

MULTI-MODAL STEREO-VISION USING INFRARED / VISIBLE CAMERA
PAIRS

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

BY

MUSTAFA YAMAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

SEPTEMBER 2014

Approval of the thesis:

**MULTI-MODAL STEREO-VISION USING INFRARED / VISIBLE CAMERA
PAIRS**

submitted by **MUSTAFA YAMAN** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

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

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

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

Examining Committee Members:

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

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

Assoc. Prof. Dr. Alptekin Temizel
Graduate School of Informatics, METU

Assist. Prof. Dr. Pınar Duygulu Şahin
Computer Engineering Dept., Bilkent Uni.

Assist. Prof. Dr. Aykut Erdem
Computer Engineering Department, Hacettepe Uni.

Date:

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

Name, Last Name: MUSTAFA YAMAN

Signature :

ABSTRACT

MULTI-MODAL STEREO-VISION USING INFRARED / VISIBLE CAMERA PAIRS

Yaman, Mustafa

Ph.D., Department of Computer Engineering

Supervisor : Assist. Prof. Dr. Sinan Kalkan

September 2014, 198 pages

In this thesis, a novel method for computing disparity maps from a multi-modal stereo-vision system composed of an infrared-visible camera pair is introduced. The method uses mutual information as the basic similarity measure where a segmentation-based adaptive windowing mechanism is proposed along with a novel mutual information computation surface for greatly enhancing the results. Besides, the method incorporates joint prior probabilities when computing the cost matrix in addition to negative mutual information measures. A novel adaptive cost aggregation method is also proposed using computed cost confidences and resulting minimum cost disparities that are confident enough are fitted planes in segments. The segments are refined by iteratively splitting and merging according to the fitted confident disparities that helps to reduce the dependence of the disparity computation to the initial segmentation. Finally, all the steps are repeated iteratively where more accurate joint probabilities are calculated by using previous iteration's disparity map. Two multi-modal stereo image datasets are generated for evaluating the method and the state of the art methods confronted in literature; the synthetically altered image pairs from the Middlebury Stereo Evaluation Dataset, and our own dataset of Kinect Device infrared- visible camera image pairs, which can function as a benchmark for multi-modal stereo-vision methods. On these datasets, it is presented that (i) the proposed method improves the quality of existing MI formulation, (ii) the proposed method outperforms state of the art methods in literature, and (iii) the proposed method can

provide depth comparable to the quality of Kinect depth data.

Keywords: multi-modal stereo-vision, mutual information, segmentation, adaptive windowing, adaptive cost aggregation, plane fitting, iterative, infrared, visible, multi-modal camera, Kinect device

ÖZ

GÖRÜNÜR VE KIZILÖTESİ KAMERA ÇİFTLERİ KULLANARAK ÇOKLU BİÇİMLİ STERYO GÖRME

Yaman, Mustafa

Doktora, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi : Yrd. Doç. Dr. Sinan Kalkan

Eylül 2014, 198 sayfa

Bu tezde, kızılötesi ve görünür bantta kamera çiftleri kullanılarak yeni bir çoklu biçimli steryo görme yöntemi önerilmiştir. Yöntem karşılıklı bilgiyi temel benzerlik ölçüsü olarak kullanmaktadır. Ancak karşılıklı bilgi değerinin üretilmesinde bölütleme temelli bir uyarlanır pencere mekanizması önerilmiş ve sonuçlar oldukça geliştirilmiştir. Ayrıca yöntem maliyet matrisini hesaplarken, negatif karşılıklı bilgi değerine ek olarak önsel olasılık değerlerini de hesaba katmaktadır. Yeni bir uyarlanır maliyet toplama yöntemi de ayrıca önerilmiş ve bu yöntemde maliyet güvenilirlikleri hesaplanarak toplama işlemi gerçekleştirilirken, en küçük maliyetle eşleştirilen piksel farklılıklarından yeterince güvenilir olanlar ise daha sonra bölütlere göre düzleme oturtulmuştur. Bölütler yinelemeli olarak güvenilir farklılık değerlerine göre birleştirilip parçalanarak düzeltilmekte ve farklılık haritası hesaplamasının ilk bölütlemeye olan bağımlılığı azaltılmaktadır. Son olarak tüm bu adımlar yinelenmektedir. Yeni adımlarda daha önce üretilmiş olan farklılık haritası da kullanılarak daha doğru ön olasılık değerleri kullanılabilir. Önerilen yöntemi ve literatürde yer alan diğer yöntemleri değerlendirmek için iki farklı çoklu biçimli steryo görüntü veri kümesi oluşturularak kullanılmıştır: Middlebury Steryo Değerlendirme veri kümesinden sentetik olarak değiştirilmiş görüntü çiftleri ve Kinect cihazından elde edilen kızılötesi ve görünür kamera görüntü çiftleri. Veri kümeleri çoklu-biçimli steryo görme yöntemleri için performans değerlendirme test veri kümesi olarak kullanılabilir şekilde oluş-

turulmuştur. Bu veri kümelerinde, (i) önerilen yöntemin varolan karşılıklı bilgi formülasyonuna göre sonuçları iyileştirdiği, (ii) önerilen yöntemin literatürde yer alan diğer yöntemlerden daha iyi performans gösterdiği, ve (iii) önerilen yöntemin Kinect derinlik verisi ile karşılaştırılabilir derecede iyi derinlik verisi elde edebildiği gösterilmiştir.

Anahtar Kelimeler: çoklu biçimli steryo görme, karşılıklı bilgi, bölütleme, uyarlanırc pencere, uyarlanırc maliyet toplama, düzlem oturtma, yinelemeli, görünür, kızılötesi, çoklu biçimli kamera, Kinect cihazı

To Desen and Özüm Deren

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor Assist. Prof. Dr. Sinan Kalkan for his constant support, guidance and friendship. It was a great honor to work with him for the last three years and hope to continue working through the rest of my career. I'd like to thank to my thesis monitoring committee members Prof. Dr. Fatoş Tünay Yarman-Vural and Assist.Prof. Dr. Pınar Duygulu Şahin for their great support and review of the study at each six months of period. Their excellent comments and suggestions have directed me towards my goal more and more closer at each semester. Besides, I'd like to thank Assoc. Prof. Dr. Alptekin Temizel and Assist. Prof. Dr. Aykut Erdem for their review and approval of the study at my final thesis defense.

I'd like to thank my lovely wife Desen, for all her support in this painful years of study. This thesis would not have been written if she was not in my life. Also the love of my daughter, Özüm Deren, kept me alive and have been my source of life since she was born.

Lastly, sincerest thanks to each of my parents and my first teachers, Emine Yaman, Veysel Yaman and my sister M. Bahar Karahan for supporting and believing in me all the way through my life.

TABLE OF CONTENTS

ABSTRACT	v
ÖZ	vii
ACKNOWLEDGMENTS	x
TABLE OF CONTENTS	xi
LIST OF TABLES	xvi
LIST OF FIGURES	xviii
LIST OF ABBREVIATIONS	xxviii

CHAPTERS

1	INTRODUCTION	1
1.1	Problem Definition and Scope of the Thesis	6
1.2	Contributions of the Thesis	7
1.3	Outline of the Thesis	9
2	BACKGROUND	11
2.1	Stereo-vision	11
2.1.1	Overview of Stereo-vision Methods	14
2.1.1.1	Sparse / Feature based vs. Dense methods	14

	2.1.1.2	Local vs. Global methods	16
2.2		Multi-modal Stereo-vision	17
	2.2.1	Similarity Measures for Multi-Modal Stereo-vision	21
	2.2.1.1	Mutual Information (MI)	21
	2.2.1.2	Mutual Information with Prior Probabilities (MIwPR)	26
	2.2.1.3	Local Self-Similarity (LSS)	28
	2.2.1.4	Histogram of Oriented Gradients (HOG)	30
	2.2.1.5	Scale Invariant Feature Transform (SIFT)	30
	2.2.1.6	Speeded-Up Robust Features (SURF) .	34
	2.2.1.7	Census Transform (CENSUS)	36
	2.2.1.8	MI of Census Transform (CENSUSMI)	37
	2.2.1.9	Binary Robust Independent Elementary Features (BRIEF)	37
	2.2.1.10	Fast Retina Keypoint (FREAK)	39
	2.2.1.11	Normalized Cross Correlation (NCC) .	42
	2.2.1.12	Sum of Square Distances - SSD	43
3		DATASETS AND PERFORMANCE EVALUATION	45
	3.1	Dataset #1 - The Middlebury Dataset	45
	3.2	Dataset #2 - The Kinect Dataset	48
4		COMPARISON OF SIMILARITY MEASURES USED FOR MULTI-MODAL STEREO-VISION	55

4.1	Performance Evaluation Using Dataset #1 - The Synth. Alt. Middlebury Dataset	56
4.1.1	Effect of Window Size	57
4.1.2	Effect of Multi-Modality	60
4.1.3	Effect of Noise	64
4.2	Performance Evaluation Using Dataset #2 - The Kinect Dataset	66
4.3	Summary and Discussion	69
5	AN ITERATIVE MULTI-MODAL STEREO-VISION METHOD . . .	73
5.1	Method	73
5.1.1	Segmentation of the IR Image	74
5.1.2	Computing the Cost Matrix	75
5.1.3	Adaptive Cost Aggregation	79
5.1.4	Computation of Disparity Planes	81
5.1.4.1	Iterative Plane Fitting and Segment Splitting Step	82
5.1.4.2	Segment Merging and Finalization of Disparity Planes Step	84
5.1.5	Iterative Refinement	85
5.2	Experiments and Results	87
5.2.1	Performance Evaluation of the Proposed Method with State of the Art Similarity Measures	87
5.2.1.1	Results on Dataset #1	87
5.2.1.2	Results on Dataset #2	89

5.2.2	Performance Evaluation of Cost Aggregation and Plane Fitting Steps of the Proposed Method	91
5.2.3	Performance Evaluation of Iterative Refinement . .	92
5.2.4	Performance Evaluation on Dataset #2- Kinect Dataset	94
6	CONCLUSION	101
	REFERENCES	105
APPENDICES		
A	PARAMETER SETTINGS USED IN EXPERIMENTS	115
B	EXPERIMENT RESULTS FOR THE PERFORMANCE EVALUATION OF SIMILARITY MEASURES	119
C	EXPERIMENT RESULTS OF THE PROPOSED METHOD	141
C.1	Adaptive Windowing Step vs. State of the Art Similarity Measures for Dataset #1	141
C.2	Adaptive Windowing vs. State of the Art Similarity Measures using Dataset #2, The Kinect Dataset	149
C.3	Dataset #2 Results of the Proposed Method All Steps	153
D	PARAMETER ANALYSIS OF THE PROPOSED METHOD	165
D.1	Segmentation	166
D.2	Parameters of Adaptive Windowing Step	172
D.2.1	The Vertical Window Size - δ_y	172
D.2.2	The Ratio of Incorporating Prior Probabilities - λ .	172
D.2.3	The Thickness of Discontinuities - ω	172
D.2.4	The Histogram Size - $Size(h_w)$	173

D.2.5	The Incrementation Constant for Histogram Computation - k	174
D.3	Parameters of Adaptive Cost Aggregation Step	175
D.3.1	The Size of the Aggregation Window - $Size(w(p, q))$	175
D.3.2	The Scaling Parameters of the Aggregation Weights - λ_{SD} and λ_{DD}	176
D.4	Parameters of Iterative Plane Fitting Step	178
D.4.1	Confidence Threshold for Inlier Disparities - τ_{ic}	178
D.4.2	Stable Segment Ratio Threshold - τ_{ir}	180
D.4.3	Distance Threshold for Outlier Disparities - τ_{od}	182
D.4.4	Confidence Threshold for Outlier Disparities - τ_{oc}	184
D.4.5	Minimum Size Threshold for Segment Splitting of Outlier Disparities - τ_{os}	188
D.5	Parameters of Segment Merging and Finalizing Step	190
D.5.1	Angle Threshold of Coplanar Disparity Planes for Segment Merging- τ_{α}	190
D.5.2	Distance Threshold of Coplanar Disparity Planes for Segment Merging- τ_{pd}	190
D.6	Experiments on RGB and Cosine-Transformed RGB Image Pairs	191
CURRICULUM VITAE		197

LIST OF TABLES

TABLES

Table 3.1 The Dataset #1 - Synthetically Altered Middlebury Stereo Evaluation Dataset.	46
Table 3.2 The Dataset #2 - Kinect Dataset	51
Table 5.1 List of notations and acronyms used for the method descriptions . . .	75
Table 5.2 Average Results on the 24 images in the Dataset #2- The Kinect Dataset for WTA, WTA of Agg. Costs and Plane Fitted Disparities vs. Kinect's native depth in 2 Iterations.	95
Table 5.3 Standard Deviations of the Results on the 24 images in the Dataset #2 for WTA, WTA of Agg. Costs and Plane Fitted Disparities vs. Kinect native depth in 2 Iterations.	95
Table A.1 Parameter Settings used for the implementation of the evaluated similarity measures	116
Table A.2 Parameter Settings of the Proposed Method Used in Dataset #1 (Synt. Altered Middlebury) Experiments.	117
Table A.3 Parameter Settings of the Proposed Method Used in Dataset #2 (Kinect) Experiments.	117
Table B.1 Results on Dataset #1 for the Similarity Measures Tested using Three Different Window Sizes	119
Table B.2 Results on Dataset #2 Selected Image Pairs for all the Similarity Measures	136
Table C.1 Results on Dataset #1 for the WTA performances of the Adaptive Windowing Algorithm (ADAPMI) of the proposed method along with the Similarity Measures Tested using Three Different Window Sizes	141

Table C.2 Results on Dataset #2 Selected Image Pairs for WTA Performances of the Adaptive Windowing Algorithm (AdapMI) of the Proposed Method and the Similarity Measures Tested	149
Table C.3 Results on Dataset #2 for Proposed Method in two iterations	154
Table D.1 Parameters of the Proposed Method.	165
Table D.2 Parameters of 10 segmentation levels used.	166
Table D.3 Results of the Proposed Method on RGB images converted to Stereo Image Pairs.	196

LIST OF FIGURES

FIGURES

Figure 1.1	Sample brain CT and MR images and the registration results (Source: [15]) <i>Left column:</i> Sample CT image <i>Middle column:</i> Sample MR image <i>Right column:</i> Registration result	1
Figure 1.2	EO/IR spectral region of the EM spectrum along with definitions of the specific regions and the primary natural sources of EO/IR radiation (Source: [8])	2
Figure 1.3	Image samples from Landsat TM imaging satellite channels along with corresponding spectral signatures (Source:[16]). (a) The Red Channel (channel#3) (b) The Near-infrared (NIR) channel (channel#4) (c) The False color composite of channels#4,3,2 where red colored regions show the vegetation areas (d) Spectral signatures of green vegetation, soil and water [Best viewed in color].	3
Figure 1.4	Composition of Optical, SAR imagery and LIDAR data showing different aspects of each data sample of the same region of interest (Source:[17]).	4
Figure 1.5	Sample COTS products from defense and security market including multi-modal imaging systems for surveillance applications (a) FLIR’s Ranger HRC product (Source:[27]) (b) L3 Wescam’s MX-RSTA system (Source:[28]) (c) L3 Wescam’s MX-25D airborne system (Source:[29]) (d) FLIR’s SEAFLIR 380HD marine surveillance system (Source:[30]) (e)-(h) Associated platforms for the imaging systems	5
Figure 1.6	Sample visible-infrared image couples from multi-modal imaging systems for surveillance applications (Source: [22]). (a,e) Visible-SWIR image couple (b,f) Visible-MWIR image couple (c,g) Visible-LWIR image couple (d,h) another Visible-LWIR image couple.	5
Figure 1.7	Sample study from [23] computing the distance to the human target using average human height statistics.	7
Figure 1.8	The Kinect device (Source: [35]).	9

Figure 2.1 Recovering depth information using stereo-vision (Source: [39])	11
Figure 2.2 Rectification to ensure the epipolar constraint (Source: [39])	13
Figure 2.3 A view from Middlebury Stereo Evaluation web page with comparison of referenced studies. (Source: [45])	15
Figure 2.4 A sketch of local- window based matching over the epipolar line.	16
Figure 2.5 One of Egnal’s experiments, where he uses an image from a camera with a red filter used on the left image and blue filter used on the right image. MI outperforms MNCC for intensity conversions while enhancing accuracy for larger windows (Source: [54]).	18
Figure 2.6 Another experiment from Egnal ([54]), including a NIR(left) and Visible/NIR(right) image couple. MI outperforms MNCC at high confidence levels.	19
Figure 2.7 Experiments of Fookes <i>et al.</i> on synthetically altered stereo image pairs (Source: [34]). (a) Negative (b) Solarized (c) Posterized (d) Simulated images. Disparities computed with (1) ZNCC (2) Rank Transform and (3) MI	20
Figure 2.8 Results from [69], showing examples of sparse depth maps from outdoor scenarios (red color in cost map corresponds to high cost values) [Best Viewed in Color]	22
Figure 2.9 Sample view from experiments of Torabi <i>et al.</i> [74]. (a) Depiction of 1D sliding window algorithm over visible-thermal stereo image pairs for human ROI identification. (b) Results for comparison of evaluated methods for human ROI detection over visible-thermal stereo image pairs.	23
Figure 2.10 Joint histogram of an MR image and its rotated version for 0, 2, 5 and 10 degrees from left to right along with computed joint entropy values at the bottom row. (Source: [1])	26
Figure 2.11 Depiction of the change in joint histogram, entropy and MI of registered and unregistered image pairs (images at top row: reference and current image; bottom row: the difference image and the joint histogram image) (a) registered image (b) unregistered image (Source: [80])	27
Figure 2.12 Reported template matching results from [70] using LSS, including a comparison with MI.	29
Figure 2.13 Depiction of computing HOG descriptors process (Source: [82])	31

Figure 2.14 Depiction of generation of DoG images for the scale space images in each octave (Source: [46])	32
Figure 2.15 Depiction of computing the SIFT local descriptor: <i>Left</i> : shows the computed gradients and orientations for each sample point in the neighbor region and the Gaussian window for weighting <i>Right</i> : 2x2 descriptor windows (the subregions) to accumulate gradients and construct 8-bin orientation histogram (Source: [46])	34
Figure 2.16 Depiction of SURF descriptor vector responses of the subregions of several patterns (Source: [47]) <i>Left</i> : homogeneous region corresponds to low values <i>Middle</i> : high frequencies in x direction corresponds to high values in $ d_x $ <i>Right</i> : increasing intensity in x direction corresponds to high values for both d_x and $ d_x $	36
Figure 2.17 Illustration of the MI of Census Transform method where multimodal image pairs are census transformed and processed for stereo correspondence using the mutual information of the transformed images	38
Figure 2.18 Illustration of the five approaches proposed for the construction of the sampling grid and the sampling pairs for encoding the BRIEF similarity measure (Source: [86]).	40
Figure 2.19 Human vision system and depiction of proposed computer vision system (Source: [87]) (a) Depiction the human visual system along with the layers transferring visual information from photo-receptors to ganglion cells which encodes and transfers data to brain (courtesy of [88]) (b) Proposed computer vision system structure from pixels to encoded binary strings for object recognition. (c) Depiction of retinal sampling pattern of receptive fields (<i>left image</i>) and sample pairings of the receptive fields (<i>right image</i>)	42
Figure 3.1 Tsukuba, Venus, Teddy and Cones stereo pairs from the Middlebury Stereo Vision Page - Evaluation Version 2 [45]. <i>Left column</i> : Synthetically altered left images. <i>Middle column</i> : The right images. <i>Right column</i> : The ground truth disparities. Note that, in the left image, important details are lost due to the cosine transformation.	47
Figure 3.2 Regions where evaluations are performed, for Tsukuba, Venus, Teddy and Cones image pairs. Only "White" pixels are included in performance evaluation calculations. (a)-(d) the "all" regions including regions of both non-occluded discontinuities (c)-(h) the "disc" regions - discontinuities (i)-(l) the "nonocc" regions - non-occluded regions.	48

Figure 3.3	The Kinect Device : (a) Kinect device built-in camera, sensors and features (Source: [91]) (b) Illustration of depth image generation process (Source: [92])	49
Figure 3.4	Depiction of the Kinect calibration process	50
Figure 3.5	Sample image pairs from Dataset #2 - Kinect Dataset <i>Left column:</i> Left (IR) camera images. <i>Middle column:</i> Right (RGB) camera images. <i>Right column:</i> Kinect's native depth computations	52
Figure 4.1	Average RMS(all) errors of all methods' "WTA" performances for three different window sizes for Dataset #1 - the synthetically altered Middlebury image pairs	57
Figure 4.2	Average Bad(all) pixels percentage errors of all methods' "WTA" performances for three different window sizes for the dataset1 - the synthetically altered Middlebury image pairs	58
Figure 4.3	Sample visual results of the leading similarity measures for the synthetically altered Tsukuba image pair in Dataset #1, for the different window sizes 9x9, 21x21 and 31x31 pixels. (includes the results of the novel CENSUSMI method along with the original CENSUS method results)	59
Figure 4.4	Figure illustrating the method for generating images of different multi-modality. <i>m</i> : the multi-modality level scale ($m=0.5$ in this case) <i>Left image:</i> Original Tsukuba image from Middlebury image database <i>Middle image:</i> Cosine transformed image <i>Right image:</i> Generated image of multi-modality level $m=0.5$	60
Figure 4.5	Average RMS(all) errors of all methods for 10 multi-modality levels for the Dataset#1 image pairs	61
Figure 4.6	Average Bad(all) pixel percentage errors of all methods for 10 multi-modality levels for the Dataset#1 image pairs	62
Figure 4.7	Sample visual results of selected similarity measures for given multi-modality levels (M9, M6, M3 and M0) of the Tsukuba image pair (local window size=21x21). 1st row shows altered left images of given multi-modality levels).	63
Figure 4.8	Different noise levels applied to left Tsukuba image in Dataset #1. (a) Noise level $n = 10$ ($\sigma = 20.0$) (b) Noise level $n = 6$ ($\sigma = 12.0$) (c) Noise level $n = 3$ ($\sigma = 6.0$) (d) Noise level $n = 0$ ($\sigma = 0.0$) the noiseless cosine transformed left image.	64

Figure 4.9 Average RMS(all) errors of all methods for 10 noise levels for the dataset#1 image pairs	65
Figure 4.10 Average Bad(all) pixel percentage errors of all methods for 10 noise levels for the dataset#1 image pairs	66
Figure 4.11 Sample visual results of some similarity measures for the added noise levels to Tsukuba left image in Dataset #1 (local window size=21x21) (noise levels: N10 ($n = 10$), N6 ($n = 6$), N3 ($n = 3$) and noiseless N0 ($n = 0$)	67
Figure 4.12 Selected Dataset #2 - Kinect Dataset Image Pairs and kinect computed depth images: <i>Left column</i> : Left (IR) camera images. <i>Middle column</i> : Right (RGB) camera images. <i>Right column</i> : Kinect's native depth computations (brighter pixels have more depth). From top to bottom, Dataset #2 Image Ids : Img#2, Img#3, Img#6,Img#10	69
Figure 4.13 Average <i>Percentage Good Depth</i> and <i>Percentage Total Coverage</i> metrics computed for similarity measures for the Dataset #2 selected image pairs.	70
Figure 4.14 Sample visual results of computed WTA disparities of the similarity measures for Kinect06 image in Dataset #2 (local window size=31x31) (brighter pixels have more disparity meaning more closer and have less depth)	71
Figure 5.1 Overview of the proposed iterative multi-modal stereo-vision method.	74
Figure 5.2 The Doll dog from a Kinect image pair in Dataset #2 - The Kinect Image Database (a) The Kinect RGB Image. (b) The Kinect IR image. [Best viewed in color].	76
Figure 5.3 Adaptive window calculation.	78
Figure 5.4 Adaptive MI computation surface using segmentation.	79
Figure 5.5 Cost Confidences for Tsukuba image pair (scaled and truncated to [0..255] range for the sake of visibility).	80

Figure 5.6 The intermediate steps of the proposed method. (a) WTA disparities of raw costs with No-Adaptive Windowing - 1st iteration (MI(woPR)). (b) WTA disparities of raw costs with Adaptive Windowing - 1st iteration. (c) WTA disparities after adaptive cost aggregation - 1st iteration. (d) Plane fitted disparities - 1st iteration. (e-h) Resultant plane fitted disparities for iterations 1-4. (i) The initial segmentation of the left image. (j-l) The input segmentations to iterations 2-4 (after segment break & merge steps are applied in the previous iteration). (m-p) Edge map of the corresponding input segmentation at each iteration. [Best viewed in color]	86
Figure 5.7 Average RMS (all) errors of "WTA" performances using the Dataset #1 for the Performance Evaluation of Adaptive Windowing Algorithm (ADAPMI) to state of the art similarity measures.	88
Figure 5.8 Average Bad (all) pixels percentage errors of "WTA" performances using the Dataset #1 for the Performance Evaluation of Adaptive Windowing Algorithm (ADAPMI) to state of the art similarity measures.	88
Figure 5.9 Sample visual results of the WTA disparity results of the Adaptive Windowing algorithm (ADAPMI) and the leading similarity measures for the synthetically altered Tsukuba image pair in Dataset #1, using different window sizes 9x9, 21x21 and 31x31 pixels.	89
Figure 5.10 Average <i>Percentage Good Depth</i> and <i>Percentage Total Coverage</i> metrics computed for similarity measures for the dataset#2 selected image pairs and the Adaptive Windowing Algorithm (ADAPMI) of the proposed method - initial iteration.	90
Figure 5.11 Sample visual results of computed WTA disparities of the Adaptive Windowing Algorithm (ADAPMI) of the proposed method and the similarity measures for Img#2 in Dataset #2 (local window size=31x31)	91
Figure 5.12 Results on Dataset #1 - Synt. Altered Middlebury Images for WTA of Adaptive Windowing Costs, WTA of Adaptively Aggregated Costs and Plane Fitting.	92
Figure 5.13 Effect of iterations on RMS and the percentage of bad pixels for "all" regions (a-b) Tsukuba, (c-d) Venus pairs.	93
Figure 5.14 Effect of iterations on RMS and the percentage of bad pixels for "all" regions (a-b) Teddy and (c-d) Cones stereo pairs.	94
Figure 5.15 Depiction of Average Results on the 24 images in the Dataset #2- The Kinect Dataset for WTA, WTA of Agg. Costs and Plane Fitted Disparities vs. Kinect's native depth in 2 Iterations	96

Figure 5.16 Sample visual results of computed disparity maps compared to native depth of Kinect, for Kinect01 image pair <i>1st row</i> : WTA disparity of aggregation results- 1st and 2nd iteration. <i>2nd row</i> : Plane fitting disparity results - 1st and 2nd iteration. <i>3rd row</i> : Kinect's native depth image (brighter pixels are farther)	97
Figure 5.17 Sample visual results of computed disparity maps compared to native depth of Kinect, for Kinect10 image pair <i>1st row</i> : WTA disparity of aggregation results- 1st and 2nd iteration. <i>2nd row</i> : Plane fitting disparity results - 1st and 2nd iteration. <i>3rd row</i> : Kinect's native depth image (brighter pixels are farther)	98
Figure 5.18 Merged 3D rendering of Kinect01 image's native depth map to proposed method's final depth map where invalid depth in the native depth map is filled with the proposed method's depth information.	99
Figure B.1 RMS(all) and Bad(all) pixels percentage errors of all methods' "WTA" performances for three different window sizes for Tsukuba image in Dataset1	122
Figure B.2 RMS(all) and Bad(all) pixels percentage errors of all methods' "WTA" performances for three different window sizes for Venus image in Dataset1	123
Figure B.3 RMS(all) and Bad(all) pixels percentage errors of all methods' "WTA" performances for three different window sizes for Teddy image in Dataset #1.	124
Figure B.4 RMS(all) and Bad(all) pixels percentage errors of all methods' "WTA" performances for three different window sizes for Cones image in Dataset #1.	125
Figure B.5 continued	127
Figure B.5 The visual results of all the similarity measures for the Tsukuba image pair, for the different window sizes 9x9, 21x21 and 31x31.	127
Figure B.6 RMS(all) and Bad(all) pixels percentage errors of all methods for 10 multi-modality levels for Tsukuba image in Dataset #1 [Best viewed in color].	128
Figure B.7 RMS(all) and Bad(all) pixels percentage errors of all methods for 10 multi-modality levels for Venus image in Dataset #1 [Best viewed in color].	129

Figure B.8 RMS(all) and Bad(all) pixels percentage errors of all methods for 10 multi-modality levels for Venus image in Dataset #1 [Best viewed in color].	130
Figure B.9 RMS(all) and Bad(all) pixels percentage errors of all methods for 10 multi-modality levels for Cones image in Dataset #1 [Best viewed in color].	131
Figure B.10 RMS(all) and Bad(all) pixels percentage errors of all methods for 10 noise levels for Tsukuba image in Dataset #1 [Best viewed in color].	132
Figure B.11 RMS(all) and Bad(all) pixels percentage errors of all methods for 10 noise levels for Venus image in Dataset #1 [Best viewed in color].	133
Figure B.12 RMS(all) and Bad(all) pixels percentage errors of all methods for 10 noise levels for Teddy image in Dataset #1 [Best viewed in color].	134
Figure B.13 RMS(all) and Bad(all) pixels percentage errors of all methods for 10 noise levels for Cones image in Dataset #1 [Best viewed in color].	135
Figure C.1 RMS(all) and Bad(all) pixels percentage errors of "WTA" performance of the Adaptive Windowing Algorithm (ADAPMI) and the state of the art similarity measures for three different window sizes for Tsukuba image in Dataset1	145
Figure C.2 RMS(all) and Bad(all) pixels percentage errors of "WTA" performance of the Adaptive Windowing Algorithm (ADAPMI) and the state of the art similarity measures for three different window sizes for Venus image in Dataset1	146
Figure C.3 RMS(all) and Bad(all) pixels percentage errors of "WTA" performance of the Adaptive Windowing Algorithm (ADAPMI) and the state of the art similarity measures for three different window sizes for Teddy image in Dataset1	147
Figure C.4 RMS(all) and Bad(all) pixels percentage errors of "WTA" performance of the Adaptive Windowing Algorithm (ADAPMI) and the state of the art similarity measures for three different window sizes for Cones image in Dataset1	148
Figure D.1 The four sample segmentation maps for the Tsukuba image from over-segmentation to under-segmentation (a) 1st segmentation level (b) 4th segmentation level (c) 7th segmentation level (d) 10th segmentation level [Best viewed in color]	166

Figure D.2 Average Results for Different Levels of Segmentation for Dataset #1.	167
Figure D.3 continued	169
Figure D.3 continued	170
Figure D.3 continued	171
Figure D.3 The results of for the Dataset #1 image pairs, for the 10 different segmentation levels from over-segmentation to under-segmentation	171
Figure D.4 Results for Increasing δ_y for Dataset #1.	173
Figure D.5 Results of Increasing λ for Dataset #1.	174
Figure D.6 Effect of increasing ω on Dataset #1.	175
Figure D.7 Effect of increasing histogram size on Dataset #1.	176
Figure D.8 Results of different k parameter values for Dataset #1.	177
Figure D.9 Effect of increasing aggregation window size on Dataset #1 (WTA cost aggregation results).	178
Figure D.10 WTA Results of Cost Aggregation Step Regarding Increasing Scaling Parameter Values for the Aggregation Weights Computation for Dataset #1.	179
Figure D.11 Average Results for different τ_{ic} as confidence threshold for determining inlier disparities to perform plane fitting.	180
Figure D.12 Results for each of the image pair in Dataset #1 for different τ_{ic} as confidence threshold for determining inlier disparities to perform plane fitting.	181
Figure D.13 Average Results for different τ_{ir} as stable segment ratio threshold.	182
Figure D.14 Results for each of the image pair in Dataset #1 for different τ_{ir} as stable segment ratio threshold	183
Figure D.15 Average Results for different τ_{od} as disparity distance threshold for determining outlier disparities to split segment after plane fitting.	184
Figure D.16 Results for each of the image pair in Dataset #1 for different τ_{od} as disparity distance threshold for determining outlier disparities to perform segment splitting after plane fitting.	185

Figure D.17 Average Results for different τ_{oc} as confidence threshold for determining outlier disparities to split segment after plane fitting.	186
Figure D.18 Results for each of the image pair in Dataset #1 for different τ_{oc} as confidence threshold for determining outlier disparities to perform segment splitting after plane fitting.	187
Figure D.19 Average Results for different τ_{os} as minimum size threshold for segment splitting	188
Figure D.20 Results for each of the image pair in Dataset #1 for different τ_{os} as minimum size threshold for selecting outlier disparity regions for segment splitting.	189
Figure D.21 Average Results for different τ_{α} as angle threshold for determining coplanar disparity planes for segment merging	191
Figure D.22 Results for each of the image pair in Dataset #1 for different τ_{α} as angle threshold for checking coplanarity of two disparity planes for segment merging.	192
Figure D.23 Average Results for different τ_{pd} as disparity distance threshold for determining the coplanar disparity planes for segment merging	193
Figure D.24 Results for each of the image pair in Dataset #1 for different τ_{pd} as distance threshold for checking coplanarity of two disparity planes for segment merging.	194
Figure D.25 The free-form RGB images converted to stereo image pairs. <i>Left column:</i> Original RGB image. <i>Middle column:</i> The cosine transformed images used as the left image. <i>Right column:</i> The shifted RGB images used as the right image.	195

LIST OF ABBREVIATIONS

2D	Two Dimension
3D	Three Dimension
BRIEF	Binary Robust Independent Elementary Feature
COTS	Commercial of the Shelf
CT	Computer Tomography
EM	Electromagnetic
EO	Electro-Optical
FREAK	Fast Retina Keypoint
FPGA	Field Programmable Gate Array
GPU	Graphics Processing Unit
HOG	Histogram of Oriented Gradients
IR	Infrared
LIDAR	Light Detection and Ranging
LRF	Laser Range Finder
LSS	Local Self Similarity
LWIR	Long-Wave Infrared
MI	Mutual Information
MR	Magnetic Resonance
MWIR	Mid-Wave Infrared
NCC	Normalized Cross Correlation
NDVI	Normalized Difference Vegetation Index
NIR	Near Infrared
PET	Positron Emission Tomography
RGB	Red, Green, Blue
RGB-D	Red, Green, Blue and Depth
ROI	Region of Interest
SAD	Sum of Absolute Differences
SAR	Synthetic Aperture Radar

SIFT	Scale Invariant Feature Transform
SSD	Sum of Square Distances
SURF	Speeded-Up Robust Features
SWIR	Short-Wave Infrared
UAV	Unmanned Aerial Vehicle
US	Ultrasound
WTA	Winner Takes All

CHAPTER 1

INTRODUCTION

Imaging systems of different modalities have been in use for a long time, especially in medical imaging [1, 2, 3, 4, 5, 6, 7] and remote sensing [8, 9, 10, 11, 12, 13, 14]. In such systems, registration and/or fusion of such imagery is a major concern since information from multiple modalities need to be combined for a solving a task. However, this is challenging since objects and surfaces look very different in different modalities in order to relate them to.

In medical imaging, registration of images from multi-modal imaging systems (e.g. CT-MR, CT-PET, MR-PET, MR-US and even CT-2D Video images [1]) is very important since complementary information can be acquired by analyzing such imagery together. For example, CT and MR imagery (Figure 1.1) should be analyzed together for the planning of radiation therapy where CT is used to calculate the radiation dose while MR provides the region of target lesion which radiation shall be imposed [6].

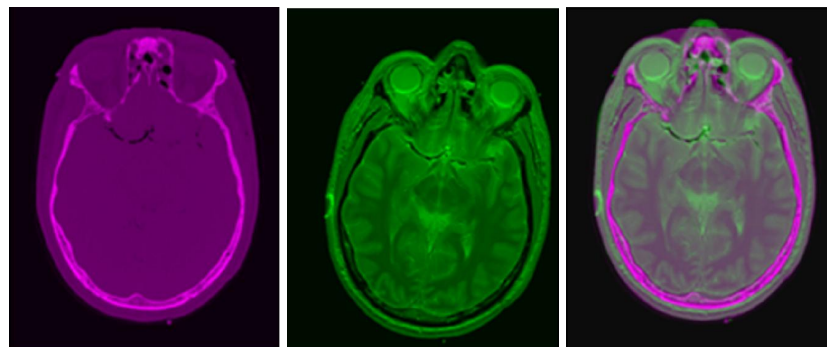


Figure 1.1: Sample brain CT and MR images and the registration results (Source: [15]) *Left column*: Sample CT image *Middle column*: Sample MR image *Right column*: Registration result

In remote sensing, imaging systems that acquire images from different band intervals of the electro-magnetic (EM) spectrum are used. Similarly, to be able to analyze a target scene of interest by combining information from different portions of the EM spectrum provides valuable information [8, 9]. Most remote sensing systems work on the EO/IR (Electro-optic/Infrared) spectral region of the EM spectrum where the wavelength ranges from 400nm to 14000nm (Figure 1.2).

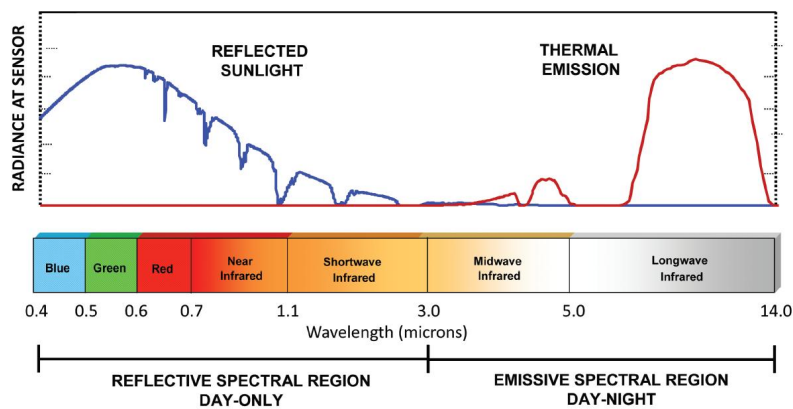


Figure 1.2: EO/IR spectral region of the EM spectrum along with definitions of the specific regions and the primary natural sources of EO/IR radiation (Source: [8])

To be able to get accurate land-cover/land-use classifications, the reflectivity and emissivity properties of objects existing in the region of interest are used. Different types of objects show different reflectance and emission characteristics at different wavelengths from visible to thermal bands in the EO/IR spectrum, such that objects can be differentiated by their so called “spectral signatures”. For instance, Figure 1.3 shows the spectral signatures of water, soil and vegetation along with corresponding band intervals in Landsat TM channels. In order to be able to distinguish vegetation from the ground, it is needed to use the Near-Infrared (NIR) channel#4 along with the visible red channel#3 which has lead to the popular NDVI (Normalized Difference Vegetation Index) measure [9] that can extract vegetation regions in multi-spectral images.

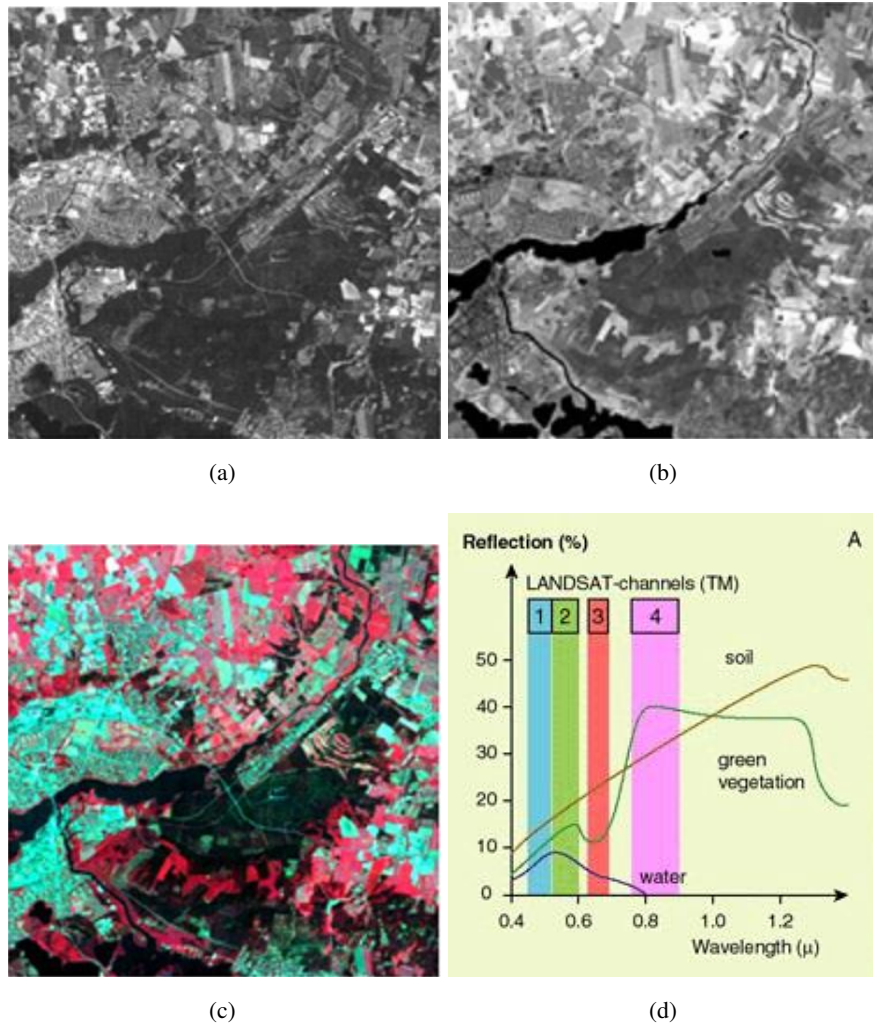


Figure 1.3: Image samples from Landsat TM imaging satellite channels along with corresponding spectral signatures (Source:[16]). (a) The Red Channel (channel#3) (b) The Near-infrared (NIR) channel (channel#4) (c) The False color composite of channels#4,3,2 where red colored regions show the vegetation areas (d) Spectral signatures of green vegetation, soil and water [Best viewed in color].

Another popular multi-modal image analysis application in remote sensing is to fuse information from passive optical and active sensors such as SAR (Synthetic Aperture Radar) and LIDAR (Light Detection and Ranging) [17, 18, 19, 20, 21]. SAR images have the ability to provide structural information but not the actual type of the land coverage where, for instance, it can be used to distinguish urban areas and the cleared areas for development when used along with a Landsat TM image [9]. Similarly, LIDAR data provides very detailed structural information of the vertical structures and

the terrain surface mapping, which brings opportunities for mapping and analyzing urban areas [19], forest areas [20] or damage assessment / change detection purposes after hazards like flooding, earthquakes etc. [21] when used along with SAR and optical (hyper-spectral or multi-spectral) sensors.

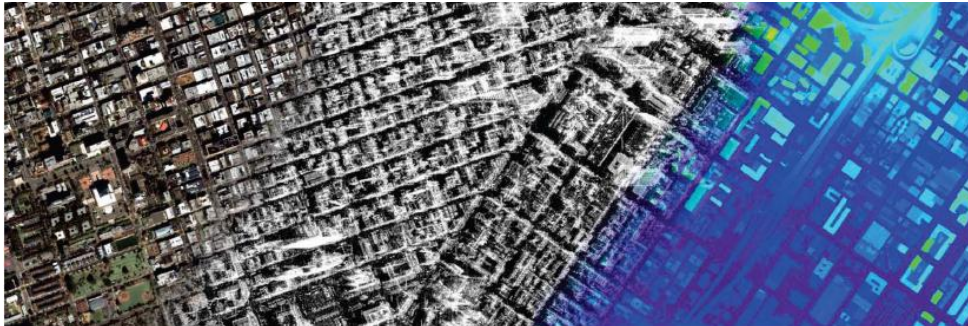


Figure 1.4: Composition of Optical, SAR imagery and LIDAR data showing different aspects of each data sample of the same region of interest (Source:[17]).

On the other hand, in recent years, using multi-modal cameras for surveillance systems has been growing in popularity [22, 23, 24, 25, 26]. Multi-modal surveillance can combine information from multiple sources provided by different types of sensors in order to be able to get the most accurate and robust interpretation of a target environment [24]. The type of sensors can range from imaging sensors to audio, thermal, infrared, vibration sensors etc. Especially, using bi-modal systems including visible and infrared/thermal cameras has become quite prevalent in such systems since surveillance should continue during daytime, night-time, under low visibility or lighting conditions [27, 28, 29, 30]. Figure 1.5 provides some of the COTS(Commercial off the Shelf) products available in the defense and security market today where there are a wide variety of imaging systems that include visible and infrared camera pairs for surveillance operations over different platforms, such as ground platforms, ground vehicle platforms, airborne platforms (UAVs, fixed-wing or rotary-wing aircrafts etc.) and marine platforms (coastal guard ships etc.). These systems also include laser range finders for accurate distance calculation to target objects, which is also an important information for surveillance operations to be able to detect an intruder human or a vehicle target.

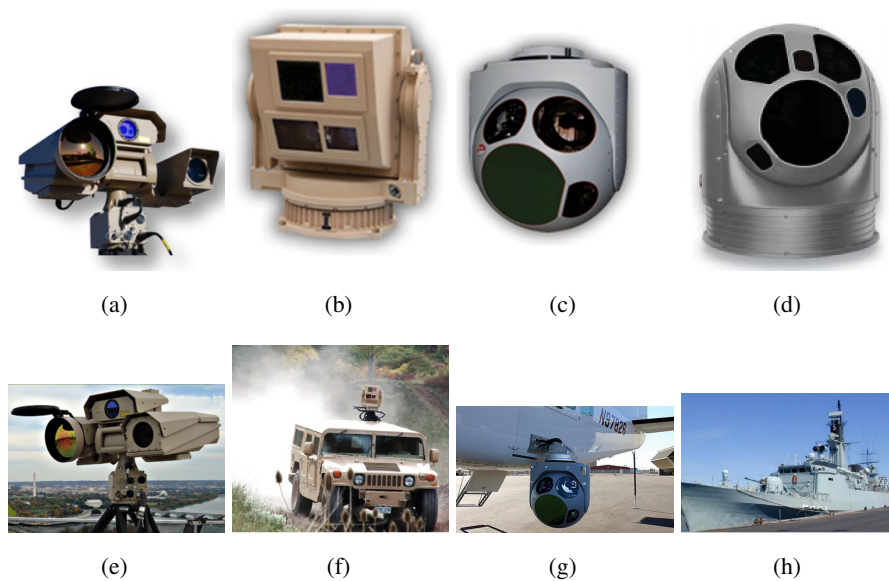


Figure 1.5: Sample COTS products from defense and security market including multi-modal imaging systems for surveillance applications (a) FLIR's Ranger HRC product (Source:[27]) (b) L3 Wescam's MX-RSTA system (Source:[28]) (c) L3 Wescam's MX-25D airborne system (Source:[29]) (d) FLIR's SEAFLIR 380HD marine surveillance system (Source:[30]) (e)-(h) Associated platforms for the imaging systems

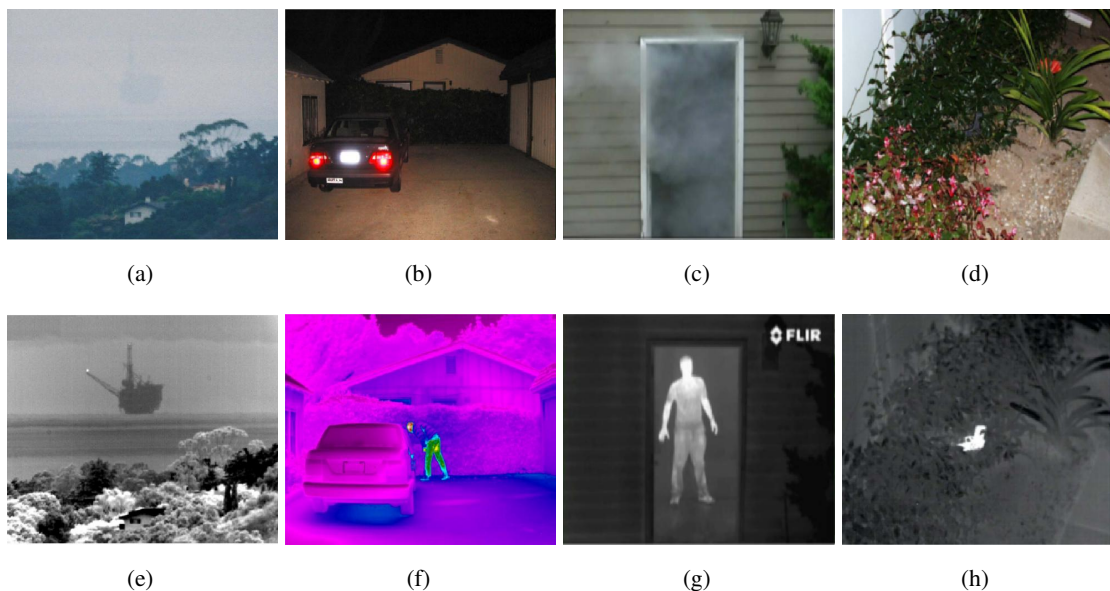


Figure 1.6: Sample visible-infrared image couples from multi-modal imaging systems for surveillance applications (Source: [22]). (a,e) Visible-SWIR image couple (b,f) Visible-MWIR image couple (c,g) Visible-LWIR image couple (d,h) another Visible-LWIR image couple.

In order to illustrate the advantages of such multi-modal imaging systems in surveillance applications, Figure 1.6 shows some representative images acquired from such surveillance systems. In Figures 1.6(a-d), an image couple from both visible and short-wave infrared cameras are shown. In Figures 1.6(e-h), additional information that can be acquired by the cameras operating in MWIR and LWIR wavelength regions (see Figure 1.2) are demonstrated. For instance, it can be observed that in the SWIR image, the low-visibility conditions due to the haze are eliminated where LWIR, MWIR images provide visibility in night, smoke and camouflaged objects of different material type.

1.1 Problem Definition and Scope of the Thesis

Computing depth / distance information by the surveillance system products is also a very important information for the operator or the soldier to detect the distance of the intruder or the vehicle or any other target to the sensor. Currently, this is done by an additional active system like a Laser Range Finder (LRF) [27, 28, 29, 30] or photogrammetrically by assuming an average height of the human, the vehicle or the vehicle tires detected on the sensor [23] (see Figure 1.7).

The other alternative is to use a unimodal stereo-vision system with an additional camera of the same modality, where stereo-vision methods are quite a well-studied area currently (Refer to Chapter 2.1 for the background on stereo-vision methods). However, currently there is no system yet computing a dense depth map of the scene from the visible and an infrared camera directly by using multi-modal stereo-vision techniques, as we know of, although the cameras are already available in these systems. When such a technique is integrated on these systems, an additional active system like an LRF or an additional camera of the same modality will not be necessary. Besides, the LRF device can only compute distance on the pointed location of the target scene which is not a dense depth map calculation either.



Figure 1.7: Sample study from [23] computing the distance to the human target using average human height statistics.

Therefore, in this thesis, the problem of whether an accurate depth information can be computed or not from multi-modal imagery is investigated. The thesis focuses on dense stereo-vision techniques, which do not require any additional device but can directly use the information from the already available visible and infrared cameras. The classical stereo-vision techniques are not applicable in this multi-modal case because of the total difference in corresponding pixel intensities which makes the problem hard and challenging to solve (see Figure 1.6).

1.2 Contributions of the Thesis

The contributions of the thesis can be summarized as below:

- A. During the thesis study, a novel multi-modal stereo-vision method is developed which can accurately generate dense disparity maps of images taken from cameras of different modalities. The method is compared to alternative methods that were confronted during the literature survey and is shown to outperform these state of the art methods. This part of the study is disseminated in [31, 32].
- B. For the performance evaluation, two image datasets are generated. Since, up to the author's knowledge, there are no multi-modal stereo-vision datasets available in the literature, this is considered as another important contribution of the thesis. The first dataset is generated from four popular images in the Middlebury Stereo Vision Page [33] (tsukuba, venus, cones and teddy) where the left images are replaced with the synthetically altered versions of these images by performing cosine transform ($\cos(\pi I/255)$) of pixel intensities just as Fookes [34] did. This

way, it was possible to compute the statistics of test results to gain more knowledge of the performance of our method and also it is now possible to compare the result metrics of developed methods with the ones on the evaluation site although they are results for the unimodal image pairs. This part of the study is disseminated in [31, 32].

The second dataset was generated by our several Kinect camera shootings. The Kinect devices (see [35] and Figure 1.8) have a built-in infrared (IR) (left), RGB (right) camera along with an IR projector. The projected infrared beams are detected by the IR camera and the depth of the scene can be acquired for several purposes like motion recognition for the associated platform, the Xbox 360 video game console [36]. Such RGB-D devices are becoming more prevalent and affordable each day and drawing more attention in computer vision studies where the built-in depth computation feature enables many vision applications and opens up important opportunities on a wide range of areas [37]. In our study, this feature enabled us to use generated depth data as some reference point to be able compare to our computed depth results. For this purpose, we have proposed a method for performance statistics computation over the Kinect's native depth computation which enabled us to evaluate our results quantitatively in addition to visual evaluation. This part of the study is disseminated [32, 38].

- C. A systematic performance evaluation of alternative similarity measures available in the literature was also performed as part of this thesis. This is important because there are no such complete study yet available in the literature, as we know of. The evaluation was performed over the datasets generated in the scope of this thesis and "Winner Takes All" (WTA) performances of similarity measures for the stereo correspondence problem was evaluated. Along with this evaluation, another novel similarity measure that is composed of a *modified version of Census Transform* was also proposed and evaluated as part of this same study, which includes computation of mutual information over the census transformed images of the initial left-right multi-modal image pairs. This part of the study is disseminated in [38].



Figure 1.8: The Kinect device (Source: [35]).

1.3 Outline of the Thesis

The rest of the thesis is structured as follows:

In Chapter 2, background information and literature survey about stereo-vision, multi-modal stereo-vision and description of methods that were proposed or tested for multi-modal stereo-vision are provided. Initially, an overview of stereo-vision techniques are provided along with the recent advances in the field. Next, literature survey on the multi-modal stereo-vision methods are provided including studies for dense or sparse techniques and related applications like human region of interest (ROI) disparity detection in thermal-visible image pairs.

In Chapter 3, the test image datasets that were prepared in the scope of this thesis and the details of the proposed and implemented performance evaluation methods are presented. Detailed descriptions on the two dataset types generated by synthetically altering the popular Middlebury Stereo-vision Evaluation Dataset and the Kinect Device infrared/visible camera image pairs are provided. The calibration and the epipolar rectification method performed for preparing the Kinect multi-modal stereo images are also presented. Besides, the description on the performance evaluation statistics calculation methods that were proposed for the datasets are also provided.

Chapter 4 provides a performance evaluation of the state of the art methods for multi-modal stereo-vision that were confronted in the literature survey. Both statistical and visual results are provided on the acquired results. For the evaluation, the datasets

generated in the scope of this thesis are used. The results are also discussed within the scope of this chapter along with explanations of how some measures were inadequate for multi-modal stereo-vision whereas some have promising results.

Chapter 5, on the other hand, includes the description of the novel dense multi-modal stereo-vision algorithm that were developed in the scope of this thesis. Besides, performance evaluation on the datasets are provided. The method is compared also the alternative state of the art similarity measures that were confronted in the literature which are evaluated in Chapter 4 and shown to outperform these methods. Both visual and statistical results are provided over the two datasets for evaluating the novel method.

Finally, in Chapter 6 conclusions and discussions are provided. The advantages and drawbacks of the proposed method are provided. Besides, future work that can be performed to enhance the method and alternative application areas that the study can be diverted or applied are discussed.

CHAPTER 2

BACKGROUND

In this chapter, initially stereo-vision is described along with an overview of the methods in the literature. Next, the problem of multi-modal stereo-vision is explained and the state of the art methods proposed and tested for multi-modal imagery are described in detail.

2.1 Stereo-vision

Stereo-vision problem is defined as computing the depth information of a scene by using images taken from two distinct viewpoints. This is generally implemented by two cameras staring at the same scene located by a defined distance from each other.

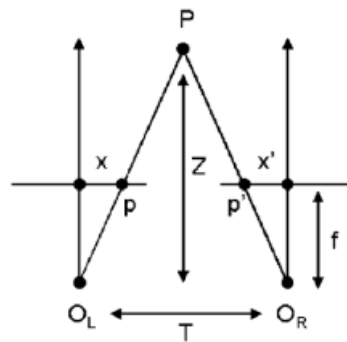


Figure 2.1: Recovering depth information using stereo-vision (Source: [39])

Figure 2.1 presents basically the geometrical relationships for the depth information

computation where T is called the stereo baseline, the distance between optical centers of the two cameras, O_L and O_R and f is the focal length of both cameras. Then from similar triangles, it is trivial to compute the depth of the 3D point P as:

$$\frac{T + x' - x}{Z - f} = \frac{T}{Z}, \quad (2.1)$$

$$Z = f \frac{T}{d}, \quad (2.2)$$

$$d = x' - x, \quad (2.3)$$

where d is called as *disparity*.

However, this computation requires three known variables; the focal lengths, the stereo baseline and the so-called disparity term d , which is the difference between the projected 2D pixel locations assuming they are on the same row (scan line) of the images. This leads to three major problems of stereo-vision; *calibration*, *correspondence*, *reconstruction*.

Calibration is the process of determining both the internal and external geometry characteristics of the cameras. The information such as the focal lengths and optical centers of cameras are among the internal geometry characteristics whereas the relative position and orientation of two cameras are external geometry characteristics to be determined. In this study, it is assumed that these information are already available and static. However, this is a well-known and long-studied problem with successful solutions [40].

Correspondence on the other hand, is a major problem that it was also needed to address in this study. Correspondence deals with determining the corresponding 2D projections (pixel locations) of a 3D scene on the images acquired from the two cameras. The difference in the locations is called the *disparity*, and the matrix of the all the corresponding disparities is called the *disparity map*. However, not all features could be visible from the left and/or right camera because some objects in the scene can occlude others partially or completely from the either camera's point of view.

This so-called occlusion problem is a big challenge and yet no general solution exists. Besides, lighting differences and texture mismatches add a certain difficulty to the correspondence problem and when all combined, correspondence problem is still a big challenge in vision. The problem is generally relaxed by using additional constraints and assumptions such as the epipolar constraint, constant brightness of the two cameras, surface smoothness etc.

Epipolar constraint is the most important of such constraints which reduces the search space for a match to a single line in the other image. The cameras should horizontally lie on the same plane and their horizontal scan lines should perfectly correspond to each other. However, mostly the case is as given in Figure 2.2. Here, a so-called epipolar plane is defined by lines T , $[PO_L]$ and $[PO_R]$ which intersects the actual image planes along the so-called epipolar lines. The epipolar lines give us the location of the projected points p and p' of any point P on the scene, i.e. they are on their respective epipolar lines. However, we can easily project the two image planes onto a common plane and this way we can get same coordinate systems for the two images where the epipolar lines lie on the same horizontal scan-line. Therefore, a 2D correspondence problem becomes a 1D problem, which greatly reduces the complexity of the problem. This method is called *image rectification* [40] and is generally a preliminary step before stereo correspondence methods are applied. In this study, it was assumed to have rectified image pairs available and the stereo correspondence problem was focused.

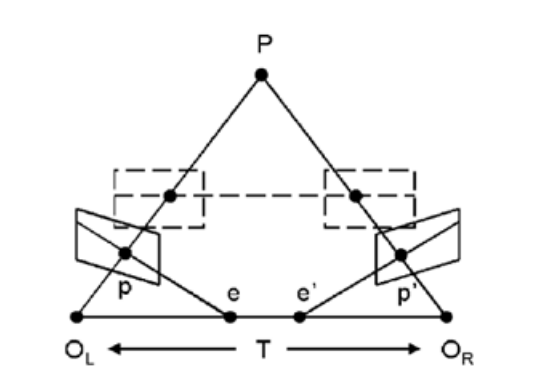


Figure 2.2: Rectification to ensure the epipolar constraint (Source: [39])

Reconstruction is indeed determining the 3D structure of a scene from the disparity

map and known camera geometries. Having the calibrated cameras setup as given in Figure 2.1 or images are rectified as given in Figure 2.2, the problem reduces directly to the equations given in Equation 2.1 to 2.3 above.

2.1.1 Overview of Stereo-vision Methods

Stereo vision is a very-well studied problem with a huge literature on the topic. Several reviews are already available where the most important ones are [41] (Dhond *et al.*) from 1980s, and [39] (Brown *et al.*) from 1990s to 2000s. Besides, the book from Hartley and Zisserman [40] provides almost all the aspects of the multiple view stereo problem. Recently, the study of Scharstein and Szeliski [42] provided a taxonomy of comparing the current state of the art solutions to the problem and provided a website¹ enabling researchers to get ready datasets with ground truth images available and compare the performance characteristics of their methods and publish them. **The website only provides unimodal images** and four popular images in the datasets are used to compare and rank the methods, the tsukuba, venus, teddy and cones (see Figure 2.3). More recent reviews are also available by [43] (Lazaros *et al.*) and [44] (Tippets *et al.*). These reviews focus on more resource-limited algorithms and hardware (for instance, FPGA or GPU based) solutions.

Stereo-vision methods are mainly clustered around two main axes:

- Sparse / Feature-based vs. Dense Methods
- Local vs. Global Methods

2.1.1.1 Sparse / Feature based vs. Dense methods

This grouping describes whether correspondences (and therefore the pixel disparities) are computed for all the pixels in the images (i.e., the dense methods), or only for some reliable features extracted from images, such as salient points, edges, corners, curves etc.

¹ <http://vision.middlebury.edu/stereo/>

Stereo Table Version 3 (beta)

Middlebury Stereo Evaluation - Version 2

[New features and main differences to version 1.](#)
[Submit and evaluate your own results.](#)

Open a new window for each link

Error Threshold = 1 Error Threshold...		Sort by nonocc			Sort by all			Sort by disc			Average percent of bad pixels (explanation)			
Algorithm	Avg.	Tsukuba ground truth			Venus ground truth			Teddy ground truth				Cones ground truth		
	Rank	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc		nonocc	all	disc
TSGO [143]	10.5	0.87 ⁴	1.13 ¹	4.66 ⁸	0.11 ⁷	0.24 ⁹	1.47 ¹⁰	5.61 ³⁸	8.09 ¹⁷	13.8 ³¹	1.67 ¹	6.16 ¹	4.95 ¹	4.06
JSOSP+GCP [151]	11.8	0.74 ¹	1.34 ⁸	3.98 ¹	0.08 ²	0.16 ¹	1.15 ²	3.96 ¹⁴	10.1 ³³	11.8 ¹⁷	2.28 ¹⁵	7.91 ²⁹	6.74 ¹⁸	4.18
ADCensus [82]	14.1	1.07 ¹⁹	1.48 ¹⁷	5.73 ²²	0.09 ³	0.25 ¹²	1.15 ²	4.10 ¹⁸	6.22 ⁷	10.9 ¹³	2.42 ²¹	7.25 ¹⁸	6.95 ²²	3.97
AdaptinqBP [16]	17.9	1.11 ²²	1.37 ¹⁰	5.79 ²⁴	0.10 ⁵	0.21 ⁸	1.44 ⁹	4.22 ¹⁸	7.06 ¹⁵	11.8 ¹⁸	2.48 ²⁵	7.92 ³¹	7.32 ³⁰	4.23
CoopRegion [39]	18.4	0.87 ⁶	1.16 ²	4.61 ⁸	0.11 ⁸	0.21 ⁸	1.54 ¹⁴	5.16 ³⁰	8.31 ²⁰	13.0 ²⁵	2.79 ⁴³	7.18 ¹⁴	8.01 ²⁰	4.41
CCRADAR [152]	22.5	1.15 ²⁵	1.42 ¹⁵	6.23 ³⁸	0.15 ¹⁹	0.27 ¹⁵	1.89 ²⁴	5.39 ³³	10.6 ³⁸	14.7 ⁴²	2.01 ²	7.37 ¹⁷	5.88 ²	4.75
RDP [87]	23.8	0.97 ¹¹	1.39 ¹²	5.00 ¹¹	0.21 ⁴⁰	0.38 ³¹	1.89 ²⁴	4.84 ²²	9.94 ³²	12.6 ²²	2.53 ²⁹	7.69 ²¹	7.38 ³¹	4.57
MultiRBF [129]	24.2	1.33 ⁴⁹	1.56 ²²	6.02 ³³	0.13 ¹¹	0.17 ³	1.84 ²¹	5.09 ²⁸	6.36 ⁸	13.4 ²⁹	2.90 ⁵¹	6.76 ⁸	7.10 ²⁷	4.39
DoubleBP [34]	24.9	0.88 ⁸	1.29 ⁶	4.76 ⁹	0.13 ¹²	0.45 ⁴⁹	1.87 ²³	3.53 ¹²	8.30 ¹⁹	9.63 ⁸	2.90 ⁵⁰	8.78 ⁶¹	7.79 ⁴²	4.19
MDPM [140]	25.2	1.15 ²⁴	1.59 ²⁵	6.14 ³⁶	0.14 ¹⁷	0.36 ²⁸	1.52 ¹³	3.79 ¹³	5.78 ⁵	11.1 ¹⁵	2.74 ³⁷	8.38 ⁴⁵	7.91 ⁴⁵	4.22
CVW-RM [148]	25.2	1.12 ²³	1.42 ¹⁴	5.99 ³²	0.16 ²⁶	0.36 ²⁹	1.40 ⁸	4.70 ²¹	6.94 ¹³	12.1 ¹⁹	2.96 ⁵⁵	7.71 ²³	7.72 ³⁹	4.38
OutlierConf [40]	25.5	0.88 ⁷	1.43 ¹⁸	4.74 ⁸	0.18 ²⁹	0.26 ¹⁴	2.40 ⁴⁰	5.01 ²⁴	9.12 ²⁸	12.8 ²⁴	2.78 ⁴²	8.57 ⁵¹	6.99 ²³	4.60
SeqAggr [146]	26.5	1.99 ⁹¹	2.39 ⁸²	8.59 ⁹⁰	0.12 ⁸	0.21 ⁷	1.68 ¹⁶	2.19 ²	3.73 ¹	7.02 ²	2.16 ⁸	6.52 ³	6.37 ⁸	3.58
AdaptiveGF [127]	30.2	1.04 ¹⁵	1.53 ¹⁸	5.62 ¹⁷	0.17 ²⁵	0.41 ³⁹	1.98 ²⁸	5.71 ⁴²	11.3 ⁴⁹	14.3 ³⁸	2.44 ²³	8.22 ³⁹	7.05 ²⁸	4.98
SOS [135]	30.3	1.45 ⁶⁰	1.63 ²⁹	7.83 ⁷⁹	0.21 ³⁸	0.32 ²¹	2.29 ³⁹	3.13 ⁸	8.45 ²²	9.74 ⁹	2.43 ²²	7.10 ¹³	7.02 ²⁴	4.30
SubPixSearch [109]	30.8	2.04 ⁹⁵	2.48 ⁸⁸	6.40 ⁴⁴	0.14 ¹⁶	0.40 ³⁸	1.74 ¹⁸	4.00 ¹⁵	6.39 ⁹	11.0 ¹⁴	2.24 ¹²	6.87 ¹⁰	6.50 ¹²	4.18
SubPixDoubleBP [29]	32.0	1.24 ³³	1.76 ⁴³	5.98 ³¹	0.12 ¹⁰	0.46 ⁵¹	1.74 ¹⁸	3.45 ¹¹	8.38 ²¹	10.0 ¹¹	2.93 ⁵³	8.73 ⁵⁸	7.91 ⁴⁴	4.39
SurfaceStereo [71]	32.1	1.28 ⁴²	1.65 ³²	6.78 ⁵³	0.19 ³¹	0.28 ¹⁹	2.61 ⁵³	3.12 ⁷	5.10 ³	8.65 ⁴	2.89 ⁴⁹	7.95 ³⁴	8.26 ⁶¹	4.06
LLR [117]	34.0	1.05 ¹⁶	1.65 ³¹	5.64 ¹⁸	0.29 ⁶⁵	0.81 ⁸³	3.07 ⁶⁸	4.56 ¹⁹	9.81 ³¹	12.2 ²⁰	2.17 ⁹	8.02 ³⁵	6.42 ¹⁰	4.64

Figure 2.3: A view from Middlebury Stereo Evaluation web page with comparison of referenced studies. (Source: [45])

As noted by Dhond and Aggarwal [41], sparse methods were much more popular at 80's initially, especially due mainly to the computational efficiency. In addition, these methods have the advantage of performing better under contrast/illumination variations, especially by the advances on feature descriptors used by image registration methods such as SIFT, SURF, Harris or many available edge or shape-based descriptors [46, 47, 48, 49, 50, 51]. However, only limited information can be extracted from features. Performance of sparse methods can be improved by using hierarchical grouping of features (e.g., [52]), or using segmentation of the images for matching [53].

