smaller λ puts more weight on the data term. Since the corresponding pixels are noisy, the data term alone is not sufficient for correct labeling. In the first row, Expansion, Swap and TRW-S can be seen to fall sharply to 0. In the second row of the figure, KZ1 is absent because its error rate is 0 for all values of λ. This is due to the additional visibility term that penalizes inconsistent labeling. Not all types of visibility constraint will result in zero error, though. This is evident from the non-zero error rate of KZ2. However, KZ2 performs slightly better than just the basic formulation e.g. its error rate decreases much more rapidly than that of Expansion.

In Table 5.1 and 5.2, the set of parameters used in the experiments and their values were listed. However, we only consider the results for $k_d = 2$. This is because the results for $k_d = 1$ and $\lambda < 5$ is similar to $k_d = 2$ and the full range of λ. They are similar in the sense that the relative performance between the different methods do not change much with k_d. To illustrate this point, sample plots for both $k_d = 1$ and 2 are shown in figure 5.2. The plots show that the relative performance between different methods is the same in both column 1 and 2. For $k_d = 1$ the optimal is achieved for small λ values (with respect to the appropriate scale of λ). However for $k_d = 2$, the

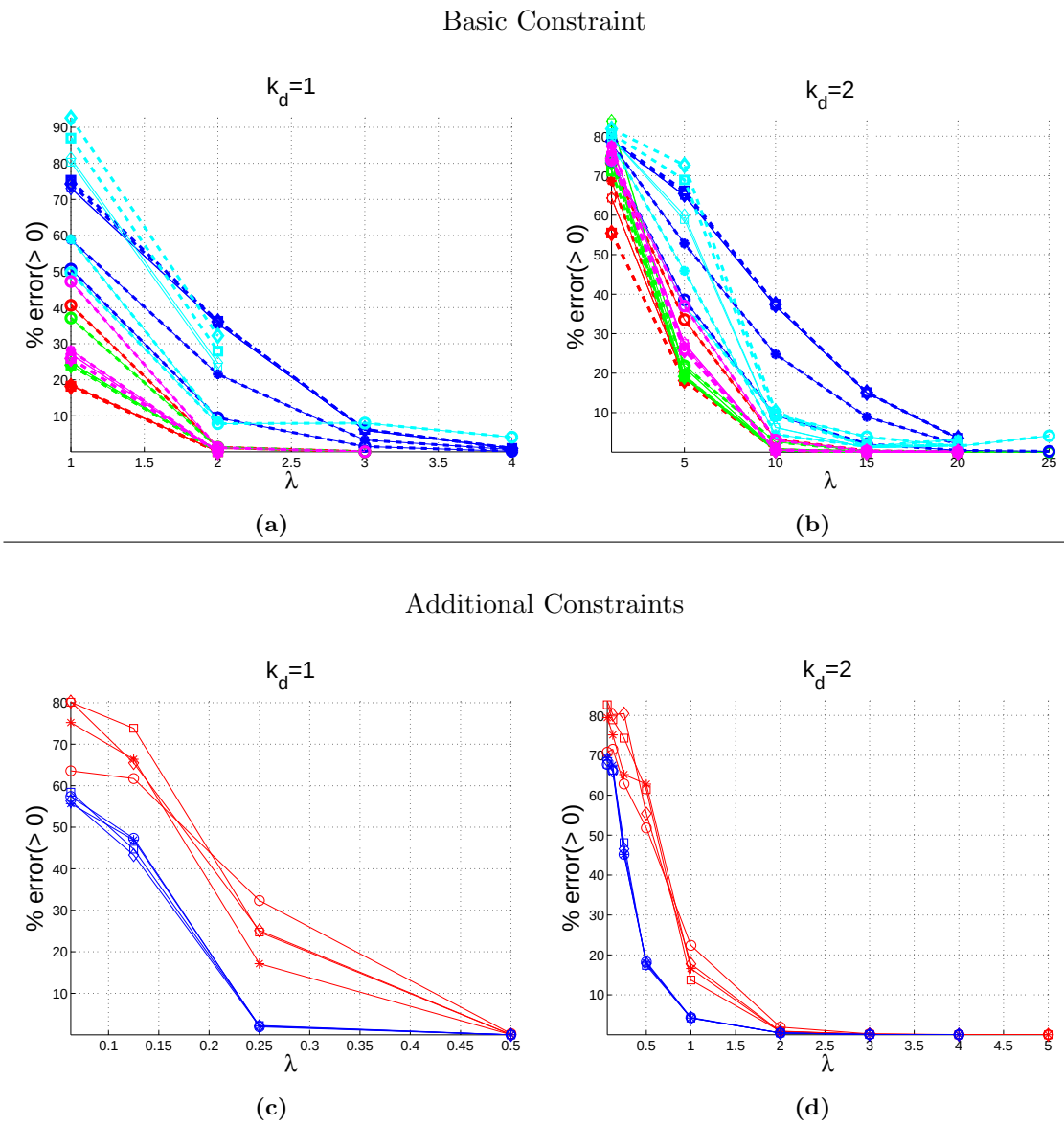
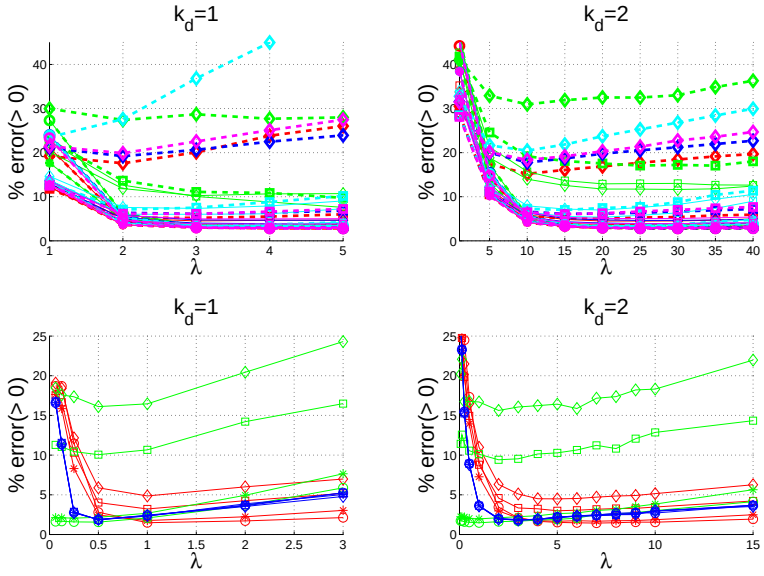
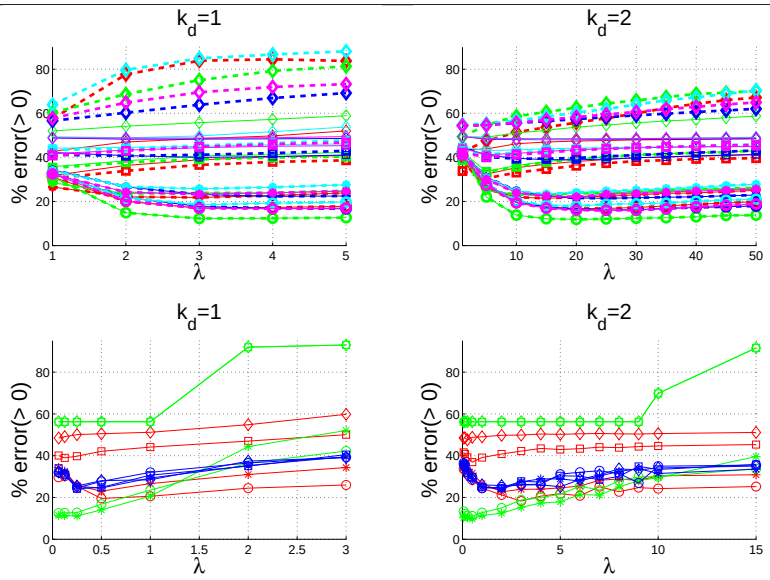


Figure 5.1: Mean error rate for scenes with only a textured background. Only the methods with non-zero error rate are shown in the figure. In the first row Expansion, Swap, BP-M, TRW-S, BP-S are colored red, green, blue, magenta and cyan respectively. $k_s = 1$ and 2 are represented by solid and dashed lines respectively. V_{max} values 1,2,10 and 100 are represented by O, *, □ and ⋄ respectively. In the second row, Expansion and KZ2 are colored red and blue respectively. The rest of the notations are the same. The λ scales are different between the two sets as was mentioned in Sec. 5.1



(a) Scene 1a



(b) Scene 3b

Figure 5.2: Comparison between $k_d = 1$ and 2 for (a) scene 1a and (b) scene 3b. The first row in each case is for the basic formulation and the second one for the additional constraints.

optimal λ range is much larger and the error rate grows slowly with λ. The fact that the relative performance between different forms of smoothness and algorithm combination

is almost the same, indicates that the form of smoothness and algorithm are more important. Since $k_d = 2$ allows a larger range of λ values, it is possible to finely tune λ to get an error rate that is better than $k_d = 1$. However the main disadvantage is the difficulty of finding an optimal λ in such a large range.

Now we investigate how this performance varies with different scenes. In the analysis that is presented below we use the plots given in pages 65 to 76. The figures show results for total error (5.4 and 5.5), off-by-one error (Figs. 5.6 and 5.7), BME (Figs. 5.8 and 5.9), BDE (Figs. 5.10 and 5.11), and BCE (Figs. 5.12 and 5.13). Finally, the energy statistics are shown in Figs. 5.14 and 5.15. In each case, results for all the scenes are shown. The columns from left to right in the figures, represent large, small and scale-invariant objects. The rows represent scenes with fixed γ from Table 4.1. In each case the y-axis scale is fixed for easier comparison. The colors and symbols carry the same meaning as before.

5.3.1 Formulation with Basic Constraints

Performance of the basic formulation based methods for different error metrics is discussed in this section. The results are from Figs. 5.4, 5.6, 5.8, 5.10, 5.12. To reduce clutter we only plot the results for Expansion algorithm with $k_s = 1$, $k_d = 2$ and different values of V_{max}. The extent of error considering all possible parameters are summarized in Tables 5.3, 5.4, 5.5, and 5.6. The figures can be used to obtain an idea about how the error changes with λ.

The tables mainly lists the minimum mean binocular error (min error), range of mean error (either binocular, BME, BDE, or BCE), range of off-by-one binocular errors (range(=1)), contribution to the binocular error (contrib.) and the range of λ for which the observations are valid. The range of errors are defined by the minimum and maximum extent of errors. They give an indication of the range of error we can expect to have for any given combination of algorithm and formulation. In the following, we discuss the performance of the methods in terms of different types of errors.

Binocular Error Table 5.3 summarizes the minimum average error, range of errors, range of off-by-one errors and the range of λ for which these observations are valid. It is based on Figures 5.4 and 5.6. It can be seen from Figure 5.4 that for each row the

general trend in the total error and the minimum average error are very similar. The error statistics depend largely on γ and depth range.

Table 5.3: Summary of Total Error Statistics for Basic Formulation. Grayed out rows represent scenes with depth range 8–32

Scene	min error	range	range (=1)	λ
1a	2.5	0 – 55	0 – 30	1 – 120
1b	3	2.5 – 70	1 – 18	1 – 120
1c	2.5	1.5 – 65	0 – 30	1 – 120
2a	16	10 – 40	12 – 35	5 – 120
2b	18	15 – 30	8 – 26	5 – 120
2c	25	8 – 55	8 – 50	5 – 120
3a	5	2 – 65	0 – 30	1 – 120
3b	12	10 – 85	1 – 20	1 – 120
3c	3	1 – 75	0 – 40	1 – 120
4a	9.5	7.5 – 30	4 – 20	10 – 120
4b	6.5	6 – 28	3 – 20	10 - 120
4c	8.5	6 – 46	1 – 20	5 – 100

In terms of the algorithms and configurations being used, smaller values of V_{max} (e.g. $V_{max} = 1$ and 2) perform well in most cases. The second (scenes 2a–c) and fourth (scenes 4a–c) rows show that $V_{max} = 10, 100$ with $k_s = 1$ perform equally well as $V_{max} = 1, 2$. In some cases, they are considered to perform equally well, either because of the large variation in error, or because the average error is approximately within 5% of the minimum error. The range of error for non scale-invariant objects that are farther away (i.e. depth range 8–32) is usually small ("range" column in Table 5.3 especially for scene 2a,b and 4a,b). However their minimum mean error is usually larger ("min error" column of the same table) than closer scenes.

Performance of the algorithms also depend on k_s for large V_{max} (e.g. $V_{max} = 10, 100$). In every case, $k_s = 1$ perform better than $k_s = 2$. For $V_{max} = 1, 2$, the performance for both k_s is the same. As for the algorithms: Expansion, BP-M, and TRW-S perform equally well in most cases.

In case of the off-by-one errors (see col. "range (=1)" of Table 5.3, and Fig. 5.6): the effect is more prominent when the objects are farther away from the camera. This

can be seen from the larger off-by-one error exhibited by scenes 2a–c and 4a–c. In general the off-by-one errors increase with λ as can be seen from Figure 5.6.

Small objects have slightly larger error rate in general (scene 4b is the only exception). Even if the off-by-one errors are ignored the behavior remains the same. In general the scale-invariant scenes have a large variation for total error as well as off-by-one error.

BME For binocular monocular errors (BME), the differences between the columns in Fig. 5.8 are considered. The figure is summarized in Table 5.4. The columns in the figure represent scenes with different sized objects: large, small and scale-invariant from left to right respectively and corresponds to gray rows in Table 5.4. The minimum error rate for small objects is usually larger than the large and scale-invariant scenes. Furthermore the contribution of this type of error is also greater than other scenes as can be seen from the contribution column of Table 5.4(the data is from Sec. A.2). The error rate can be correlated with the percentage of binocular pixels shown in Fig. 4.4. If the percentage of binocular pixel is small then the error rate and contribution is more and the opposite is true for larger fraction of binocular pixels.

Table 5.4: Summary of Binocular Monocular Error Statistics. Grayed out rows represent scenes with small objects

Scene	range	contrib.	λ
1a	0 – 6	10 – 55	10 – 120
1b	1.5 – 30	35 –75	1 – 120
1c	0 – 10	5 – 50	1 – 100
2a	0.5 – 2	3 – 8	5 – 120
2b	2 – 5.5	12 – 21	1 – 120
2c	0.8 – 2.6	2 – 10	1 – 100
3a	1 – 8	8 – 35	1 – 120
3b	7 – 35	35 – 65	1 – 120
3c	0.5 – 10	10 – 55	1 – 120
4a	1 – 3	10 – 15	10 – 120
4b	2 – 12	30 – 45	10 – 120
4c	1 – 4.5	8 – 20	5 – 100

BDE The columns in Fig. 5.10 are also compared for binocular discontinuity error. Table 5.5 summarizes the results from the figure. As before the errors are larger for small object scenes (middle column in the figure and grayed out rows in the table). The contribution of the error for smaller objects is also much greater than the large and scale-invariant ones (left and right columns).

Table 5.5: Summary of Binocular Discontinuity Error Statistics. Grayed out rows represent scenes with small objects.

Scene	range	contrib.	λ
1a	0.5 – 40	20 – 100	1 – 120
1b	2 – 70	65 – 100	1 – 120
1c	1 – 40	15 – 100	5 – 100
2a	2 – 8	12 – 30	15 – 120
2b	5 – 16	35 – 85	15 – 120
2c	2 – 10	10 – 40	15 – 100
3a	2 – 40	10 – 75	15 – 120
3b	7 – 80	50 – 95	15 – 120
3c	1 – 50	10 – 80	20 – 120
4a	2.5 – 15	20 – 65	10 – 120
4b	4 – 24	55 – 90	20 – 120
4c	3 – 20	15 – 70	10 – 100

BCE Table 5.6 summarizes the results for binocular continuity error shown in Fig. 5.12. In this case, the small object scenes (middle column in the figure and grayed rows in the table) have relatively smaller error. The contribution to the total error is also small.

Energy The energy plots are shown in Figure 5.14. It is evident in each case that the ground truth energy is much larger than the optimal energy found by the algorithms for different forms of energy terms. The reason behind this is that, in most of the scenes there are a good number of monocular pixels (Fig. 4.4). In the basic formulation there are no constraints for assigning labels to them. As a result, labels which

Table 5.6: Summary of Binocular Continuity Error Statistics. Grayed out rows represent scenes with small objects.

Scene	range	contrib.	λ
1a	0 – 10	0 – 80	1 – 120
1b	0 – 1.5	0 – 25	1 – 120
1c	0 – 35	0 – 80	1 – 100
2a	10 – 30	65 – 85	10 – 120
2b	2 – 15	5 – 55	1 – 120
2c	5 – 45	55 – 90	15 – 100
3a	0.5 – 35	20 – 80	15 – 120
3b	0.5 – 10	1 – 25	15 – 120
3c	0.5 – 45	15 – 80	10 – 100
4a	5 – 10	30 – 70	10 – 120
4b	0.5 – 3.5	5 – 40	10 – 120
4c	1 – 30	30 – 80	10 – 120

minimizes the overall energy is chosen, resulting in a large gap between ground truth and reconstruction.

5.3.2 Formulation with Visibility Constraint

As was mentioned before only KZ1 and KZ2 methods are considered under the visibility constraint formulation. The implementation is from [4]. Both the implementations use Expansion for finding a solution to the energy function. For comparison purposes the results for Expansion algorithm for the basic formulation are also shown. In this section, by Expansion algorithm we always mean the basic formulation with Expansion. Since this section considers three different formulations (i.e. one basic and two with different visibility constraints), only the relative performance between the formulations are compared instead of the range of errors and error contribution. The objective of the analysis (as before) is to understand how the error rate varies between scenes for different range of parameters and if these variations can be explained by the underlying statistics of the scenes. Figs. 5.5, 5.7, 5.9, 5.11, 5.13 are used in the analysis.

