CHALMERS

Objective Evaluation of 3D Video Quality

Master of Science Thesis in Communication Engineering

Usman Hakeem

Department of Signals and Systems
Chalmers University of Technology
Göteborg, Sweden, 2011

Report No: EX102/2011

Abstract

The change of interest in the field of 3D video from time of advent of 3D technology to recent times has been significant. From standardization committees to broadcasters, many international organizations have been active to make advancements in the field of stereoscopic video. Standardization committees are working towards standards for 3D video, whereas broadcasters want better quality of 3D TV to be provided for their viewers. In the current scenario it becomes essential that a step towards standardizing an objective quality model for 3D video is taken.

In this thesis existing objective quality models for 3D video were studied and a new model was developed which evaluates the quality of 3D video based on the major 3D artifacts. The developed model works by extracting essential features needed to evaluate the stereoscopic 3D video quality. As numerous artifacts were present, specific evaluation of different artifacts was done using separate algorithms and at the end individual scores were combined to form the total quality score.

A subjective test was also performed which served as the basis for training and validation of the quality model. A linear function was formed after training on subjective video data to serve as the fundamental model for validation. The results show that the objective quality model correlates nicely with the subjective test data. The use of a nonlinear function shows further improvement but due to lack of subjective test data a nonlinear function was not used.

Table of Contents

CHAPTER 1 1

1 INTRODUCTION 1

1.1 Background 1
1.2 Objective and Scope 2
1.3 Methodology 2
1.4 Related Work 2
1.5 Organization of Thesis 3

CHAPTER 2 5

2 3D ARTIFACTS 5

2.1 Introduction 5
2.2 Artifacts 5

CHAPTER 3 9

3 SUBJECTIVE TEST 9

3.1 Introduction 9
3.2 Test Description 9
3.3 Analysis of 3D subjective test 12

CHAPTER 4 21

4 2D IMPLEMENTATION METHODS 21

4.1 Introduction 21
4.2 PSNR 21
4.3 SSIM 21
4.4 LU Factorization 22
4.5 Opticom's Video Quality Measure PEVQ 23
4.6 Comparison of 2D Methods 24

CHAPTER 5 28

5 3D PARAMETERS 28

5.1 Introduction 28
5.2 Initial disparity estimation process 28
5.3 SIFT Algorithm 29
5.4 Disparity Estimation Algorithm 33
5.5 Comparison of disparity map methods 41

CHAPTER 6 44

6 OBJECTIVE 3D QUALITY MODEL 44

6.1 Introduction 44
6.2 Block Diagram 44
6.3 Calculate 2D score 45
6.4 Detect asymmetric coding 45
6.5 Difference in Luminance and Contrast 46
6.6 3D Parameters 46

6.7 SELECTION OF PARAMETERS 47

CHAPTER 7 50

7 RESULTS 50

7.1 INTRODUCTION 50
7.2 METHOD AND RESULTS 50
7.3 SUMMARY OF RESULTS 53

CHAPTER 8 55

8 CONCLUSION AND FUTURE WORK 55

REFERENCES 57

List of Figures

Figure 1: Factors affecting the 3D Video....5
Figure 2: Coding artifacts....6
Figure 3: Vertical disparity....6
Figure 4: Color disturbance impact....7
Figure 5: Asymmetric coding....7
Figure 6: Voting scale for 3D subjective test....9
Figure 7: Test content....10
Figure 8: Distribution of subjective test for all votes....14
Figure 9: Mean scores with confidence intervals....15
Figure 10: Coding artifacts....16
Figure 11: Varying baseline effect....16
Figure 12: Vertical disparity effect....17
Figure 13: Color disturbance effect....18
Figure 14: Asymmetric coding....18
Figure 15: Average score per content for all test conditions....19
Figure 16: LU factorization model....23
Figure 17: PEVQ model....23
Figure 18: QP Coding levels vs. Objective scores for 2D Methods....24
Figure 19: Linear mapping of 2D methods....26
Figure 20: Disparity estimates using SIFT matching algorithm....28
Figure 21: DOG pyramid....29
Figure 22: Pixel comparison in DOG pyramid to find maxima and minima....30
Figure 23: Keypoints for Balloon and Newspaper sequence using SIFT....32
Figure 24: Keypoint matching using SIFT descriptor matching....33
Figure 25: Disparity between left and right image pairs....34
Figure 26: Overlapped left/right image pair with its ground truth disparity map....34
Figure 27: Local window matching algorithm....35
Figure 28: Disparity maps using different window sizes....35
Figure 29: Left: Ground truth disparity map, Right: Disparity map using adaptive window method....37
Figure 30: Left: Disparity Map using geodesic method, Right: Disparity Map using Fast Cost-Volume..37
Figure 31: Slanted surface problem....38
Figure 32: Left: Ground truth Right: Shape and correspondence method disparity maps....40
Figure 33: Balloon sequence disparity maps using shape and correspondence stereo algorithm....40
Figure 34: Middlebury stereo evaluation results....41
Figure 35: Left: Disparity Map Cost Volume Filtering, Right: Disparity Map Stereo Correspondence....42
Figure 36: Objective evaluation model for 3D video....44
Figure 37 Asymmetric coding of left and right image pairs....45
Figure 38: PCA analysis for baseline changed sequences....47
Figure 39: PCA analysis for asymmetric sequences....48
Figure 40: Predicted MOS vs. MOS and Residuals for linear regression....51
Figure 41: Predicted MOS vs. MOS and Residuals non-linear regression....51
Figure 42: Training and validation results using linear regression....52

List of Tables

Table 1: Optimal disparity for the six video sequences 12
Table 2: Pearson correlation coefficients with MOS for nine subjects 13
Table 3: Statistical parameters for artifacts 13
Table 4: Output parameters of PEVQ 24
Table 5: Different 2D methods scores for color disturbance (Only left pair) 25
Table 6: Average scores for baseline 1 sequences for several 2D methods 25
Table 7: Asymmetric coding scores for Lovebird sequence 45
Table 8: Luminance disturbance scores 46
Table 9: List of Parameters used for objective evaluation 48
Table 10: Correlation between Subjective MOS and Predicted MOS 51

List of Abbreviations

MVC Multiview Video Coding 1
OpenCV **Open** Source **C**omputer **V**ision 2
PEVQ Perceptual Evaluation of Video Quality 3
ITU-T Telecommunication Standardization Sector 3
ACR-HR Absolute Category Rating with Hidden Reference 9
QP Quantization Parameter 11
CI Confidence Interval 12
MOS Mean Opinion Score 12
MSE Mean Squared Error 21
SSIM Structural Similarity 21
SIFT Scale Invariant Feature Transform 28
DOG Difference of Gaussian 29
DSCORE Differential score 45
PCA Principal Component Analysis 47
SURF Speeded Up Robust Feature 55

Acknowledgments

This report is part of master thesis for the Communication Engineering Master program at Chalmers University of Technology and is carried out at Ericsson's Visual Technology department in Kista, Stockholm.

I would like to pay my sincere gratitude to my Ericsson supervisor Martin Pettersson who has helped me throughout my thesis. I am deeply indebted to him for his time, assistance and the knowledge he has shared with me during my thesis. I would also like to thank all the Ericsson researchers and manager of Visual Technology department Per Fröjdh for taking the time out of their schedule to take part in the subjective test.

I would like to thank my Chalmers examiner Erik Ström for taking the time out to evaluate my thesis and being one of the best professors during my Master's program. I am also grateful to my Chalmers supervisor Lotfollah Beygi for his time in keeping track of the work during the course of my thesis. Finally I would like to thank my parents and friends who have helped me in so many ways. Without their contribution this would not have been possible.

Usman Hakeem, Stockholm, September 2011.

Chapter 1

Introduction

1.1 Background

3D is one of the lesser understood forms of video making in present times. The first work on 3D stereoscopy dates back to 1838 when Charles Wheatstone showed that the difference in left and right image is interpreted by brain as a unified single object of three dimensions. When we see the world around us with our two eyes, we experience binocular stereopsis. It is the ability of our brain to fuse two images of slightly different perspective, enabling us to perceive depth. Depending on the distance between the two images, some of the objects appear closer from the screen. While, other objects may appear further away from the screen.

The conventional form of stereoscopy is to use two 2D images with each providing a different perspective to the brain. A slight change in perspective in horizontal direction allows the brain to perceive depth. The horizontal difference between the two 2D images, i.e., the left and right views is called disparity. Disparity is a very important cue in the perceived spatial depth of the object in stereoscopic vision. To view the stereoscopic 3D video the viewer usually has to wear 3D glasses unless the source splits the images directionally into the viewer's eyes. The newer technology in 3D displays is autostereoscopic, which does not require glasses and this is the reason for it being called "glasses free 3D". In autostereoscopic displays for multiple viewers, several views are used to generate the 3D video allowing more flexibility in viewing angle thus enabling multiple viewers to watch the 3D video.

Apart from the advances in display technologies in 3D video, work is also carried out to standardize coding schemes for multiview 3D. Multiview Video Coding (MVC) [1] enables efficient encoding of sequences captured from multiple cameras, e.g., two or three views are transmitted and at the receiver end additional intermediate views can be synthesized for free viewpoint TV which utilizes up to 28 views for new displays. In all a lot of work is being put in 3D video, so that viewers can use 3D video in different applications. On the other hand there has not been much work done to evaluate the quality of 3D video being produced from different displays.

Evaluation of video data in general can be done based on subjective or objective tests. In subjective test several viewers are shown the test video and are asked to rate it on a scale. The process of subjectively evaluating the quality of a video is expensive and time consuming contrary to objective tests. On the other hand accurate subjective tests serve as benchmark for evaluation of objective tests. There are many objective methods available which can be used to evaluate the quality of 2D video with high accuracy. In case of 3D video there is lack of work both in terms of subjective tests done and research on objective measures until recently, where performance evaluation of 3D video using subjective and objective measures has been carried out which will be discussed in section 1.4.

1.2 Objective and Scope

With increasing demand in 3D video it has become necessary that viewers can see high quality 3D videos. For that it is necessary that there should be objective methods which can evaluate 3D quality accurately. The objective of this thesis is the implementation of objective quality measures which can evaluate the quality of the 3D video data. The scope of this thesis is to build a full reference model to evaluate the 3D video quality. The emphasis of the 3D video quality method is on capture/calibration factors that include vertical/horizontal disparities, color distortions and camera baseline distances. Apart from the above mentioned disturbances coding artifacts have also been considered as well.

1.3 Methodology

The work in this thesis was carried out in different phases. The initial phase was the study of literature regarding the basics of the technology being worked in this area. The study included the 3D video technical aspects and its relation with the human visual system. Then few 2D video quality evaluation methods were implemented, the 2D quality evaluation is an integral part of the complete 3D model. Then 3D artifacts were identified based on which subjective test was carried out. After subjective test was carried out the most important phase was the development of objective metrics which could evaluate the quality of 3D artifacts specifically. The final phase of the thesis was to combine 2D quality metrics with 3D metrics to compute the objective scores for video sequences used in subjective test and find the correlation between them. The implementation was mostly done in C using **Open** Source **C**omputer **V**ision (OpenCV) [2] library and Matlab. Statistical analysis was done in Microsoft Excel where needed.

1.4 Related Work

Existing work in the area of objective evaluation of 3D video is of limited scope until recently. Lately there have been several papers which propose different models for the quality evaluation of 3D video. Initial work [3] done in this area was to use the existing well developed 2D video quality models to evaluate the quality of 3D video. The differences in 2D and 3D video are significant in terms of human perception and thus the 2D methods do not provide an even near optimal solution for 3D video evaluation. The next step in 3D evaluation was to fuse disparity metric along with 2D quality metric to evaluate the quality of 3D video. [4], [5] and [6] use this criterion for evaluation of 3D video. In other literature particular 3D artifacts have been considered and based on the specific artifact the evaluation of 3D video is done. [7] presents a quality assessment algorithm based on average luminance and contrast of left/right views for 3D video. In [8] color and edge distortion metrics were used for evaluation of 3D video. Feature extraction from disparity map was used in [9] to find the best features that correlate with the subjective tests.

This thesis addresses all the important 2D and 3D artifacts evaluation metrics, which have been considered in the work done previously. Apart from that, several other artifacts like vertical disparity and asymmetric view effects have been considered in this thesis as well.

1.5 Organization of Thesis

The thesis is divided into 8 chapters. The next chapter presents the artifacts considered for subjective test, whereas chapter 3 describes the methodology for subjective test and analysis is also provided. In chapter 4, several 2D methods that are implemented are presented and compared to Perceptual Evaluation of Video Quality (PEVQ) which is a Telecommunication Standardization Sector (ITU-T) J.247 standard metric [10]. In chapter 5 the 3D algorithms used to extract 3D parameters from the two stereoscopic views that are used in the final model are discussed. Chapter 6 provides acquaintance with the complete 3D model developed in this thesis for evaluation of the artifacts mentioned in chapter 2. The results of the objective evaluation are provided in chapter 7. The final chapter contains a brief summary of the work in this thesis, moreover some ideas are provided on which future work can be based upon.

Chapter 2

3D Artifacts

2.1 Introduction

In 2D video there are a number of artifacts that need to be considered. When it comes to 3D video the number of artifacts increases further. In different phases of 3D video different artifacts are introduced. Several phases in 3D video along with the artifacts or factors that affect the 3D video quality are given in the Figure 1 below.

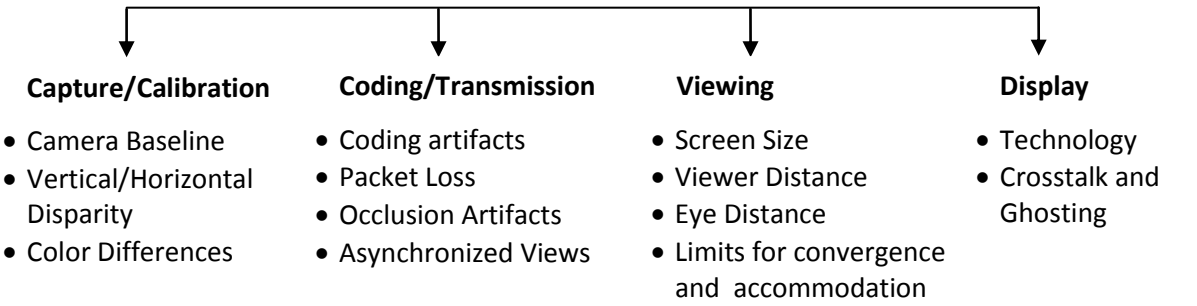


Figure 1: Factors affecting the 3D Video

This thesis considers the artifacts that mainly occur due to capture/calibration and coding. The next section describes the specific 3D artifacts in a bit more detail.