Dense methods on the other hand attracted most researchers during the last decade. Along with significant improvements on computer hardware performances, the motivation is reported to come from the needs of current applications such as view syn-

thesis or image-based rendering requiring disparity maps of all pixels including the occluded or textureless regions. Scharstein and Szeliski totally ignored sparse methods for their comparative study [42].

2.1.1.2 Local vs. Global methods

Local methods use only the local neighborhood information for finding stereo correspondences. These methods usually perform a WTA (winner takes all) over the computed costs using this neighborhood information while performing matching (see Figure 2.4). Dense methods performing a window-based matching over pixels in this neighborhood are grouped as local methods. As the similarity metric, Sum of Squared Distances (SSD), Sum of Absolute Distances (SAD), Normalized Cross Correlation (NCC) and Mutual Information (MI) are widely used [54, 55, 56, 57]. Of course, the window size brings an important limitation to the solutions. This limitation can be levitated by using windows whose sizes are adaptively changed, e.g. [58]. Sparse or feature-based methods can also be local [39, 52].

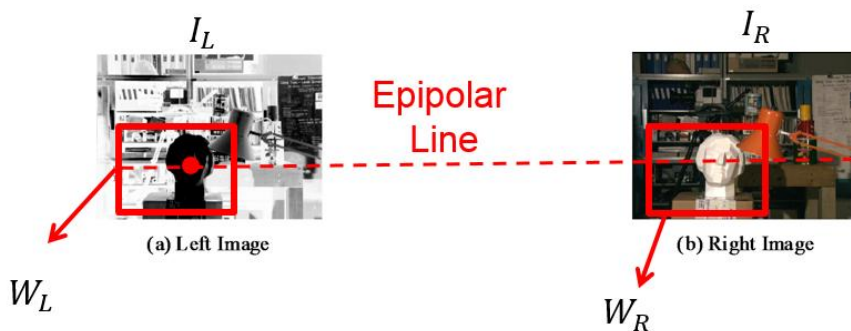


Figure 2.4: A sketch of local- window based matching over the epipolar line.

Global methods, on the other hand, use global constraints to correct wrong correspondences that are otherwise not possible locally. Although computational complexity is much higher, significant improvements in the accuracy of disparity maps can be achieved. Dynamic Programming techniques, and several energy minimization or global optimization techniques have been employed for this purpose. Dynamic Programming techniques make use of ordering constraint and ensure a minimum cost is achieved on the specific epipolar line [58, 59, 60]. Other energy minimization ap-

proaches add smoothness, uniqueness or occlusion constraints for defining a global energy function, and employ methods such as Graph cuts, Simulated Annealing and Belief Propagation to minimize this energy function [42, 61, 62, 63, 64].

2.2 Multi-modal Stereo-vision

Multi-modal stereo-vision is defined as performing stereo-vision using two camera pairs of different modalities, for instance an infrared-visible camera pair.

Stereo-vision from multi-modal camera pairs has not been studied much until 2000s. The earliest of such studies, up to the our knowledge, is from Egnal [54], who, affected from Viola's studies of multi-modal registration [65], applied mutual information (MI) as the basic similarity measure for stereo correspondence. Egnal tested his method both on unimodal images and also on red / blue filtered, multimodal (an NIR and Visible/NIR image couple) and differently lighted images and compared to a modified NCC algorithm as a baseline. The results were promising and revealed the power of MI compared to standard correlation-based methods especially on images with different spectral characteristics for the same scene. However, using MI still had low quality. Figure 2.5 and 2.6 show sample results from his studies.

Fookes *et al.* extended the MI-based approach with adaptive windowing [66] and integrated prior probabilities using a 2D matching surface [34]. The 2D match surface is simply performed by computing MI costs for every possible combination of left and right pixels in the match window and putting on a 2D surface. Then, first, maximum of one row is found. It is compared by all the costs on the same column and if it is also maximum of the column, then determined as a valid match. This is claimed to enforce left-right and uniqueness constraints. Prior probability incorporation is performed by computing a joint histogram from all the intensities in the stereo image pair and using this as prior probabilities added to the joint probability of matching windows along with a weighting constant. The results were taken from synthetically altered unimodal images actually, by first computing the negative, solarized, posterized and simulated (by $\cos(I * \pi/255)$) versions of one of the image pairs and compared to results from NCC and rank transform where MI outperforms all these methods (see

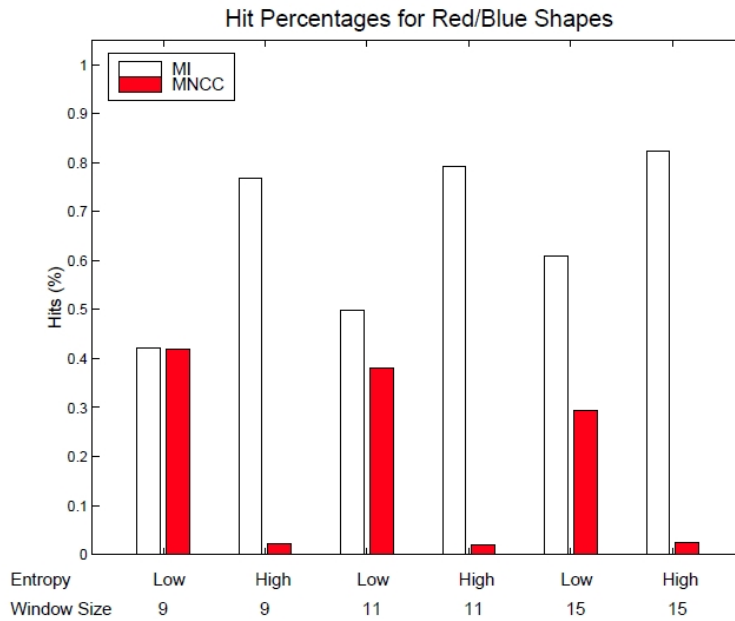
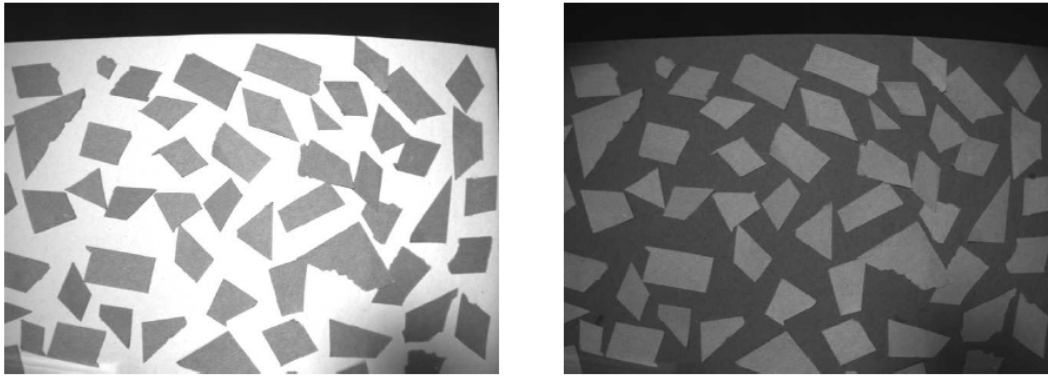


Figure 2.5: One of Egnal’s experiments, where he uses an image from a camera with a red filter used on the left image and blue filter used on the right image. MI outperforms MNCC for intensity conversions while enhancing accuracy for larger windows (Source: [54]).

Figure 2.7). This study is important to show that stereo-vision results using MI could be significantly enhanced when combined with other state-of-the-art stereo-vision techniques, however, the results were taken only from synthetically altered unimodal images, which do not actually include different segmentation / edge characteristics that multi-modal images may have.

Krotosky and Trivedi [26, 67, 68] used mutual information for an infrared-visible camera pair for pedestrian detection and tracking. They applied mutual information for stereo correspondence within region of interests (ROI) including the human bod-

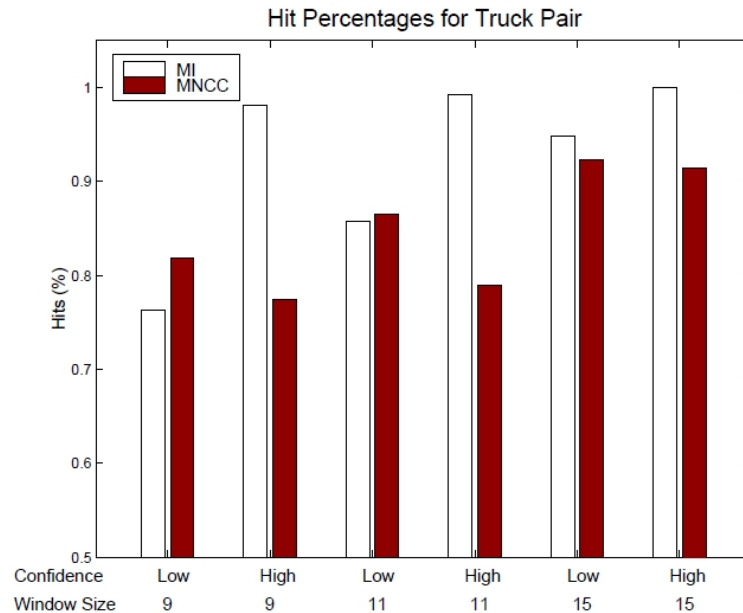


Figure 2.6: Another experiment from Egnal ([54]), including a NIR(left) and Visible/NIR(right) image couple. MI outperforms MNCC at high confidence levels.

ies, and propose a disparity voting method for computing the final depth information for the corresponding regions as a significant restriction. Finally, this depth information is used to accurately register the multi-modal images for the ROIs.

Campo *et al.* [69] proposed an MI-based method where the similarity measures were extended using the gradient information. They developed a multi-modal stereo rig (with thermal and visible cameras) and a database. The 3D depth results presented in their work are quite sparse for the scenes tested; however, their results are promising since they show that stereo-vision is possible from images with very distinct spectral bands. Figure 2.8 shows results from [69], presenting examples of sparse depth maps from outdoor scenarios (red color in cost map corresponds to high cost values).

Recently, a measure, called Local Self Similarity (LSS), originally proposed for im-

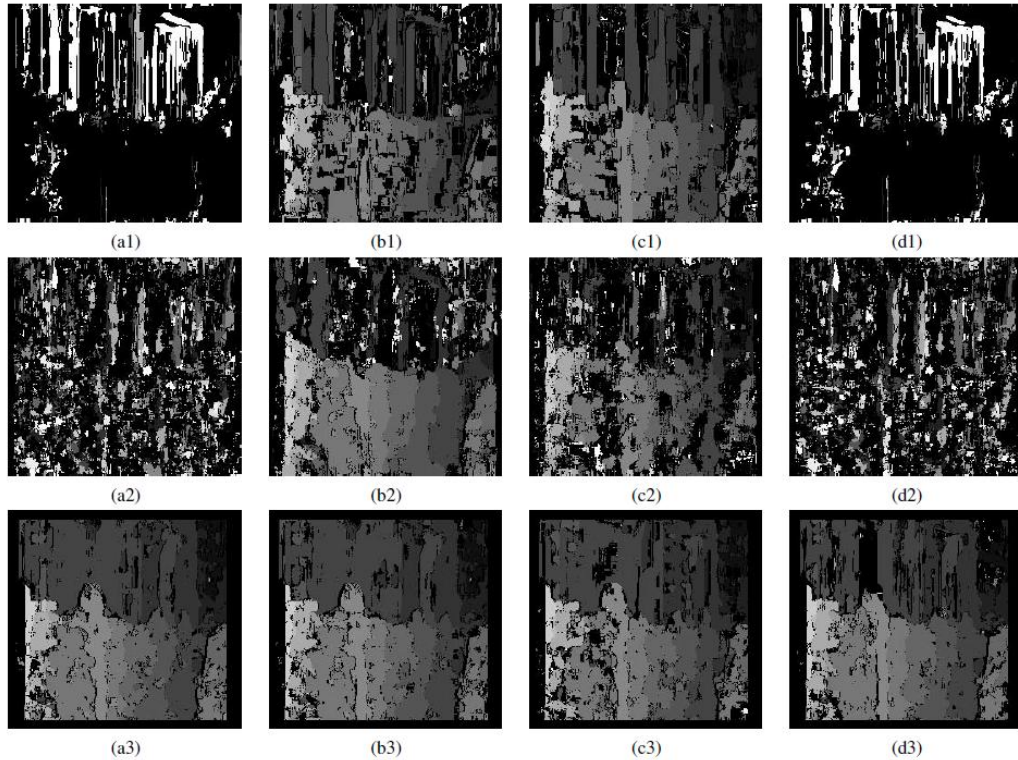


Figure 2.7: Experiments of Fookes *et al.* on synthetically altered stereo image pairs (Source: [34]). (a) Negative (b) Solarized (c) Posterized (d) Simulated images. Disparities computed with (1) ZNCC (2) Rank Transform and (3) MI

age template matching [70], has been applied as a thermal-visible stereo correspondence measure by Torabi and Bilodeau [71]. They implemented a ROI-based image matching by tracking people in the scene according to their silhouettes, and compared it against MI-based similarity descriptors. In their first publication [72], they showed that LSS measures outperform MI and HoG (Histogram of Oriented Gradients). Later, they used LSS measure in an energy minimization framework, enhancing the results compared to their previous work [73].

In their latest publication [74], with more data (about 300 images), they compared LSS and MI with (i) traditional descriptors such as SIFT, SURF, HOG, (ii) binary descriptors such as Census, Fast RETina Keypoint (FREAK) or Binary Robust Independent Elementary Feature (BRIEF) and (iii) direct comparisons of windows based on SSD, NCC. In this study, MI and LSS were shown to be the leading measures for ROI-based image matching of human silhouettes (see Figure 2.9). MI outperformed LSS showing that it is still the best choice for multi-modal image windows

matching; however, for smaller window sizes where the objects of interest were small or segmented into small fragments or there were many occlusions between objects, LSS performed better. On the other hand, LSS measure is not yet tested for a dense disparity map estimation and still requiring large window sizes. Moreover, it is computationally more expensive, and performs badly on uniform regions or small regions at salient points that are dissimilar to neighboring regions [72]. Such regions constitute non-informative descriptors and are eliminated in the beginning of the proposed method which makes the method sparse, i.e., not suitable for dense disparity map calculation.

2.2.1 Similarity Measures for Multi-Modal Stereo-vision

In this section, the similarity measures that were confronted in the literature that are proposed and tested as multi-modal stereo correspondence measures are described in detail. The performances of these similarity measures are evaluated in the scope of this thesis in Chapter 4.

2.2.1.1 Mutual Information (MI)

Mutual information (MI) was invented by Shannon [75] in 1948 in the field of Information Theory. In this study, the concepts of *entropy* as well as the *MI* was introduced. The aim was to propose a measure for counting the information content in a received data over a random variable x having the probability distribution $p(x)$. For events that are highly probable and was certain to happen, this should lead to small values or to a value representing no information, but for events that are less probable and unlikely occur carrying more information should lead to higher values of the measure. Therefore, Shannon first proposed a function $h()$ which is defined as:

$$h(x) = -\log(p(x)), \quad (2.4)$$

which is a monotonic function of $p(x)$.

By this function, it can be observed that if two events x and y are unrelated, then for the joint probability condition $p(x, y) = p(x)p(y)$ of independent variables, the

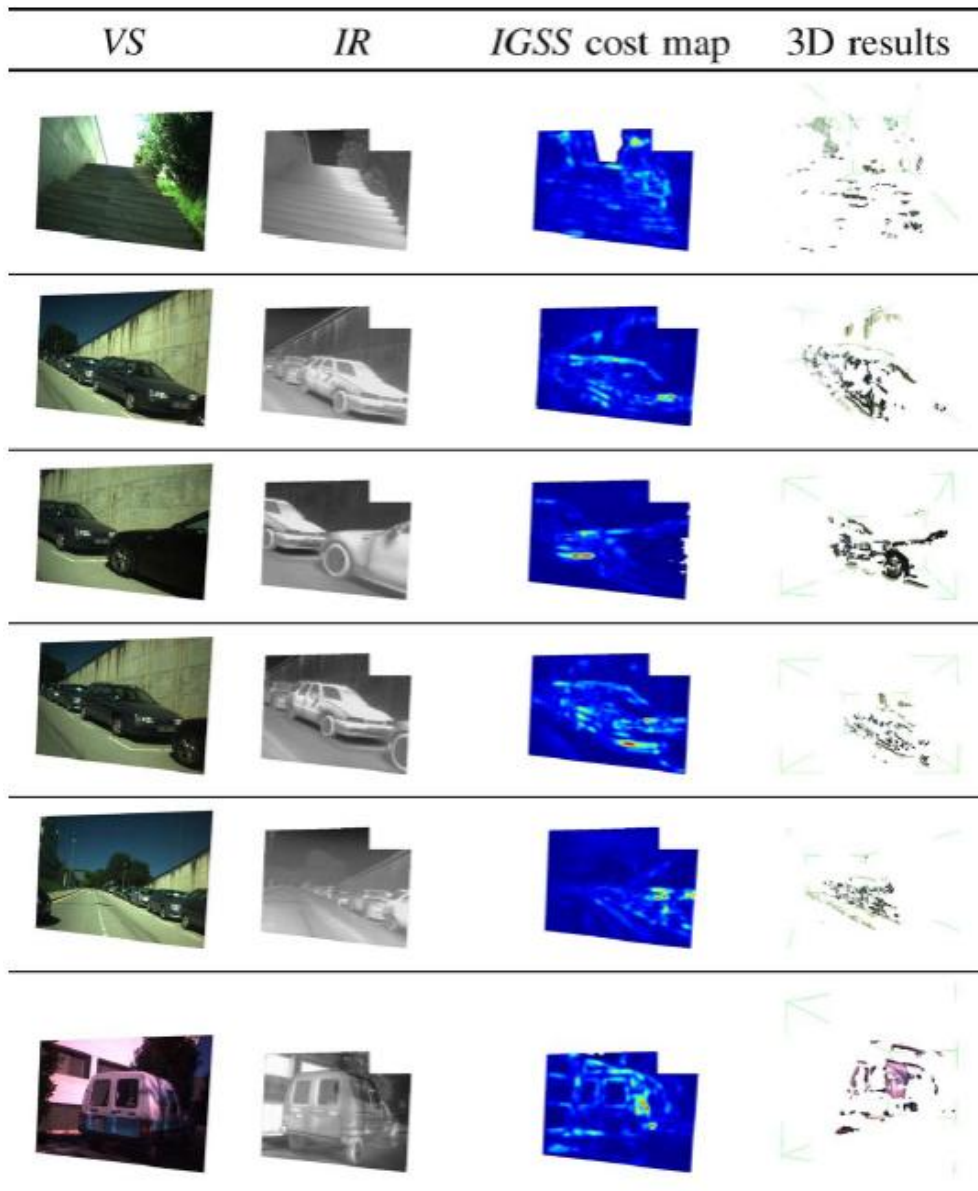
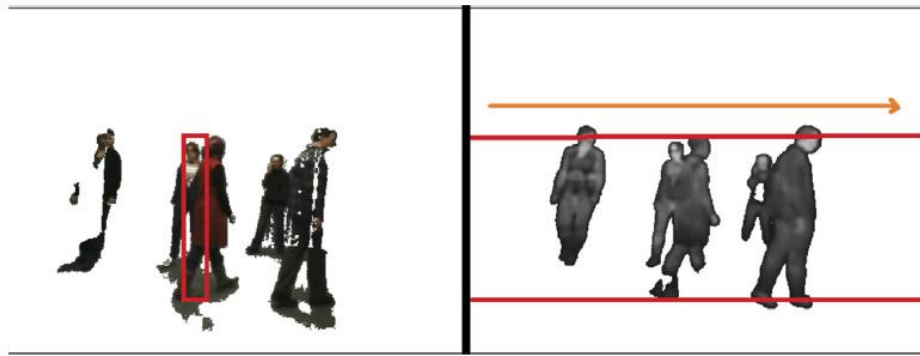


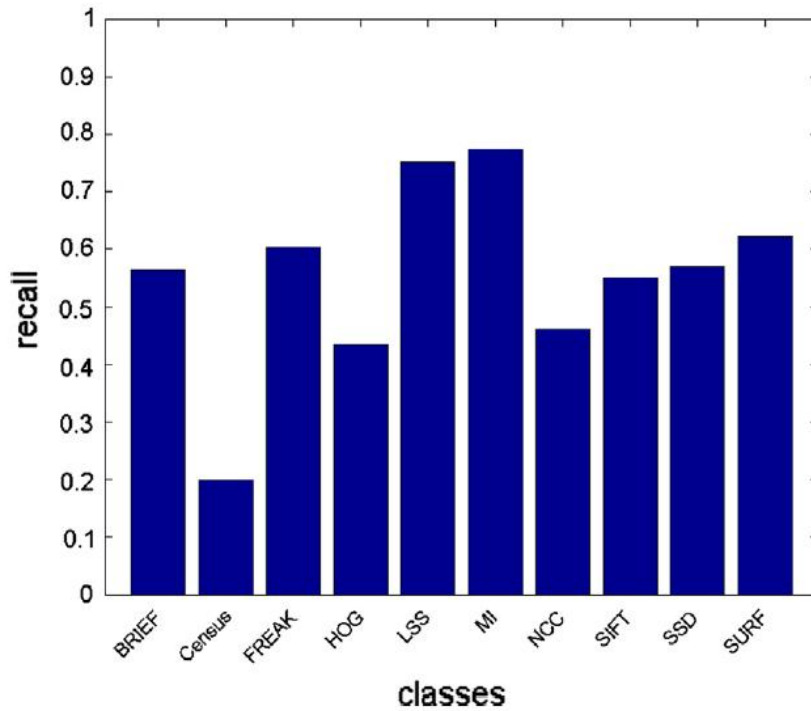
Figure 2.8: Results from [69], showing examples of sparse depth maps from outdoor scenarios (red color in cost map corresponds to high cost values) [Best Viewed in Color]

total information content will be $h(x, y) = h(x) + h(y)$, which is the sum of the two information that are unrelated.

Next, Shannon proposed the concept *entropy* as the average amount of information that can be calculated before transmitting the data of the random variable x wrt dis-



(a)



(b)

Figure 2.9: Sample view from experiments of Torabi *et al.* [74]. (a) Depiction of 1D sliding window algorithm over visible-thermal stereo image pairs for human ROI identification. (b) Results for comparison of evaluated methods for human ROI detection over visible-thermal stereo image pairs.

tribution $p(x)$ as:

$$H(x) = - \sum_x p(x) \log(p(x)). \quad (2.5)$$

Using this concept, Shannon proposed *noiseless coding theorem* which claims that the entropy constitutes the lower bound on the number of bits which is needed to

transmit the state of a random variable [75, 76].

Shannon also defined the *joint entropy* $H(x, y)$ and the *conditional entropy* $H(x|y)$ and showed the relations between them as:

$$H(x, y) = H(p(x, y)) = - \sum \sum (p(x, y) \log(p(x, y))), \quad (2.6)$$

$$H(x|y) = H(p(x|y)) = - \sum \sum (p(x|y) \log(p(x|y))), \quad (2.7)$$

where joint entropy can also be defined as:

$$H(x, y) = H(x) + H(y|x) = H(y) + H(x|y), \quad (2.8)$$

which yields the *mutual information* as follows:

$$MI(x, y) = H(x) - H(y|x) = H(y) - H(x|y) = H(x) + H(y) - H(x, y). \quad (2.9)$$

Joint entropy is defined as the information describing x and y which corresponds to information needed to describe x alone and additional information required to specify y given x , i.e. the conditional entropy. Conditional entropy measures the uncertainty when, for instance, the received signal is known but the actual sent signal is unknown due to noise etc.

Mutual Information (MI) on the other hand, is computed as the subtraction of the information needed to describe x alone and the additional information required to specify y given x , which corresponds to the reduction in the information content of x when y is known. Therefore, if their joint entropies are high (meaning joint probabilities are low) or conditional entropies (conditional probabilities are low) are high then the mutual information shall be low and opposite in the reverse condition. Intuitively, it measures the mutual amount of information they share.

Another definition of MI uses Kullback-Leibler distance measure [77], which measures the additional information required for defining x using another approximating distribution $q(x)$ instead of $p(x)$ which is assumed unknown.

$$KL(p||q) = - \sum_x p(x) \log(q(x)) - \left(- \sum_x p(x) \log(p(x)) \right), \quad (2.10)$$

which is then equal to:

$$KL(p||q) = - \sum_x p(x) \log \left(\frac{q(x)}{p(x)} \right). \quad (2.11)$$

$KL(p||q)$ is also known as *Kullback-Leibler divergence* or *relative entropy* [76]. Hence, when we consider two random variables x and y given $p(x, y)$, if they are independent, then $p(x, y) = p(x)p(y)$. But if they are not, we can get their mutual information from KL divergence as:

$$KL(p(x, y)||p(x)p(y)) = - \sum_y \sum_x p(x, y) \log \left(\frac{p(x)p(y)}{p(x, y)} \right) = MI(x, y) \quad (2.12)$$

From sum and product rules it is obvious to show the equality of Eqn. 2.12 to 2.9.

Regarding computer vision, one can notice that the entropy concept can be applied to an image pixel values where probability distribution of an image $I(x, y)$ defined as $p(I)$ can be computed from an histogram $hist(I(x, y))$ of the pixel intensities. An image with less intensity levels shall have high numbers in the histogram and high probabilities leading to low entropy values. A high entropy value will be computed if image has more intensity levels with less quantities (which means more information content). Therefore, it can be concluded that the entropy measure stands for the unsmoothness or irregularity in an image where images with big regions of small variance in intensity levels shall have low entropy values. It can be stated that entropy measures the dispersion in the probability distribution of the image pixel intensities [1].

Similar conclusions can be deduced for joint probabilities of images. If joint probability of two images are high, this will yield a low joint entropy and vice versa. This idea has led to using entropy and mutual information measures for the image registration purposes. First, Collignon *et al.* [78] and Studholme *et al.* [79] suggested to use entropy and mutual information as a measure of image registration, especially for multi-modal medical images registration. This followed Viola and Well's [65] study on multi-modal image registration using mutual information in a maximization problem which drew significant attention.

Later, effected from Viola's studies Egnal [54] first proposed mutual information for the multi-modal stereo correspondence problem. Later, the studies of [26, 34, 66, 67, 68, 69] are published using mutual information for multi-modal stereo-vision [34, 69, 66] and related problems such as human ROI tracking on multi-modal stereo image pairs [26, 67, 68, 74]. See Chapter 2.2 for detailed literature survey on multi-modal stereo-vision.

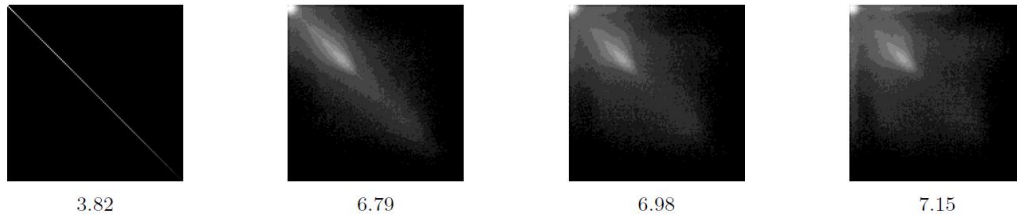


Figure 2.10: Joint histogram of an MR image and its rotated version for 0, 2, 5 and 10 degrees from left to right along with computed joint entropy values at the bottom row. (Source: [1])

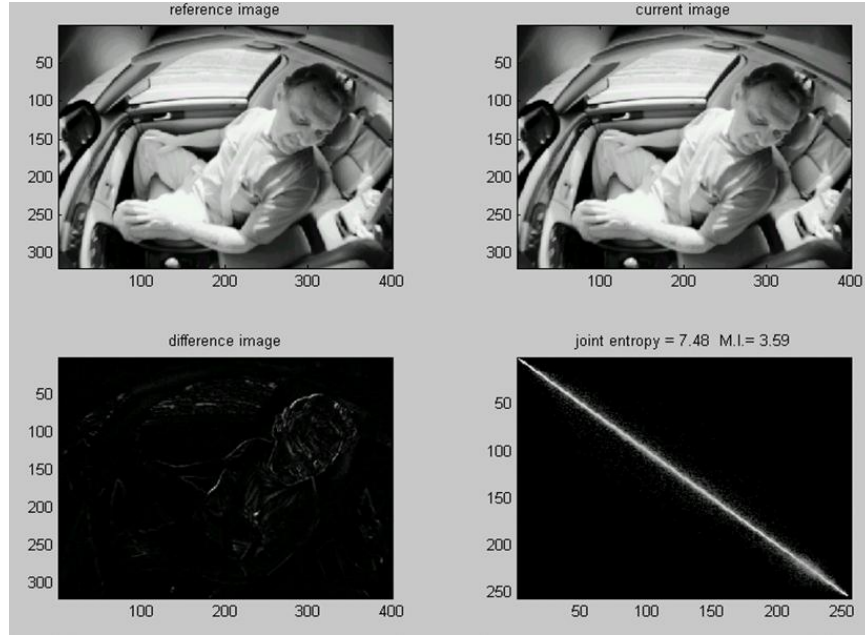
Figure 2.10 shows joint histogram of an image over its unrotated and rotated versions along with the joint entropies calculated. As can be observed, when the dispersion in the joint histogram is low, this means the pixel intensity values in one image corresponds to same or similar intensity values in the other image, which yields low joint entropy values. Similarly, Figure 2.11 presents another experiment showing the change in the joint histogram as well as the joint entropy and MI for the registered and unregistered images.

The advantage of MI over joint entropy is that MI also includes entropies of the two images (hence the marginal probability distributions) in addition to joint entropy. This reduces the misalignment problem for the low entropy locations such as background areas. Having marginal entropies in the calculation, this increases the mutual information when pixels containing structural information of objects in the images also align well [1].

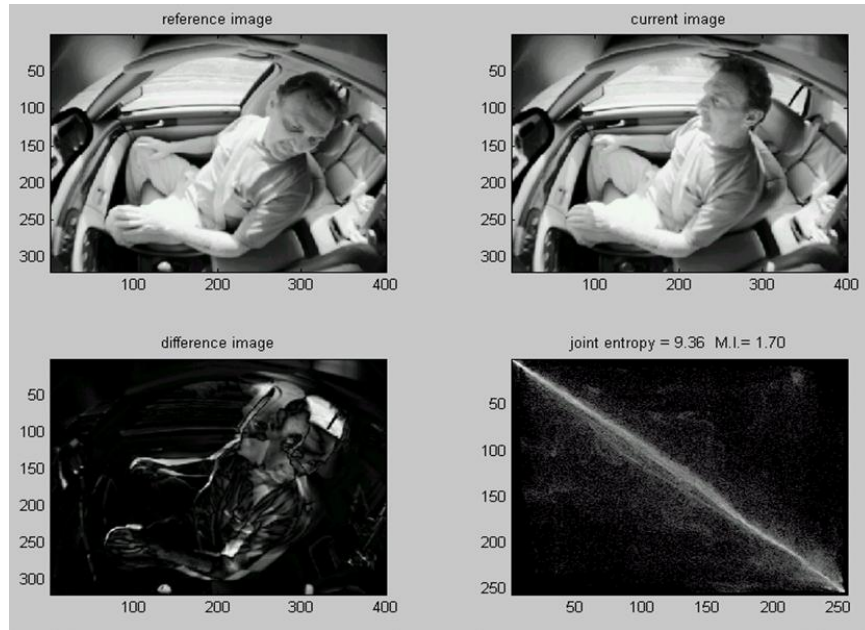
2.2.1.2 Mutual Information with Prior Probabilities (MIwPR)

Incorporating prior probabilities to MI calculation was proposed by Fookes *et al.* [34] for multi-modal image matching and stereo correspondence. The aim is to increase the statistical discriminability of joint probability calculation of the two local matching windows. In order to accomplish this, joint prior probabilities computed from the whole image is added to the local joint probability calculation.

MI can be calculated from the two local windows W_L and W_R extracted from the left



(a)



(b)

Figure 2.11: Depiction of the change in joint histogram, entropy and MI of registered and unregistered image pairs (images at top row: reference and current image; bottom row: the difference image and the joint histogram image) (a) registered image (b) unregistered image (Source: [80])

image L and right image R as:

$$MI(W_L, W_R) = \sum_{I_l \in W_L} \sum_{I_r \in W_R} P(I_l, I_r) \log \frac{P(I_l, I_r)}{P(I_l)P(I_r)}, \quad (2.13)$$

where $P(I_l, I_r)$ corresponds to joint probability for the left and right image patches W_L and W_R , $P(I_l)$ and $P(I_r)$ are the marginal probabilities of the pixel intensities.

On the other hand, it is possible to change the formulation as:

$$MI_{(wPR)}(W_L, W_R) = \sum_{I_l \in W_L} \sum_{I_r \in W_R} P(I_l, I_r) \log \frac{P^*(I_l, I_r)}{P(I_l)P(I_r)}, \quad (2.14)$$

where $P^*(I_l, I_r)$ can be defined as:

$$P^*(I_l, I_r) = \lambda P(I_l, I_r) + (1 - \lambda) P_{prior}(I_l, I_r), \quad (2.15)$$

and $P_{prior}(I_l, I_r)$ is the joint prior probability computed from the joint histogram of the whole images as:

$$P_{prior}(I_l, I_r) = \frac{hist(I_l, I_r)}{\sum_{l,r} hist(I_l, I_r)}, \quad (2.16)$$

for all corresponding pixels I_l in left image L and I_r in right image R .

The λ corresponds to the degree of this incorporation of prior probabilities into the joint probability calculation. This modification to MI calculation was shown to increase the performance of MI as a similarity measure in stereo correspondence problem in [34].

2.2.1.3 Local Self-Similarity (LSS)

Local self similarity (LSS) is a similarity measure proposed initially by Shechtman and Irani [70] for image template matching.

The method simply extracts a small patch (e.g. 5x5) from the center pixel q of a larger window (e.g. 40x40) and sum of squared distances (SSD) between the small patch and the surrounding larger region is computed. Next, the SSD costs are normalized by maximum value of small image patch variance and a noise term generating a *correlation surface* as given in below equation:

$$Sq(x, y) = \exp \left(- \frac{SSD_q(x, y)}{\max(var_{noise}, var_{auto}(q))} \right). \quad (2.17)$$

Finally, the LSS descriptor is produced by a partitioned log-polar representation of this correlation surface like e.g. 20 angles and 4 radial intervals = 80 bins. Figure 2.12

taken from [70] shows some spectacular results of using LSS including a comparison with MI.

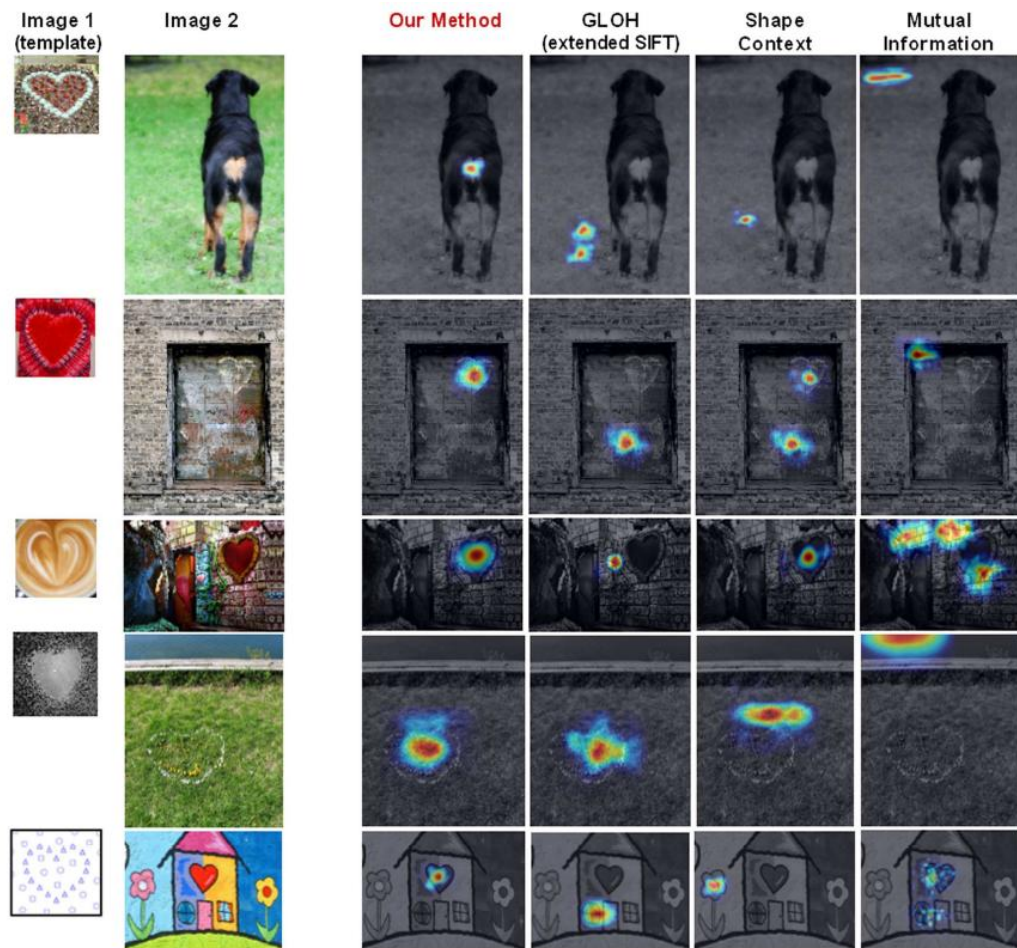


Figure 2.12: Reported template matching results from [70] using LSS, including a comparison with MI.

Torabi and Bilodeau introduced LSS measure for thermal-visible image pairs stereo correspondence problem where they implemented an ROI based image matching framework for human tracking and computing depth of the human ROI in thermal-visible image pairs [71, 73]. In their recent studies, they compared this measure with other similarity measures like MI, HOG, Census, SIFT, SURF, BRIEF, FREAK etc. [72, 73] where LSS was successful for smallest window size that were tested but MI was still outperforming LSS for the other two window sizes.

2.2.1.4 Histogram of Oriented Gradients (HOG)

Histogram of Oriented Gradients (HOG) similarity measure was proposed by Dalal and Triggs [81] for human detection.

The method first divides the image patch extracted around the center pixel q into a grid of *cells*. Each cell accumulates the local 1-D histogram of gradient directions. For robustness to illumination effects, *blocks* composed of several of these cells accumulate the local histograms and they are used for normalization of local cells. The normalized descriptor blocks are called Histogram of Oriented Gradients (HOG). The HOG descriptors of each pixel in a local detection window can then be combined to have a feature vector for human detection. Computation of the descriptors are expressed mathematically as below:

$$HOG_q(k) = \sum_{(x,y) \in W_q} T\left(\frac{\Theta(x,y)}{\gamma}\right), \quad (2.18)$$

where $HOG_q(k)$ corresponds to k^{th} bin in the histogram of K bins, $\Theta(x,y)$ is the gradient at pixel (x,y) and γ is a scaling constant and T is the function defined as:

$$T(u) = \begin{cases} 1 & \text{if } u = 0 \\ 0 & \text{otherwise.} \end{cases} \quad (2.19)$$

Figure 2.13 summarizes the whole process of computing the HOG descriptors.

HOG was used by Torabi and Bilodeau for comparing their proposed method of human ROI detection in thermal-visible stereo image pairs using LSS [72] and also compare to other alternative similarity measures like MI, SIFT, SURF, Census etc. in [74].

2.2.1.5 Scale Invariant Feature Transform (SIFT)

Scale Invariant Feature Transform (SIFT) was proposed by David Lowe [46] for object detection. The method generates sets of descriptive features from images that are claimed to be invariant to rotation and scaling and partially to camera viewpoint and illumination changes. The method has drawn significant attention since was proposed

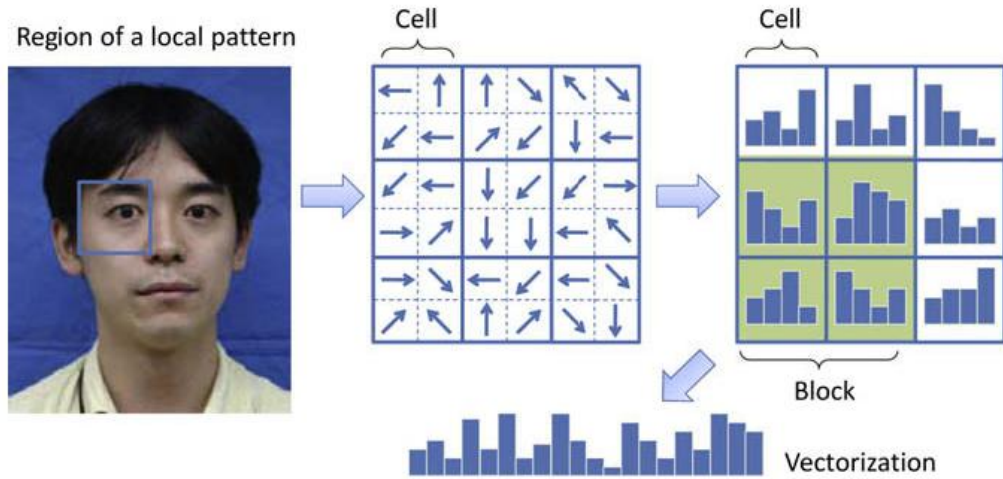


Figure 2.13: Depiction of computing HOG descriptors process (Source: [82])

at 2004 and become very popular for object detection, template matching and related applications.

The method is composed of below major steps:

A. *Scale-space extrema detection:*

In this step, the aim is to identify locations which are invariant to scale and orientation changes. To accomplish this task, the image is transformed into several scales where at each scale level (called an *octave*) the image is convolved consecutively with Gaussians. Mathematically, for the image $I(x, y)$, the scale space of the image is defined as a function $L(x, y, \sigma)$ as:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \quad (2.20)$$

where $*$ represents the convolution operation and G is the Gaussian function with the standard deviation σ .

If the image consecutively convolved with Gaussians of incremented scales as $k\sigma$ and the resultant images are subtracted, then *Difference of Gaussians (DoG)* $D(x, y, \sigma)$ images are generated which enables to efficiently detect stable keypoints.

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma). \quad (2.21)$$

In Figure 2.14, the generation of DoG images in the scale space is depicted. The method uses s intervals in each octave, where $k = 2^{1/s}$. At each octave, the Gaussian images are resampled by halving the resolution of the image.

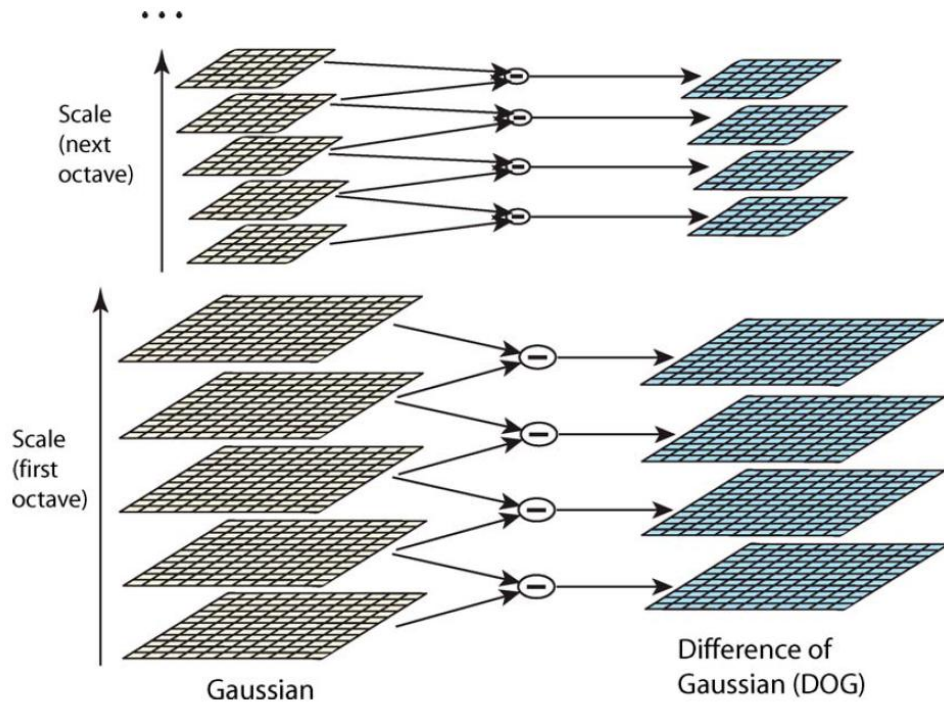


Figure 2.14: Depiction of generation of DoG images for the scale space images in each octave (Source: [46])

Finally, the local extrema are detected in the scale space DoG images generated. The method compares each sample point in the DoG image by its eight neighbors in the current scale and nine neighbors in the corresponding images of scale levels higher and lower by one level. The sample point is selected if it is greater than or lower than all these neighbor pixels.

B. Accurate Keypoint Localization:

In this step, the keypoint candidates are accurately localized by fitting a 3D quadratic function to the neighboring data. This method was proposed by Brown and Lowe in [83]. The Taylor expansion of the D function (Eqn. 2.21) is used for this purpose as:

$$D(x) = D + \frac{\partial D^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 D}{\partial x^2} x, \quad (2.22)$$

where the local extrema, \hat{x} are accurately localized by taking the derivative of this

function and setting to zero:

$$\hat{x} = -\frac{\partial^2 D}{\partial x^2} \frac{\partial D}{\partial x}. \quad (2.23)$$

The vectors comprised by the location, scale and orientation are generated for each keypoint. Next, the keypoints having low contrast with the neighbor data are eliminated at this step by computing the D value at the extremum, $D(\hat{x})$, as below:

$$D(\hat{x}) = D + \frac{1}{2} \frac{\partial D^T}{\partial x} \hat{x}. \quad (2.24)$$

However, this is not sufficient since edge responses should also be eliminated which have high DoG values although weakly determined. These locations are detected by computing principal curvatures from a 2x2 Hessian Matrix, H and computing the ratio and thresholding as:

$$\frac{Tr(H)^2}{Det(H)} < \frac{(r+1)^2}{r}, \quad (2.25)$$

where $Tr(H)$ is the trace of the H , $Det(H)$ is the determinant and the r is the ratio between the largest eigenvalue and smaller one.

C. *Assignment of Orientation:*

The orientations of the keypoints are also assigned by forming a 36-bin histogram which was determined experimentally. The histogram entries are weighted by the gradient magnitude and by a Gaussian weighted circular window. This Gaussian window has a σ that is 1.5 times of the keypoint's own scale.

Later, the local peaks in this orientation histogram are also inspected. Any local peak greater than 80% of the highest peak is added as another keypoint with this new orientation.

Finally, a parabola is fit to the three values closest to the peaks in the orientation histogram to detect an accurate position for the orientation values assigned.

D. *Generating the Local Descriptor:*

In this final stage, the local descriptors at the salient points detected by the method are generated.

The process starts by computing image gradients and orientations around the local region of each keypoint. Next, the gradients and orientations are weighted by a

Gaussian function with the σ as the half of the descriptor window determining the 2×2 or 4×4 subregions around the keypoint. The gradients and orientations are accumulated within these descriptor windows where the orientations are binned to 8 directions within each subregion using the gradient values. Figure 2.15 shows this procedure where a 2×2 subregions of 8-bin orientation histograms are depicted for illustrative purposes. In the original article examples, [46], 4×4 subregions are used yielding a total of $4 \times 4 \times 8 = 128$ entries for the descriptor vector of the keypoints.

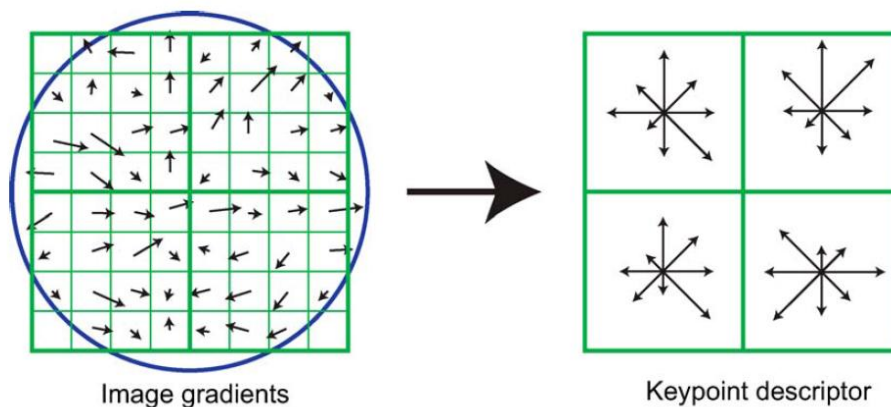


Figure 2.15: Depiction of computing the SIFT local descriptor: *Left*: shows the computed gradients and orientations for each sample point in the neighbor region and the Gaussian window for weighting *Right*: 2×2 descriptor windows (the subregions) to accumulate gradients and construct 8-bin orientation histogram (Source: [46])

Finally, to have a more robust descriptor to illumination changes, the descriptor vectors are normalized and some large values greater than 0.2 (determined experimentally) are removed and then renormalized again. This way the over-effect of large magnitude gradients are reduced due to non-linear illumination changes.

2.2.1.6 Speeded-Up Robust Features (SURF)

The Speeded-Up Robust Features (SURF) was proposed by Bay *et al.* [47]. The aim was to develop a faster similarity measure in contrast to SIFT with comparable performance.

The method is based on the Hessian matrix defined as:

$$H(x, y, \sigma) = \begin{bmatrix} L_{xx}(x, y, \sigma) & L_{xy}(x, y, \sigma) \\ L_{xy}(x, y, \sigma) & L_{yy}(x, y, \sigma) \end{bmatrix}, \quad (2.26)$$

where $L(x, y, \sigma)$ is the convolution of the Gaussian to image I same as in Eqn. 2.20, L_{xx} , L_{xy} and L_{yy} are the second order derivatives. For computing the Gaussian second order derivatives for this purpose, 9x9 approximations (the *box filters*) are used which increases the speed of the total method significantly.

The scale space analysis on the other hand is performed by the method by up-sampling the box filters instead of down-sampling the image. The box filters of size 9x9, 15x15, 21x21 and 27x27 etc. are used. As scales are enlarged, the step sizes between scales are also scaled. Next, the interest points are localized by using non-maximum suppression in a 3x3x3 neighborhood over the scales of the image. Similar to SIFT, Brown and Lowe's method in [83] is used for interpolation in scale space over the Hessian matrix for the maxima of the determinant.

For achieving rotational invariance, Haar-wavelet responses are computed in x and y direction in a circular neighborhood of the interest points with radius $6s$ where s is the scale of the interest point detected. The responses are weighted by the Gaussian of scale $\sigma = 2.5s$ centered on the interest point. The dominant orientation is then computed by the sum of the responses using a sliding orientation window having an angle of $\frac{\pi}{3}$ determined experimentally.

Finally, for generating the descriptor vector, a square region around the interest point oriented in the dominant orientation is constructed. The size of this window is set to $20s$. Next, this window is split into 4x4 subregions. The subregions are sampled by 5x5 sample points and the wavelet responses d_x , d_y are summed up over each subregion. The absolute values of the responses are also summed, $|d_x|$ and $|d_y|$. Therefore, the feature vector is comprised by $v = (\sum d_x, \sum |d_x|, \sum d_y, \sum |d_y|)$ for each subregion which yields totally $4x4x4 = 64$ sized descriptor vector.

The descriptor vector is normalized for invariance to contrast changes.

Figure 2.16 shows sample descriptor vector responses for a subregion of given patterns.

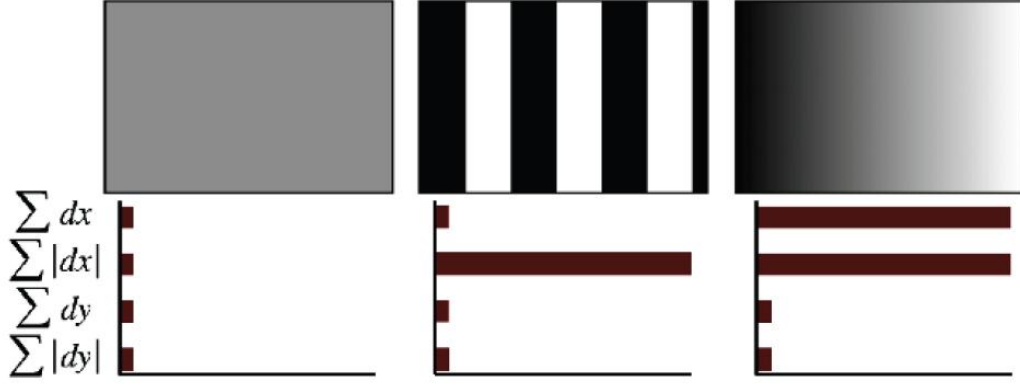


Figure 2.16: Depiction of SURF descriptor vector responses of the subregions of several patterns (Source: [47]) *Left*: homogeneous region corresponds to low values *Middle*: high frequencies in x direction corresponds to high values in $|d_x|$ *Right*: increasing intensity in x direction corresponds to high values for both d_x and $|d_x|$

2.2.1.7 Census Transform (CENSUS)

Census Transform was proposed by Zabih *et al.* [84] which is composed of a non-parametric local transform. The method simply uses relative ordering of pixel intensities in the image patch over the center pixel and transforms it into a binary encoded string as the descriptor information. Mathematically, this is expressed as:

$$C(W_u) = \otimes T(W_u; u, v), \quad (2.27)$$

where W_u is the image patch extracted from center pixel u , v denotes all neighbor pixels in W_u , \otimes is the concatenation operation and the T function is defined as:

$$T(W_u; u, v) = \begin{cases} 0 & \text{if } W_u(v) < W_u(u) \\ 1 & \text{otherwise.} \end{cases} \quad (2.28)$$

Therefore, the neighbor pixels are checked if their value is greater than the center pixel which leads to a 0 in the binary string and 1 otherwise if smaller.

The descriptor is composed of the binary string encoded using this T function as shown in Eqn. 2.27 and 2.28.

The *hamming* distance [85] is used to compare the encoded binary strings, that counts

the number of bits that differ:

$$hamming(s_1, s_2) = \sum_{i \in (0, N)} T_h(s_1(i), s_2(i)), \quad (2.29)$$

where T_h function is defined as:

$$T_h(c_1, c_2) = \begin{cases} 1 & \text{if } c_1 \neq c_2 \\ 0 & \text{otherwise} \end{cases} \quad (2.30)$$

and where s_1 and s_2 are the two binary strings to be compared, $s(i)$ corresponds to the i th bit in the string s .

Census was used as one of the similarity measures to compare to alternative measures in their study of Torabi and Bilodeau [74] for human ROI detection in thermal-visible stereo image pairs.

2.2.1.8 MI of Census Transform (CENSUSMI)

This similarity measure is a novel method proposed in the scope of this thesis [38].

The idea is to use mutual information as the similarity function instead of hamming distance of the Census Transformed image patches which is claimed to eliminate the issues in multi-modal image pairs having different intensity responses. Mathematically, it can be expressed as:

$$CENSUSMI(W_L, W_R) = MI(C(W_L), C(W_R)), \quad (2.31)$$

where $C(W_L), C(W_R)$ corresponds to census transform of the left and right image local windows W_L, W_R to be matched, as given in Eqn. 2.27.

Figure 2.17 illustrates the method where instead of the two initial multi-modal image pairs having different gray level intensities the census transformed image patches are sought for matching using the mutual information of the transformed image patches.

2.2.1.9 Binary Robust Independent Elementary Features (BRIEF)

Binary Robust Independent Elementary Features (BRIEF) is a similarity measure proposed by Calonder *et al.* [86] based on features as encoded binary strings over an

Multi-Modal Left – Right Stereo Image Pairs

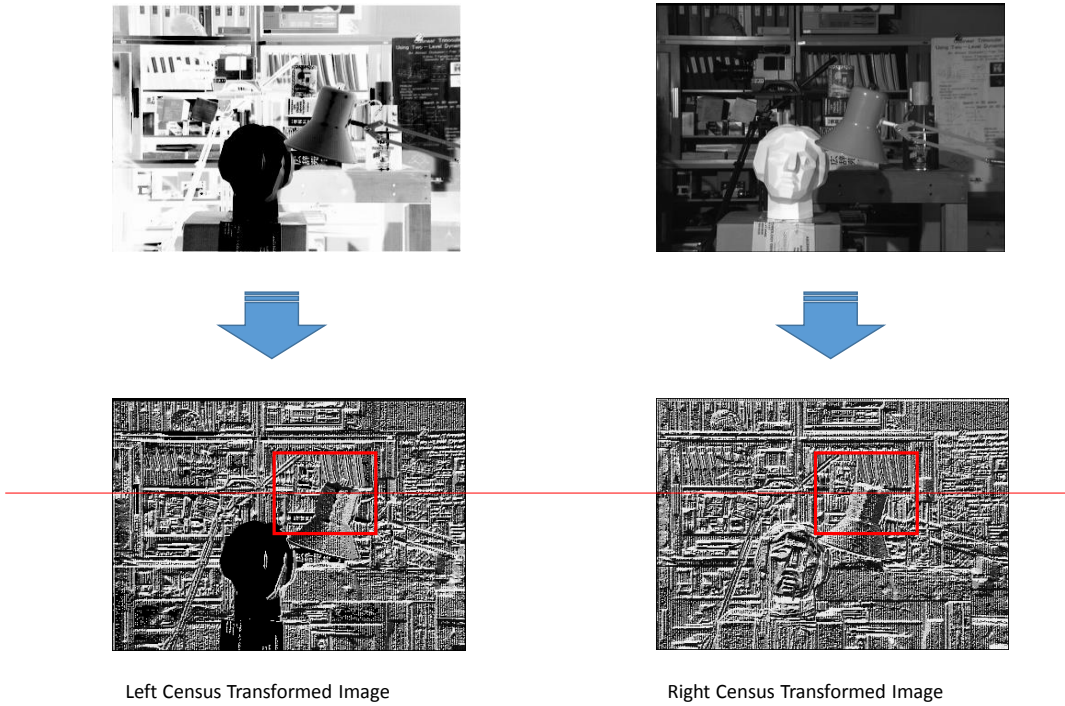


Figure 2.17: Illustration of the MI of Census Transform method where multi-modal image pairs are census transformed and processed for stereo correspondence using the mutual information of the transformed images

image patch.

The method is composed of three steps:

- A Sampling grid of points in a defined pattern is generated around the region of the pixel interested.
- List of sampling pairs of points from the sampling grid is computed.
- The binary string is encoded from the sampling pairs.

The binary string is encoded using the T function as in Eqn. 2.32, for an image patch p of size $S \times S$:

$$T(p; x, y) = \begin{cases} 1 & \text{if } p(x) < p(y) \\ 0 & \text{otherwise} \end{cases} \quad (2.32)$$

The pixel intensities of the image patch are smoothed using Gaussian kernels in order not to effect from noise.

Regarding the construction of the sampling grid and computing the sampling pairs (x, y) , a number of methods were proposed in [86] as:

- **(G I)**: points are evenly distributed, pairs are randomly selected.
- **(G II)**: points are sampled using a Gaussian distribution and pairs are randomly selected from this distribution of points, which means points near the center are preferred.
- **(G III)**: the first location x is sampled from a Gaussian centered around the origin, the other point is sampled from another Gaussian centered around the x , which creates more local pairs.
- **(G IV)**: A coarse polar grid is used and pairs are randomly selected from this grid.
- **(G V)**: A coarse polar grid is used and pairs are selected as $x = (0, 0)$ at the origin and y is randomly selected.

Figure 2.18 illustrates these five approaches proposed for the construction of the sampling grid and the sampling pairs.

The *hamming* distance [85] is used to compare the encoded binary strings as given in Eqn. 2.29.

BRIEF was used as one of the similarity measures to compare to alternative measures in their study of Torabi and Bilodeau [74] for human ROI detection in thermal-visible stereo image pairs.

2.2.1.10 Fast Retina Keypoint (FREAK)

Fast Retina Keypoint (FREAK) is yet another similarity measure proposed recently by Alahi *et al.* [87] based on features as encoded binary strings over a neighborhood of a pixel. Like other competitiveness (such as BRIEF, BRISK etc.), the binary strings

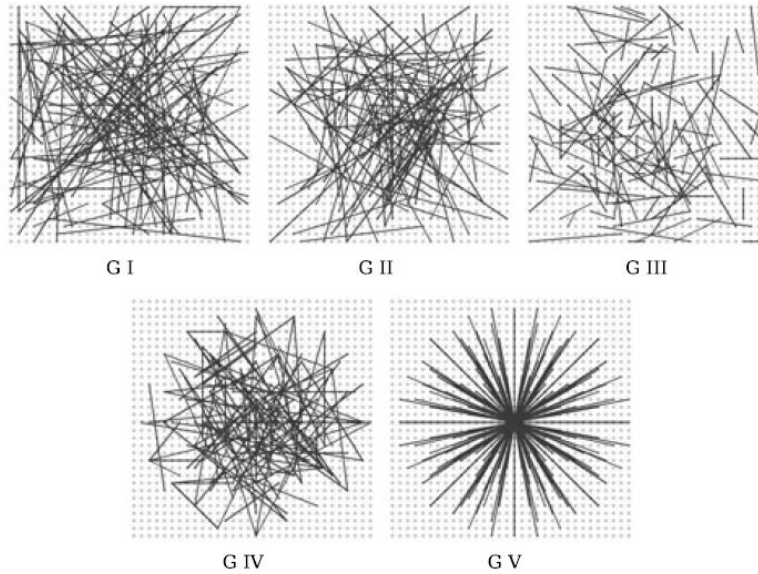


Figure 2.18: Illustration of the five approaches proposed for the construction of the sampling grid and the sampling pairs for encoding the BRIEF similarity measure (Source: [86]).

are generated over a sampling grid, but the difference of FREAK is that the method was inspired by the human visual system and the topology and spatial relationship of receptor cells in the retina. The sampling grid was proposed to have a similar pattern to retinal system.

Briefly, the retinal structure is composed of layers of the photo-receptors (rods and cones), the inner cells (horizontal, bipolar and amacrine cells) and the ganglion cells [88] (see Figure 2.19-a). The layers transfer visual information between them where finally the ganglion cells actually encode the visual information as action potentials and transfers to the several parts of the brain. The number of ganglion cells decrease exponentially by the distance from the foveal (where cones are densely packed but no rods). In the foveal region, it becomes almost 1:1 to have a ganglion cell to each cone. Therefore, highest resolution of information acquired from the foveal whereas lowest resolution is acquired from the perifoveal where many photo-receptors influence less ganglion cells.

Influenced by this structure, it is proposed to have a similar structure and transformation of image pixels to meaningful binary patterns as depicted in Figure 2.19-b.

Accordingly, the retinal sampling grid is proposed to be circular but more number of points are near the center of the grid. The density of the sample points are decreased exponentially by the distance to the center. For each sampling point, the associated Gaussian kernel size is also changed comprising bigger receptive fields on the periphery (i.e. lower resolution). Besides, the receptive fields do overlap which experimentally found to increase performance.(see Figure 2.19-c)

Using this sampling grid of receptive fields, a binary descriptor is constructed by first pairing the receptive fields and then thresholding the difference between the receptive fields, as below, as a sequence of one-bit Difference of Gaussians (DoG):

$$FREAK = \sum_{0 \leq a < N} 2^a T(P_a), \quad (2.33)$$

where T function is defined as:

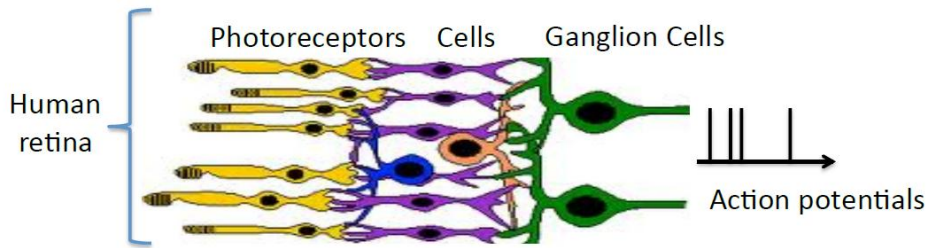
$$T(P_a) = \begin{cases} 1 & \text{if } I(P_a^{r_1}) - I(P_a^{r_2}) > 0 \\ 0 & \text{otherwise,} \end{cases} \quad (2.34)$$

and P_a is the pair of receptive fields, N is the size of the descriptor binary string.

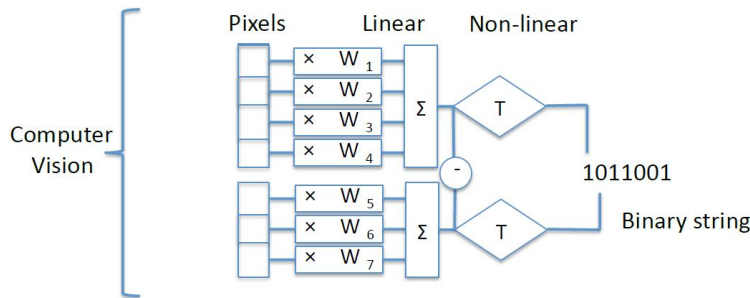
In order to select the pairings a learning phase is performed over the data which yields automatically a coarse-to-fine ordering of DoGs similar to human vision system. Later at the recognition step, it is preferred to search the binary strings in several levels where initially first 16 bits are compared which represents the coarser information and after locating the candidate matchings a more detailed search is performed over the rest of the string for a higher resolution matching. This mechanism is also claimed to mimic how the human visual recognition is performed [87].

The *hamming* distance [85] is used to compare the encoded binary strings as given in Eqn. 2.29, like other binary encoded string based similarity measures like Census and BRIEF.

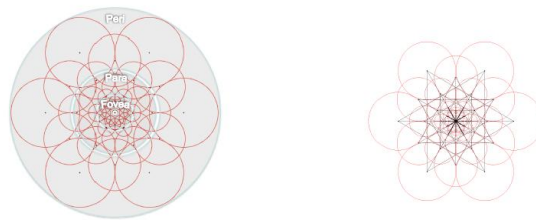
FREAK was used as one of the similarity measures to compare to alternative measures in their study of Torabi and Bilodeau [74] for human ROI detection in thermal-visible stereo image pairs.



(a)



(b)



(c)

Figure 2.19: Human vision system and depiction of proposed computer vision system (Source: [87]) (a) Depiction the human visual system along with the layers transferring visual information from photo-receptors to ganglion cells which encodes and transfers data to brain (courtesy of [88]) (b) Proposed computer vision system structure from pixels to encoded binary strings for object recognition. (c) Depiction of retinal sampling pattern of receptive fields (*left image*) and sample pairings of the receptive fields (*right image*).

2.2.1.11 Normalized Cross Correlation (NCC)

Normalized Cross Correlation (NCC) is also another traditional similarity measure [39, 42, 89]. The pixel-wise cross-correlation of the two matching windows are computed and normalized by the overall intensity difference. It is expected to have higher similarity value when similar patterns of intensities exist during the matching.

NCC similarity measure is defined as:

$$NCC(W_L, W_R, d) = \frac{\sum_{x,y} (W_L(x, y) - \bar{W}_L)(W_R(x - d, y) - \bar{W}_R)}{\sqrt{\sum_{x,y} (W_L(x, y) - \bar{W}_L)^2 (W_R(x - d, y) - \bar{W}_R)^2}}, \quad (2.35)$$

where W_L is a local window around the pixel (x, y) in the left image L corresponding to the pixel in the x th column and y th row of this left image. W_R is either of the matching windows from the same row y in the right image for the candidate disparities $d \in [0, d_{max}]$ which are tested for maximum similarity (or namely minimum cost in this case). \bar{W}_L, \bar{W}_R are the mean pixel intensities of the local windows W_L and W_R respectively.

NCC was used as the basic similarity measure to compare the proposed methods for multi-modal stereo-vision in studies: [54, 74, 72].

2.2.1.12 Sum of Square Distances - SSD

Sum of Square Distances (SSD) is the most basic similarity measure for stereo-vision [39, 42, 89], which is simply composed of computing the sum of squares of intensity differences of the pixels in the two matching windows W_L and W_R from the left L and right R images respectively.

The similarity measure is defined as:

$$SSD(W_L, W_R, d) = \sum_{x,y} (W_L(x, y) - W_R(x - d, y))^2, \quad (2.36)$$

where W_L is a local window around the pixel (x, y) in the left image L corresponding to the pixel in the x th column and y th row of this left image. W_R is either of the matching windows from the same row y in the right image for the candidate disparities $d \in [0, d_{max}]$ which are tested for maximum similarity (or namely minimum cost in this case).

CHAPTER 3

DATASETS AND PERFORMANCE EVALUATION

In this chapter, the multi-modal stereo image datasets that were generated and prepared in the scope of the thesis are described. Besides, the methods that were proposed and used for the performance evaluation over each dataset are provided.