Expansion for Basic Formulation First we consider the performance of the Expansion algorithm for the basic formulation and ensure whether it gives the same error rate as before. From the total error (Fig. 5.5) the characteristics of the result (e.g. shape of the curve, range of values, and ordering of V_{max} etc.) are similar. Like before $V_{max} = 1$ performs better in general and this is followed by $V_{max} = 2, 10$ and 100. The off-by-one error (Fig. 5.7) is more for farther objects than for nearer objects. For BME and BDE (Figs. 5.9, and 5.11),the error for small objects (middle column) is more than the large and scale-invariant objects (left and right columns). For BCE (Fig. 5.13) small objects have lower error rate than large and scale-invariant scenes. This ensures that the Expansion algorithm with the basic formulation has the same behavior as before. Furthermore it shows the consistency of the results across different implementations. Since the Expansion algorithm performs well in general in the first set of experiments and it has the same characteristics in the second set, comparing KZ1 and KZ2 with the Expansion algorithm is sufficient for comparing the two types of formulations.

KZ2 For KZ2 there is no dependence on V_{max} or k_s (see Eq. 2.16) in the smoothness term. Therefore, all forms of smoothness term have the same error rate (see Fig. 5.5). The performance of the algorithm depends only on the type of scene and the chosen λ parameter. In most cases, KZ2 has the same minimum error rate as Expansion. For scenes with $\gamma = 0.1$ and depth range 2–8 (i.e. scenes 3a–c in Figs. 5.5), the difference in error rate is $\lesssim 2$. For these cases, KZ2 also has larger off-by-one error. For $\lambda < 5$, the error is primarily due to the background as can be seen from Fig. 5.1d. The above observations are also true in the case of BME and BDE errors (Figs. 5.9 and 5.11). In the case of continuity error (Fig. 5.13), KZ2's performance is significantly better even for scenes 3a–c.

KZ1 Like the basic formulation, KZ1 performs best for small V_{max} especially for $V_{max} = 1$. $V_{max} = 10, 100$ only performs well for large and scale-invariant objects that are far away (scenes 2a,2c, and 3c in particular). Larger V_{max} performs particularly worse than smaller V_{max} for scenes with objects closer to the viewer (e.g. scenes 1a–c, 3a–c).

KZ1 performs particularly well for small λ values. Most other methods have large error for these values. This is partly due to it being relatively insensitive to noise and texture pattern as was seen in Fig. 5.1d. For larger λ the error increases but the rate of increase is usually slow. From the off-by-one error plots, in most cases KZ1 has smaller off-by-one error for small λ (except in scenes 2b and 4b). This would explain the relatively good performance of KZ1 for small λ values.

In most cases the total error for KZ1 is comparable to Expansion. However in cases where objects are far away and have significant depth discontinuities (e.g. scene 2b, 4b), KZ1's performance is relatively worse than Expansion (but always within $\lesssim 2$). In fact the performance for BME, BDE and BCE are also not better for those cases. These scenes also cause larger off-by-one errors for KZ1 but the error rate does not improve even when $\tau = 1$ is considered (see Sec. A.3).

In the case of binocular monocular error, similar to the basic constraints KZ1 has larger error for smaller objects than for large or scale-invariant objects. Compared to the Expansion algorithm, it performs well for scene 3b ($\lambda = 0.1$ and depth range 2–8). This scene usually has larger error in all the error metric. In the case of basic formulation Swap algorithm has the best error rate for this case (≈ 6). It can be seen that the error rate for Expansion is almost the same as in the previous set of experiments. Therefore, KZ1 (error ≈ 5) marginally performs better than the basic formulation. For other scenes with small objects, KZ1 does not perform better than Expansion. For large and scale-invariant scenes KZ1's BME performance is as good as Expansion or slightly better.

For BME, KZ1 exhibits exactly the same trend in performance as BCE but with larger error.

In the case of binocular continuity, KZ1's performance is comparable to Expansion and KZ2 in most cases. In general KZ1 does not perform very well when $\gamma = 0.1$. Scenes 3a, 4a, and 4c give examples where binocular continuity error for KZ1 is particularly large.

Energy Fig. 5.15 shows the energy plots with ground truth energy. It is important to note that for visibility constraints the difference between the ground truth and output energy is smaller. This experimentally shows that visibility constraints are able

to model the underlying statistics better. Furthermore, the plots show that the KZ1 formulation is much closer to the true model than any other formulation.

5.4 Discussion

In this section, we summarize the results of the experiments, generalize some of the observations and answer the questions that were posed at the beginning of this thesis in Sec. 1.4.

To summarize, we have significantly reduced the choice of parameters that can be used in practice to obtain good stereo reconstruction of cluttered scenes. For the data term, we found that $k_d = 1$ and 2 give similar result except the λ ranges are different. $k_d = 2$ gives good result for a wider range of λ values i.e. error increases very slowly. This motivated us to use only $k_d = 2$ in all the analysis.

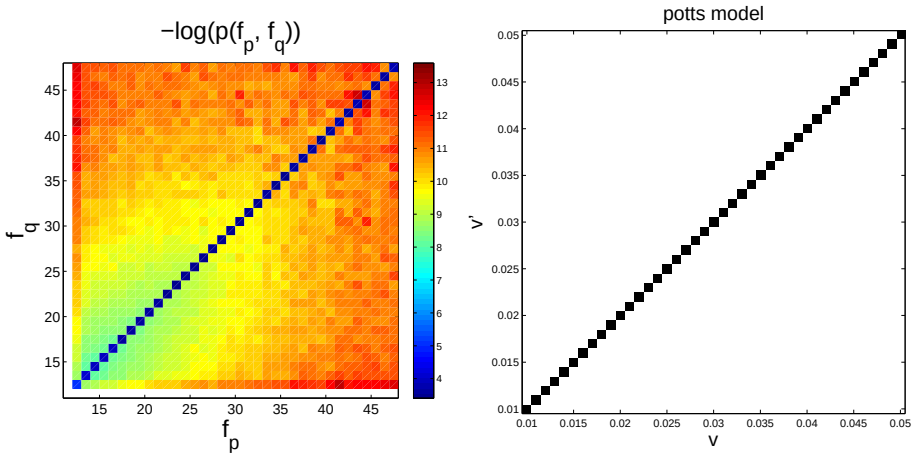


Figure 5.3: 1) $-\log p$ of joint probability for scene 1a. 2) $V_{max} = 1$ or Potts model.

For the smoothness term, we found $k_s = 1$ to perform better than $k_s = 2$ in almost all cases. As for the maximum smoothness penalty, smaller V_{max} especially $V_{max} = 1$ always perform better than larger V_{max} (i.e. ≥ 10). This is surprising considering the fact that the joint probability of disparity decreases smoothly away from the diagonal, whereas in case of $V_{max} = 1$ its a sharp increase (Figure 5.3). It can partly be justified by the negative log probability of disparity difference statistics (see 3rd row of Figs. 4.5

to 4.8). The rise in penalty is usually linear within a small range of disparity differences. After that it is relatively flat and rises sharply for very large disparities. This implies that very large disparity differences are very unlikely. A certain range of disparity difference is equally likely and the probability of small disparity differences change very rapidly. Since the linear rise happens only within a small disparity difference (in most cases $|f_p - f_q| \leq 2$), larger V_{max} does not improve the result and in some cases worsens it.

The off-by-one error mainly occurs when objects are far away. It should be noted that this error is not due to the range of disparity being decreased. For instance the range of disparity decreases when baseline is decreased (Fig. A.1). But it does not exhibit large ($\gtrsim 5$) off-by-one error. Another possibility could be the noisy texture pattern. But the amount of noise is independent of depth. The texture pattern however, depends on depth. The color gradient for smaller objects (caused due to foreshortening) is larger than that of larger objects. For a single large texture it can be seen that the off-by-one error is relatively small for algorithms like Expansion, Swap, TRW-S, and BP-S. So, it is likely that the color gradient contributes to the off-by-one error to some extent.

Naturally if the percentage of binocular pixel decreases, then the total error (considering all pixel) increases. This is true because the percentage of monocular pixels increases, and not all methods are able to handle monocular pixels properly (middle column of Fig. 5.16 and 5.17). But does it affect the fraction of mislabeled binocular pixels? We found that, the percentage of binocular pixels does not affect the binocular error rate. From Fig. 4.4 we can see that the percentage of binocular pixels can be small for scenes with smaller objects. The smallest percentage is for scene 1b ($\gamma = 0.54$, $r = 0.025$ and depth range 2–8). If we look at the binocular error rate we can see that it is not the largest. Furthermore, the first set of scenes (i.e. scene 1a–c) performs better in general in both sets of experiments (i.e. with basic and additional constraints). On the contrary, scenes 3a–c which have $\gamma = 0.1$ and depth 2–8 do not perform well in all cases. Most of the error for these scenes is caused by the visible background. We argue that this is primarily because the disparity difference probability assumed in the smoothness term (e.g. Potts model) does not capture the true probability. For instance, when the objects are between 2–8, the disparity range is 12–48. The disparity for the background is 0. Because of this, there will be significant number of neighbors

with disparity difference 12–48. This can be seen in the $-\log p(f_p - f_q)$ plot in Fig. 4.7, where the penalty suddenly drops near $|f_p - f_q| = 12$ as was discussed in Sec. 4.3. When the objects are farther away the background is less visible and there is no abrupt drop in penalty. Since this type of change is not assumed by any of the models, the binocular error rate for these scenes is usually worse than all other scenes. In fact it is not possible to capture this behavior without using some type of adaptive term. It can be seen that the cluttered scene model also does not capture this abrupt change.

Is there any correlation between the scene statistics and error statistics? The answer is yes. We have seen that there is a correlation between the underlying statistics and the error statistics. In fact the error rate mainly depends on the γ and depth range. For example, even when the baseline is decreased for scenes 3a–b, the error rates are similar (Figs. A.2, A.3).

The correlation between the scene statistics and error rate is an important characteristic. It implies that if we can model a natural scene using synthetic scenes then it might be possible to find the set of parameters that works well for the synthetic scenes and will also work well for the real scene.

As for methods using additional constraints like the visibility constraint, we naturally expected KZ1 and KZ2 to perform significantly better than Expansion especially in the case of large discontinuities. However experimentally we did not observe any significant advantage if we only consider binocular pixels. Different values of K might give better results. However we did not try varying K. An unique characteristic of KZ1 is that it performs very well for a single background proving that it is less sensitive to texture pattern or noise. The most important characteristic that we observed is that, for KZ1, the ground truth and output energy difference is smaller than the any other methods (Fig. 5.15). Now the question is how does it affect the performance? If we only consider binocular pixels then the performance does not differ largely from the other methods as mentioned above. But if we consider all pixels then we can see a big difference in performance between KZ1 and other methods (Fig. 5.17). The reason behind this is that the visibility constraint implicitly enforces correct labeling of monocular pixels. As a result, the overall error (i.e. error considering all pixels) is much smaller than other methods. This conforms with the observation made in [3, 4].

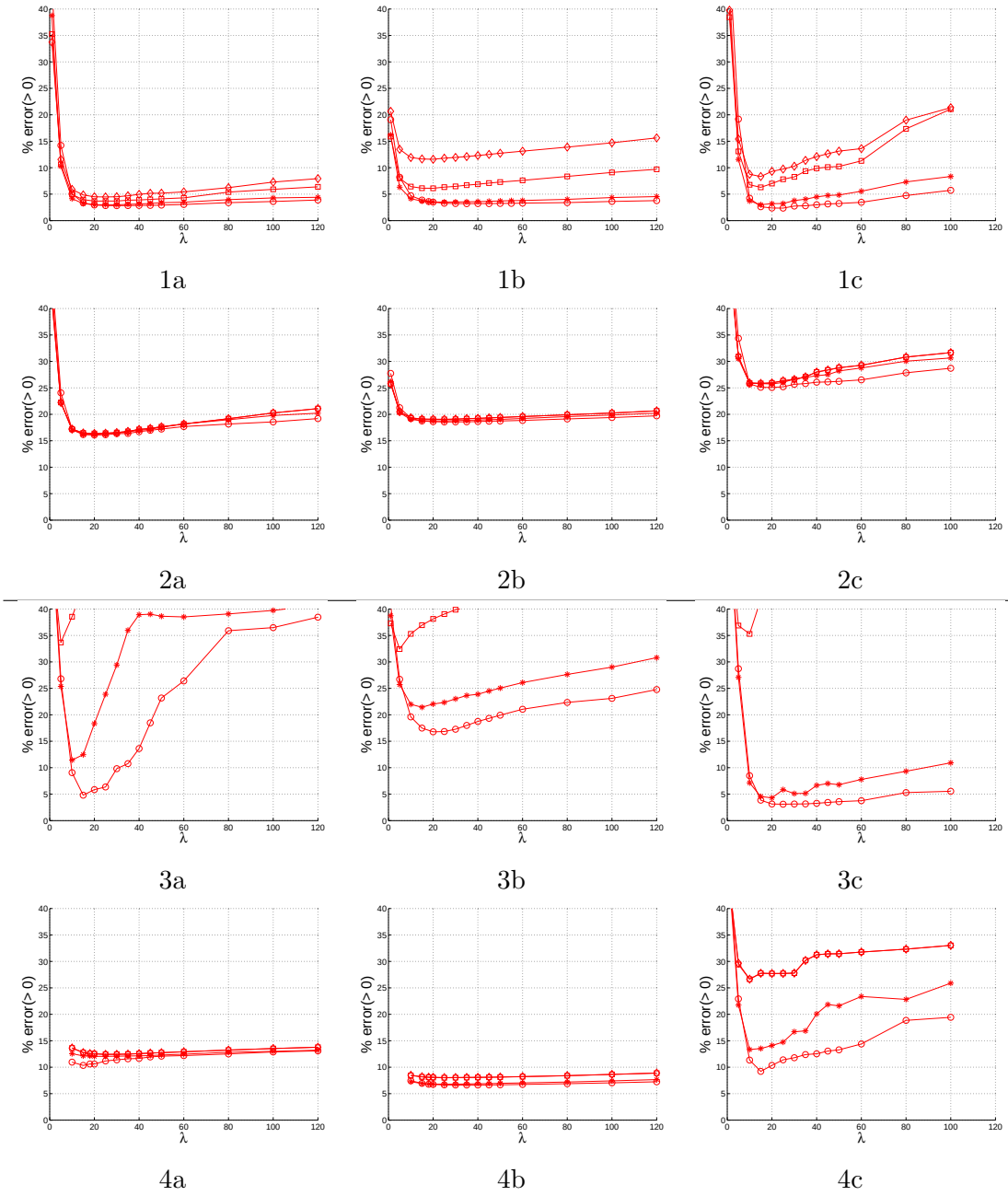


Figure 5.4: Non-zero error (i.e. error > 0) statistics for methods using the basic formulation. Only errors between 0 – 40% for Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ⋄ respectively, are shown.

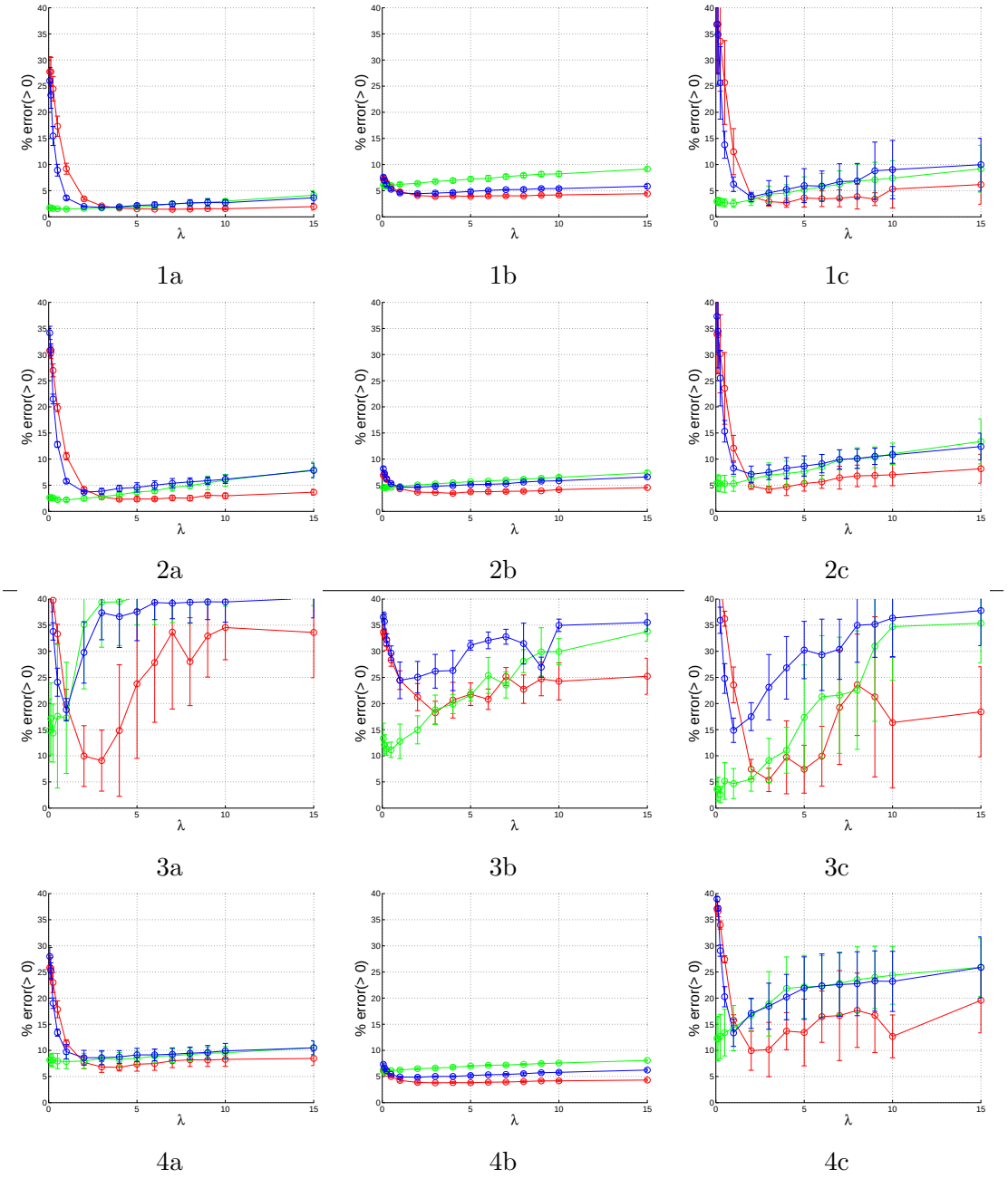


Figure 5.5: Non-zero error (i.e. error > 0) statistics for methods using visibility constraints. Only errors between 0 – 40% are shown for $k_s = 1$, $k_d = 2$ and $V_{max} = 1$. Expansion, KZ1, KZ2 are colored red, green, and blue respectively.