2.2 Artifacts

As mentioned in section 2.1, there are numerous stages in which different errors can occur in 3D video. From acquisition to display every intermediate stage can introduce anomalies in the 3D video. As acquisition of new 3D sequences was not possible, this meant restriction in the number of artifacts that could be introduced in this thesis. The following artifacts were considered in the 3D subjective test.

✓ Coding artifacts due to MVC/H.264,

- Blocking – Block-based video coding produces artifacts known as blocking artifacts which may occur due to transforms and quantization used for compressing the video.
- Cardboard effect – Cardboard effect is the phenomenon in which objects on the screen appear to be unnaturally flat, i.e., with no depth.
- Blurring – Blurring is also mainly a 2D artifact coming from low bit rate encoding in combination with the deblocking filter. Blurring in 3D video may occur due to differences in brightness (difference of lightning among two cameras) and compression between views.
- Pixelation – Pixelation is an artifact typical in 2D video which results in individual pixels becoming visible. In 3D this artifact may be visible if the objects appear too close to the viewer.
- Ringing – Ringing artifact occurs in 2D due to video coding and in case of 3D, depth ringing may occur due to coding of depth map.
- Depth inconsistencies – Depth inconsistencies may also occur due to coding of depth map.

✓ Some of the above coding artifacts are common in 2D video but cardboard effect and depth inconsistency are specific artifacts related to 3D video. These artifacts occur due to the coding of the depth map of a certain scene. Figure 2 below shows the blocking/blurring effect due to coding.

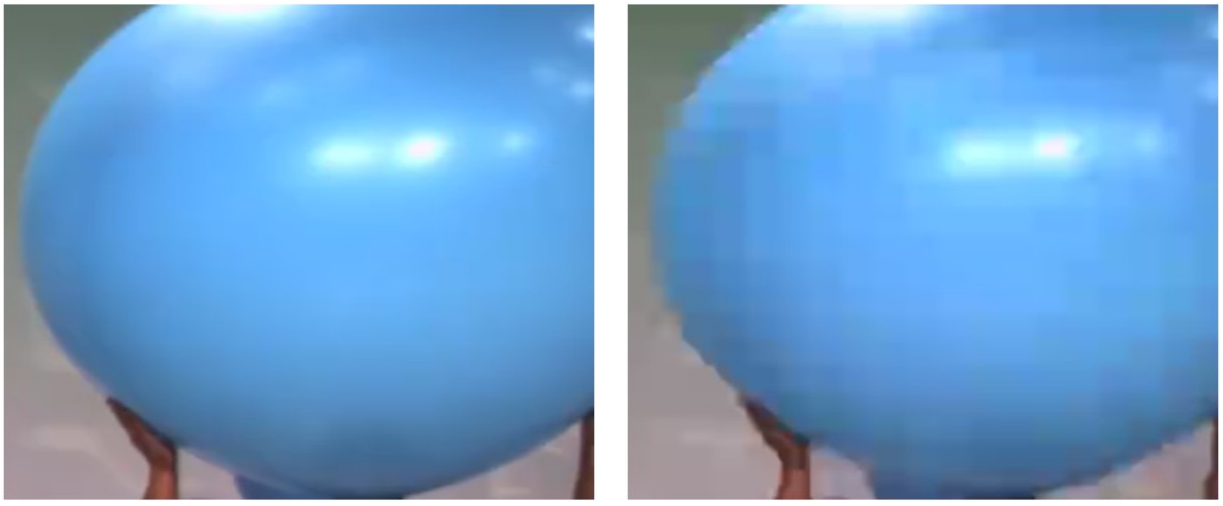


Figure 2: Coding artifacts

✓ Puppet theatre effect due to large camera baseline distances which results in excessive positive parallax. As the camera baseline distance increases the depth effect also increases. From a certain point onwards too much depth becomes annoying and can cause headache to the viewer. Thus it is of utmost importance that an optimal baseline distance between the two views is used in 3D video.

✓ Vertical disparity occurs when one stereo pair is shifted vertically than the other stereo pair. This artifact occurs during the acquisition phase of 3D video and usually adjustments are done to mitigate this affect. Slight changes in vertical disparity are not noticeable but large variations have an impact on the 3D experience. Figure 3 below shows the left and right stereo pair with a vertical shift between them. To get some idea of the 3D effect and the artifact the viewer can look at the Figure from a proper distance by crossing the eyes.

Figure 3: Vertical disparity

✓ Negative or cross eyed parallax is when right image is switched with left one. This affect is more noticeable for certain video sequences than the others as the subjective test suggests. With large camera baseline distances this affect is quite apparent. Nearer objects tend to appear further away from the viewer as opposed to the general 3D experience.

✓ Color changes like fading or luminance disturbances occur due to calibration errors between the two cameras. A slight change in luminance and contrast can decline the quality of 3D video.

Figure 4: Color disturbance impact

✓ Apart from the above artifacts the affect of asymmetric coding was also evaluated through subjective scores. When one of the views is slightly more compressed than the other view then it is called asymmetric coding. The severity of the reduction in 3D quality depends on the difference in the compression ratio between the two views. Some human beings are left eyed while others are right eyed [11], so this factor also influences the experience of 3D in this case.

Figure 5: Asymmetric coding

Chapter 3

Subjective Test

3.1 Introduction

The methodology used for performing 3D subjective tests is not fully mature. Research is ongoing to date on how to perform 3D subjective tests and what conditions are best suitable for tests. ITU-T has recommended requirements on how to perform the subjective video tests in [12] for multimedia applications. This chapter outlines the details of the subjective test performed and then analysis on the results is also provided.

3.2 Test Description

3.2.1 Test Method

The subjective test method used was Absolute Category Rating with Hidden Reference (ACR - HR). In this method the viewer is shown processed as well as reference (hidden) video sequence. In ACR-HR viewer does not have any knowledge when the reference video sequence is shown. In subjective test carried out hidden reference was shown twice to the viewer. The length of all video sequences was 10 seconds and further 5 seconds were provided to subject for voting. Subject was asked to rate perceived total quality of the presented sequence without the knowledge of reference sequence on the following scale:

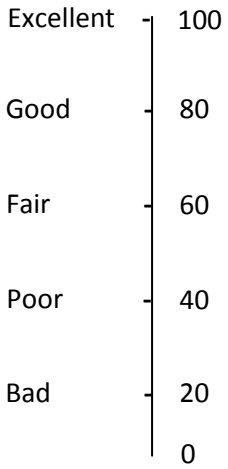


Figure 6: Voting scale for 3D subjective test

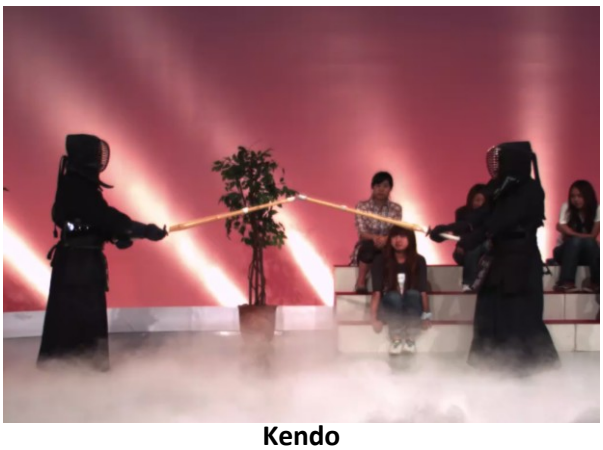
The subject could position the slider anywhere on the scale and the corresponding registered score would be between 0 and 100. A single scale for voting was used as it becomes easier to analyze the results in this way. After the subjective test each subject was asked to fill in a short questionnaire which would help in building the model for objective evaluation.

3.2.2 Test Content

The selected video sequences are among the multi-view sequences being used in the MPEG 3DV standardization. Each sequence had duration of 10 seconds. The six video sequences used for the test were

1). Balloon [Resolution 1024 x 768. Frame Rate 30 FPS]
2). Kendo [Resolution 1024 x 768. Frame Rate 30 FPS]
3). GT_FLY [Resolution 1920 x 1088.Frame Rate 25 FPS]
4). Lovebird1 [Resolution 1024 x 768. Frame Rate 30 FPS]
5). Newspaper [Resolution 1024 x 768. Frame Rate 30 FPS]
6). Poznan Street [Resolution 1920 x 1088. Frame Rate 25 FPS]

Balloon



Kendo

GT_FLY

Lovebird1

Newspaper

Poznan Street

Figure 7: Test content

3.2.3 Test Conditions

MVC was used for coding all the sequences except for asymmetric case where H.264 was used. Some short definitions of the terms used below are given.

Disparity: Disparity is the horizontal difference in image location of an object seen by left and right eyes. When stereoscopic video is recorded it is the horizontal distance between two points in stereo pair images. Further explanation of disparity is provided in section 5.4.1.

Baseline distance: In stereoscopy a single object is captured from two view points. The distance between the two viewpoints is called baseline distance. The optimal baseline distance is the distance between the left and right eye.

Quantization parameter: Quantization in MVC is controlled by quantization parameter. The lower its value the better is the quality of output.

Total number of video content = 6

Uncoded video sequences

Baseline = 1, Disparity = 0 [06x2 (shown twice)]	=	12 Sequences

Baseline/Disparity changes:

Baseline = 0, Disparity = 0, 2 Quantization Parameter (QP) levels [2D case]		
QP Level 32 & 42	=	12 Sequences
Baseline = 1, Disparity = 0, 4 QP levels		
QP Level 27, 32, 37 & 42	=	24 Sequences
Baseline = 2, Disparity = 0, 2 QP levels		
QP Level 32 & 42	=	12 Sequences
Baseline = 4, Disparity = 0, 2 QP levels		
QP Level 32 & 42	=	12 Sequences
Baseline = 2, Disparity = Optimal, 2 QP levels		
QP Level 32 & 42	=	12 Sequences
Baseline = 4, Disparity = Optimal, 2 QP levels		
QP Level 32 & 42	=	12 Sequences
Total Sequences	=	84 Sequences

Vertical disparity effect tests:

Vertical shifts: Pixel shift of 6, 12 and 24 were used.

Number of vertically shifted video sequences for each video = 3		
Baseline = 1, Disparity = 0, 2 QP levels		
QP Level 32 & 42	=	36 Sequences

Color disturbance effect tests:

Color 1 disturbance: Scale red component 1.5 times.

Color 2 disturbance: Scale luminance 1.5 times.

No. of color changed video sequences for each video	= 2	
Baseline = 1, Disparity = 0, 1 QP level		
QP Level 32	=	12 Sequences

Asymmetric coding effect tests:
QP Pairs: [27-27], [27-32], [27-38], [37-37], [37-42], [37-48] were used.
Number of fixed QP's for left view = 2
Number of changing QP's for right view = 3
Baseline = 1, Disparity = 0, Changing QP Levels = 36 Sequences

Negative parallax effect tests:
Baseline = -1, Disparity = 0, 1 QP level
QP Level 32 = 06 Sequences

Total = 186 Sequences

For baseline 2 & 4, the disparities chosen to be optimal for different sequences by visual inspection are provided in the table below.

	Optimal Disparity [Baseline 2]	**Optimal Disparity [Baseline 4]**
Balloons	35	70
GT_FLY	20	40
Kendo	25	50
Lovebird	15	30
Newspaper	70	140
Poznan	60	120

Table 1: Optimal disparity for the six video sequences

3.2.4 Test Subjects

The subjects used for subjective evaluation were experts in the field of 3D, who have worked with 3D video. This allowed getting better results. Nine experts performed subjective test according to the ACR-HR criteria mentioned above.

3.3 Analysis of 3D subjective test

The results of the subjective test were analyzed using Microsoft Excel with statistical functions. All plots in the analysis are shown with a 95% confidence interval (CI). The main parameters considered were correlation coefficient, standard deviation, confidence intervals, kurtosis and skewness.

The correlation coefficient was used to determine the correlation with Mean Opinion Score (MOS) for each test subject, so that any outlier can be determined and results can be discarded for that subject if needed. Standard deviation was used to look at the variety/diversity in the results. The CI was used to determine the range of interval in which the value of the results could be considered true and the amount of uncertainty in the results. Kurtosis tells us about the infrequent deviations in the results compared with the normal distribution, whereas skewness determines the degree of asymmetry of a distribution with respect to its mean. The mean and standard deviation of a distribution could be same even if two distributions are different, so these other parameters (kurtosis/skewness) are used to distinguish between them.

For the purpose of analysis results were categorized according to different criteria. Below defined are the different parameters evaluated from the subjective test results.

3.3.1 Pearson correlation coefficient

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9
Correlation Coefficient with MOS	0.861	0.845	0.840	0.838	0.820	0.755	0.800	0.830	0.760

Table 2: Pearson correlation coefficients with MOS for nine subjects

As we can see from the results that the correlation of the subjects with MOS is high for all subjects, so there was no need to screen the results.