Two types of datasets are generated. The synthetically altered stereo image pairs from the Middlebury Stereo Evaluation Dataset [45] and the visible-infrared image pairs captured from a Kinect device [35].

3.1 Dataset #1 - The Middlebury Dataset

This dataset contains the four *popular* image pairs (Tsukuba, Venus, Cones and Teddy) in the Middlebury Stereo Evaluation Dataset [45], where the left images are replaced with the synthetically altered ones by using a cosine transform ($\cos(\pi I/255)$) of pixel intensities just as Fookes did [34]. Table 3.1 provides the list of the image pairs that comprises the dataset along with several properties of the stereo images. Figure 3.1 presents the image pairs generated and used in the experiments.

This dataset enabled to compute the statistics of test results for gaining more knowledge of the performance of the evaluated methods and it became possible for any current or future method to be able to be compared for the result metrics regarding the ones on the evaluation site although they are results for the unimodal image pairs.

Note that, in the left images, important details are lost due to the cosine transformation.

Table 3.1: The Dataset #1 - Synthetically Altered Middlebury Stereo Evaluation Dataset.

Dataset	Image No	Image Name	Resolution	Max. Disparity
Dataset #1	1	Tsukuba	384×288	15
Dataset #1	2	Venus	434×383	19
Dataset #1	3	Teddy	450×375	59
Dataset #1	4	Cones	450×375	59

In the experiments, for the statistics computation, the prepared non-occluded regions and discontinuity regions are used "as is" as provided by the Middlebury page [45]. Regarding the "all" regions, we performed clipping on the left border for the region that do not exist in the right image since it is not in the scope of the thesis to perform any extrapolation for those regions. Besides, the image borders are also excluded by 32 pixels because of the limitations of the similarity measures evaluated. In addition, for the window-based methods half of the used window sizes at the borders are also discarded when computing performance statistics for a fair comparison between methods.

Refer to Figure 3.2 for the "all", "disc" and "nonocc" regions in accordance with the the Middlebury page [45], where the white pixels show the regions that the performance evaluations are performed.

The performance evaluations for dataset#1 are performed using two types of metrics: the *RMS* - Root Mean Square distances between computed disparities and the ground truth disparity map and the *Bad* - percentage of bad pixels at which the distance between computed disparity and the ground truth disparity is greater than the designated threshold. This error threshold δ_d for the *Bad* pixels metric is set to 1.5 disparity distance as was performed in Middlebury Stereo Vision Evaluation Page and suggested in the description page [90] for non-integer subpixel disparities unless rounded.

These metrics are computed as follows [42]:

$$RMS = \sqrt{\frac{1}{N} \sum_{(x,y)} |d_C(x,y) - d_T(x,y)|^2}, \quad (3.1)$$

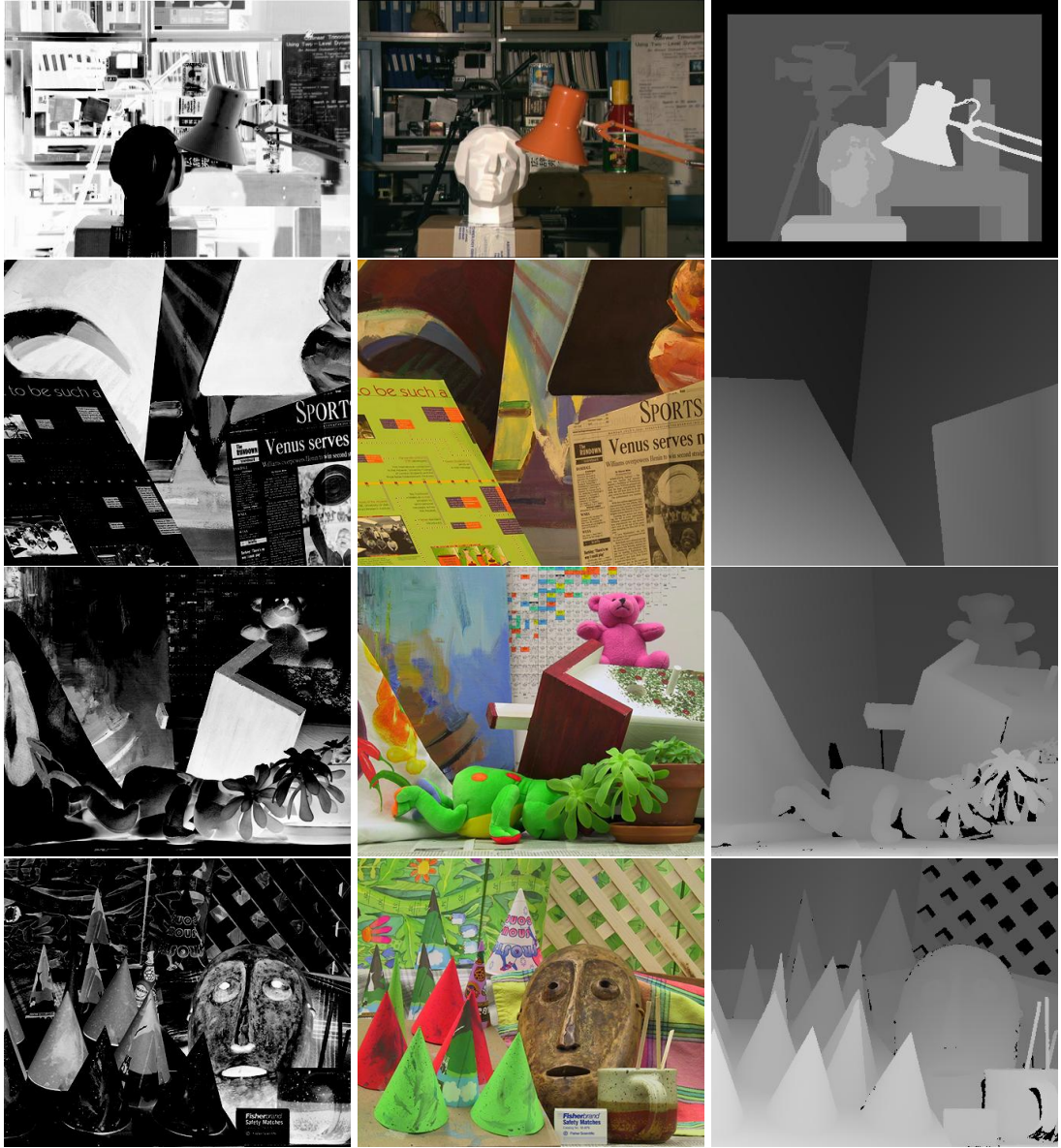


Figure 3.1: Tsukuba, Venus, Teddy and Cones stereo pairs from the Middlebury Stereo Vision Page - Evaluation Version 2 [45]. *Left column:* Synthetically altered left images. *Middle column:* The right images. *Right column:* The ground truth disparities. Note that, in the left image, important details are lost due to the cosine transformation.

$$Bad = \frac{1}{N} \sum_{(x,y)} (|d_C(x,y) - d_T(x,y)| > \delta_d). \quad (3.2)$$

where $d_C(x,y)$ is the computed disparity map, $d_T(x,y)$ is the the ground truth disparity map, δ_d is the error threshold ($=1.5$), N is the number of pixels accounted for metrics computation.

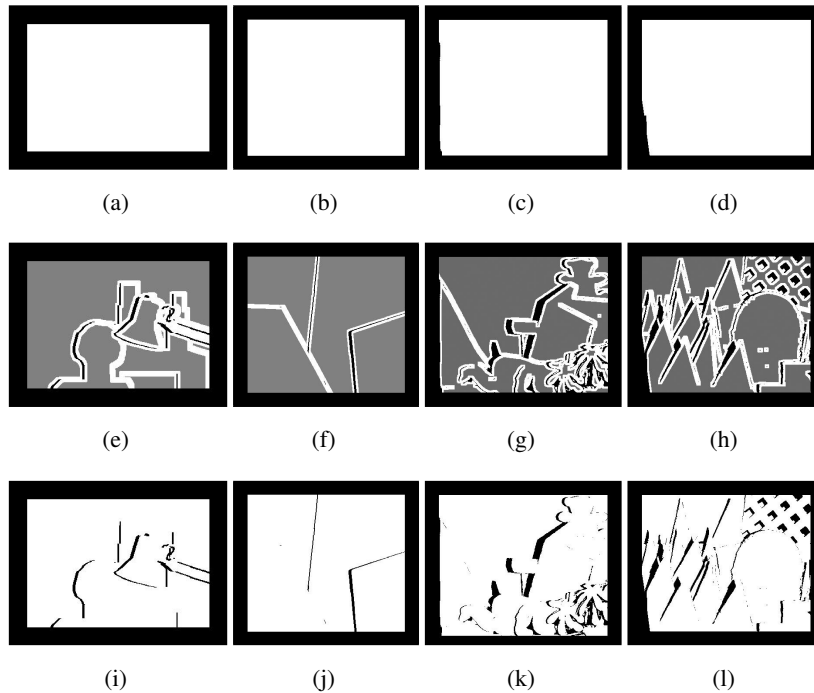


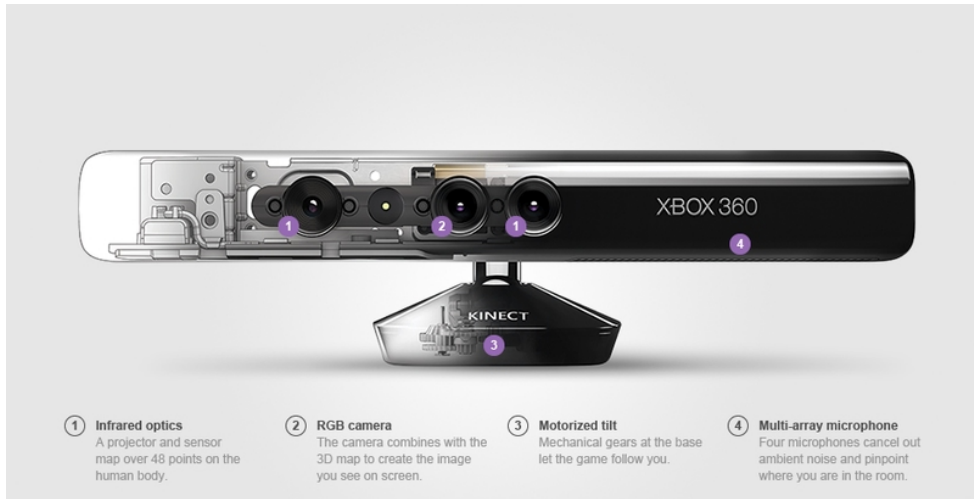
Figure 3.2: Regions where evaluations are performed, for Tsukuba, Venus, Teddy and Cones image pairs. Only "White" pixels are included in performance evaluation calculations. (a)-(d) the "all" regions including regions of both non-occluded discontinuities (c)-(h) the "disc" regions - discontinuities (i)-(l) the "nonocc" regions - non-occluded regions.

3.2 Dataset #2 - The Kinect Dataset

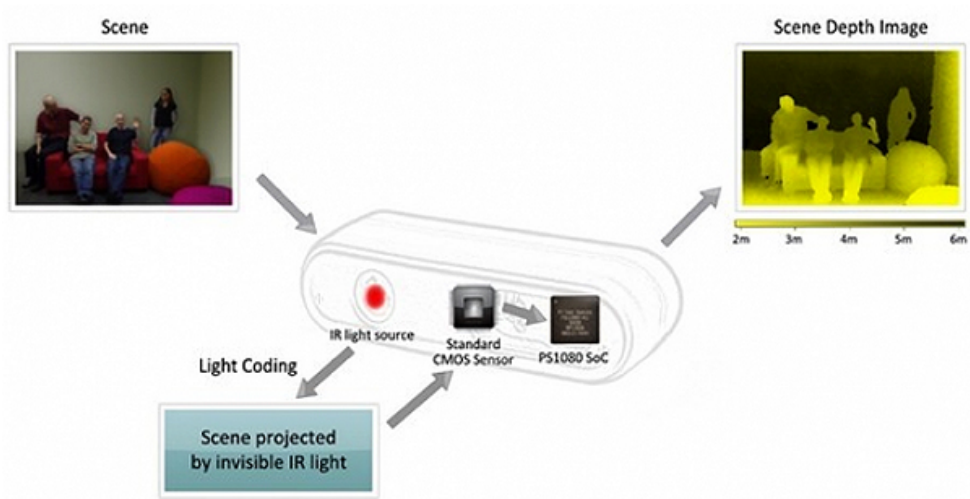
The Kinect dataset contains infrared (left) and visible (right) images captured from a Kinect device. The Kinect Device was introduced by Microsoft for the Xbox 360 game console [36] which enabled the user use his/her own body as the game controller.

As shown in Figure 3.3, the device has a built-in RGB camera and an infrared camera and projector couple. The infrared projector sends beams to the scene and the beams are sensed on the infrared camera which enables the device to generate a 3D depth map of the scene where the intention is the human body.

To be able to use the built-in infrared and visible camera images for multi-modal stereo-vision, it is needed to perform stereo rectification on the two cameras so that epipolar constraint (refer to 2) is satisfied. This is accomplished by using RGBDemo



(a)



(b)

Figure 3.3: The Kinect Device : (a) Kinect device built-in camera, sensors and features (Source: [91]) (b) Illustration of depth image generation process (Source: [92])

software with OpenNI backend [93] with a set of images including a checkerboard of around 50 poses) to find the extrinsic and intrinsic parameters of the IR and RGB cameras. The software uses OpenCv camera calibration and 3D reconstruction (*calib3d*) module for this purpose [94].

Figure 3.4 depicts this process by showing the sample chessboard images taken by the Kinect infrared and visible cameras, the detected chessboard grid points for the computation of the rectification parameters and the achieved stereo rectification re-

sults.

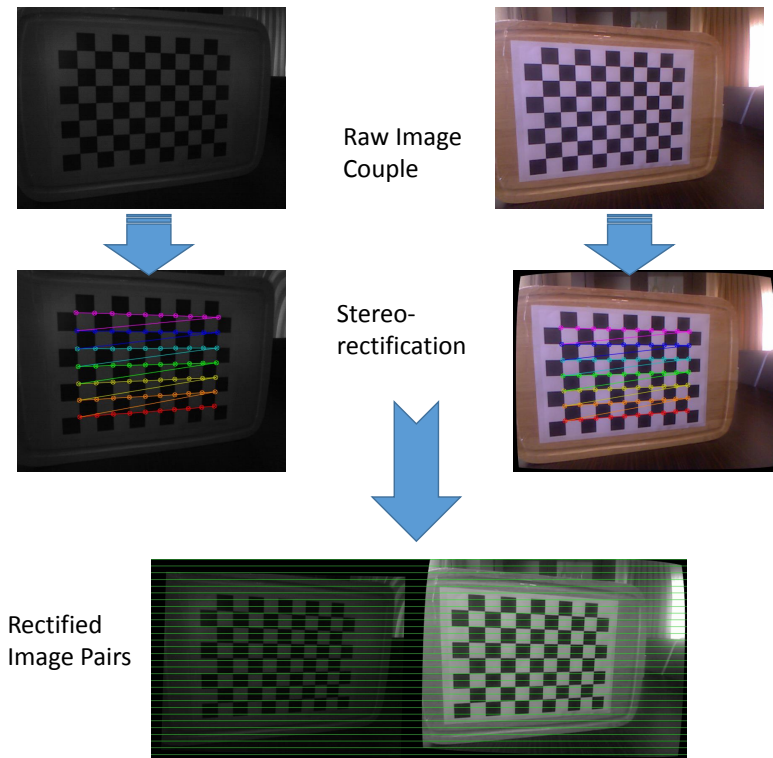


Figure 3.4: Depiction of the Kinect calibration process

The wavelength of the infrared camera in Kinect refers to around 830nm which is defined in the NIR region (refer to Figure 1.2 for the EO/IR spectrum).

Using the Kinect device, a stereo vision evaluation dataset composed of infrared and visible image pairs is constructed. For this purpose, several scenes of indoor environments such as the office cubicles and living room corners with several objects located in the scene which have different reflectance properties are prepared and recorded by the infrared and visible camera of the device. The images are stereo-rectified and totally 24 image pairs are stored in the dataset.

Table 3.2 provides the list of the image pairs that comprises the dataset along with the properties of the images acquired and stereo rectified for multi-modal stereo-vision performance evaluations.

Figure 3.5 provides visuals of sample image pairs from the dataset.

Table 3.2: The Dataset #2 - Kinect Dataset

Dataset	Image No	Image Name	Resolution	Max. Disparity
Dataset #2	1	Kinect01	640×480	36
Dataset #2	2	Kinect02	640×480	21
Dataset #2	3	Kinect03	640×480	20
Dataset #2	4	Kinect04	640×480	20
Dataset #2	5	Kinect05	640×480	27
Dataset #2	6	Kinect06	640×480	25
Dataset #2	7	Kinect07	640×480	20
Dataset #2	8	Kinect08	640×480	16
Dataset #2	9	Kinect09	640×480	16
Dataset #2	10	Kinect10	640×480	33
Dataset #2	11	Kinect11	640×480	23
Dataset #2	12	Kinect12	640×480	33
Dataset #2	13	Kinect13	640×480	33
Dataset #2	14	Kinect14	640×480	33
Dataset #2	15	Kinect15	640×480	23
Dataset #2	16	Kinect16	640×480	23
Dataset #2	17	Kinect17	640×480	23
Dataset #2	18	Kinect18	640×480	23
Dataset #2	19	Kinect19	640×480	23
Dataset #2	20	Kinect20	640×480	23
Dataset #2	21	Kinect21	640×480	20
Dataset #2	22	Kinect22	640×480	20
Dataset #2	23	Kinect23	640×480	30
Dataset #2	24	Kinect24	640×480	15

The *performance evaluation* on the Kinect image pairs are performed by using the depth data that Kinect computes. Two types of metrics are proposed for performance evaluation on this dataset:

- (i) "Percentage Good Depth" (PGD): Percentage of estimates z_c that are close to the Kinect depth z_k for different thresholds δ_z (namely, 10, 20, 30 or 40 cm) for only valid z_k . Note that Kinect's depth is limited to $(0., 5.0]$ meters.

$$PGD = \frac{1}{N} \sum_{x,y} LT(|z_c(x,y) - z_k(x,y)|, \delta_z) : \{\forall(x,y) | z_k(x,y) \in (0., 5]\}, \quad (3.3)$$

$$LT(a, b) = \begin{cases} 1 & \text{if } a < b \\ 0 & \text{otherwise} \end{cases} \quad (3.4)$$

- (ii) "Percentage Total Coverage" (PTC): The percentage added to PGD where Kinect does not provide an estimation ($z_k \notin (0., 5]$ meters) but the evaluated method provides an estimation in the range $(0., 5]$ meters.

$$PTC = PGD + \frac{1}{N} \sum_{x,y} R(z_c(x, y), 0, 5) : \{\forall(x, y) | z_k(x, y) \notin (0., 5]\}, \quad (3.5)$$

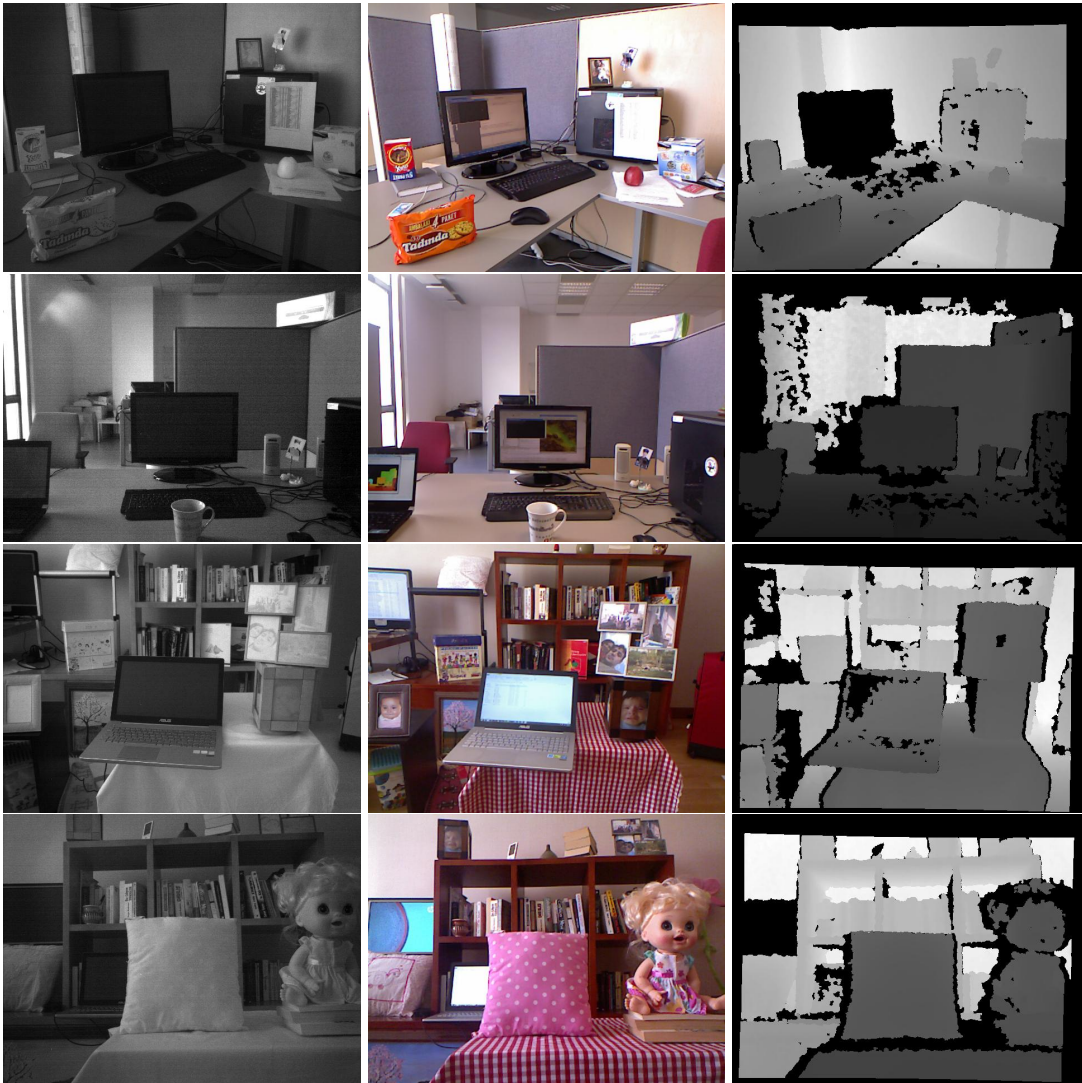


Figure 3.5: Sample image pairs from Dataset #2 - Kinect Dataset *Left column*: Left (IR) camera images. *Middle column*: Right (RGB) camera images. *Right column*: Kinect's native depth computations

$$R(a, b, c) = \begin{cases} 1 & \text{if } a \in (b, c] \\ 0 & \text{otherwise} \end{cases} \quad (3.6)$$

Note that, since there is no ground truth available, it is not possible to provide a quantitative evaluation of which estimation is better, other than these two criteria.

CHAPTER 4

COMPARISON OF SIMILARITY MEASURES USED FOR MULTI-MODAL STEREO-VISION

In this chapter, a list of similarity measures that are widely used in the literature (see Chapter 2.2) are evaluated for their performances using the datasets that were generated in the scope of this thesis (see Chapter 3 for the description of datasets).

The similarity measures can be grouped into three categories similar to [74]:

1. *Local Window Based Methods*: These methods perform calculation of the measures by using local windows extracted around the compared pixels from left and right images. The SSD, NCC and MI (with and without Prior Probability) measures fall into this category. The similarity measures are computed as below, where the details are provided in Section 4:

$$SSD(W_L, W_R) = \sum_{x,y} (W_L(x, y) - W_R(x, y))^2, \quad (4.1)$$

$$NCC(W_L, W_R) = \frac{\sum_{x,y} (W_L(x, y) - \bar{W}_L)(W_R(x, y) - \bar{W}_R)}{\sqrt{\sum_{x,y} (W_L(x, y) - \bar{W}_L)^2 (W_R(x, y) - \bar{W}_R)^2}}, \quad (4.2)$$

$$MI_{(woPR)}(W_L, W_R) = \sum_{X \in W_L} \sum_{Y \in W_R} P(X, Y) \log \frac{P(X, Y)}{P(X)P(Y)}, \quad (4.3)$$

$$MI_{(wPR)}(W_L, W_R) = \sum_{X \in W_L} \sum_{Y \in W_R} P^*(X, Y) \log \frac{P^*(X, Y)}{P(X)P(Y)}, \quad (4.4)$$

where W_L and W_R are the two matching local windows around left and right pixels to be tested for correspondence and \bar{W}_L , \bar{W}_R are the mean pixel intensities of the local windows, $P(X, Y)$ is the joint probability distribution, $P(X)$ and $P(Y)$ are the

prior probabilities of the two matching windows and $P^*(X, Y)$ are the joint probabilities incorporated with joint prior probabilities, i.e. $P^*(X, Y) = \lambda P(X, Y) + (1 - \lambda)P_{prior}(X, Y)$. P_{prior} is computed from the joint histogram of corresponding pixels through the whole image (Refer to Section 4 for the details).

2. *Measures as a Collection of Feature Vectors*: These methods are composed of initially calculating feature vectors densely for each pixel. The LSS, HOG, SIFT and SURF measures fall into this category. To compute the stereo correspondences, a similarity measure using the sum of distances of each corresponding feature vector of the pixels within the local windows around the matching left and right image pixels are used, which is defined as below:

$$SM(W_L, W_R) = \sqrt{\sum_{x,y} (f_L(x, y) - f_R(x, y))^2}, \quad (4.5)$$

where f_L and f_R are the feature vectors of each pixel in the two matching windows W_L and W_R .

3. *Measures based on Binary Comparisons*: These methods are based on binary descriptors for each pixel. CENSUS, CENSUSMI, BRIEF, FREAK measures fall into this category. *Hamming* distance of the matching windows are applied to the binary descriptors as the similarity measure (see Eqn. 2.29).

4.1 Performance Evaluation Using Dataset #1 - The Synth. Alt. Middlebury Dataset

In this section, the performance evaluation of the similarity measures are performed using the Dataset #1 - The Synthetically Altered Middlebury Dataset. After computing the similarity measures for the matching pixels, the "WTA" - Winner Takes All disparities are computed by selecting the best disparity having the maximum similarity value over candidate disparities. The performance evaluations are performed as described in Chapter 3.

Three different experiments are conducted in the scope of this section, as provided in the following subsections. Initially, the measures are tested using three different window sizes, 9x9, 21x21 and 31x31. Next, the measures are tested for different

multi-modality levels of the left image. Finally, several levels of Gaussian noise are added to the left image and the measures are tested for increasing noise levels.

Appendix A provides the parameter settings used in the experiments for each of the similarity measure method.

4.1.1 Effect of Window Size

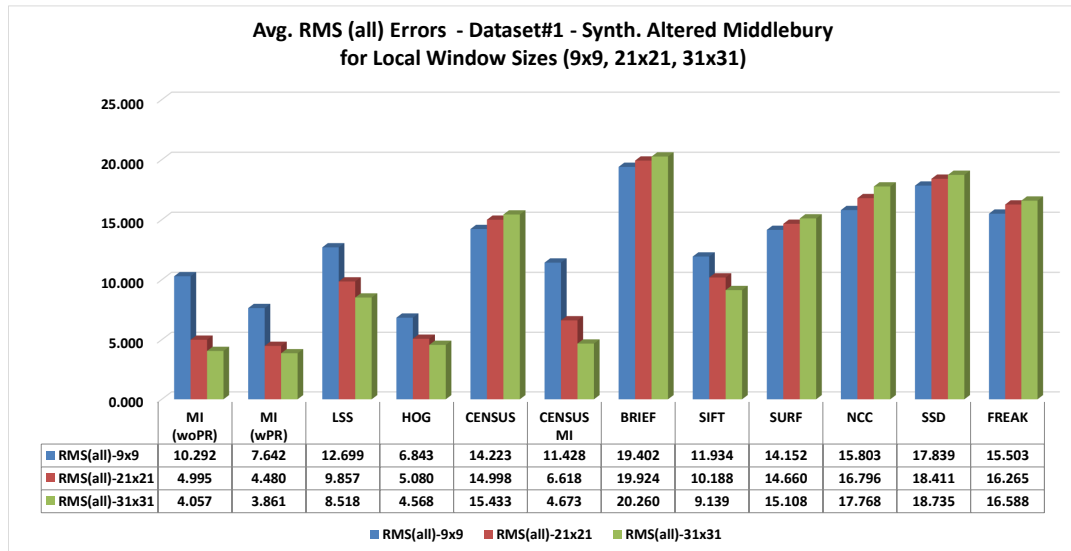


Figure 4.1: Average RMS(all) errors of all methods' "WTA" performances for three different window sizes for Dataset #1 - the synthetically altered Middlebury image pairs

The experiments in this part consist of performance evaluation over the computed disparity maps by using three different local window sizes, 9x9, 21x21 and 31x31 pixels for the similarity measure computations. The table of performance statistics computed are provided in Table B.1 in Appendix B along with the whole set of visual results.

On the other hand, Figure 4.1 and Figure 4.2 show the average RMS and Bad pixel percentage errors for the "all" regions of performance evaluation where Figures B.1, B.2, B.3 and B.4 in Appendix B show the RMS and Bad pixel performances for each image separately, i.e. Tsukuba, Venus, Teddy and Cones. The Figure 4.3 shows sample visual results of the "WTA" disparity maps obtained from the Tsukuba image

for the leading similarity measures tested.

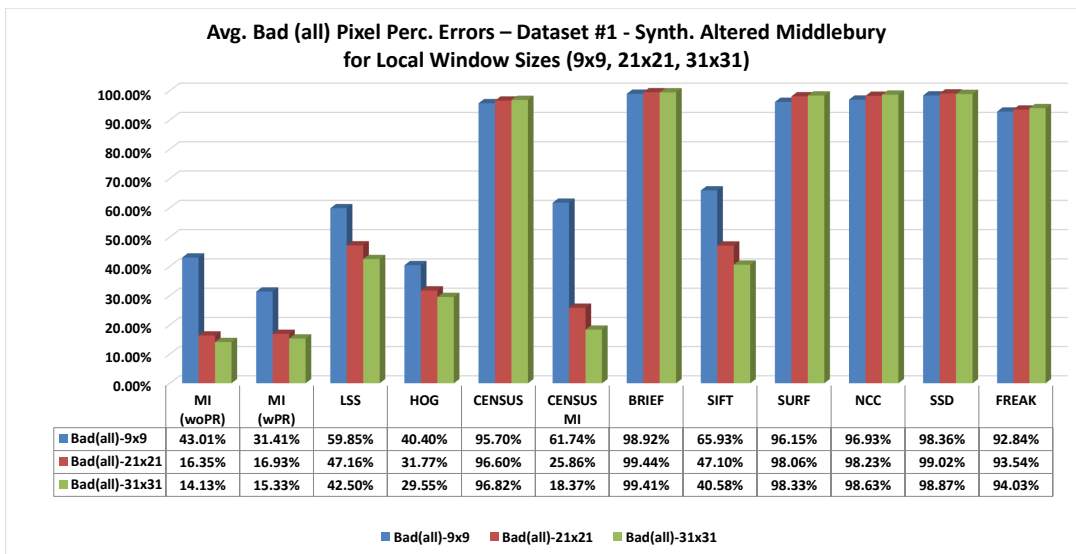


Figure 4.2: Average Bad(all) pixels percentage errors of all methods' "WTA" performances for three different window sizes for the dataset1 - the synthetically altered Middlebury image pairs

It can be observed from the results that the average RMS and Bad pixel percentage errors are obtained as the smallest for the MI(wPR) and MI(woPR) similarity measures where incorporation of the prior probabilities enhanced the results of MI measure. Besides, MI of Census Transform measure that was proposed in the scope of this thesis enhanced the Census Transform results significantly. HOG is ranked right after the MI results which shows us the effect of using gradient information in the multi-modal image pairs help to match the image patches for stereo correspondence problem. LSS is following these measures using the spatial information in the images and SIFT is following LSS which also effectively uses the gradient information in several scale levels.

On the other hand, the similarity measures SURF, CENSUS, BRIEF, FREAK, NCC and SSD are totally confused for this set of image pairs which depend more on similarity in the intensity levels and texture.

When the results are evaluated for the effect of local window sizes extracted for computing the similarity measures, it is observed that as the size of the windows increase, the performances also increase but by decreasing distances for the measures that can


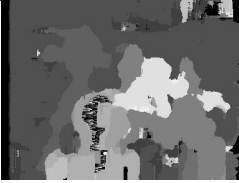


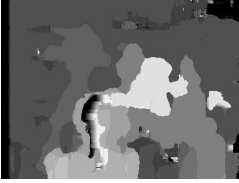

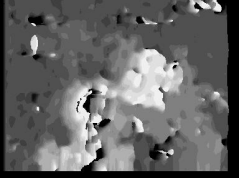
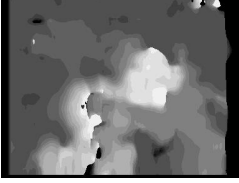
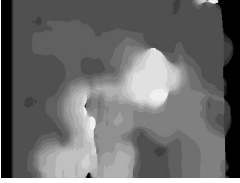



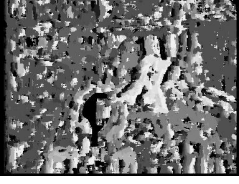
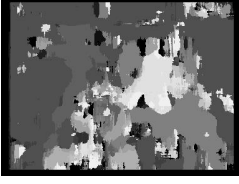




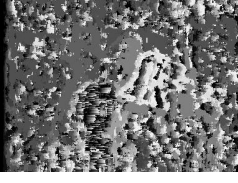
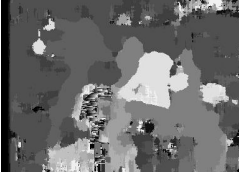
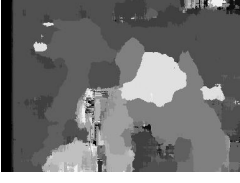
	9x9	21x21	31x31
MI (woPR)			
MI (wPR)			
HOG			
LSS			
SIFT			
CENSUS			
CENSUSMI			

Figure 4.3: Sample visual results of the leading similarity measures for the synthetically altered Tsukuba image pair in Dataset #1, for the different window sizes 9x9, 21x21 and 31x31 pixels. (includes the results of the novel CENSUSMI method along with the original CENSUS method results)

represent multi-modal image patches to some extent (i.e. MI, HOG, CENSUSMI, LSS, SIFT). Especially for teddy and cones images having bigger and curved objects

are affected more by the local window size. HOG affected less than all other measures by the window size where it provides best result in average for the smallest window size 9x9. This is because of the similarity measure computing method since an inner window is used to compute each feature vector for each pixel in the compared local window yielding a greater window in total.

4.1.2 Effect of Multi-Modality

The experiments in this part aims to evaluate the performances of the similarity measures for different multi-modality levels. To accomplish this task, the below equation is proposed (Eqn.4.6), which generates image pairs where the left images range from the synthetically-altered cosine transformed image in the dataset#1 to the original unimodal left image in the Middlebury image database.

$$I_m(x, y) = I_{orig}(x, y)(1 - m) + I_{cos}(x, y)(m), \quad (4.6)$$

where I_{orig} is the original image from the Middlebury Image Database, I_{cos} is the cosine transformed image as ($I_{cos} = \cos(\pi I_{orig}/255)$) and m ($m \in [0, 1]$) is the multi-modality level for generating I_m image. Therefore, when $m = 1$, I_m shall be equal to I_{cos} and when $m = 0$, I_m shall be equal to I_{orig} . Figure 4.4 illustrates this method of how the images of different modality are generated synthetically by this equation.



Figure 4.4: Figure illustrating the method for generating images of different multi-modality. m : the multi-modality level scale ($m=0.5$ in this case) *Left image*: Original Tsukuba image from Middlebury image database *Middle image*: Cosine transformed image *Right image*: Generated image of multi-modality level $m=0.5$

Regarding the experiments presented in this section, 10 multi-modality levels are generated and the performance statistics of each level are computed for each of the similarity measure. The experiments are held with local window sizes set to 21x21

pixels.

Figure 4.5 and Figure 4.6 shows the average RMS and Bad pixel percentage errors for the "all" regions of performance evaluation, for the 10 levels of multi-modality of Tsukuba, Venus, Teddy and Cones images in the Dataset #1 where M10 (corresponds to $m = 1$) stands for the cosine transformed left image in Dataset #1 and M0 (corresponds to $m = 0$) stands for the original left image in the Middlebury image database.

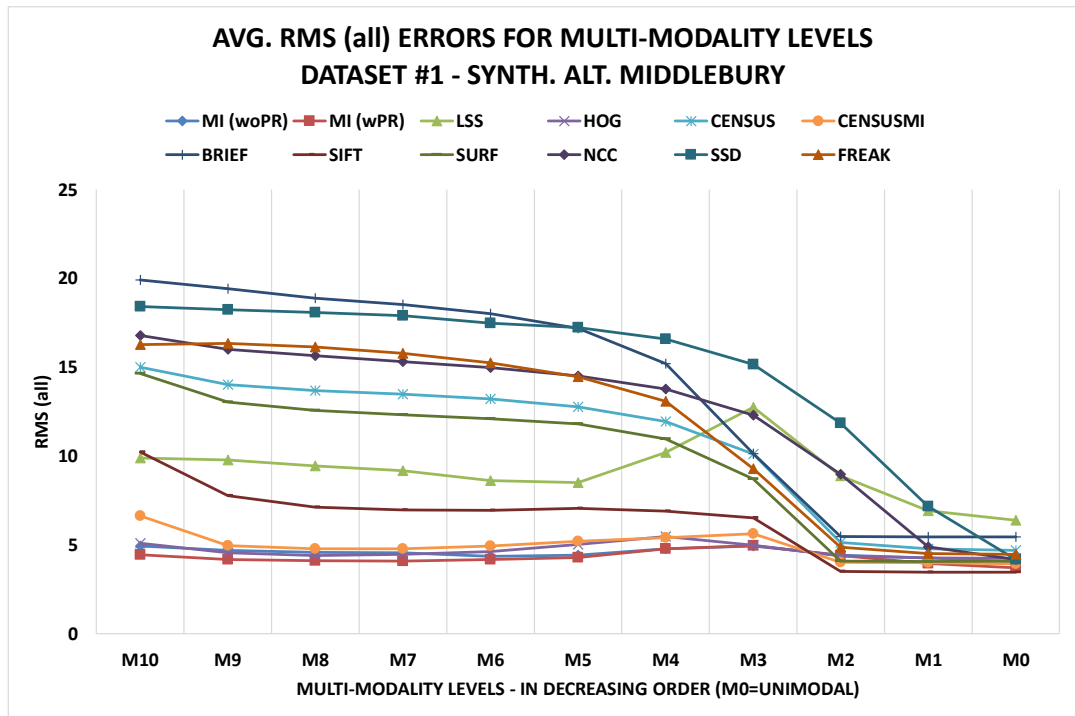


Figure 4.5: Average RMS(all) errors of all methods for 10 multi-modality levels for the Dataset#1 image pairs

The Figures B.6, B.7, B.8 and B.9 in Appendix B provide the RMS(all) and Bad(all) pixels percentage errors of all similarity measures tested for the 10 multi-modality levels for each image pair in Dataset #1 separately, i.e. Tsukuba, Venus, Teddy and Cones. Figure 4.7 shows sample visual results for the disparities generated by some similarity measures for computed multi-modality levels.

As can be observed from the obtained results, the similarity measures are clustered into three groups. The 1st group is MI(woPR), MI(wPR) and CENSUSMI where no significant change occurs from the multi-modal to unimodal image pairs. HOG can

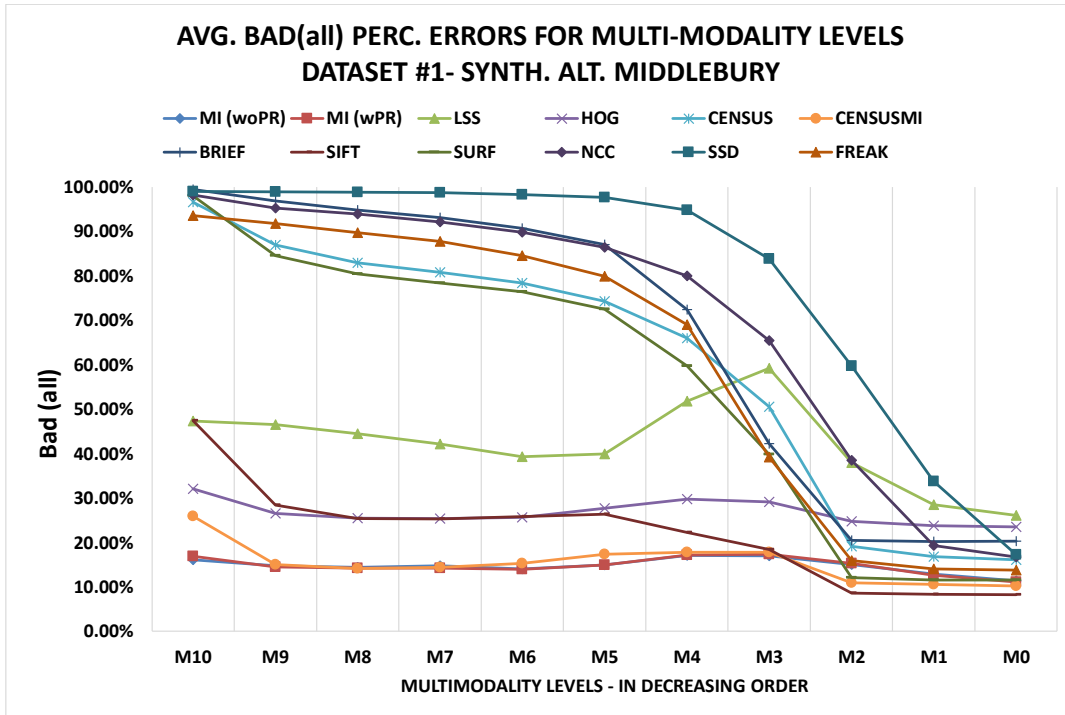


Figure 4.6: Average Bad(all) pixel percentage errors of all methods for 10 multi-modality levels for the Dataset#1 image pairs

also be added into this group which has worse results in bad pixels percentage but the change in the performance curve is still not significant.

The 2nd group is composed of LSS and SIFT which present moderate results for the multi-modal case although yield good results in unimodal case. LSS makes an increase in error at the M3 level which is concluded to result from the disappearance of some of the spatial features at this level.

The 3rd group includes the rest of the measures which are SURF, CENSUS, BRIEF, FREAK, NCC and SSD. Among these, the SURF, FREAK, CENSUS and BRIEF starts having good results at and after M2nd level and NCC at the M1st level. SSD needs to wait until the original unimodal case is configured (M0) as can be expected.

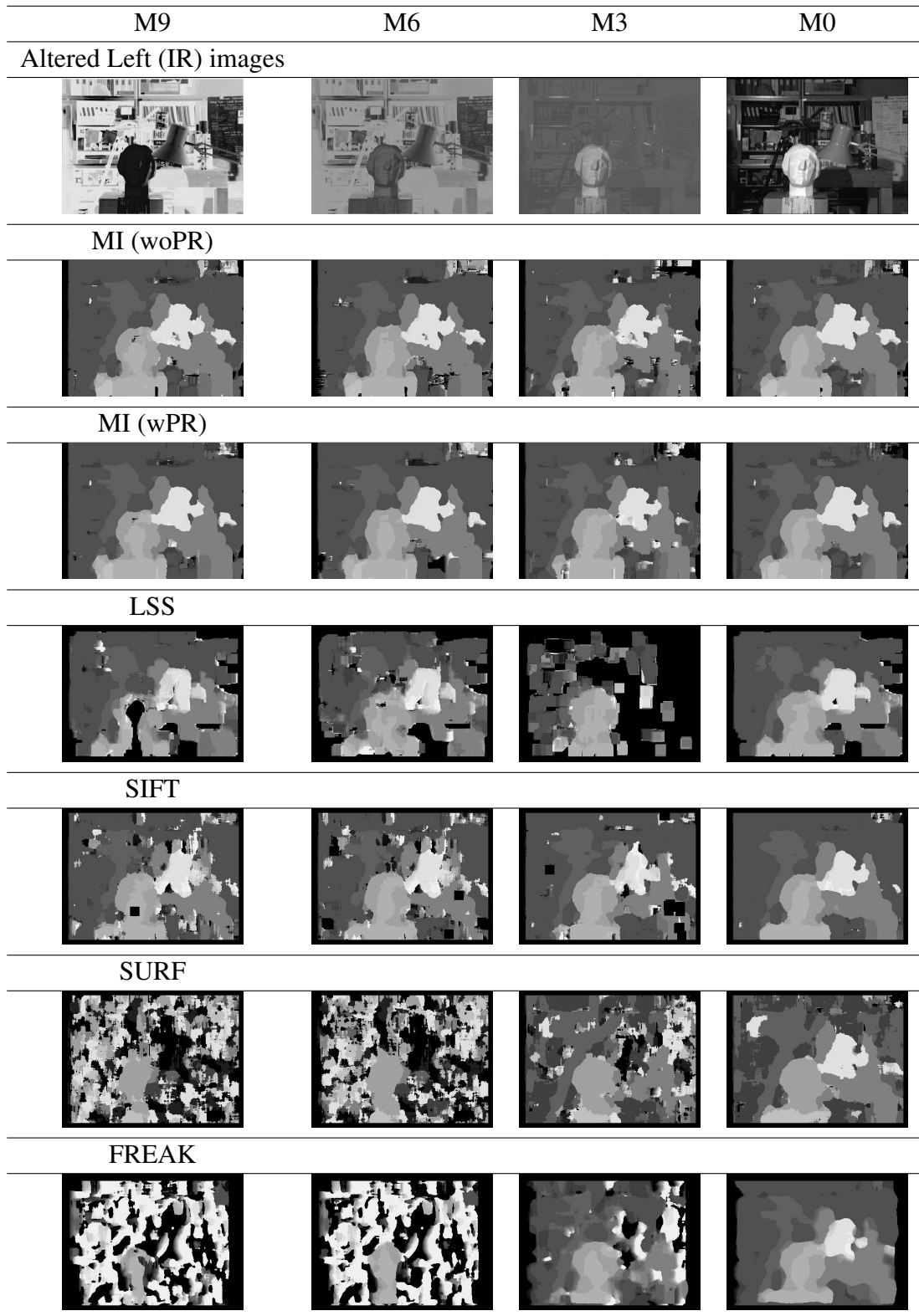


Figure 4.7: Sample visual results of selected similarity measures for given multi-modality levels (M9, M6, M3 and M0) of the Tsukuba image pair (local window size=21x21). 1st row shows altered left images of given multi-modality levels).

4.1.3 Effect of Noise

The experiments in this part aims to evaluate the performances of the similarity measures for different noise levels. To accomplish this, a noise image is generated using a normal distribution of random values of gray levels z determined for the mean $\mu = 0$ and for standard deviation σ as:

$$N(z, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}, \quad (4.7)$$

for all pixels (x, y) and added to the initial cosine transformed left image as:

$$I_{noisy}(x, y, z) = I_{cos}(x, y, z) + N(z, \mu, \sigma). \quad (4.8)$$



Figure 4.8: Different noise levels applied to left Tsukuba image in Dataset #1. (a) Noise level $n = 10$ ($\sigma = 20.0$) (b) Noise level $n = 6$ ($\sigma = 12.0$) (c) Noise level $n = 3$ ($\sigma = 6.0$) (d) Noise level $n = 0$ ($\sigma = 0.0$) the noiseless cosine transformed left image.

The left images of different noise levels are generated by changing the σ in the range $[20.0, 0.0]$ where regarding the experiments presented in this section, 10 noise levels

are generated for $\sigma = 20.0$ for noise level $n = 10$ and decreasing by 2.0 at each level up to $\sigma = 2.0$ for $n = 1$ and $\sigma = 0.0$ for $n = 0$ which is the noiseless image. The performance statistics of each level are computed for each of the similarity measure. The experiments are held with local window sizes set to 21x21 pixels. Figure 4.8 shows several of the images with decreasing noise levels.

Figure 4.9 and Figure 4.10 shows the average RMS and Bad pixel percentage errors for the "all" regions of performance evaluation, for the 10 levels of noise in the left images of Tsukuba, Venus, Teddy and Cones images in the dataset#1 where N10 stands for the noise level $n = 10$ and N0 stands for noiseless left image.

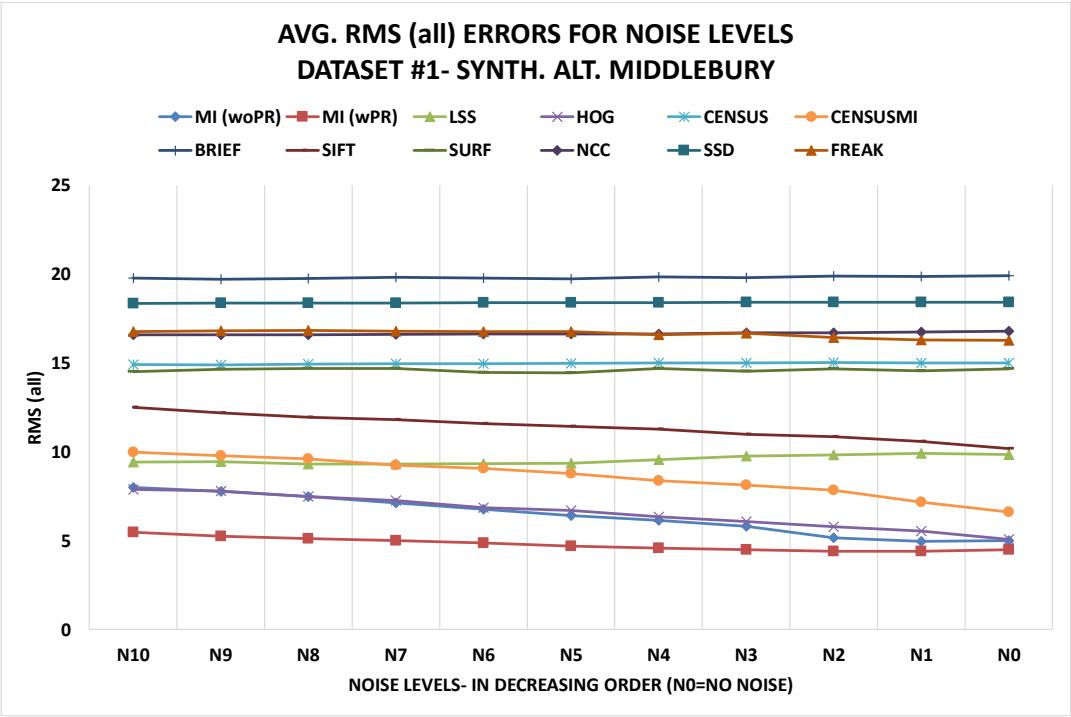


Figure 4.9: Average RMS(all) errors of all methods for 10 noise levels for the dataset#1 image pairs

Figures B.10, B.11, B.12 and B.13 in Appendix B show the RMS and Bad pixel performances for each image pair separately, i.e. Tsukuba, Venus, Teddy and Cones, for the noise levels.

Figure 4.11 shows sample visual results for the disparities generated by some similarity measures for computed noise levels.

In this type of experiments, it is intended to check the vulnerability of the similarity

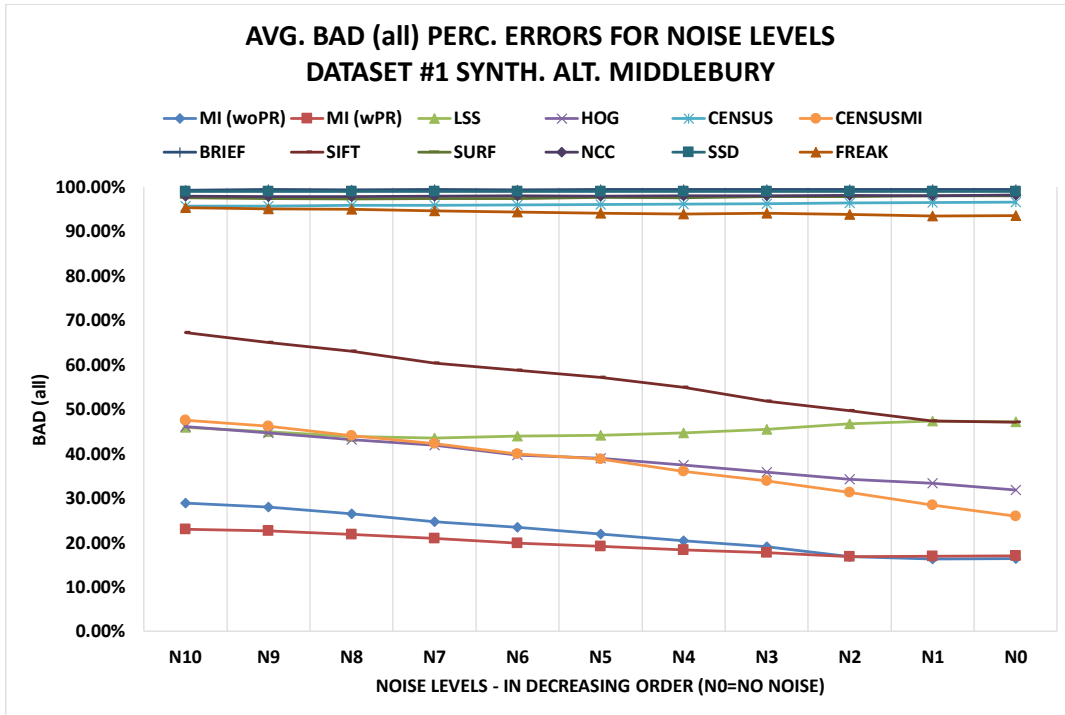


Figure 4.10: Average Bad(all) pixel percentage errors of all methods for 10 noise levels for the dataset#1 image pairs

measures which have promising results for multi-modal image pairs. The results of the measures classified as the 1st and 2nd group, i.e. MI, CENSUSMI, HOG, LSS and SIFT should be focused. As can be observed from the obtained results, CENSUSMI, SIFT and HOG are concluded as the most vulnerable measures to noise. MI (woPR) also increase the performance as noise is decreased. MI(wPR) is concluded as the most robust method to noise. On the other hand, LSS is not affected by the noise and even has a small shift in error upwards. It is concluded that this behavior was due to the small increase in spatial correlation of homogeneous segments due to added noise, as can be observed from the results in venus, teddy and cones alone given in Figures B.10, B.11, B.12 and B.13.

4.2 Performance Evaluation Using Dataset #2 - The Kinect Dataset

In this section, the performance evaluation of the similarity measures are performed using the Dataset #2 - The Kinect Dataset. Four representative image pairs are selected from the dataset which includes properties like fronto-planar, tilted surfaces as

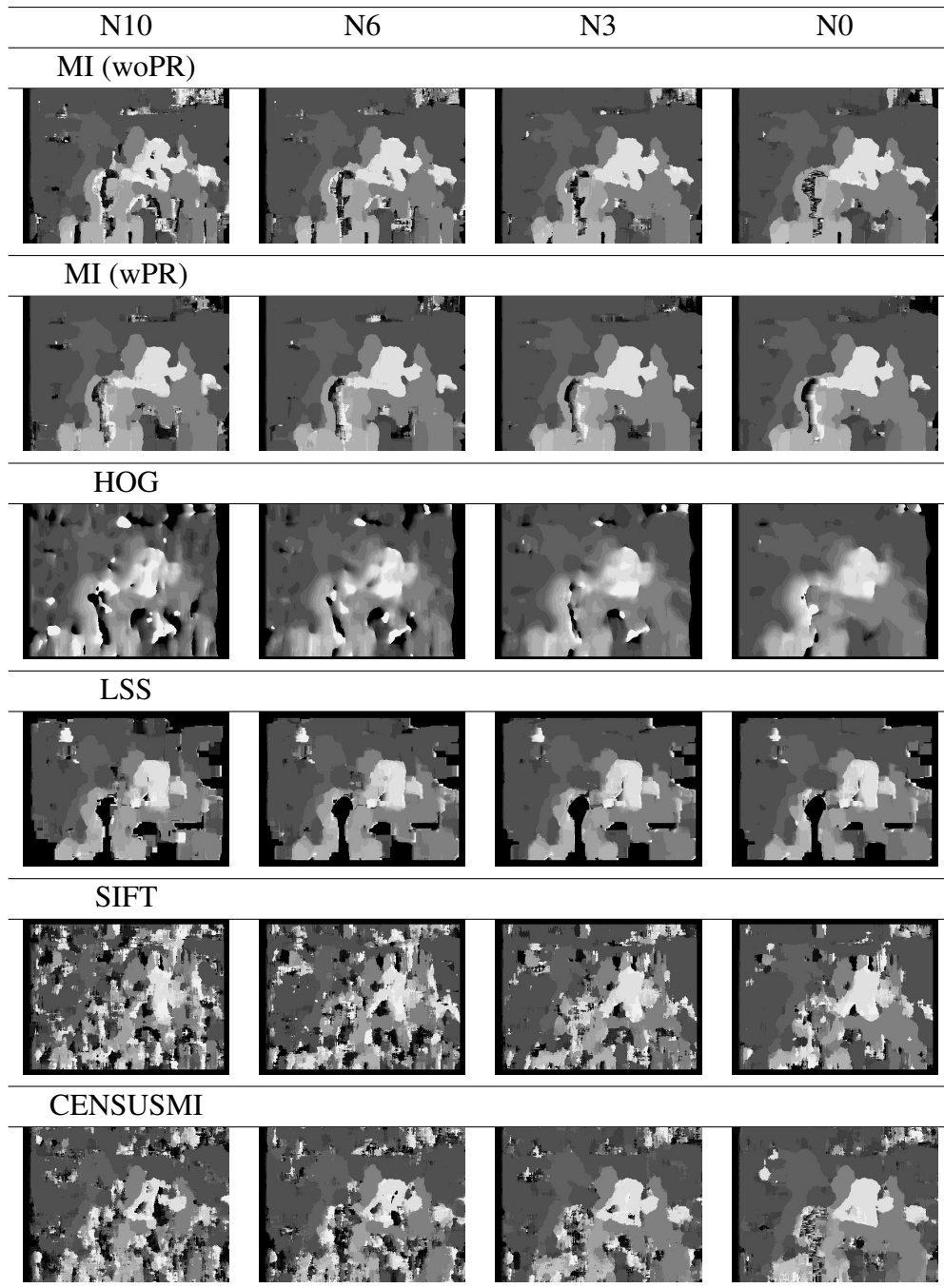


Figure 4.11: Sample visual results of some similarity measures for the added noise levels to Tsukuba left image in Dataset #1 (local window size=21x21) (noise levels: N10 ($n = 10$), N6 ($n = 6$), N3 ($n = 3$) and noiseless N0 ($n = 0$))

well as objects composed of curved and irregular surfaces.

Same as the previous section, the experiments are held for the "WTA" - Winner Takes All disparities that are computed by selecting the best disparity having the maxi-

imum similarity value over candidate disparities. The performance evaluations on the Dataset #2 are performed using the evaluation method and the metrics as described in Section 3.2.

The measures are computed using the local window sizes of 31x31, since the resolution of these images are higher than Dataset #1 images. Appendix A provides the parameter settings used in the experiments for each of the similarity measure method where same settings are used as the Dataset #1 experiments.

The table of performance statistics computed are provided in Appendix B in Table B.2.

Figure 4.12 shows selected images from Dataset #2 for the experiments to be held for the performance evaluation of all similarity measures along with the left (IR) and right (RGB) image pairs and the kinect built-in depth image which is used as the ground truth depth for computing the performance metrics.

Figure 4.13 shows average *Percentage Good Depth* and *Percentage Total Coverage* metrics computed for similarity measures for the Dataset #2 selected image pairs. The metrics are computed using the final depth data reconstructed from the disparity maps and comparing the depth data to Kinect generated depth data for thresholds of 10cm, 20cm, 30cm and 40 cm for the depth range of the scenes $< 500cm$.

Figure 4.14 shows visual results of generated disparities for the dataset image Kinect06 for all the similarity measures tested. See Figure 4.12 for the Kinect generated depth data (note that brighter pixels have more depth in the depth image in contrast to disparity image results provided.)

Experiments over the Kinect infrared-visible image pairs show that SURF and FREAK in addition to the leading performing measures in the Dataset #1 - Synthetically Altered Middlebury Dataset (MI, CENSUSMI, HOG, SIFT and LSS) also performed well over the metrics computed from Kinect generated depth image. This is explained by the low multi-modality in Near-Infrared and visible image pairs where only cloth textures and monitors behave well different than the rest of the scene objects where almost all measures failed to match well. Besides, the performance metrics of none of the similarity measures were greater than %55 even for the 40 cm threshold where

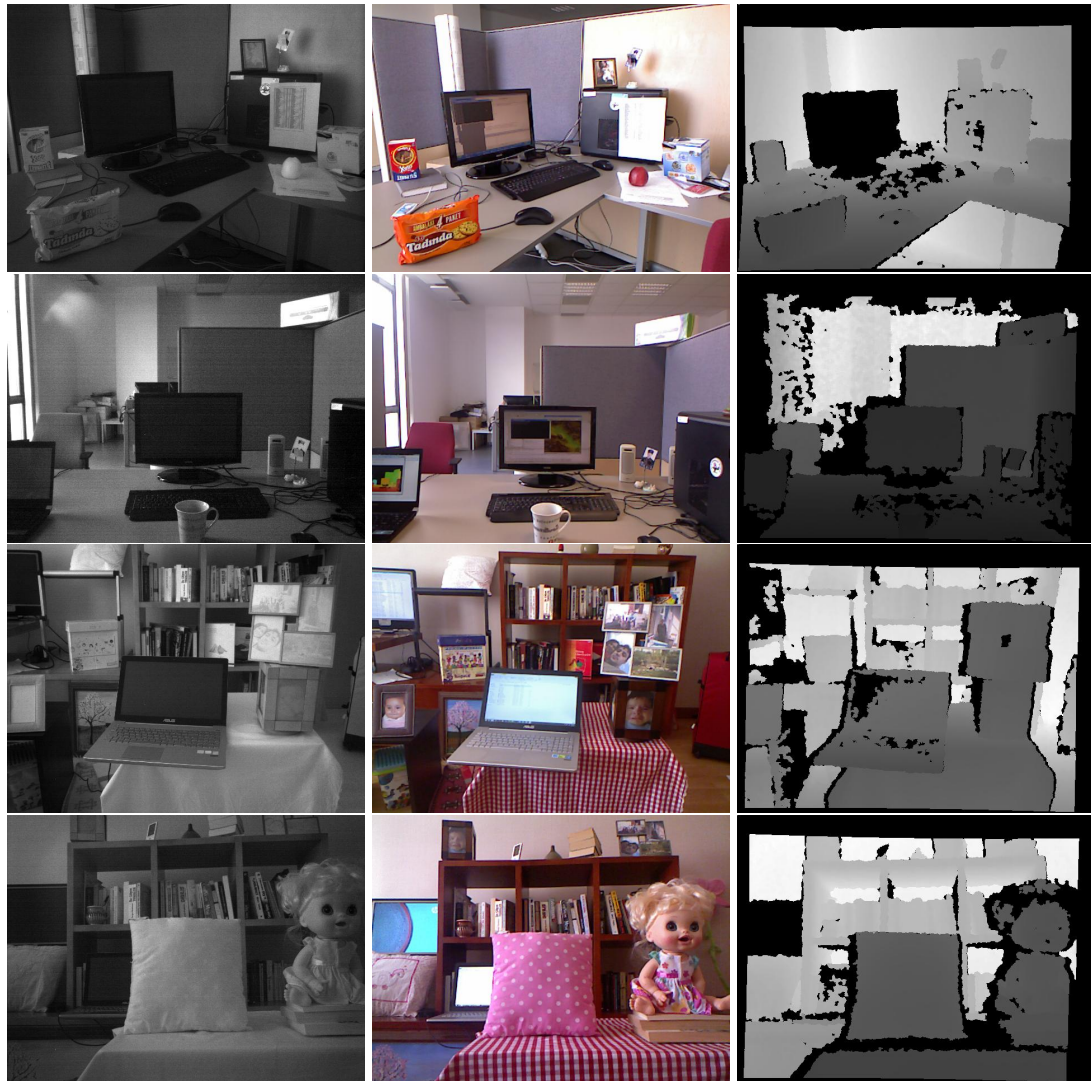


Figure 4.12: Selected Dataset #2 - Kinect Dataset Image Pairs and kinect computed depth images: *Left column*: Left (IR) camera images. *Middle column*: Right (RGB) camera images. *Right column*: Kinect's native depth computations (brighter pixels have more depth). From top to bottom, Dataset #2 Image Ids : Img#2, Img#3, Img#6,Img#10

most of the measures were below %50 compared to Kinect depth data.

4.3 Summary and Discussion

In this chapter, a list of similarity measures that are widely used in the literature (see Chapter 2.2) are evaluated for their performances using the datasets that were

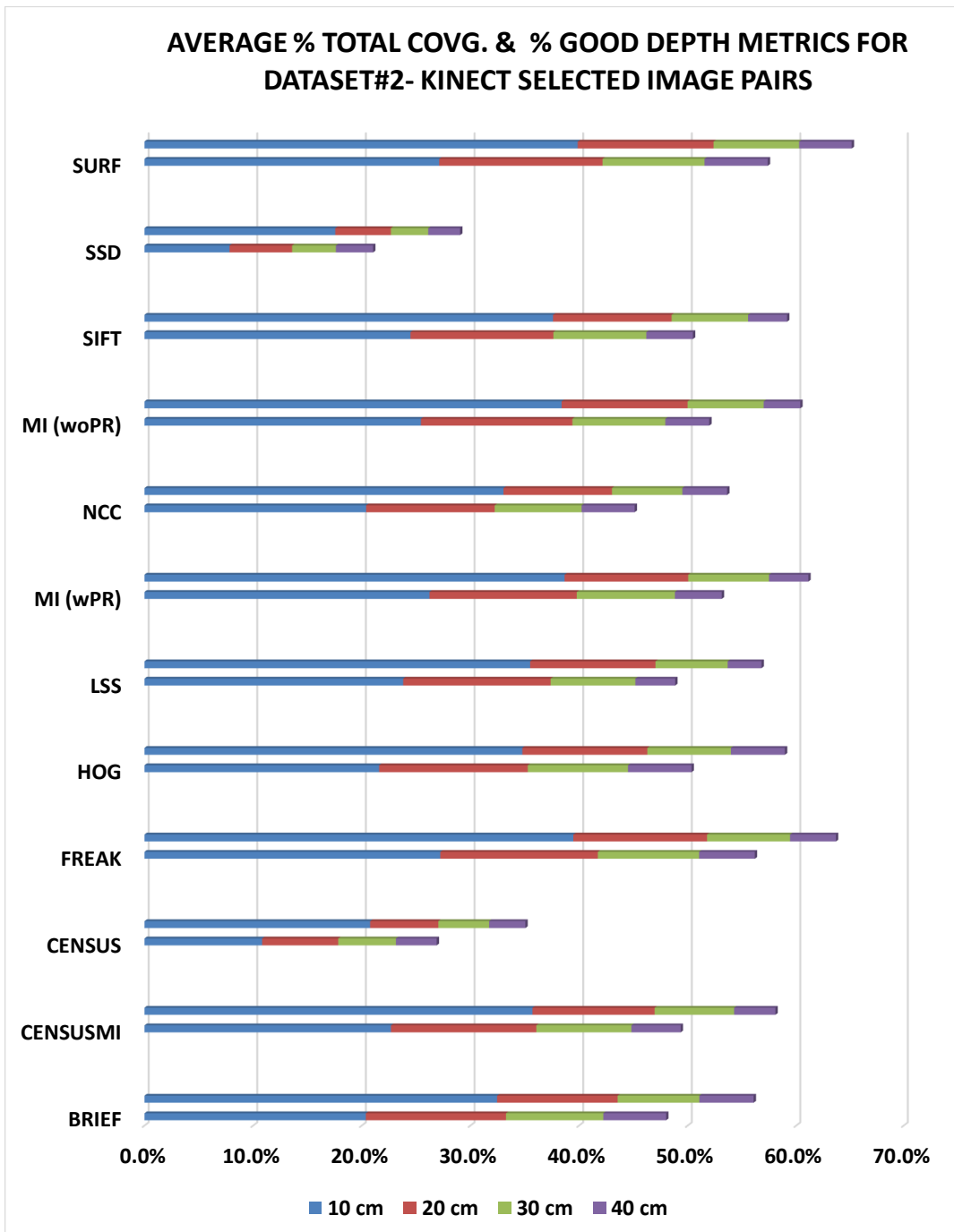


Figure 4.13: Average *Percentage Good Depth* and *Percentage Total Coverage* metrics computed for similarity measures for the Dataset #2 selected image pairs.

generated in the scope of this thesis (see Chapter 3 for the description of datasets).