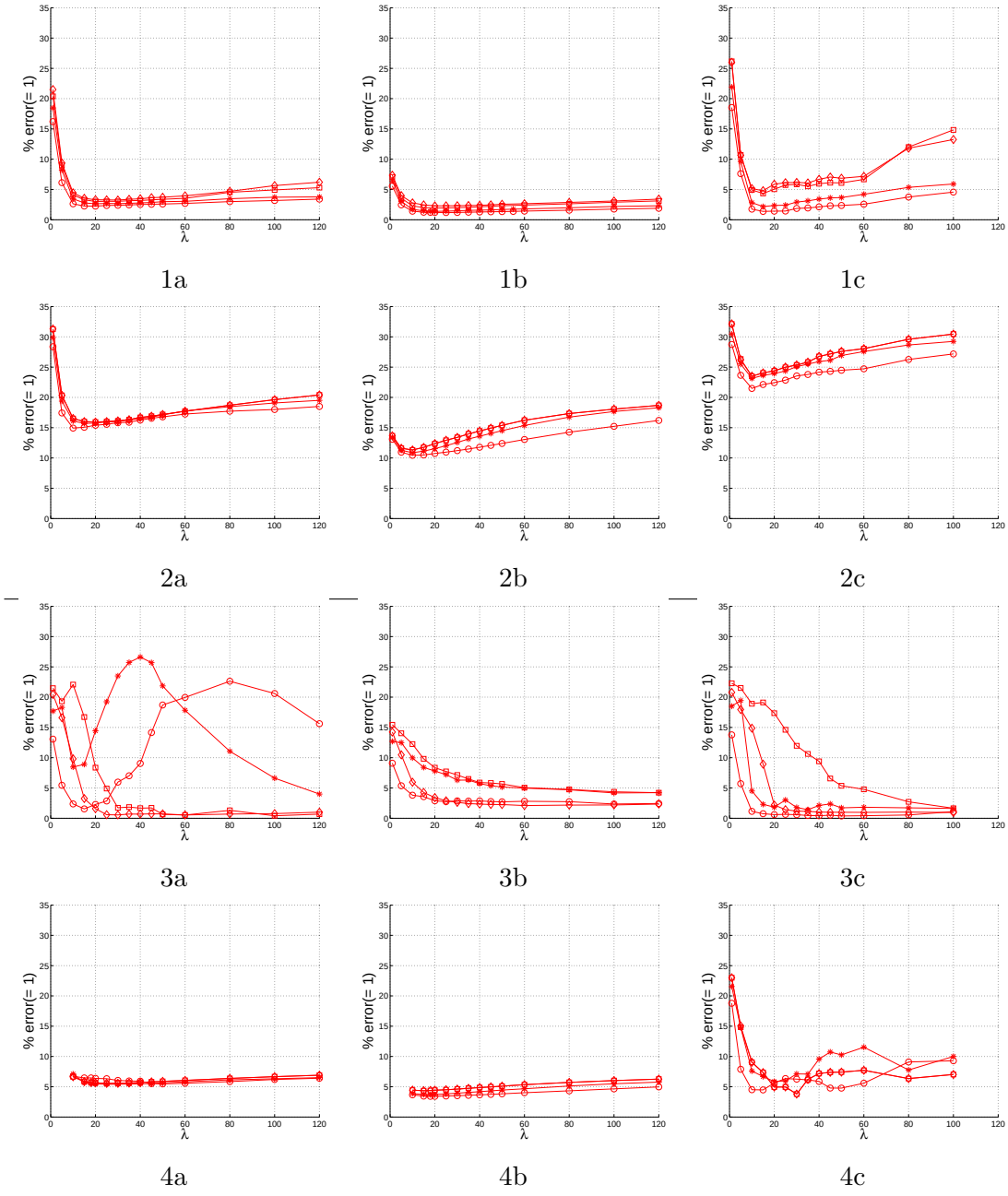


Figure 5.6: Error statistics of mislabeled pixels that differ by exactly 1 from the ground truth (off-by-one error) for basic formulation. Only errors between 0 – 35% for Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ◇ respectively, are shown.

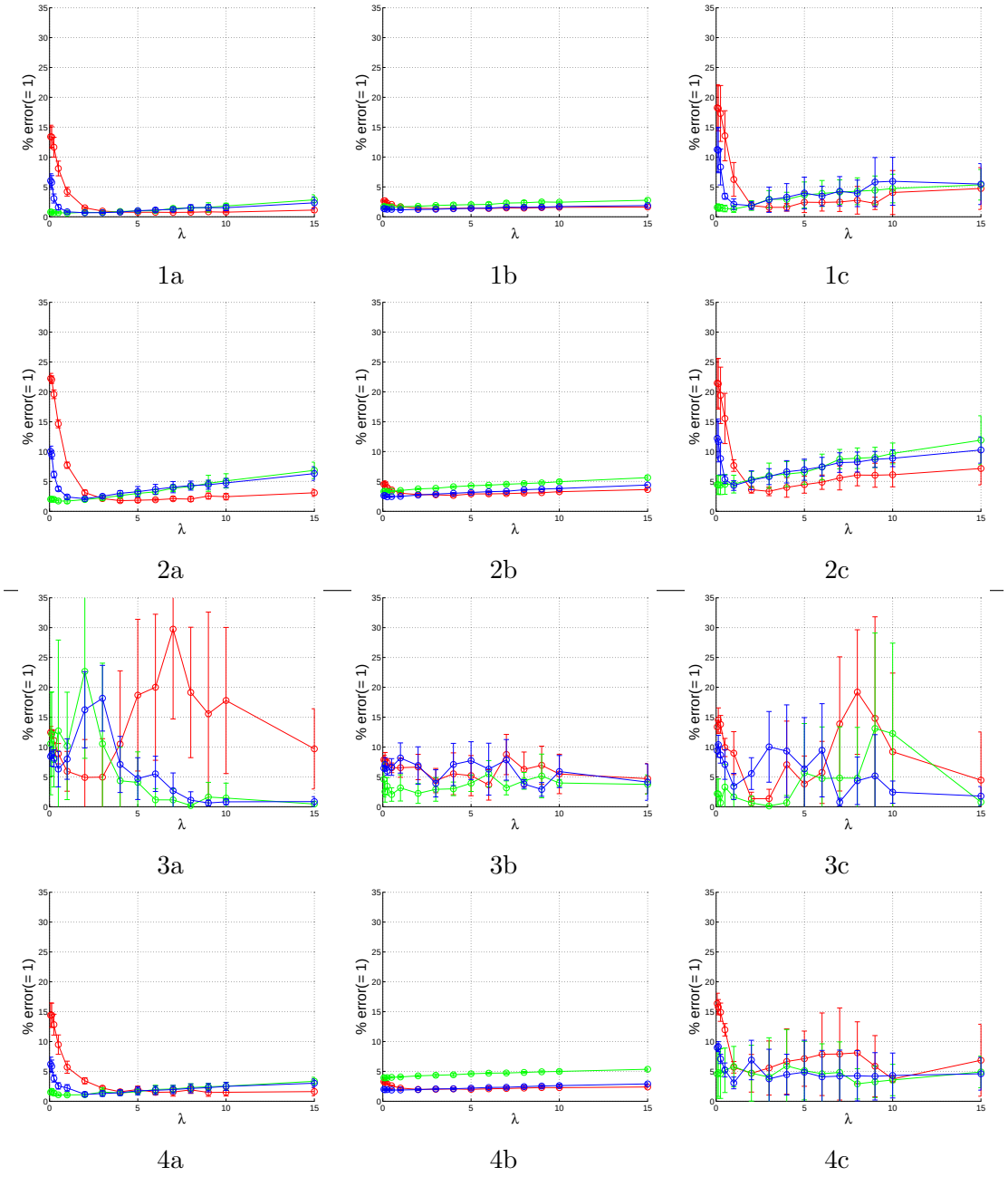


Figure 5.7: Error statistics of mislabeled pixels that differ by exactly 1 from the ground truth for visibility formulation. Only errors between 0 – 35% are shown for $k_s = 1$, $k_d = 2$ and $V_{max} = 1$. Expansion, KZ1, KZ2 are colored red, green, and blue respectively.

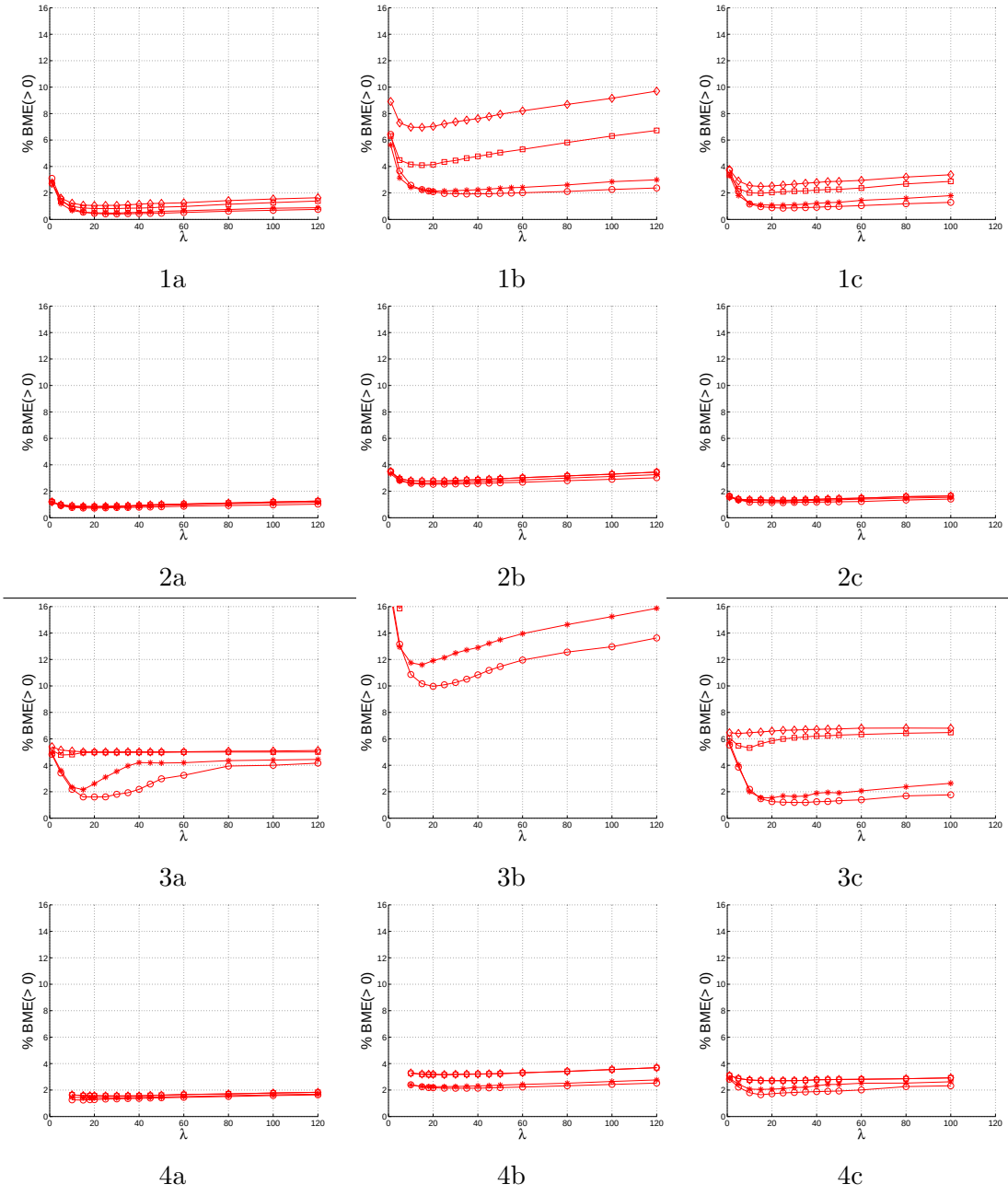


Figure 5.8: Binocular monocular error statistics for basic formulation. Only errors between 0 – 16% for Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ⋄ respectively, are shown.

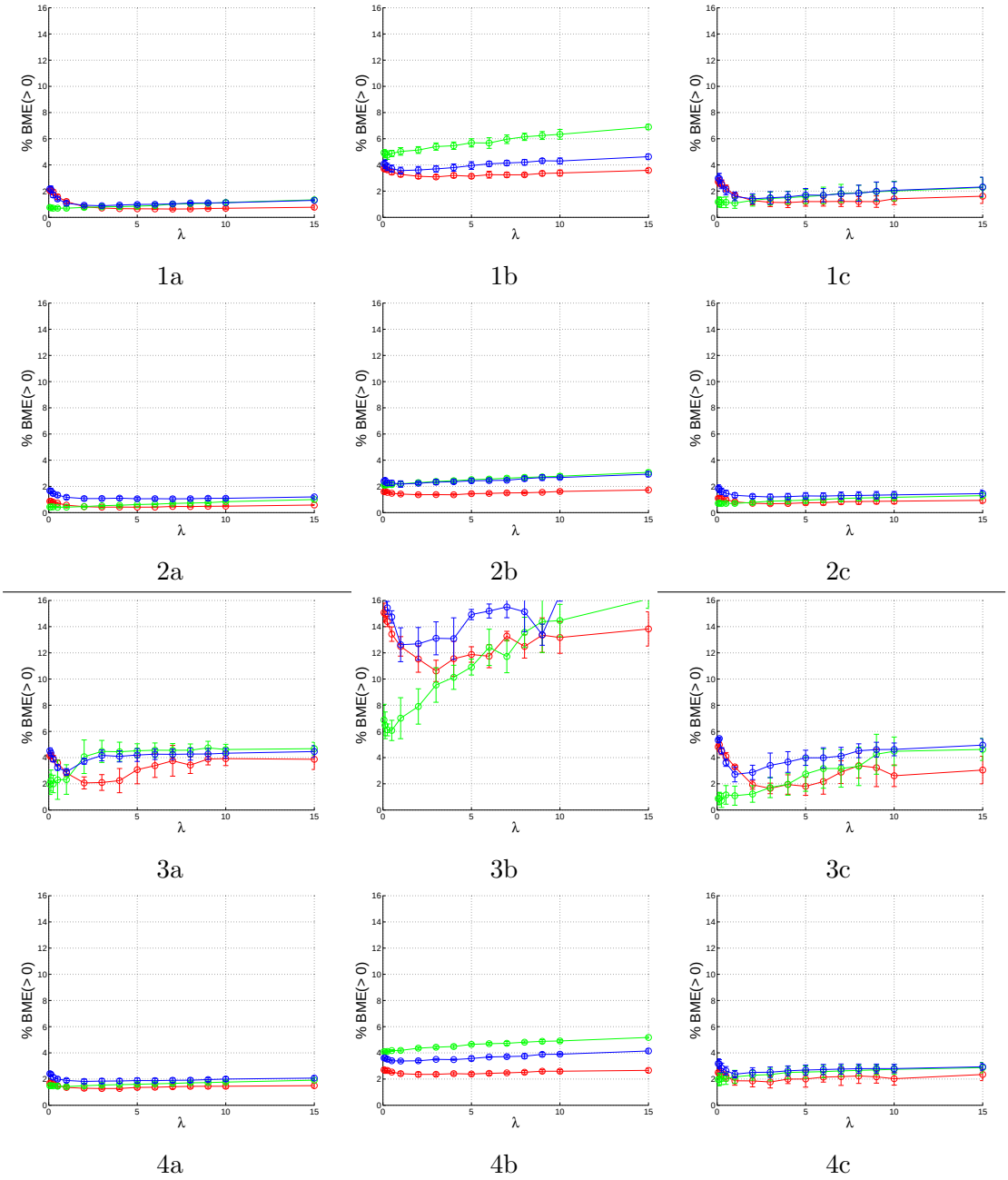


Figure 5.9: Binocular monocular boundary error statistics for additional constraint. Only errors between 0 – 16% are shown for $k_s = 1$, $k_d = 2$ and $V_{max} = 1$. Expansion, KZ1, KZ2 are colored red, green, and blue respectively.