3.3.2 Statistical parameters

	Mean	Confidence [95%]	Standard Deviation	Count	Kurt	Skew
Uncoded	72,23	4,81	18,03	54,0	0,48	-0,81
2D-QP32	59,72	4,89	18,32	54,0	0,35	-0,13
2D-QP42	22,94	3,38	12,67	54,0	-0,68	0,33
BL1-D0-QP27	68,04	4,62	17,33	54,0	-0,08	-0,56
BL1-D0-QP32	61,30	4,59	17,22	54,0	0,02	-0,31
BL1-D0-QP37	46,09	4,92	18,43	54,0	0,34	0,32
BL1-D0-QP42	25,13	2,95	11,05	54,0	-0,22	0,38
BL2-D0-QP32	53,57	6,55	24,56	54,0	-0,75	-0,25
BL2-D0-QP42	24,13	3,43	12,86	54,0	-0,87	0,28
BL4-D0-QP32	39,11	7,41	27,80	54,0	-1,35	0,16
BL4-D0-QP42	17,11	3,33	12,48	54,0	-0,28	0,70
Color1-BL1-D0-QP32	52,17	5,29	19,64	53,0	-0,61	-0,43
Color2-BL1-D0-QP32	45,46	6,13	22,98	54,0	-0,88	-0,04
Negative-BL1-D0-QP32	50,61	5,94	22,27	54,0	-0,73	-0,27
Vertical6-BL1-D0-QP32	46,35	6,43	24,10	54,0	-0,96	-0,20
Vertical6-BL1-D0-QP42	23,72	3,62	13,43	53,0	-0,92	0,33
Vertical12-BL1-D0-QP32	54,46	5,22	19,57	54,0	-0,50	-0,39
Vertical12-BL1-D0-QP42	23,50	3,38	12,55	53,0	-0,32	0,31
Vertical24-BL1-D0-QP32	37,14	6,41	23,57	52,0	-0,71	0,49
Vertical24-BL1-D0-QP42	17,06	3,45	12,94	54,0	-0,01	0,78
Assy_BL1_D0_QP27-27	68,99	4,57	16,96	53,0	-0,20	-0,45
Assy_BL1_D0_QP27-32	67,70	4,64	17,38	54,0	0,35	-0,68
Assy_BL1_D0_QP27-38	50,25	4,42	16,40	53,0	0,30	-0,57
Assy_BL1_D0_QP37-37	49,11	4,03	15,10	54,0	-0,16	-0,19
Assy_BL1_D0_QP37-42	36,21	3,89	14,43	53,0	-0,58	0,33
Assy_BL1_D0_QP37-48	19,49	3,05	11,23	52,0	-0,69	0,05
BL2_DOpt_QP32	65,56	4,99	18,71	54,0	-0,19	-0,59
BL2_DOpt_QP42	26,06	3,36	12,59	54,0	-0,39	0,35
BL4_DOpt_QP32	61,37	6,10	22,66	53,0	-0,59	-0,51
BL4_DOpt_QP42	25,52	3,81	14,30	54,0	-0,94	0,33

Table 3: Statistical parameters for artifacts

In Table 3 shown above there are some key points which are noticeable. In the standard deviation column, we can see that, the baseline 2 & 4 without disparity compensation has larger deviation. In addition, baseline 4 with optimal disparity has a high deviation. The reason for this may have been due

to the fact that some subjects liked the sensation of greater depth, while for others it was quite annoying. Another reason for this could be the variation in the content in terms of the perceived depth, so in some sequences excessive depth didn't had as annoying effects compared with some of the other sequences.

3.3.3 General results

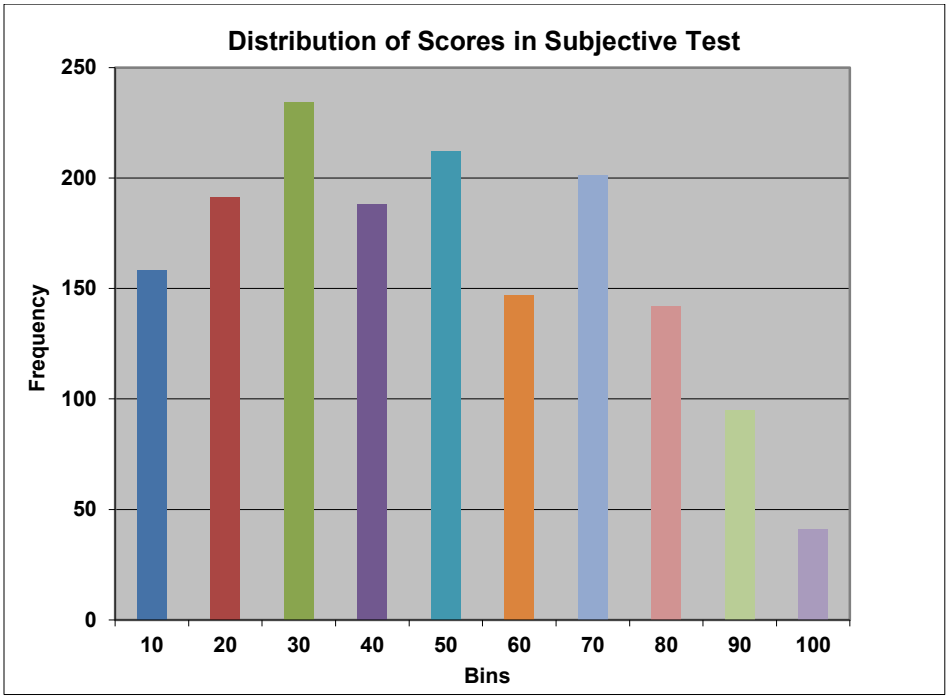


Figure 8: Distribution of subjective test for all votes

The distribution of the subjective ACR test is shown in Figure 8. The distribution is somewhat close to a normal distribution. The average skewness turns out to be -0.04 (close to zero) which supports above argument to some extent. From this we assume in further analysis that subjective test results are normally distributed.

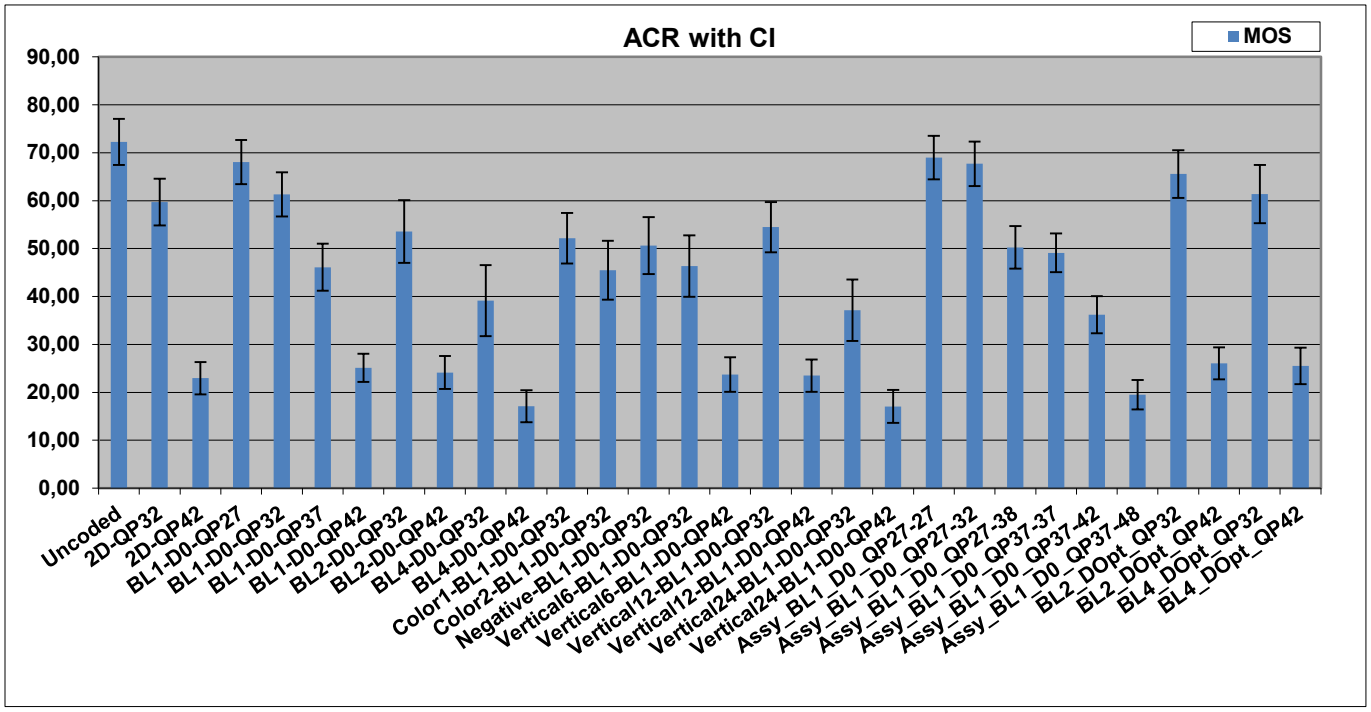


Figure 9: Mean scores with confidence intervals

Figure 9 and Table 3 above show the results of subjective test for all the artifacts. It can be seen from the graph that uncoded video sequences are highest scored with a score of 72.2. This seems intuitive as there are no coding or any other artifacts present in these video sequences. Apart from the generic trend, which is decrease in score with increasing QP level, there are some other noticeable 3D artifact scores which will be discussed below. One noticeable result is that vertical disparity with higher shift, i.e., 12 pixel shift is scored better than the lower vertical disparity sequence with a vertical shift of 6 pixels. One justification for this abnormal behavior can be characterized to the fact that the difference between 6 & 12 pixel shifts is not very perceptible.

Apart from that it can be concluded that color disturbances also have an effect on the 3D experience. Color disturbance 1 with change in red color component has a lesser effect on the 3D viewing experience than the luminance variation. Similarly negative parallax in which left and right view pair is switched also has a depreciating effect on the 3D video quality. Further graphical analysis on the results is done later.

3.3.4 Analysis of artifacts

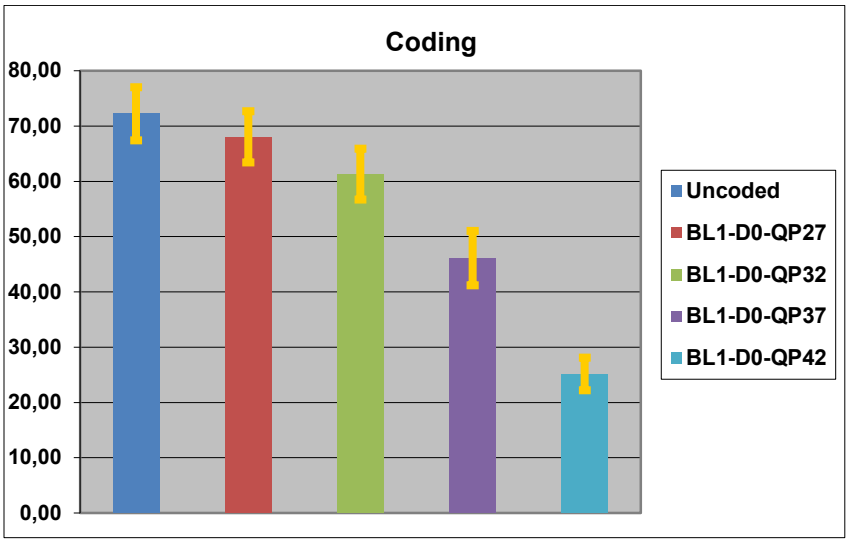


Figure 10: Coding artifacts

Figure 10 shows the plots for subjective scores based on the QP levels used for encoding the sequences. There is a conclusive trend in the scores for QP levels. The blurring/blockiness affect become more apparent as QP level is increased. The artifacts are not very noticeable when QP level is 32 but when QP level is changed to higher levels then the quality of the video in general decreases, which also deteriorates the 3D video experience. The answers to questionnaire that was conducted after the subjective test show that subjects had varying opinion about the coding artifacts. Six out of the nine subjects said that blurring/blockiness either had the same effect as in case of 2D or it had very less effect on the whole 3D experience.

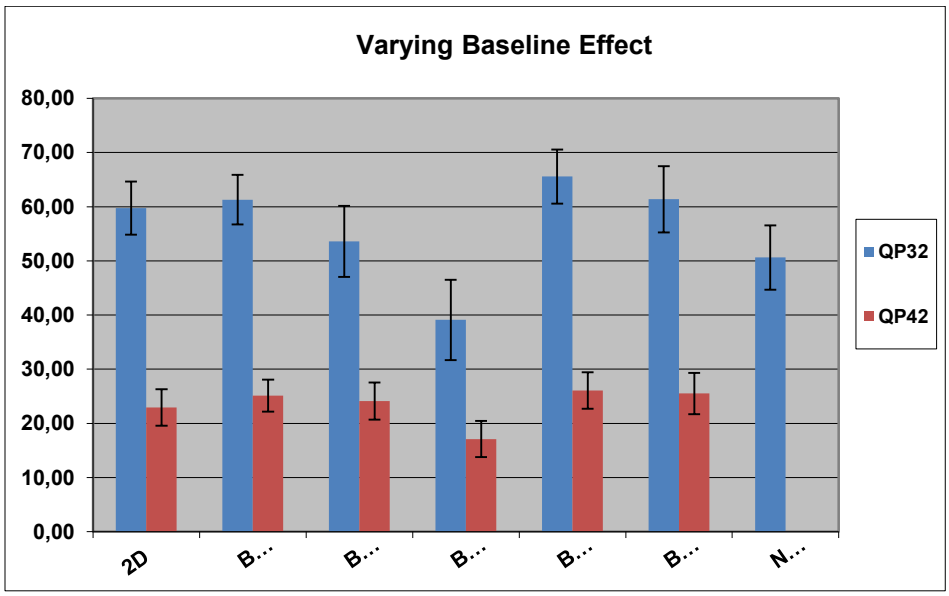


Figure 11: Varying baseline effect

From Figure 11 it is easy to conclude that for most of the subjects baseline 2 & 4 sequences (BL2_D0 and BL4_D0) with no disparity compensation were aggravating. The high CI for these sequences also

shows the uncertainty in the score, which in turn depends on the content as well as the user likeliness for depth. On the other hand compensated disparity sequences were rated better, which was the expected outcome. Some of the subjects rated 2D and baseline 1 sequences based on their quality (i.e., high) which is determined by the QP level used, while others rated them in terms of the 3D experience (i.e., low). Negative parallax sequences have a mean score of around 50 with fairly high CI. From the scoring trend per content, it is deducible that for some sequences the negative parallax effect was more noticeable than the others, e.g., Newspaper and Lovebird sequences were rated lowest, while GT_Fly was rated quite high. At QP 42 the scoring trend is almost the same as compared to QP 32. The overall scores are low due to lower quality of video sequences.