The experiments conducted using Dataset #1 - Synthetically Altered Middlebury Images provides a good evaluation of the measures regarding the effect of multi-

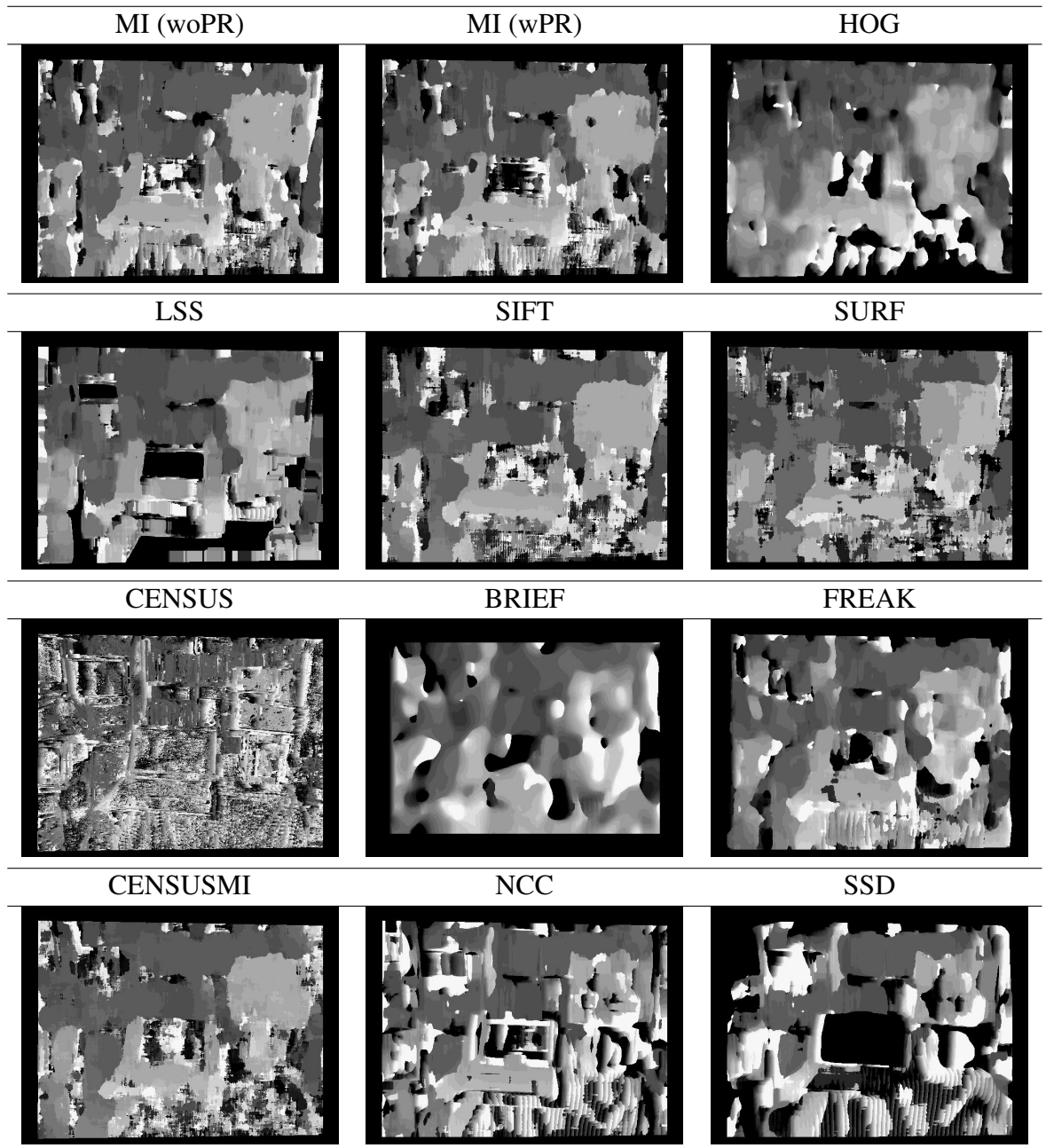


Figure 4.14: Sample visual results of computed WTA disparities of the similarity measures for Kinect06 image in Dataset #2 (local window size=31x31) (brighter pixels have more disparity meaning more closer and have less depth)

modality, noise and the local window sizes. MI, CENSUSMI and HOG are shown to have better performance than the rest of the measures for multi-modal imagery where LSS and SIFT yields moderate results. Among these, CENSUSMI, SIFT and HOG are shown to have vulnerability to noise. Finally, the local window sizes increase the

performance figures as the sizes increase, however, at the expense of blurring in the resultant disparity image as expected. The upward shift in performance is smaller than the difference in the growing window sizes.

Experiments over Kinect image pairs show that although FREAK and SURF provides good performance metrics on the computed depth images, the regions where the multi-modality is high like cloth textures or monitor screens which does not provide any image in infrared band interval are still challenging to match for almost all methods. Overall metrics show that there is still a significant space for improvement on WTA results of these measures.

CHAPTER 5

AN ITERATIVE MULTI-MODAL STEREO-VISION METHOD

In this chapter, a new MI-based multi-modal stereo-vision method, composed of several consecutive steps that are iteratively refined, is proposed. The method starts with a segmentation of the IR image, estimates disparities using windows that are adapted to the sizes of the segments, improves the estimated disparities with segment merging-splitting, and re-iterates these steps to get even better estimates.

The implemented method was applied to the two datasets that were presented in Chapter 3: Namely, Dataset #1 - the Synthetically Altered Middlebury Stereo Evaluation Dataset and Dataset #2 - the Kinect Dataset.

5.1 Method

The overview of the proposed method is depicted in Figure 5.1. The method takes as input a pair of rectified multi-modal images, satisfying the epipolar line constraint such that correspondences can be found on horizontal scanlines. The initial step is to segment the left (IR) image. Next, the cost matrix for all candidate matching pixels in each scanline of the rectified image pair according to a predefined maximum disparity is computed by the MI computation algorithm using adaptive windowing method proposed uniquely in this study. Later, the raw costs are adaptively aggregated using confidence metrics and the segmentation information. Next, the disparity planes corresponding to segments are computed from the stable pixels where the outliers of each disparity plane are inspected for splitting the segment. Finally, the segments are inspected for merging with a neighboring segment by comparing the similarities

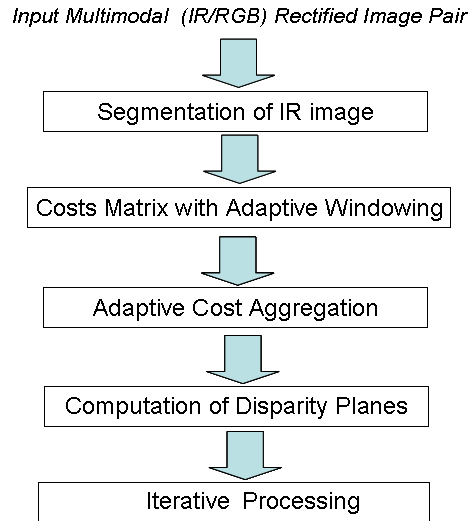


Figure 5.1: Overview of the proposed iterative multi-modal stereo-vision method.

between the associated disparity planes. The new iteration uses refined segmentation and current disparity map for the new disparity plane computation. In the following subsections, each of these steps are explained in more detail. Table 5.1 provides the definitions of the symbols used throughout this chapter for the description of the proposed method.

5.1.1 Segmentation of the IR Image

As the first step in the method, the IR image is segmented where the rest of the processing relies on this segmentation. The reason for segmenting only the IR image is that the surfaces in IR images are also common in the RGB images but the reverse is not true (see Figure 5.2) since RGB images contain more detailed and textured surfaces which do not exist in the IR images in our datasets. With this step, non-overlapping segments representing homogeneous regions in the IR image is generated (see Figure 5.6). It is assumed that each segment corresponds to a planar surface in the scene, which is a common assumption in segmentation-based stereo-vision techniques [95, 96, 97, 98, 99, 100, 101].

For segmenting the IR image, Synergistic Image Segmentation (SIM) algorithm [102] is used. The SIM method incorporates an edge magnitude/confidence map into the

Table 5.1: List of notations and acronyms used for the method descriptions

Symbol	Definition	Symbol	Definition
L	Left (IR) image	l_c	current center pixel in left image
R	Right (RGB) image	$Conf$	Confidence map regarding calculated costs
(i)	Iteration number ($i \in [0, N]$)	c_1	Min. cost of the candidate disparities
S	Segmentation	c_2	Second min. cost of the candidate disparities
C	Cost matrix	ρ	Ceiling value for the maximum confidence
D	Disparity map	w	Weights for performing cost aggregation
MI	Mutual Information	b	Half-size of the window for cost aggregation
WTA	Winner Takes All	SD	Spatial Distance
x	Column number of a pixel	DD	Disparity Distance
y	Row number of a pixel	λ_{SD}	Designated scaling constant for spatial distance
d	Disparity in range $[0, d_{max}]$	λ_{DD}	Scaling constant for disparity distance
p	Current pixel	f	Function for subpixel disparity computation
q	Neighbor pixel	τ_{ic}	Confident inlier disparity threshold
s	Segment in ($s \in S$)	τ_{ir}	Stable segment ratio threshold
I_l	Intensities of left image pixels	τ_{od}	Outlier disparity distance threshold
I_r	Intensities of right image pixels	τ_{os}	Outlier disparities size threshold
W	Local window of computation for a center pixel	τ_{oc}	Confident outlier disparity threshold
ω	Assumed thickness of discontinuities in images	$Plane$	Set of disparity planes
P	Joint Probability	α	Angle between two disparity planes
P_{prior}	Prior Joint Prob. of Left & Right Images	τ_α	Angle threshold for parallel planes
P_{window}	Joint Prob. of Left & Right Images	τ_{pd}	Plane to plane distance threshold
λ	Ratio of incorporating prior prob. to joint prob.	h_s	Spatial bandwidth in mean-shift segm.
h_w	Histogram computed for the adaptive window	h_r	Feature (range) bandwidth in mean-shift segm.
$T()$	Counter function for hist. computation	M	Minimum segment size in mean-shift segm.
$L1$	L1 distance	n	Size of the grad. window in syn. image segm.
k	Increment of counts for histograms	a_{ij}	Mixture parameter in syn. image segm.
t_e	Threshold value for the edge computation		

mean-shift segmentation algorithm [103] enhancing the results on especially weak edges, and hence, separating the objects better. The algorithm makes use of the parameters of the mean shift segmentation algorithm; the spatial bandwidth h_s , the feature (range) bandwidth h_r and the minimum segment size M as well as the size of the gradient window n , the mixture parameter for blending of the gradient magnitude a_{ij} and the threshold for the discontinuities t_e - see [102] for the details.

5.1.2 Computing the Cost Matrix

Computing the cost matrix is the key step of the method, and a significant part of the contributions in the thesis (see Algorithm 1). The inputs to the algorithm are the left (IR) image L and the right (RGB) image R , the segmentation $S^{(i)}$ (computed from the left image for the initial iteration and modified at the previous iteration for the subsequent iterations) and the disparity map $D^{(i)}$ ($D^{(0)} = 0$, and otherwise, $D^{(i)}$ is the disparity map generated in the previous iteration).

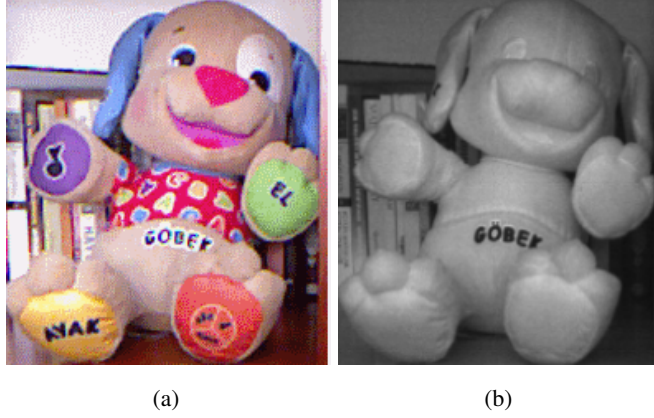


Figure 5.2: The Doll dog from a Kinect image pair in Dataset #2 - The Kinect Image Database (a) The Kinect RGB Image. (b) The Kinect IR image. [Best viewed in color].

Algorithm 1 Cost matrix computation.

Inputs: L : Left (IR) Image
 R : Right (RGB) Image
 $S^{(i)}$: Input segmentation ($i \in [0, N]$: iteration)
 $D^{(i)}$: Input disparity map ($D^{(0)}$ is zero)
Outputs: $C^{(i)}$: The cost matrix

```

1: Compute  $P_{prior}^{(i)}(L, R, D^{(i)})$  //see Eqn. 5.1
2: for  $y = 0$  to height do
3:   for  $x = 0$  to width do
4:     for  $d = 0$  to  $d_{max}$  do
5:        $C^{(i)}(x, y, d) \leftarrow -M(W_L(x, y), W_R(x - d, y), S^{(i)}, P_{prior}^{(i)})$ 
           // see Eq. 5.8 for  $M()$ 
6:     end for
7:   end for
8: end for
9: return  $C^{(i)}$ 

```

Algorithm 1 first computes joint prior probabilities for all corresponding pixel intensities in left and right images using the current disparity map available (for the sake of simplicity, in the rest of the section, the current iteration superscript “ (i) ” is omitted since all variables correspond to their values in iteration i):

$$P_{prior}(I_l, I_r) = \frac{h(I_l, I_r)}{\sum_{l,r} h(I_l, I_r)}, \quad (5.1)$$

where I_l, I_r are respectively the intensities of the pixels $l(i, j) \in L$ and the corresponding pixel $r(i, j - D(l)) \in R$. Prior probabilities are computed using $h()$, the 2D histogram of all the corresponding pixel intensities.

Having defined the prior probability, computation of the cost matrix using MI (the negative of the MI measure is used as the cost) is performed as follows:

$$W_L(x, y) = L(x_{min} : x_{max}, y_{min} : y_{max}), \quad (5.2)$$

$$x_{min} = x - \delta x_l - \omega, \quad (5.3)$$

$$x_{max} = x + \delta x_r + \omega, \quad (5.4)$$

$$y_{min} = y - \delta y, \quad (5.5)$$

$$y_{max} = y + \delta y, \quad (5.6)$$

where δx_l and δx_r are distances to the border of the segment which the current pixel (x, y) belongs to; and the window is enlarged by ω ; the assumed thickness of discontinuity at the images on the segment border (Figure 5.3). δy similarly provides the window size in vertical direction, and it is currently an empirically determined parameter ($\delta y \leq 4$ pixels for the Middlebury database). The segment borders in the vertical direction are not considered since the segment plane may not be a fronto-planar surface and will confuse cost calculation. The same window is applied to the right image by moving the window for each candidate disparity d :

$$W_R(x, y) = R(x_{min} - d : x_{max} - d, y_{min} : y_{max}). \quad (5.7)$$

Given two windows, the MI measure between them is computed using the segment information and the prior probabilities as:

$$M(W_L, W_R, S, P_{prior}) = \sum_W P(I_l, I_r) \ln \frac{P(I_l, I_r)}{P(I_l)P(I_r)}, \quad (5.8)$$

where joint probabilities are computed using the adaptive correlation surface, and the prior probabilities are incorporated, just like Fookes did [34], as follows:

$$P(I_l, I_r) = \lambda P_{window}(I_l, I_r) + (1 - \lambda) P_{prior}(I_l, I_r). \quad (5.9)$$

The correlation surface used in finding the joint probability is another key contribution of the thesis for the MI cost calculation, where the joint histogram is derived by

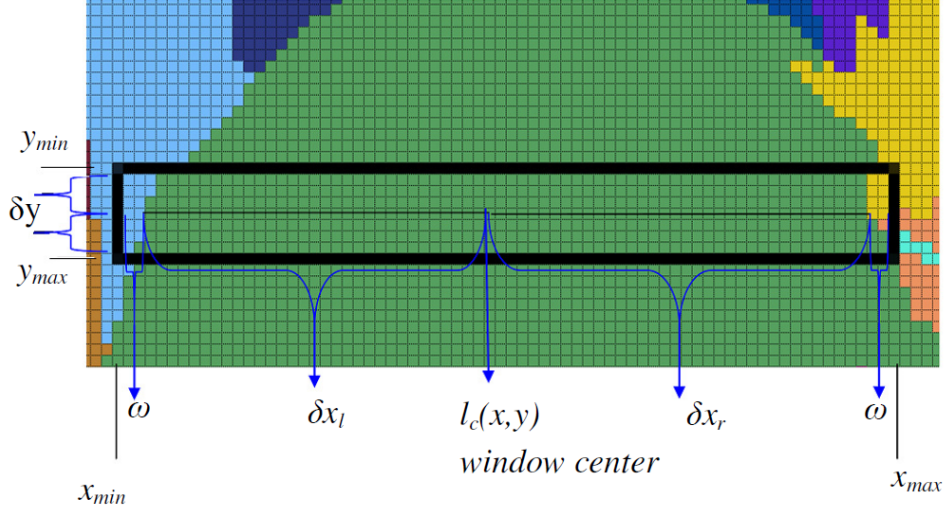


Figure 5.3: Adaptive window calculation.

considering pixels within the current segment in the window and the pixels nearby the edge of the segment as:

$$P_{window}(I_l, I_r, S) = \frac{h_w(I_l, I_r, S)}{\sum_w h_w(I_l, I_r, S)}, \quad (5.10)$$

$$h_w(I_l, I_r, S) = \sum_w T(I_l, I_r, S), \quad (5.11)$$

$$(5.12)$$

where the $T()$ function is defined as follows:

$$T(I_l, I_r, S) = \begin{cases} k & \text{if } S(l) = S(l_c) \ \& \ L1 > \omega \\ k - \frac{k}{\exp(L1)} & \text{elif } S(l) = S(l_c) \ \& \ L1 \leq \omega \\ \frac{k}{\exp(L1)} & \text{elif } L1 \leq \omega \\ 0 & \text{otherwise} \end{cases} \quad (5.13)$$

where $L1 = \|l - S(l_c)\|$ is the L1 distance between the neighbor pixel l and the border of the segment that the current pixel being matched l_c belongs to; and ω is the assumed thickness of the segment border as defined in Equation 5.2.

The usage of L1 distance (Figure 5.4) in Eq. 5.13 incorporates the pixels near the segment borders to MI calculation with some penalty due to possible occlusions around borders and this way, it was possible to consider both the segment and the edges excluding other segments within the rectangular window in MI measure computation.

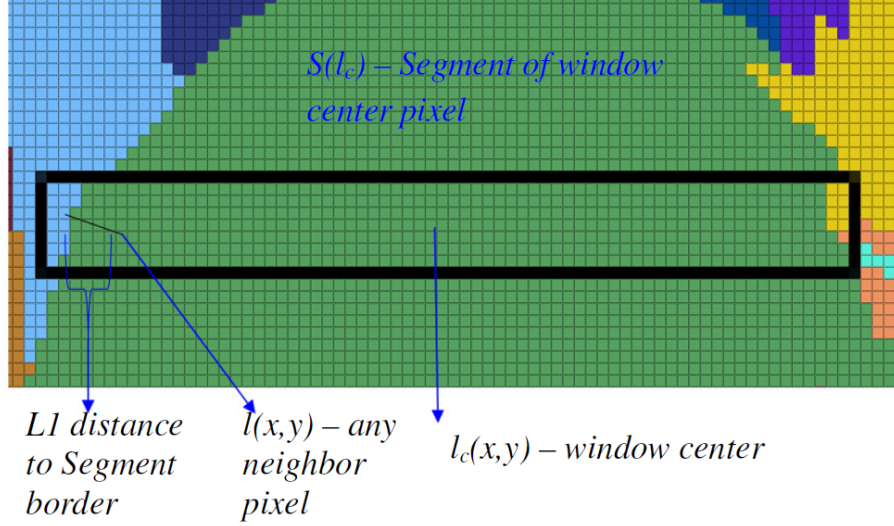


Figure 5.4: Adaptive MI computation surface using segmentation.

5.1.3 Adaptive Cost Aggregation

In this step, the major concern is to detect, revise and reduce the *un-confident* cost measures from the previous step that cause a majority of wrong disparities in the WTA step.

The cost confidences are used to detect *un-confident* costs computed for a pair of corresponding pixels and for this purpose, a modified version of the confidence measure that was used in [99] is proposed:

$$Conf(x, y) = \min \left(\frac{|c_1 - c_2|}{|c_1|}, \rho \right), \quad (5.14)$$

where c_1 is the minimum cost within the disparity range $[0..d_{max}]$, and c_2 is the second minimum cost. The ratio of the minimum and the second minimum cost value margin to the minimum cost value is used as the confidence measure and the obtained values are truncated with respect to a predetermined value ρ . This way, higher confidence values are prevented from dominating the cost aggregation step. See Figure 5.5 for the confidence values on an example.

Cost aggregation is performed by visiting all the pixels p in the initially computed cost matrix $C^{(i)}(p, d)$ (see Algorithm 1) and aggregating the costs according to the follow-



Figure 5.5: Cost Confidences for Tsukuba image pair (scaled and truncated to $[0..255]$ range for the sake of visibility).

ing weights within a local neighborhood for all the disparities d in range $[0..d_{max}]$:

$$C_{agg}^{(i)}(p, d) = \sum_{q \in w_p} w(p, q) C^{(i)}(q, d), \quad (5.15)$$

where w_p is a square support window of size $(2b + 1) \times (2b + 1)$ (b : half window size) where the costs of the same disparity in the neighborhood are aggregated by a weighting mechanism incorporating the current segmentation effectively as:

$$w(p, q) = \begin{cases} Conf^{(i)}(q) & \text{if } S^{(i)}(p) = S^{(i)}(q) \\ Conf^{(i)}(q) \exp\left(-\left(\frac{SD(p, q)}{\lambda_{SD}} + \frac{DD(p, q)}{\lambda_{DD}}\right)\right) & \text{if } S^{(i)}(p) \neq S^{(i)}(q) \end{cases} \quad (5.16)$$

where $Conf^{(i)}(q)$ is the confidence of aggregating pixel q (see Eqn. 5.14); $SD(p, q)$ is the spatial distance between pixels p and q ; $DD(p, q)$ is the WTA (winner takes all) disparity distance of initial costs $C^{(i)}$ between pixels p and q ; λ_{SD} and λ_{DD} are the designated scaling constants for the spatial distance and the disparity distance.

To sum up, the proposed scheme aggregates costs of neighboring pixels within the same segment according to the confidences of the pixels and penalizes aggregating weights according to the spatial distance and the WTA disparity distances of initial costs for the pixels in the other segments around.

5.1.4 Computation of Disparity Planes

In this step, the main idea is to fit planes to the segments using disparities of the confident pixels. Algorithm 2 shows the major steps of the method proposed in this step. The inputs to the algorithm are the segmentation for current iteration $S^{(i)}$ and the aggregated cost matrix $C_{agg}^{(i)}$ computed as described in the previous step (refer to Section 5.1.3). Below, each step of the algorithm is described in detail:

Algorithm 2 Computation of disparity planes.

Inputs: $S^{(i)}$: Segments of current iteration, ($i \in [0, N]$: iteration)
 $C_{agg}^{(i)}$: Aggregated cost matrix, (refer to Section 5.1.3)

Outputs: $S_{final}^{(i)}$: Revised segmentation,
 $D_{final}^{(i)}$: Disparity map computed from the fitted planes

- 1: $D_{agg}^{(i)} \leftarrow$ WTA disparity map corresponding to $C_{agg}^{(i)}$ aggregated cost matrix
- 2: $D_{aggsub}^{(i)} \leftarrow D_{agg}^{(i)} + f(C_{agg}^{(i)})$ //estimate subpixel disparities - See Equation 5.18
- 3: $D_{aggsub,m}^{(i)} \leftarrow med_{W_m}(D_{aggsub}^{(i)})$ //perform 3x3 median filter to subpixel disparities
- 4: $Conf_{agg}^{(i)} \leftarrow$ Compute confidences for aggregated cost matrix $C_{agg}^{(i)}$ // See Equation 5.14
- 5: $(S_{split}^{(i)}, P_{split}^{(i)}) \leftarrow$ Perform iterative segment splitting step
using $(D_{aggsub,m}^{(i)}, S^{(i)}, Conf_{agg}^{(i)})$ // See Algorithm 3
- 6: $(S_{final}^{(i)}, P_{final}^{(i)}, D_{final}^{(i)}) \leftarrow$ Perform segment merging & finalization step
using $(D_{aggsub,m}^{(i)}, S_{split}^{(i)}, P_{split}^{(i)}, Conf_{agg}^{(i)})$ // See Algorithm 4
- 7: **return** $(S_{final}^{(i)}, D_{final}^{(i)})$

1. **WTA of aggregated costs:** The Winner Takes All (WTA) disparities $D_{agg}^{(i)}$ corresponding to the aggregated cost matrix $C_{agg}^{(i)}$ (see Equation 5.15) are computed as the first step.
2. **Subpixel disparity computation:** As the next step, the subpixel disparity estimates ($D_{aggsub}^{(i)}$) are computed using the aggregated cost matrix $C_{agg}^{(i)}$ by finding the minimum of the parabola fitted to the minimum cost disparity in $D_{agg}^{(i)}$ and the two neighboring cost values.

Let us use d to denote the integer disparity of minimum cost (the WTA disparity) in the cost matrix C within the disparity range d_0 to d_{max} :

$$d = \arg \min_{d_i \in \{d_0, \dots, d_{max}\}} C(d_i), \quad (5.17)$$

the subpixel disparity estimate is defined as:

$$d_{sub} = d + f(C(d-1), C(d), C(d+1)), \quad (5.18)$$

where f is the function for parabolic interpolation:

$$f = \frac{C(d-1) - C(d+1)}{2(C(d-1) - 2C(d) + C(d+1))}. \quad (5.19)$$

This yields floating-point disparities, having more smooth transitions within a segment.

3. **Median filtering:** Next, a median filter is applied to the subpixel disparities so that outliers are eliminated (yielding $D_{aggsub,m}^{(i)}$), if there is any. This step may affect disparity plane fitting.
4. **Compute confidences of the aggregated cost matrix:** Confidences ($Conf_{agg}^{(i)}$) of the aggregated cost matrix $C_{agg}^{(i)}$ are computed in order to rely on only the confident pixels in the following steps.
5. **Iterative segment splitting:** In the next step, the confident disparities within the segments are fitted planes, and the outlier disparities are re-evaluated by splitting the segments. The step is iteratively re-applied for the new segmentation map ($S_{split}^{(i)}$) until no further segment splitting can occur - see Section 5.1.4.1.
6. **Segment merging & finalization:** Finally, the split segments ($S_{split}^{(i)}$) and the corresponding disparity planes ($P_{split}^{(i)}$) are inspected for finalization by (i) merging neighboring segments that are co-planar at the same disparity level and (ii) refining unstable segments that may be generated during the segment splitting operations or which may have inadequate number of confident pixels to be able compute a disparity plane (*i.e.*, when the number of pixels is less than 4). The final disparity map is computed from the resultant segmentation and the corresponding segment plane equations - see Section 5.1.4.2 for the details of this step.

5.1.4.1 Iterative Plane Fitting and Segment Splitting Step

This step revises the input segmentation according to the confident outlier disparities within the corresponding segments once plane fitting is performed - see Algorithm 3. This way, the dependency of the performance of the algorithm on the initial segmentation is reduced.

Algorithm 3 Iterative plane fitting & segment splitting.

Inputs: D : Input disparity map
 S : Initial segmentation map
 $Conf$: Confidences of the disparities

Outputs: S : Revised segmentation,
 $Plane_S$: Fitted disparity plane equations for each segment

- 1: **repeat**
- 2: **for all** segment $s \in S$ **do**
- 3: **repeat**
- 4: $Cloud \leftarrow \{(p, d) \mid \forall p \in s, d = D(p), Conf(p) > \tau_{ic}\}$ //extract the point cloud of confident pixels p in s
- 5: **if** $size(Cloud) < 4$ or $size(Cloud)/size(s) < \tau_{ir}$ **then**
- 6: $stable(s) \leftarrow FALSE$
- 7: **else**
- 8: $stable(s) \leftarrow TRUE$
- 9: Fit plane $Plane_S(s)$ to $Cloud$ using RANSAC
- 10: $OutCloud \leftarrow \{(p, d) \mid \forall p \in s : d = D(p), |d - Plane_S(s, p)| > \tau_{od}\}$ //extract outlier point cloud of disparities according to fitted plane
- 11: **if** $(size(OutCloud) > \tau_{os})$ **then**
- 12: $OutCloud2 \leftarrow \{(p, d) \mid \forall (p, d) \in OutCloud, Conf(p) > \tau_{oc}\}$
- 13: Split segment s for all the connected subsets of $OutCloud2$
- 14: Append splitted segments to segments list S
- 15: **end if**
- 16: **end if**
- 17: **until** segment s is not splitted
- 18: **end for**
- 19: Re-compute segments map S //since new segments can break bigger segments to two or more disconnected sub-segments
- 20: **until** no segment splitting performed
- 21: **return** $(S, Plane_S)$

In the algorithm, the disparity plane for each segment in the current segmentation map S is computed from the confident disparities only. Plane fitting is performed using RANSAC (RANDOM SAMPLE CONSENSUS) [104]. Next, the outlier disparities are analyzed in each plane fitted segment and checked whether they constitute connected regions of a significant size; if so, the outlier region is split out. This operation is performed iteratively until the segment is no longer split. Finally, in the outer loop, the segmentation map is re-computed and the above described steps are re-applied since the segments might break more than once. This way, the segmentations and the plane fits are revised iteratively until no further segment splits can be performed. The algorithm returns the revised segmentation map along with the fitted disparity plane equations to the segments.

Algorithm 3 makes use of several thresholds effectively to perform its goal. τ_{ic} is the disparity confidence value threshold to be able to construct the initial point cloud of disparities from the segment disparities. τ_{ir} is the stable segment ratio threshold which determines whether a segment is stable or not by checking the ratio of the size of the confident disparities point cloud and the segment size. If the size of the cloud is also smaller than 4 pixels then it will not be possible to fit a plane and therefore such segments are marked as unstable and left for the next step for correction. τ_{od} threshold is used for determining the outlier points of the fitted plane which is designed to be greater than the RANSAC distance threshold parameter used for plane fitting. τ_{os} determines the minimum size of the outlier point cloud of disparities to continue splitting operation and τ_{oc} is the confidence threshold for the outlier points which are to be selected for segment splitting. Therefore, to be able to create a new segment by splitting from the original segment, a connected region whose size is greater than a designated threshold should be available.

5.1.4.2 Segment Merging and Finalization of Disparity Planes Step

This step computes the final segmentation and the disparity map of the scene - see Algorithm 4. The step is composed of three phases: In the first phase, all the stable segments are inspected along with their neighbors and merged if they are coplanar. The coplanarity of two disparity planes are defined as:

$$cop(s, s') = \begin{cases} 1, & \text{if } (\alpha(s, s') < \tau_\alpha) \text{ and } (\|s - s'\| < \tau_{pd}) \\ 0, & \text{otherwise} \end{cases} \quad (5.20)$$

which checks if the normal of the planes are parallel (the difference in their normals α is smaller than a threshold τ_α) and if they are at the same disparity level (the distance between planes is smaller than a threshold τ_{pd}). Moreover, the segments that were marked as unstable are re-evaluated by decrementing the confidence threshold iteratively.

In the second phase, the equations for the disparity planes are recomputed for the merged segments and finally, in the third phase, the disparity of each pixel is computed from the disparity plane equations, except for the still-unstable segments where the input disparity map is accepted as is for those pixels.

Algorithm 4 Segment Merging and Finalization

```

Inputs:    $D$            : Input disparity map
            $S$            : Input segmentation map
            $Conf$         : Confidences of the disparities
            $Plane_S$      : Fitted disparity planes for segments
Outputs:  $S$            : Revised segmentation by merged segments
            $Plane_S$      : Fitted disparity plane equations for each segment
            $D_{Plane}$      : Disparity map computed from fitted disparity plane equations

// Phase 1: merge stable segments & retry for unstable segments
for all segment  $s \in S$  do
  if  $stable(s) = \text{TRUE}$  /* See Algorithm 3 */ then
    for all  $s' \in \Omega(s)$  { $\Omega(s)$ : neighboring segments of  $s$ } do
      if  $cop(s, s') = \text{TRUE}$  /* Coplanar planes - See Equation 5.20 */ then
         $s \leftarrow Merge(s, s')$  //Segments are merged
         $merged(s) \leftarrow \text{TRUE}$ 
         $S \leftarrow S - s'$  //remove  $s'$  from set  $S$  since it is merged with  $s$ 
      end if
    end for
  else if  $stable(s) = \text{FALSE}$  then
    repeat
       $\tau_{ic2} \leftarrow \tau_{ic} * \lambda_{\tau_{ic}}$  //Decrement confidence threshold by  $\lambda_{\tau_{ic}} \in (0, 1)$ 
      Re-compute plane fitting  $Plane_S(s)$  for  $Cloud$  where:  $Cloud \leftarrow \{(p, d) \mid p \in s, d = D(p), Conf(p) > \tau_{ic2}\}$ 
      if  $size(Cloud) < 4$  or  $size(Cloud)/size(s) < \tau_{ir}$  then
         $stable(s) \leftarrow \text{FALSE}$ 
      else
         $stable(s) \leftarrow \text{TRUE}$ 
      end if
       $decrement(\lambda_{\tau_{ic}}, \gamma)$  //Decrement  $\lambda_{\tau_{ic}}$  by  $\gamma \in (0, 1)$ 
    until ( $\lambda_{\tau_{ic2}} = 0$ ) or ( $stable(s) = \text{TRUE}$ )
    end if
  end for

// Phase 2: recompute plane fits for merged segments
for all (segment  $s \in S$ ) and ( $stable(s) = \text{TRUE}$ ) and ( $merged(s) = \text{TRUE}$ ) do
  Re-compute plane fitting  $Plane_S(s)$  for  $Cloud$  where  $Cloud \leftarrow \{(p, d) \mid p \in s, d = D(p), Conf(p) > \tau_c\}$ 
end for

//Phase 3: compute final disparity map
for all segment  $s \in S$  do
  if  $stable(s) = \text{TRUE}$  then
    Compute disparities  $D_{Plane(s)}$  for pixels in segment  $s$  using fitted plane equation  $Plane(s)$ 
  else if  $stable(s) = \text{FALSE}$  then
    Set disparities  $D_{Plane(s)}$  for pixels in segment  $s$  to original disparities in input  $D(s)$  // for segments which are still unstable
  end if
end for
return ( $S, Plane_S, D_{Plane}$ )

```

5.1.5 Iterative Refinement

As a result of the previous steps, an updated segmentation map is generated along with the corresponding disparity map. With this segmentation map as the new segmentation map, the same steps are repeated as a new iteration. The new iteration uses the current disparity map for better estimation of the joint prior probabilities (see Equations 5.1 and 5.9) along with some adjustments that can be performed with such a priori data. Therefore, this step starts with the segmentation $S^{(i+1)}$ set to resul-

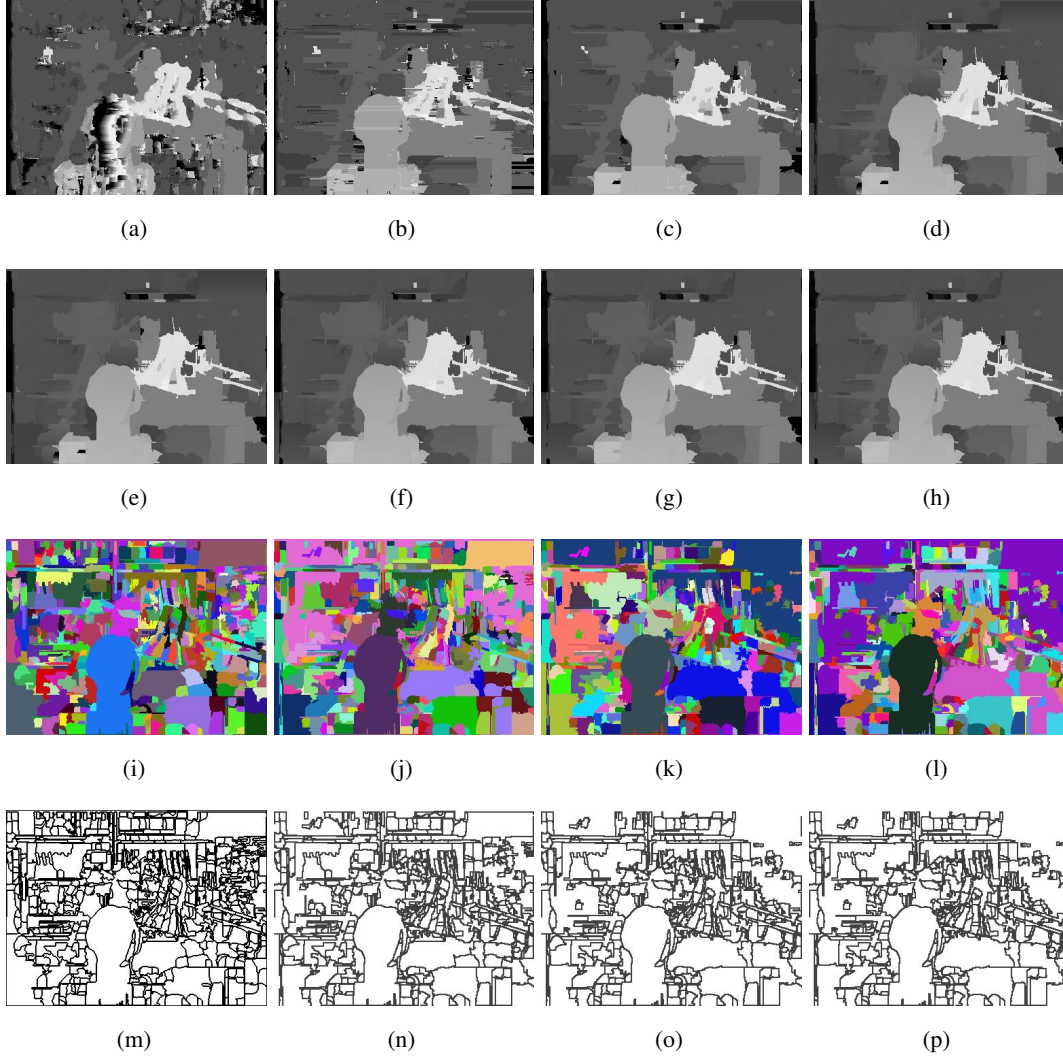


Figure 5.6: The intermediate steps of the proposed method. (a) WTA disparities of raw costs with No-Adaptive Windowing - 1st iteration (MI(woPR)). (b) WTA disparities of raw costs with Adaptive Windowing - 1st iteration. (c) WTA disparities after adaptive cost aggregation - 1st iteration. (d) Plane fitted disparities - 1st iteration. (e-h) Resultant plane fitted disparities for iterations 1-4. (i) The initial segmentation of the left image. (j-l) The input segmentations to iterations 2-4 (after segment break & merge steps are applied in the previous iteration). (m-p) Edge map of the corresponding input segmentation at each iteration. [Best viewed in color]

tant segmentation of the current iteration $S_{final}^{(i)}$, and disparities $D^{(i+1)}$ set to resultant disparity map $D_{final}^{(i)}$ for the current iteration (see Algorithm 2).

Moreover, for iterations after the first iteration (*i.e.*, $i \geq 1$), there is now the opportunity to adjust the adaptive window calculation method presented in Equation 5.2,

where x_{max} value can be moved back in the direction of the center pixel if the right neighboring segment has a disparity level greater than the current segment, and the shifting value is determined by the difference in the disparity levels of segments. This enables to not fetch the pixels in the right segment when the same window is applied to right image for correspondence matching.

5.2 Experiments and Results

The performance evaluation of the proposed method is performed using the two datasets that were generated in the scope of this thesis; *i.e.*, the Dataset #1 - the Synthetically Altered Middlebury Stereo Evaluation Dataset and Dataset #2 - the Kinect Dataset and the performance evaluation methods proposed for the datasets (see Chapter 3 for the description of datasets and the performance evaluation methods). The following sections include the experiments and results over these datasets along with the evaluation of the results. Besides, in Appendix D a detailed analysis of the parameters of the proposed method is provided.

5.2.1 Performance Evaluation of the Proposed Method with State of the Art Similarity Measures

In this section, the 1st iteration "WTA" performance of the proposed method is evaluated by comparing the performance metrics of the similarity measures that were already evaluated in Chapter 4. Only the first step, *i.e.*, negative MI costs computed by the adaptive windowing algorithm described in Section 5.1.2 is compared to the similarity measures for the WTA disparity maps generated in the experiments.

5.2.1.1 Results on Dataset #1

Figures 5.7 and 5.8 show the average RMS and Bad pixel percentage errors for the "all" regions of performance evaluation. The WTA performances of the adaptive windowing algorithm for the three window sizes are compared with the selected similarity measures with the leading performances in the experiments held in Section 4.1.1.

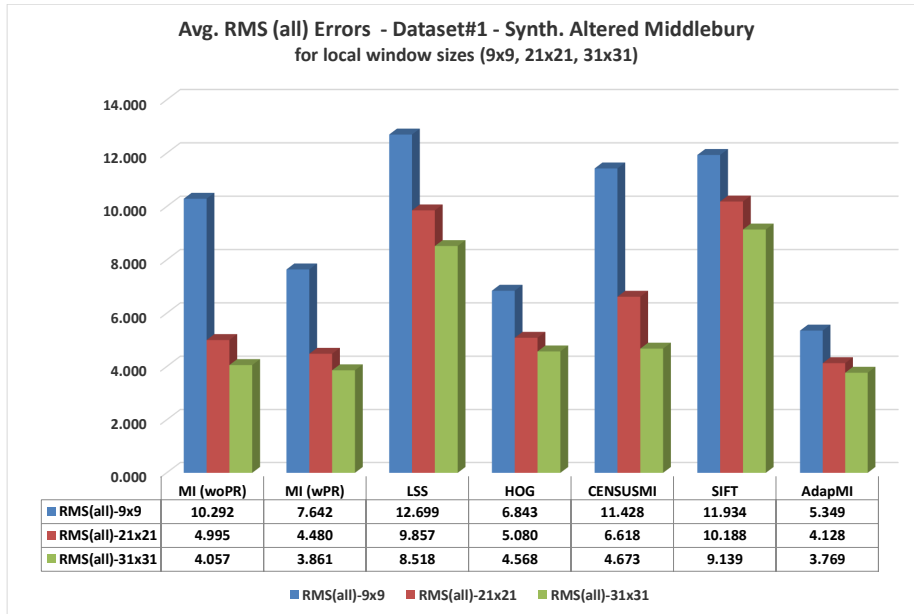


Figure 5.7: Average RMS (all) errors of “WTA” performances using the Dataset #1 for the Performance Evaluation of Adaptive Windowing Algorithm (ADAPMI) to state of the art similarity measures.

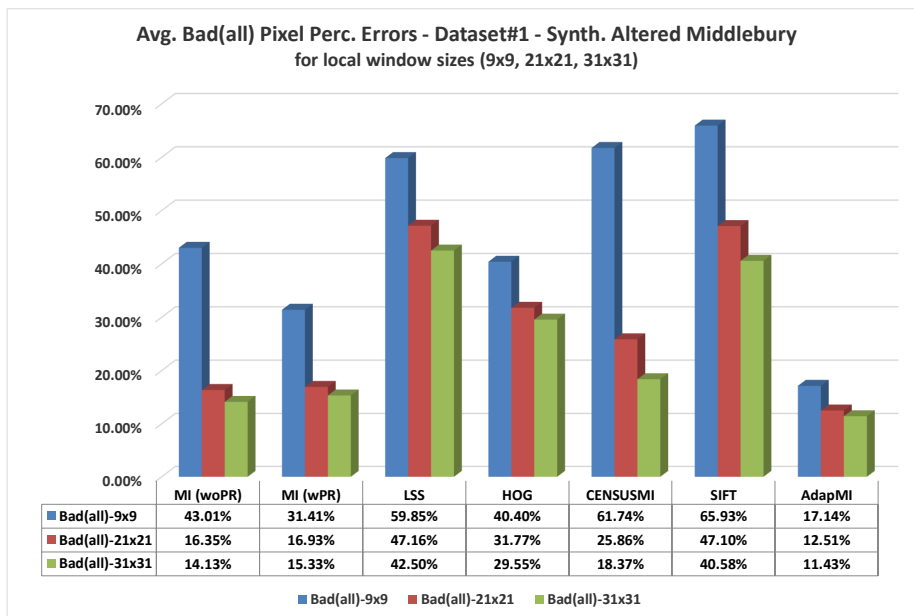


Figure 5.8: Average Bad (all) pixels percentage errors of "WTA" performances using the Dataset #1 for the Performance Evaluation of Adaptive Windowing Algorithm (ADAPMI) to state of the art similarity measures.

Figures C.1, C.2, C.3 and C.4 in Appendix C show the RMS and Bad pixel performances for each image separately, *i.e.*, Tsukuba, Venus, Teddy and Cones. The

Figure 5.9 shows sample visual results of the “WTA” disparity maps obtained from the Tsukuba image. The whole table of performance statistics computed are provided in Appendix C in Table C.1.




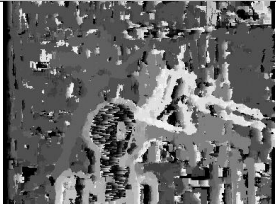
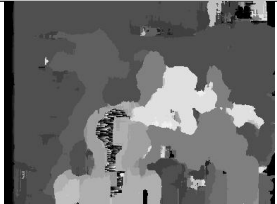

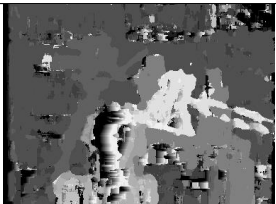
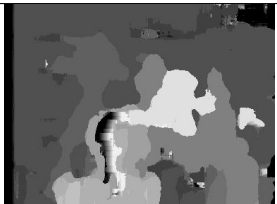

	9x9	21x21	31x31
ADAPMI			
MI (woPR)			
MI (wPR)			

Figure 5.9: Sample visual results of the WTA disparity results of the Adaptive Windowing algorithm (ADAPMI) and the leading similarity measures for the synthetically altered Tsukuba image pair in Dataset #1, using different window sizes 9x9, 21x21 and 31x31 pixels.

As can be observed from the results, the proposed method outperforms in this dataset in all the tested similarity measures even at the initial phase of the computation, *i.e.*, computing the cost matrix using the developed adaptive windowing algorithm for MI computation.

5.2.1.2 Results on Dataset #2

In this section, the results obtained from the selected image pairs in Dataset #2 are provided similarly for the WTA disparities of the tested similarity measures and the proposed method’s initial step - the adaptive windowing algorithm. The same set of images are used with the ones used in Section 4.2.

Figure 5.10 shows average *Percentage Good Depth* and *Percentage Total Coverage* metrics, and Table C.2 provides all the experiment results over the dataset. Figure 5.11 shows visual results of generated WTA disparities for the dataset image Kinect02 for leading similarity measures tested and the adaptive windowing algorithm of the proposed method. See Figure 4.12 for the Kinect-generated depth data (note that brighter pixels have more depth in the depth image in contrast to disparity image results provided).

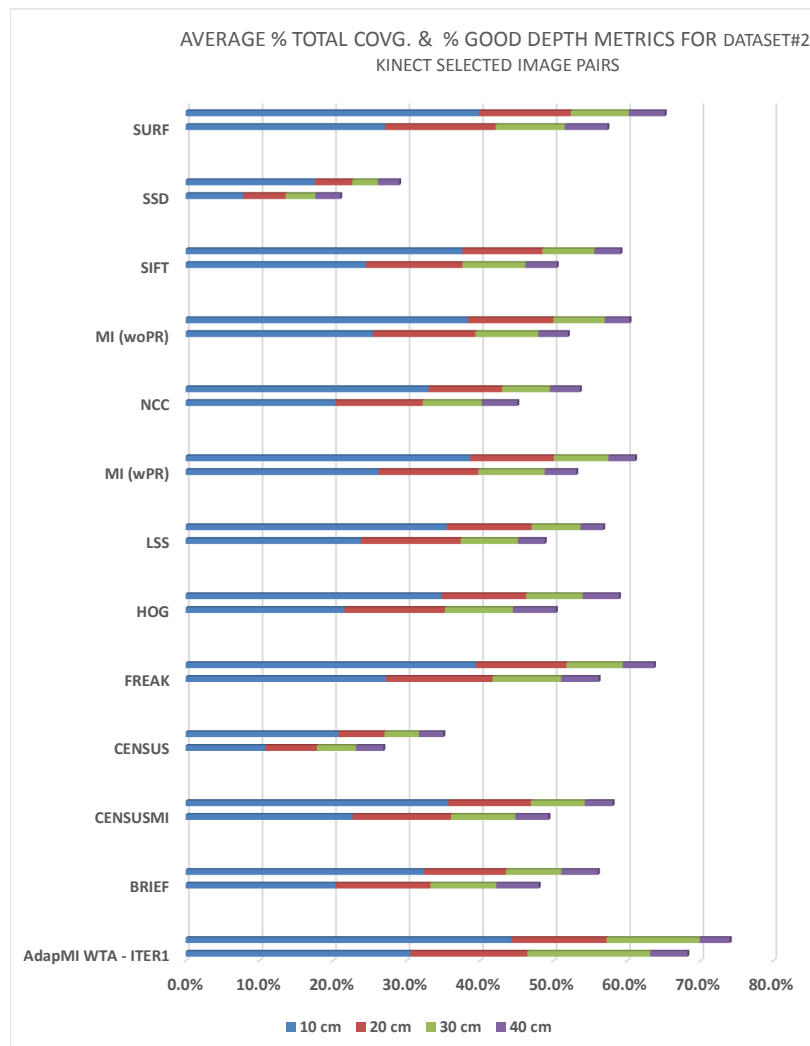


Figure 5.10: Average *Percentage Good Depth* and *Percentage Total Coverage* metrics computed for similarity measures for the dataset#2 selected image pairs and the Adaptive Windowing Algorithm (ADAPMI) of the proposed method - initial iteration.

As can be observed from these results, the proposed method outperforms in the Kinect

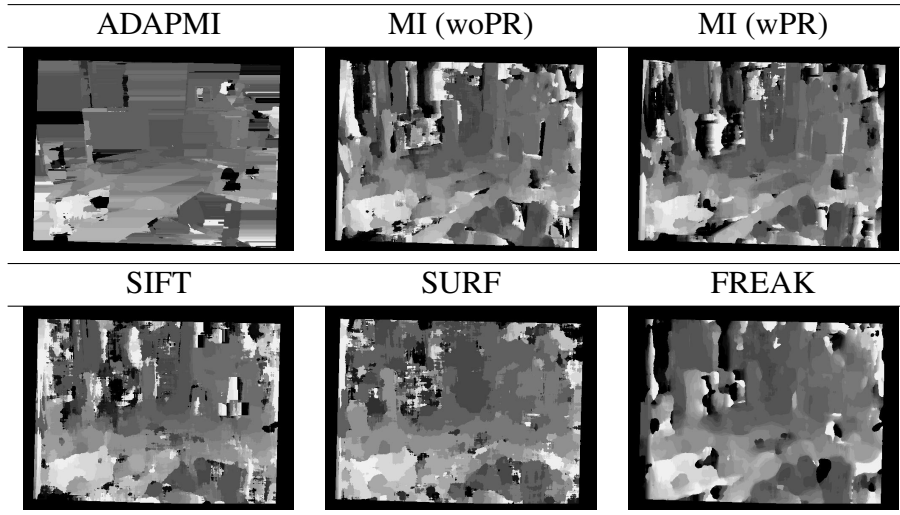


Figure 5.11: Sample visual results of computed WTA disparities of the Adaptive Windowing Algorithm (ADAPMI) of the proposed method and the similarity measures for Img#2 in Dataset #2 (local window size=31x31)

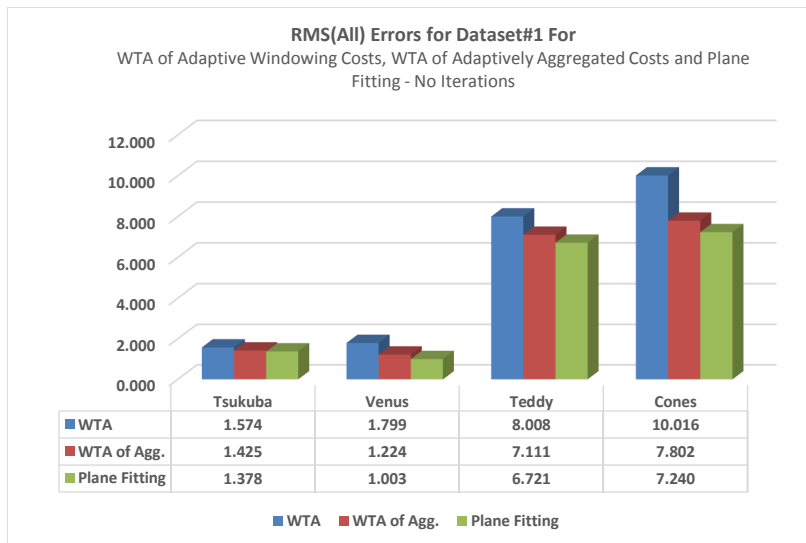
dataset also, when compared using all the tested similarity measures, even if method is at the initial phase of the computation, *i.e.*, computing the cost matrix using the developed adaptive windowing algorithm for MI computation.

5.2.2 Performance Evaluation of Cost Aggregation and Plane Fitting Steps of the Proposed Method

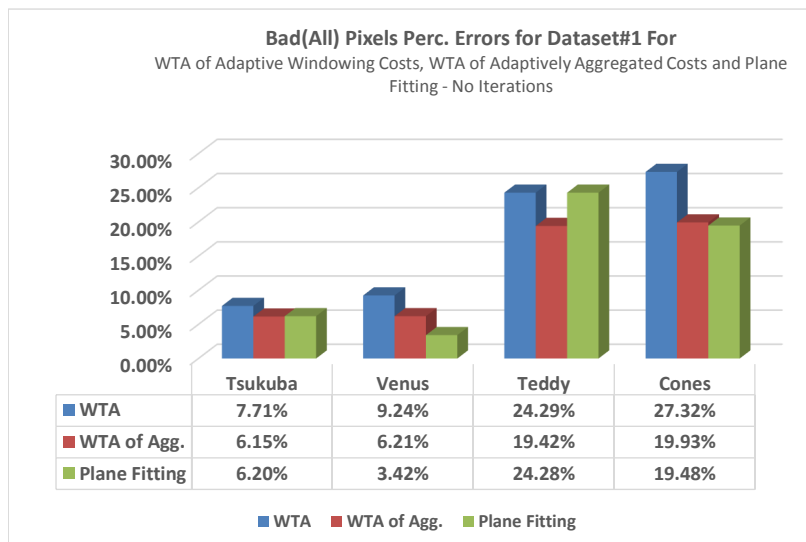
In this section, the WTA results of the initial step of the method (the adaptive windowing algorithm) are taken as the baseline and are used to show the enhancement in performances by adaptive cost aggregation (Section 5.1.3) and using disparity planes (Section 5.1.4), *i.e.*, with the subsequent steps of the method without any iterative refinements yet.

The results in Figure 5.12 show that WTA with cost aggregation improves the disparity estimation of the WTA with adaptive windowing method. Plane fitting, however, improves the RMS values of the estimated disparities in all image pairs whereas the bad matching percentage is almost the same with WTA with cost aggregation in Tsukuba image pair (composed of fronto-parallel surfaces) and worse in Teddy image pair (including curved surfaces mostly) whereas, for the Venus (totally composed of

planar surfaces), the bad matching percentage improves significantly.



(a)



(b)

Figure 5.12: Results on Dataset #1 - Synt. Altered Middlebury Images for WTA of Adaptive Windowing Costs, WTA of Adaptively Aggregated Costs and Plane Fitting.

5.2.3 Performance Evaluation of Iterative Refinement

Finally, the effect of iterative refinement on the estimated disparities is analyzed in this section. Figures 5.13 and 5.14 show the qualities of the estimated disparities in 10 iterations following Section 5.1.5. It is observed from the figures that the RMS and

bad matching percentage decrease drastically in the second iteration and the values more or less stabilize. This suggests that iterating twice over the disparity estimation steps as suggested in Section 5.1.5 is sufficient.

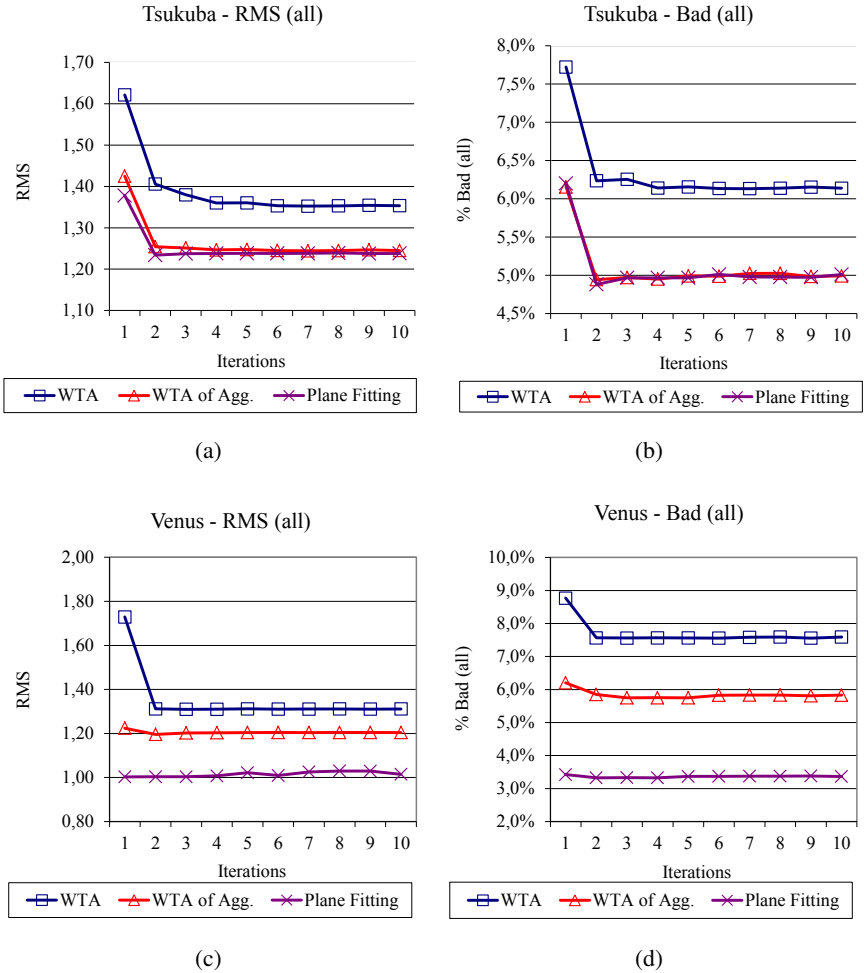


Figure 5.13: Effect of iterations on RMS and the percentage of bad pixels for “all” regions (a-b) Tsukuba, (c-d) Venus pairs.

Figure 5.6 in Section 5.1.5 shows the resultant WTA disparity maps of non-adaptive vs. adaptive windowing costs, the aggregated costs and plane fitted disparities as well as the resultant first 4 iterations of disparity maps, the segmentation maps computed for the Tsukuba image. Analyzing the results of all the image pairs in the dataset, it is concluded that two iterations is an ideal stop.

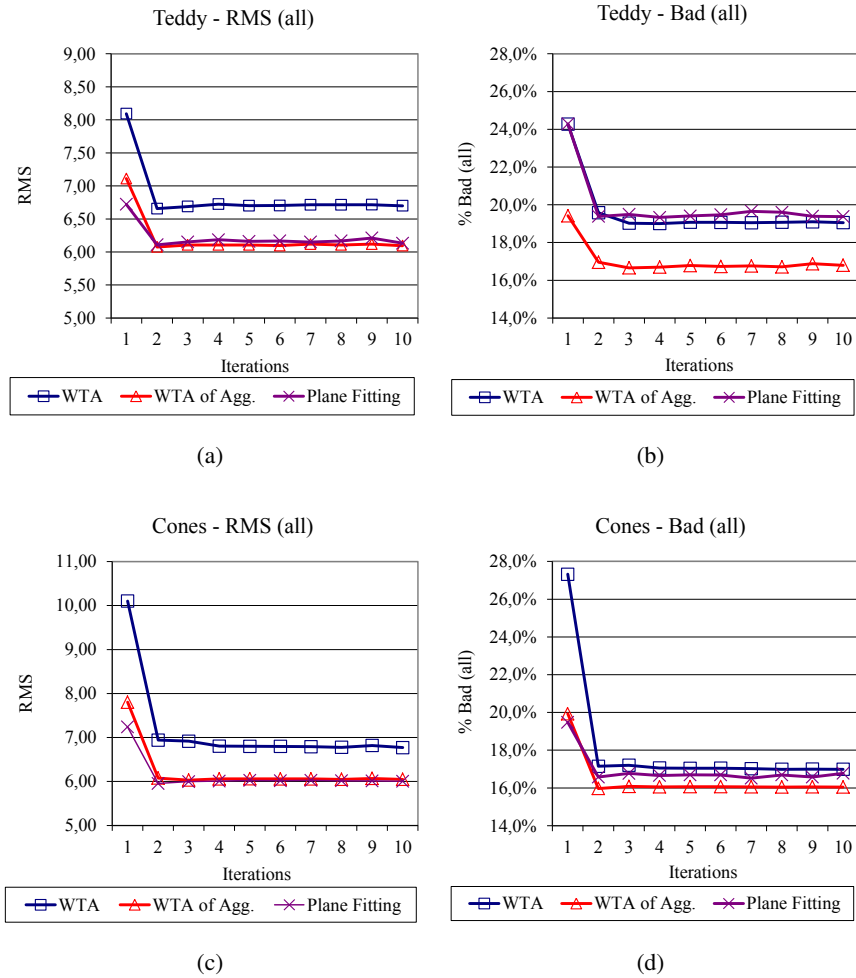


Figure 5.14: Effect of iterations on RMS and the percentage of bad pixels for “all” regions (a-b) Teddy and (c-d) Cones stereo pairs.

5.2.4 Performance Evaluation on Dataset #2- Kinect Dataset

In previous subsections, it was shown that the proposed method outperforms state of the art similarity measures with no iterations yet and that the subsequent steps (cost aggregation, plane fitting and iterative refinement) enhance the results significantly. In this section, the results obtained from the whole Dataset #2 are provided for the proposed method. The iterations are stopped after the 2nd iteration due to the conclusion given in Section 5.2.3 that two iterations are sufficient.

Table 5.2 provides the mean of the resultant depth maps generated by the proposed method when compared to the Kinect’s native depth map and Table 5.3 lists the stan-

standard deviation of the results over the mean performance values. The whole table of performance statistics computed are provided Table C.3 in Appendix C.

Table 5.2: Average Results on the 24 images in the Dataset #2- The Kinect Dataset for WTA, WTA of Agg. Costs and Plane Fitted Disparities vs. Kinect’s native depth in 2 Iterations.

Method	Metric	10cm	20cm	30 cm	40 cm	Total
WTA - Iter1	Perc. Good Depth	31%	15%	11%	5%	63%
	Perc. Total Covg.	43%	13%	9%	5%	69%
WTA - Iter2	Perc. Good Depth	33%	16%	11%	5%	65%
	Perc. Total Covg.	44%	13%	9%	5%	71%
Agg. - Iter1	Perc. Good Depth	35%	17%	11%	6%	68%
	Perc. Total Covg.	46%	14%	9%	5%	74%
Agg. - Iter2	Perc. Good Depth	36%	17%	11%	5%	69%
	Perc. Total Covg.	48%	14%	9%	5%	75%
PFIT - Iter1	Perc. Good Depth	39%	17%	9%	6%	71%
	Perc. Total Covg.	50%	14%	8%	5%	76%
PFIT - Iter2	Perc. Good Depth	41%	16%	9%	6%	72%
	Perc. Total Covg.	52%	13%	7%	5%	77%

Table 5.3: Standard Deviations of the Results on the 24 images in the Dataset #2 for WTA, WTA of Agg. Costs and Plane Fitted Disparities vs. Kinect native depth in 2 Iterations.

Method	Metric	10cm	20cm	30 cm	40 cm	Total
WTA - Iter1	Perc. Good Depth	10%	5%	6%	2%	23%
	Perc. Total Covg.	10%	4%	4%	2%	20%
WTA - Iter2	Perc. Good Depth	10%	5%	6%	2%	24%
	Perc. Total Covg.	10%	5%	4%	2%	20%
Agg. - Iter1	Perc. Good Depth	10%	6%	5%	2%	24%
	Perc. Total Covg.	10%	5%	4%	2%	20%
Agg. - Iter2	Perc. Good Depth	11%	5%	5%	3%	24%
	Perc. Total Covg.	9%	4%	4%	2%	19%
PFIT - Iter1	Perc. Good Depth	11%	6%	4%	3%	24%
	Perc. Total Covg.	10%	5%	3%	2%	20%
PFIT - Iter2	Perc. Good Depth	12%	6%	5%	3%	25%
	Perc. Total Covg.	10%	5%	3%	2%	20%

Figure 5.15 provides the same figures obtained in a bar chart graphics representation.

Figures 5.16 and 5.17 are sample visual results of computed disparity maps compared

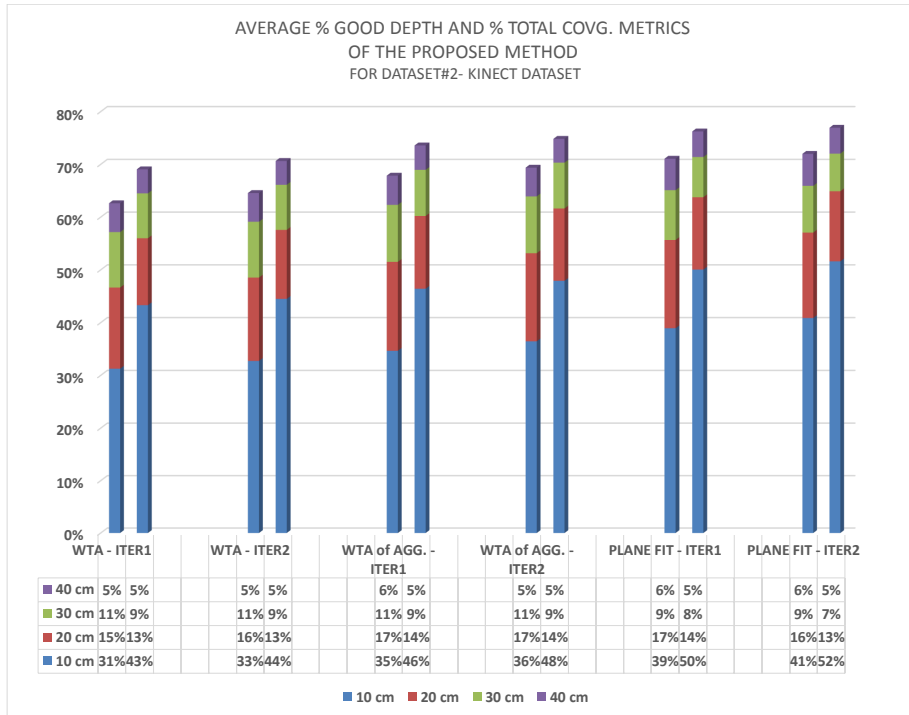


Figure 5.15: Depiction of Average Results on the 24 images in the Dataset #2- The Kinect Dataset for WTA, WTA of Agg. Costs and Plane Fitted Disparities vs. Kinect’s native depth in 2 Iterations

to the native depth map of Kinect for the two iterations performed.

Finally, Figure 5.18 shows a merged 3D rendered view of the Kinect’s native depth map and the proposed method’s final depth map for a sample (Img #1) in Dataset #2. As one can observe, Kinect’s native depth map is not successful in such cases, since the device can not generate depth on reflective surfaces when the sent infrared beams do not return back to the sensor. The figure shows that the proposed method’s depth generation method can be used to fill in empty depth information in the acquired scene.

To sum up, from both the statistical and visual evaluation, this section showed that depth map generated by the proposed method is comparable to Kinect native depth. Besides, the method can compute depth information on edges and non fronto-planar surfaces where Kinect’s depth generation fails. Therefore, the method can also be used in combination with Kinect to get a better coverage of the scene.