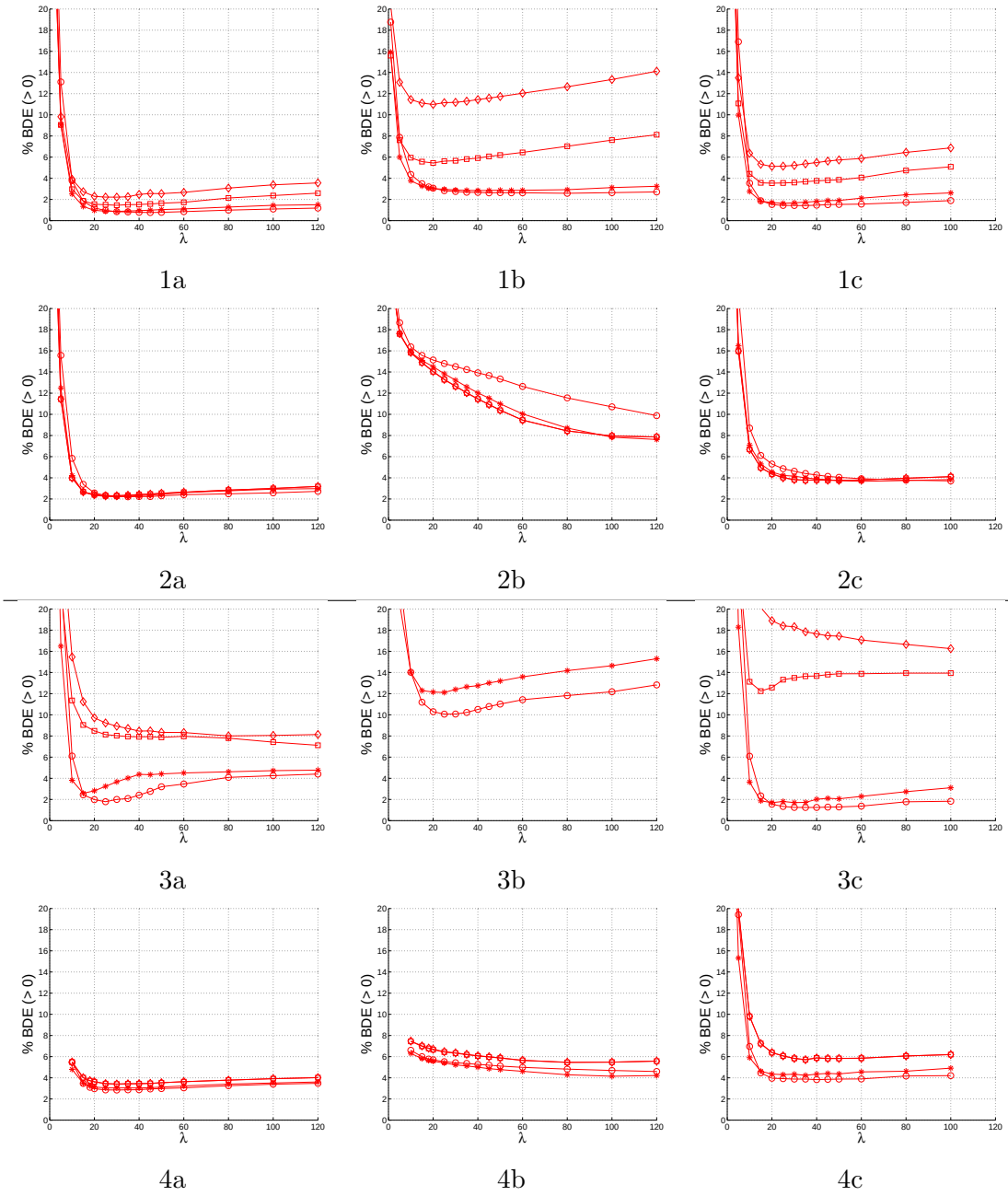


Figure 5.10: Binocular discontinuity error statistics for basic formulation. Only errors between 0 – 20% for Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ⋄ respectively, are shown.

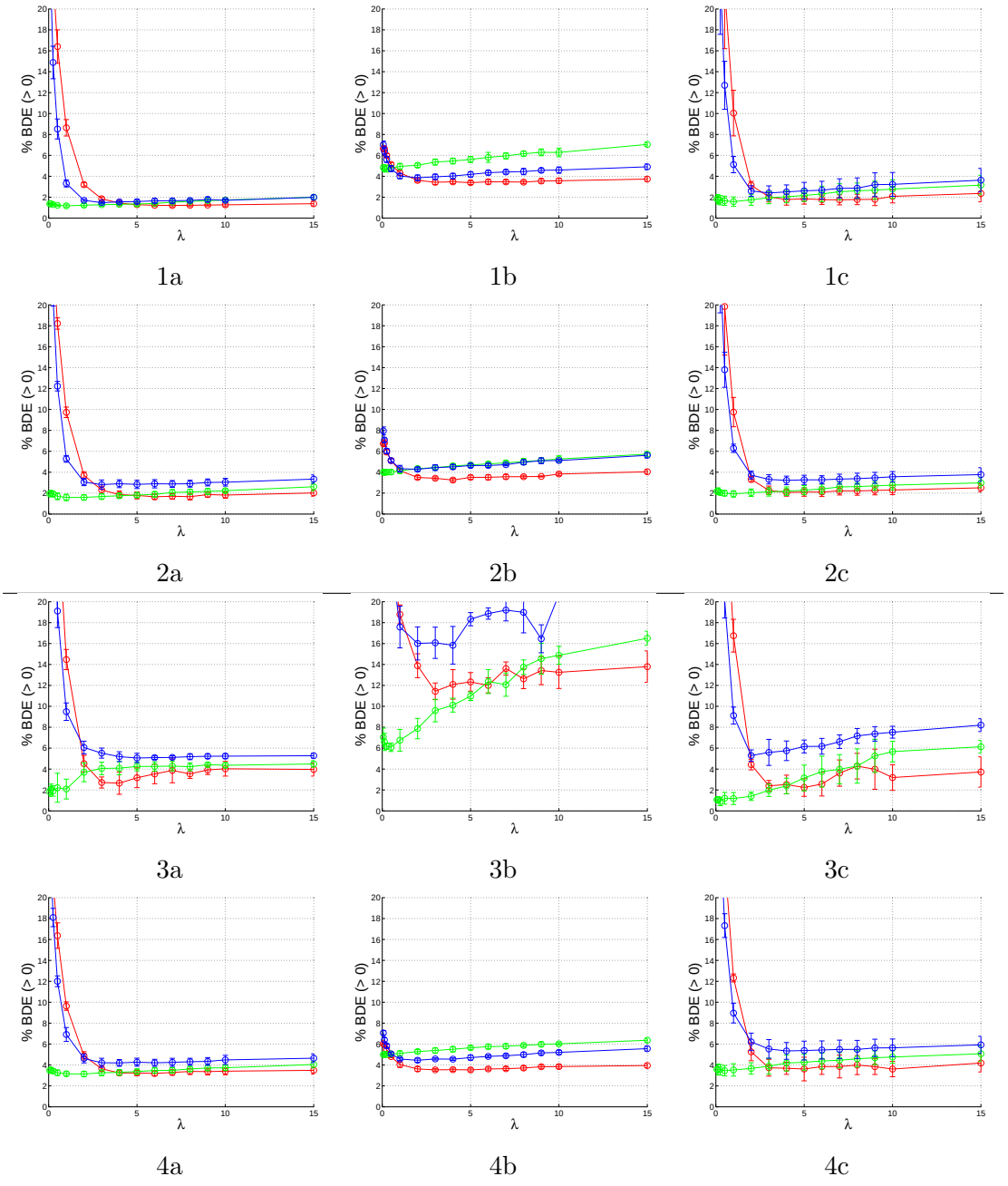


Figure 5.11: Binocular discontinuity error statistics for additional constraint. Only errors between 0 – 20% are shown for $k_s = 1$, $k_d = 2$ and $V_{max} = 1$. Expansion, KZ1, KZ2 are colored red, green, and blue respectively.

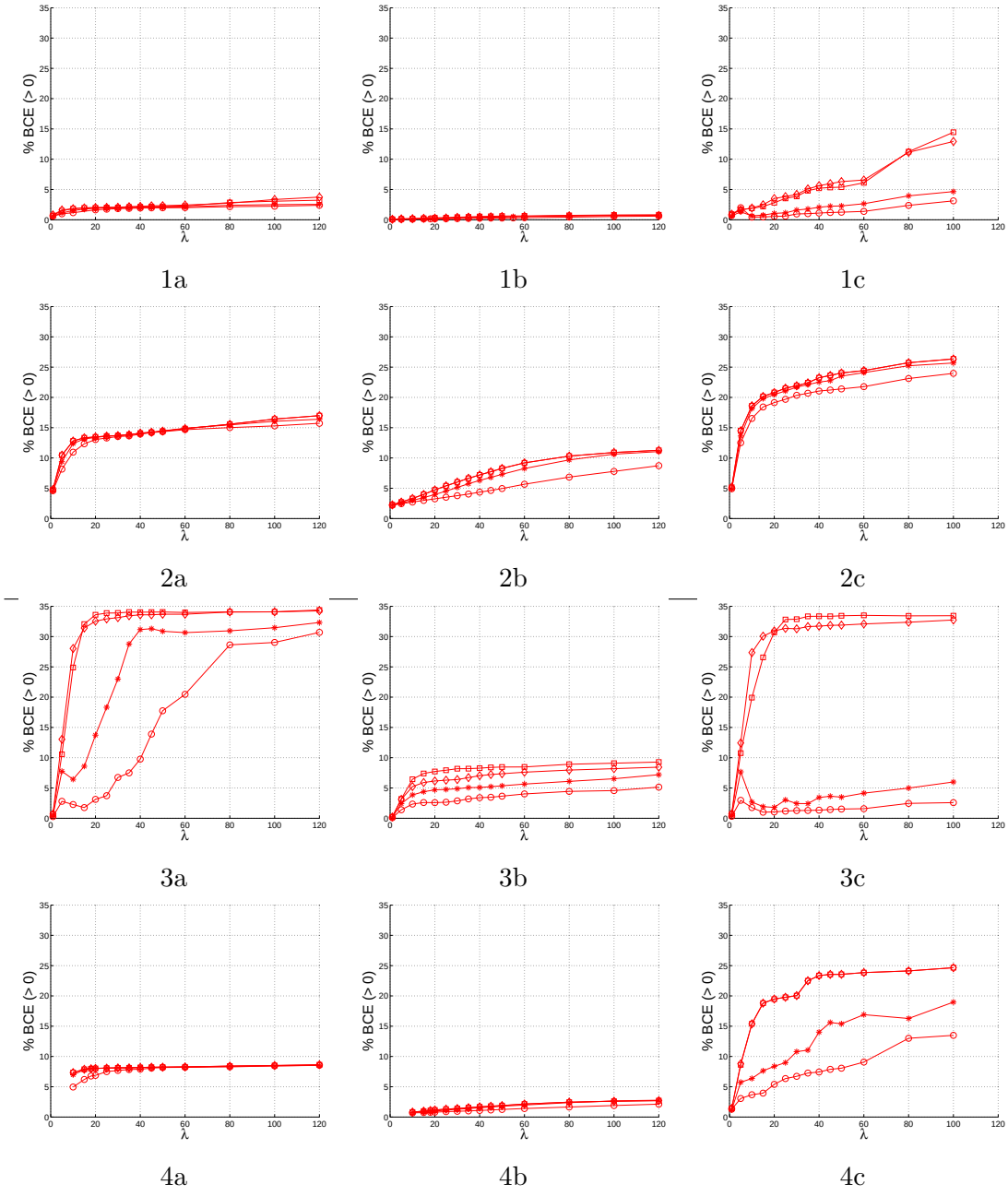


Figure 5.12: Binocular continuity error statistics for basic formulation. Only errors between 0 – 35% for Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ◇ respectively, are shown.

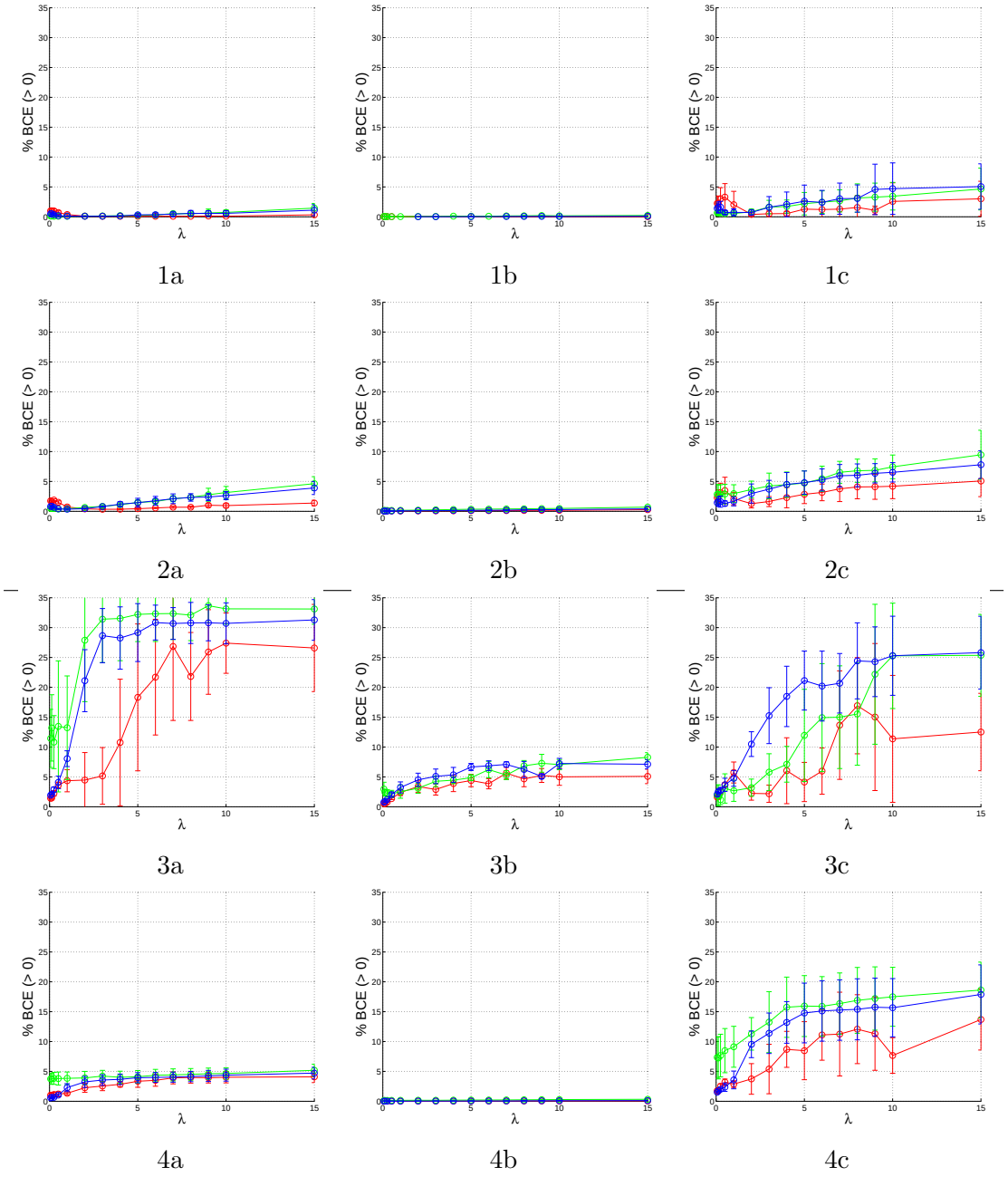


Figure 5.13: Binocular continuity error statistics for additional constraint. Only errors between 0 – 35% are shown for $k_s = 1$, $k_d = 2$ and $V_{max} = 1$. Expansion, KZ1, KZ2 are colored red, green, and blue respectively.

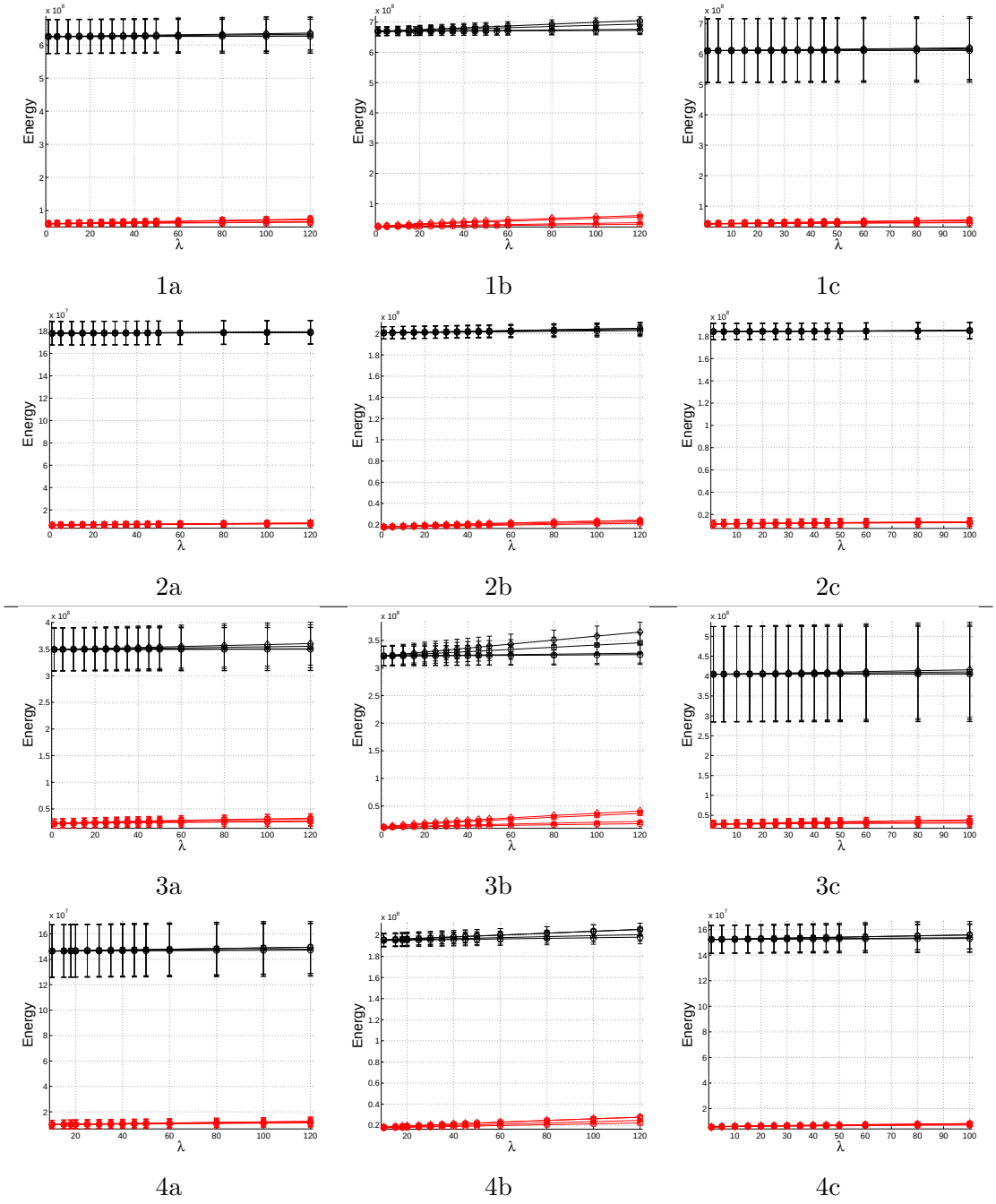


Figure 5.14: Energy statistics for basic formulation. Only Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ◇ respectively, is shown. The ground truth is represented by the black curves.