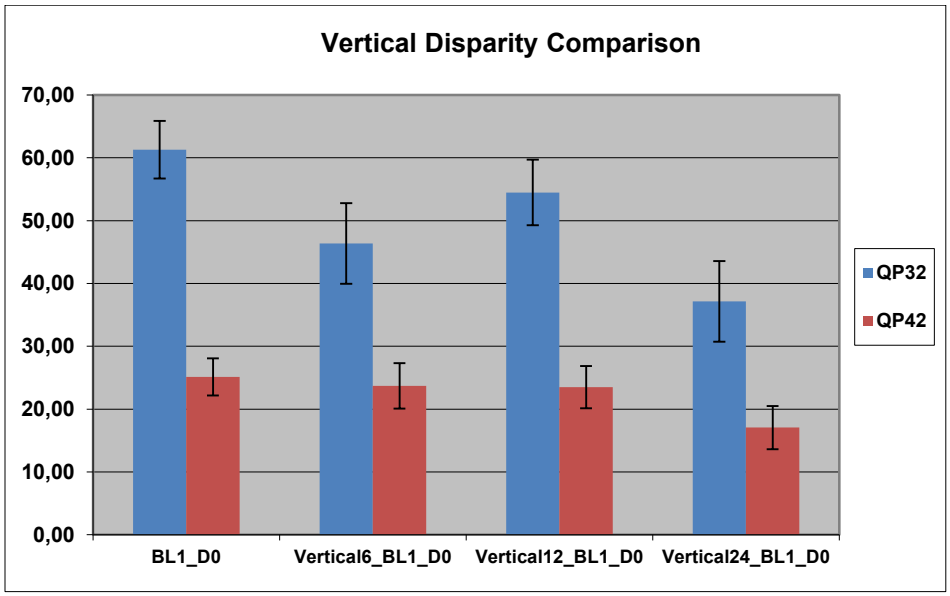


Figure 12: Vertical disparity effect

For vertical disparity three levels of vertical shifts were used in subjective test. The lowest vertical disparity introduced was a shift of 6 pixels, whereas 12 and 24 pixel shifts were the other levels of introduced vertical disparities. The trend shown in the plot for vertical disparity is not quite normal at QP 32, as 6 pixel vertically shifted sequences were rated lower on average than the higher vertical disparity sequences. After further analysis on the subjective scores, we see that average score for five out of six sequences follow the trend shown in graph for vertical disparity 6 and 12 at QP level 32. While for only one sequence which is Balloons, vertical disparity 6 was rated higher than vertical disparity 12. The only reason for this out of norm trend could be due to negligible difference in vertical disparities at these two levels. For higher vertical disparity 24, we can see that average score is less than lower vertically shifted sequences, which follows the norm.

For QP level 42 the scoring pattern is normal as opposed to QP level 32. Vertical disparity 6 & 12 are scored almost the same, which supports the argument of less noticeable difference in these two vertical disparity levels, but it is difficult to conclude this at this very high QP level and lesser number of subjects.

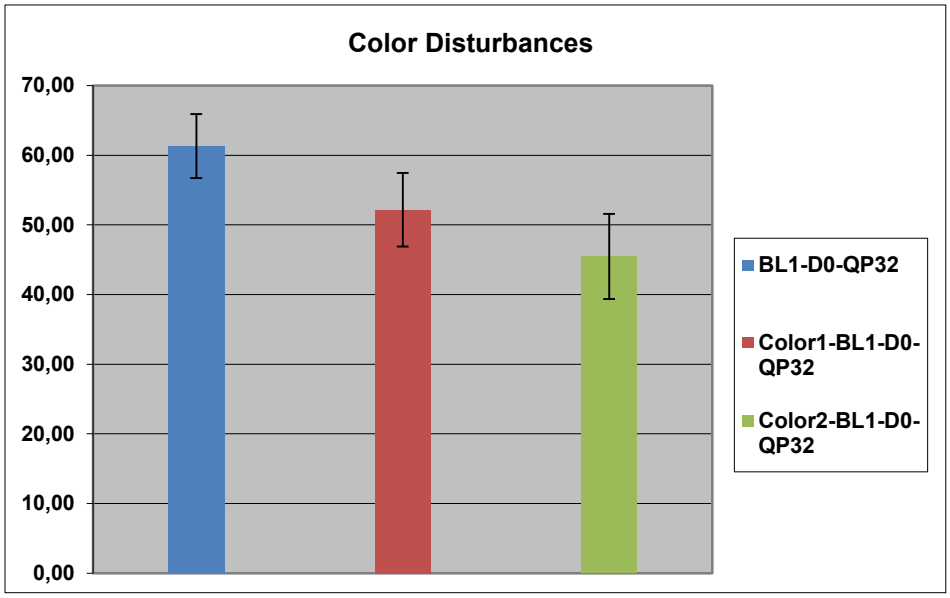


Figure 13: Color disturbance effect

It can be observed from Figure 13 that color disturbances do have an effect on the quality of 3D experience. Out of the two color disturbances the change in luminance (color2) was more annoying compared to scaling of red color component (color1). Analyzing in terms of content, we see that color disturbance 1 was least annoying in 'Balloons' video sequence as average score for this sequence is high for this color disturbance. Whereas, luminance effect was irritating in Newspaper sequence which is reflected by the score.

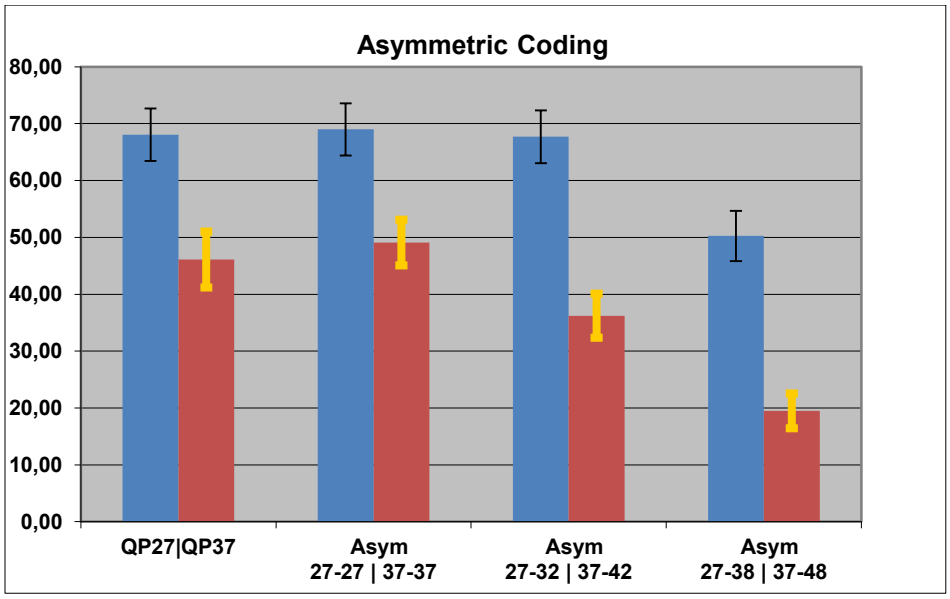


Figure 14: Asymmetric coding

In asymmetric coding the views were encoded using H.264/MPEG4-AVC rather than MVC (amendment to H.264/MPEG4-AVC) that uses prediction between the views. This is the reason that asymmetric sequences with QP 27 and 37 are scored slightly above the other corresponding MVC coded sequences as can be seen in Figure 14. Other than that, when there is a slight difference in QP levels for the left/right view pairs and overall QP is good, then the asymmetric coding effect is not

observable to a high degree. When the difference in QP level for left/right view increases, then overall 3D experience deteriorates and this is evident from the scores in the above plot. Among all the test subjects, the difference in left/right eye quality was irritating for only one subject according to the questionnaire.

3.3.5 Analysis of content

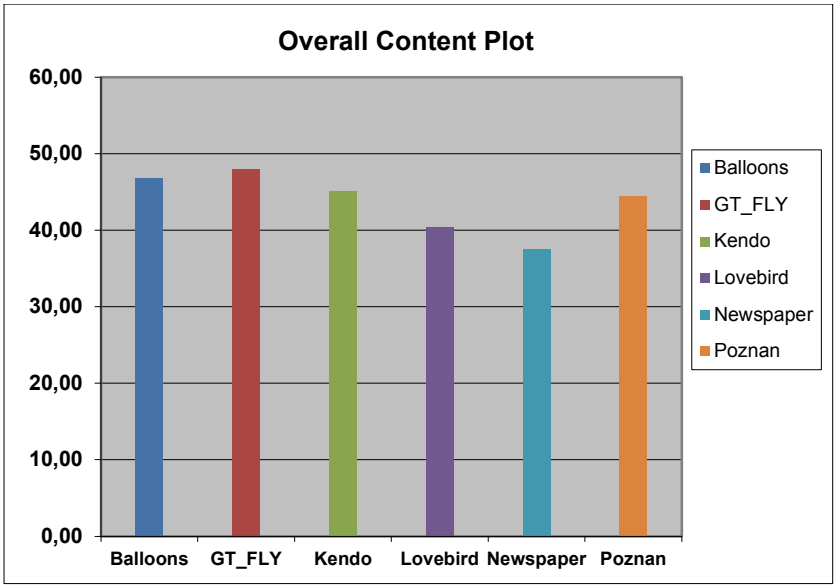


Figure 15: Average score per content for all test conditions

Figure 15 indicates that GT_Fly is the top scored sequence among the video content shown to the subjects in the test, whereas Newspaper and Lovebird are the two lowest scored sequences. If we consider the subjective test results, then it becomes apparent that Newspaper sequence was one of the worst for 3D experience. However in questionnaire, the content that was rated as one of the least favorable for 3D experience was GT_Fly, which is completely contradictory to the results of Figure 15. The reason for this could be mainly due to the fact that 3D artifacts were not easy distinguishable in this sequence, so it became difficult for subjects to grade the different qualities. Blurring/blockiness and aliasing effects were quite visible in this sequence but these anomalies were not as annoying as they were for some other sequences in terms of the 3D experience.

Newspaper and Lovebird sequences contain 3D occlusion artifacts at the frame borders, which agitates the 3D experience. For Newspaper sequence even small camera baseline distance creates high depth sensation which could be annoying for subjects. Apart from that the luminance disturbance is also relatively evident in Newspaper sequence. Due to these reasons these two sequences were the two lowest rated sequences.

Chapter 4

2D Implementation Methods

4.1 Introduction

In this chapter descriptions of several 2D methods that were implemented are provided. In a complete 3D objective model, the 2D quality plays an important role in the overall score. So, it is important that a good method is used for 2D evaluation. A comparison is done between the several 2D methods to show which method was preferred over the others.

4.2 PSNR

Peak Signal-to-Noise Ratio (PSNR) is a quality measure used to evaluate 2D images and it is based on finding the ratio between the maximum value of a pixel in image to noise. It is worth mentioning that PSNR is far from a complete description of the image quality. However, by combining PSNR with the visual observation one can somewhat determine the image quality. Usually, images with PSNR values higher than 30dB are considered as acceptable quality. The PSNR between the original image x and the reconstructed image y is defined as:

$$PSNR = 10\ log_{10}(\frac{Max_x^2}{MSE}),$$

where, Max is the dynamic range of the image (e.g., for an image x with B = 8 bits/pixel unsigned char, Max = 2B − 1 = 255), and Mean Squared Error (MSE) is defined as,

$$\mathrm{MSE} = \frac{1}{\mathrm{mn}}\sum_{\mathrm{i=0}}^{\mathrm{m-1}}\ \sum_{\mathrm{j=0}}^{\mathrm{n-1}}[\mathrm{x(i,j)\text{-}y(i,j)}]^2\ ,$$

x and y are grayscale images to be compared.

4.3 SSIM

The Structural Similarity (SSIM) metric measures the similarity between two images, for example, the original image and the compressed image. The SSIM measure is usually considered better than PSNR since it takes into account the image structure; however, computing SSIM is more complicated than PSNR. SSIM is designed to improve on traditional methods like PSNR or MSE which are known to be less consistent with human visual perception.

The SSIM metric is defined for two images x and y (or windowed image parts x and y) as

$$SSIM\ = \frac{(2\mu_x\mu_y + \ c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + \ c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where,

μ_x is the average of x, μ_y is the average of y, σ_x^2 is the variance of x, σ_y^2 is the variance of y, σ_{xy} is the covariance of x and y,

$c_1 = (k_1 L)^2, c_2 = (k_2 L)^2$ are two variables to stabilize the division with weak denominator,

L is the dynamic range of the pixel-values (for an image with B-bits/pixel, L = 2^B − 1),

$\mathrm{k}_1 = 0.01$ and $\mathrm{k}_2 = 0.03$ by default.

Mean SSIM (MSSIM) is used in this thesis. In MSSIM, SSIM is calculated for one block and block is shifted by one pixel at a time. For this reason computational complexity of this method is very high.

4.4 LU Factorization

MSSIM is a rather good objective quality measure for 2D images. The problem with MSSIM is that its computational cost is very high. LU Factorization [13] is a method which works on separating a block into lower and upper triangular matrix. The LU factorization can be expressed as

$$A = LU,$$

$$\det(A) = \det(L) \times \det(U) = \prod_{i=1}^{N} u_{i,}$$

Where, det is determinant of matrix A is the image block, L and U are the lower and upper triangular matrices. Lower triangular matrix contains the diagonal elements equal to one. After this the distortion metric is calculated which is defined as

$$D_j^u = \left\| u_j - \hat{u}_j \right\|_2$$

Here $\| \|_2$ is Euclidean norm, $u_j$and $\hat{u}_j$are the j-th reference and distorted image blocks, respectively, which consists of diagonal elements of matrices U and $\hat{U}$. Finally the objective quality metric MLU is computed as

$$MLU = \frac{1}{(W/N) \times (H/N)} \sum_{j=1}^{(W/N)\times(H/N)} D_j^U - D_{med}^U$$

where, W and H denote width and height of image respectively, N is the block size used and D_{med}^Urepresents the median value of D_j^U. The following block diagram represents the LU factorization method.