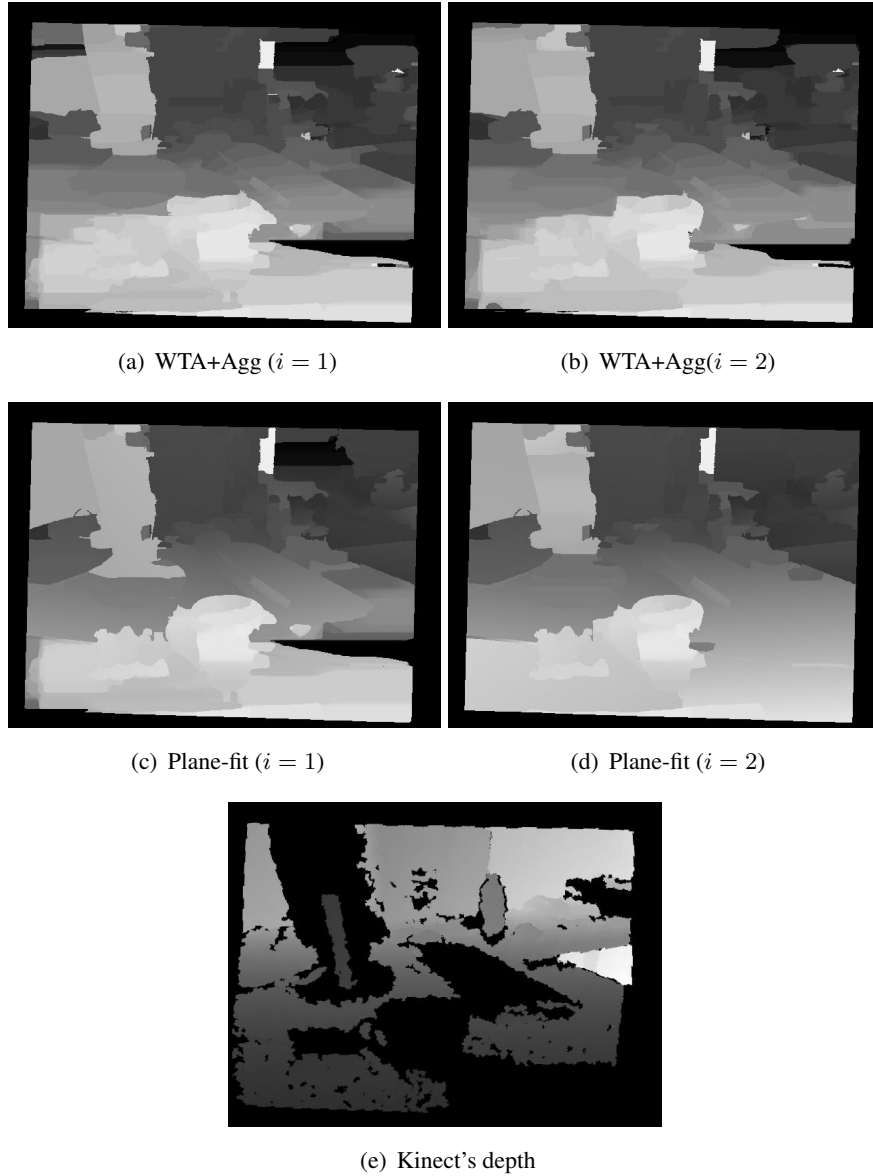


Figure 5.16: Sample visual results of computed disparity maps compared to native depth of Kinect, for Kinect01 image pair *1st row*: WTA disparity of aggregation results- 1st and 2nd iteration. *2nd row*: Plane fitting disparity results - 1st and 2nd iteration. *3rd row*: Kinect's native depth image (brighter pixels are farther)

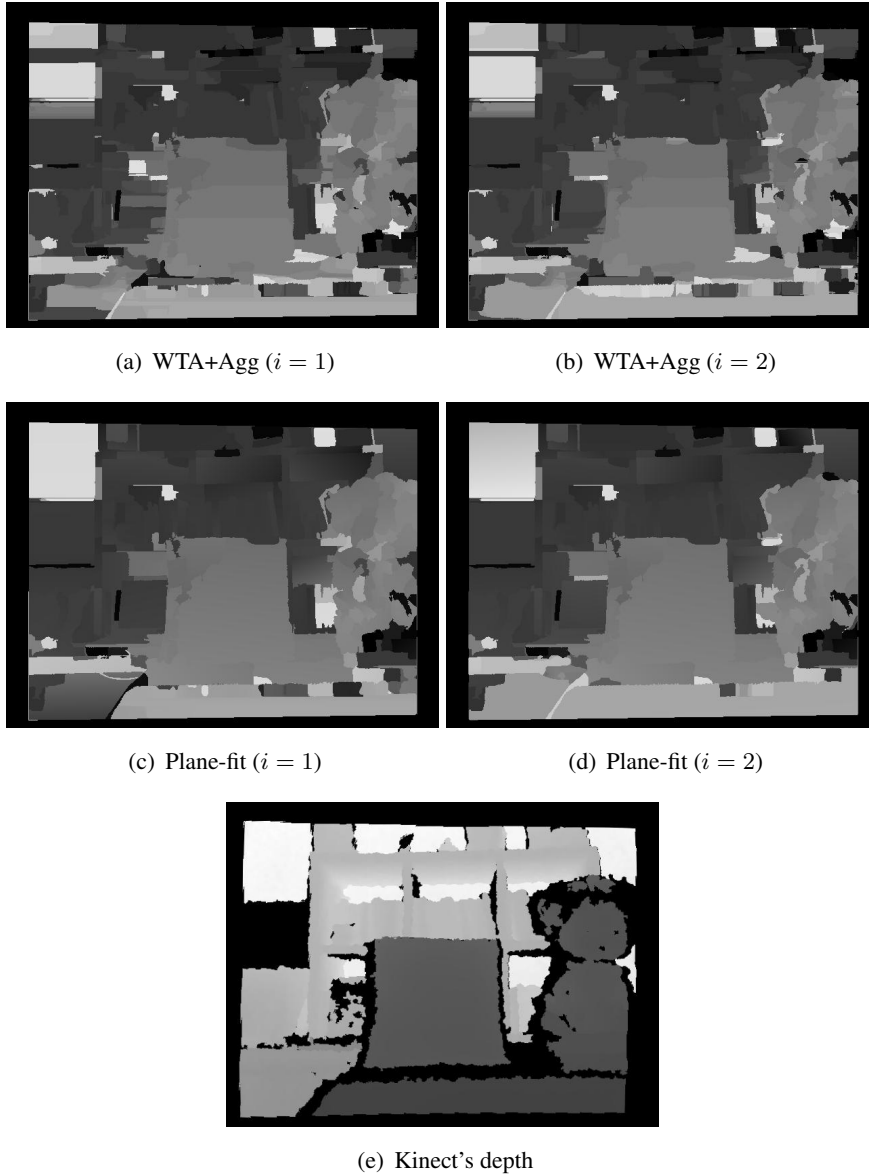


Figure 5.17: Sample visual results of computed disparity maps compared to native depth of Kinect, for Kinect10 image pair *1st row*: WTA disparity of aggregation results- 1st and 2nd iteration. *2nd row*: Plane fitting disparity results - 1st and 2nd iteration. *3rd row*: Kinect's native depth image (brighter pixels are farther)

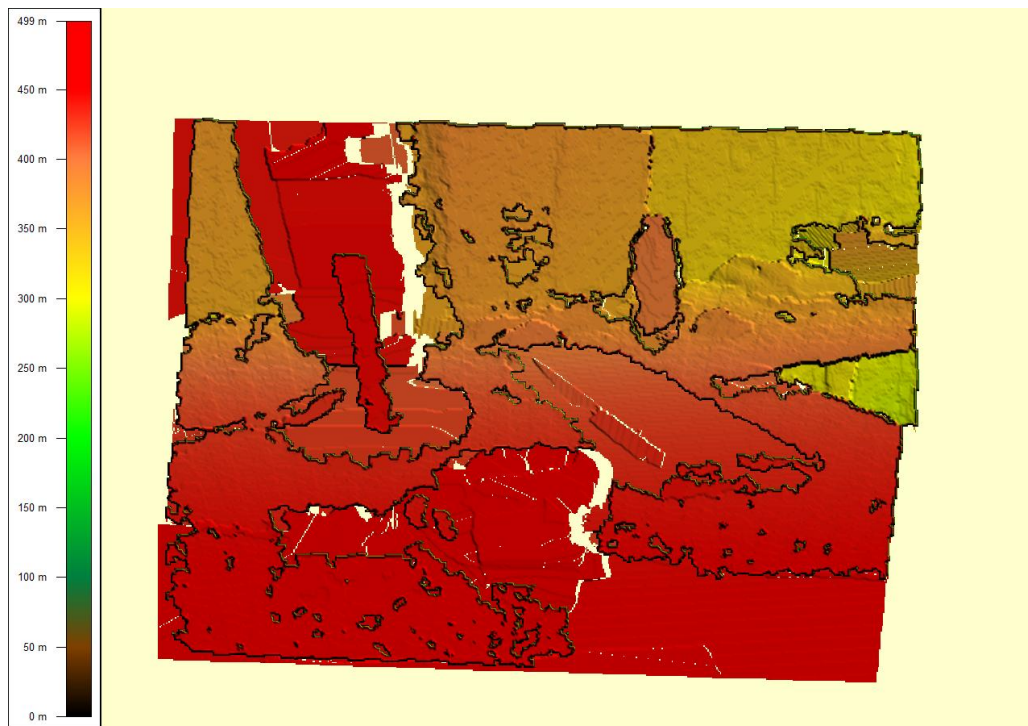


Figure 5.18: Merged 3D rendering of Kinect01 image's native depth map to proposed method's final depth map where invalid depth in the native depth map is filled with the proposed method's depth information.

CHAPTER 6

CONCLUSION

In this thesis, a novel dense multi-modal stereo-vision method is introduced which (i) is iterative, (ii) uses adaptive windowing and (iii) adaptive cost aggregation along with (iv) iteratively refined disparity plane fitting. The method uses mutual information as the basic similarity measure and was tested on multi-modal stereo images from two image datasets generated in the scope of the thesis; the synthetically altered image pairs from the Middlebury Stereo Evaluation Dataset, and our own dataset of Kinect Device infrared-visible camera image pairs. The datasets are also used for evaluating the state of the art methods in literature.

On these datasets, it is presented that (i) the proposed method improves the quality of existing MI formulation, (ii) the proposed method outperforms state of the art methods in literature, and (iii) the proposed method can provide depth comparable to the quality of Kinect depth data

The results show that significant increase in performance is achieved by the initial step of the proposed method, the adaptive windowing step, when compared to a non-adaptive local window MI calculation scheme as well as the other state of the art similarity measures in the literature for multi-modal stereo-vision. Moreover, the adaptively aggregated costs enhance the results while smoothing out the disparity maps whereas plane fitting enables to get more clean disparity maps although it depends on the current segmentation. This dependence is levitated by performing iterative segment splitting/merging over confident disparities and finally the whole method is re-applied in the next iteration where an initial disparity map is available now to incorporate more accurately the prior probabilities into joint probability calculation.

The results show that two iterations are sufficient to converge to reasonable results.

Regarding the evaluations on the Kinect dataset; from the quantitative and visual evaluation, it is observed that the depth map generated by the method is comparable to Kinect native depth and the proposed method can compute depth information on edges and non fronto-planar surfaces where Kinect's depth estimation fail due to insufficient reflectance of infrared beams on such surfaces. Another potential application of the method can also be to use in combination with Kinect to get a better depth coverage of the scene.

The proposed method is limited only to planar surfaces, though it provides reasonable estimations on curved surfaces as well. Moreover, the method does not run in real-time (the computational complexity is $O(Ndw)$, where N is the number of pixels in the image, d is maximum designated disparity and w is the maximum segment size in number of pixels in the image segmentation); rather the focus was to develop an accurate method for multi-modal stereo-vision.

A systematic performance evaluation of alternative similarity measures available in the literature is also performed as part of this thesis. The evaluated measures are MI with and without incorporating prior probabilities, i.e. MI(woPR) and MI(wPR), LSS, HOG, Census, MI of Census, SIFT, SURF, BRIEF, FREAK, NCC and SSD. Besides, a modified version of Census Transform which is based on computing mutual information similarity over the locally transformed image patches is introduced.

Mutual information and its derivatives are concluded to be the best performing measures in all cases of multi-modality levels where HOG, MI of Census and SIFT showed promising results. On the Kinect dataset, SURF and BRIEF are also performing well in the whole image however, in the local regions corresponding to more diverged intensity levels of the NIR band to visible band, these two measures still lack in performance. Regarding the noise experiments; MI of Census, SIFT and HOG are concluded as the most vulnerable measures and MI(wPR) is concluded as the most robust method to noise.

The future work regarding this study can be defined as (i) to develop a hierarchical processing method (ii) evaluating the method on thermal-visible camera pairs (iii)

evaluating alternative segmentation algorithms including texture segmentation algorithms.

Hierarchical processing can overcome the high computational cost of the proposed method while enhancing the results. Besides, within the hierarchical steps, the iterative method can still be applied for reducing the dependency of the method on initial segmentation.

The thermal-visible camera pairs are widely used in surveillance products since thermal cameras provide enhanced visibility under low visibility conditions like low light or night conditions, smoke and camouflaged human, vehicles or weapons. Since thermal cameras work at the emission bands of the EM spectrum, performing stereo correspondence is challenging. In this study, performing the cosine transformation over the unimodal images, these challenging conditions were tested to some extent where the proposed method outperformed all the alternative methods providing promising results for the thermal-visible camera pairs also.

Initial segmentation of the left IR images is very important for the performance of the algorithm although the dependency is shown to be reduced to some extent. Therefore, alternative segmentation algorithms can be applied for better performance of the method including texture segmentation. These algorithms can also enable to evaluating the segments of both left and right images and compromise to a one set of segmentation map to increase the details in the final disparity map.

A final future work can be to continue comparing the method with more alternative similarity measures available in the literature like the multi-modal version of the SIFT.

REFERENCES

- [1] J.P. Pluim, J. A. Maintz, and M. A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004, 2003.
- [2] P.A. Van den Elsen, Pol E-J.D., and Viergever. Medical image matching – a review with classification.
- [3] A. Roche, G. Malandain, X. Pennec, and N. Ayache. The correlation ratio as a new similarity measure for multimodal image registration. In *Medical Image Computing and Computer-Assisted Intervention, MICCAI-98*, pages 1115–1124. Springer, 1998.
- [4] M. Mellor and M. Brady. Non-rigid multimodal image registration using local phase. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004*, pages 789–796. Springer, 2004.
- [5] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, 1997.
- [6] C. B. Fookes. Medical image registration and stereo vision using mutual information.
- [7] S. Periaswamy and H. Farid. Medical image registration with partial data. *Medical Image Analysis*, 10(3):452–464, June 2006.
- [8] M. T. Eismann. *Hyperspectral remote sensing*, volume PM210. SPIE Press Monograph.
- [9] J. A. Richards. *Remote Sensing Digital Image Analysis*, volume PM210. Springer-Verlag, 5 edition.
- [10] P. E Anuta. Spatial registration of multispectral and multitemporal digital imagery using fast fourier transform techniques. *IEEE Transactions on Geoscience Electronics*, 8(4):353–368, 1970.
- [11] E. Rignot, R. Kwok, J. Curlander, and S. Pang. Automated multisensor registration: Requirements and techniques. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 945–948. IEEE, 1990.

- [12] M. A. Ali and D. A. Clausi. Automatic registration of sar and visible band remote sensing images. In *IEEE International Geoscience and Remote Sensing Symposium*, volume 3, pages 1331–1333. IEEE, 2002.
- [13] X. Liu, J. Yang, and H. Shen. Automatic image registration by local descriptors in remote sensing. *Optical Engineering*, 47(8):087206–087206, 2008.
- [14] A. Wong and D. A. Clausi. Arrsi: automatic registration of remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 45(5):1483–1493, 2007.
- [15] A. Wong and P. Fieguth. Fast phase-based registration of multi modal image data. *Signal Processing*, 89:724–737, June 2009.
- [16] Esa eduspace: Remote sensing in depth. http://www.esa.int/SPECIALS/Eduspace_EN/SEM7IQ3Z2OF_0.html. Accessed: 20 Aug 2014.
- [17] C. Berger, M. Voltersen, R. Eckardt, J. Eberle, T. Heyer, N. Salepci, S. Hese, C. Schmullius, J. Tao, S. Auer, R. Bamler, K. Ewald, M. Gartley, J. Jacobson, A. Buswell, Q. Du, and F. Pacifici. Multi-modal and multi-temporal data fusion: outcome of the 2012 grss data fusion contest. *IEEE Journal of Selected Topics In Applied Earth Observations and Remote Sensing*, 6(3), June 2013.
- [18] M. Hasan, M.R. Pickering, and X. Jia. Multi-modal registration of sar and optical satellite images. pages 447–453. *Digital Image Computing: Techniques and Applications*, 2009.
- [19] A. Brook, E. Ben-Dor, and R Richter. Fusion of hyperspectral images and lidar data for civil engineering structure monitoring. In *2nd Workshop Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, pages 1–5. WHISPERS, June 2009.
- [20] J. M. Kellndorfer, W. S. Walker, E. LaPoint, K. Kirsch, and J. Bishop. Statistical fusion of lidar, insar, and optical remote sensing data for forest stand height characterization: A regional-scale method based on lvis, srtm, landsat etm plus, and ancillary data sets. *Journal of Geophysical Research: Biogeosciences*, 115, June 2010.
- [21] N. Longbotham, F. Pacifici, T. Glenn, A. Zare, M. Volpi, D. Tuia, E. Christophe, J. Michel, J. Inglada, J. Chanussot, and Q. Du. Multimodal change detection, application to the detection of flooded areas: Outcome of the 2009–2010 data fusion contest. *IEEE Journal of Selected Topics In Applied Earth Observations and Remote Sensing*, 5(1):331–342, February.
- [22] A. A. Richards. *Alien Vision: Exploring the Electromagnetic Spectrum with Imaging Technology*, volume PM205. SPIE Press Monograph, 2011.

- [23] T. P. Breckon, A. Gaszczak, J. Han, M. L. Eichner, and S. E. Barnes. Multi-modal target detection for autonomous wide area search and surveillance. In *Proceedings of SPIE: Emerging Technologies in Security and Defence; and Quantum Security II; and Unmanned Sensor Systems X*, volume 8899. SPIE.
- [24] Z. Zhu and T. S. Huang, editors. *Multimodal Surveillance: Sensors, Algorithms and Systems*. Artech House, 2007.
- [25] C. Beyan and A. Temizel. Mean-shift tracking for surveillance applications using thermal and visible band data fusion. In *Proceedings of SPIE: Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications VIII*, volume 8020. SPIE.
- [26] S. Krotosky and M. Trivedi. Mutual information based registration of multi-modal stereo videos for person tracking. *Computer Vision and Image Understanding*, 106(2):270–287, 2007.
- [27] Ranger hrc: Portable, long range thermal imaging surveillance system with multi-sensor option. <http://gs.flir.com/surveillance-products/ranger-imagers/ms-hrc>. Accessed: 20 Aug 2014.
- [28] Mx-rsta: A multi-sensor, multi-spectral imaging system. <http://www.wescam.com/index.php/products-services/ground-market/mx-rsta/>. Accessed: 20 Aug 2014.
- [29] Mx-25d: Fully digital, high definition, ultra long-range multi-sensor, multi-spectral imaging and targeting systems. <http://www.wescam.com/index.php/products-services/airborne-targeting/mx-25d/>. Accessed: 20 Aug 2014.
- [30] Seaflir 380hd: The only all-digital, full hd system. <http://gs.flir.com/surveillance-products/seaflir/seaflir-380-hd>. Accessed: 20 Aug 2014.
- [31] M. Yaman and S. Kalkan. Multimodal stereo vision using mutual information with adaptive windowing. In *13th IAPR International Conference on Machine Vision Applications*. IAPR, 2013.
- [32] M. Yaman and S. Kalkan. An iterative adaptive multi-modal stereo-vision method using mutual information. *Journal of Visual Communication and Image Representation*, August 2014 (Major Revision).
- [33] The middlebury stereo vision page. <http://vision.middlebury.edu/stereo/>. Accessed: 20 Aug 2013.
- [34] C. Fookes, A. Maeder, S. Sridharan, and J. Cook. Multi-spectral stereo image matching using mutual information. In *International Symposium on 3D Data Processing, Visualization and Transmission*, pages 961–968. IEEE, 2004.

- [35] Microsoft's kinect for windows. <http://www.microsoft.com/en-us/kinectforwindows/>. Accessed: 6 Oct 2013.
- [36] The xbox 360 video game console. <http://www.xbox.com>. Accessed: 6 Oct 2013.
- [37] M. Beetz, D. Cremers, J. Gall, W. Li, Z. Liu, D. Pangercic, J. Sturm, and Y.-W. Tai. Special issue on visual understanding and applications with rgb-d cameras. *Journal of Visual Communication and Image Representation*, 25(1):1–238, 2014.
- [38] M. Yaman and S. Kalkan. A performance evaluation of similarity measures for dense multi-modal stereo-vision applications. *Journal of Visual Communication and Image Representation*, Sep 2014 (Initial Submission).
- [39] Z. B. Myron, B. Darius, and D. H. Gregory. Advances in computational stereo. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 25(8):993–1008, 2003.
- [40] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge Univ Press, 2000.
- [41] U. R. Dhond and J. K. Aggarwal. Structure from stereo—a review. *IEEE Transactions on Systems, Man and Cybernetics*, 19(6):1489–1510, 1989.
- [42] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.
- [43] N. Lazaros, G. C. Sirakoulis, and A. Gasteratos. Review of stereo vision algorithms: from software to hardware. *International Journal of Optomechatronics*, 2(4):435–462, 2008.
- [44] B. Tippetts, D. J. Lee, K. Lillywhite, and J. Archibald. Review of stereo vision algorithms and their suitability for resource-limited systems. *Journal of Real-Time Image Processing*, pages 1–21, 2008.
- [45] Middlebury stereo evaluation - version 2. <http://vision.middlebury.edu/stereo/eval/>. Accessed: 20 Aug 2013.
- [46] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [47] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision—ECCV*, pages 404–417. Springer, 2006.
- [48] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003.

- [49] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Manchester, UK, 1988.
- [50] C. Schmid, R. Mohr, and C. Bauckhage. Comparing and evaluating interest points. In *Sixth International Conference on Computer Vision*, pages 230–235. IEEE, 1998.
- [51] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *European Conference on Computer Vision*, pages 128–142. Springer, 2002.
- [52] V. Venkateswar and R. Chellappa. Hierarchical stereo and motion correspondence using feature groupings. *International Journal of Computer Vision*, 15(3):245–269, 1995.
- [53] S. Birchfield and C. Tomasi. Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision*, 35(3):269–293, 1999.
- [54] G. Egnal. Mutual information as a stereo correspondence measure. *Technical Report MS-CIS-00-20, University of Pennsylvania*, page 113, 2000.
- [55] M. J. Hannah. *Computer matching of areas in stereo images*. PhD thesis, Stanford University, 1974.
- [56] P. Aschwanden and W. Guggenbuhl. Experimental results from a comparative study on correlation-type registration algorithms. *Robust computer vision*, pages 268–289, 1992.
- [57] K. Ambrosch, W. Kubinger, M. Humenberger, and A. Steininger. Flexible hardware-based stereo matching. *EURASIP Journal on Embedded Systems*, 2008(2), 2008.
- [58] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):920–932, 1994.
- [59] C. S. Park and H. W. Park. A robust stereo disparity estimation using adaptive window search and dynamic programming search. *Pattern Recognition*, 34(12):2573–2576, 2001.
- [60] O. Veksler. Stereo correspondence by dynamic programming on a tree. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 384–390. IEEE, 2005.
- [61] C. Cassisa. Local vs global energy minimization methods: application to stereo matching. In *IEEE International Conference on Progress in Informatics and Computing (PIC)*, volume 2, pages 678–683. IEEE, 2010.

- [62] J. Marroquin, S. Mitter, and T. Poggio. Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association*, 82(397):76–89, 1987.
- [63] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *IEEE International Conference on Computer Vision*, volume 2, pages 508–515. IEEE, 2001.
- [64] J. Sun, N. Zheng, and H. Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):787–800, 2003.
- [65] P. Viola and W. M. Wells III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.
- [66] C. Fookes, A. Lamanna, and M. Bennamoun. A new stereo image matching technique using mutual information. *International Conference on Computer, Graphics and Imaging*, 2001.
- [67] S. Krotosky and M. Trivedi. Multimodal stereo image registration for pedestrian detection. In *IEEE Intelligent Transportation Systems Conference*, pages 109–114. IEEE, 2006.
- [68] S. Krotosky and M. Trivedi. Registration of multimodal stereo images using disparity voting from correspondence windows. In *IEEE International Conference on Video and Signal Based Surveillance*, pages 91–91. IEEE, 2006.
- [69] F. Barrera Campo, F. Lumbreras Ruiz, and A.D. Sappa. Multimodal stereo vision system: 3d data extraction and algorithm evaluation. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):437–446, 2012.
- [70] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.
- [71] A. Torabi and G-A Bilodeau. Local self-similarity as a dense stereo correspondence measure for thermal-visible video registration. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 61–67. IEEE, 2011.
- [72] A. Torabi, M. Najafianrazavi, and G-A Bilodeau. A comparative evaluation of multimodal dense stereo correspondence measures. In *IEEE International Symposium on Robotic and Sensors Environments (ROSE)*, pages 143–148. IEEE, 2011.
- [73] A. Torabi and G-A Bilodeau. A lss-based registration of stereo thermal–visible videos of multiple people using belief propagation. *Computer Vision and Image Understanding*, 117(12):1736–1747, 2013.

- [74] G-A Bilodeau, A. Torabi, P.-L. St-Charles, and D. Riahi. Thermal-visible registration of human silhouettes: a similarity measure performance evaluation. *Infrared Physics & Technology*, 64:79–86, 2014.
- [75] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 1948.
- [76] C. M. Bishop, editor. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [77] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [78] P. Suetens A. Collignon, D. Vandermeulen and G. Marchal. 3d multi-modality medical image registration using feature space clustering. In *Proceedings of 1st International Conference On: Computer Vision, Virtual Reality, and Robotics in Medicine*, volume 905, page 195–204, April 1995.
- [79] D. L. G. Hill C. Studholme and D. J. Hawkes. Multiresolution voxel similarity measures for mr-pet registration. In *Information Processing in Medical Imaging*, page 287–298, 1995.
- [80] Mutual information for image registration and feature selection. www.cse.msu.edu/~cse902/S03/mut_info.ppt. Accessed: 20 August 2014.
- [81] B. Triggs N. Dalal. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, page 886–893. IEEE, 2005.
- [82] M. Yuasa T. Kozakaya, T. Shibata and O. Yamaguchi. Facial feature localization using weighted vector concentration approach. *Image and Vision Computing*, 28(5):772–780, 2010.
- [83] M. Brown and D. Lowe. Invariant features from interest point groups. In *British Machine Vision Conference*, pages 656–665, 2002.
- [84] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *Computer Vision ECCV '94, Lecture Notes in Computer Science*, volume 801, page 151–158. Springer, 1994.
- [85] R. W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160, 1950.
- [86] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: binary robust independent elementary features. In *Computer Vision ECCV 2010, Lecture Notes in Computer Science*, volume 6314, page 778–792. Springer, 2010.

- [87] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: fast retina keypoint. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, page 510–517. IEEE, 2012.
- [88] Avian visual cognition. <http://pigeon.psy.tufts.edu/avc/>, September 2001. Accessed: 20 Aug 2014.
- [89] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1582–1599, September 2009.
- [90] Middlebury stereo evaluation - version 2 new features and main differences to version 1. <http://vision.middlebury.edu/stereo/eval/newFeatures.html>. Accessed: 20 Aug 2013.
- [91] Company behind microsofts kinect sensor sold to apple for 345 million. <http://www.winbeta.org/news/company-behind-microsofts-kinect-sensor-sold-apple-345-million>. Accessed: 20 August 2014.
- [92] How it works: Xbox kinect. <http://www.jameco.com/jameco/workshop/howitworks/xboxkinect.html>. Accessed: 20 August 2014.
- [93] Rgbdemo software - calibrating kinect with openni backend. <http://labs.manctl.com/rgbdemo/index.php/Documentation/Calibration>. Accessed: 6 Oct 2013.
- [94] Opencv api reference: Camera calibration and 3d reconstruction. http://docs.opencv.org/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html. Accessed: 20 August 2014.
- [95] A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *International Conference on Pattern Recognition*, volume 3, pages 15–18. IEEE, 2006.
- [96] Y. Taguchi, B. Wilburn, and C. L. Zitnick. Stereo reconstruction with mixed pixels using adaptive over-segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.
- [97] H. Tao, H.S. Sawhney, and R. Kumar. A global matching framework for stereo computation. In *IEEE International Conference on Computer Vision*, volume 1, pages 532–539. IEEE, 2001.
- [98] Z. Wang and Z. Zheng. A region based stereo matching algorithm using cooperative optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

- [99] Q. Yang, L. Wang, R. Yang, H. Stewénus, and D. Nistér. Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):492–504, 2009.
- [100] C. L. Zitnick and S.B. Kang. Stereo for image-based rendering using image over-segmentation. *International Journal of Computer Vision*, 75(1):49–65, 2007.
- [101] C. L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics (TOG)*, 23(3):600–608, 2004.
- [102] C. M. Christoudias, B. Georgescu, and P. Meer. Synergism in low-level vision. In *16th International Conference on Pattern Recognition*, pages 150–155, 2002.
- [103] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.
- [104] M. A. Fischler and R.C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

APPENDIX A

PARAMETER SETTINGS USED IN EXPERIMENTS

Table A.1 provides the configured parameters of the similarity measures implemented for the performance evaluation experiments in Chapter 4, over the datasets provided.

Table A.2 provides the configured parameters of the proposed method applied to Dataset #1 for the experiments given in Chapter 5.

Table A.3 provides the configured parameters of the proposed method applied to Dataset #2 for the experiments given in Chapter 5.

Table A.1: Parameter Settings used for the implementation of the evaluated similarity measures

Method Name	Parameter Name	Value
MI(woPR)	<i>binsize(hist)</i>	40
MI(wPR)	λ	0.3
LSS	<i>size(smallpatch)</i>	5
	<i>size(largepatch)</i>	41
	<i>num(angles)</i>	20
HOG	K	9
	γ	1
	<i>size(detectionwindow)</i>	16
	<i>size(cells)</i>	8
SIFT	σ	3
	<i>num(octaves)</i>	4
	<i>num(octaveintervals)</i>	3
	<i>size(descriptor)</i>	128
SURF	σ	3.3
	<i>num(octaves)</i>	4
	<i>num(octaveintervals)</i>	2
	<i>size(descriptor)</i>	64
CENSUS	<i>size(window)</i>	3
	<i>size(descriptorstrbits)</i>	8
CENSUSMI	<i>binsize(hist)</i>	40
BRIEF	<i>size(descriptorbits)</i>	32
	<i>size(patch)</i>	48
	<i>size(kernel)</i>	9
	<i>gridtype</i>	<i>GIV</i>
FREAK	<i>num(octaves)</i>	4
	<i>num(scales)</i>	64
	<i>num(pairs)</i>	512
	<i>num(oripairs)</i>	45

Table A.2: Parameter Settings of the Proposed Method Used in Dataset #1 (Synt. Altered Middlebury) Experiments.

Segmentation	h_s 7	h_r 6	M 50	n 7	a_{ij} 0.5	t_e 0.2
Adaptive Windowing	δ_y 4,10,15	λ 0.3	ω 1	$Size(h_w)$ 40	k 5	
Adaptive Cost Aggregation	ρ 0.25	λ_{SD} 1	λ_{DD} 1	$Size(w(p,q))$ 17x17		
Iterative Plane Fitting	τ_{ic} 0.007	τ_{ir} 0.25	τ_{od} 1.0	τ_{os} 20	τ_{oc} 0.014	
Segment Merging & Finalizing	$\tau_\alpha (^{\circ})$ 0.1	τ_{pd} 0.15	γ 0.25			

Table A.3: Parameter Settings of the Proposed Method Used in Dataset #2 (Kinect) Experiments.

Segmentation	h_s 7	h_r 4	M 300	n 2	a_{ij} 0.3	t_e 0.4*,0.6
Adaptive Windowing	δ_y 15	λ 0.4	ω 2	$Size(h_w)$ 40	k 5	
Adaptive Cost Aggregation	ρ 0.25	λ_{SD} 1	λ_{DD} 1	$Size(w(p,q))$ 37x37		
Iterative Plane Fitting	τ_{ic} 0.0015	τ_{ir} 0.25	τ_{od} 2.0	τ_{os} 200	τ_{oc} 0.004	
Segment Merging & Finalizing	$\tau_\alpha (^{\circ})$ 0.1	τ_{pd} 0.15	γ 0.25			

*: for the kinect IR images having low contrast

APPENDIX B

EXPERIMENT RESULTS FOR THE PERFORMANCE EVALUATION OF SIMILARITY MEASURES

Table B.1 provides the experiment results of the similarity measures tested over Dataset #1 using three different window sizes.

Table B.1: Results on Dataset #1 for the Similarity Measures Tested using Three Different Window Sizes

Image	Method	RMS	RMS	RMS	Bad	Bad	Bad
		(all) (9x9)*	(all) (21x21)	(all) (31x31)	(all) (9x9)	(all) (21x21)	(all) (31x31)
Avg. All	MI (woPR)	10,292	4,995	4,057	43,01%	16,35%	14,13%
	MI (wPR)	7,642	4,480	3,861	31,41%	16,93%	15,33%
	LSS	12,699	9,857	8,518	59,85%	47,16%	42,50%
	HOG	6,843	5,080	4,568	40,40%	31,77%	29,55%
	CENSUS	14,223	14,998	15,433	95,70%	96,60%	96,82%
	CENSUSMI	11,428	6,618	4,673	61,74%	25,86%	18,37%
	BRIEF	19,402	19,924	20,260	98,92%	99,44%	99,41%
	SIFT	11,934	10,188	9,139	65,93%	47,10%	40,58%
	SURF	14,152	14,660	15,108	96,15%	98,06%	98,33%
	NCC	15,803	16,796	17,768	96,93%	98,23%	98,63%
	SSD	17,839	18,411	18,735	98,36%	99,02%	98,87%
	FREAK	15,503	16,265	16,588	92,84%	93,54%	94,03%
Tsukuba	MI (woPR)	3,695	2,223	1,930	31,02%	12,93%	11,96%

Continued on next page

Table B.1 – *Continued from previous page*

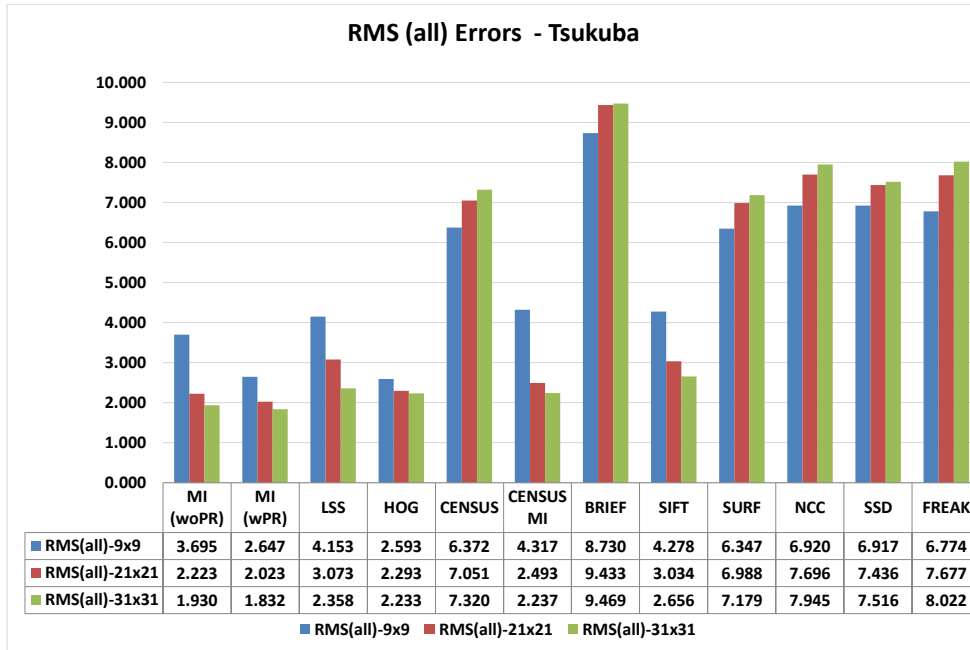
Image	Method	RMS (all) (9x9)*	RMS (all) (21x21)	RMS (all) (31x31)	Bad (all) (9x9)	Bad (all) (21x21)	Bad (all) (31x31)
	MI (wPR)	2,647	2,023	1,832	20,28%	12,53%	12,29%
	LSS	4,153	3,073	2,358	38,63%	25,64%	21,70%
	HOG	2,593	2,293	2,233	27,25%	27,59%	28,21%
	CENSUS	6,372	7,051	7,320	93,95%	95,33%	95,64%
	CENSUSMI	4,317	2,493	2,237	47,63%	18,77%	16,41%
	BRIEF	8,730	9,433	9,469	98,67%	99,34%	99,28%
	SIFT	4,278	3,034	2,656	46,91%	29,06%	24,91%
	SURF	6,347	6,988	7,179	94,79%	98,88%	99,51%
	NCC	6,920	7,696	7,945	93,29%	96,42%	97,34%
	SSD	6,917	7,436	7,516	97,17%	98,19%	98,05%
	FREAK	6,774	7,677	8,022	88,04%	89,79%	91,28%
Venus	MI (woPR)	5,607	2,783	2,084	38,11%	13,77%	9,26%
	MI (wPR)	4,153	2,915	2,462	28,50%	15,21%	11,77%
	LSS	7,842	6,092	5,308	69,41%	49,62%	41,01%
	HOG	3,887	2,338	1,774	37,44%	25,78%	20,41%
	CENSUS	8,108	8,603	8,826	93,43%	94,13%	94,36%
	CENSUSMI	6,342	3,615	1,959	60,65%	24,50%	10,73%
	BRIEF	11,965	12,801	12,906	99,17%	99,94%	99,99%
	SIFT	6,583	5,423	4,635	66,66%	50,16%	43,32%
	SURF	8,273	8,558	8,817	94,79%	95,51%	95,81%
	NCC	9,195	9,848	10,426	96,61%	98,09%	98,78%
	SSD	9,663	9,675	9,734	99,58%	99,94%	100,00%
	FREAK	9,344	10,094	10,384	92,28%	93,15%	93,40%
Teddy	MI (woPR)	15,840	8,433	6,415	55,65%	20,40%	15,94%
	MI (wPR)	11,210	7,019	6,206	39,49%	20,60%	17,49%
	LSS	18,474	15,853	14,545	69,29%	60,37%	57,87%
	HOG	11,886	9,104	8,489	50,93%	38,67%	35,55%
	CENSUS	20,471	21,910	22,633	97,53%	98,37%	98,56%

Continued on next page

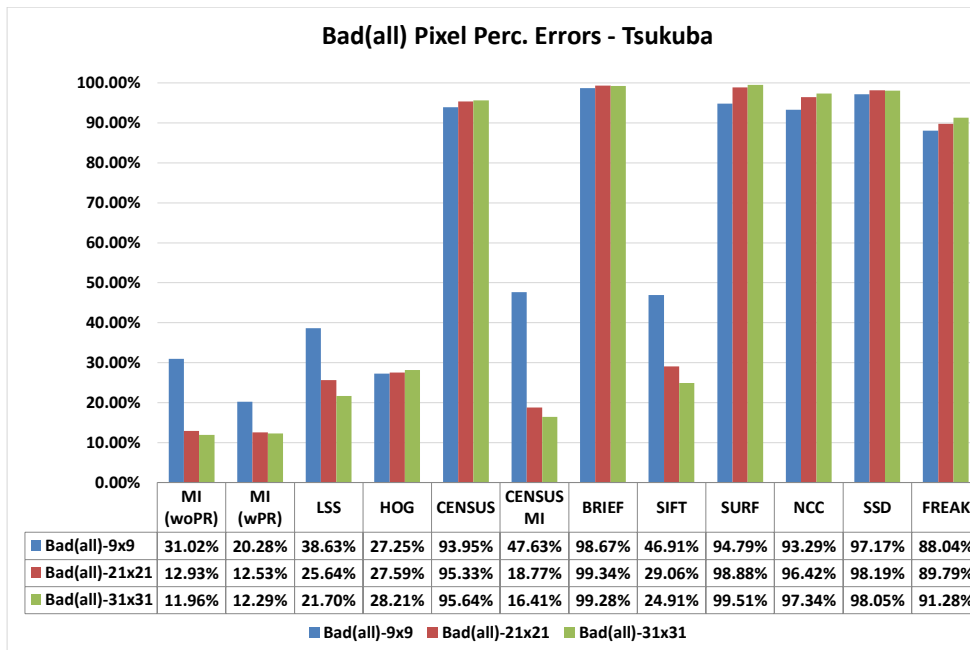
Table B.1 – Continued from previous page

Image	Method	RMS (all) (9x9)*	RMS (all) (21x21)	RMS (all) (31x31)	Bad (all) (9x9)	Bad (all) (21x21)	Bad (all) (31x31)
	CENSUSMI	17,527	10,909	7,920	72,81%	30,93%	21,39%
	BRIEF	28,744	29,732	30,222	99,21%	99,64%	99,52%
	SIFT	18,196	16,087	14,649	74,86%	56,53%	48,49%
	SURF	20,464	21,266	21,906	97,56%	98,65%	98,82%
	NCC	22,869	24,716	26,368	98,76%	99,50%	99,29%
	SSD	25,317	26,673	27,573	98,57%	99,19%	99,19%
	FREAK	22,329	22,703	23,025	94,49%	93,32%	93,06%
Cones	MI (woPR)	16,025	6,540	5,797	47,3%	18,3%	19,4%
	MI (wPR)	12,557	5,964	4,944	37,3%	19,4%	19,8%
	LSS	20,328	14,407	11,862	62,1%	53,0%	49,4%
	HOG	9,006	6,586	5,776	46,0%	35,0%	34,0%
	CENSUS	21,941	22,430	22,954	97,9%	98,6%	98,7%
	CENSUSMI	17,525	9,454	6,576	65,9%	29,3%	25,0%
	BRIEF	28,168	27,728	28,445	98,6%	98,9%	98,9%
	SIFT	18,678	16,210	14,616	75,3%	52,7%	45,6%
	SURF	21,525	21,826	22,530	97,5%	99,2%	99,2%
	NCC	24,229	24,924	26,333	99,1%	98,9%	99,1%
	SSD	29,457	29,859	30,118	98,1%	98,8%	98,2%
	FREAK	23,565	24,586	24,922	96,6%	97,9%	98,4%

Figures B.1, B.2, B.3 and B.4 depicts these RMS and Bad pixel performances for each image separately, i.e. Tsukuba, Venus, Teddy and Cones for the three window sizes tested over Dataset #1.

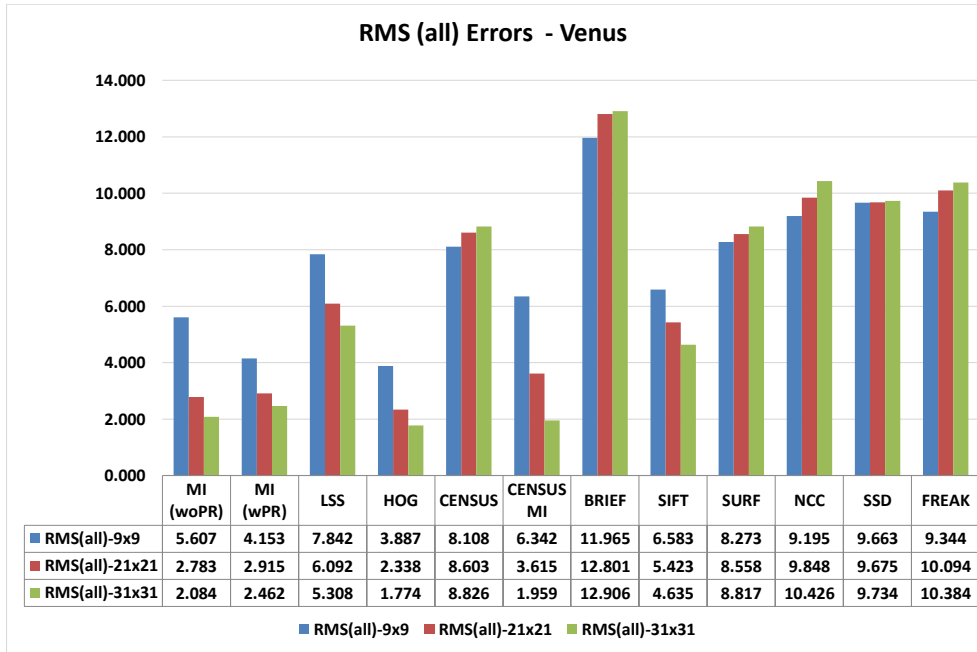


(a)

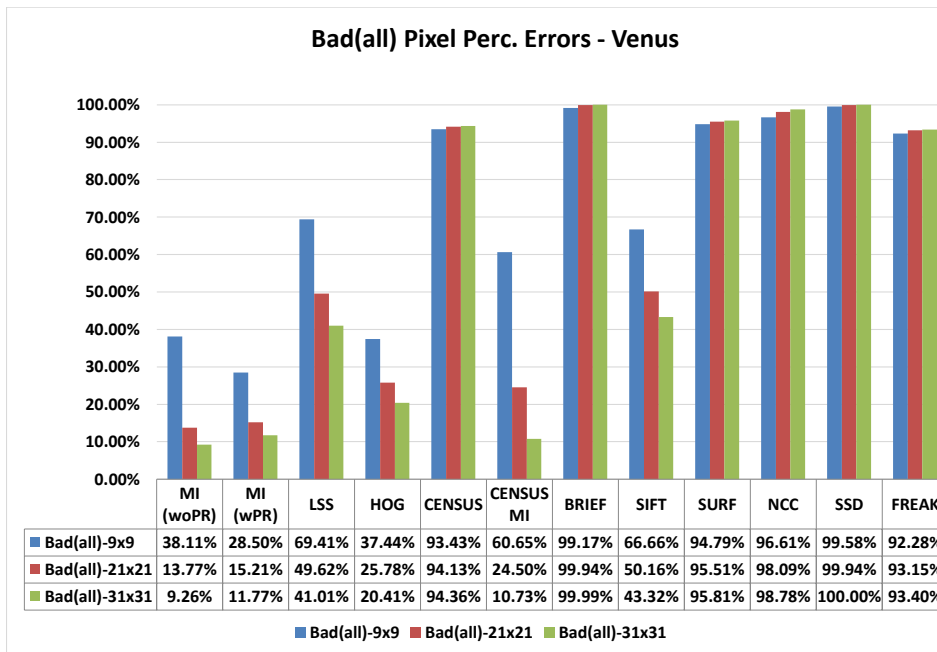


(b)

Figure B.1: RMS(all) and Bad(all) pixels percentage errors of all methods' "WTA" performances for three different window sizes for Tsukuba image in Dataset1

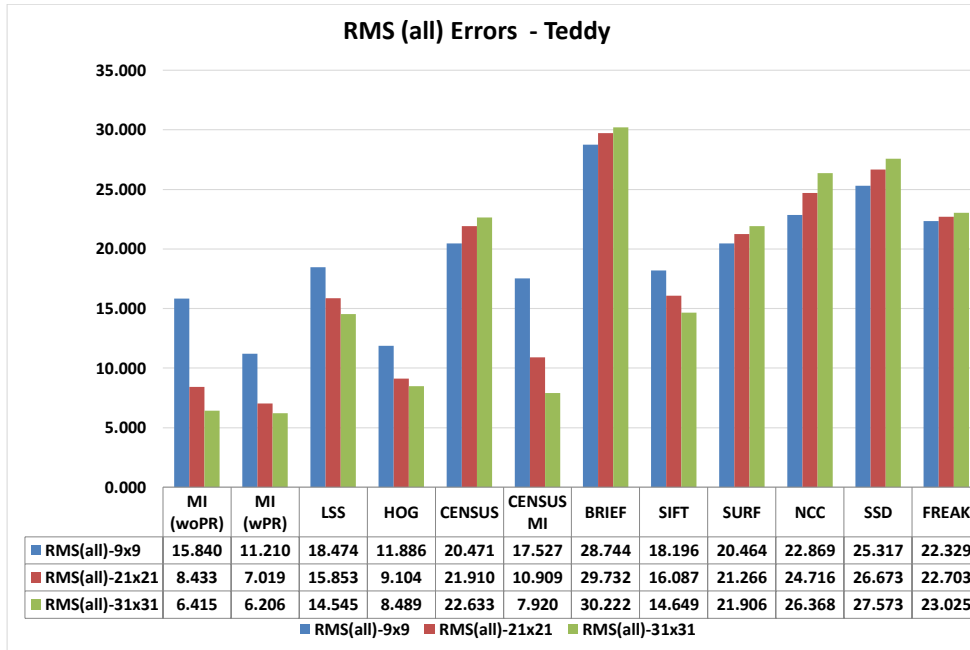


(a)

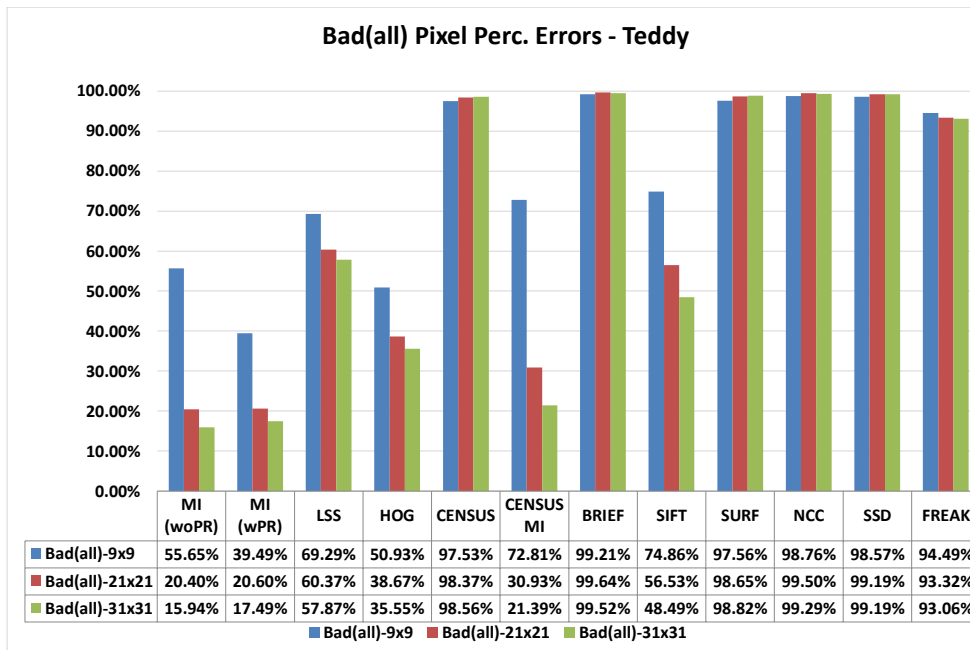


(b)

Figure B.2: RMS(all) and Bad(all) pixels percentage errors of all methods' "WTA" performances for three different window sizes for Venus image in Dataset1

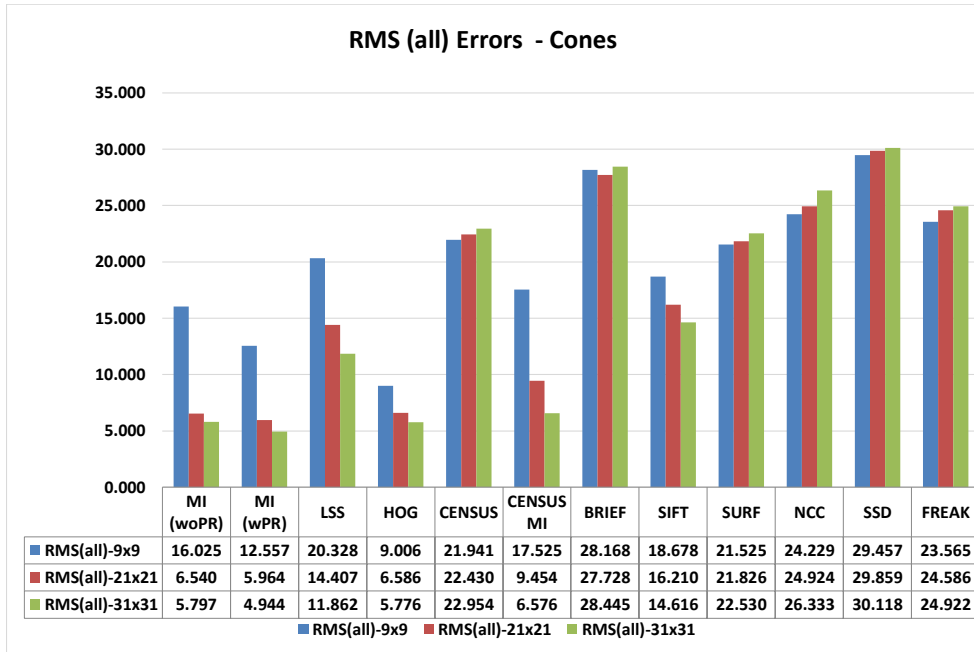


(a)

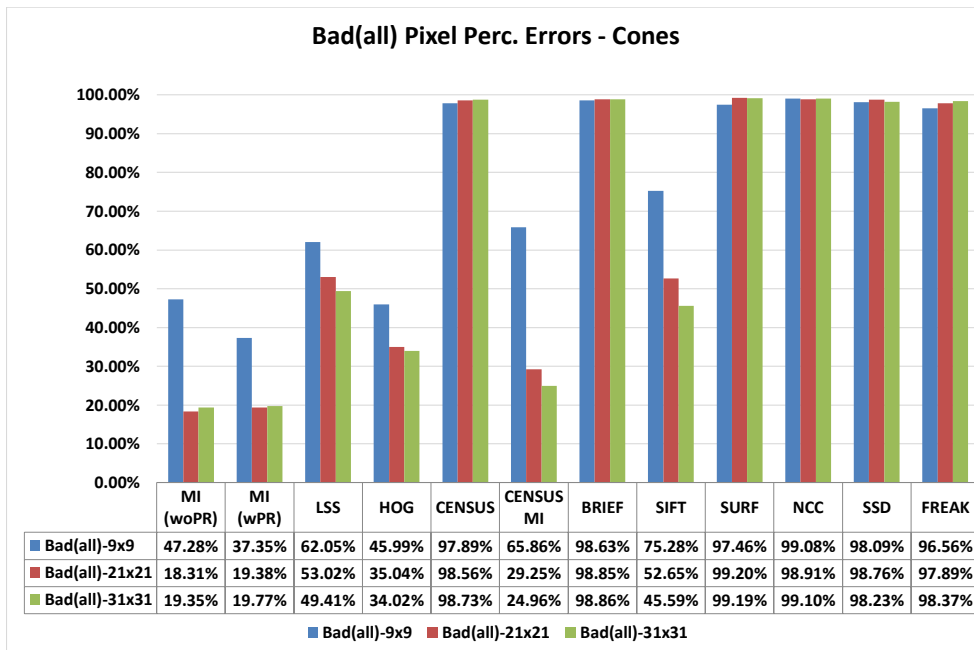


(b)

Figure B.3: RMS(all) and Bad(all) pixels percentage errors of all methods' "WTA" performances for three different window sizes for Teddy image in Dataset #1.



(a)



(b)

Figure B.4: RMS(all) and Bad(all) pixels percentage errors of all methods' "WTA" performances for three different window sizes for Cones image in Dataset #1.

In the below part, Figure B.5 shows all the visual results for the "WTA" disparities generated by the similarity measures using Tsukuba image pair in Dataset #1 for three different window sizes, 9x9, 21x21 and 31x31 pixels.

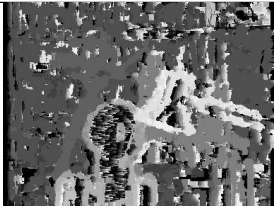
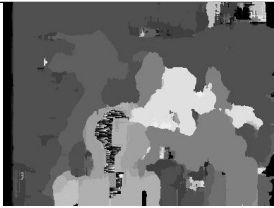
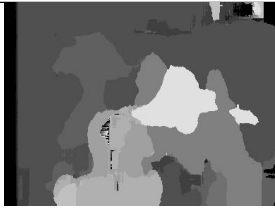

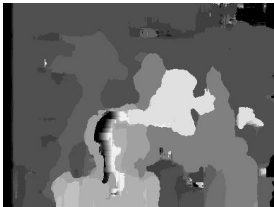




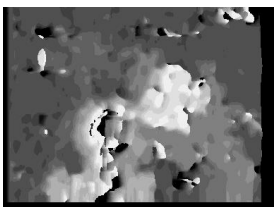
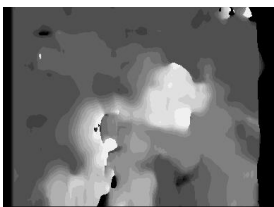
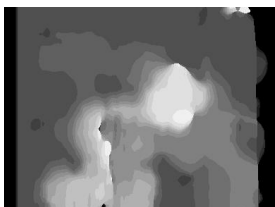
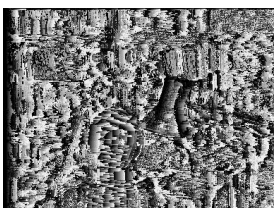

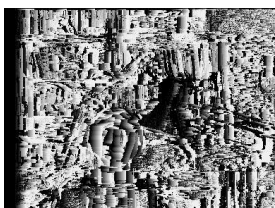
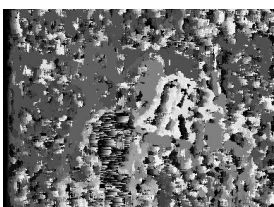
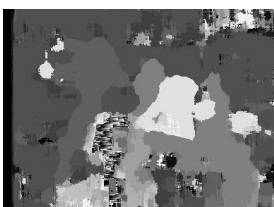




Method	9x9	21x21	31x31
MI (woPR)			
MI (wPR)			
LSS			
HOG			
CENSUS			
CENSUSMI			
BRIEF			

Figure B.5: continued

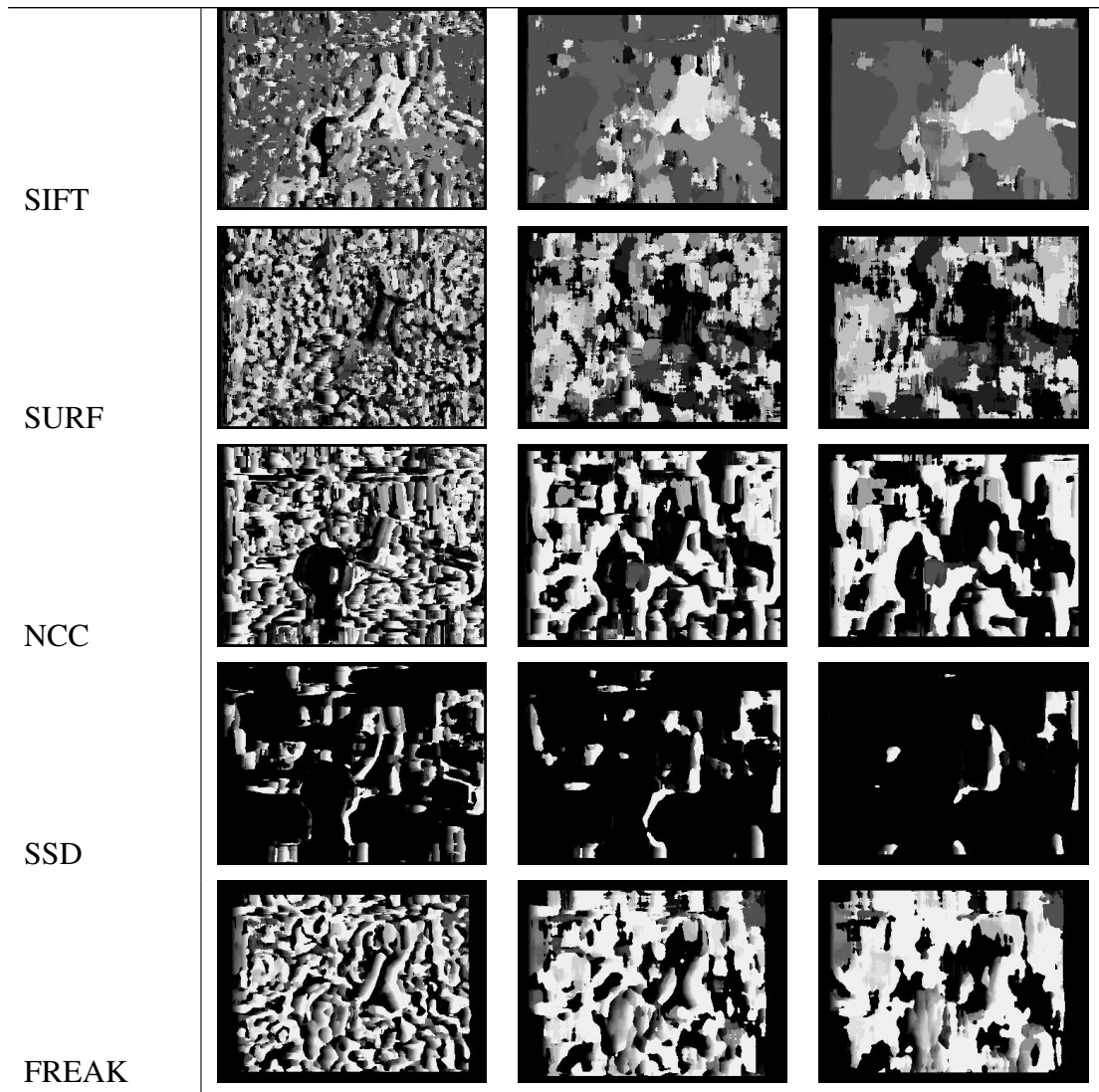
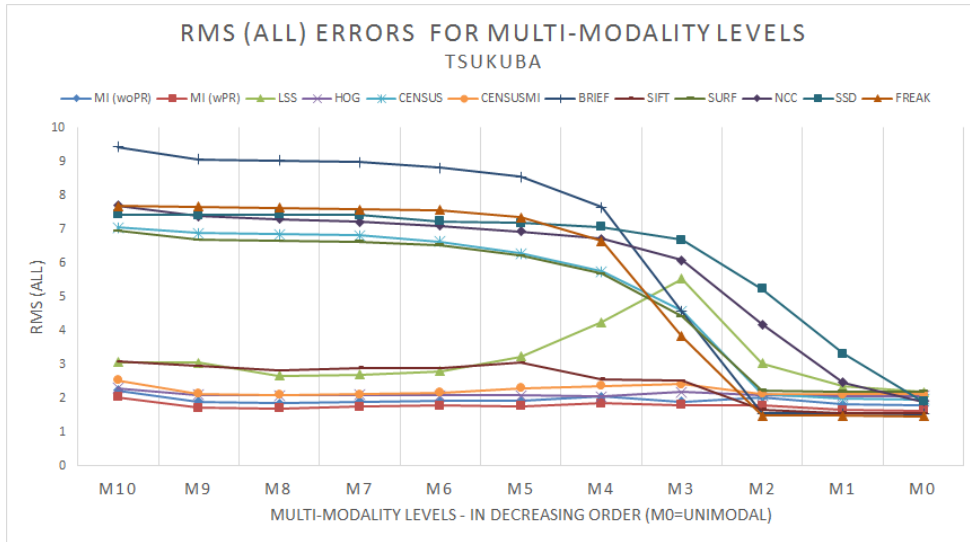
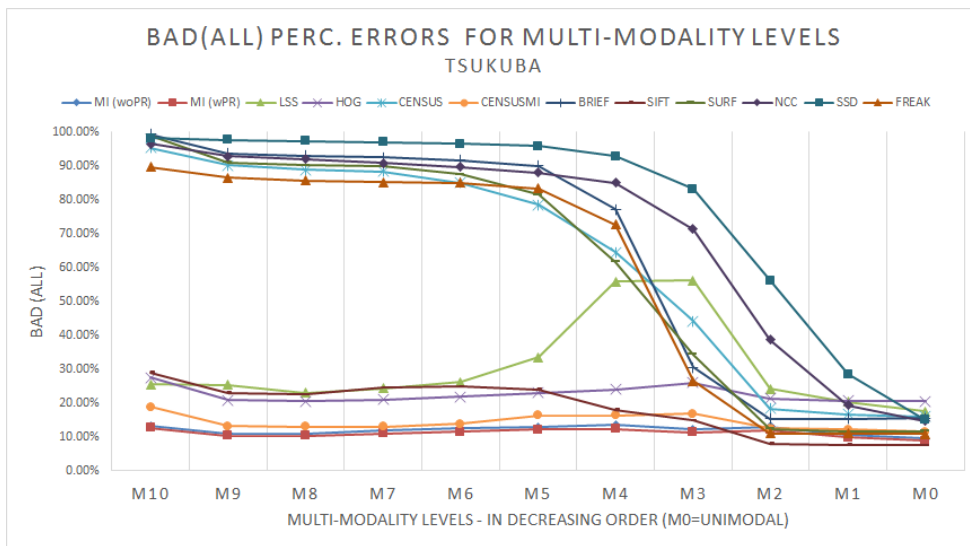


Figure B.5: The visual results of all the similarity measures for the Tsukuba image pair, for the different window sizes 9x9, 21x21 and 31x31.

The Figures B.6, B.7, B.8 and B.9 provide the RMS(all) and Bad(all) pixels percentage errors of all similarity measures tested for the 10 multi-modality levels for each image pair in Dataset #1 separately, i.e. Tsukuba, Venus, Teddy and Cones.

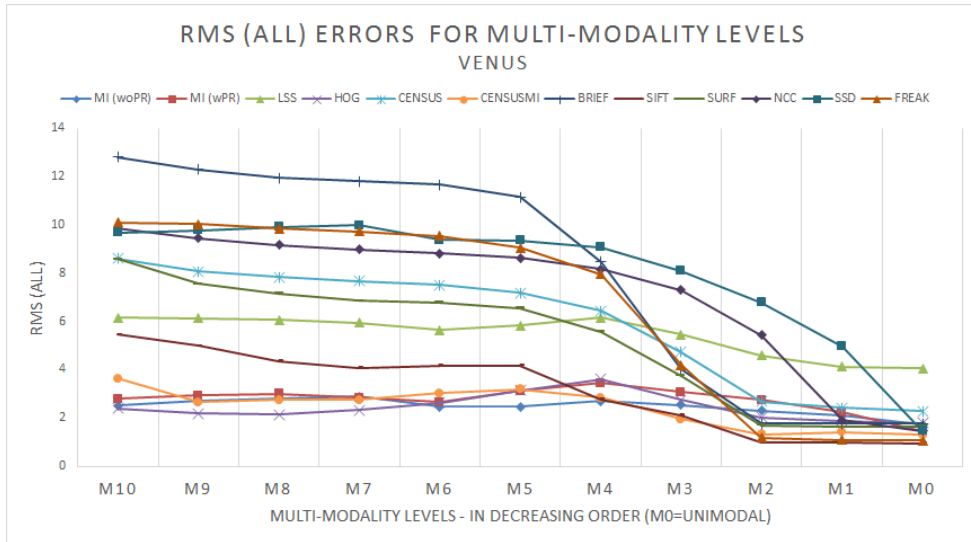


(a)

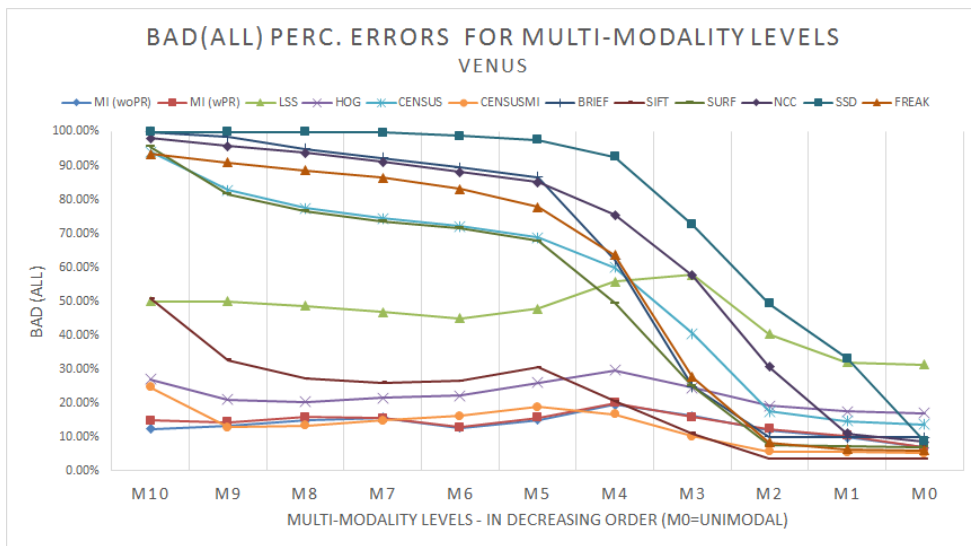


(b)

Figure B.6: RMS(all) and Bad(all) pixels percentage errors of all methods for 10 multi-modality levels for Tsukuba image in Dataset #1 [Best viewed in color].