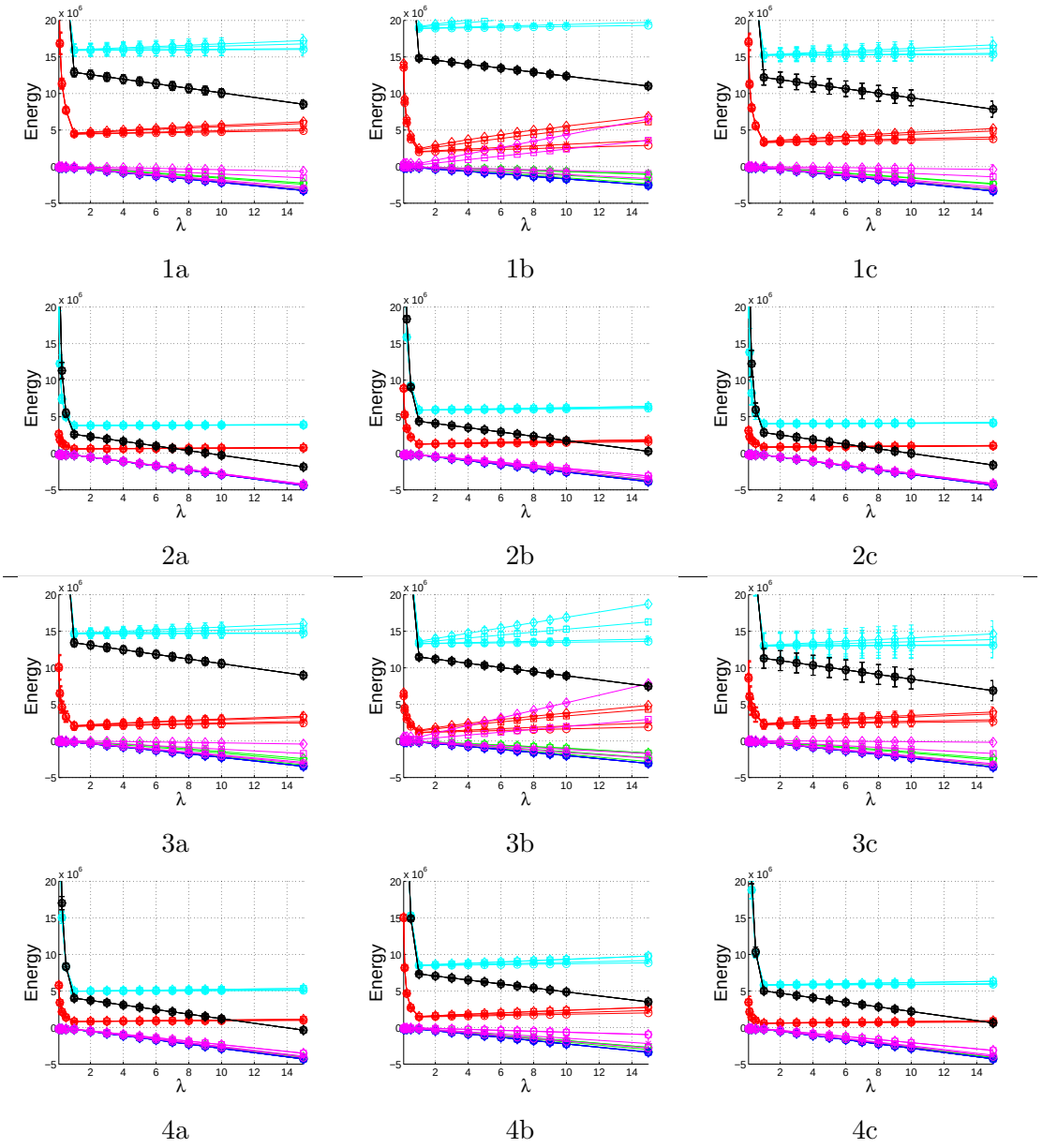


Figure 5.15: Energy statistics for additional constraint based formulations. The colors used for output energy for Expansion, KZ1, KZ2 are red, green, and blue respectively. For ground truth energy the corresponding colors are cyan, magenta, and black.

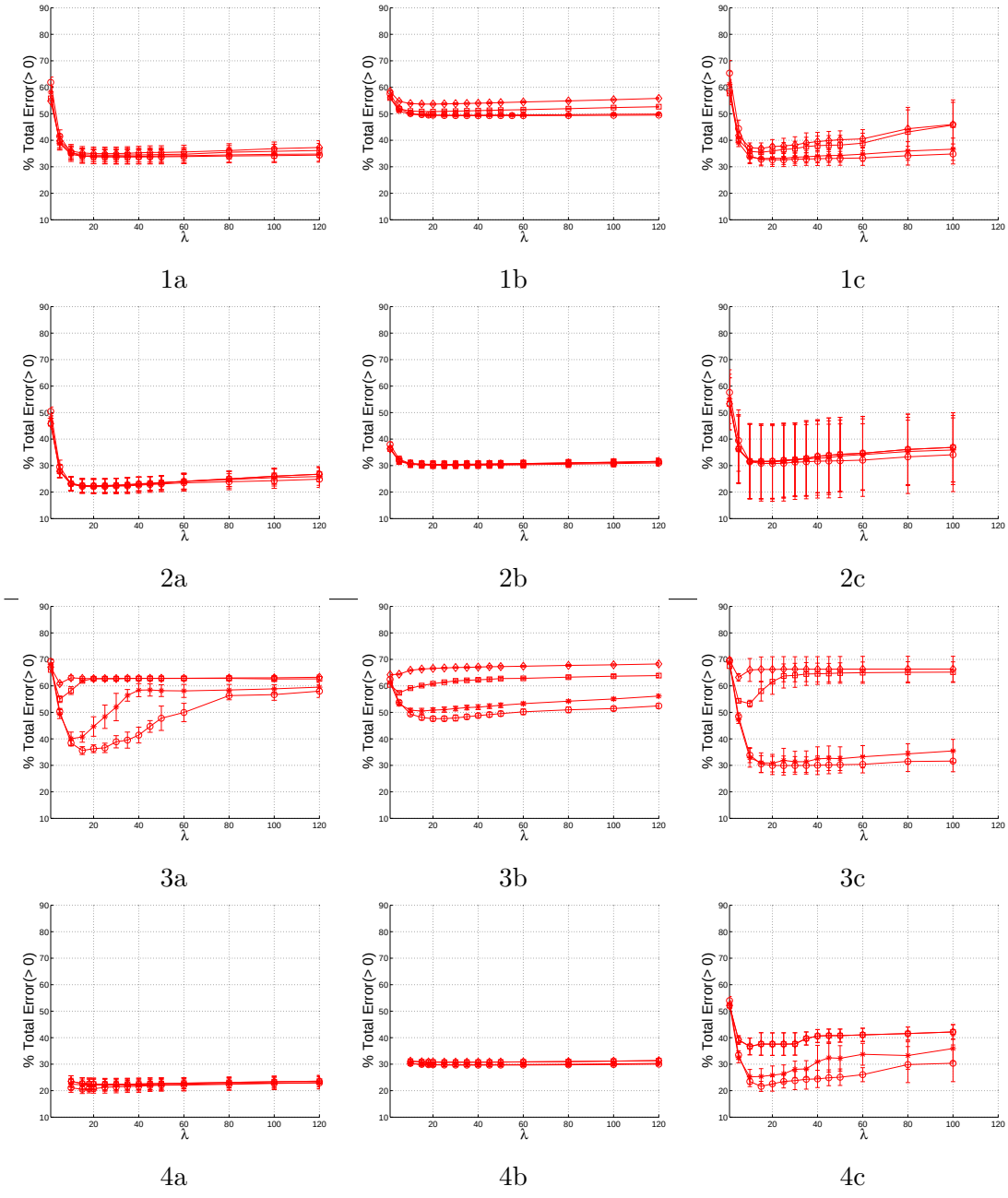


Figure 5.16: All pixel (i.e. both binocular and monocular) non-zero error statistics for methods using the basic formulation. Only errors between 10 – 90% for Expansion with $k_s = 1$, $k_d = 2$ and V_{max} values 1,2,10 and 100 represented by O, *, □ and ⋄ respectively, are shown.

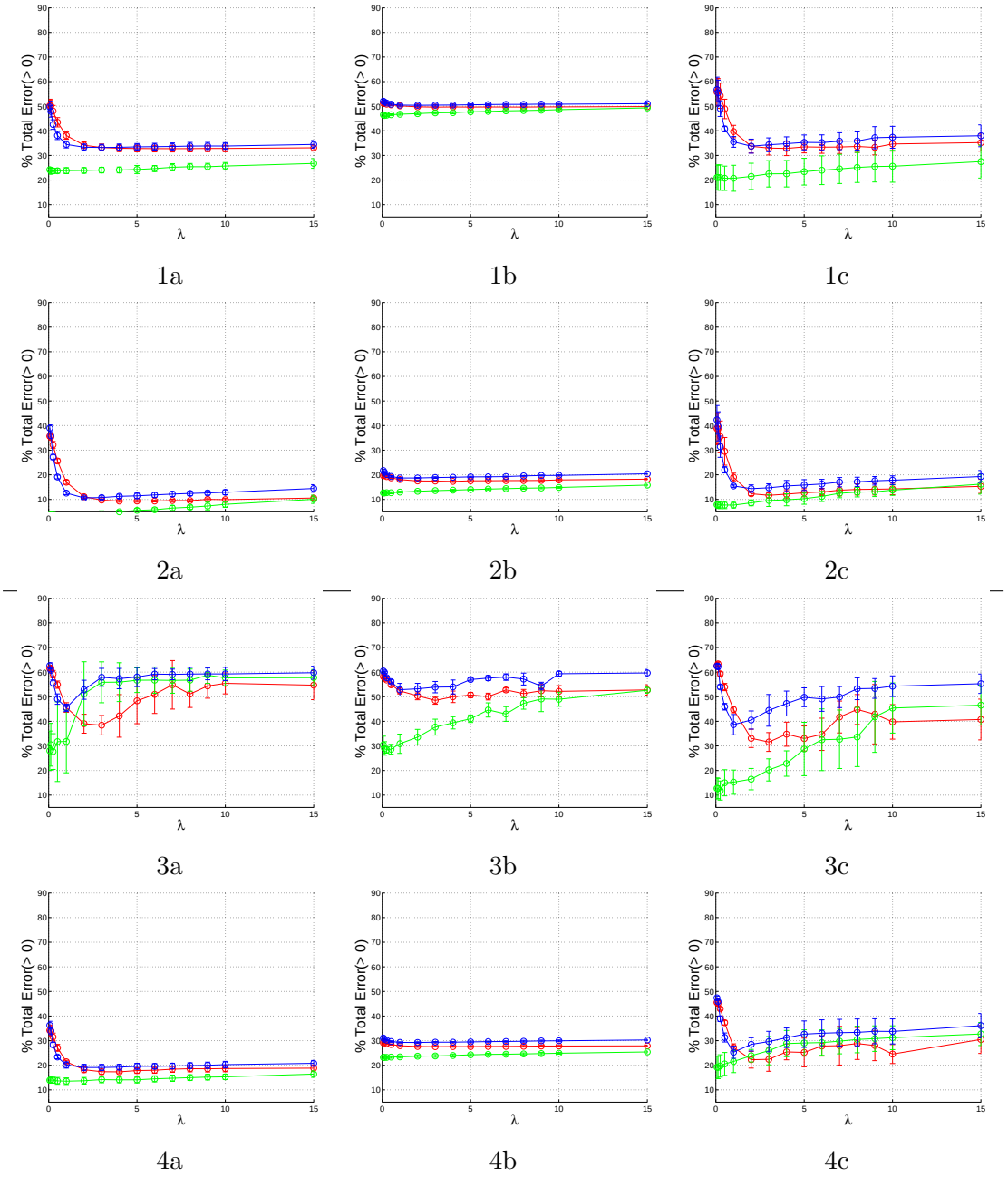


Figure 5.17: All pixel (i.e. both binocular and monocular) non-zero error statistics for methods using additional constraints. Only errors between 5 – 90% are shown for $k_s = 1$, $k_d = 2$ and $V_{max} = 1$. Expansion, KZ1, KZ2 are colored red, green, and blue respectively.

Chapter 6

Conclusion

Stereo reconstruction has been of interest to the computer vision community for a long time. But its application to cluttered scenes has never been directly addressed before. As a result, the performance of existing algorithms, especially the ones based on MRF, is unknown for these scenes. In the introduction of this thesis, we discussed how important this category of natural scene is. Not only that, there are also fields that can benefit from stereo reconstruction methods for such scenes. This motivated us to investigate how a class of techniques performs for cluttered scenes. In our investigation we limited ourselves to the MAP-MRF based stereo formulation. Our goal was to find how different forms of energy and optimization techniques would perform for different types of cluttered scenes.

In this chapter, the overall approach is summarized in Sec. 6.1. Sec. 6.2 summarizes the main observations, and the conclusions that we make from them. In Sec. 6.3 the contribution of this thesis is discussed. We address some of the open questions in Sec. 6.4 and finally discuss some possible future directions of this work in Sec. 6.5.

6.1 Summary of Our Approach

To this end, we reviewed the underlying concepts behind the MAP-MRF stereo formulation, and the optimization algorithms used for solving the MAP problem. We looked at how natural scenes have been modeled by different authors, and the justification behind those models. We also discussed the cluttered scene model that is considered in the thesis, the justification behind using it, its derivations and charac-

teristics. For evaluating algorithms, we require a benchmark dataset. Currently there are no stereo datasets for cluttered scenes. Therefore, we generated synthetic scenes for our experiments. We discussed how the synthetic scenes were generated, the difficulties in the generation process and how they were resolved. After that we focused on the types of scenes that are needed for the experiments. We generated different types of cluttered scenes and experimentally verified their underlying statistics. Finally we applied different algorithms with different parameter settings on the dataset, observed the performance for each case and generalized the results.

6.2 Summary of Observations and Conclusions

We found Expansion, TRW-S, BP-M, and Swap to perform equally well. Despite the variation in the scenes, the Potts model (i.e. $V_{max} = 1$) or small V_{max} in general performs best in most cases. We found the form of the data term to be less important than the form of smoothness term, because the relative performance between different forms of smoothness and algorithm combination changed very little with the data term. If only binocular pixel errors are considered then the methods with visibility constraints do not necessarily perform significantly better than those with basic constraints. However, if all pixels (i.e. binocular and monocular) are considered then methods with visibility constraint (e.g. KZ1) can perform significantly better. This conforms to previous observations made in [2, 3, 4]. This performance differences are also naturally reflected in the energy values. For basic formulation the gap between ground truth and output energy is very large, whereas for visibility formulation the gap is usually smaller and for KZ1 it is the smallest.

The key question now is, is there any room for further improvement, especially in the case of cluttered scenes? In terms of total error considering all pixels, basic formulations are insufficient for cluttered scene stereo. Even though KZ2 uses additional constraints like the uniqueness constraint, its performance is not on par with KZ1. If total error is considered then KZ1 is clearly the winner. If only binocular pixels are considered then we did not see any significant performance improvement for KZ1. In fact in some cases Expansion and KZ2 perform better. However the performance gap is not big enough ($\lesssim 5$) to warrant for an improvement. Also if we look at the error rates, we see that in most cases the minimum mean error is less than 5%. But for certain scenes the error

for even the best performing method can be close to 10%. We argued that this happens because of the non-contiguous nature of $-\log(p(f_p - f_q))$. It should be noted that this type of cases can also occur in natural scenes where large part of the background is visible through sparse bush or foliage. Therefore, it is important to improve the forms of prior for these type of cases.

Furthermore, we have seen that monocular pixels make the most contribution to the total error for all pixels. So, if our goal is to reduce the total error for all pixels then there is a vast room for improvement. We saw that KZ1 performs better than other methods but for very large number of monocular pixels (e.g. scene 1b in Fig. 5.17) the error rate for KZ1 is close to 50%. So there is still room for improving KZ1, especially for cluttered scenes.

6.3 Contributions

This thesis contributes to the current field of stereo vision in a number of ways. In the following we list some of the major and minor contributions.

6.3.1 Comparative Study of Cluttered Scenes

The main contribution of the thesis is in evaluating some of the fundamental MRF based techniques for cluttered scenes. In the past there have been performance analysis studies for generic scenes. However no such study has been carried out for cluttered scenes. Cluttered scenes occur very frequently in the natural environment and there are very promising applications of cluttered scene stereo reconstruction. This thesis contributes to the field of vision, by investigating how some of the widely used optimization techniques in conjunction with different types of energy formulation, work for cluttered scenes.