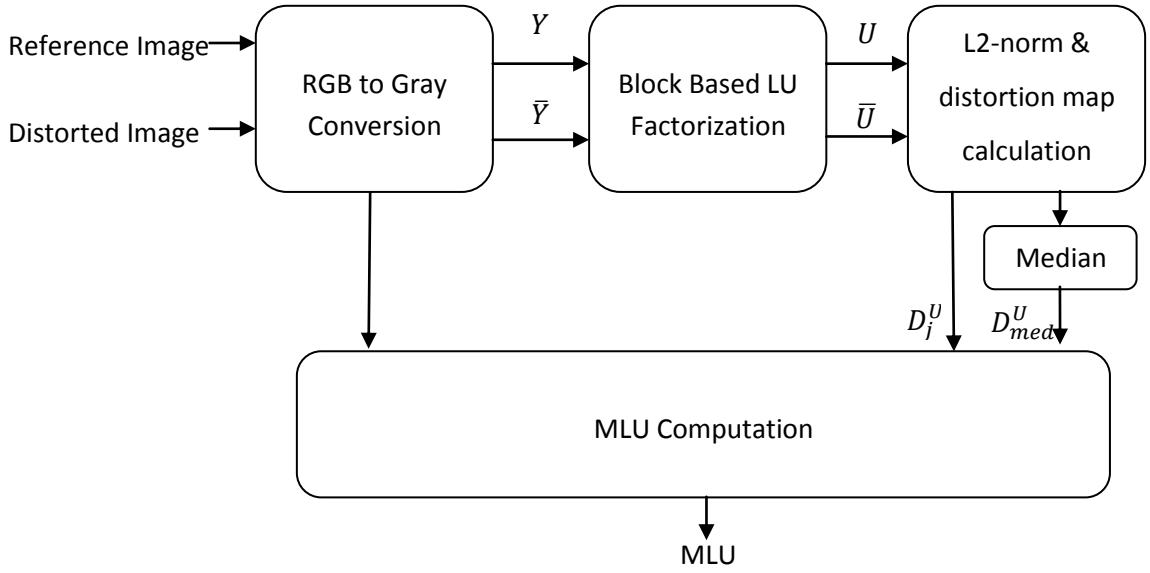


Figure 16: LU factorization model

4.5 Opticom's Video Quality Measure PEVQ

Perceptual Evaluation of Video Quality is a full reference 2D video quality evaluation model which is a standardized method described in ITU-T J.247 standard for resolutions up to 640x480. The main advantage of this method apart from accuracy is that it provides scores for different indicators as perceived by the human visual system. PEVQ model uses a number of criterions to evaluate the performance of video which include temporal, spatial, luminance and chrominance disturbances. After all the individual scores of separate stages in PEVQ have been computed, they are combined to form a final MOS.

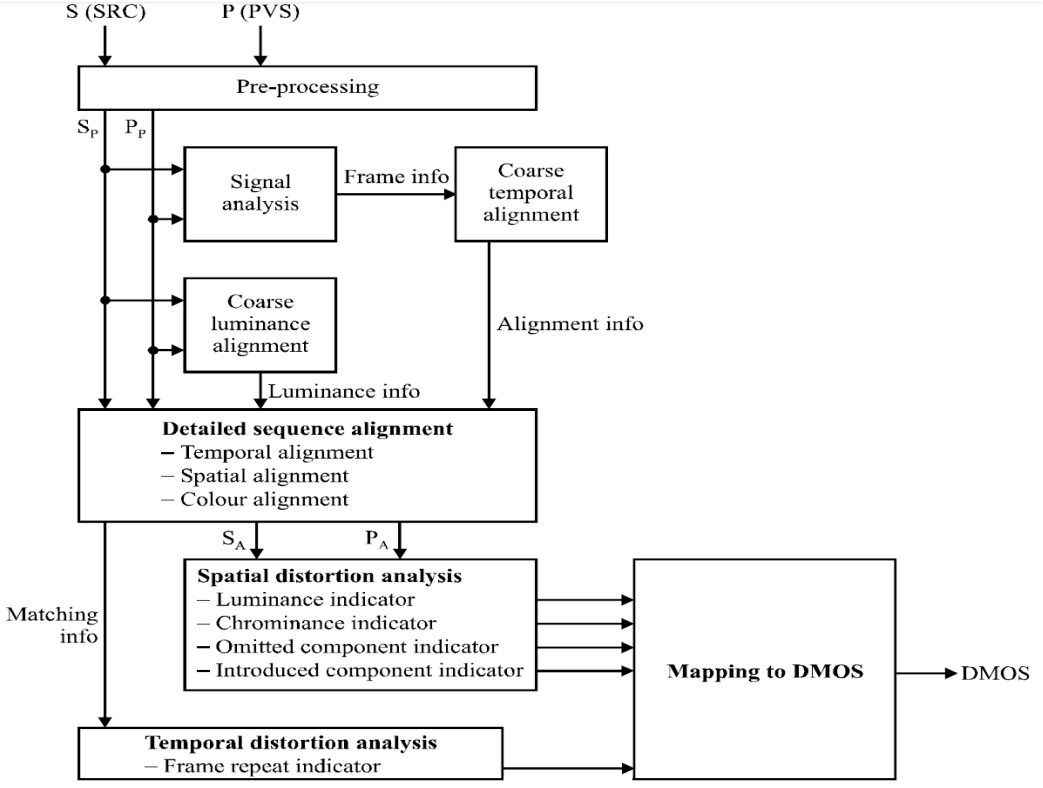


Figure 17: PEVQ model [10] Reprinted with permission from ITU

Figure 17 above shows the overview of the complete PEVQ model. As can be seen, the final estimated MOS is a combination of several individual scores. PEVQ output contains final estimated MOS along with individual scores as well. The output parameters for PEVQ are listed below in Table 4.

DSCORE	Jerkiness	Mean Temporal Dist	Temporal Activity Ref	Spatial Complexity Ref	Brightness Reference	Contrast Reference	Blockiness	PSNR Cb	Luminance Indicator	Chrominance Indicator
SCORE	Blur	Worst Temporal Dist	Temporal Activity Test	Spatial Complexity Test	Brightness Test	Contrast Test	PSNR Y	PSNR Cr	Correlation Indicator	

Table 4: Output parameters of PEVQ

4.6 Comparison of 2D Methods

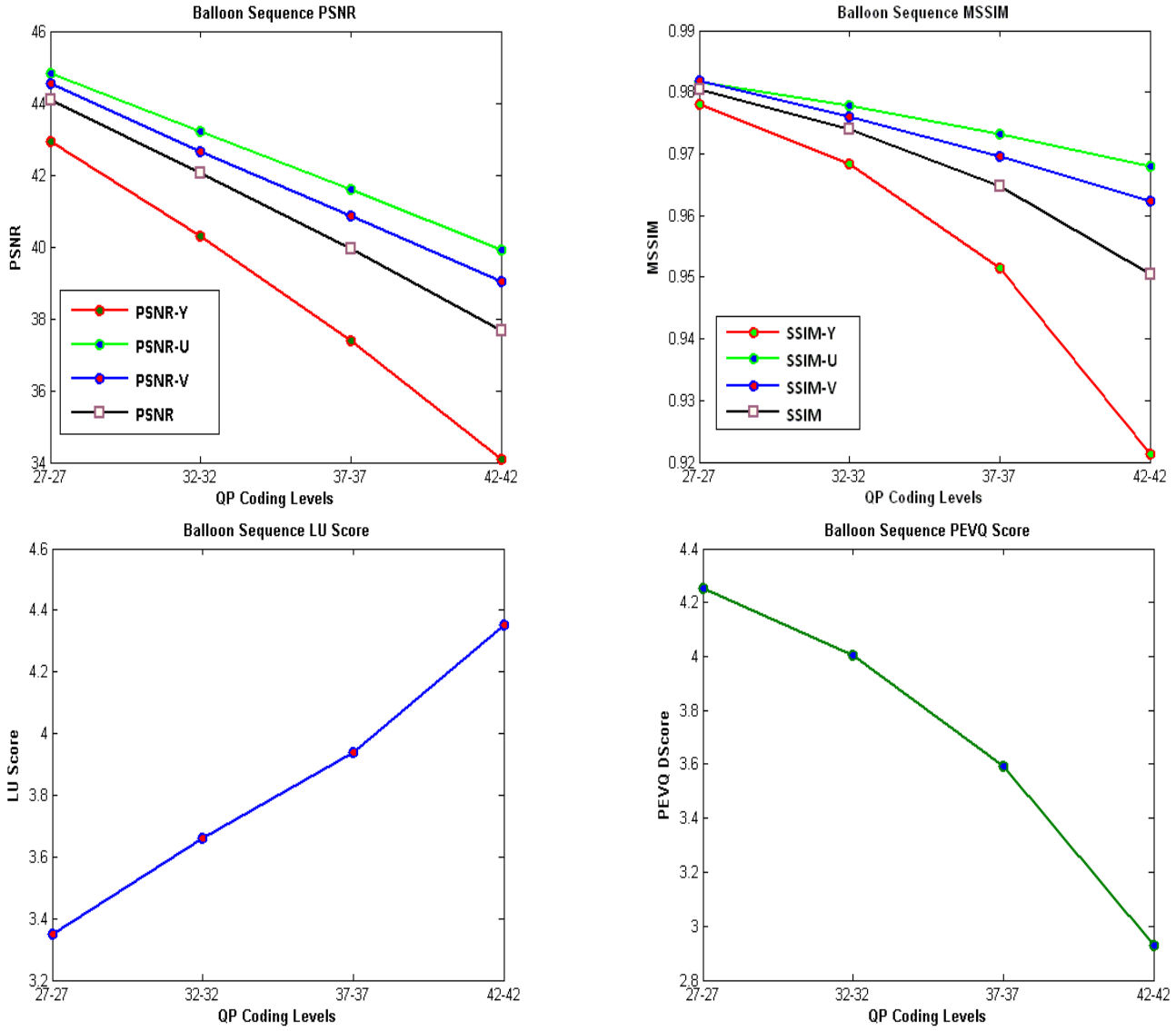


Figure 18: QP Coding levels vs. Objective scores for 2D Methods

Figure 18 above shows the score for left view balloon sequence for four different QP levels. PSNR and LU decomposition score vary almost linearly with increasing QP levels. For MSSIM and PEVQ the gradient changes between QP levels.

	Y		U		V		Mean	
	Color1	Color2	Color1	Color2	Color1	Color2	Color1	Color2
PSNR	25.30	14.81	29.70	30.16	20.62	29.04	25.21	24.67
MSSIM	0.95	0.87	0.97	0.95	0.92	0.94	0.95	0.89
LU	-	-			-		4.76	11.84
PEVQ	-	-			-		3.77	0.67

Table 5: Different 2D methods scores for color disturbance (Only left pair)

Table 5 shows the results of 2D methods for color disturbances. From Figure 13 it was concluded that Color2 disturbance was more annoying for subjects, so 2D methods comply with the subjective test results. For balloon sequence subjective test shows that the second color disturbance effect was more annoying than any other sequence. For this case PSNR shows only a slight difference in score for two color disturbances.

Further analysis was done to check which method best represents the quality of 3D video by checking the correlation between the subjective and objective 2D scores. To do fair comparison only subjective scores without change in baseline were used for this. Scores for this case are presented in Table 6.

	Baseline 1 sequences		**MOS**	**LU**	**MSSIM**	**PEVQ**	**PSNR**
Color 1 Disturbance	Balloon	QP-32	60,00	4,28	0,95	3,89	30,61
	GT_FLY	QP-32	52,78	6,40	0,91	3,51	30,07
	Kendo	QP-32	48,33	4,26	0,96	3,87	29,17
	Lovebird	QP-32	45,22	7,88	0,93	3,65	31,77
	Newspaper	QP-32	49,67	5,53	0,93	3,65	29,76
	Poznan	QP-32	57,00	4,58	0,90	3,29	29,33
Color 6 Disturbance	Balloon	QP-32	34,78	7,82	0,91	2,34	25,37
	GT_FLY	QP-32	64,44	6,16	0,89	3,53	25,81
	Kendo	QP-32	45,44	8,23	0,94	2,82	25,27
	Lovebird	QP-32	48,44	11,96	0,90	3,42	27,14
	Newspaper	QP-32	26,89	10,92	0,89	2,63	24,86
	Poznan	QP-32	52,78	7,76	0,87	2,91	24,37
Baseline 1 sequences	Balloon	QP-27	70,44	3,42	0,97	4,26	41,45
		QP-32	65,67	3,68	0,96	4,00	38,76
		QP-37	42,78	3,94	0,95	3,61	35,76
		QP-42	25,89	4,38	0,92	2,97	32,46
	GT_FLY	QP-27	69,11	6,74	0,94	3,79	38,74
		QP-32	58,22	6,98	0,92	3,51	36,39
		QP-37	48,11	7,08	0,89	3,05	34,05
		QP-42	22,11	7,17	0,84	2,30	31,44
	Kendo	QP-27	67,33	2,64	0,97	4,19	42,06
		QP-32	65,11	2,80	0,97	3,91	39,63
		QP-37	42,22	2,94	0,96	3,52	36,89
		QP-42	28,89	3,26	0,94	2,98	33,73
	Lovebird	QP-27	70,33	7,20	0,96	4,01	39,73
		QP-32	59,00	7,40	0,94	3,64	36,64

		QP-37	43,00	7,56	0,90	3,21	33,78
		QP-42	21,33	7,96	0,86	2,63	30,90
	Newspaper	QP-27	58,22	4,79	0,96	4,06	40,06
		QP-32	50,33	4,93	0,94	3,71	37,44
		QP-37	44,78	5,19	0,91	3,32	34,66
		QP-42	26,56	5,50	0,88	2,80	31,71
	Poznan	QP-27	72,78	4,65	0,93	3,68	38,37
		QP-32	69,44	4,66	0,90	3,31	36,26
		QP-37	55,67	4,83	0,87	2,86	34,19
		QP-42	26,00	5,12	0,83	2,23	31,76

Table 6: Average scores for baseline 1 sequences for several 2D methods

Linear mapping of 2D methods on to the subjective scores was done to find out which method better represents the quality of the video sequences. Figure 19 below shows the result of the linear mapping.