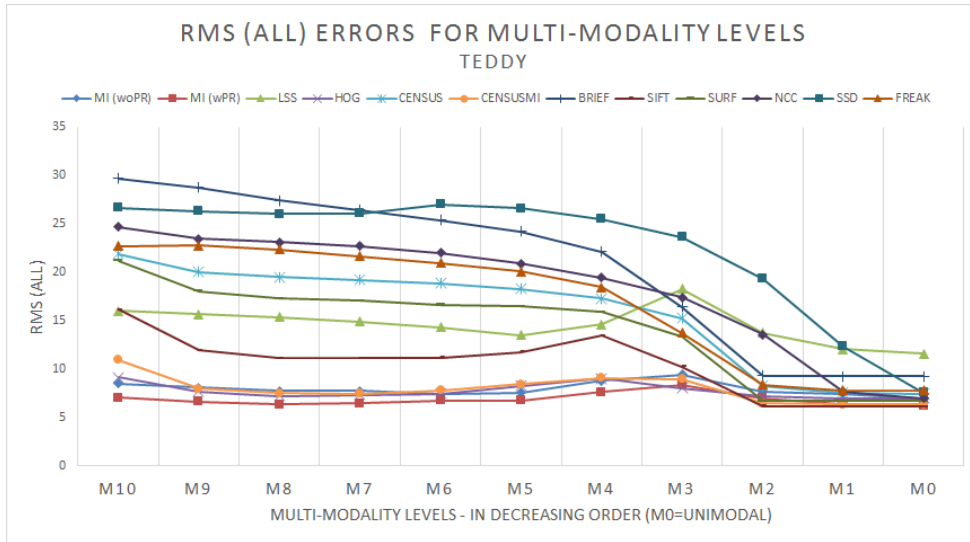


(a)

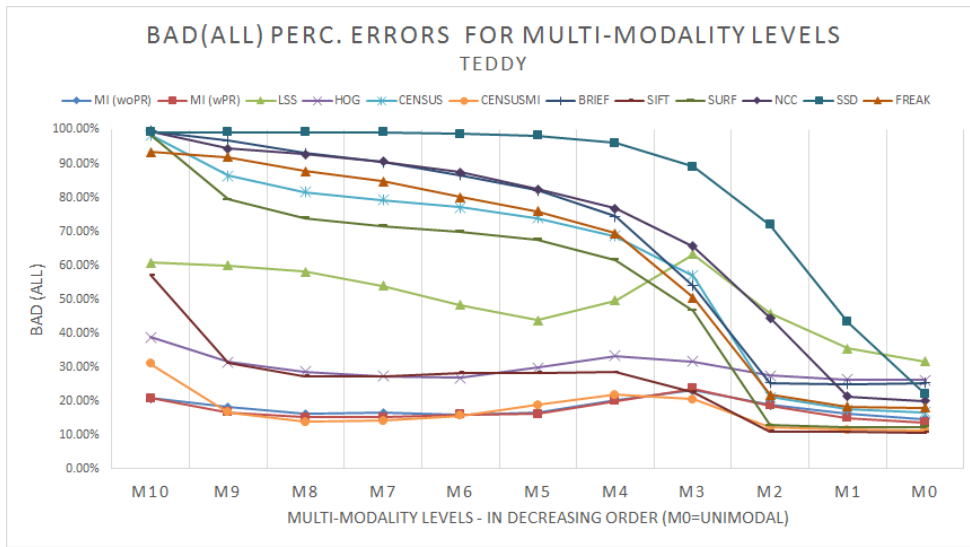


(b)

Figure B.7: RMS(all) and Bad(all) pixels percentage errors of all methods for 10 multi-modality levels for Venus image in Dataset #1 [Best viewed in color].

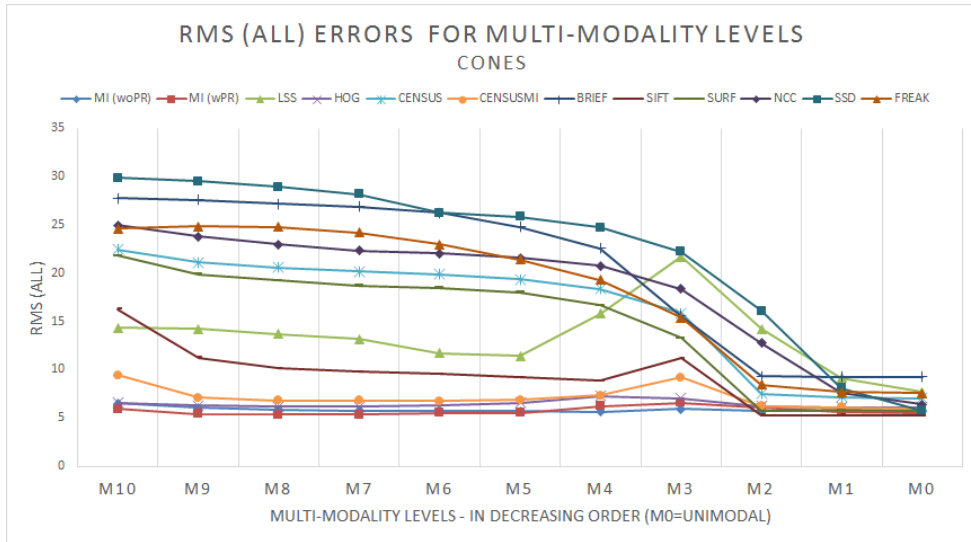


(a)

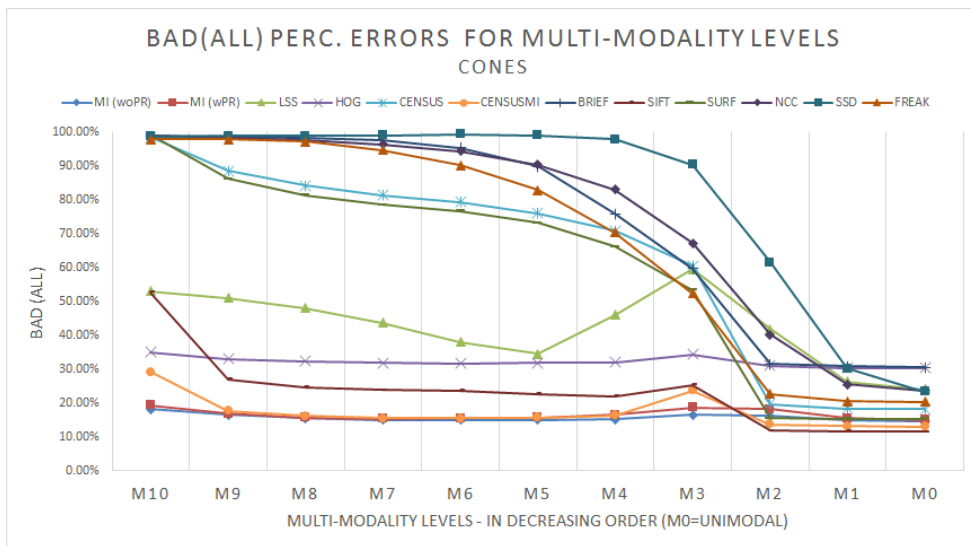


(b)

Figure B.8: RMS(all) and Bad(all) pixels percentage errors of all methods for 10 multi-modality levels for Venus image in Dataset #1 [Best viewed in color].



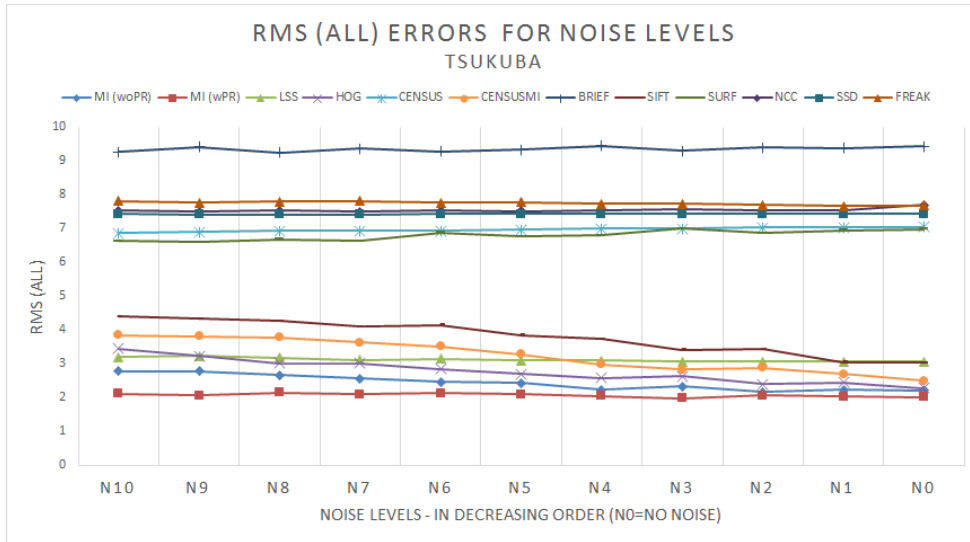
(a)



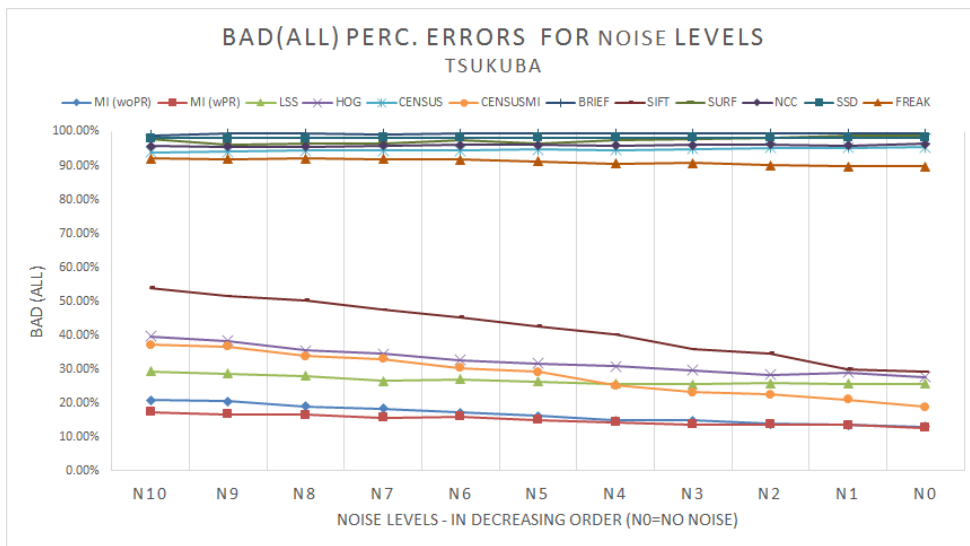
(b)

Figure B.9: RMS(all) and Bad(all) pixels percentage errors of all methods for 10 multi-modality levels for Cones image in Dataset #1 [Best viewed in color].

Figures B.10, B.11, B.12 and B.13 show the RMS and Bad pixel performances for each image pair separately, i.e. Tsukuba, Venus, Teddy and Cones, for the noise levels.

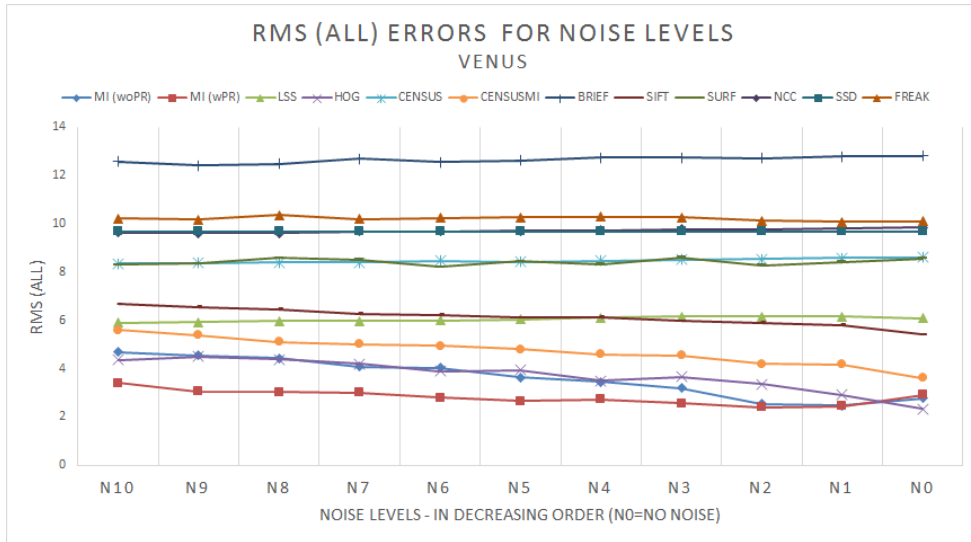


(a)

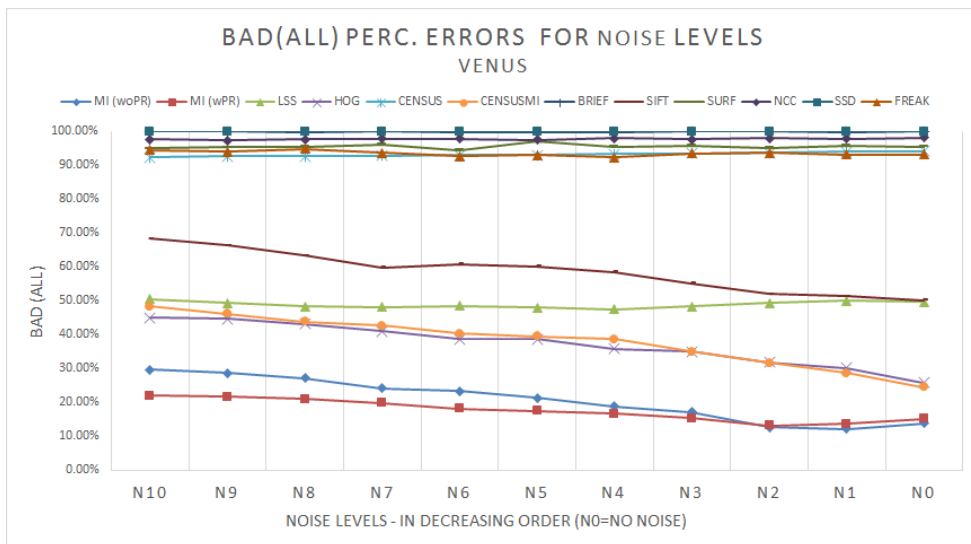


(b)

Figure B.10: RMS(all) and Bad(all) pixels percentage errors of all methods for 10 noise levels for Tsukuba image in Dataset #1 [Best viewed in color].

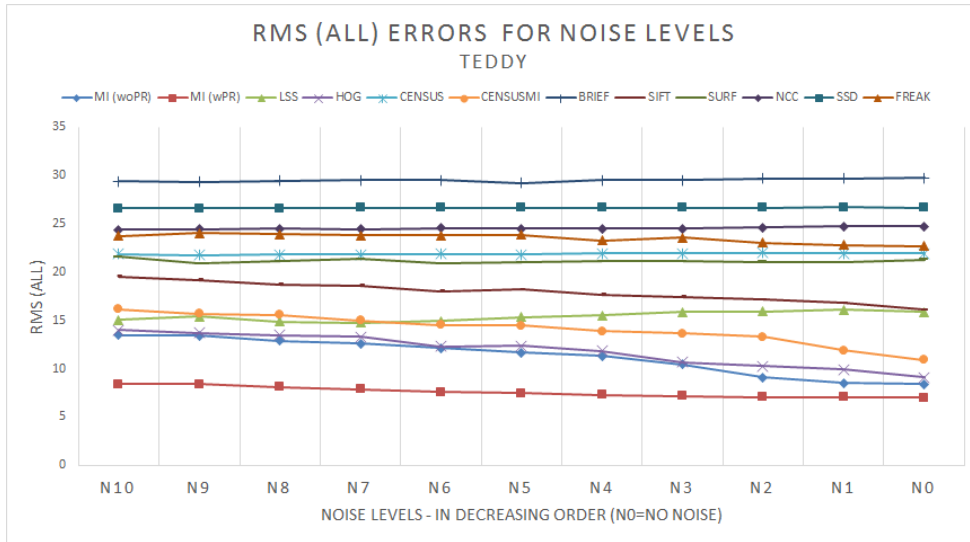


(a)

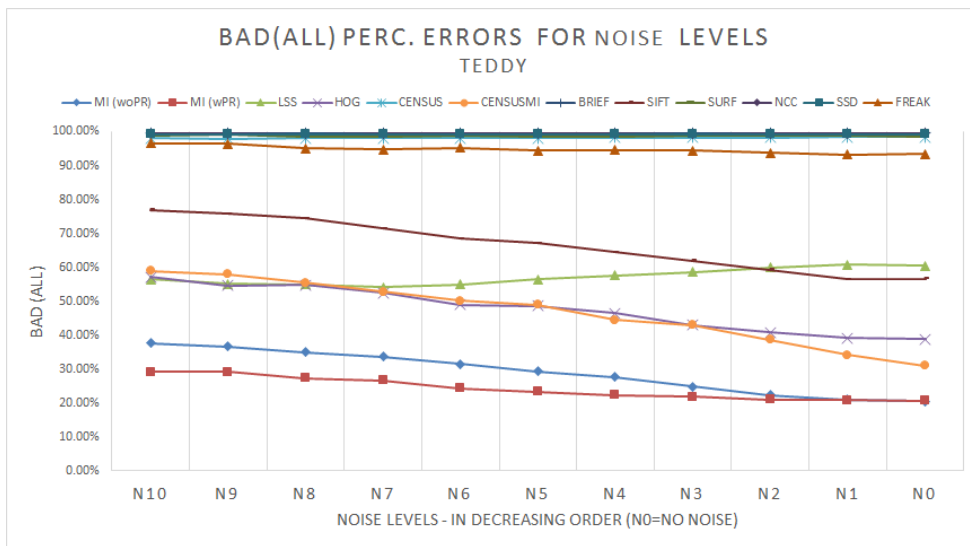


(b)

Figure B.11: RMS(all) and Bad(all) pixels percentage errors of all methods for 10 noise levels for Venus image in Dataset #1 [Best viewed in color].

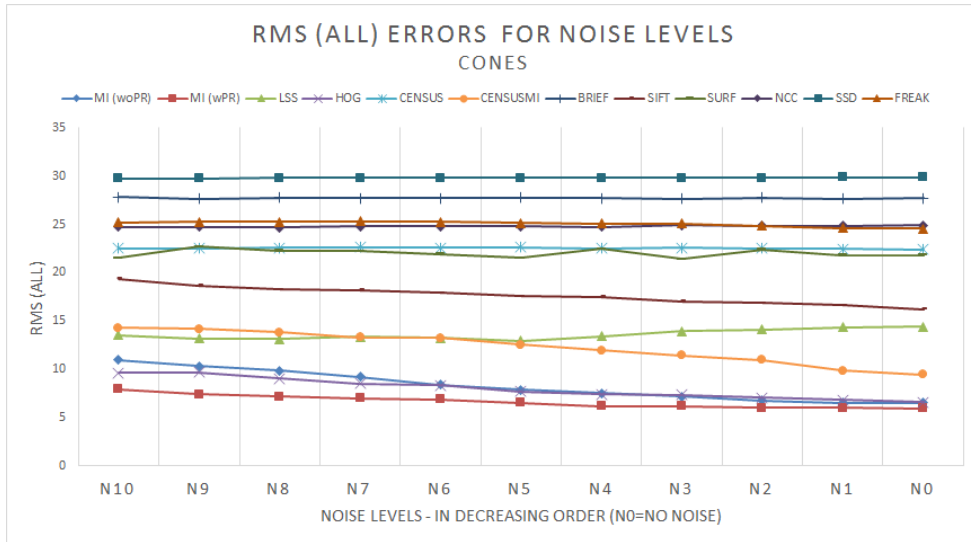


(a)

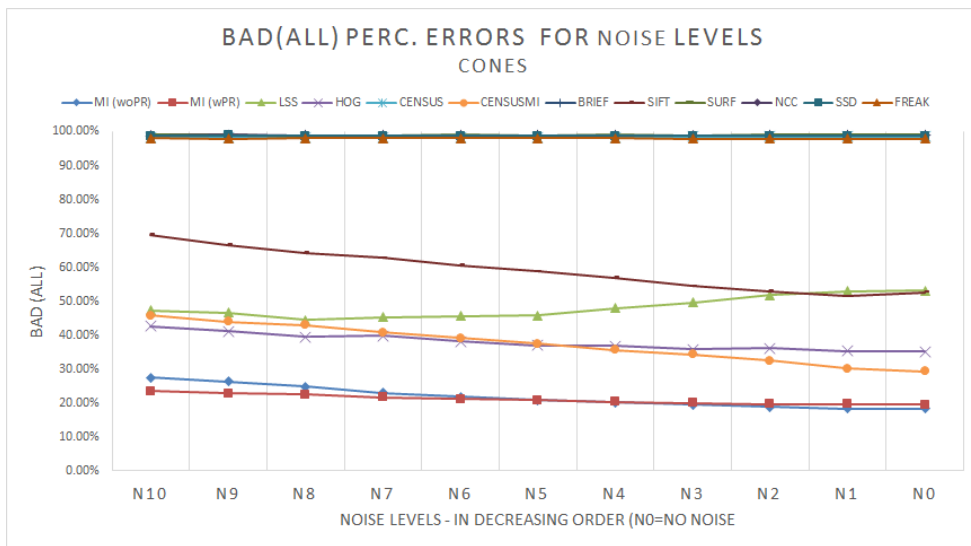


(b)

Figure B.12: RMS(all) and Bad(all) pixels percentage errors of all methods for 10 noise levels for Teddy image in Dataset #1 [Best viewed in color].



(a)



(b)

Figure B.13: RMS(all) and Bad(all) pixels percentage errors of all methods for 10 noise levels for Cones image in Dataset #1 [Best viewed in color].

In the below part, Table B.2 provides the experiment results of the similarity measures tested over Dataset #2, the Kinect Dataset.

Table B.2: Results on Dataset #2 Selected Image Pairs for all the Similarity Measures

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
Avg. All	BRIEF	Perc. Good Depth	20.4%	12.9%	9.0%	5.8%	48.0%
		Perc. Total Covg.	32.5%	11.1%	7.5%	5.0%	56.1%
	CENSUSMI	Perc. Good Depth	22.7%	13.4%	8.8%	4.5%	49.4%
		Perc. Total Covg.	35.7%	11.2%	7.3%	3.8%	58.1%
	CENSUS	Perc. Good Depth	10.8%	7.0%	5.3%	3.8%	26.9%
		Perc. Total Covg.	20.8%	6.3%	4.7%	3.3%	35.1%
	FREAK	Perc. Good Depth	27.2%	14.5%	9.3%	5.1%	56.2%
		Perc. Total Covg.	39.5%	12.3%	7.6%	4.3%	63.7%
	HOG	Perc. Good Depth	21.6%	13.7%	9.2%	5.9%	50.4%
		Perc. Total Covg.	34.8%	11.5%	7.7%	5.0%	59.0%
	LSS	Perc. Good Depth	23.8%	13.6%	7.8%	3.7%	48.9%
		Perc. Total Covg.	35.5%	11.5%	6.6%	3.1%	56.8%
	MI (wPR)	Perc. Good Depth	26.2%	13.6%	9.0%	4.3%	53.1%
		Perc. Total Covg.	38.7%	11.4%	7.4%	3.6%	61.1%
	NCC	Perc. Good Depth	20.4%	11.8%	8.0%	4.9%	45.1%
		Perc. Total Covg.	33.1%	10.0%	6.5%	4.1%	53.6%
	MI (woPR)	Perc. Good Depth	25.4%	13.9%	8.6%	4.0%	52.0%
		Perc. Total Covg.	38.4%	11.6%	7.0%	3.4%	60.4%
	SIFT	Perc. Good Depth	24.5%	13.2%	8.6%	4.3%	50.5%
		Perc. Total Covg.	37.6%	10.9%	7.0%	3.6%	59.2%
SSD	Perc. Good Depth	7.8%	5.8%	4.0%	3.4%	21.0%	
	Perc. Total Covg.	17.6%	5.1%	3.4%	2.9%	29.1%	
SURF	Perc. Good Depth	27.1%	15.0%	9.4%	5.8%	57.4%	
	Perc. Total Covg.	39.9%	12.5%	7.9%	4.9%	65.2%	
Kinect02	BRIEF	Perc. Good Depth	20.7%	16.7%	9.5%	6.1%	53.0%
		Perc. Total Covg.	29.0%	14.9%	8.5%	5.5%	57.9%
	CENSUSMI	Perc. Good Depth	26.6%	14.6%	9.2%	4.8%	55.2%
		Perc. Total Covg.	34.5%	13.0%	8.2%	4.3%	60.0%

Continued on next page

Table B.2 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	CENSUS	Perc. Good Depth	11.4%	7.8%	5.2%	3.6%	28.1%
		Perc. Total Covg.	17.6%	7.2%	4.9%	3.4%	33.1%
	FREAK	Perc. Good Depth	30.7%	18.5%	9.5%	5.0%	63.7%
		Perc. Total Covg.	38.0%	16.5%	8.5%	4.5%	67.5%
	HOG	Perc. Good Depth	25.6%	15.5%	9.0%	5.1%	55.0%
		Perc. Total Covg.	33.1%	13.9%	8.1%	4.5%	59.6%
	LSS	Perc. Good Depth	28.5%	13.7%	7.3%	3.5%	52.9%
		Perc. Total Covg.	34.4%	12.6%	6.7%	3.2%	56.8%
	MI (wPR)	Perc. Good Depth	26.7%	17.7%	8.9%	4.5%	57.8%
		Perc. Total Covg.	34.0%	16.0%	8.0%	4.1%	62.0%
	NCC	Perc. Good Depth	26.6%	16.3%	8.4%	5.0%	56.3%
		Perc. Total Covg.	33.8%	14.7%	7.6%	4.5%	60.6%
	MI (woPR)	Perc. Good Depth	26.4%	17.7%	8.6%	3.7%	56.4%
		Perc. Total Covg.	34.7%	15.7%	7.6%	3.3%	61.3%
	SIFT	Perc. Good Depth	31.0%	15.7%	7.7%	3.6%	58.0%
		Perc. Total Covg.	38.5%	14.0%	6.9%	3.2%	62.6%
	SSD	Perc. Good Depth	3.5%	3.2%	1.6%	1.2%	9.6%
		Perc. Total Covg.	4.9%	3.2%	1.6%	1.2%	10.8%
	SURF	Perc. Good Depth	25.1%	19.7%	14.7%	6.7%	66.2%
		Perc. Total Covg.	33.9%	17.4%	13.0%	5.9%	70.2%
Kinect03	BRIEF	Perc. Good Depth	12.0%	10.9%	11.4%	4.6%	38.9%
		Perc. Total Covg.	36.6%	7.9%	8.2%	3.3%	56.0%
	CENSUSMI	Perc. Good Depth	9.6%	13.5%	9.5%	4.8%	37.3%
		Perc. Total Covg.	37.0%	9.4%	6.6%	3.4%	56.4%
	CENSUS	Perc. Good Depth	6.0%	6.2%	6.7%	3.9%	22.9%
		Perc. Total Covg.	28.0%	4.8%	5.1%	3.0%	40.9%
	FREAK	Perc. Good Depth	16.2%	12.3%	12.8%	5.9%	47.3%
		Perc. Total Covg.	43.1%	8.4%	8.7%	4.0%	64.2%
	HOG	Perc. Good Depth	7.2%	13.9%	10.5%	4.9%	36.5%
		Perc. Total Covg.	35.2%	9.7%	7.3%	3.4%	55.7%

Continued on next page

Table B.2 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	LSS	Perc. Good Depth	10.6%	15.2%	8.1%	4.3%	38.2%
		Perc. Total Covg.	35.6%	11.0%	5.9%	3.1%	55.5%
	MI (wPR)	Perc. Good Depth	12.4%	14.4%	12.2%	4.1%	43.1%
		Perc. Total Covg.	39.7%	9.9%	8.4%	2.8%	60.8%
	NCC	Perc. Good Depth	11.5%	12.8%	13.8%	5.1%	43.2%
		Perc. Total Covg.	39.0%	8.8%	9.5%	3.5%	60.8%
	MI (woPR)	Perc. Good Depth	11.6%	15.2%	11.5%	4.0%	42.4%
		Perc. Total Covg.	39.7%	10.4%	7.9%	2.7%	60.7%
	SIFT	Perc. Good Depth	10.4%	15.1%	10.8%	4.4%	40.6%
		Perc. Total Covg.	39.3%	10.2%	7.3%	3.0%	59.8%
	SSD	Perc. Good Depth	6.0%	5.7%	6.6%	5.3%	23.5%
		Perc. Total Covg.	31.5%	4.1%	4.8%	3.8%	44.2%
	SURF	Perc. Good Depth	15.5%	15.7%	8.0%	5.2%	44.4%
		Perc. Total Covg.	41.7%	10.8%	5.5%	3.6%	61.7%
Kinect06	BRIEF	Perc. Good Depth	14.2%	14.7%	7.8%	8.2%	45.0%
		Perc. Total Covg.	22.0%	13.4%	7.1%	7.4%	49.9%
	CENSUSMI	Perc. Good Depth	22.6%	15.9%	10.6%	5.7%	54.8%
		Perc. Total Covg.	30.9%	14.2%	9.4%	5.1%	59.7%
	CENSUS	Perc. Good Depth	9.0%	7.5%	5.5%	5.4%	27.4%
		Perc. Total Covg.	14.8%	7.0%	5.2%	5.1%	32.0%
	FREAK	Perc. Good Depth	20.9%	17.8%	9.0%	6.8%	54.5%
		Perc. Total Covg.	28.9%	16.0%	8.1%	6.1%	59.1%
	HOG	Perc. Good Depth	20.8%	16.3%	10.0%	9.1%	56.2%
		Perc. Total Covg.	29.7%	14.5%	8.9%	8.1%	61.2%
	LSS	Perc. Good Depth	22.4%	17.2%	9.7%	5.1%	54.5%
		Perc. Total Covg.	30.5%	15.5%	8.7%	4.6%	59.2%
	MI (wPR)	Perc. Good Depth	25.4%	14.3%	10.1%	6.9%	56.7%
		Perc. Total Covg.	33.4%	12.8%	9.0%	6.1%	61.3%
NCC	Perc. Good Depth	15.0%	11.2%	6.1%	7.3%	39.7%	
	Perc. Total Covg.	23.3%	10.2%	5.5%	6.6%	45.6%	

Continued on next page

Table B.2 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	MI (woPR)	Perc. Good Depth	25.4%	15.1%	9.4%	6.5%	56.3%
		Perc. Total Covg.	33.4%	13.5%	8.4%	5.8%	61.0%
	SIFT	Perc. Good Depth	26.1%	13.8%	10.0%	6.2%	56.0%
		Perc. Total Covg.	33.7%	12.4%	8.9%	5.6%	60.6%
	SSD	Perc. Good Depth	7.5%	9.4%	5.3%	5.6%	27.8%
		Perc. Total Covg.	14.1%	8.7%	5.0%	5.2%	32.9%
	SURF	Perc. Good Depth	25.3%	14.6%	11.5%	8.7%	60.2%
		Perc. Total Covg.	33.8%	12.9%	10.2%	7.8%	64.7%
Kinect10	BRIEF	Perc. Good Depth	34.5%	9.3%	7.1%	4.2%	55.2%
		Perc. Total Covg.	42.3%	8.2%	6.3%	3.7%	60.5%
	CENSUSMI	Perc. Good Depth	32.1%	9.5%	5.8%	2.8%	50.2%
		Perc. Total Covg.	40.4%	8.4%	5.1%	2.4%	56.3%
	CENSUS	Perc. Good Depth	16.8%	6.5%	3.8%	2.0%	29.2%
		Perc. Total Covg.	22.8%	6.1%	3.5%	1.9%	34.3%
	FREAK	Perc. Good Depth	41.0%	9.4%	6.1%	2.7%	59.2%
		Perc. Total Covg.	47.9%	8.3%	5.3%	2.4%	64.0%
	HOG	Perc. Good Depth	32.9%	9.1%	7.4%	4.4%	53.7%
		Perc. Total Covg.	41.1%	8.0%	6.4%	3.8%	59.4%
	LSS	Perc. Good Depth	33.8%	8.1%	6.0%	1.8%	49.8%
		Perc. Total Covg.	41.5%	7.2%	5.3%	1.6%	55.6%
	MI (wPR)	Perc. Good Depth	40.4%	7.8%	4.9%	1.8%	55.0%
		Perc. Total Covg.	47.6%	6.9%	4.3%	1.6%	60.3%
	NCC	Perc. Good Depth	28.4%	7.0%	3.7%	2.1%	41.2%
		Perc. Total Covg.	36.0%	6.2%	3.3%	1.9%	47.5%
	MI (woPR)	Perc. Good Depth	38.3%	7.7%	4.8%	2.0%	52.8%
		Perc. Total Covg.	45.8%	6.8%	4.2%	1.8%	58.6%
	SIFT	Perc. Good Depth	30.5%	8.1%	5.8%	3.0%	47.3%
		Perc. Total Covg.	38.9%	7.1%	5.1%	2.6%	53.6%
SSD	Perc. Good Depth	14.2%	4.8%	2.6%	1.6%	23.2%	
	Perc. Total Covg.	19.8%	4.5%	2.4%	1.5%	28.2%	

Continued on next page

Table B.2 – *Continued from previous page*

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	SURF	Perc. Good Depth	42.5%	10.3%	3.3%	2.7%	58.7%
		Perc. Total Covg.	50.1%	8.9%	2.8%	2.3%	64.2%

APPENDIX C

EXPERIMENT RESULTS OF THE PROPOSED METHOD

C.1 Adaptive Windowing Step vs. State of the Art Similarity Measures for Dataset #1

Table C.1 provides the WTA performance results of the Adaptive Windowing Algorithm (ADAPMI) of the proposed method along with the similarity measures tested, using Dataset #1 for three different window sizes.

Table C.1: Results on Dataset #1 for the WTA performances of the Adaptive Windowing Algorithm (ADAPMI) of the proposed method along with the Similarity Measures Tested using Three Different Window Sizes

Image	Method	RMS (all) (9x9)*	RMS (all) (21x21)	RMS (all) (31x31)	Bad (all) (9x9)	Bad (all) (21x21)	Bad (all) (31x31)
Avg. All	AdapMI	5,349	4,128	3,769	17,14%	12,51%	11,43%
	MI (woPR)	10,292	4,995	4,057	43,01%	16,35%	14,13%
	MI (wPR)	7,642	4,480	3,861	31,41%	16,93%	15,33%
	LSS	12,699	9,857	8,518	59,85%	47,16%	42,50%
	HOG	6,843	5,080	4,568	40,40%	31,77%	29,55%
	CENSUS	14,223	14,998	15,433	95,70%	96,60%	96,82%
	CENSUSMI	11,428	6,618	4,673	61,74%	25,86%	18,37%
	BRIEF	19,402	19,924	20,260	98,92%	99,44%	99,41%

Continued on next page

Table C.1 – Continued from previous page

Image	Method	RMS (all) (9x9)*	RMS (all) (21x21)	RMS (all) (31x31)	Bad (all) (9x9)	Bad (all) (21x21)	Bad (all) (31x31)
	SIFT	11,934	10,188	9,139	65,93%	47,10%	40,58%
	SURF	14,152	14,660	15,108	96,15%	98,06%	98,33%
	NCC	15,803	16,796	17,768	96,93%	98,23%	98,63%
	SSD	17,839	18,411	18,735	98,36%	99,02%	98,87%
	FREAK	15,503	16,265	16,588	92,84%	93,54%	94,03%
Tsukuba	AdapMI	1,574	1,457	1,433	7,71%	6,41%	5,44%
	MI (woPR)	3,695	2,223	1,930	31,02%	12,93%	11,96%
	MI (wPR)	2,647	2,023	1,832	20,28%	12,53%	12,29%
	LSS	4,153	3,073	2,358	38,63%	25,64%	21,70%
	HOG	2,593	2,293	2,233	27,25%	27,59%	28,21%
	CENSUS	6,372	7,051	7,320	93,95%	95,33%	95,64%
	CENSUSMI	4,317	2,493	2,237	47,63%	18,77%	16,41%
	BRIEF	8,730	9,433	9,469	98,67%	99,34%	99,28%
	SIFT	4,278	3,034	2,656	46,91%	29,06%	24,91%
	SURF	6,347	6,988	7,179	94,79%	98,88%	99,51%
	NCC	6,920	7,696	7,945	93,29%	96,42%	97,34%
	SSD	6,917	7,436	7,516	97,17%	98,19%	98,05%
	FREAK	6,774	7,677	8,022	88,04%	89,79%	91,28%
Venus	AdapMI	1,799	1,157	1,008	9,24%	6,53%	6,01%
	MI (woPR)	5,607	2,783	2,084	38,11%	13,77%	9,26%
	MI (wPR)	4,153	2,915	2,462	28,50%	15,21%	11,77%
	LSS	7,842	6,092	5,308	69,41%	49,62%	41,01%
	HOG	3,887	2,338	1,774	37,44%	25,78%	20,41%
	CENSUS	8,108	8,603	8,826	93,43%	94,13%	94,36%
	CENSUSMI	6,342	3,615	1,959	60,65%	24,50%	10,73%
	BRIEF	11,965	12,801	12,906	99,17%	99,94%	99,99%
	SIFT	6,583	5,423	4,635	66,66%	50,16%	43,32%
	SURF	8,273	8,558	8,817	94,79%	95,51%	95,81%

Continued on next page

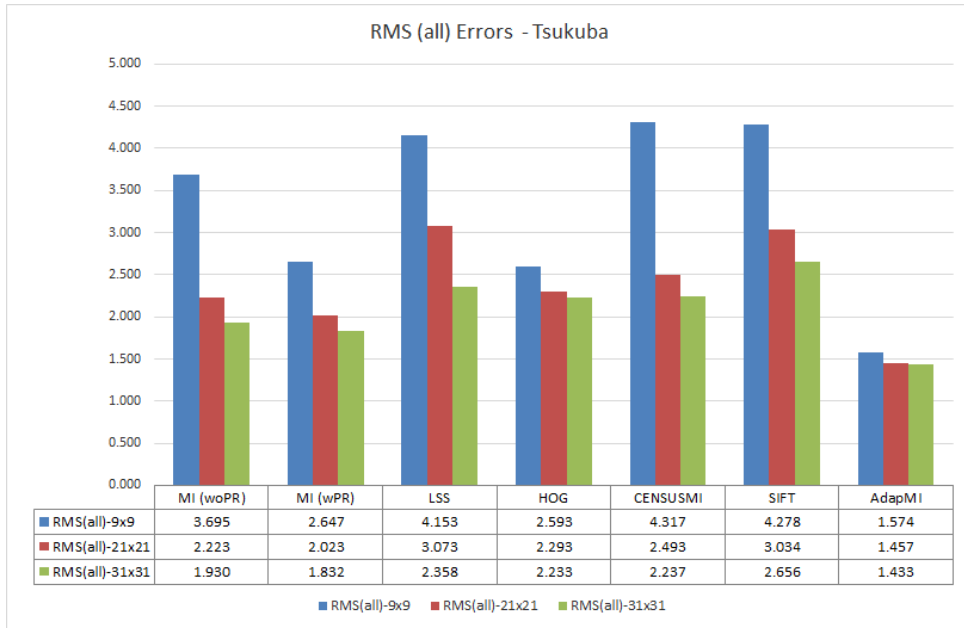
Table C.1 – Continued from previous page

Image	Method	RMS (all) (9x9)*	RMS (all) (21x21)	RMS (all) (31x31)	Bad (all) (9x9)	Bad (all) (21x21)	Bad (all) (31x31)
	NCC	9,195	9,848	10,426	96,61%	98,09%	98,78%
	SSD	9,663	9,675	9,734	99,58%	99,94%	100,00%
	FREAK	9,344	10,094	10,384	92,28%	93,15%	93,40%
Teddy	AdapMI	8,008	7,037	6,650	24,29%	18,24%	17,11%
	MI (woPR)	15,840	8,433	6,415	55,65%	20,40%	15,94%
	MI (wPR)	11,210	7,019	6,206	39,49%	20,60%	17,49%
	LSS	18,474	15,853	14,545	69,29%	60,37%	57,87%
	HOG	11,886	9,104	8,489	50,93%	38,67%	35,55%
	CENSUS	20,471	21,910	22,633	97,53%	98,37%	98,56%
	CENSUSMI	17,527	10,909	7,920	72,81%	30,93%	21,39%
	BRIEF	28,744	29,732	30,222	99,21%	99,64%	99,52%
	SIFT	18,196	16,087	14,649	74,86%	56,53%	48,49%
	SURF	20,464	21,266	21,906	97,56%	98,65%	98,82%
	NCC	22,869	24,716	26,368	98,76%	99,50%	99,29%
	SSD	25,317	26,673	27,573	98,57%	99,19%	99,19%
	FREAK	22,329	22,703	23,025	94,49%	93,32%	93,06%
Cones	AdapMI	10,016	6,860	5,985	27,3%	18,9%	17,2%
	MI (woPR)	16,025	6,540	5,797	47,3%	18,3%	19,4%
	MI (wPR)	12,557	5,964	4,944	37,3%	19,4%	19,8%
	LSS	20,328	14,407	11,862	62,1%	53,0%	49,4%
	HOG	9,006	6,586	5,776	46,0%	35,0%	34,0%
	CENSUS	21,941	22,430	22,954	97,9%	98,6%	98,7%
	CENSUSMI	17,525	9,454	6,576	65,9%	29,3%	25,0%
	BRIEF	28,168	27,728	28,445	98,6%	98,9%	98,9%
	SIFT	18,678	16,210	14,616	75,3%	52,7%	45,6%
	SURF	21,525	21,826	22,530	97,5%	99,2%	99,2%
	NCC	24,229	24,924	26,333	99,1%	98,9%	99,1%
	SSD	29,457	29,859	30,118	98,1%	98,8%	98,2%

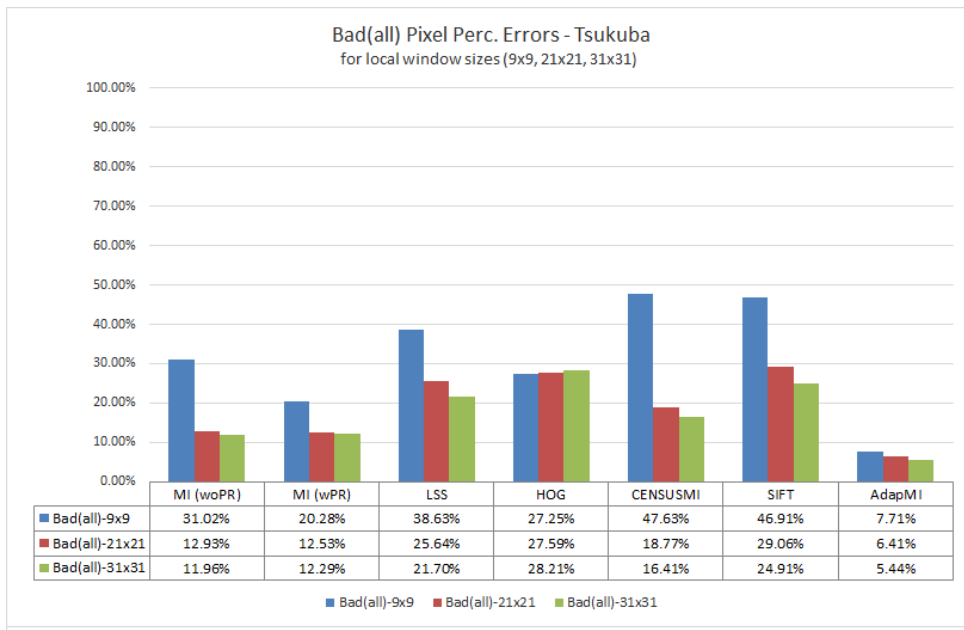
Continued on next page

Table C.1 – *Continued from previous page*

Image	Method	RMS (all) (9x9)*	RMS (all) (21x21)	RMS (all) (31x31)	Bad (all) (9x9)	Bad (all) (21x21)	Bad (all) (31x31)
	FREAK	23,565	24,586	24,922	96,6%	97,9%	98,4%



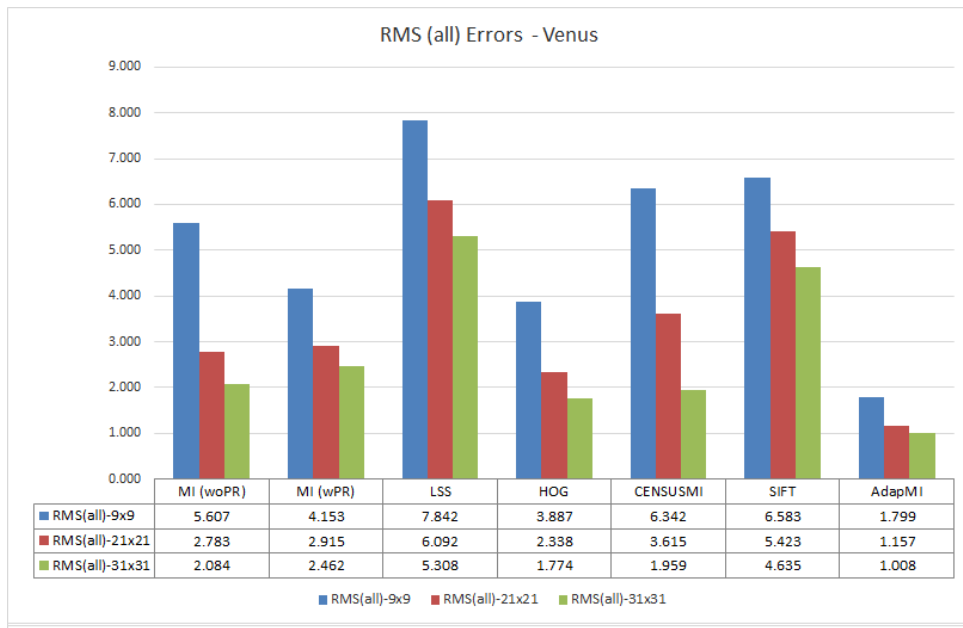
(a)



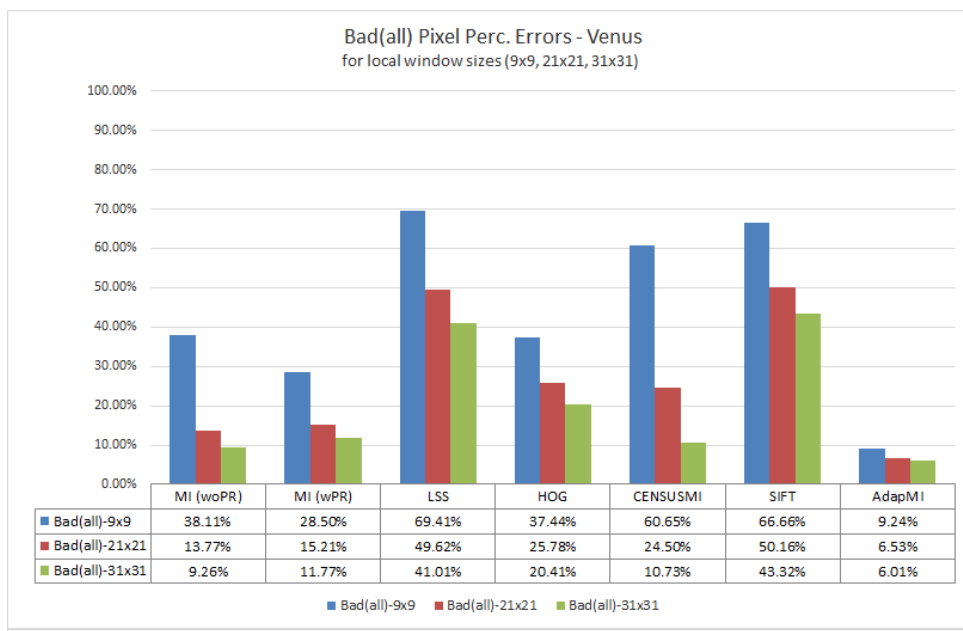
(b)

Figure C.1: RMS(all) and Bad(all) pixels percentage errors of "WTA" performance of the Adaptive Windowing Algorithm (ADAPMI) and the state of the art similarity measures for three different window sizes for Tsukuba image in Dataset1

Figures C.1, C.2, C.3 and C.4 depict these RMS and Bad pixel performances for each image separately in Dataset #1, i.e. Tsukuba, Venus, Teddy and Cones for the three window sizes and along with the leading similarity measures.

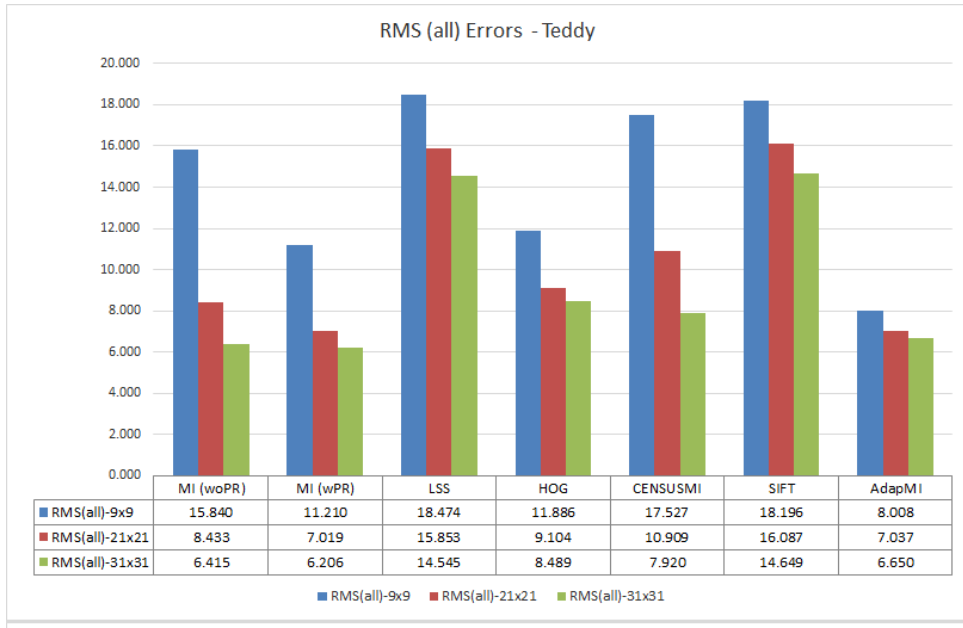


(a)

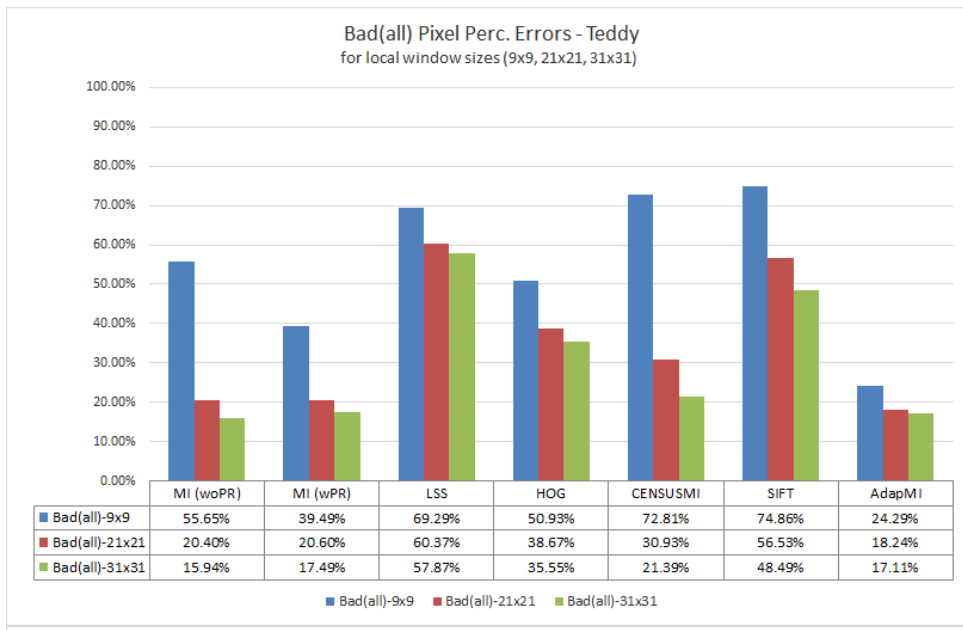


(b)

Figure C.2: RMS(all) and Bad(all) pixels percentage errors of "WTA" performance of the Adaptive Windowing Algorithm (ADAPMI) and the state of the art similarity measures for three different window sizes for Venus image in Dataset1

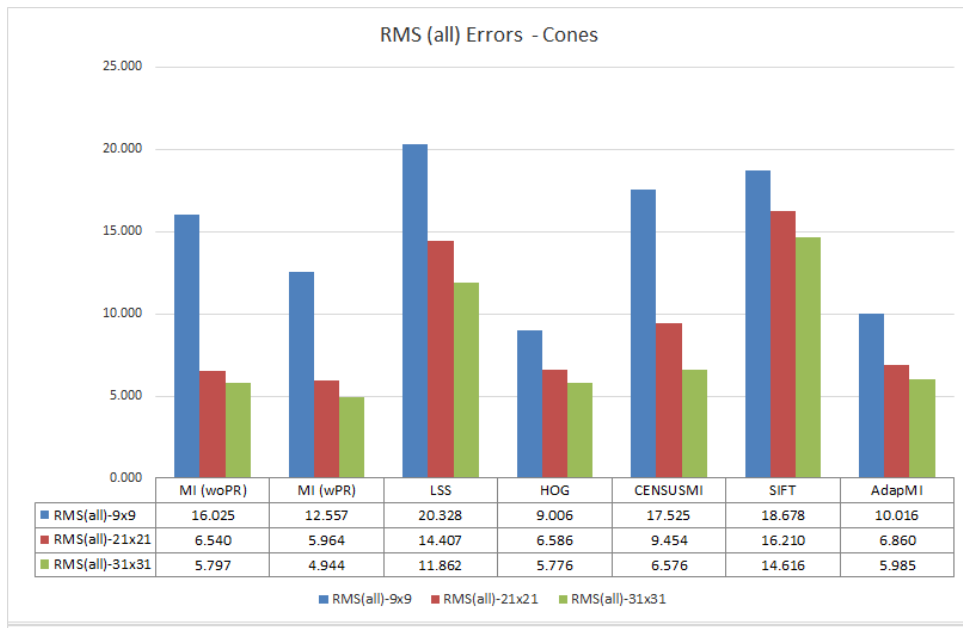


(a)

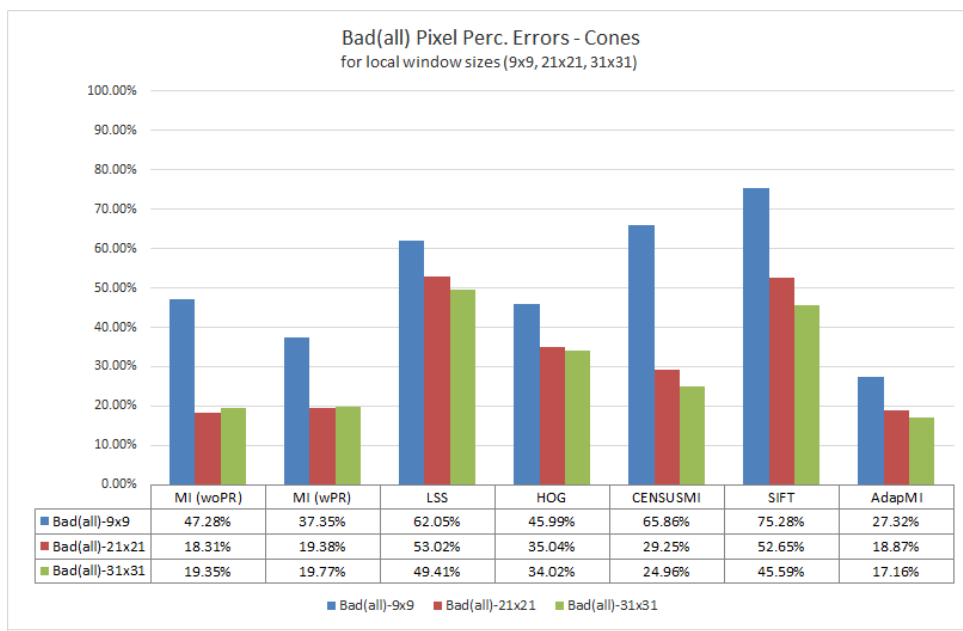


(b)

Figure C.3: RMS(all) and Bad(all) pixels percentage errors of "WTA" performance of the Adaptive Windowing Algorithm (ADAPMI) and the state of the art similarity measures for three different window sizes for Teddy image in Dataset1



(a)



(b)

Figure C.4: RMS(all) and Bad(all) pixels percentage errors of "WTA" performance of the Adaptive Windowing Algorithm (ADAPMI) and the state of the art similarity measures for three different window sizes for Cones image in Dataset1

C.2 Adaptive Windowing vs. State of the Art Similarity Measures using Dataset #2, The Kinect Dataset

Table C.2 provides the experiment results of WTA performance obtained from the adaptive windowing algorithm (AdapMI) of the proposed method for the 1st iteration and the similarity measures tested over Dataset #2, the Kinect Dataset.

Table C.2: Results on Dataset #2 Selected Image Pairs for WTA Performances of the Adaptive Windowing Algorithm (AdapMI) of the Proposed Method and the Similarity Measures Tested

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
Avg. All	AdapMI	Perc. Good Depth	30.5%	15.9%	16.7%	5.1%	68.3%
		Perc. Total Covg.	44.3%	12.9%	12.7%	4.2%	74.1%
	BRIEF	Perc. Good Depth	20.4%	12.9%	9.0%	5.8%	48.0%
		Perc. Total Covg.	32.5%	11.1%	7.5%	5.0%	56.1%
	CENSUSMI	Perc. Good Depth	22.7%	13.4%	8.8%	4.5%	49.4%
		Perc. Total Covg.	35.7%	11.2%	7.3%	3.8%	58.1%
	CENSUS	Perc. Good Depth	10.8%	7.0%	5.3%	3.8%	26.9%
		Perc. Total Covg.	20.8%	6.3%	4.7%	3.3%	35.1%
	FREAK	Perc. Good Depth	27.2%	14.5%	9.3%	5.1%	56.2%
		Perc. Total Covg.	39.5%	12.3%	7.6%	4.3%	63.7%
	HOG	Perc. Good Depth	21.6%	13.7%	9.2%	5.9%	50.4%
		Perc. Total Covg.	34.8%	11.5%	7.7%	5.0%	59.0%
	LSS	Perc. Good Depth	23.8%	13.6%	7.8%	3.7%	48.9%
		Perc. Total Covg.	35.5%	11.5%	6.6%	3.1%	56.8%
	MI (wPR)	Perc. Good Depth	26.2%	13.6%	9.0%	4.3%	53.1%
		Perc. Total Covg.	38.7%	11.4%	7.4%	3.6%	61.1%
	NCC	Perc. Good Depth	20.4%	11.8%	8.0%	4.9%	45.1%
		Perc. Total Covg.	33.1%	10.0%	6.5%	4.1%	53.6%
	MI (woPR)	Perc. Good Depth	25.4%	13.9%	8.6%	4.0%	52.0%
		Perc. Total Covg.	38.4%	11.6%	7.0%	3.4%	60.4%

Continued on next page

Table C.2 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	SIFT	Perc. Good Depth	24.5%	13.2%	8.6%	4.3%	50.5%
		Perc. Total Covg.	37.6%	10.9%	7.0%	3.6%	59.2%
	SSD	Perc. Good Depth	7.8%	5.8%	4.0%	3.4%	21.0%
		Perc. Total Covg.	17.6%	5.1%	3.4%	2.9%	29.1%
	SURF	Perc. Good Depth	27.1%	15.0%	9.4%	5.8%	57.4%
		Perc. Total Covg.	39.9%	12.5%	7.9%	4.9%	65.2%
Kinect02	AdapMI	Perc. Good Depth	28.1%	22.7%	10.9%	6.9%	68.7%
		Perc. Total Covg.	37.7%	19.7%	9.4%	6.0%	72.8%
	BRIEF	Perc. Good Depth	20.7%	16.7%	9.5%	6.1%	53.0%
		Perc. Total Covg.	29.0%	14.9%	8.5%	5.5%	57.9%
	CENSUSMI	Perc. Good Depth	26.6%	14.6%	9.2%	4.8%	55.2%
		Perc. Total Covg.	34.5%	13.0%	8.2%	4.3%	60.0%
	CENSUS	Perc. Good Depth	11.4%	7.8%	5.2%	3.6%	28.1%
		Perc. Total Covg.	17.6%	7.2%	4.9%	3.4%	33.1%
	FREAK	Perc. Good Depth	30.7%	18.5%	9.5%	5.0%	63.7%
		Perc. Total Covg.	38.0%	16.5%	8.5%	4.5%	67.5%
	HOG	Perc. Good Depth	25.6%	15.5%	9.0%	5.1%	55.0%
		Perc. Total Covg.	33.1%	13.9%	8.1%	4.5%	59.6%
	LSS	Perc. Good Depth	28.5%	13.7%	7.3%	3.5%	52.9%
		Perc. Total Covg.	34.4%	12.6%	6.7%	3.2%	56.8%
	MI (wPR)	Perc. Good Depth	26.7%	17.7%	8.9%	4.5%	57.8%
		Perc. Total Covg.	34.0%	16.0%	8.0%	4.1%	62.0%
	NCC	Perc. Good Depth	26.6%	16.3%	8.4%	5.0%	56.3%
		Perc. Total Covg.	33.8%	14.7%	7.6%	4.5%	60.6%
	MI (woPR)	Perc. Good Depth	26.4%	17.7%	8.6%	3.7%	56.4%
		Perc. Total Covg.	34.7%	15.7%	7.6%	3.3%	61.3%
	SIFT	Perc. Good Depth	31.0%	15.7%	7.7%	3.6%	58.0%
		Perc. Total Covg.	38.5%	14.0%	6.9%	3.2%	62.6%
	SSD	Perc. Good Depth	3.5%	3.2%	1.6%	1.2%	9.6%
		Perc. Total Covg.	4.9%	3.2%	1.6%	1.2%	10.8%

Continued on next page

Table C.2 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	SURF	Perc. Good Depth	25.1%	19.7%	14.7%	6.7%	66.2%
		Perc. Total Covg.	33.9%	17.4%	13.0%	5.9%	70.2%
Kinect03	AdapMI	Perc. Good Depth	12.2%	18.4%	35.2%	4.8%	70.6%
		Perc. Total Covg.	42.1%	12.2%	23.2%	3.1%	80.6%
	BRIEF	Perc. Good Depth	12.0%	10.9%	11.4%	4.6%	38.9%
		Perc. Total Covg.	36.6%	7.9%	8.2%	3.3%	56.0%
	CENSUSMI	Perc. Good Depth	9.6%	13.5%	9.5%	4.8%	37.3%
		Perc. Total Covg.	37.0%	9.4%	6.6%	3.4%	56.4%
	CENSUS	Perc. Good Depth	6.0%	6.2%	6.7%	3.9%	22.9%
		Perc. Total Covg.	28.0%	4.8%	5.1%	3.0%	40.9%
	FREAK	Perc. Good Depth	16.2%	12.3%	12.8%	5.9%	47.3%
		Perc. Total Covg.	43.1%	8.4%	8.7%	4.0%	64.2%
	HOG	Perc. Good Depth	7.2%	13.9%	10.5%	4.9%	36.5%
		Perc. Total Covg.	35.2%	9.7%	7.3%	3.4%	55.7%
	LSS	Perc. Good Depth	10.6%	15.2%	8.1%	4.3%	38.2%
		Perc. Total Covg.	35.6%	11.0%	5.9%	3.1%	55.5%
	MI (wPR)	Perc. Good Depth	12.4%	14.4%	12.2%	4.1%	43.1%
		Perc. Total Covg.	39.7%	9.9%	8.4%	2.8%	60.8%
	NCC	Perc. Good Depth	11.5%	12.8%	13.8%	5.1%	43.2%
		Perc. Total Covg.	39.0%	8.8%	9.5%	3.5%	60.8%
	MI (woPR)	Perc. Good Depth	11.6%	15.2%	11.5%	4.0%	42.4%
		Perc. Total Covg.	39.7%	10.4%	7.9%	2.7%	60.7%
	SIFT	Perc. Good Depth	10.4%	15.1%	10.8%	4.4%	40.6%
		Perc. Total Covg.	39.3%	10.2%	7.3%	3.0%	59.8%
	SSD	Perc. Good Depth	6.0%	5.7%	6.6%	5.3%	23.5%
		Perc. Total Covg.	31.5%	4.1%	4.8%	3.8%	44.2%
SURF	Perc. Good Depth	15.5%	15.7%	8.0%	5.2%	44.4%	
	Perc. Total Covg.	41.7%	10.8%	5.5%	3.6%	61.7%	
Kinect06	AdapMI	Perc. Good Depth	35.8%	15.4%	14.2%	6.0%	71.4%
		Perc. Total Covg.	43.4%	13.6%	12.5%	5.3%	74.8%

Continued on next page

Table C.2 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	BRIEF	Perc. Good Depth	14.2%	14.7%	7.8%	8.2%	45.0%
		Perc. Total Covg.	22.0%	13.4%	7.1%	7.4%	49.9%
	CENSUSMI	Perc. Good Depth	22.6%	15.9%	10.6%	5.7%	54.8%
		Perc. Total Covg.	30.9%	14.2%	9.4%	5.1%	59.7%
	CENSUS	Perc. Good Depth	9.0%	7.5%	5.5%	5.4%	27.4%
		Perc. Total Covg.	14.8%	7.0%	5.2%	5.1%	32.0%
	FREAK	Perc. Good Depth	20.9%	17.8%	9.0%	6.8%	54.5%
		Perc. Total Covg.	28.9%	16.0%	8.1%	6.1%	59.1%
	HOG	Perc. Good Depth	20.8%	16.3%	10.0%	9.1%	56.2%
		Perc. Total Covg.	29.7%	14.5%	8.9%	8.1%	61.2%
	LSS	Perc. Good Depth	22.4%	17.2%	9.7%	5.1%	54.5%
		Perc. Total Covg.	30.5%	15.5%	8.7%	4.6%	59.2%
	MI (wPR)	Perc. Good Depth	25.4%	14.3%	10.1%	6.9%	56.7%
		Perc. Total Covg.	33.4%	12.8%	9.0%	6.1%	61.3%
	NCC	Perc. Good Depth	15.0%	11.2%	6.1%	7.3%	39.7%
		Perc. Total Covg.	23.3%	10.2%	5.5%	6.6%	45.6%
	MI (woPR)	Perc. Good Depth	25.4%	15.1%	9.4%	6.5%	56.3%
		Perc. Total Covg.	33.4%	13.5%	8.4%	5.8%	61.0%
	SIFT	Perc. Good Depth	26.1%	13.8%	10.0%	6.2%	56.0%
		Perc. Total Covg.	33.7%	12.4%	8.9%	5.6%	60.6%
SSD	Perc. Good Depth	7.5%	9.4%	5.3%	5.6%	27.8%	
	Perc. Total Covg.	14.1%	8.7%	5.0%	5.2%	32.9%	
SURF	Perc. Good Depth	25.3%	14.6%	11.5%	8.7%	60.2%	
	Perc. Total Covg.	33.8%	12.9%	10.2%	7.8%	64.7%	
Kinect10	AdapMI	Perc. Good Depth	45.9%	7.2%	6.6%	2.6%	62.4%
		Perc. Total Covg.	54.2%	6.1%	5.6%	2.2%	68.1%
	BRIEF	Perc. Good Depth	34.5%	9.3%	7.1%	4.2%	55.2%
		Perc. Total Covg.	42.3%	8.2%	6.3%	3.7%	60.5%
	CENSUSMI	Perc. Good Depth	32.1%	9.5%	5.8%	2.8%	50.2%
		Perc. Total Covg.	40.4%	8.4%	5.1%	2.4%	56.3%

Continued on next page

Table C.2 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	CENSUS	Perc. Good Depth	16.8%	6.5%	3.8%	2.0%	29.2%
		Perc. Total Covg.	22.8%	6.1%	3.5%	1.9%	34.3%
	FREAK	Perc. Good Depth	41.0%	9.4%	6.1%	2.7%	59.2%
		Perc. Total Covg.	47.9%	8.3%	5.3%	2.4%	64.0%
	HOG	Perc. Good Depth	32.9%	9.1%	7.4%	4.4%	53.7%
		Perc. Total Covg.	41.1%	8.0%	6.4%	3.8%	59.4%
	LSS	Perc. Good Depth	33.8%	8.1%	6.0%	1.8%	49.8%
		Perc. Total Covg.	41.5%	7.2%	5.3%	1.6%	55.6%
	MI (wPR)	Perc. Good Depth	40.4%	7.8%	4.9%	1.8%	55.0%
		Perc. Total Covg.	47.6%	6.9%	4.3%	1.6%	60.3%
	NCC	Perc. Good Depth	28.4%	7.0%	3.7%	2.1%	41.2%
		Perc. Total Covg.	36.0%	6.2%	3.3%	1.9%	47.5%
	MI (woPR)	Perc. Good Depth	38.3%	7.7%	4.8%	2.0%	52.8%
		Perc. Total Covg.	45.8%	6.8%	4.2%	1.8%	58.6%
	SIFT	Perc. Good Depth	30.5%	8.1%	5.8%	3.0%	47.3%
		Perc. Total Covg.	38.9%	7.1%	5.1%	2.6%	53.6%
	SSD	Perc. Good Depth	14.2%	4.8%	2.6%	1.6%	23.2%
		Perc. Total Covg.	19.8%	4.5%	2.4%	1.5%	28.2%
	SURF	Perc. Good Depth	42.5%	10.3%	3.3%	2.7%	58.7%
		Perc. Total Covg.	50.1%	8.9%	2.8%	2.3%	64.2%

C.3 Dataset #2 Results of the Proposed Method All Steps

In the below part, Table C.3 provides the experiment results obtained from the whole Dataset #2 - the Kinect Dataset (which includes 24 multi-modal stereo image pairs) for the steps of the proposed method along with two iterations applied.