6.3.2 Classifying Cluttered Scenes

Cluttered scene categorization is an unexplored area. We took a preliminary step towards classifying cluttered scenes based on the underlying statistics. Based on the model, we chose parameters that would affect different statistical properties of the scenes and experimentally verified those effects. In Chapter 5 we have seen that similar underlying statistics causes similar error statistics. Therefore, if we know the set of

good parameters for a particular class, we can try to use that for other scenes of the same class.

6.3.3 Synthetic Cluttered Stereo Pair Generation

The lack of cluttered scene benchmark dataset and the need for such dataset for performance evaluation, drove us to generate synthetic cluttered scene stereo images. While a single depth map is sufficient for statistical analysis, for stereo evaluation, stereo pairs are required. These stereo pairs and the corresponding ground truth disparity maps have to be rendered pixel accurately. There are certain challenges involved in this process. Chapter. 4 identifies some of the subtle issues related to round-off error such as consistency in size of projection, consistency between theoretical disparity value and the actual value, and proposes an approach for overcoming those issues. The synthetic stereo pairs were generated using the proposed approach and their correctness was verified.

6.3.4 Best Performing Parameters

From a practitioner's point of view, it is important to know the set of parameters that can produce good results for cluttered scene stereo problems. Finding such a set of parameters is a time consuming process. This thesis gives an overview of how some of the parameters affect the error in scenes with certain underlying statistics. This would simplify the process of choosing the set of parameters for a scene whose statistical model is close to the ones we have in our experiment. Furthermore, we conducted our experiments with a wide range of parameters and experimentally showed which set of parameters work best in general and which do not. Therefore, for a completely new scene those set of parameters should be the first choice to try.

6.4 Issues and Open Questions

Some of the scenes exhibit greater sensitivity to off-by-one error. We were only able to identify where this happens and indicated that they might happen due to the texture pattern. However we did not conclusively show this to be the case. Since they have more impact on objects that are farther away, for certain scenes (e.g. 2a–c,4a–c) minimizing this type of error is advantageous.

We have seen that small V_{max} works best because of the $-\log p(f_p - f_q)$ curve. We showed how to choose parameters that affect certain properties of the scene like probability density of disparity, discontinuity, etc. But the question is, is it possible to generate scenes with certain probability of disparity difference? The reason why it is somewhat difficult is that it is an average over disparity differences for different disparities. As a result, generating scenes with a certain disparity difference probability is non-trivial.

6.5 Future Work

A natural progression would be to consider non front-parallel scenes or scenes with curved surfaces. These scenes break some of the assumptions made by the methods that are currently used in the thesis and can be better modeled using higher-order priors [34].

In this thesis, we mostly emphasized on binocular pixels because ideally these pixels should have the correct disparity labels. We have seen that in terms of total error with all pixels, monocular pixel errors make the most contribution. Future work will focus on how the monocular errors vary with different scenes and stereo methods.

The scope of the thesis was limited to methods using the basic formulation and a couple of methods that uses additional constraints. There are several other methods that model visibility and occlusion. They can be applied to cluttered scene reconstruction to see how well they perform.

The most important future step would be to apply the results that we obtained in this thesis to real scenes. Real scenes entail certain other challenges such as shadows, intensity variation, textureless regions, etc. These challenges need to be addressed for successful application of stereo algorithms for natural cluttered scenes.

Appendix A

List of Results

In this appendix, we provide the additional plots that can be referenced to get a better idea about the results. In Sec. A.1 the scene and error statistics for baseline, $T_x = 0.1$ is given. Sec. A.2 gives plots showing contribution of different types of errors. Plots for greater than 1 errors given in Sec. A.3.

A.1 Results for $T_x = 0.1$, $\gamma = 0.1$ and $r = 0.1$

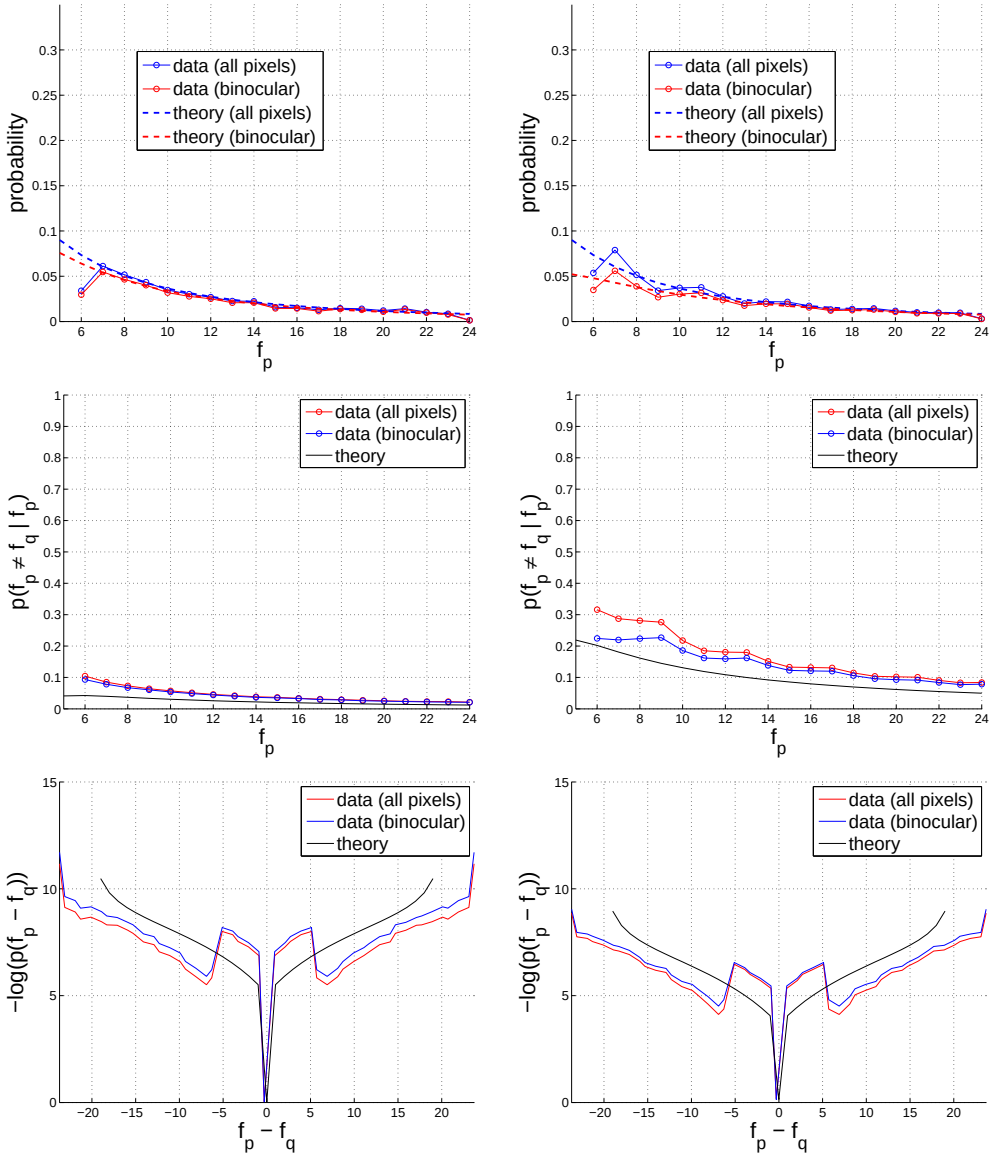
In this section, we first show the statistical properties of scenes with small baseline. We only consider the scenes with $\gamma = 0.1$ and $r = 0.1$. This is primarily because this type of scenes exhibit more error. We are interested in finding out if the same scene with a smaller baseline exhibits the same characteristics.

A.1.1 Scene Statistics

Fig. A.1 shows the scene statistics. This set of scenes has basically the same parameters as the third set (i.e. scenes 3a–c) except the baseline is set to 0.1. This increases binocular visibility. As a result, the difference between binocular and all pixel curve is small. The range of disparity value is smaller because of the small baseline. This makes the probability of all the statistics greater than that of set 3.

A.1.2 Error Statistics

The error statistics shown in Figs. A.2 and A.3 are similar to that of Scenes 3a–c.



(5a) (5b)

Figure A.1: Disparity and Neighbor Statistics for Data and Model with $\gamma \approx 0.1$, z-range between 2 and 8 and baseline 0.1

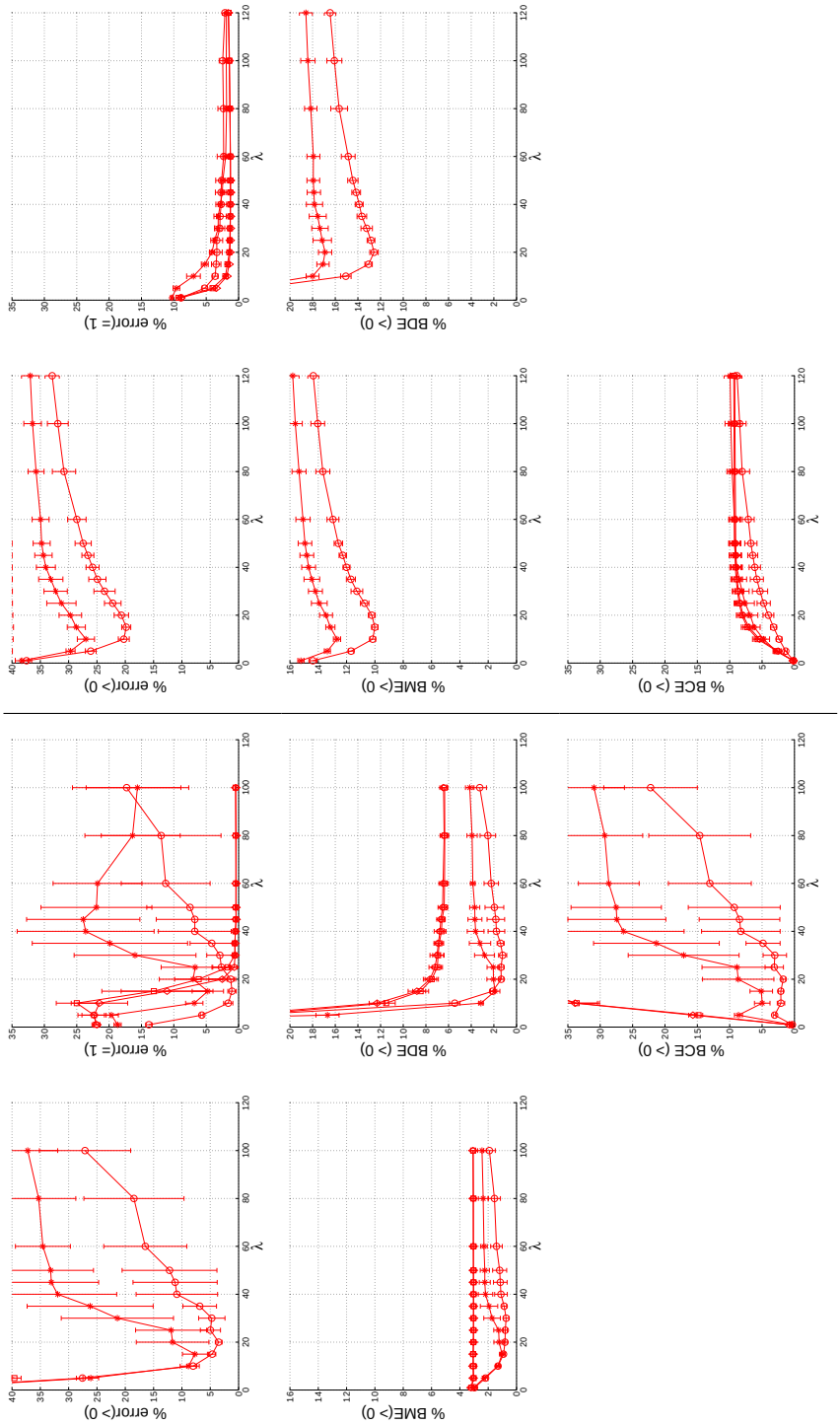


Figure A.2: Error statistics for basic formulation for scenes 5a (left half), and 5b (right half)

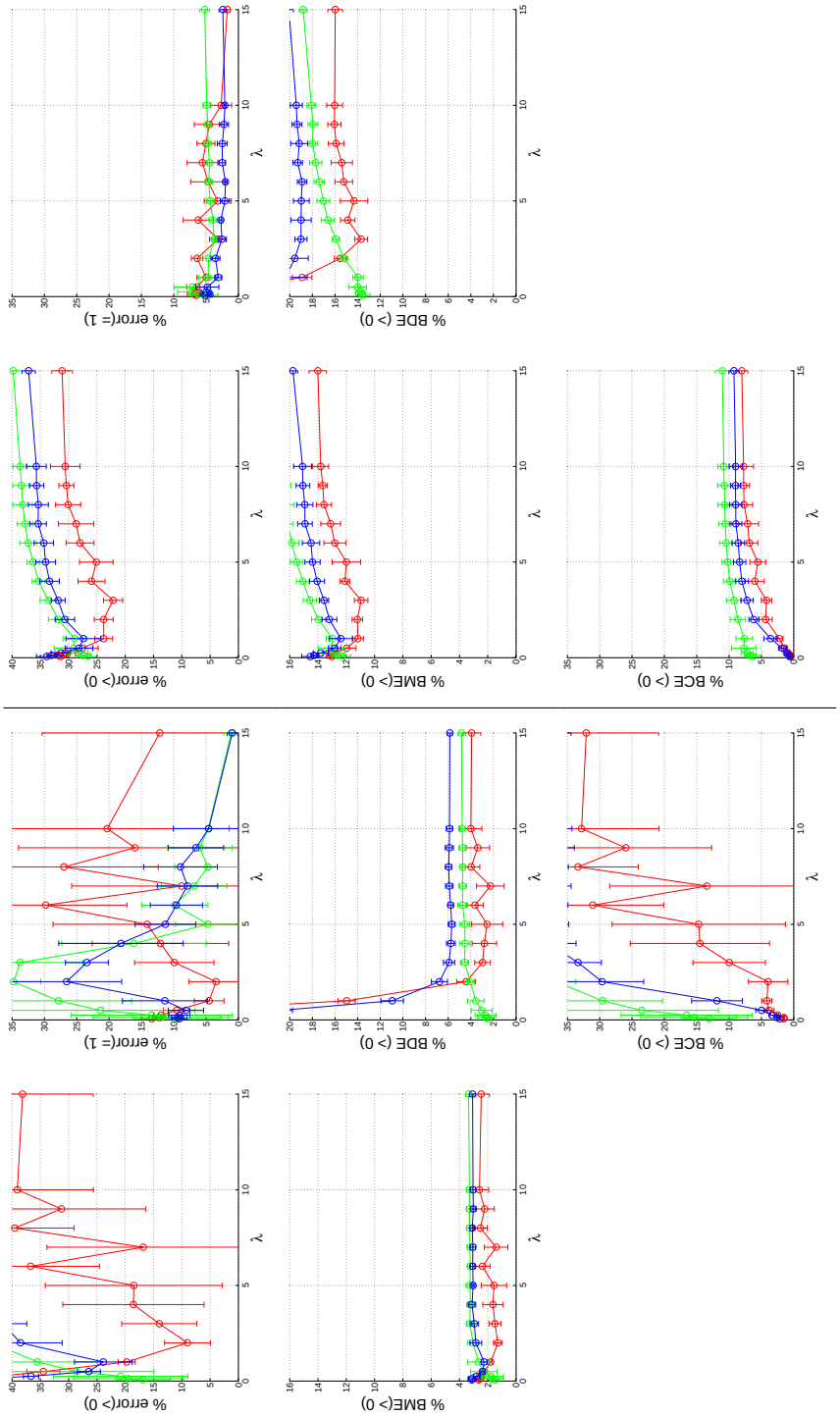


Figure A.3: Error statistics for formulations with visibility constraints for scenes 5a (left half), and 5b (right half)

A.2 Error Contribution Plots

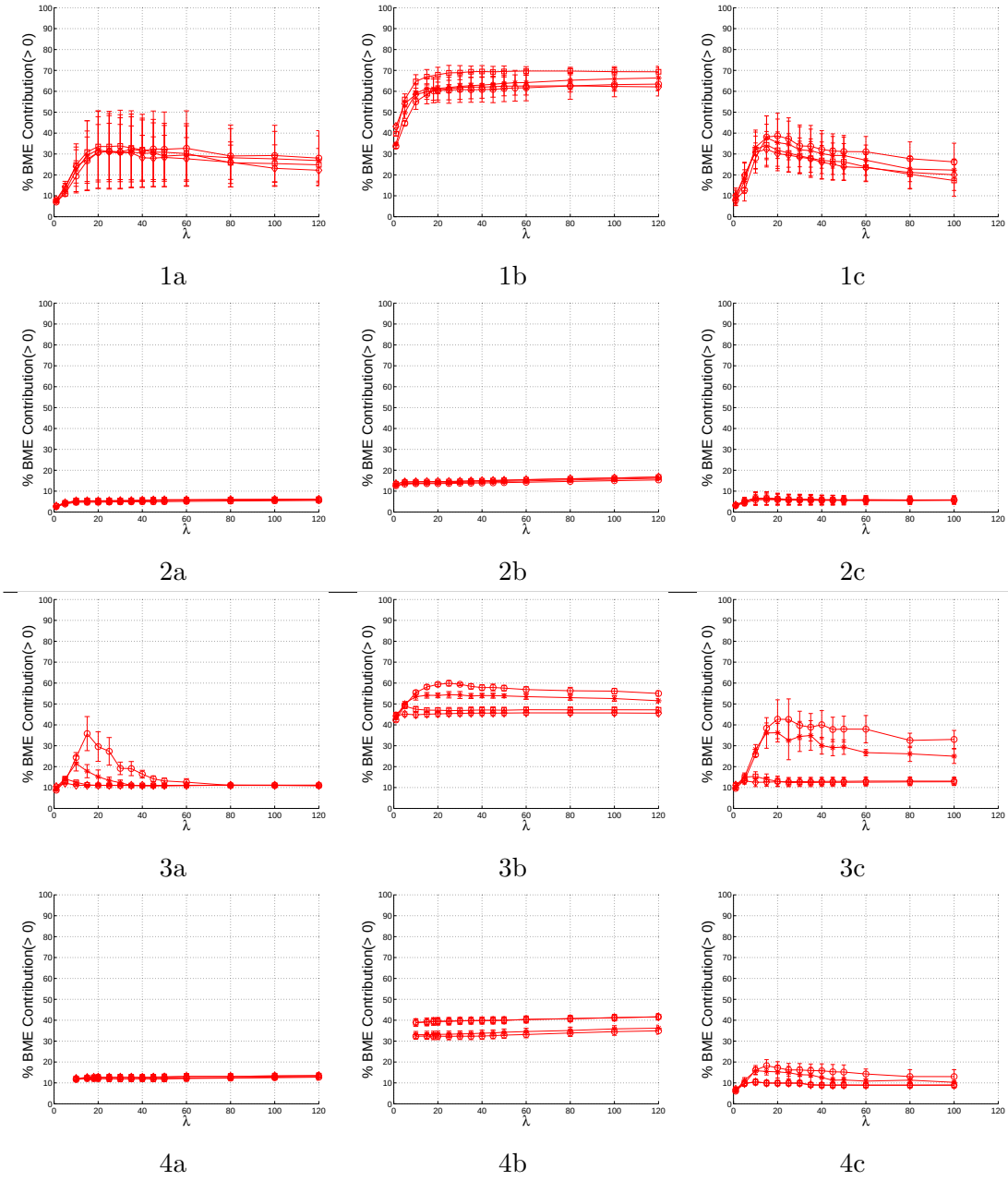


Figure A.4: Contribution of binocular monocular error for basic formulation.