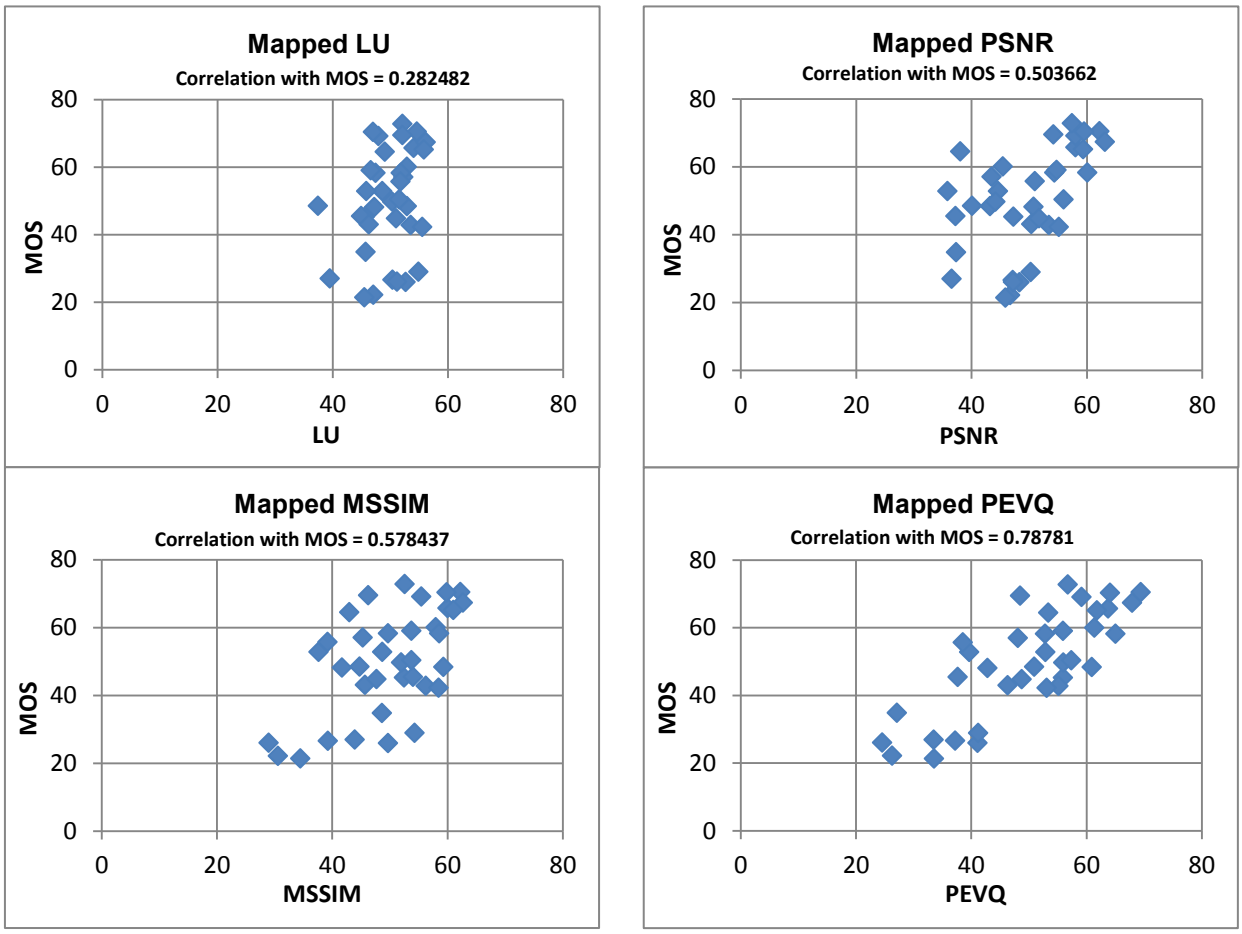


Figure 19: Linear mapping of 2D methods

It can be seen from Figure 19 that PEVQ is clearly the highest correlated method with MOS. MSSIM which is considered as one of the better 2D quality methods has a better correlation with the subjective scores relative to PSNR and LU. LU plot shows that it has a very low correlation with MOS. Due to highest accuracy and flexibility in available parameters for PEVQ; it was chosen as the method to be used in the 3D objective evaluation model.

Chapter 5

3D Parameters

5.1 Introduction

When evaluating the quality of 3D video, parameters related to 3D video along with 2D needs to be considered for a better estimate of the 3D experience. Several of the parameters for stereo video are novel and are specific for 3D. Among these aspects disparity or depth is one of the important criteria that have an impact on the quality of stereo video most. Another aspect of stereo video that affects the 3D experience is occluded regions. As left and right views are shifted with respect to each other, some of the objects near the sides might be missing in one of the views. This creates issues especially in multiview coding when a virtual view is synthesized from other views, then occluded regions pose a problem as predicting an object is a tedious task. This case is not considered in this thesis because synthesis artifacts were not the main focus.

The estimation of 3D parameters in this thesis was done using two main algorithms. An initial estimate of disparity was made using feature matching algorithm Scale Invariant Feature Transform (SIFT) and this initial estimate was then used as an input to disparity map algorithm for dense stereo correspondence. How these two algorithms work is provided in the sections to follow.

5.2 Initial disparity estimation process

The block diagram below shows the process of disparity estimation using SIFT matching algorithm.

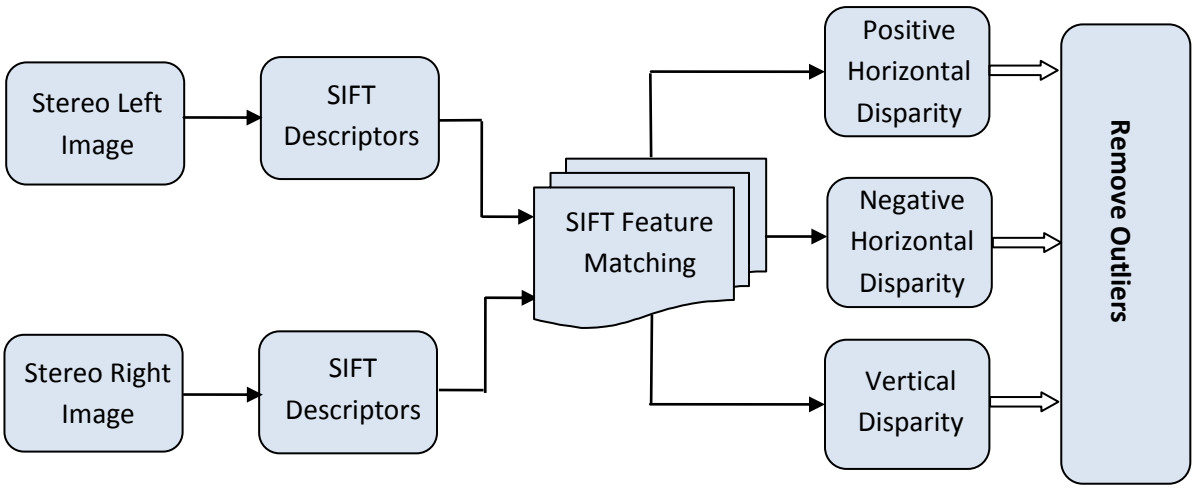


Figure 20: Disparity estimates using SIFT matching algorithm

First of all SIFT is applied to left/right image pairs one by one. Once the important features are detected for images, then a matching algorithm is applied on the detected features to find the common features across image pairs. After feature matching is complete, horizontal and vertical disparities are computed using the coordinates of the matched feature points. The last step of the process is to apply an outlier removal algorithm on calculated disparities to get rid of the false estimates. The next section describes the blocks of Figure 20 individually.

5.3 SIFT Algorithm

SIFT was developed in 1999 by David Lowe [14]. It is one of the most powerful algorithms for feature detection and matching. The objective of this algorithm is to find features in one image that can be used for matching features that appear in another image. The features extracted are invariant to rotation and image scaling, and they are also partially independent to viewpoint and illumination changes. The images are represented by SIFT features, so that these features can be reliably identified in other images if common features are present there as well. The major steps involved in SIFT are:

- ✓ Scale-space extrema detection
- ✓ Keypoint Localization
- ✓ Orientation Assignment
- ✓ Keypoint Descriptor

Details of these steps are provided in next subsections.

5.3.1 Scale-space extrema detection

The first stage of SIFT is to identify the main features in an image by applying a Gaussian filter to different scales of image. By cascading the different scales of an image, those features that are most stable over all scales are identified. The scale-space kernel used is Gaussian kernel which is the most suitable one. The variable scale Gaussian $\mathrm{G}(\mathrm{x}, \mathrm{y}, \sigma)$, is convolved with input image $\mathrm{I}(\mathrm{x}, \mathrm{y})$

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

Where, $G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$ and $*$ is the convolution. The implementation is done efficiently by blurring the input images and then finding the Difference of Gaussian (DOG) between two nearby scales separated by multiplicative factor k which is a constant. Mathematical representation for that is as follows

$$D(x, y, \sigma) = \big(G(x, y, k\sigma) - G(x, y, \sigma)\big) * I(x, y)$$
$$= L(x, y, k\sigma) - L(x, y, \sigma)$$

The DOG operator is approximately equivalent to Laplacian which is computationally expensive process. An example of DOG pyramids formed is as shown below.

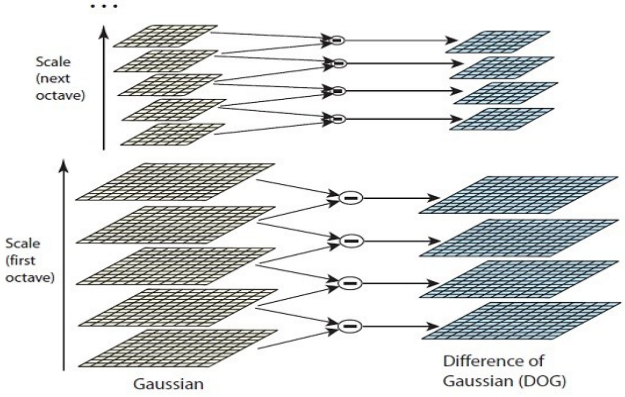


Figure 21: DOG pyramid

The left pyramids in above Figure contain the Gaussian blurred images and right pyramids contain the DOG of blurred images. In each octave the images are down sampled by a factor of 2 and then same process is repeated. After DOG pyramids are produced then local minima/maxima detection is applied to find the keypoints. The first step in to locate the maxima/minima in DOG images. For locating the maxima/minima each pixel in iterated and is compared with its neighboring pixels in same scale and also scales above and below.

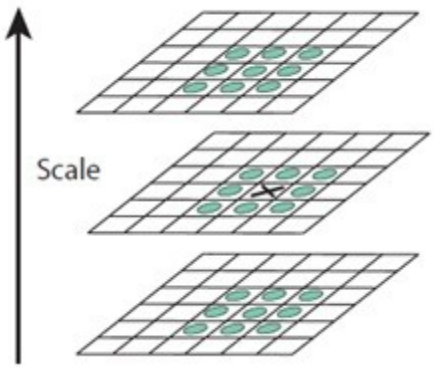


Figure 22: Pixel comparison in DOG pyramid to find maxima and minima

For a pixel marked with x, the pixels that will be checked in DOG are shown in green color. A point x is marked as keypoint depending on whether it is greatest or lowest among the compared pixels.

5.3.2 Keypoint Localization

In previous step the keypoints of the image were detected. The next step is to refine the keypoints, so that only those keypoints that are least susceptible to location, scale and curvature changes are preserved. This stage involves applying second order Taylor series expansion. The steps in this stage are as follows.

1). 2nd order Taylor expansion of D at (x, y, ρ).

$$D(\Delta\vec{x}) = D(\vec{x}) + \left(\frac{\partial D}{\partial \vec{x}}\right)^T . \Delta\vec{x} + \frac{1}{2}(\Delta\vec{x})^T . \frac{\partial^2 D}{\partial \vec{x}^2}(\Delta\vec{x})$$

2). Take derivatives with respect to $(\Delta\vec{x})$

$$\frac{\partial D}{\partial(\Delta\vec{x})} = (\frac{\partial D}{\partial \vec{x}})^T + \frac{\partial^2 D}{(\partial \vec{x}^2)(\Delta\vec{x})}$$

3). For extrema, put derivative equal to zero and solve for $\frac{\partial D}{\partial(\Delta\vec{x})}$=0.

$$(\Delta\vec{x}) = -(\frac{\partial^2 D}{\partial \vec{x}^2})^{-1} . \left(\frac{\partial D}{\partial \vec{x}}\right)^T$$

So, the keypoint is refined to new location$(x + \Delta x, y + \Delta y, \rho + \Delta\rho)$.

In localization stage the next step is to eliminate low contrast keypoints and edge responses. Low contrast points in the detected keypoints are eliminated by computing second order Taylor expansion at offset $\vec{x}$ calculated in previous step. If the value is less than 0.03, then the keypoint being considered is discarded otherwise it is retained with a new position containing a shift of $\vec{x}$ in original position and same scale.

The DOG has strong response along edges in those cases when it is determined poorly and thus is prone to very small amount of noise as well. Principal curvature is high across the edge for poorly defined peaks in DOG and small in perpendicular direction. The principal curvature can be computed by calculating the Hessian matrix corresponding to a specific keypoint.

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{yx} & D_{yy} \end{bmatrix}$$

eigenvalues of Hessian matrices are proportional to principal curvature. The ratio of two eigenvalues α (larger value) and β is equal to γ = α/β. If for a candidate keypoint the ratio or absolute difference in eigenvalues is higher than a certain threshold then it is rejected, as the eigenvalues are proportional to principal curvature. The higher the principal curvature, the more unstable the keypoint is.

5.3.3 Orientation Assignment

The next step assigns the orientation to each keypoint based on the local image properties. By assigning orientation, the SIFT algorithm makes sure that it is invariant to rotational changes. This step is integral part of SIFT but stereoscopic case does not have rotation between the images.

Firstly Gaussian smoothed images are selected based on the closest scale to the keypoint being considered. The gradient magnitude $\mathrm{m}(\mathrm{x},\mathrm{y})$ and orientation $\theta(\mathrm{x},\mathrm{y})$ are pre-computed using pixel differences for each image sample. For image sample point $L(x,y)$

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2}$$

$$\theta(x,y) = tan^{-1}\left(\frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}\right)$$

The orientations are found for all the pixels around the keypoint within a certain region. From gradient orientations an orientation histogram is formed. The histogram contains 36 bins divided into intervals of 10, corresponding to a 360 degree rotation. The highest peak corresponds to dominant local gradients in image. Any other peak not less than 20% below the highest peak is also assigned a keypoint. Thus a location maybe assigned multiple keypoints with different orientations.

5.3.4 Keypoint Descriptor

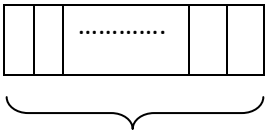
The final stage of SIFT is to assign vectors to each keypoint so that it is invariant to remaining variations like illumination and changes in viewpoint. For this we need to calculate gradient of 16x16 patches around each sample point in image $\mathrm{I}*\mathrm{G}_{\sigma_\mathrm{i}}$ centered at $(\mathrm{x_i},\mathrm{y_i})$. A weight is assigned to each image sample point using Gaussian weighting function with σ equal to half the width of the descriptor window. After this, gradient orientation relative to keypoint is computed using the following equation

$$\theta(x,y) = tan^{-1}\left[\frac{\partial(I * G_{\sigma_i})}{\partial y} \Big/ \frac{\partial(I * G_{\sigma_i})}{\partial x}\right]$$

Next the orientation histogram of each 4x4 pixel block is created. Histogram contains 8 bins each covering 45 degrees. The length of orientation defines the magnitude of that bin in the histogram. As 4x4 pixel block is used with 8 bins in histogram, the feature vector contains 4x4x8 = 128 features for each keypoint. For luminance invariance the keypoint feature vector is normalized to unit length and clamping is also performed to mitigate nonlinear effects.