Table C.3: Results on Dataset #2 for Proposed Method in two iterations

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
Avg. All	WTA - Iter1	Perc. Good Depth	31%	15%	11%	5%	63%
		Perc. Total Covg.	43%	13%	9%	5%	69%
	WTA - Iter2	Perc. Good Depth	33%	16%	11%	5%	65%
		Perc. Total Covg.	44%	13%	9%	5%	71%
	Agg. - Iter1	Perc. Good Depth	35%	17%	11%	6%	68%
		Perc. Total Covg.	46%	14%	9%	5%	74%
	Agg. - Iter2	Perc. Good Depth	36%	17%	11%	5%	69%
		Perc. Total Covg.	48%	14%	9%	5%	75%
	PFIT - Iter1	Perc. Good Depth	39%	17%	9%	6%	71%
		Perc. Total Covg.	50%	14%	8%	5%	76%
	PFIT - Iter2	Perc. Good Depth	41%	16%	9%	6%	72%
		Perc. Total Covg.	52%	13%	7%	5%	77%
Kinect01	WTA - Iter1	Perc. Good Depth	40.1%	19.7%	7.8%	4.0%	71.7%
		Perc. Total Covg.	62.5%	12.4%	4.9%	2.5%	82.3%
	WTA - Iter2	Perc. Good Depth	43.8%	18.0%	7.5%	3.3%	72.7%
		Perc. Total Covg.	64.9%	11.3%	4.7%	2.1%	82.9%
	Agg. - Iter1	Perc. Good Depth	44.8%	20.1%	8.5%	3.9%	77.3%
		Perc. Total Covg.	65.7%	12.5%	5.3%	2.5%	85.9%
	Agg. - Iter2	Perc. Good Depth	48.1%	19.1%	7.0%	3.8%	78.1%
		Perc. Total Covg.	67.7%	11.9%	4.4%	2.4%	86.4%
	PFIT - Iter1	Perc. Good Depth	46.7%	19.3%	7.7%	4.2%	77.9%
		Perc. Total Covg.	66.9%	12.0%	4.8%	2.6%	86.3%
	PFIT - Iter2	Perc. Good Depth	57.3%	16.3%	7.1%	3.1%	83.8%
		Perc. Total Covg.	73.6%	10.1%	4.4%	1.9%	90.0%
Kinect02	WTA - Iter1	Perc. Good Depth	28.1%	22.7%	10.9%	6.9%	68.7%
		Perc. Total Covg.	37.7%	19.7%	9.4%	6.0%	72.8%
	WTA - Iter2	Perc. Good Depth	29.4%	22.8%	11.8%	7.6%	71.5%
		Perc. Total Covg.	38.7%	19.8%	10.2%	6.6%	75.3%

Continued on next page

Table C.3 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	Agg. - Iter1	Perc. Good Depth	31.5%	24.4%	12.0%	6.7%	74.6%
		Perc. Total Covg.	40.7%	21.1%	10.4%	5.8%	78.0%
	Agg. - Iter2	Perc. Good Depth	32.5%	24.8%	12.4%	7.3%	76.9%
		Perc. Total Covg.	41.6%	21.5%	10.7%	6.3%	80.0%
	PFIT - Iter1	Perc. Good Depth	34.2%	24.7%	14.7%	6.5%	80.2%
		Perc. Total Covg.	43.2%	21.4%	12.7%	5.6%	82.9%
	PFIT - Iter2	Perc. Good Depth	34.9%	25.4%	13.1%	6.5%	80.0%
		Perc. Total Covg.	43.7%	22.0%	11.3%	5.7%	82.7%
Kinect03	WTA - Iter1	Perc. Good Depth	12.2%	18.4%	35.2%	4.8%	70.6%
		Perc. Total Covg.	42.1%	12.2%	23.2%	3.1%	80.6%
	WTA - Iter2	Perc. Good Depth	14.5%	20.1%	36.3%	5.0%	75.8%
		Perc. Total Covg.	42.8%	13.4%	24.3%	3.3%	83.8%
	Agg. - Iter1	Perc. Good Depth	14.7%	25.9%	32.7%	4.7%	78.1%
		Perc. Total Covg.	45.7%	16.5%	20.8%	3.0%	86.0%
	Agg. - Iter2	Perc. Good Depth	15.5%	27.5%	33.9%	4.2%	81.1%
		Perc. Total Covg.	45.9%	17.6%	21.7%	2.7%	87.9%
	PFIT - Iter1	Perc. Good Depth	18.8%	27.1%	23.4%	12.4%	81.8%
		Perc. Total Covg.	48.3%	17.3%	14.9%	7.9%	88.4%
	PFIT - Iter2	Perc. Good Depth	16.0%	30.7%	27.9%	9.2%	83.7%
		Perc. Total Covg.	46.5%	19.5%	17.7%	5.8%	89.7%
Kinect04	WTA - Iter1	Perc. Good Depth	19.7%	11.5%	9.6%	8.8%	49.6%
		Perc. Total Covg.	52.1%	6.9%	5.7%	5.3%	69.9%
	WTA - Iter2	Perc. Good Depth	20.5%	11.7%	9.5%	10.2%	51.9%
		Perc. Total Covg.	52.7%	7.0%	5.7%	6.1%	71.4%
	Agg. - Iter1	Perc. Good Depth	22.5%	12.6%	9.7%	8.3%	53.1%
		Perc. Total Covg.	54.9%	7.3%	5.6%	4.8%	72.7%
	Agg. - Iter2	Perc. Good Depth	23.0%	12.3%	9.6%	11.2%	56.2%
		Perc. Total Covg.	55.4%	7.1%	5.6%	6.5%	74.6%
	PFIT - Iter1	Perc. Good Depth	22.5%	11.6%	16.4%	9.8%	60.2%
		Perc. Total Covg.	55.5%	6.6%	9.4%	5.6%	77.2%

Continued on next page

Table C.3 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	PFIT - Iter2	Perc. Good Depth	23.4%	11.4%	10.3%	15.4%	60.6%
		Perc. Total Covg.	54.8%	6.7%	6.1%	9.1%	76.7%
Kinect05	WTA - Iter1	Perc. Good Depth	22.9%	16.1%	10.3%	2.5%	51.8%
		Perc. Total Covg.	50.8%	10.3%	6.6%	1.6%	69.2%
	WTA - Iter2	Perc. Good Depth	22.8%	14.2%	12.3%	2.5%	51.9%
		Perc. Total Covg.	51.2%	9.0%	7.8%	1.6%	69.6%
	Agg. - Iter1	Perc. Good Depth	25.6%	16.5%	9.1%	2.1%	53.2%
		Perc. Total Covg.	52.6%	10.5%	5.8%	1.3%	70.2%
	Agg. - Iter2	Perc. Good Depth	26.3%	14.3%	11.0%	1.6%	53.3%
		Perc. Total Covg.	53.3%	9.1%	7.0%	1.0%	70.4%
	PFIT - Iter1	Perc. Good Depth	31.5%	15.1%	5.4%	4.0%	55.9%
		Perc. Total Covg.	56.7%	9.5%	3.4%	2.5%	72.1%
	PFIT - Iter2	Perc. Good Depth	32.1%	12.5%	6.4%	6.4%	57.3%
		Perc. Total Covg.	57.2%	7.9%	4.0%	4.0%	73.1%
Kinect06	WTA - Iter1	Perc. Good Depth	35.8%	15.4%	14.2%	6.0%	71.4%
		Perc. Total Covg.	43.4%	13.6%	12.5%	5.3%	74.8%
	WTA - Iter2	Perc. Good Depth	38.3%	15.0%	13.1%	5.8%	72.3%
		Perc. Total Covg.	45.6%	13.2%	11.6%	5.2%	75.6%
	Agg. - Iter1	Perc. Good Depth	41.6%	13.9%	14.6%	7.2%	77.3%
		Perc. Total Covg.	48.6%	12.2%	12.8%	6.4%	80.0%
	Agg. - Iter2	Perc. Good Depth	43.0%	12.5%	14.4%	6.3%	76.3%
		Perc. Total Covg.	49.9%	11.0%	12.6%	5.6%	79.1%
	PFIT - Iter1	Perc. Good Depth	39.6%	19.7%	11.0%	5.3%	75.7%
		Perc. Total Covg.	46.8%	17.4%	9.7%	4.7%	78.6%
	PFIT - Iter2	Perc. Good Depth	44.7%	16.1%	9.6%	5.3%	75.7%
		Perc. Total Covg.	51.4%	14.2%	8.4%	4.7%	78.7%
Kinect07	WTA - Iter1	Perc. Good Depth	30.1%	21.3%	13.8%	6.2%	71.4%
		Perc. Total Covg.	43.5%	17.2%	11.1%	5.0%	76.9%
	WTA - Iter2	Perc. Good Depth	31.0%	22.4%	14.4%	6.1%	74.0%
		Perc. Total Covg.	44.4%	18.1%	11.6%	4.9%	79.0%

Continued on next page

Table C.3 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	Agg. - Iter1	Perc. Good Depth	34.4%	22.9%	14.1%	6.4%	77.8%
		Perc. Total Covg.	47.3%	18.4%	11.3%	5.2%	82.2%
	Agg. - Iter2	Perc. Good Depth	35.2%	23.1%	14.0%	6.1%	78.4%
		Perc. Total Covg.	47.9%	18.6%	11.2%	4.9%	82.6%
	PFIT - Iter1	Perc. Good Depth	40.5%	23.3%	10.2%	6.9%	80.8%
		Perc. Total Covg.	52.1%	18.7%	8.2%	5.5%	84.5%
	PFIT - Iter2	Perc. Good Depth	40.1%	24.1%	9.9%	6.8%	81.0%
		Perc. Total Covg.	51.9%	19.4%	8.0%	5.5%	84.7%
Kinect08	WTA - Iter1	Perc. Good Depth	34.1%	17.8%	14.8%	5.6%	72.3%
		Perc. Total Covg.	44.1%	15.1%	12.5%	4.8%	76.5%
	WTA - Iter2	Perc. Good Depth	35.8%	18.0%	14.4%	5.4%	73.6%
		Perc. Total Covg.	45.6%	15.2%	12.2%	4.5%	77.6%
	Agg. - Iter1	Perc. Good Depth	39.6%	17.0%	14.5%	5.7%	76.8%
		Perc. Total Covg.	49.0%	14.4%	12.2%	4.8%	80.4%
	Agg. - Iter2	Perc. Good Depth	39.9%	17.8%	14.6%	5.9%	78.2%
		Perc. Total Covg.	49.2%	15.0%	12.4%	5.0%	81.6%
	PFIT - Iter1	Perc. Good Depth	43.7%	17.0%	12.0%	5.5%	78.1%
		Perc. Total Covg.	52.6%	14.3%	10.1%	4.6%	81.6%
	PFIT - Iter2	Perc. Good Depth	48.1%	15.5%	12.1%	4.8%	80.5%
		Perc. Total Covg.	56.3%	13.0%	10.2%	4.1%	83.6%
Kinect09	WTA - Iter1	Perc. Good Depth	26.6%	25.8%	11.6%	6.3%	70.3%
		Perc. Total Covg.	35.7%	22.6%	10.1%	5.5%	74.0%
	WTA - Iter2	Perc. Good Depth	28.8%	27.5%	10.0%	6.4%	72.7%
		Perc. Total Covg.	37.6%	24.1%	8.8%	5.6%	76.1%
	Agg. - Iter1	Perc. Good Depth	32.4%	26.3%	13.4%	6.4%	78.5%
		Perc. Total Covg.	40.9%	23.0%	11.7%	5.6%	81.2%
	Agg. - Iter2	Perc. Good Depth	38.2%	25.4%	10.5%	6.3%	80.4%
		Perc. Total Covg.	46.0%	22.2%	9.2%	5.5%	82.9%
	PFIT - Iter1	Perc. Good Depth	44.5%	25.7%	8.1%	5.9%	84.2%
		Perc. Total Covg.	51.6%	22.4%	7.1%	5.1%	86.2%

Continued on next page

Table C.3 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	PFIT - Iter2	Perc. Good Depth	46.8%	23.6%	8.1%	5.5%	84.0%
		Perc. Total Covg.	53.5%	20.6%	7.1%	4.8%	86.0%
Kinect10	WTA - Iter1	Perc. Good Depth	45.9%	7.2%	6.6%	2.6%	62.4%
		Perc. Total Covg.	54.2%	6.1%	5.6%	2.2%	68.1%
	WTA - Iter2	Perc. Good Depth	45.7%	8.7%	7.1%	2.4%	63.9%
		Perc. Total Covg.	54.1%	7.4%	6.0%	2.0%	69.5%
	Agg. - Iter1	Perc. Good Depth	48.5%	8.8%	6.4%	2.6%	66.2%
		Perc. Total Covg.	56.6%	7.4%	5.4%	2.2%	71.6%
	Agg. - Iter2	Perc. Good Depth	47.4%	9.5%	7.2%	2.4%	66.6%
		Perc. Total Covg.	55.8%	8.0%	6.1%	2.1%	71.9%
	PFIT - Iter1	Perc. Good Depth	50.7%	9.9%	6.9%	3.2%	70.7%
		Perc. Total Covg.	58.7%	8.3%	5.8%	2.7%	75.4%
	PFIT - Iter2	Perc. Good Depth	52.6%	9.2%	6.2%	2.5%	70.4%
		Perc. Total Covg.	60.4%	7.7%	5.2%	2.1%	75.3%
Kinect11	WTA - Iter1	Perc. Good Depth	41.9%	9.0%	7.2%	2.6%	60.6%
		Perc. Total Covg.	51.0%	7.6%	6.0%	2.2%	66.8%
	WTA - Iter2	Perc. Good Depth	42.9%	9.7%	7.3%	2.7%	62.6%
		Perc. Total Covg.	51.6%	8.2%	6.2%	2.3%	68.3%
	Agg. - Iter1	Perc. Good Depth	43.9%	10.8%	6.3%	2.6%	63.6%
		Perc. Total Covg.	53.0%	9.0%	5.3%	2.2%	69.5%
	Agg. - Iter2	Perc. Good Depth	44.9%	10.2%	6.8%	2.2%	64.2%
		Perc. Total Covg.	53.9%	8.6%	5.7%	1.9%	70.1%
	PFIT - Iter1	Perc. Good Depth	49.5%	8.6%	5.3%	3.8%	67.1%
		Perc. Total Covg.	57.5%	7.2%	4.4%	3.2%	72.3%
	PFIT - Iter2	Perc. Good Depth	49.9%	9.2%	4.9%	3.9%	67.9%
		Perc. Total Covg.	58.3%	7.7%	4.1%	3.2%	73.3%
Kinect12	WTA - Iter1	Perc. Good Depth	38.0%	11.1%	10.5%	4.1%	63.7%
		Perc. Total Covg.	47.7%	9.3%	8.9%	3.5%	69.4%
	WTA - Iter2	Perc. Good Depth	39.6%	10.6%	10.1%	3.9%	64.2%
		Perc. Total Covg.	49.0%	8.9%	8.6%	3.3%	69.8%

Continued on next page

Table C.3 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	Agg. - Iter1	Perc. Good Depth	40.5%	12.1%	10.4%	3.5%	66.6%
		Perc. Total Covg.	50.0%	10.2%	8.7%	2.9%	71.9%
	Agg. - Iter2	Perc. Good Depth	42.9%	11.7%	9.7%	3.3%	67.6%
		Perc. Total Covg.	51.9%	9.8%	8.2%	2.8%	72.7%
	PFIT - Iter1	Perc. Good Depth	45.9%	10.6%	8.1%	5.2%	69.7%
		Perc. Total Covg.	54.5%	8.9%	6.8%	4.3%	74.5%
	PFIT - Iter2	Perc. Good Depth	46.4%	9.8%	8.3%	4.9%	69.4%
		Perc. Total Covg.	54.7%	8.3%	7.1%	4.1%	74.1%
Kinect13	WTA - Iter1	Perc. Good Depth	44.0%	9.1%	9.7%	3.9%	66.7%
		Perc. Total Covg.	52.1%	7.8%	8.3%	3.3%	71.5%
	WTA - Iter2	Perc. Good Depth	45.4%	10.3%	9.1%	3.8%	68.7%
		Perc. Total Covg.	53.2%	8.8%	7.8%	3.3%	73.2%
	Agg. - Iter1	Perc. Good Depth	48.1%	10.3%	11.4%	2.8%	72.6%
		Perc. Total Covg.	55.6%	8.8%	9.8%	2.4%	76.6%
	Agg. - Iter2	Perc. Good Depth	50.0%	11.0%	11.1%	2.4%	74.4%
		Perc. Total Covg.	57.2%	9.4%	9.5%	2.0%	78.1%
	PFIT - Iter1	Perc. Good Depth	53.2%	11.3%	9.5%	4.8%	78.8%
		Perc. Total Covg.	60.0%	9.7%	8.1%	4.1%	81.9%
	PFIT - Iter2	Perc. Good Depth	54.4%	11.6%	8.1%	4.8%	78.9%
		Perc. Total Covg.	61.1%	9.9%	6.9%	4.1%	82.0%
Kinect14	WTA - Iter1	Perc. Good Depth	22.5%	11.6%	6.8%	9.6%	50.4%
		Perc. Total Covg.	29.4%	10.6%	6.2%	8.7%	54.9%
	WTA - Iter2	Perc. Good Depth	24.1%	10.6%	7.3%	8.9%	50.9%
		Perc. Total Covg.	30.8%	9.6%	6.7%	8.2%	55.2%
	Agg. - Iter1	Perc. Good Depth	28.3%	13.6%	9.8%	9.6%	61.2%
		Perc. Total Covg.	34.8%	12.3%	8.9%	8.7%	64.8%
	Agg. - Iter2	Perc. Good Depth	28.4%	13.6%	9.3%	9.9%	61.2%
		Perc. Total Covg.	35.0%	12.3%	8.5%	9.0%	64.8%
	PFIT - Iter1	Perc. Good Depth	28.9%	19.1%	12.6%	9.7%	70.2%
		Perc. Total Covg.	35.2%	17.4%	11.4%	8.8%	72.8%

Continued on next page

Table C.3 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	PFIT - Iter2	Perc. Good Depth	31.9%	15.4%	9.5%	10.1%	66.9%
		Perc. Total Covg.	38.1%	14.0%	8.7%	9.2%	69.9%
Kinect15	WTA - Iter1	Perc. Good Depth	28.2%	14.0%	7.0%	4.6%	53.9%
		Perc. Total Covg.	35.8%	12.6%	6.3%	4.1%	58.7%
	WTA -Iter2	Perc. Good Depth	30.1%	12.6%	7.2%	4.9%	54.9%
		Perc. Total Covg.	37.5%	11.3%	6.4%	4.4%	59.6%
	Agg. - Iter1	Perc. Good Depth	30.6%	14.5%	6.9%	5.2%	57.3%
		Perc. Total Covg.	38.0%	13.0%	6.2%	4.7%	61.8%
	Agg. - Iter2	Perc. Good Depth	33.3%	13.0%	7.1%	5.2%	58.7%
		Perc. Total Covg.	40.5%	11.6%	6.4%	4.6%	63.2%
	PFIT - Iter1	Perc. Good Depth	31.6%	11.5%	5.9%	9.5%	58.5%
		Perc. Total Covg.	38.8%	10.3%	5.3%	8.5%	62.9%
	PFIT - Iter2	Perc. Good Depth	36.1%	8.7%	6.7%	9.8%	61.3%
		Perc. Total Covg.	43.0%	7.8%	6.0%	8.8%	65.5%
Kinect16	WTA - Iter1	Perc. Good Depth	23.3%	13.7%	8.7%	7.8%	53.4%
		Perc. Total Covg.	28.5%	12.7%	8.1%	7.2%	56.6%
	WTA -Iter2	Perc. Good Depth	25.7%	15.3%	7.9%	7.3%	56.2%
		Perc. Total Covg.	31.1%	14.2%	7.3%	6.8%	59.4%
	Agg. - Iter1	Perc. Good Depth	25.6%	14.8%	11.1%	8.2%	59.8%
		Perc. Total Covg.	30.8%	13.8%	10.4%	7.6%	62.6%
	Agg. - Iter2	Perc. Good Depth	27.4%	16.5%	10.0%	8.6%	62.5%
		Perc. Total Covg.	32.6%	15.4%	9.3%	8.0%	65.2%
	PFIT - Iter1	Perc. Good Depth	30.2%	15.7%	10.5%	5.2%	61.5%
		Perc. Total Covg.	34.7%	14.7%	9.8%	4.9%	64.0%
	PFIT - Iter2	Perc. Good Depth	29.6%	18.0%	10.0%	6.2%	63.9%
		Perc. Total Covg.	34.4%	16.8%	9.4%	5.8%	66.4%
Kinect17	WTA - Iter1	Perc. Good Depth	17.0%	13.5%	10.7%	5.2%	46.4%
		Perc. Total Covg.	28.1%	11.7%	9.3%	4.5%	53.6%
	WTA -Iter2	Perc. Good Depth	18.5%	14.9%	11.2%	4.6%	49.3%
		Perc. Total Covg.	29.6%	12.9%	9.7%	4.0%	56.1%

Continued on next page

Table C.3 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	Agg. - Iter1	Perc. Good Depth	20.4%	17.2%	10.8%	5.1%	53.5%
		Perc. Total Covg.	31.5%	14.8%	9.3%	4.4%	60.0%
	Agg. - Iter2	Perc. Good Depth	21.5%	17.3%	11.0%	4.0%	53.8%
		Perc. Total Covg.	32.4%	14.9%	9.5%	3.4%	60.3%
	PFIT - Iter1	Perc. Good Depth	22.9%	16.4%	10.0%	5.1%	54.3%
		Perc. Total Covg.	33.7%	14.1%	8.6%	4.4%	60.7%
	PFIT - Iter2	Perc. Good Depth	26.2%	19.6%	8.8%	5.0%	59.6%
		Perc. Total Covg.	36.6%	16.8%	7.6%	4.3%	65.3%
Kinect18	WTA - Iter1	Perc. Good Depth	26.9%	18.8%	9.5%	5.2%	60.4%
		Perc. Total Covg.	37.2%	16.1%	8.1%	4.5%	66.0%
	WTA - Iter2	Perc. Good Depth	28.8%	19.3%	9.1%	4.8%	61.9%
		Perc. Total Covg.	38.7%	16.6%	7.8%	4.1%	67.3%
	Agg. - Iter1	Perc. Good Depth	28.2%	21.2%	9.2%	4.1%	62.7%
		Perc. Total Covg.	38.4%	18.2%	7.9%	3.5%	68.0%
	Agg. - Iter2	Perc. Good Depth	32.4%	19.0%	9.2%	3.3%	63.9%
		Perc. Total Covg.	41.8%	16.4%	7.9%	2.8%	68.9%
	PFIT - Iter1	Perc. Good Depth	34.1%	12.1%	13.5%	4.6%	64.4%
		Perc. Total Covg.	43.5%	10.4%	11.6%	4.0%	69.5%
	PFIT - Iter2	Perc. Good Depth	38.0%	13.1%	9.2%	3.9%	64.2%
		Perc. Total Covg.	46.8%	11.2%	7.9%	3.3%	69.3%
Kinect19	WTA - Iter1	Perc. Good Depth	37.1%	11.2%	7.2%	4.2%	59.6%
		Perc. Total Covg.	47.6%	9.3%	6.0%	3.5%	66.4%
	WTA - Iter2	Perc. Good Depth	37.0%	12.3%	7.1%	4.2%	60.7%
		Perc. Total Covg.	47.5%	10.2%	5.9%	3.5%	67.2%
	Agg. - Iter1	Perc. Good Depth	40.1%	11.1%	6.6%	4.4%	62.2%
		Perc. Total Covg.	50.3%	9.2%	5.5%	3.7%	68.7%
	Agg. - Iter2	Perc. Good Depth	40.8%	12.4%	7.2%	3.4%	63.8%
		Perc. Total Covg.	50.8%	10.3%	6.0%	2.8%	70.0%
	PFIT - Iter1	Perc. Good Depth	45.6%	8.8%	5.1%	3.8%	63.2%
		Perc. Total Covg.	54.9%	7.3%	4.2%	3.1%	69.5%

Continued on next page

Table C.3 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	PFIT - Iter2	Perc. Good Depth	45.3%	9.4%	5.0%	4.1%	63.8%
		Perc. Total Covg.	54.7%	7.8%	4.1%	3.4%	70.0%
Kinect20	WTA - Iter1	Perc. Good Depth	26.2%	10.6%	7.5%	7.7%	52.0%
		Perc. Total Covg.	33.6%	9.5%	6.8%	6.9%	56.8%
	WTA - Iter2	Perc. Good Depth	27.3%	10.6%	7.8%	6.6%	52.3%
		Perc. Total Covg.	33.8%	9.7%	7.1%	6.0%	56.5%
	Agg. - Iter1	Perc. Good Depth	28.7%	13.6%	8.3%	8.7%	59.3%
		Perc. Total Covg.	36.0%	12.2%	7.4%	7.8%	63.4%
	Agg. - Iter2	Perc. Good Depth	30.0%	14.4%	8.1%	7.6%	60.1%
		Perc. Total Covg.	37.1%	13.0%	7.3%	6.8%	64.2%
	PFIT - Iter1	Perc. Good Depth	31.7%	15.0%	6.5%	7.9%	61.1%
		Perc. Total Covg.	38.8%	13.5%	5.8%	7.1%	65.1%
	PFIT - Iter2	Perc. Good Depth	32.0%	15.6%	7.5%	7.3%	62.5%
		Perc. Total Covg.	39.0%	14.0%	6.8%	6.6%	66.3%
Kinect21	WTA - Iter1	Perc. Good Depth	35.9%	14.4%	8.4%	5.4%	64.1%
		Perc. Total Covg.	42.5%	12.9%	7.5%	4.8%	67.8%
	WTA - Iter2	Perc. Good Depth	39.3%	17.7%	8.8%	5.3%	71.1%
		Perc. Total Covg.	45.5%	15.9%	7.9%	4.8%	74.1%
	Agg. - Iter1	Perc. Good Depth	38.1%	13.8%	10.3%	6.1%	68.3%
		Perc. Total Covg.	44.5%	12.4%	9.2%	5.5%	71.6%
	Agg. - Iter2	Perc. Good Depth	43.6%	16.8%	9.9%	5.8%	76.0%
		Perc. Total Covg.	49.4%	15.0%	8.9%	5.2%	78.5%
	PFIT - Iter1	Perc. Good Depth	47.0%	15.8%	7.3%	6.5%	76.6%
		Perc. Total Covg.	52.4%	14.2%	6.6%	5.8%	79.0%
	PFIT - Iter2	Perc. Good Depth	47.6%	15.8%	7.3%	5.8%	76.4%
		Perc. Total Covg.	53.0%	14.2%	6.5%	5.2%	78.9%
Kinect22	WTA - Iter1	Perc. Good Depth	27.4%	17.5%	10.3%	7.1%	62.3%
		Perc. Total Covg.	37.3%	15.1%	8.9%	6.1%	67.5%
	WTA - Iter2	Perc. Good Depth	27.6%	17.8%	10.0%	7.6%	63.0%
		Perc. Total Covg.	37.6%	15.3%	8.6%	6.5%	68.1%

Continued on next page

Table C.3 – Continued from previous page

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	Agg. - Iter1	Perc. Good Depth	31.3%	19.1%	10.7%	7.7%	68.7%
		Perc. Total Covg.	41.2%	16.3%	9.1%	6.6%	73.2%
	Agg. - Iter2	Perc. Good Depth	32.0%	18.4%	11.0%	7.8%	69.2%
		Perc. Total Covg.	41.7%	15.8%	9.5%	6.7%	73.7%
	PFIT - Iter1	Perc. Good Depth	33.4%	21.7%	5.2%	7.4%	67.7%
		Perc. Total Covg.	43.2%	18.5%	4.4%	6.3%	72.5%
	PFIT - Iter2	Perc. Good Depth	34.7%	20.9%	4.9%	7.0%	67.5%
		Perc. Total Covg.	44.2%	17.8%	4.2%	6.0%	72.3%
Kinect23	WTA - Iter1	Perc. Good Depth	57.0%	12.6%	6.1%	2.9%	78.6%
		Perc. Total Covg.	62.0%	11.2%	5.4%	2.6%	81.1%
	WTA - Iter2	Perc. Good Depth	59.1%	13.2%	6.1%	3.0%	81.4%
		Perc. Total Covg.	63.9%	11.6%	5.4%	2.6%	83.6%
	Agg. - Iter1	Perc. Good Depth	61.2%	13.8%	5.4%	3.8%	84.1%
		Perc. Total Covg.	65.9%	12.1%	4.7%	3.3%	86.1%
	Agg. - Iter2	Perc. Good Depth	63.1%	14.6%	6.4%	3.9%	88.0%
		Perc. Total Covg.	67.6%	12.8%	5.6%	3.4%	89.5%
	PFIT - Iter1	Perc. Good Depth	68.6%	14.2%	3.7%	1.1%	87.7%
		Perc. Total Covg.	72.6%	12.4%	3.2%	1.0%	89.2%
	PFIT - Iter2	Perc. Good Depth	68.7%	15.1%	3.9%	1.4%	89.1%
		Perc. Total Covg.	72.7%	13.2%	3.4%	1.2%	90.5%
Kinect24	WTA - Iter1	Perc. Good Depth	29.5%	26.1%	8.0%	7.1%	70.7%
		Perc. Total Covg.	40.4%	22.0%	6.8%	6.0%	75.2%
	WTA - Iter2	Perc. Good Depth	29.9%	26.1%	8.3%	7.9%	72.1%
		Perc. Total Covg.	40.7%	22.0%	7.1%	6.7%	76.4%
	Agg. - Iter1	Perc. Good Depth	31.2%	30.4%	7.1%	7.2%	75.8%
		Perc. Total Covg.	42.0%	25.6%	5.9%	6.1%	79.6%
	Agg. - Iter2	Perc. Good Depth	35.3%	25.7%	6.9%	8.1%	76.0%
		Perc. Total Covg.	45.5%	21.7%	5.8%	6.8%	79.8%
	PFIT - Iter1	Perc. Good Depth	39.4%	27.1%	8.2%	4.4%	79.1%
		Perc. Total Covg.	49.2%	22.7%	6.9%	3.7%	82.5%

Continued on next page

Table C.3 – *Continued from previous page*

Image	Method	Metric	10cm	20cm	30cm	40cm	Total
	PFIT - Iter2	Perc. Good Depth	44.2%	21.7%	7.6%	5.7%	79.3%
		Perc. Total Covg.	53.2%	18.2%	6.4%	4.8%	82.6%

APPENDIX D

PARAMETER ANALYSIS OF THE PROPOSED METHOD

This appendix is dedicated to the detailed analysis of the parameters that the proposed method uses (see Section 5.1). Regarding the experiments presented in Section 5.2, the experimentally determined set of values were already provided in Tables A.2 and A.3. In the following subsections, the parameters of each of the proposed method steps are analyzed separately for their effects on the results. In the experiments, only the analyzed parameter’s value is changed and the other parameters are fixed to the values in Tables A.2 and A.3. Below, Table D.1 shows the list of parameters analyzed along with the corresponding steps of the proposed method they are used in (see Table 5.1 for the description of each symbol). Note that the table and the following subsections contains all the parameters except for the ρ parameter (of the Adaptive Cost Aggregation step) and the γ parameter (of Segment Merging and Finalizing step). The former is the truncation value of confidence map and is evaluated as not being a significant parameter. The latter defines how fast the inlier confidence threshold is decreased in Algorithm 4, which does not significantly affect the results.

Table D.1: Parameters of the Proposed Method.

Segmentation	h_s	h_r	M	n	a_{ij}	t_e
Adaptive Windowing	δ_y	λ	ω	$Size(h_w)$	k	
Adaptive Cost Aggregation	λ_{SD}	λ_{DD}	$Size(w(p, q))$			
Iterative Plane Fitting	τ_{ic}	τ_{ir}	τ_{od}	τ_{os}	τ_{oc}	
Segment Merging & Finalizing	τ_α ($^\circ$)	τ_{pd}				

D.1 Segmentation

The effect of initial segmentation on the performance of the proposed method is analyzed in this section. Table D.2 shows the parameters of the 10 different segmentation configurations, which yielded segmentations from over-segmentation to under-segmentation of the left (IR) images. Figure D.1 shows the four sample segmentation maps for the Tsukuba image tested.

Table D.2: Parameters of 10 segmentation levels used.

Segmentation Level (S)	h_s	h_r	M	n	a_{ij}	t_e
1	7	1	50	7	0.5	0.1
2	7	2	50	7	0.5	0.2
3	7	4	50	7	0.5	0.2
4	7	4	50	7	0.5	0.4
5	7	6	50	7	0.5	0.2
6	7	6	50	7	0.5	0.6
7	7	9	50	7	0.5	0.6
8	7	9	50	7	0.5	0.9
9	7	10	50	7	0.5	1.0
10	7	12	50	7	0.5	1.0

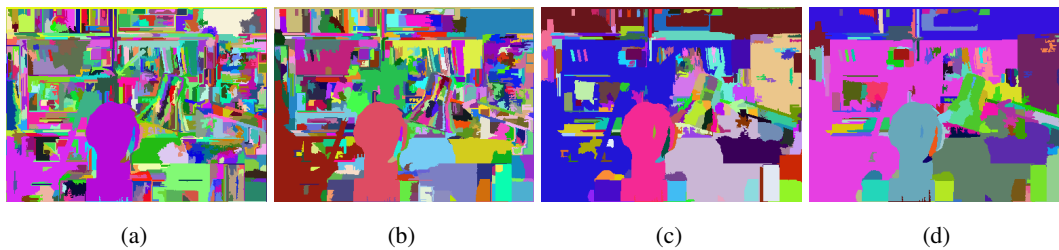
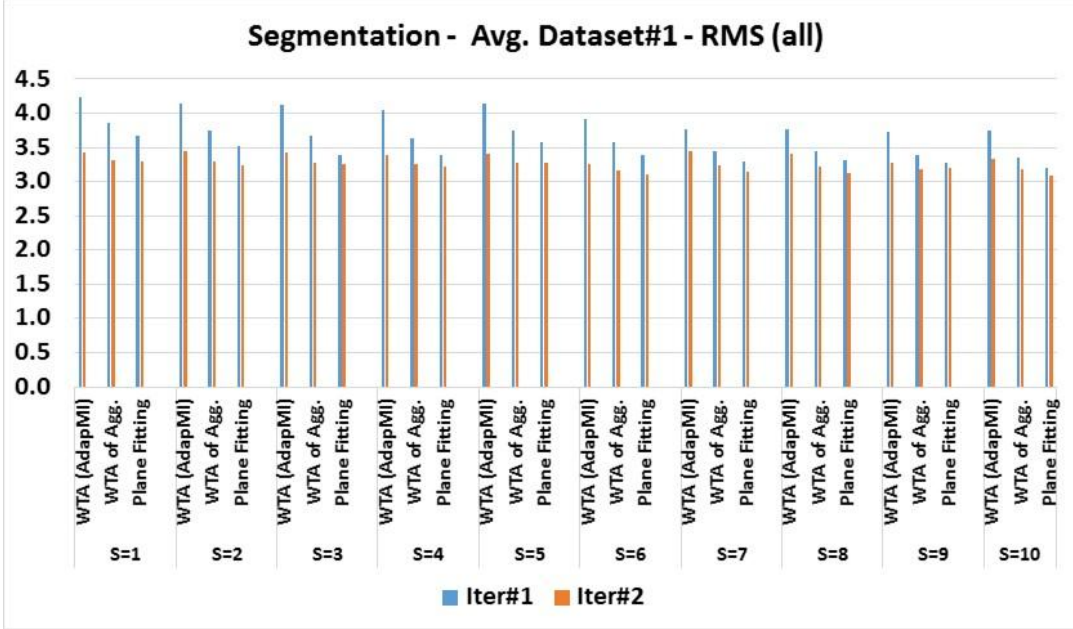


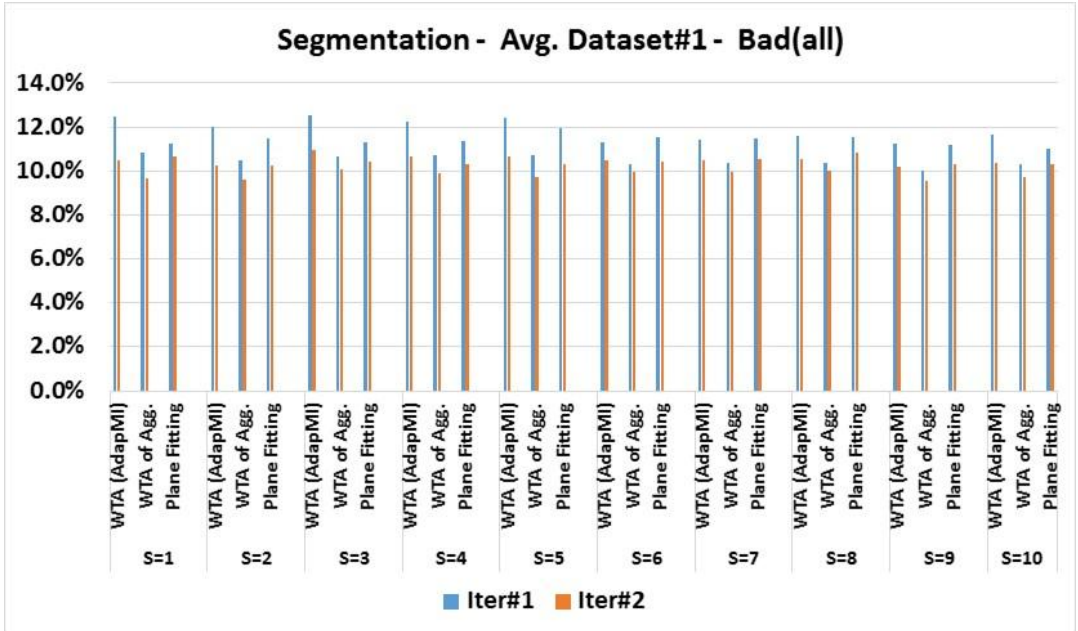
Figure D.1: The four sample segmentation maps for the Tsukuba image from over-segmentation to under-segmentation (a) 1st segmentation level (b) 4th segmentation level (c) 7th segmentation level (d) 10th segmentation level [Best viewed in color]

The average results are provided in Figure D.2, showing that the dependence of the proposed method on segmentation parameters are low. However, when the separate results for each of the image in Dataset #1 are inspected as given in Figure D.3, it is observed that better results are achieved when the level of segmentation is more in accordance with the existing surfaces in the scene. For instance, results on Tsukuba are better with under-segmentation since Tsukuba has many small surfaces; on the other

hand, for Venus, one requires larger segments for better disparity map computation. Another conclusion is that, for most of the adjacent segmentation levels, the second iteration results are more closer than the first iteration results, showing the decreased dependency on the initial segmentation by iterations.



(a)



(b)

Figure D.2: Average Results for Different Levels of Segmentation for Dataset #1.

Graph	Result																																																																																																								
Tsukuba	<p>Segmentation - Tsukuba - RMS (all)</p> <table border="1"> <thead> <tr> <th>Scene Size</th> <th>Method</th> <th>Iter#1</th> <th>Iter#2</th> </tr> </thead> <tbody> <tr><td rowspan="3">S=1</td><td>WTA (AdapMI)</td><td>1.52</td><td>1.48</td></tr> <tr><td>WTA of Agg.</td><td>1.40</td><td>1.43</td></tr> <tr><td>Plane Fitting</td><td>1.43</td><td>1.44</td></tr> <tr><td rowspan="3">S=2</td><td>WTA (AdapMI)</td><td>1.55</td><td>1.46</td></tr> <tr><td>WTA of Agg.</td><td>1.42</td><td>1.38</td></tr> <tr><td>Plane Fitting</td><td>1.44</td><td>1.36</td></tr> <tr><td rowspan="3">S=3</td><td>WTA (AdapMI)</td><td>1.52</td><td>1.45</td></tr> <tr><td>WTA of Agg.</td><td>1.38</td><td>1.39</td></tr> <tr><td>Plane Fitting</td><td>1.38</td><td>1.33</td></tr> <tr><td rowspan="3">S=4</td><td>WTA (AdapMI)</td><td>1.48</td><td>1.44</td></tr> <tr><td>WTA of Agg.</td><td>1.44</td><td>1.34</td></tr> <tr><td>Plane Fitting</td><td>1.40</td><td>1.40</td></tr> <tr><td rowspan="3">S=5</td><td>WTA (AdapMI)</td><td>1.48</td><td>1.44</td></tr> <tr><td>WTA of Agg.</td><td>1.44</td><td>1.38</td></tr> <tr><td>Plane Fitting</td><td>1.44</td><td>1.38</td></tr> <tr><td rowspan="3">S=6</td><td>WTA (AdapMI)</td><td>1.44</td><td>1.41</td></tr> <tr><td>WTA of Agg.</td><td>1.42</td><td>1.41</td></tr> <tr><td>Plane Fitting</td><td>1.44</td><td>1.44</td></tr> <tr><td rowspan="3">S=7</td><td>WTA (AdapMI)</td><td>1.44</td><td>1.41</td></tr> <tr><td>WTA of Agg.</td><td>1.40</td><td>1.40</td></tr> <tr><td>Plane Fitting</td><td>1.42</td><td>1.42</td></tr> <tr><td rowspan="3">S=8</td><td>WTA (AdapMI)</td><td>1.50</td><td>1.44</td></tr> <tr><td>WTA of Agg.</td><td>1.42</td><td>1.41</td></tr> <tr><td>Plane Fitting</td><td>1.44</td><td>1.44</td></tr> <tr><td rowspan="3">S=9</td><td>WTA (AdapMI)</td><td>1.50</td><td>1.44</td></tr> <tr><td>WTA of Agg.</td><td>1.38</td><td>1.38</td></tr> <tr><td>Plane Fitting</td><td>1.44</td><td>1.44</td></tr> <tr><td rowspan="3">S=10</td><td>WTA (AdapMI)</td><td>1.52</td><td>1.44</td></tr> <tr><td>WTA of Agg.</td><td>1.42</td><td>1.42</td></tr> <tr><td>Plane Fitting</td><td>1.44</td><td>1.44</td></tr> </tbody> </table>	Scene Size	Method	Iter#1	Iter#2	S=1	WTA (AdapMI)	1.52	1.48	WTA of Agg.	1.40	1.43	Plane Fitting	1.43	1.44	S=2	WTA (AdapMI)	1.55	1.46	WTA of Agg.	1.42	1.38	Plane Fitting	1.44	1.36	S=3	WTA (AdapMI)	1.52	1.45	WTA of Agg.	1.38	1.39	Plane Fitting	1.38	1.33	S=4	WTA (AdapMI)	1.48	1.44	WTA of Agg.	1.44	1.34	Plane Fitting	1.40	1.40	S=5	WTA (AdapMI)	1.48	1.44	WTA of Agg.	1.44	1.38	Plane Fitting	1.44	1.38	S=6	WTA (AdapMI)	1.44	1.41	WTA of Agg.	1.42	1.41	Plane Fitting	1.44	1.44	S=7	WTA (AdapMI)	1.44	1.41	WTA of Agg.	1.40	1.40	Plane Fitting	1.42	1.42	S=8	WTA (AdapMI)	1.50	1.44	WTA of Agg.	1.42	1.41	Plane Fitting	1.44	1.44	S=9	WTA (AdapMI)	1.50	1.44	WTA of Agg.	1.38	1.38	Plane Fitting	1.44	1.44	S=10	WTA (AdapMI)	1.52	1.44	WTA of Agg.	1.42	1.42	Plane Fitting	1.44	1.44
Scene Size	Method	Iter#1	Iter#2																																																																																																						
S=1	WTA (AdapMI)	1.52	1.48																																																																																																						
	WTA of Agg.	1.40	1.43																																																																																																						
	Plane Fitting	1.43	1.44																																																																																																						
S=2	WTA (AdapMI)	1.55	1.46																																																																																																						
	WTA of Agg.	1.42	1.38																																																																																																						
	Plane Fitting	1.44	1.36																																																																																																						
S=3	WTA (AdapMI)	1.52	1.45																																																																																																						
	WTA of Agg.	1.38	1.39																																																																																																						
	Plane Fitting	1.38	1.33																																																																																																						
S=4	WTA (AdapMI)	1.48	1.44																																																																																																						
	WTA of Agg.	1.44	1.34																																																																																																						
	Plane Fitting	1.40	1.40																																																																																																						
S=5	WTA (AdapMI)	1.48	1.44																																																																																																						
	WTA of Agg.	1.44	1.38																																																																																																						
	Plane Fitting	1.44	1.38																																																																																																						
S=6	WTA (AdapMI)	1.44	1.41																																																																																																						
	WTA of Agg.	1.42	1.41																																																																																																						
	Plane Fitting	1.44	1.44																																																																																																						
S=7	WTA (AdapMI)	1.44	1.41																																																																																																						
	WTA of Agg.	1.40	1.40																																																																																																						
	Plane Fitting	1.42	1.42																																																																																																						
S=8	WTA (AdapMI)	1.50	1.44																																																																																																						
	WTA of Agg.	1.42	1.41																																																																																																						
	Plane Fitting	1.44	1.44																																																																																																						
S=9	WTA (AdapMI)	1.50	1.44																																																																																																						
	WTA of Agg.	1.38	1.38																																																																																																						
	Plane Fitting	1.44	1.44																																																																																																						
S=10	WTA (AdapMI)	1.52	1.44																																																																																																						
	WTA of Agg.	1.42	1.42																																																																																																						
	Plane Fitting	1.44	1.44																																																																																																						
RMS(all)																																																																																																									
Tsukuba	<p>Segmentation - Tsukuba- Bad(all)</p> <table border="1"> <thead> <tr> <th>Scene Size</th> <th>Method</th> <th>Iter#1</th> <th>Iter#2</th> </tr> </thead> <tbody> <tr><td rowspan="3">S=1</td><td>WTA (AdapMI)</td><td>7.0%</td><td>6.0%</td></tr> <tr><td>WTA of Agg.</td><td>6.0%</td><td>5.5%</td></tr> <tr><td>Plane Fitting</td><td>6.5%</td><td>6.8%</td></tr> <tr><td rowspan="3">S=2</td><td>WTA (AdapMI)</td><td>6.5%</td><td>5.5%</td></tr> <tr><td>WTA of Agg.</td><td>5.5%</td><td>5.0%</td></tr> <tr><td>Plane Fitting</td><td>5.5%</td><td>4.5%</td></tr> <tr><td rowspan="3">S=3</td><td>WTA (AdapMI)</td><td>6.5%</td><td>5.5%</td></tr> <tr><td>WTA of Agg.</td><td>5.5%</td><td>5.0%</td></tr> <tr><td>Plane Fitting</td><td>5.5%</td><td>5.5%</td></tr> <tr><td rowspan="3">S=4</td><td>WTA (AdapMI)</td><td>6.0%</td><td>5.5%</td></tr> <tr><td>WTA of Agg.</td><td>6.0%</td><td>5.5%</td></tr> <tr><td>Plane Fitting</td><td>6.0%</td><td>5.5%</td></tr> <tr><td rowspan="3">S=5</td><td>WTA (AdapMI)</td><td>6.0%</td><td>5.5%</td></tr> <tr><td>WTA of Agg.</td><td>5.5%</td><td>5.0%</td></tr> <tr><td>Plane Fitting</td><td>5.5%</td><td>5.0%</td></tr> <tr><td rowspan="3">S=6</td><td>WTA (AdapMI)</td><td>6.5%</td><td>6.0%</td></tr> <tr><td>WTA of Agg.</td><td>6.5%</td><td>6.0%</td></tr> <tr><td>Plane Fitting</td><td>6.5%</td><td>6.0%</td></tr> <tr><td rowspan="3">S=7</td><td>WTA (AdapMI)</td><td>6.5%</td><td>6.0%</td></tr> <tr><td>WTA of Agg.</td><td>6.5%</td><td>6.0%</td></tr> <tr><td>Plane Fitting</td><td>6.5%</td><td>6.0%</td></tr> <tr><td rowspan="3">S=8</td><td>WTA (AdapMI)</td><td>7.0%</td><td>6.5%</td></tr> <tr><td>WTA of Agg.</td><td>7.0%</td><td>6.5%</td></tr> <tr><td>Plane Fitting</td><td>7.0%</td><td>6.5%</td></tr> <tr><td rowspan="3">S=9</td><td>WTA (AdapMI)</td><td>6.5%</td><td>6.0%</td></tr> <tr><td>WTA of Agg.</td><td>6.5%</td><td>6.0%</td></tr> <tr><td>Plane Fitting</td><td>6.5%</td><td>6.0%</td></tr> <tr><td rowspan="3">S=10</td><td>WTA (AdapMI)</td><td>6.5%</td><td>6.0%</td></tr> <tr><td>WTA of Agg.</td><td>6.5%</td><td>6.0%</td></tr> <tr><td>Plane Fitting</td><td>6.5%</td><td>6.0%</td></tr> </tbody> </table>	Scene Size	Method	Iter#1	Iter#2	S=1	WTA (AdapMI)	7.0%	6.0%	WTA of Agg.	6.0%	5.5%	Plane Fitting	6.5%	6.8%	S=2	WTA (AdapMI)	6.5%	5.5%	WTA of Agg.	5.5%	5.0%	Plane Fitting	5.5%	4.5%	S=3	WTA (AdapMI)	6.5%	5.5%	WTA of Agg.	5.5%	5.0%	Plane Fitting	5.5%	5.5%	S=4	WTA (AdapMI)	6.0%	5.5%	WTA of Agg.	6.0%	5.5%	Plane Fitting	6.0%	5.5%	S=5	WTA (AdapMI)	6.0%	5.5%	WTA of Agg.	5.5%	5.0%	Plane Fitting	5.5%	5.0%	S=6	WTA (AdapMI)	6.5%	6.0%	WTA of Agg.	6.5%	6.0%	Plane Fitting	6.5%	6.0%	S=7	WTA (AdapMI)	6.5%	6.0%	WTA of Agg.	6.5%	6.0%	Plane Fitting	6.5%	6.0%	S=8	WTA (AdapMI)	7.0%	6.5%	WTA of Agg.	7.0%	6.5%	Plane Fitting	7.0%	6.5%	S=9	WTA (AdapMI)	6.5%	6.0%	WTA of Agg.	6.5%	6.0%	Plane Fitting	6.5%	6.0%	S=10	WTA (AdapMI)	6.5%	6.0%	WTA of Agg.	6.5%	6.0%	Plane Fitting	6.5%	6.0%
Scene Size	Method	Iter#1	Iter#2																																																																																																						
S=1	WTA (AdapMI)	7.0%	6.0%																																																																																																						
	WTA of Agg.	6.0%	5.5%																																																																																																						
	Plane Fitting	6.5%	6.8%																																																																																																						
S=2	WTA (AdapMI)	6.5%	5.5%																																																																																																						
	WTA of Agg.	5.5%	5.0%																																																																																																						
	Plane Fitting	5.5%	4.5%																																																																																																						
S=3	WTA (AdapMI)	6.5%	5.5%																																																																																																						
	WTA of Agg.	5.5%	5.0%																																																																																																						
	Plane Fitting	5.5%	5.5%																																																																																																						
S=4	WTA (AdapMI)	6.0%	5.5%																																																																																																						
	WTA of Agg.	6.0%	5.5%																																																																																																						
	Plane Fitting	6.0%	5.5%																																																																																																						
S=5	WTA (AdapMI)	6.0%	5.5%																																																																																																						
	WTA of Agg.	5.5%	5.0%																																																																																																						
	Plane Fitting	5.5%	5.0%																																																																																																						
S=6	WTA (AdapMI)	6.5%	6.0%																																																																																																						
	WTA of Agg.	6.5%	6.0%																																																																																																						
	Plane Fitting	6.5%	6.0%																																																																																																						
S=7	WTA (AdapMI)	6.5%	6.0%																																																																																																						
	WTA of Agg.	6.5%	6.0%																																																																																																						
	Plane Fitting	6.5%	6.0%																																																																																																						
S=8	WTA (AdapMI)	7.0%	6.5%																																																																																																						
	WTA of Agg.	7.0%	6.5%																																																																																																						
	Plane Fitting	7.0%	6.5%																																																																																																						
S=9	WTA (AdapMI)	6.5%	6.0%																																																																																																						
	WTA of Agg.	6.5%	6.0%																																																																																																						
	Plane Fitting	6.5%	6.0%																																																																																																						
S=10	WTA (AdapMI)	6.5%	6.0%																																																																																																						
	WTA of Agg.	6.5%	6.0%																																																																																																						
	Plane Fitting	6.5%	6.0%																																																																																																						
Bad(all)																																																																																																									
Venus																																																																																																									

Figure D.3: continued

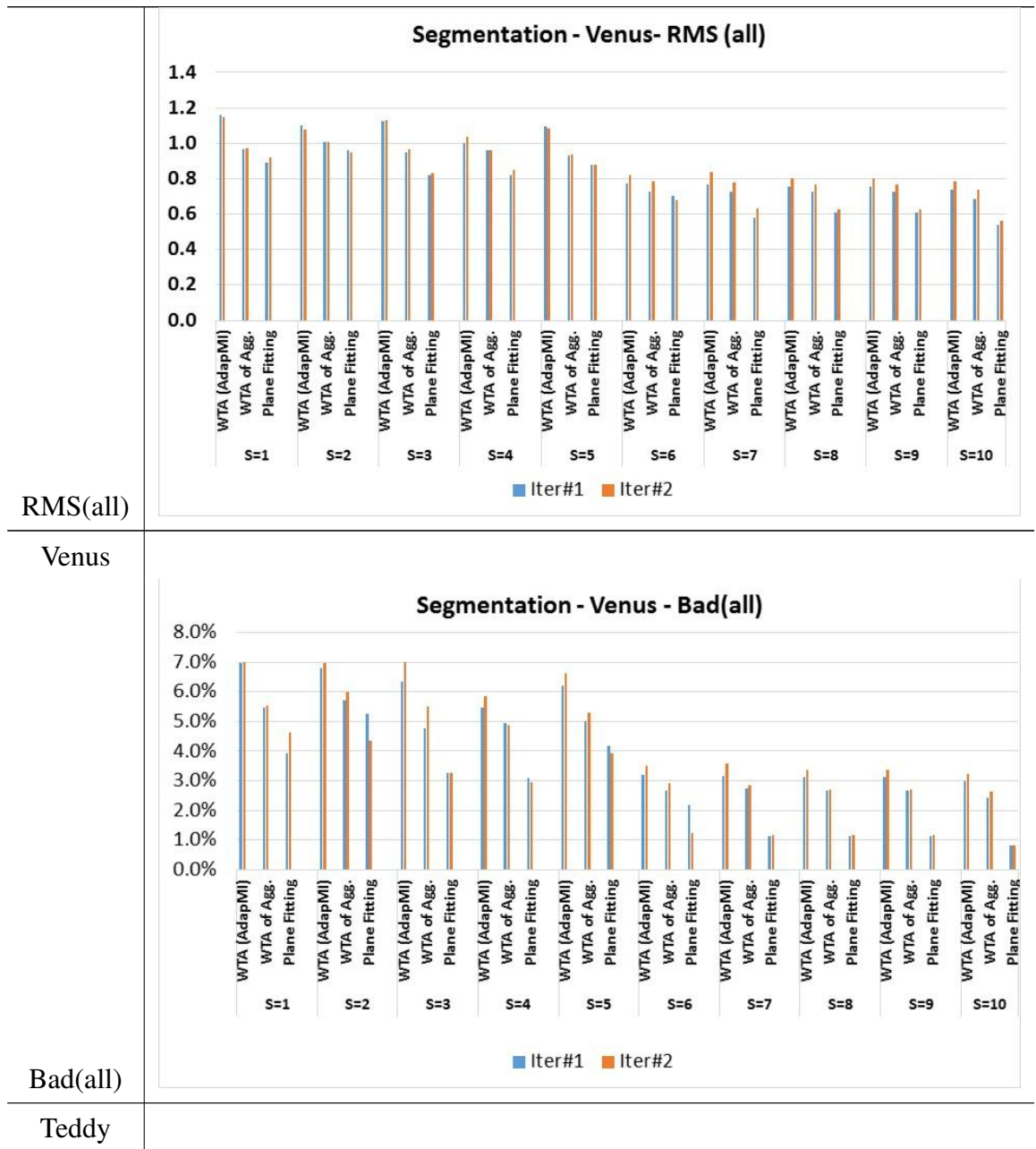


Figure D.3: continued

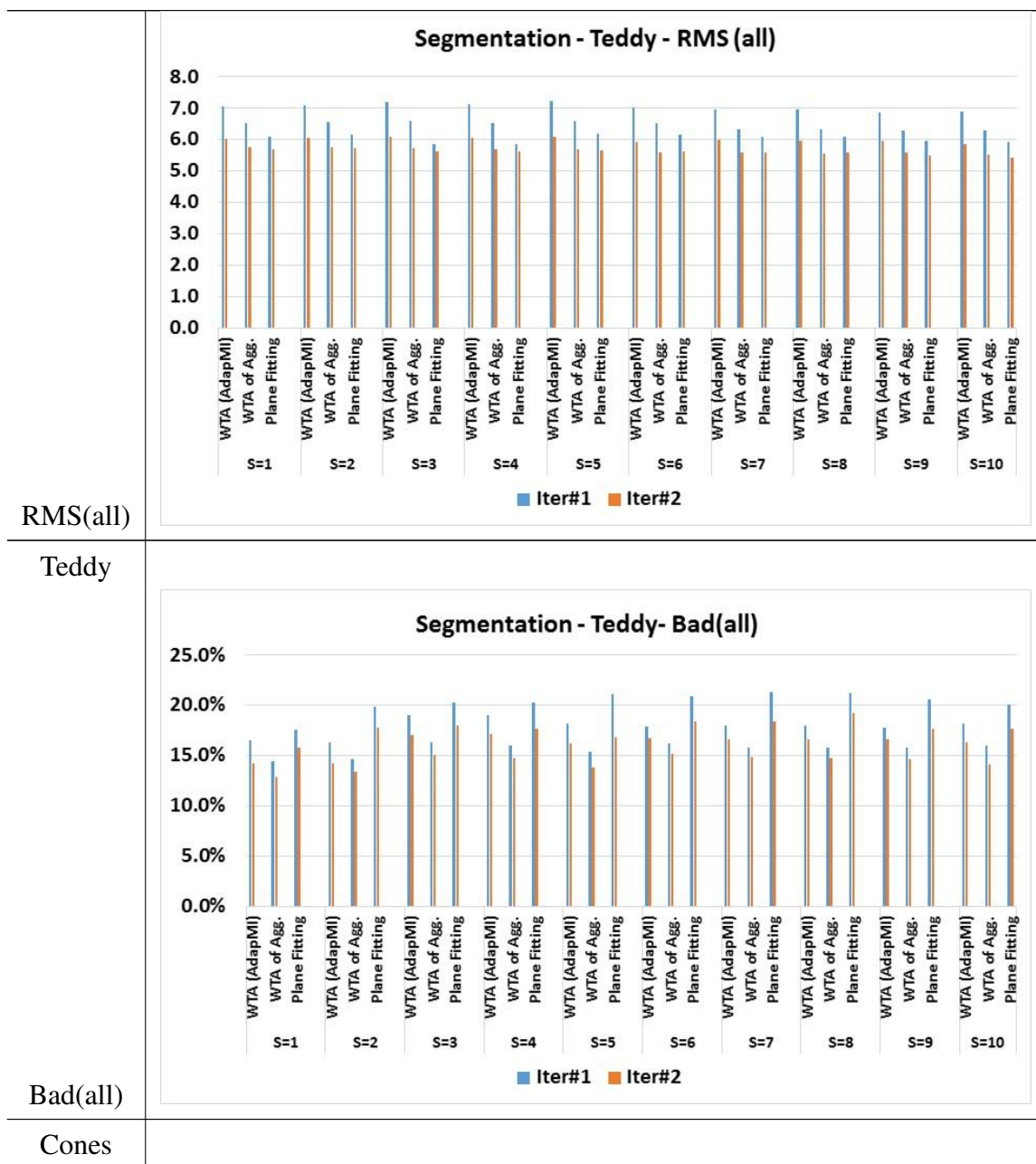


Figure D.3: continued

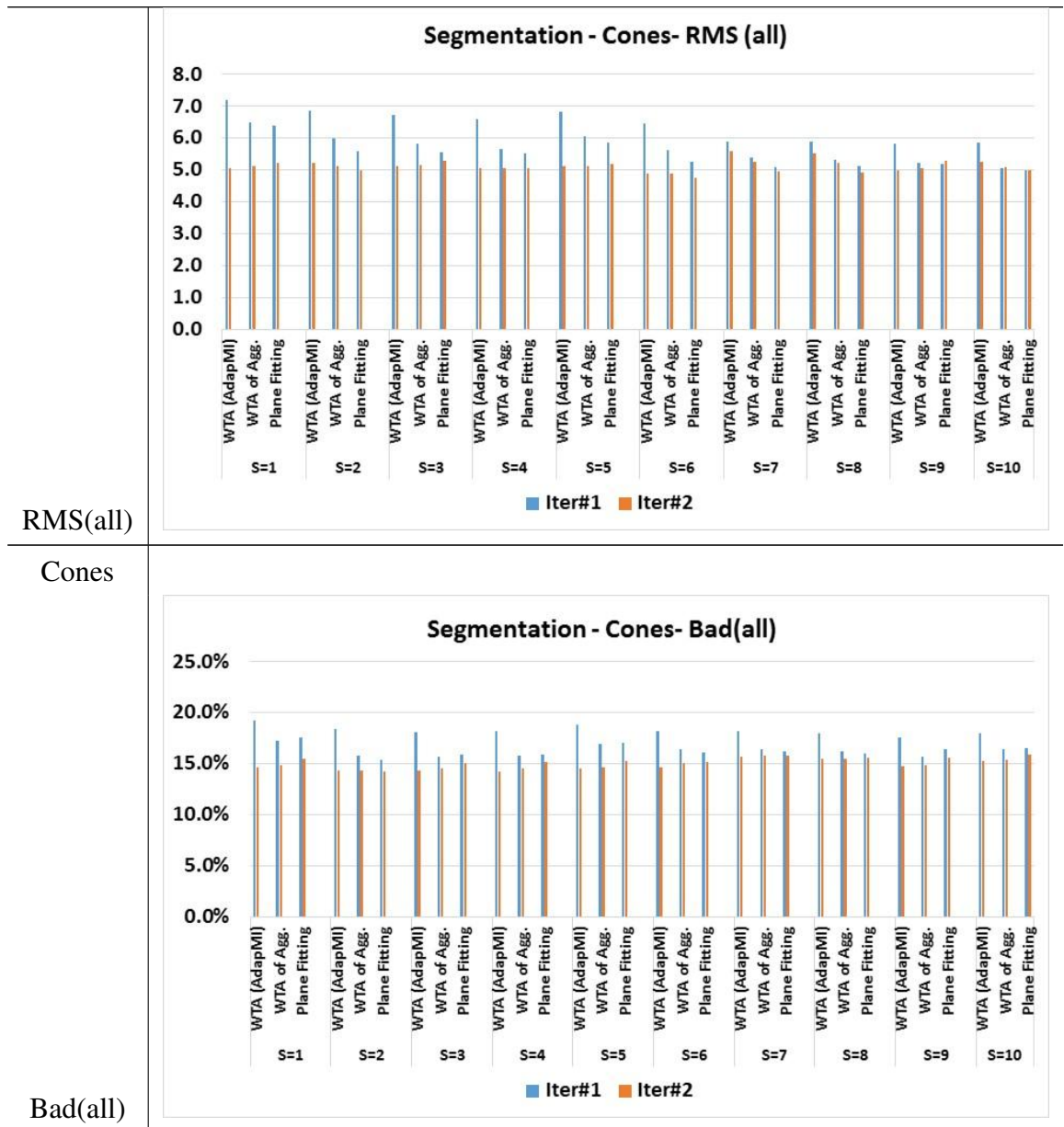


Figure D.3: The results of for the Dataset #1 image pairs, for the 10 different segmentation levels from over-segmentation to under-segmentation

D.2 Parameters of Adaptive Windowing Step

In this part, the parameters of the Adaptive Windowing step of the proposed method are analyzed.

D.2.1 The Vertical Window Size - δ_y

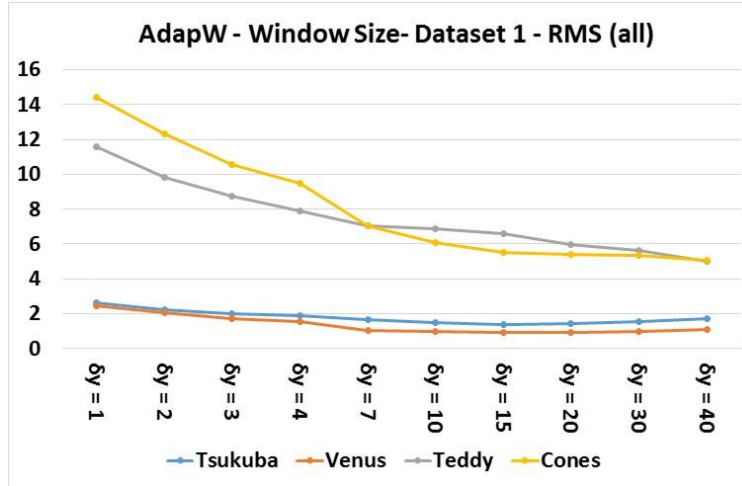
Regarding the window size used in the vertical direction of the adaptively-computed windows, values from [1-40] are used, and the results are provided in Figure D.4. As can be observed from the results, the vertical window size affects the results to some extent, which should be adjusted according to the resolution of the images and the size of the segments in the images. However, after some point, e.g., $\delta_y = 10$ yielding a 21-pixel window size in the vertical direction, the results converge and even get worse along with more running time.