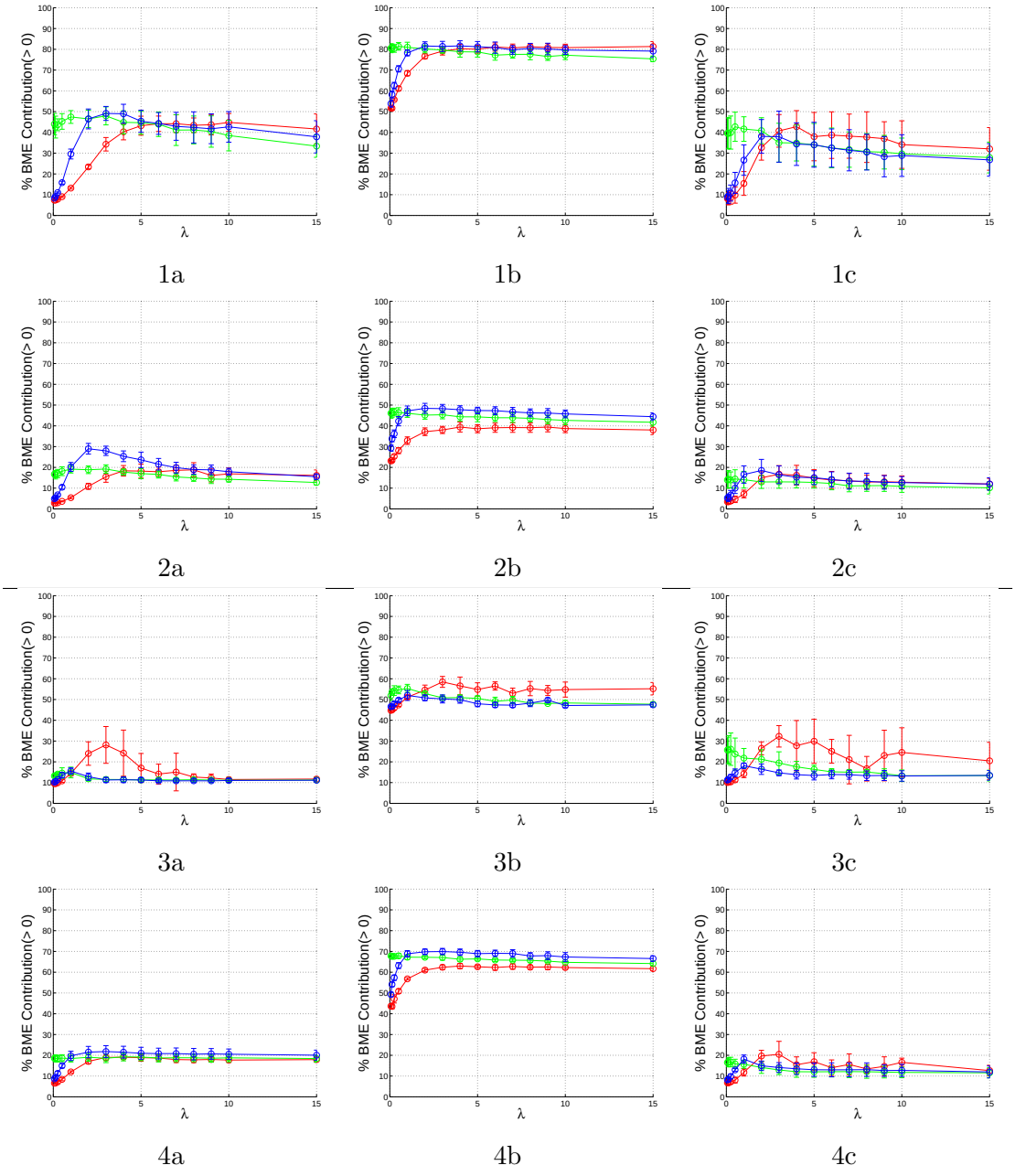


Figure A.5: Contribution of binocular monocular boundary error for formulations with additional constraint

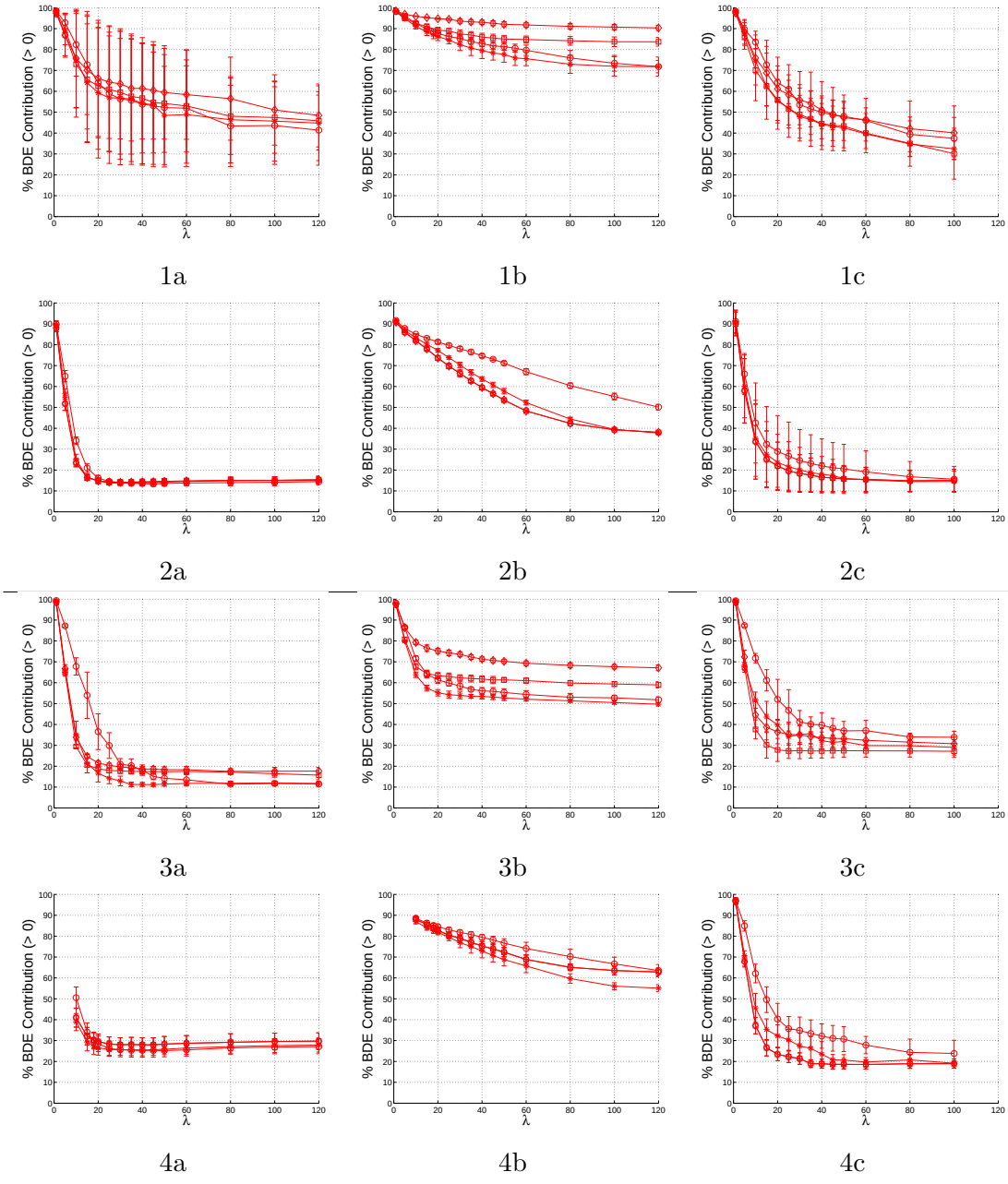


Figure A.6: Contribution of binocular discontinuity error for formulations with basic constraints

A. LIST OF RESULTS

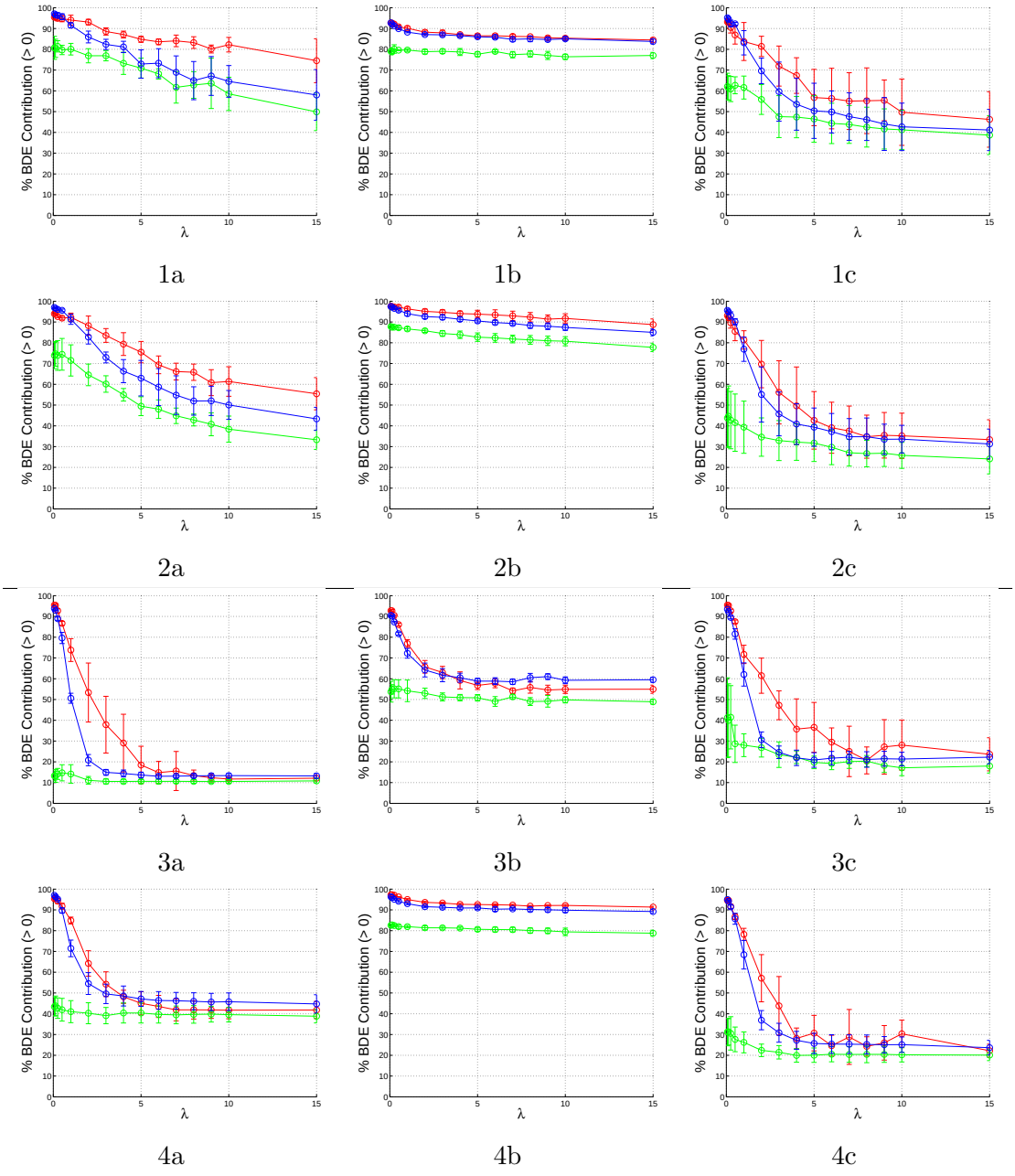


Figure A.7: Contribution of binocular discontinuity error for formulations with additional constraint

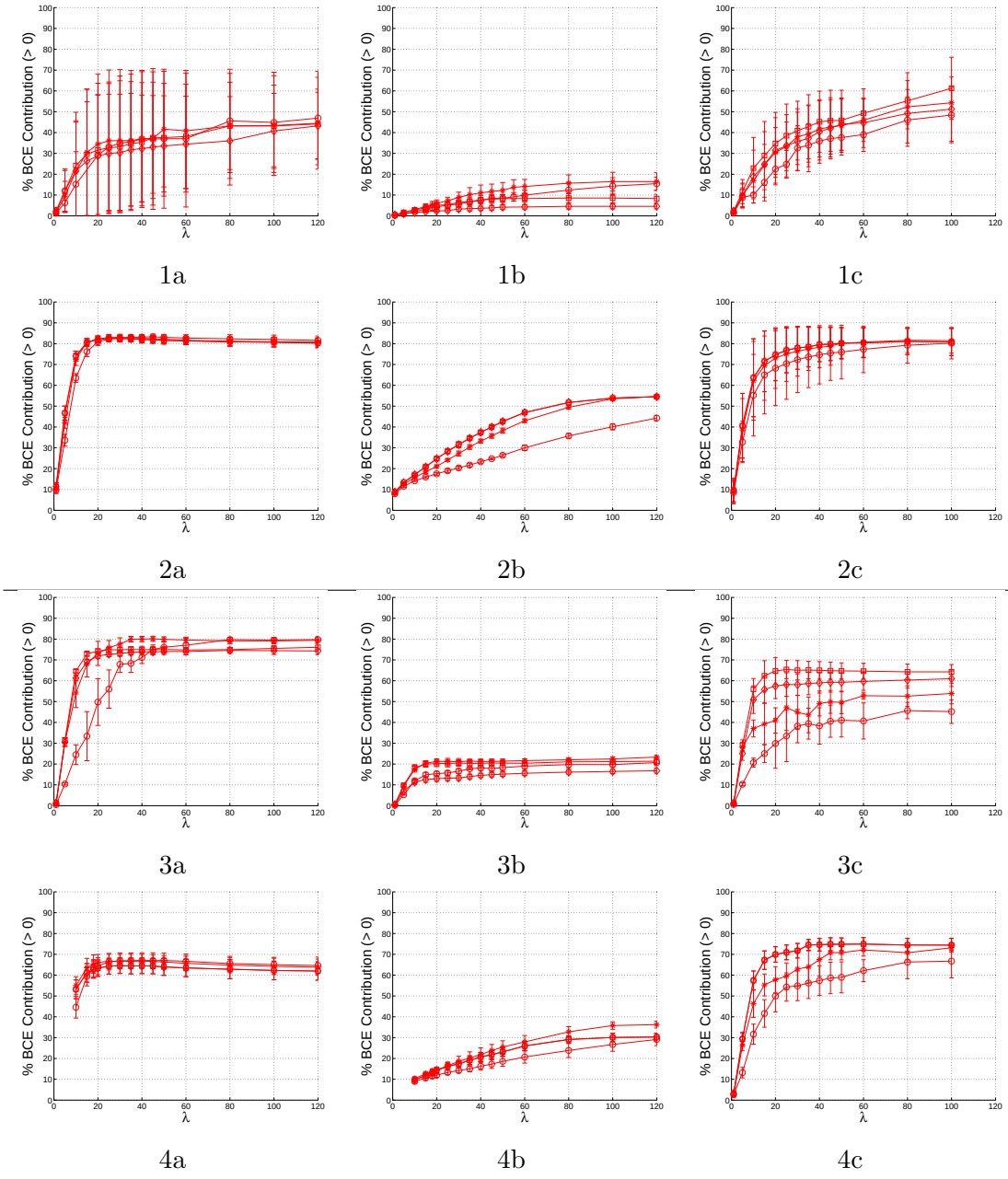


Figure A.8: Contribution of binocular continuity error for formulations with basic constraints