The input to the SIFT is the image and output produced by SIFT is a set of keypoints and a set of feature vectors corresponding to the keypoints.

$P_i = (x_i, y_i, \sigma_i, \theta_i)$ Set of keypoints.

$f_i =$  Feature vector.

Set of features defining patch around keypoint P_i

Example Images with a set of keypoints identified are shown in Figure 23.

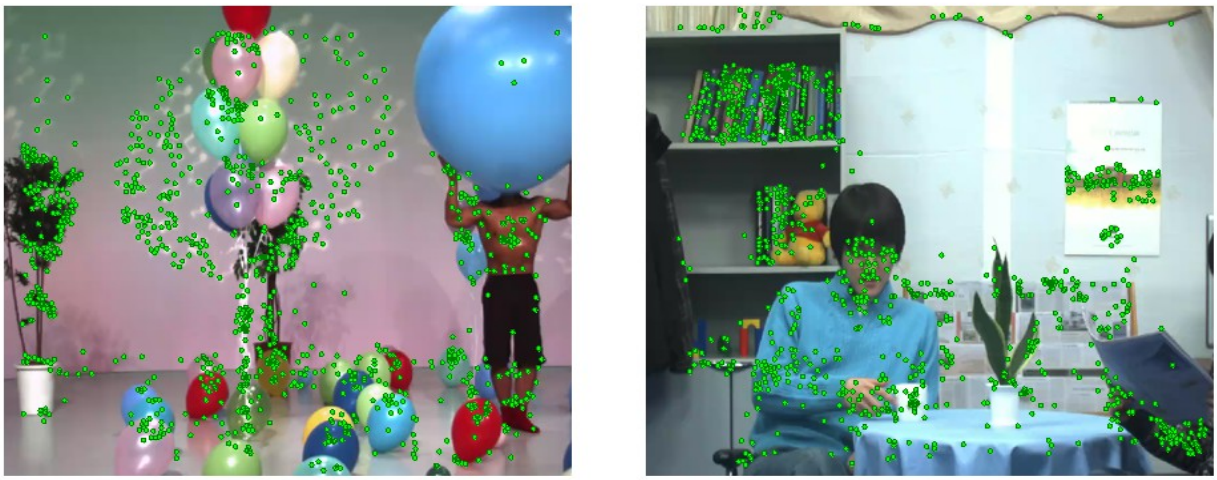


Figure 23: Keypoints for Balloon and Newspaper sequence using SIFT

5.3.5 SIFT Matching and Outlier Removal Algorithm

The matching algorithm used for SIFT feature matching is linear time histogram metric [15]. Open source implementation by Ofir Pele is used in this thesis which gives very accurate matching results.

Based on the SIFT distance metric between the SIFT descriptors of two images a SIFT distance ratio is calculated between descriptors. The higher the value of this ratio the better is the keypoint match. In this way the best keypoint matches between two images is found. Figure 24 shows two cascaded images showing some matched keypoints between them. The next step is to find the horizontal and

vertical disparities from the matched keypoints. The difference in x and y coordinates for the matched sample point correspond to the horizontal and vertical disparities. Two vectors containing the horizontal and vertical disparities are obtained where each vector point gives the disparity for a certain keypoint.

As can be seen from Figure 24 an outlier, i.e., incorrectly matched keypoint is also present. These outliers are then removed from disparity vectors using a simple and efficient outlier removal algorithm called Thompson Tau method. This method works by finding absolute difference of sample points from sample mean and standard deviation.

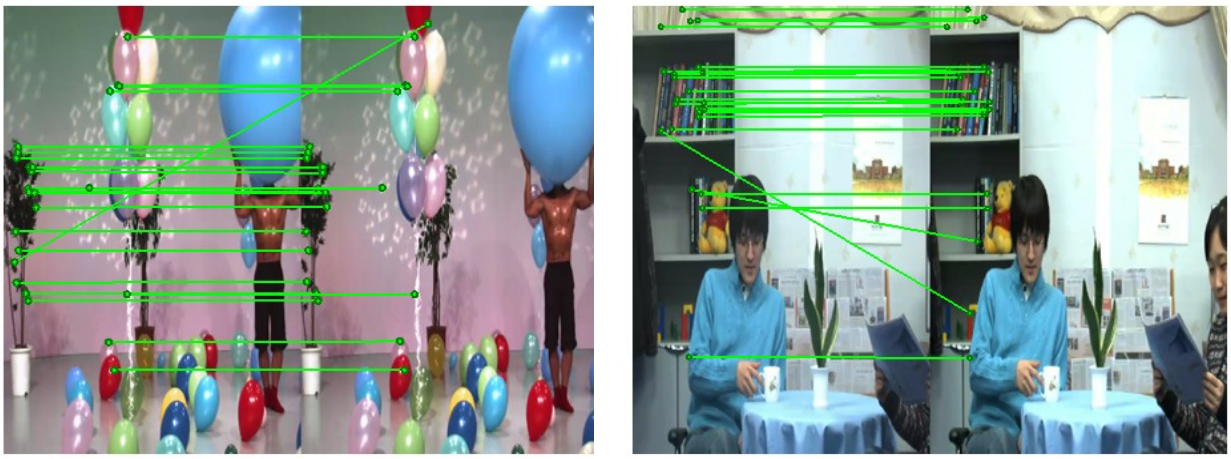


Figure 24: Keypoint matching using SIFT descriptor matching

5.4 Disparity Estimation Algorithm

Disparity map generation is one of the tedious tasks in 3D technology. Disparity map is used to find the depth information of objects that are present in image frames. In some applications that use depth information, a rough estimate of depth is what is needed but in the field of 3D a good quality depth map is often required. For high resolution images/videos the process of disparity estimation is very slow and thus makes it difficult to be used in real time. A lot of research has been done on finding a suitable algorithm which could produce desirable results. There are different techniques used for disparity estimation from local methods to global methods. In this thesis several disparity map algorithms from very basic local method to methods that take into account different aspects of 3D were implemented. The disparity estimation problem is a vast field and is not the focus of thesis, so those methods which were less complex were implemented. First a brief description about disparity is provided and then an overview of algorithms implemented is provided.

5.4.1 What is Disparity?

Disparity is the amount of pixel displacement between the two left and right image pairs.

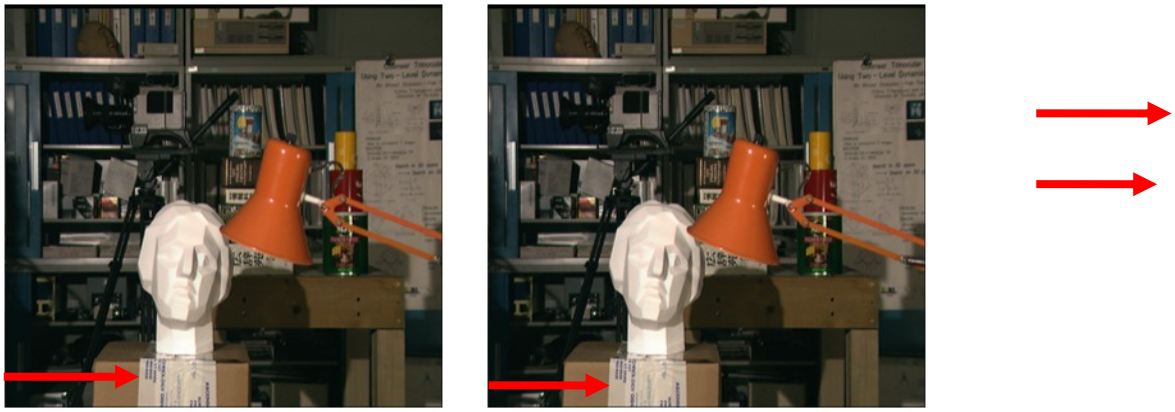


Figure 25: Disparity between left and right image pairs

Figure 25 shows the left and right image pairs for a 3D image. The two image pairs are shifted in horizontal direction with respect to each other. The amount of shift between the two images can be seen by the magnitude of two arrows. This translation in horizontal direction when expressed in terms of difference in pixel gray values is called horizontal disparity. When two left/right image pairs overlap each other then this shift in horizontal direction is seen as depth and thus enabling the viewer to see a 3D effect. The left and right image pairs overlapped on top of each other are shown in Figure 26 along with its ground truth disparity map which is the ideal disparity map.

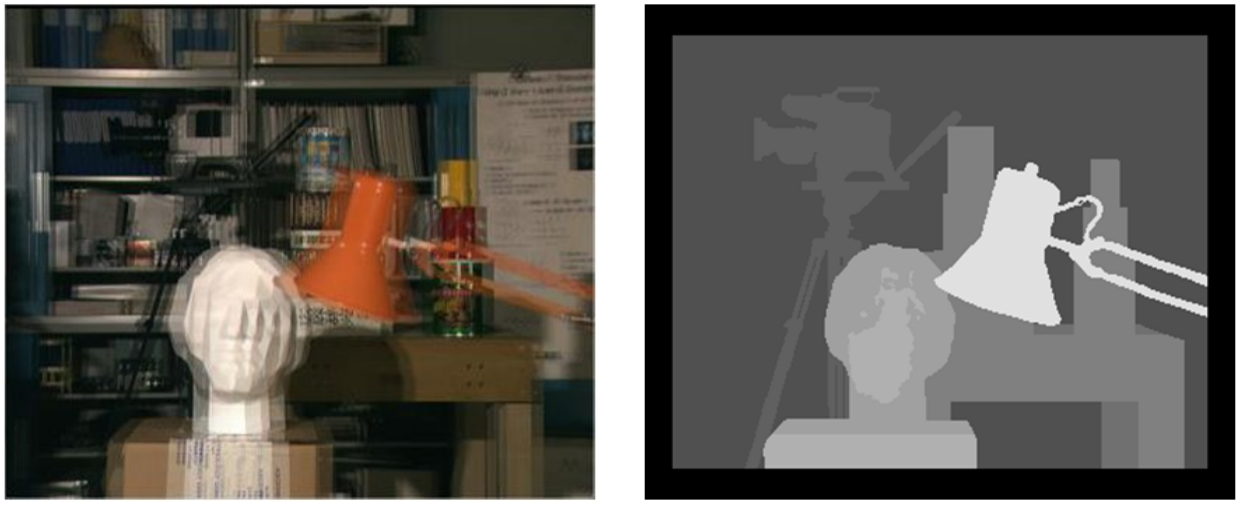


Figure 26: Overlapped left/right image pair with its ground truth disparity map

5.4.2 Local Window Matching Method

The most basic methods for finding the disparity map is to take a window from one image pair and slide it over the other image pair. The window with least amount of difference in color values is chosen as the correct match. All pixels within that window have the same disparity. The example of how window sliding method works is shown in Figure 27.

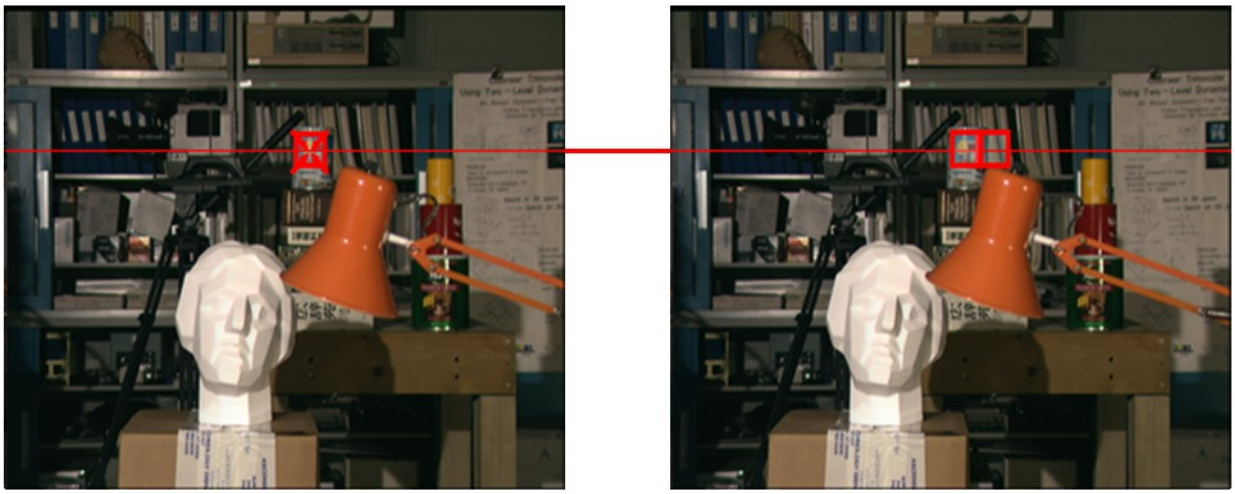


Figure 27: Local window matching algorithm

The mathematical form for finding disparity in a window using this method [16] is

$$d_p = \underset{0 \le d \le d_{max}}{\operatorname{argmin}} \sum_{q \in W_p} C(q, q-d)$$

where,

$\mathrm{d_p}$ Is the disparity for pixel p,

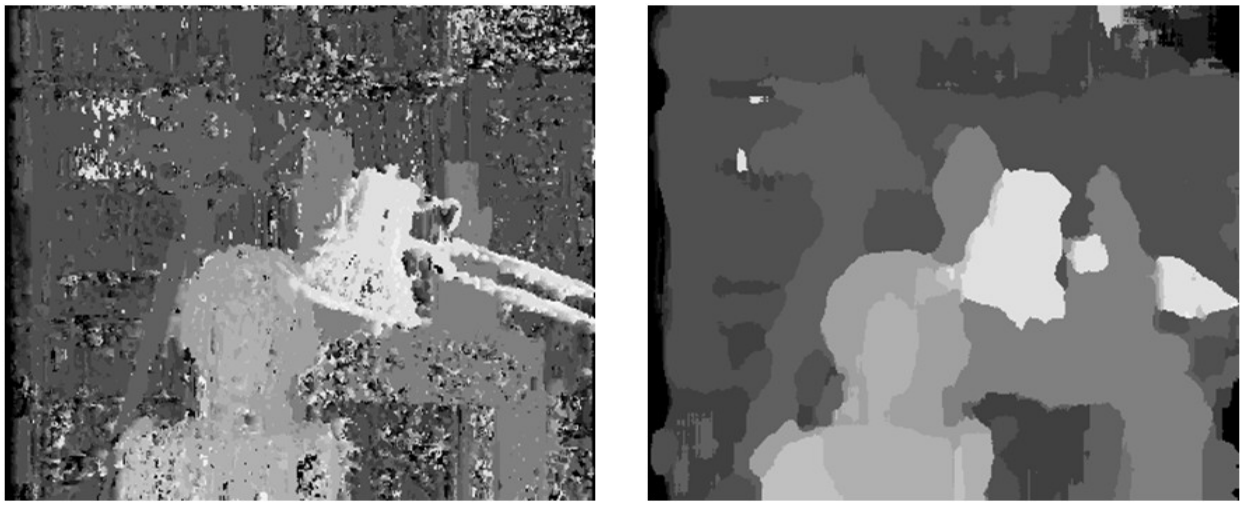
argmin returns the minimizing argument of the function,

$\mathrm{d_{max}}$ is the maximum disparity value possible between two images.