D.2.2 The Ratio of Incorporating Prior Probabilities - λ

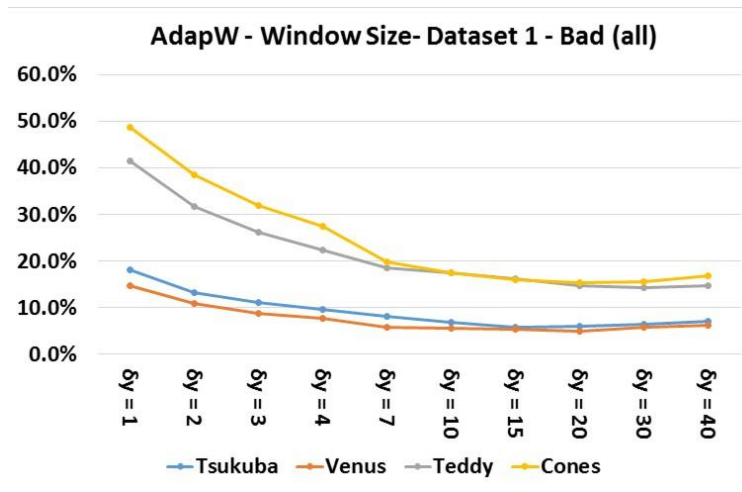
This parameter adjusts how much the joint prior probabilities will be incorporated to the MI calculation - see Eqn. 5.9. The results in Figure D.5 show that a reasonable incorporation ratio is around $\lambda = 0.5$ where, towards the edges, all results get worse. This also proves incorporation of prior probabilities improves the results.

D.2.3 The Thickness of Discontinuities - ω

The ω parameter defines the assumed thickness of discontinuities when the size of the adaptive window is determined in the horizontal direction (see Eqn. 5.2) and Figure 5.3). The results in Figure D.6 show that a small ω value, e.g., values smaller than 3 pixels, is enough. Increasing ω does not drastically affect the results since pixels away from the segment borders are weighted by inverse exponential distance to border (see Eqn. 5.13).



(a)

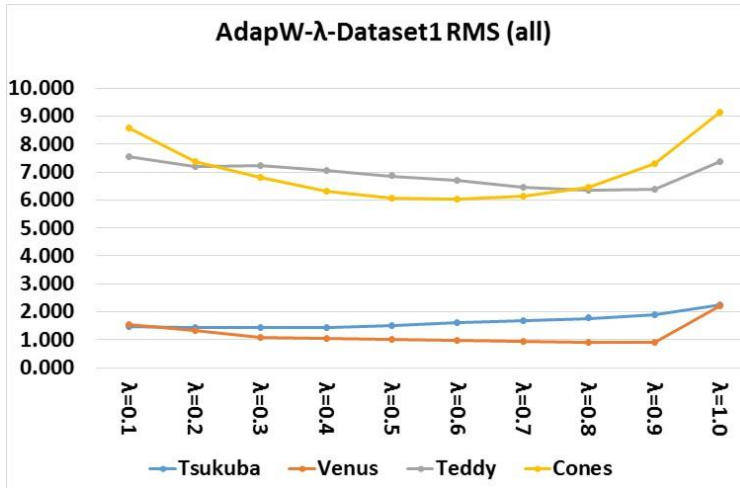


(b)

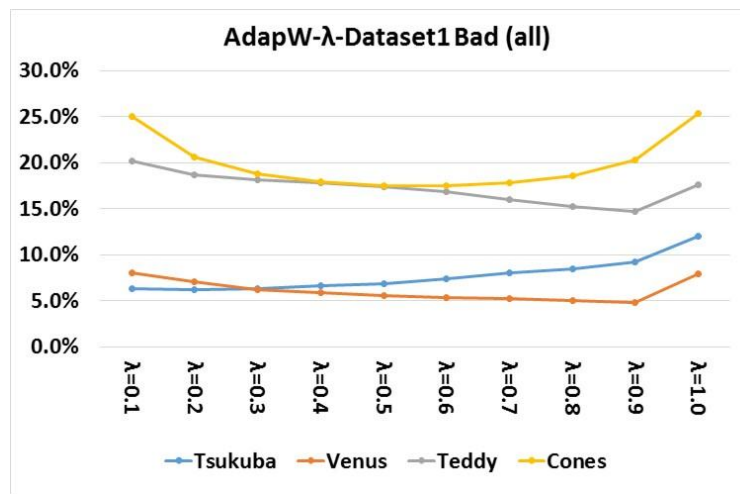
Figure D.4: Results for Increasing δ_y for Dataset #1.

D.2.4 The Histogram Size - $Size(h_w)$

The size of the intensity bins used when constructing the histogram is an important parameter when computing the joint probabilities. Figure D.7 shows results obtained by changing the bin sizes in the interval $[10 - 255]$ (10 bins correspond to a bin size of 25 intensity levels whereas 255 bins correspond to a separate bin for each intensity level). The results show that around 40 bins is adequate for matching two local windows in accordance with the Fookes' study in [34].



(a)



(b)

Figure D.5: Results of Increasing λ for Dataset #1.

D.2.5 The Incrementation Constant for Histogram Computation - k

The k parameter used in $T()$ function in Eqn. 5.13 enables incrementing corresponding histogram bin values. The incrementation is not performed by a constant value but rather adaptively decreased by the distance to the segment border if neighbor pixel is around the segment border. However, as the results given in Figure D.8 shows, the selection of any k value does not change the ordering of candidate disparities but only the scale of computed values changes. In the experiments, $k = 5$ value was used only for easy debugging of the method when computing the histogram counts.

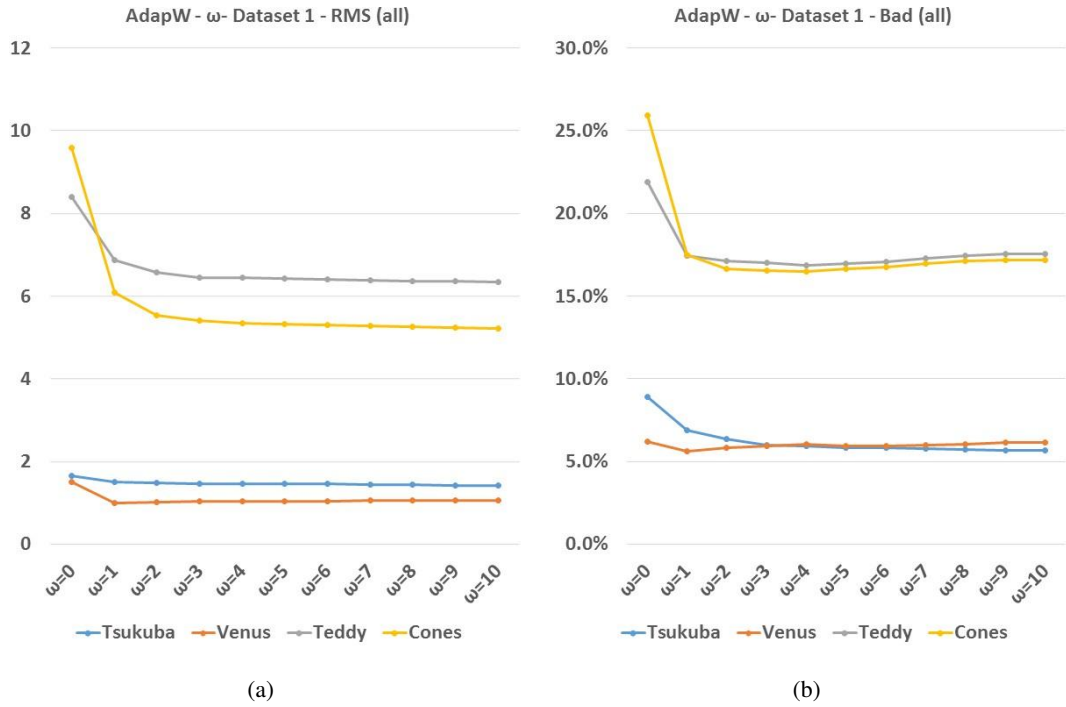


Figure D.6: Effect of increasing ω on Dataset #1.

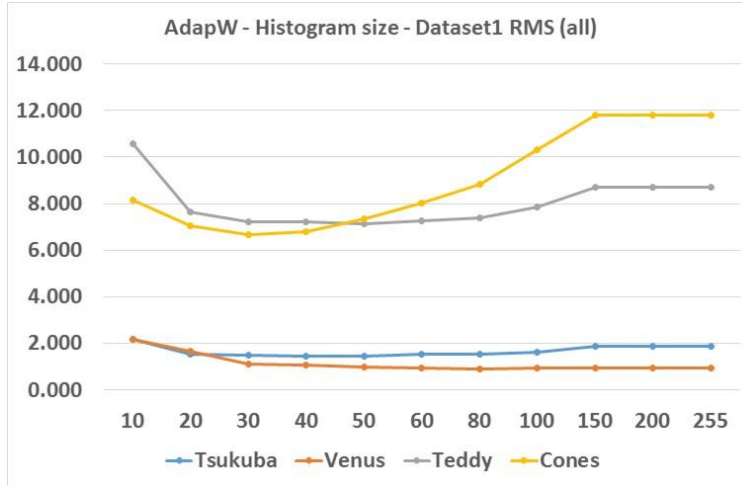
D.3 Parameters of Adaptive Cost Aggregation Step

In this section, the parameters of the Adaptive Cost Aggregation step of the proposed method is analyzed.

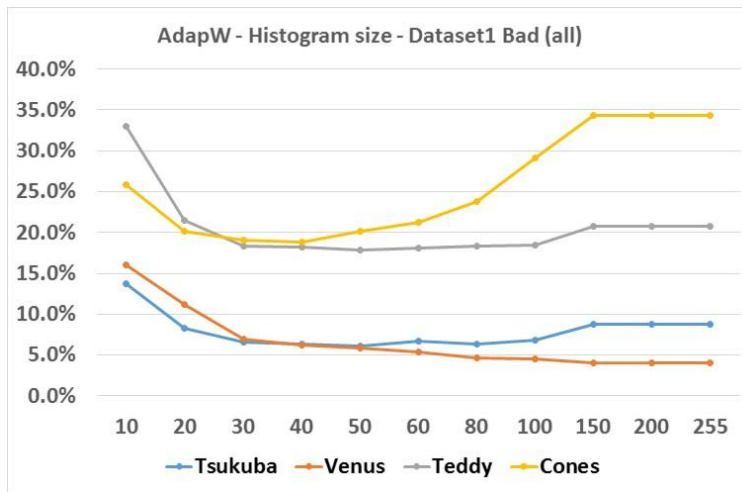
D.3.1 The Size of the Aggregation Window - $Size(w(p, q))$

This parameter defines the size of the local window of cost aggregation step explained in Section 5.1.3 (see Eqn. 5.15). The results are provided in Figure D.9 for the window sizes ranging from 3x3 to 81x81.

As can be observed from the results, the window size should be adjusted according to the resolution of the image pair and the objects in the scene where, after some point, the blurring of the output disparity maps occurs and the performance of the method degrades.



(a)

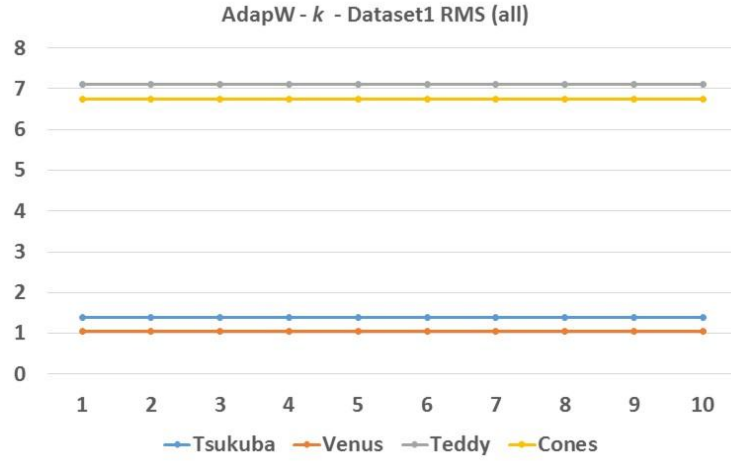


(b)

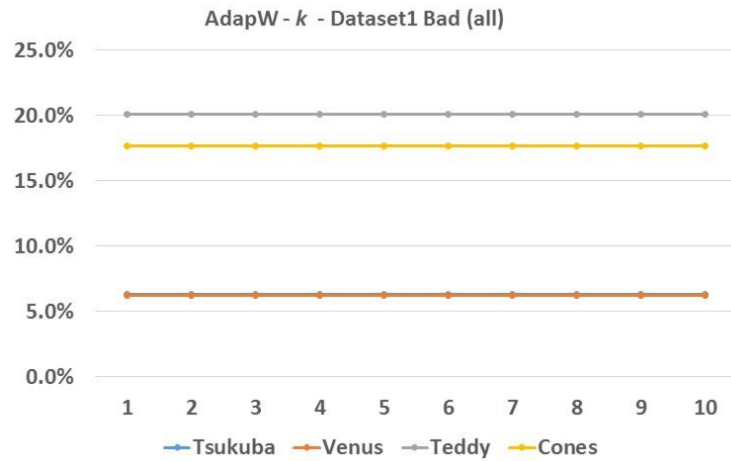
Figure D.7: Effect of increasing histogram size on Dataset #1.

D.3.2 The Scaling Parameters of the Aggregation Weights - λ_{SD} and λ_{DD}

The λ_{SD} and λ_{DD} parameters in Eqn. 5.16 were used as scaling constants in the equation for the pixels outside the current segment within the neighborhood of aggregation. λ_{SD} parameter scales the spatial distance and λ_{DD} scales the disparity distance. In the experiments, each of the parameters are set to values from 1 to 20 where the other is fixed at 1 as default. Increasing λ_{SD} means decreasing the importance of spatial distance and increasing the importance of disparity distance for the cost aggregation weights computation regarding the pixels outside the current segment, and vice versa.



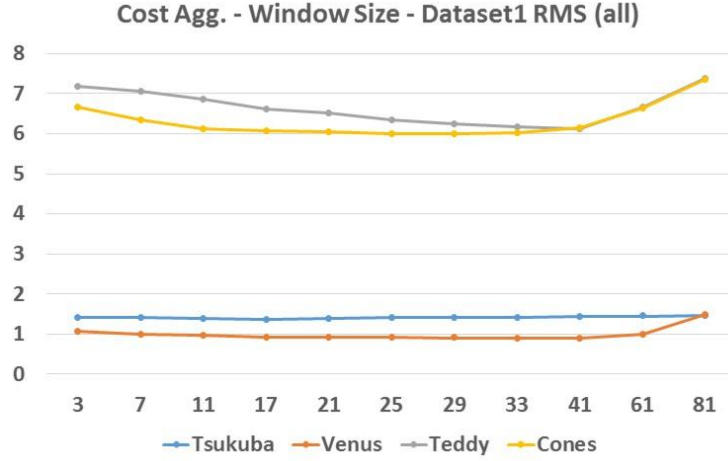
(a)



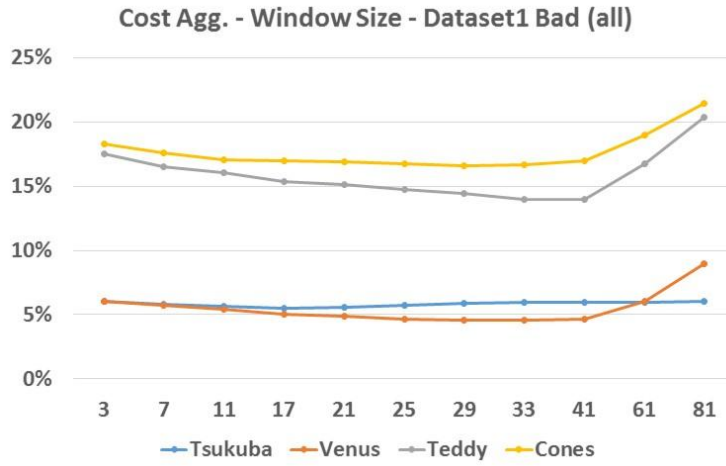
(b)

Figure D.8: Results of different k parameter values for Dataset #1.

When the results given in Figure D.10 are analyzed together, it is concluded that, for some image pairs (like Tsukuba and Cones), increasing the effect of either parameter can affect the results and for some others, no significant effect occurs. Tsukuba has smaller segments and softer disparity decreases in neighboring segments, which yields worse results when spatial distance is more incorporated into the weight calculation. On the other hand, Cones has larger segments and steeper disparity changes between neighboring regions that yields better results when disparity distance is more incorporated. Therefore, the user can make use of these parameters according to the image scene worked on.



(a)



(b)

Figure D.9: Effect of increasing aggregation window size on Dataset #1 (WTA cost aggregation results).

D.4 Parameters of Iterative Plane Fitting Step

In this section, the parameters of the Iterative Plane Fitting step of the proposed method is analyzed.

D.4.1 Confidence Threshold for Inlier Disparities - τ_{ic}

This parameter is used in Algorithm 3 while selecting confident pixels for the segment to be plane-fitted using the inlier disparities only. Figure D.11 shows the average

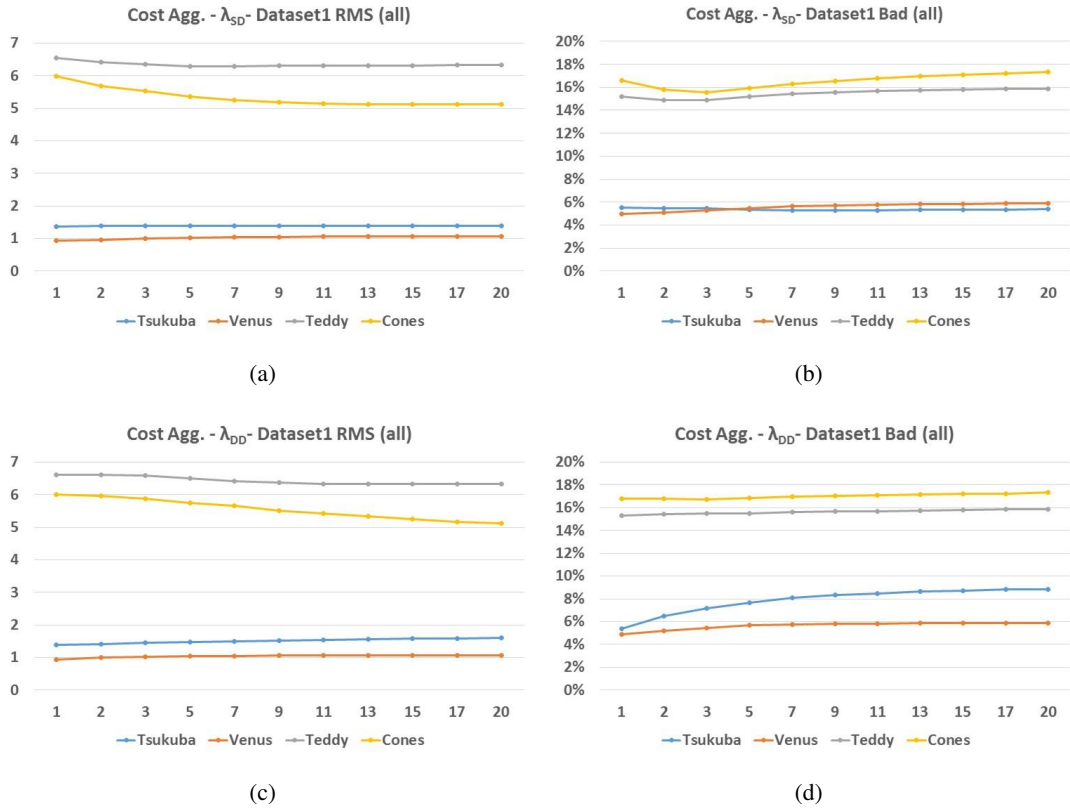
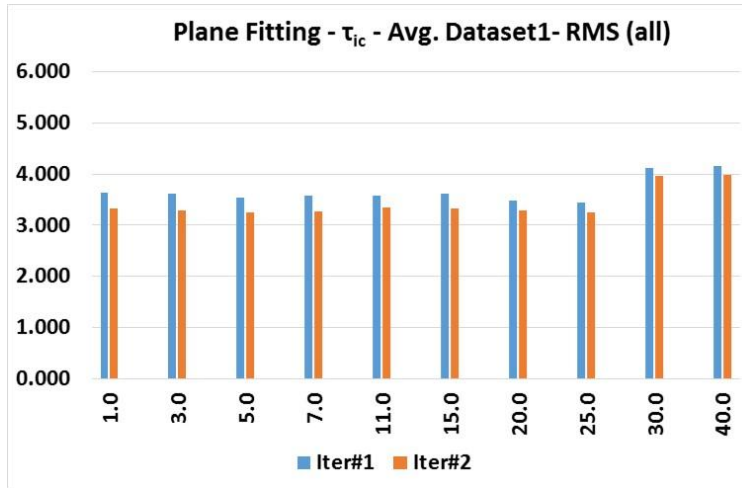


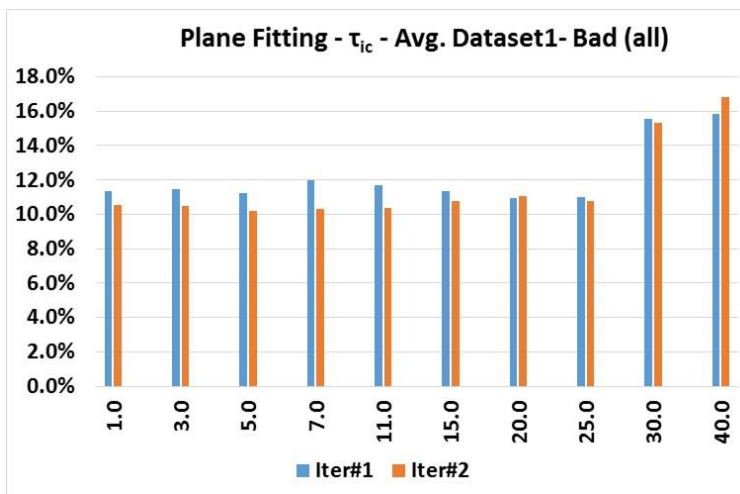
Figure D.10: WTA Results of Cost Aggregation Step Regarding Increasing Scaling Parameter Values for the Aggregation Weights Computation for Dataset #1.

results obtained for two iterations using Dataset #1 for the increasing τ_{ic} parameters given in units of 10s of the percentage (see Eqn. 5.14). Figure D.12 provides the separate results for each of the image in Dataset #1.

When the average results are inspected, as the threshold is increased, more confident but less number of disparities are extracted, which, in turn, affects the fitted plane's accuracy. When analyzed along with the separate results for each image, although the threshold can be determined separately for each image for better optimization, a value smaller than 0.10% as a confidence metric can be used for all the images in Dataset #1 experimentally.



(a)



(b)

Figure D.11: Average Results for different τ_{ic} as confidence threshold for determining inlier disparities to perform plane fitting.

D.4.2 Stable Segment Ratio Threshold - τ_{ir}

This parameter is used in Algorithm 3 when selecting the pixels for the segment which is to be plane fitted using the inlier disparities only. The selected disparities are marked as stable if the ratio of the number of confident pixels over the segment size is greater than this threshold and the plane fitting is performed over the stable segment.

Figure D.13 shows the average results obtained for two iterations using Dataset #1 for the increasing τ_{ir} parameters given in percentage units. Figure D.14 provides the separate results for each of the image in Dataset #1. When the average results are

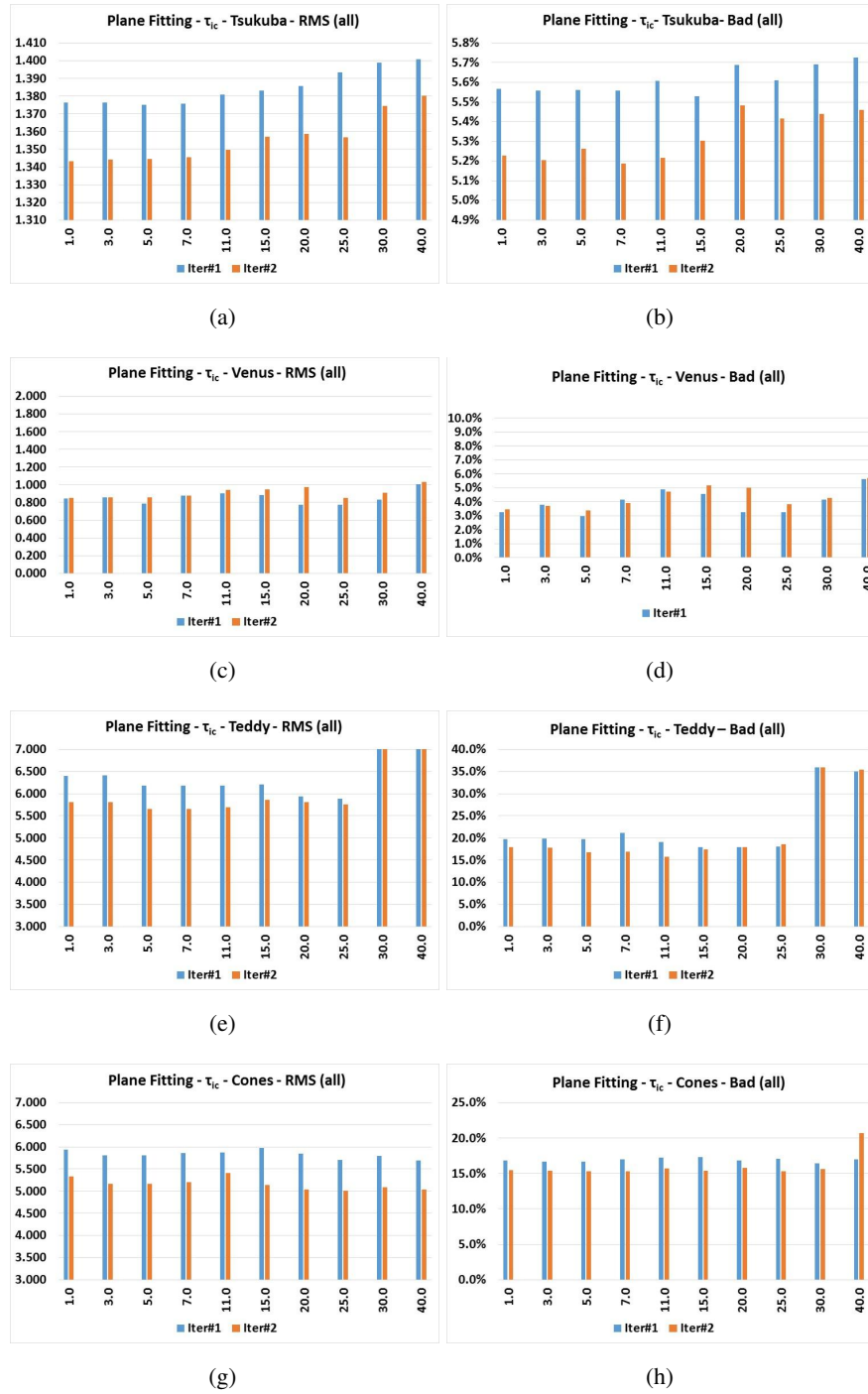
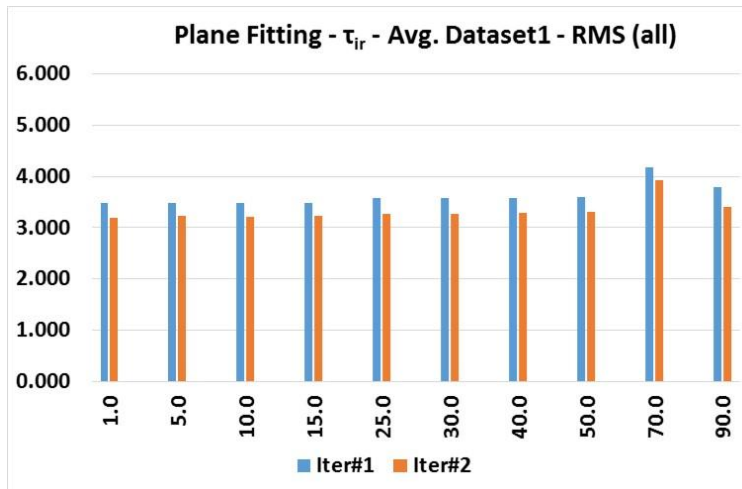


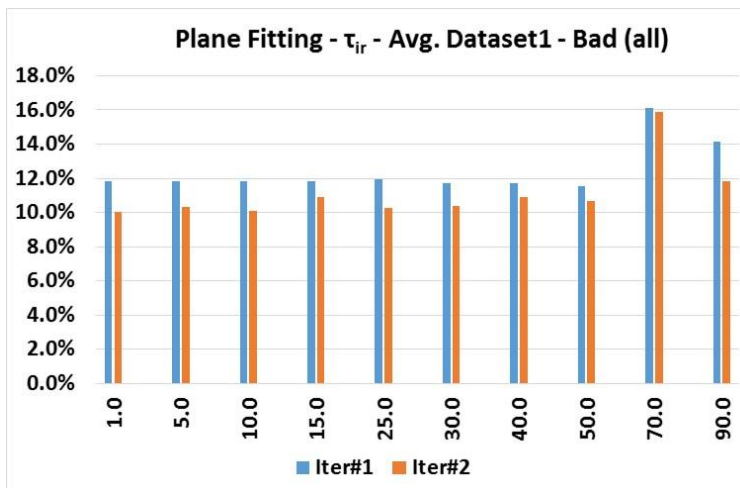
Figure D.12: Results for each of the image pair in Dataset #1 for different τ_{ic} as confidence threshold for determining inlier disparities to perform plane fitting.

inspected, it is seen that, as the threshold increases, more stable but less number of segments are marked as stable, which affects the total performance of the method. When analyzed along with the separate results of each image, although the threshold

can be determined separately for each image for better optimization, a value around 25% as the stable segment ratio threshold can be used for all the images in Dataset #1 experimentally.



(a)



(b)

Figure D.13: Average Results for different τ_{ir} as stable segment ratio threshold.

D.4.3 Distance Threshold for Outlier Disparities - τ_{od}

This parameter is used in Algorithm 3 when selecting the outlier pixels after a disparity plane is fitted to the segment. These pixels are later inspected for splitting segments from the initial segment. Figure D.15 shows the average results obtained for two iterations on Dataset #1 for increasing τ_{od} values, given in units of disparities.

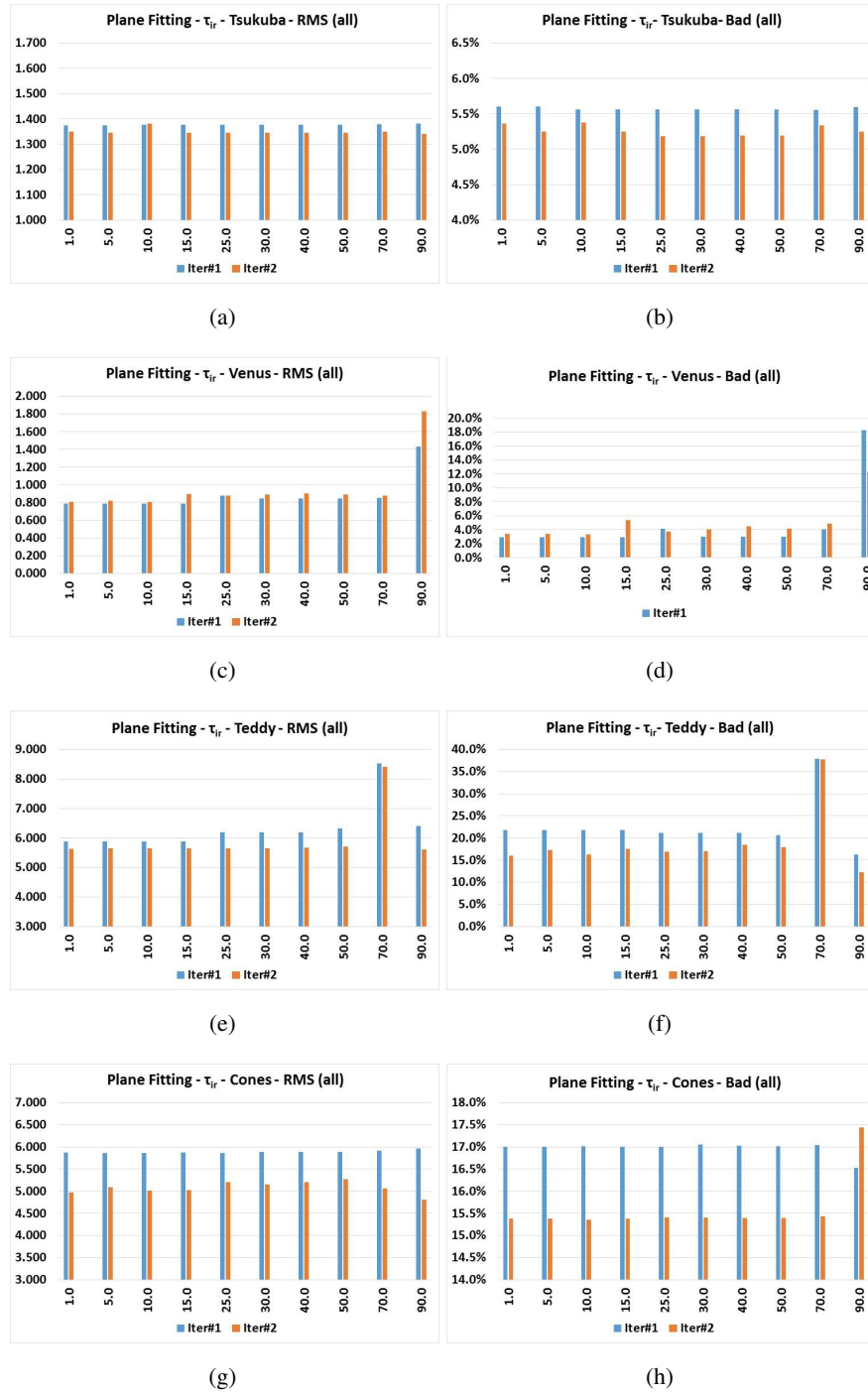
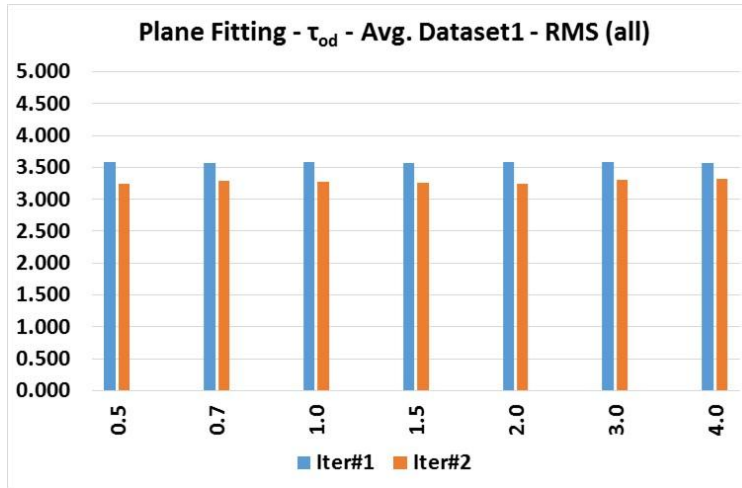


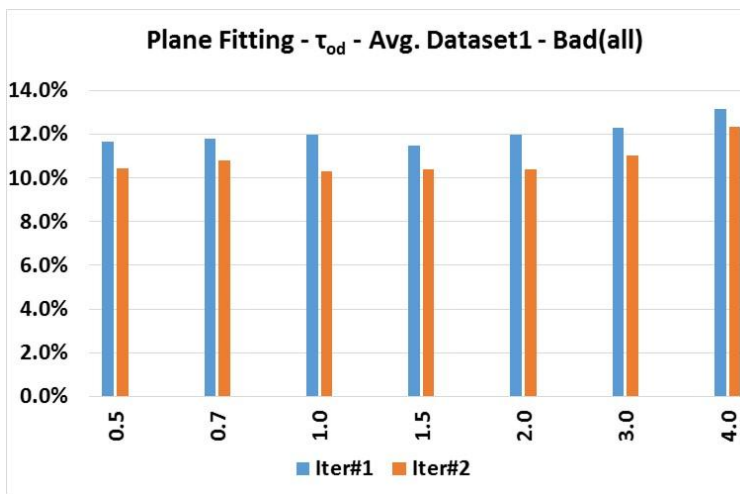
Figure D.14: Results for each of the image pair in Dataset #1 for different τ_{ir} as stable segment ratio threshold

Figure D.16 provides the separate results for each of the image in Dataset #1.

When the average results are inspected, as the threshold increases, more distant but less disparities are extracted, which affects the accuracy in the results due to split seg-



(a)



(b)

Figure D.15: Average Results for different τ_{od} as disparity distance threshold for determining outlier disparities to split segment after plane fitting.

ments. When the separate results are also analyzed, an optimal threshold is concluded to be in the interval $[1.0, 2.0]$ disparities.

D.4.4 Confidence Threshold for Outlier Disparities - τ_{oc}

This parameter is used in Algorithm 3 when constructing the outlier disparities that are confident enough after a disparity plane is fitted to the segment. These disparities are later inspected for splitting segments from the initial segment. Figure D.17 shows the average results obtained for two iterations using Dataset #1 for the increasing τ_{oc}

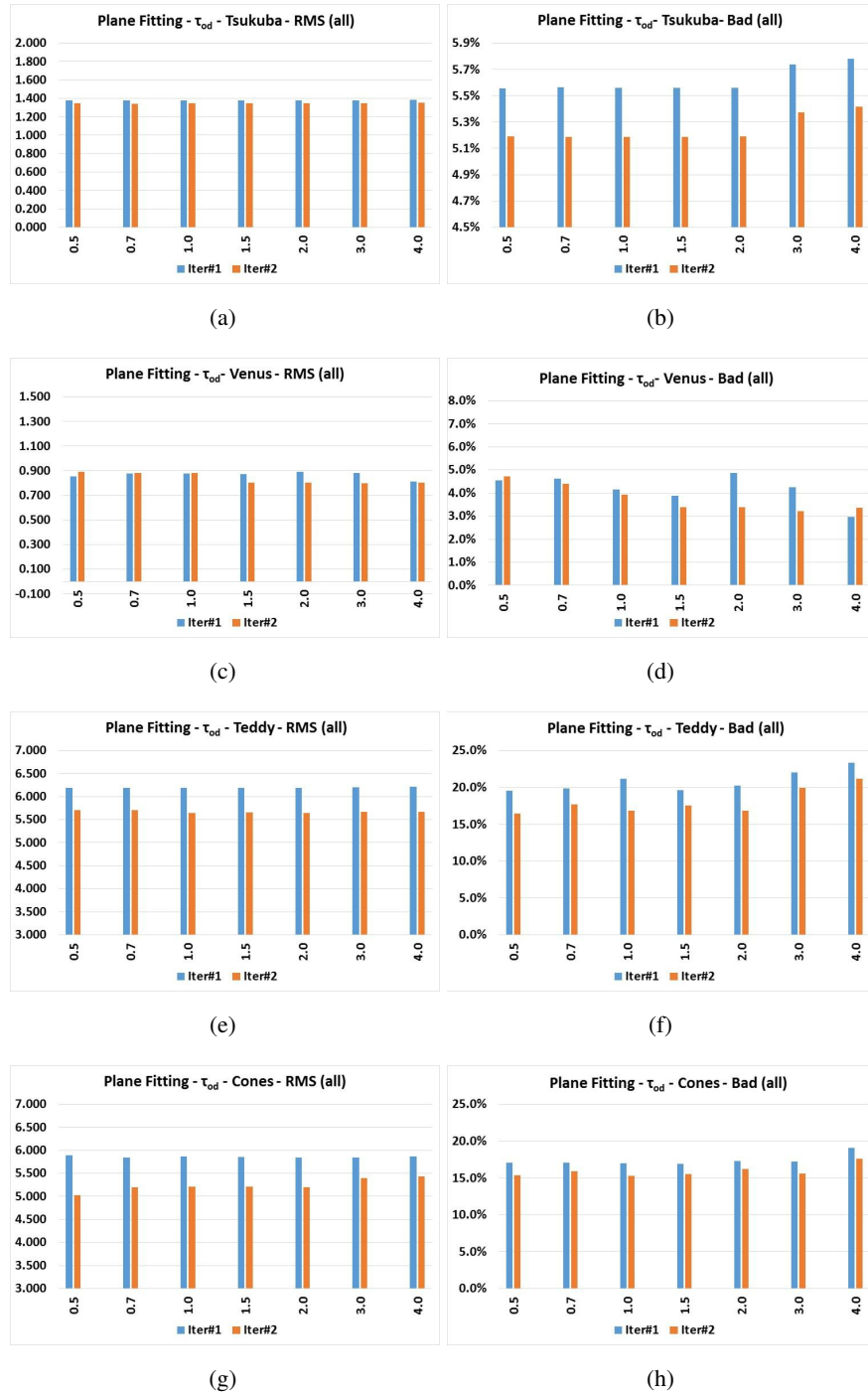


Figure D.16: Results for each of the image pair in Dataset #1 for different τ_{od} as disparity distance threshold for determining outlier disparities to perform segment splitting after plane fitting.

values given in units of 10s of the percentage (see Eqn. 5.14). Figure D.18 provides the separate results for each of the image in Dataset #1.



(a)



(b)

Figure D.17: Average Results for different τ_{oc} as confidence threshold for determining outlier disparities to split segment after plane fitting.

Inspecting the average results shows that, as the threshold increases, more confident but less disparities are extracted, which yields decreased performance in the results due to less split segments. When analyzed along with the separate results of each image, although the threshold can be determined separately for each image for better optimization, a value around 0.14% as the outlier confidence threshold can be used for all the images in Dataset #1 experimentally.

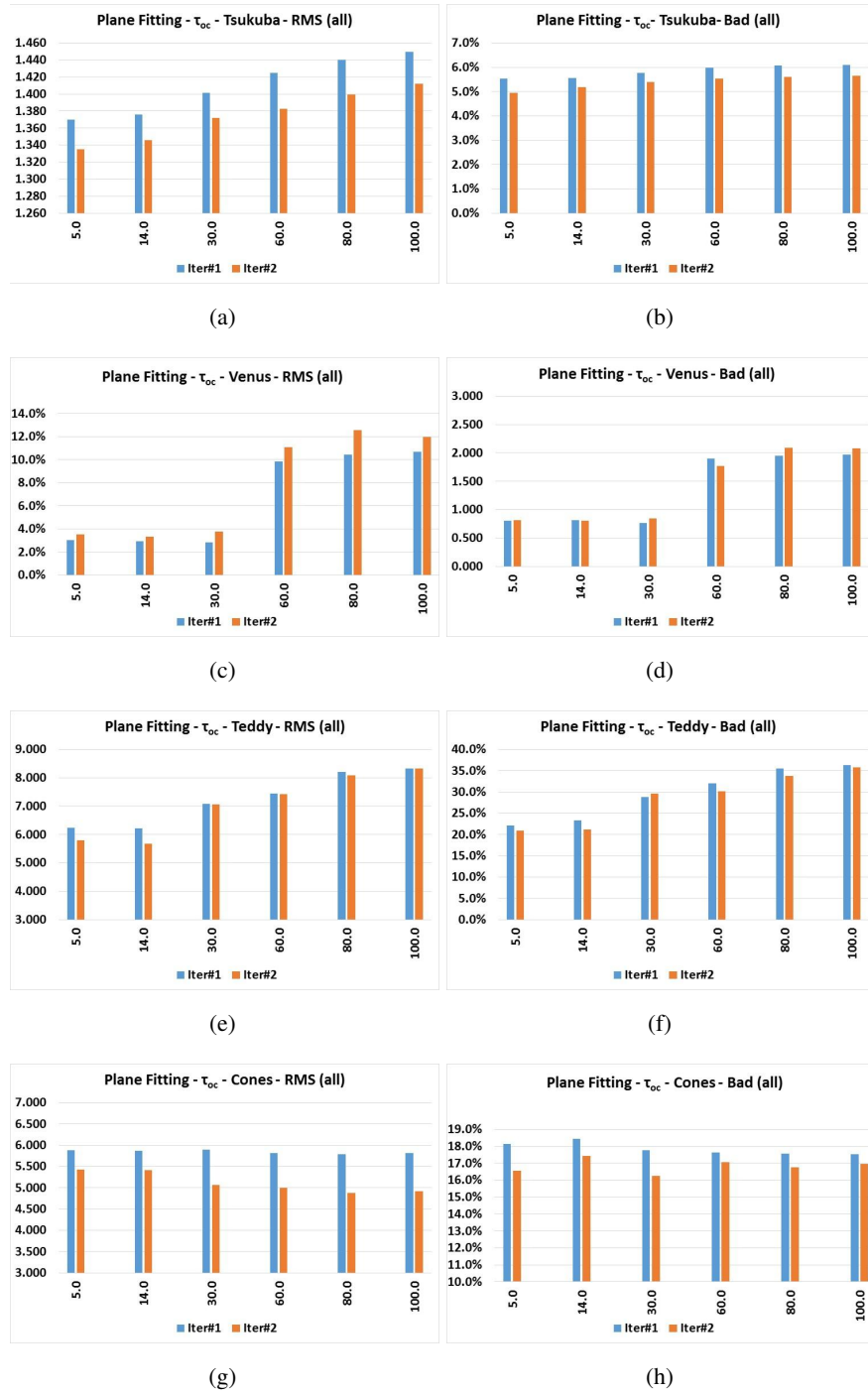


Figure D.18: Results for each of the image pair in Dataset #1 for different τ_{oc} as confidence threshold for determining outlier disparities to perform segment splitting after plane fitting.

D.4.5 Minimum Size Threshold for Segment Splitting of Outlier Disparities -

$$\tau_{os}$$

This parameter is used in Algorithm 3 when determining whether or not the selected outlier disparities of a plane fitted segment should continue for the segment split operation. Figure D.19 shows the average results obtained for two iterations using Dataset #1 for the increasing τ_{os} values given in units of number of pixels. Figure D.20 provides the separate results for each of the image in Dataset #1.



(a)



(b)

Figure D.19: Average Results for different τ_{os} as minimum size threshold for segment splitting

From the average results, it is seen that, as the threshold increases, less outlier dispar-

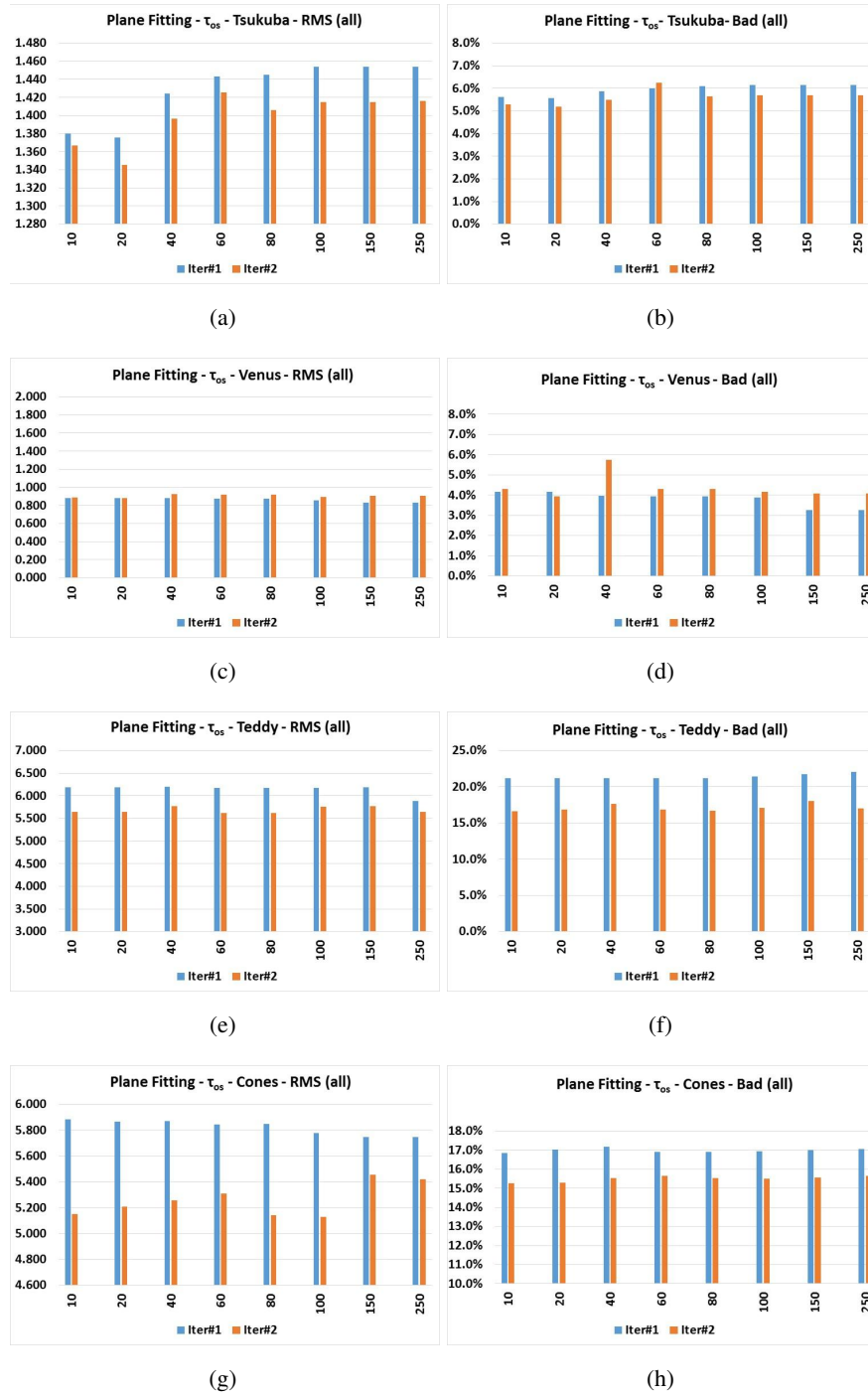


Figure D.20: Results for each of the image pair in Dataset #1 for different τ_{os} as minimum size threshold for selecting outlier disparity regions for segment splitting.

ity regions are selected for segment splitting, affecting the performance of the method. When analyzed along with the separate results of each image, although the threshold can be determined separately for each image for better optimization, a value around

20 pixels is concluded to be used for all the images in Dataset #1 experimentally.

D.5 Parameters of Segment Merging and Finalizing Step

In this section, the parameters of the Segment Merging and Finalizing step of the proposed method is analyzed.

D.5.1 Angle Threshold of Coplanar Disparity Planes for Segment Merging- τ_α

This parameter is used in Algorithm 4 when determining the coplanarity of two disparity planes - by checking whether the corresponding disparity planes are parallel. Figure D.21 shows the average results obtained for two iterations using Dataset #1 for the increasing τ_α parameters given in units of degrees. Figure D.22 provides the separate results for each of the image in Dataset #1.

When the average results are inspected, it is seen that, as the angle threshold increases, more non-coplanar segments are merged affecting the performance of the method negatively. When analyzed along with the separate results of each image, although the threshold can be determined separately for each image for better optimization, a small value less than 0.5 degrees as the angle difference between plane normals is concluded to be an optimal threshold for all the images in Dataset #1 experimentally.

D.5.2 Distance Threshold of Coplanar Disparity Planes for Segment Merging-

τ_{pd}

This parameter is used in Algorithm 4 in determining the coplanarity of two disparity planes, used as a threshold for the distance between the planes. Figure D.23 shows the average results obtained for two iterations using Dataset #1 for the increasing τ_{pd} values given in units of disparities. Figure D.24 provides the separate results for each of the image in Dataset #1.

When the average results are inspected, it is seen that, as the threshold increases, more non-coplanar segments are merged affecting the performance of the method



(a)



(b)

Figure D.21: Average Results for different τ_α as angle threshold for determining coplanar disparity planes for segment merging

negatively. When analyzed along with the separate results of each image, although the threshold can be determined separately for each image for better optimization, a small value such as 0.20 disparity distance between planes is concluded to be an optimal threshold for all the images in Dataset #1 experimentally.

D.6 Experiments on RGB and Cosine-Transformed RGB Image Pairs

In this section, the applicability of the defined cosine transform ($\cos(\pi I/255)$) that was applied to generate Dataset #1 for multi-modal stereo image correspondence

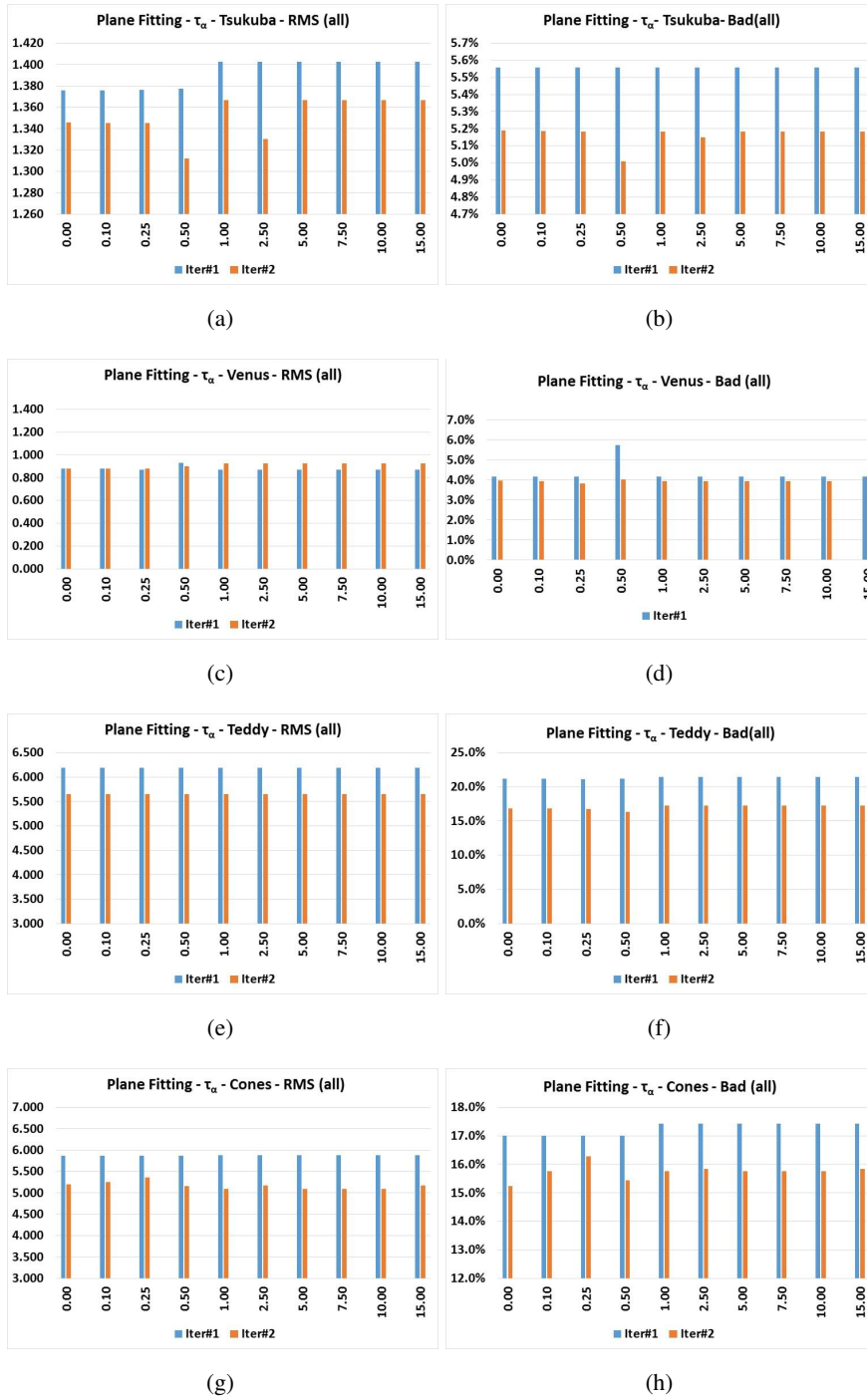
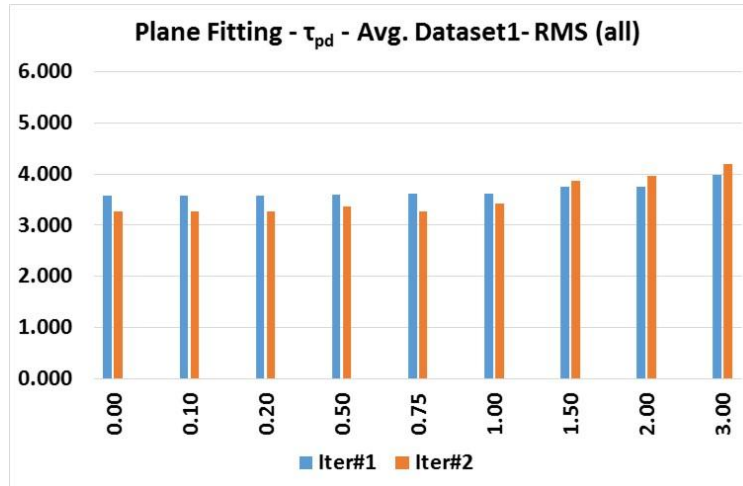
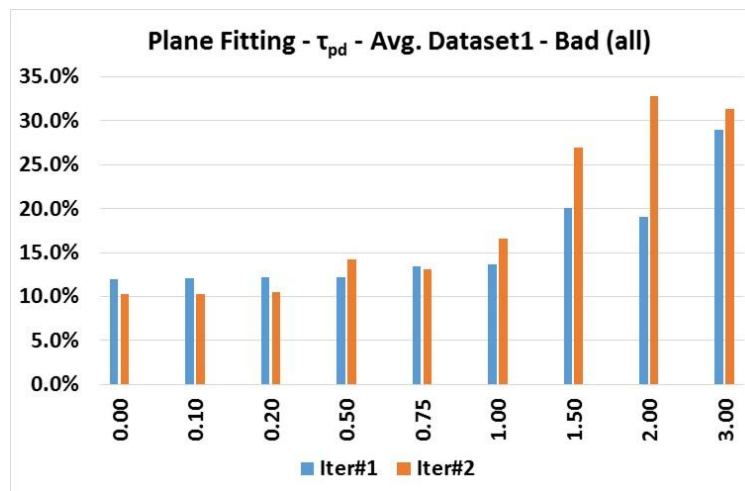


Figure D.22: Results for each of the image pair in Dataset #1 for different τ_α as angle threshold for checking coplanarity of two disparity planes for segment merging.

matching is analyzed over arbitrary RGB images. The goal is to see the performance of the proposed method in different scene characteristics. An image in this set is cosine-transformed and shifted by a fixed disparity. Figure D.25 provides the four



(a)



(b)

Figure D.23: Average Results for different τ_{pd} as disparity distance threshold for determining the coplanar disparity planes for segment merging

images that this experiment was applied. The ground truth disparities are 12, 11, 15 and 7 pixels respectively. The results are provided in Table D.3 for two iterations. As can be observed, proposed method is totally successful estimating the disparities of these images.

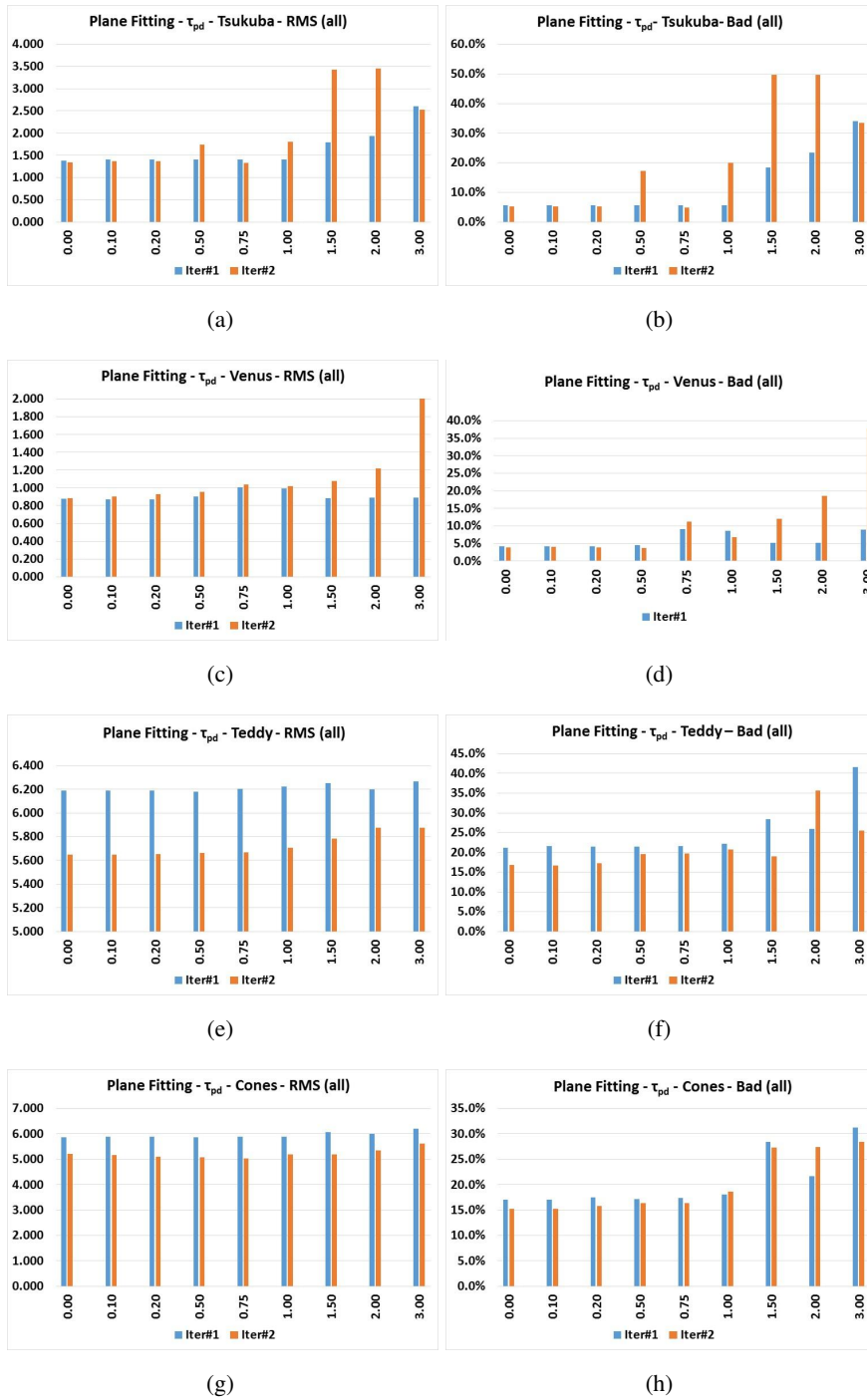


Figure D.24: Results for each of the image pair in Dataset #1 for different τ_{pd} as distance threshold for checking coplanarity of two disparity planes for segment merging.

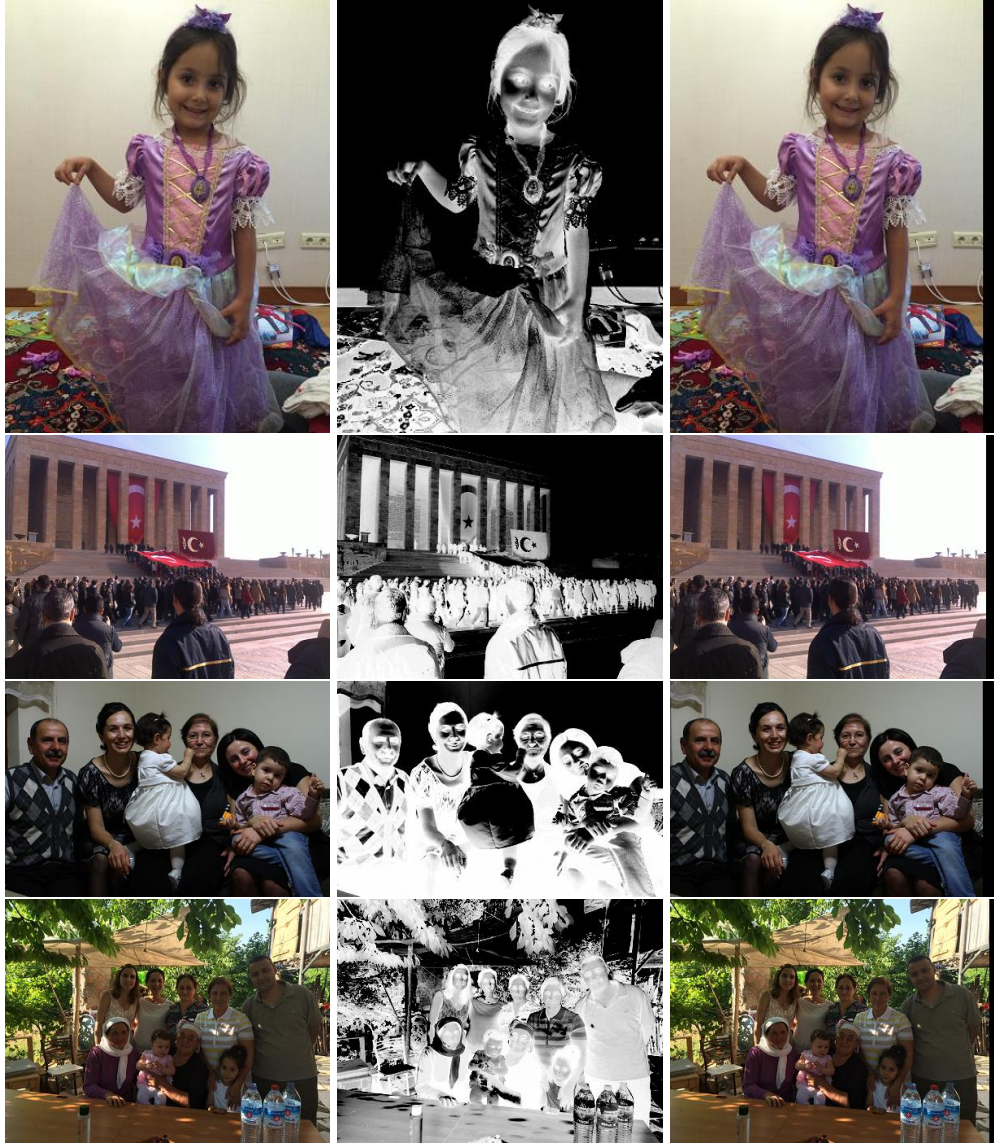


Figure D.25: The free-form RGB images converted to stereo image pairs. *Left column*: Original RGB image. *Middle column*: The cosine transformed images used as the left image. *Right column*: The shifted RGB images used as the right image.

Table D.3: Results of the Proposed Method on RGB images converted to Stereo Image Pairs.

Image	Method	Iter1	Iter2	Iter1	Iter2
		RMS (all)	RMS (all)	Bad (all)	Bad (all)
RGB#1	WTA of Adap.W.	0.985	0.608	2.0%	0.7%
	WTA of Agg.	0.678	0.413	0.5%	0.1%
	Plane Fitting	0.321	0.313	0.1%	0.0%
RGB#2	WTA of Adap.W.	0.269	0.0	0.0%	0.0%
	WTA of Agg.	0.229	0.0	0.0%	0.0%
	Plane Fitting	0.013	0.01	0.0%	0.0%
RGB#3	WTA of Adap.W.	0.0	0.0	0.0%	0.0%
	WTA of Agg.	0.0	0.0	0.0%	0.0%
	Plane Fitting*	0.015	0.09	0.0%	0.0%
RGB#4	WTA of Adap.W.	0.0	0.0	0.0%	0.0%
	WTA of Agg.	0.001	0.0	0.0%	0.0%
	Plane Fitting*	0.01	0.01	0.0%	0.0%

* Due to the subpixel disparity computation, plane fitting may yield RMS values > 0 even when WTA of Agg. has no errors.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: YAMAN, MUSTAFA

Nationality: Turkish (TC)

Date and Place of Birth: 18.05.1978, Boyabat/SİNOP

Marital Status: Married

Phone: 0 312 4737688

EDUCATION

Degree	Institution	Year of Graduation
M.S.	Computer Engineering Dept.,METU	2003
B.S.	Computer Engineering Dept.,METU	1999
High School	Fethiye-Kemal Mumcu Anadolu Lisesi	1995

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
1999-2003	Yön Ltd.	Software Engineer
2004-2013	AYESAŞ	Senior Software Engineer
2013-present	HAVELSAN	Senior Software Engineer

PUBLICATIONS

International Conference Publications

Yaman, M., Atalay V., Turker M., Çetin A.E., Gerek Ö.N., Texture Segmentation by using Adaptive Polyphase Subband Decomposition. "Proceedings of the 17nd International Symposium on Computer and Information Sciences", Orlando, Florida, USA (2002), p.81-85.

Yaman, M., Atalay V., Turker M, Texture Discrimination and Segmentation of Remote Sensing Images by using Adaptive Polyphase Subband Decomposition. "3rd International Symposium Remote Sensing of Urban Areas", 2, Istanbul, Turkey (2002), p.641-648.

Yaman, M., Kalkan S., "Multimodal Stereo Vision Using Mutual Information with Adaptive Windowing", 13th IAPR Conference on Machine Vision and Applications, Kyoto, Japan, 2013.

MS Thesis

Yaman, M., MS. Thesis, Texture Discrimination and Segmentation of Remotely Sensed Imagery By Using Adaptive Polyband Subband Decomposition, METU-CENG, Jan. 2003.