A. LIST OF RESULTS

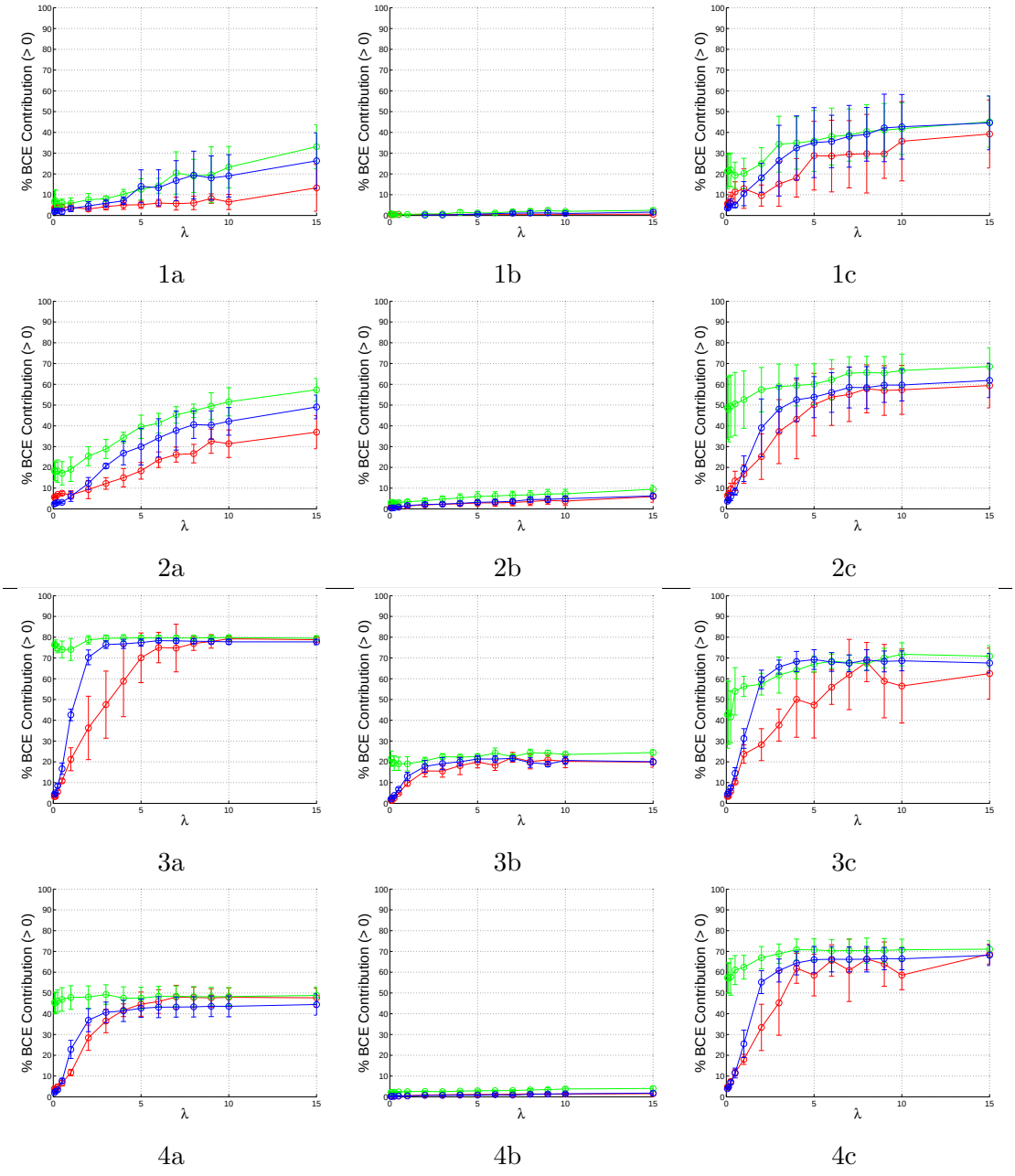


Figure A.9: Contribution of binocular continuity error for formulations with additional constraint

A.3 Total Errors (> 1)

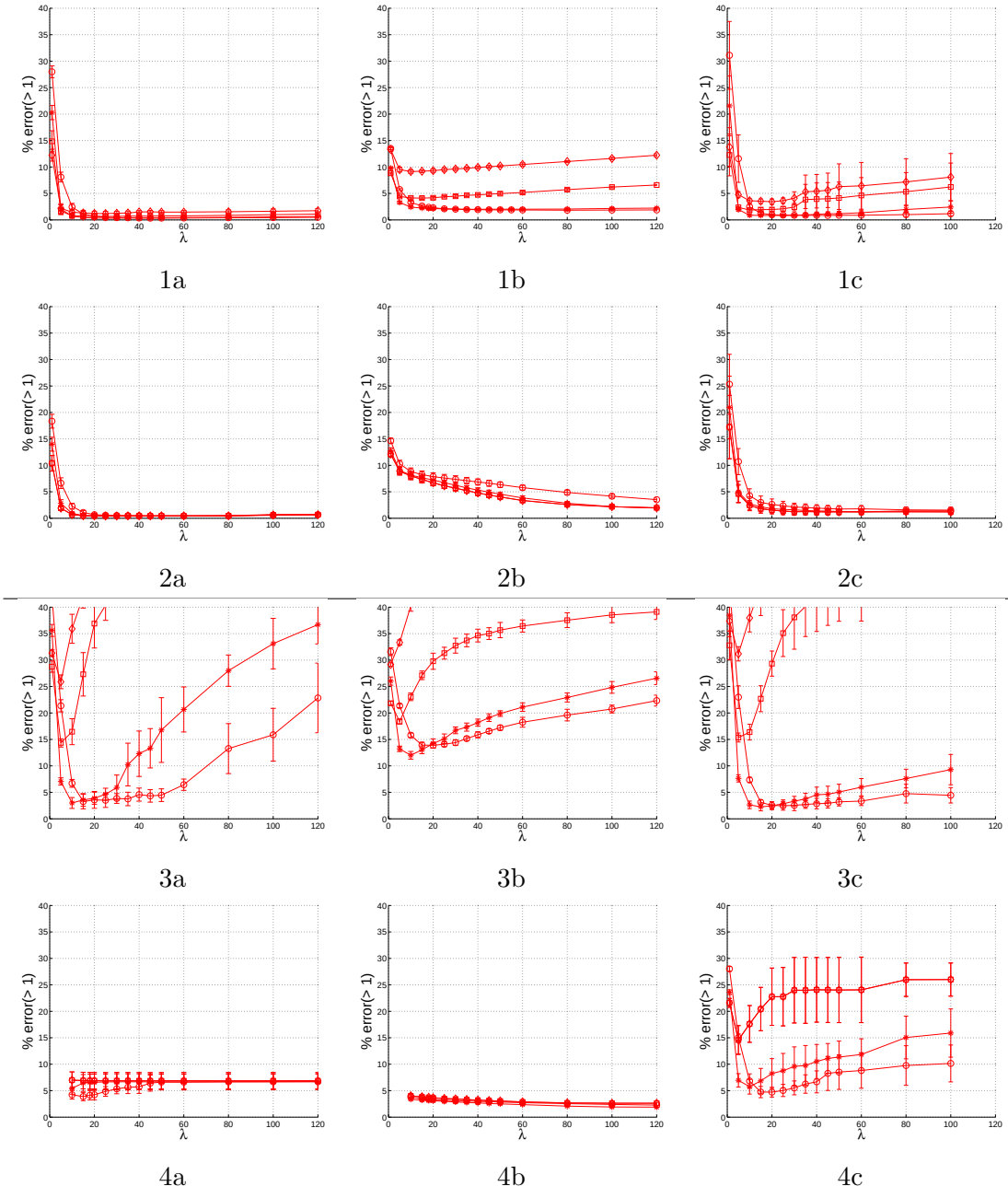


Figure A.10: Error statistics (> 1) for methods using the basic formulation

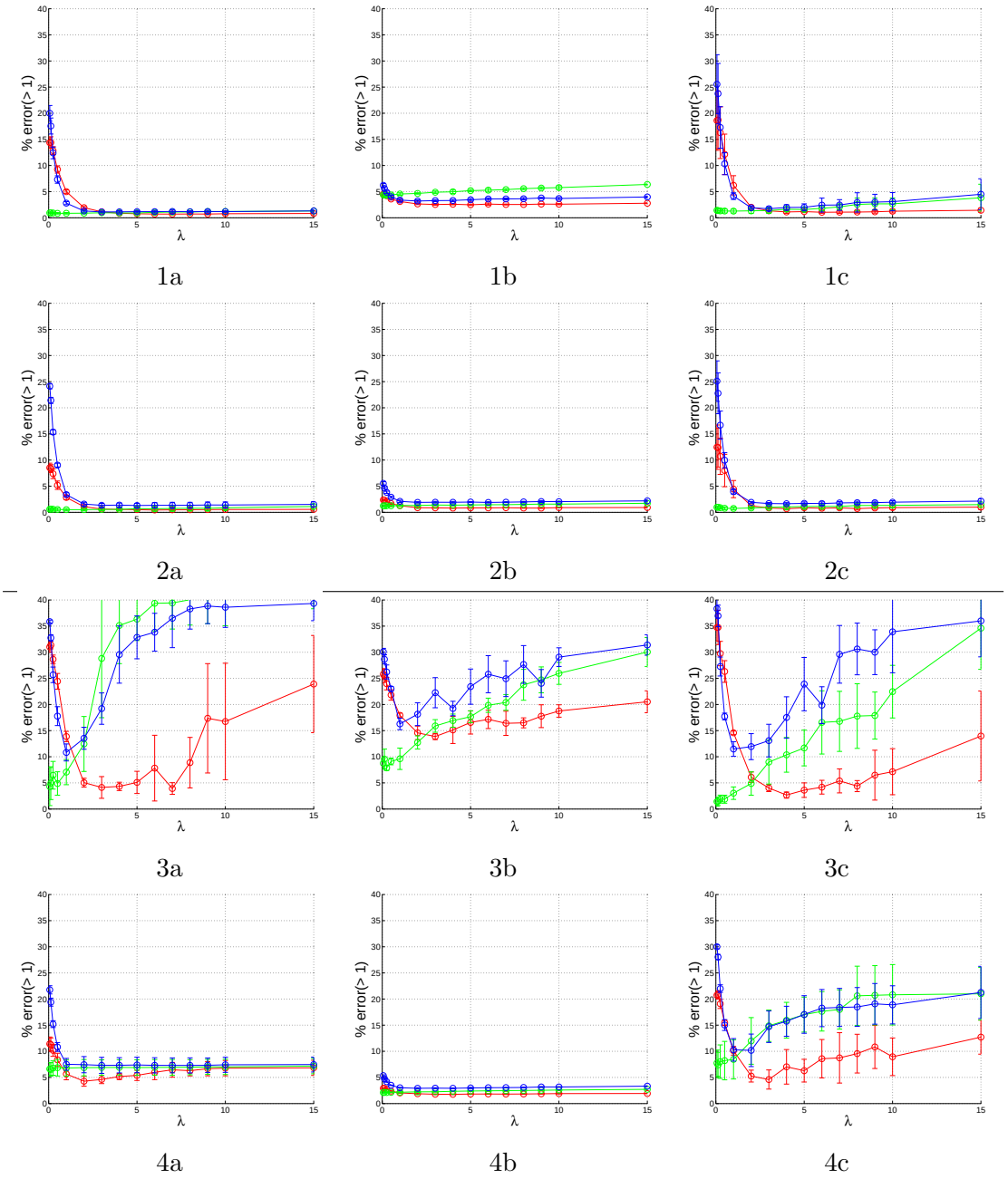


Figure A.11: Error statistics (> 1) for methods using additional constraints

References

[1] Michael S. Langer. **Surface Visibility Probabilities in 3D Cluttered Scenes**. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 401–412, Berlin, Heidelberg, 2008. Springer-Verlag. ii, 21, 22, 24, 26, 29, 32

[2] Vladimir Kolmogorov and Ramin Zabih. **Computing Visual Correspondence with Occlusions via Graph Cuts**. In *ICCV*, pages 508–515, 2001. ii, iv, 16, 17, 48, 80

[3] Vladimir Kolmogorov and Ramin Zabih. **Multi-camera Scene Reconstruction via Graph Cuts**. In Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, editors, *ECCV (3)*, **2352** of *Lecture Notes in Computer Science*, pages 82–96. Springer, 2002. ii, iv, 16, 48, 64, 80

[4] Vladimir Kolmogorov, Ramin Zabih, and Steven J. Gortler. **Generalized Multi-camera Scene Reconstruction Using Graph Cuts**. In Anand Rangarajan, Mário A. T. Figueiredo, and Josiane Zerubia, editors, *EMMCVPR*, **2683** of *Lecture Notes in Computer Science*, pages 501–516. Springer, 2003. ii, iv, 16, 17, 18, 48, 49, 59, 64, 80

[5] A. B. Lee, D. Mumford, and J. Huang. **Occlusion Models for Natural Images: A Statistical Study of a Scale-Invariant Dead Leaves Model**. *Int. J. Comput. Vision*, **41**(1-2):35–59, 2001. vii, 3, 21, 22

[6] S.Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer Publishing Company, Incorporated, 2009. 5

[7] Stuart Geman and Donald Geman. **Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images**. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **PAMI-6**(6):721 – 741, nov. 1984. 7

[8] J. Besag. **On the statistical analysis of dirty pictures**. *Journal of the Royal Statistical Society*, **B-48**:259–302, 1986. 7

[9] D. M. Greig, B. T. Porteous, and A. H. Seheult. **Exact Maximum A Posteriori Estimation for Binary Images**. *Journal of the Royal Statistical Society. Series B (Methodological)*, **51**(2):271–279, 1989. 7, 8

[10] Y. Boykov, O. Veksler, and R. Zabih. **Markov Random Fields with Efficient Approximations**. In *CVPR '98: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 648, Washington, DC, USA, 1998. IEEE Computer Society. 7

[11] Hiroshi Ishikawa and Davi Geiger. **Occlusions, Discontinuities, and Epipolar Lines in Stereo**. In *ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume I*, pages 232–248, London, UK, 1998. Springer-Verlag. 7, 8

[12] Sébastien Roy and Ingemar J. Cox. **A Maximum-Flow Formulation of the N-Camera Stereo Correspondence Problem**. In *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, page 492, Washington, DC, USA, 1998. IEEE Computer Society. 7, 8

[13] Sébastien Roy. **Stereo Without Epipolar Lines: A Maximum-Flow Formulation**. *Int. J. Comput. Vision*, **34**(2-3):147–161, 1999. 7, 8

[14] Y. Boykov, O. Veksler, and R. Zabih. **Fast approximate energy minimization via graph cuts**. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **23**(11):1222–1239, Nov 2001. 7, 8, 11, 48

[15] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. 7, 48

[16] V. Kolmogorov and R. Zabin. **What energy functions can be minimized via graph cuts?** *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **26**(2):147–159, Feb. 2004. 8, 10, 47

[17] M.J. Wainwright, T.S. Jaakkola, and A.S. Willsky. **MAP estimation via agreement on trees: message-passing and linear programming**. *Information Theory, IEEE Transactions on*, **51**(11):3697 – 3717, nov. 2005. 13, 48

[18] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. **A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors**. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **30**(6):1068–1080, June 2008. 13, 14, 20, 47, 48

[19] V. Kolmogorov. **Convergent Tree-Reweighted Message Passing for Energy Minimization**. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **28**(10):1568 –1583, oct. 2006. 14, 48

[20] Stan Birchfield and Carlo Tomasi. **A Pixel Dissimilarity Measure That Is Insensitive to Image Sampling**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**:401–406, 1998. 15

[21] V. Kolmogorov. *Graph based algorithms for scene reconstruction from two or more views*. PhD thesis, Cornell University, 2004. 17, 18

[22] D. Scharstein and R. Szeliski. **A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms**. *Int. J. Comput. Vision*, **47**(1-3):7–42, 2002. 19

REFERENCES

[23] Marshall F. Tappen and William T. Freeman. **Comparison of Graph Cuts with Belief Propagation for Stereo, using Identical MRF Parameters**. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 900, Washington, DC, USA, 2003. IEEE Computer Society. 19, 48, 50, 51

[24] D. Scharstein and R. Szeliski. **A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms**. *Int. J. Comput. Vision*, **47**(1-3):7–42, 2002. 20

[25] D. L. Ruderman and William Bialek. **Statistics of natural images: Scaling in the woods**. *Phys. Rev. Lett.*, **73**(6):814–817, Aug 1994. 21, 22

[26] G. J. Burton and Ian R. Moorhead. **Color and spatial structure in natural scenes**. *Appl. Opt.*, **26**(1):157–170, 1987. 22

[27] David J. Field. **Relations between the statistics of natural images and the response properties of cortical cells**. *J. Opt. Soc. Am. A*, **4**(12):2379–2394, 1987. 22

[28] J. Huang, A.B. Lee, and D. Mumford. **Statistics of range images**. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, **1**, pages 324–331 vol.1, 2000. 22

[29] G. Matheron. *Random sets and integral geometry*. Wiley New York,, 1974. 22

[30] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu. **On Advances in Statistical Modeling of Natural Images**. *J. Math. Imaging Vis.*, **18**(1):17–33, 2003. 22, 34

[31] Z. Chi. *Probability models for complex systems*. PhD thesis, Providence, RI, USA, 1998. Adviser-Gonan, Stuart. 22

[32] Michael S. Langer. **Visibility and smoothness probabilities in 3D cluttered scenes**. 23, 29

[33] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. **Understanding belief propagation and its generalizations**. pages 239–269, 2003. 48

[34] Oliver Woodford, Philip Torr, Ian Reid, and Andrew Fitzgibbon. **Global Stereo Reconstruction under Second-Order Smoothness Priors**. *IEEE Trans. Pattern Anal. Mach. Intell.*, **31**(12):2115–2128, 2009. 83