$\mathrm{W_p}$ contains the pixels within the window being compared.

$\mathrm{C(q, q\text{-}d)}$ computes the rgb color difference between pixel q in left image to a shifted pixel $q-d$ in the right image.

The results from this algorithm are shown in Figure 28 for two different window sizes.



3x3 **21x21**

Figure 28: Disparity maps using two window sizes

It can be seen from above figure that the results of this method are not suitable. This is due to the fact that constant size windows don't take into account small changes in textures within a window in neither horizontal nor vertical direction. For this method to work there should be a constant change in texture which is non-repetitive. When there is a repetitive change then a point in one image can match the wrong point due to similarity in different points.

5.4.3 Adaptive Window Method

In adaptive window method unlike local fixed window methods, the window weights are calculated based on the surrounding pixels. The weights are assigned considering only those pixels which have the same disparity in the current window. The mathematical form of the adaptive window method becomes

$$A_{p,d} = \sum_{q \in Wp} w(p,q).c(q, q-d)$$

The weight function decides whether the pixels p and q lie on the same disparity or not. The weight computation done in [17] is based on color dissimilarity as well as distance between the pixels. Depending on these two parameters the function $w(p,q)$ is calculated using the following equation

$$w(p,q) = \exp(-(\frac{\Delta c_{p,q}}{\gamma_c} + \frac{\Delta g_{p,q}}{\gamma_g}))$$

where,

$\Delta c_{\boldsymbol{p,q}}$ is the color difference, $\Delta c_{\boldsymbol{p,q}} = \sqrt{(L_p - L_q)^2 + (a_p - a_q)^2 + (b_p - b_q)^2}$

L, a and b represent the colors in CIELab color space.

$\Delta g_{\boldsymbol{p,q}}$ is the spatial distance such as Euclidean distance.

γ_c and γ_g are constants which are calculated based on gestalt principal of proximity and similarity.

The final disparity for pixel p is calculated using $d_p = \text{argmin}_{d \in S_d} E(p, \bar{p}_d)$. Here $E(p, \bar{p}_d)$ is dissimilarity measure between pixel p in left image and disparity d shifted pixel $\bar{p}_d$ in right image calculated as follows.

$$E(p, \bar{p}_d) = \frac{\sum_{q \in N_p, \bar{q}_d \in N_{p_d}} w(p,q)\, w(\bar{p}_d, \bar{q}_d) e_o(q, \bar{q}_d)}{\sum_{q \in N_p, \bar{q}_d \in N_{p_d}} w(p,q)\, w(\bar{p}_d, \bar{q}_d)}$$

p, q are pixels from left image and $\bar{p}_d$, $\bar{q}_d$ are from right image. Here q represents all the other pixels around p in the window being considered. $e_o(q, \bar{q}_d)$ is summation of absolute color difference between RGB components of a pixel.

The results of this method on the Tsukuba image is shown below along with disparity map for comparison.

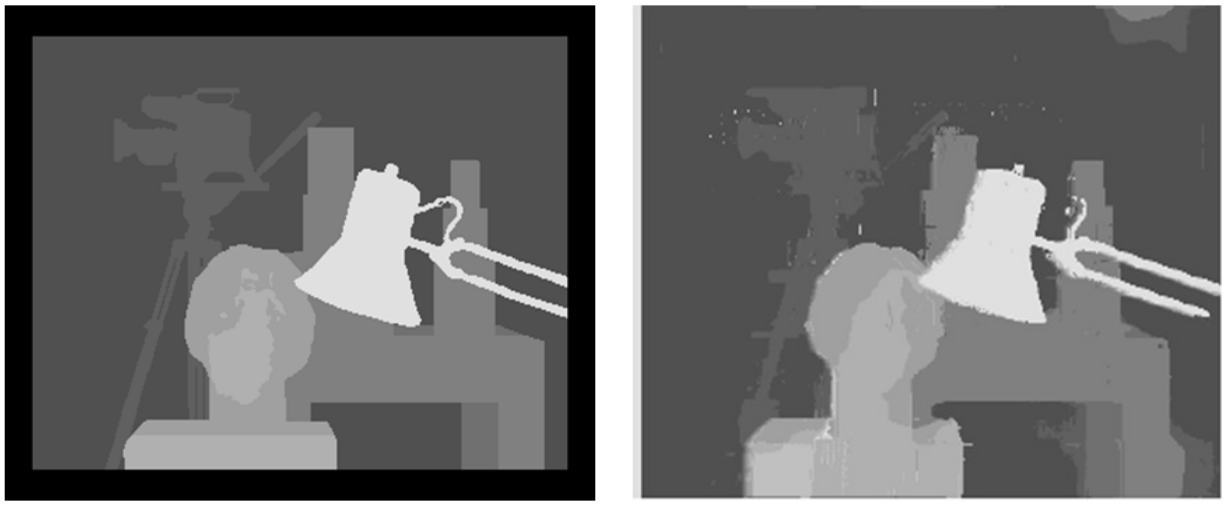


Figure 29: Left: Ground truth disparity map, Right: Disparity map using adaptive window method

The results of this method were very promising on the smaller resolution images but the results on the video sequences used in this thesis were not near optimal. Another reason for not using this method was the computation complexity for calculating weights for each window which increased the amount of time considerably on higher resolution images.

One other similar method implemented that relied on the same principle of assigning weights to square windows using geodesic distances is described in [18]. The geodesic distance $D(\rho, c)$ between a center pixel c and a pixel ρ in support window is defined as the shortest path connecting them in terms of their color volume. $D(\rho, c) = \min_{P \epsilon P_{p,c}} d(P)$, where $P \epsilon P_{p,c}$ is the set of all possible paths between ρ and c. The cost of a path is calculated using sum of Euclidean distance between the RGB color components for all pixels in the path. In this way weights are computed and matching costs are derived. Finally the lowest matching cost disparity among the possible disparities is chosen. The result of this method is shown in Figure below.

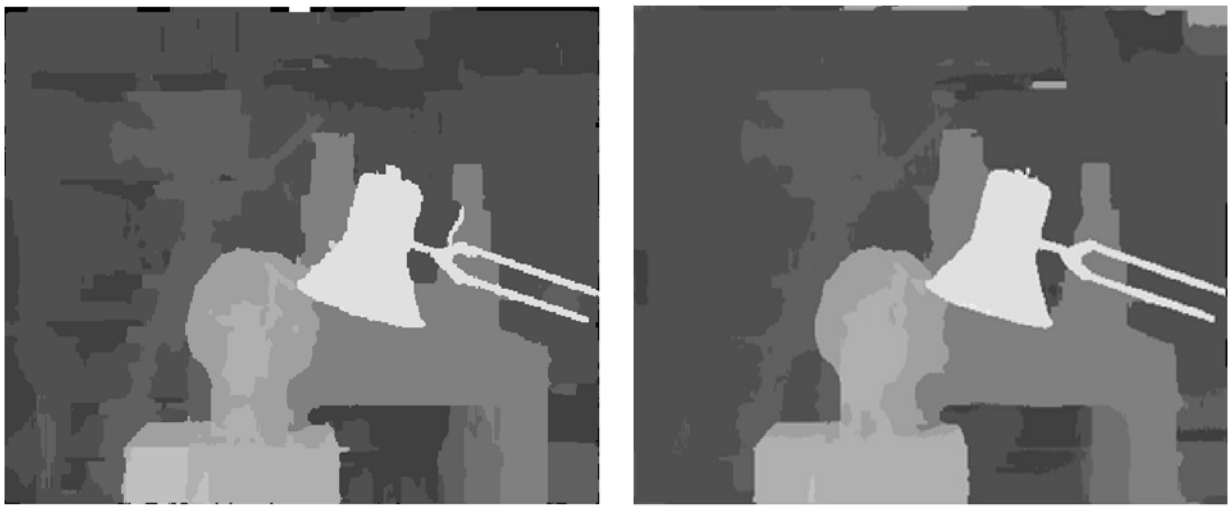


Figure 30: Left: Disparity Map using geodesic method, Right: Disparity Map using Fast Cost-Volume

Another method tested for disparity map generation which uses the principle of fast cost volume filtering is described in [19]. This method relies on minimizing the energy cost function and it takes into account occlusions as well. The implementation provided by the authors of this paper was used to compare the results. The result of this method on Tsukuba image pairs is shown in Figure 30.

5.4.4 Disparity Map Generation Using Shape and Stereo Correspondence Method

The disparity map method finally used in this thesis is based on paper shape and stereo correspondence problem [20]. This method produced the best results on the test video sequences among the methods implemented. The paper addresses a disparity map method which is piecewise continuous with minimal discontinuities. The basic criterion behind this is similar to matching regions of left image pair with right image pair. This method takes into account the horizontally and vertically slanted surfaces and deals with its issues.

In complex images there are lots of slanted surfaces and the local methods explained before don't take care of the slant problems. Suppose there is a surface that is horizontally slanted as shown in Figure 31 with line AB. C1 and C2 are two cameras with baseline distance t between them. The projection of two points A and B on to the camera is shown by lines L1 and L2. Clearly we can see that these two lines are not equal. Mathematically,

$$L1 = \frac{X_B}{Z_B} - \frac{X_A}{Z_A}$$

$$L2 = \frac{(X_B - t)}{Z_B} - \frac{(X_A - t)}{Z_A}$$

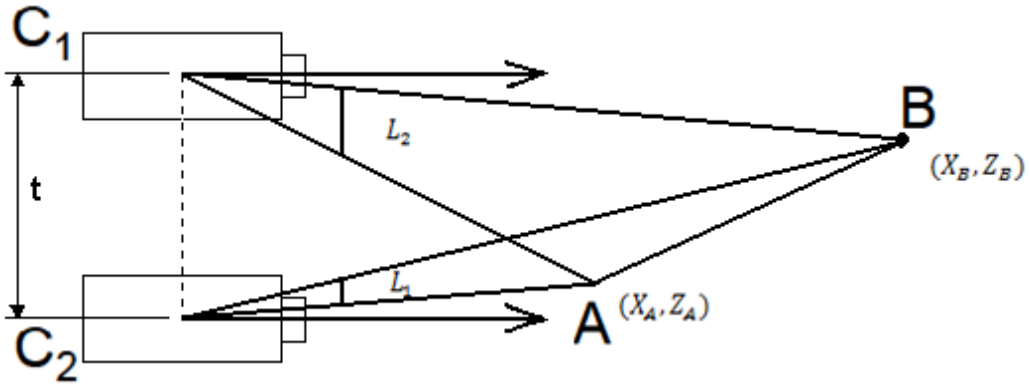


Figure 31: Slanted surface problem

The problem for horizontally slanted surfaces is that the objects will always project onto the two stereo cameras with different lengths and thus N number of pixels in one image can correspond to M number of pixels in other image. To counter this problem the algorithm should allow unequal number of pixel matching between image pairs.

For horizontally slanted surfaces pixel intensity differences cannot be found in similar manner as for frontal-parallel lines. A very useful method for matching pixel intensities is provided by Birchfield and Tomasi and it is used by many modern stereo correspondence algorithms. This method works very well for non-slanted surfaces but does not work for slanted surfaces. For applying this method it is

mandatory to stretch left image pair by a factor of m, where $x_R = mx_L + d, x_R \; and \; x_L$ are left and right image pair pixels and d is the disparity. After this is done the Birchfield-Tomasi method can be applied in the following way. For two scanlines $\mathrm{I_L(x)}$ and $\mathrm{I_R(x)}$, $\mathrm{I_L}\left(\mathrm{x_L}\text{-}\frac{1}{2}\right)$, $\mathrm{I_L}\left(\mathrm{x_L}+\frac{1}{2}\right)$, $\mathrm{I_R}\left(\mathrm{x_R}\text{-}\frac{1}{2}\right)$, $\mathrm{I_R}\left(\mathrm{x_R}+\frac{1}{2}\right)$ are found by using interpolation. Then the following values are found from above values

$$I_L^{min} = \min\left(I_L\left(x_L - \frac{1}{2}\right), I_L(x_L), I_L\left(x_L + \frac{1}{2}\right)\right)$$

$$I_L^{max} = \max\left(I_L\left(x_L - \frac{1}{2}\right), I_L(x_L), I_L\left(x_L + \frac{1}{2}\right)\right)$$

$$I_R^{min} = \min\left(I_R\left(x_R - \frac{1}{2}\right), I_R(x_R), I_R\left(x_R + \frac{1}{2}\right)\right)$$

$$I_R^{max} = \max\left(I_R\left(x_R - \frac{1}{2}\right), I_R(x_R), I_R\left(x_R + \frac{1}{2}\right)\right)$$

Then,

$$d_L = \max(0, I_L(x_L) - I_R^{max}, I_R^{min} - I_L(x_L))$$

$$d_R = \max(0, I_R(x_R) - I_L^{max}, I_L^{min} - I_R(x_R))$$

The final value of absolute intensity difference is equal to,

$$d = \min(d_L, d_R)$$

The problem that occurs due to horizontal slant and a simple way to tackle with it is described above.

The algorithm assigns horizontal disparities to scan lines rather than individual pixels. For a point $\mathrm{x_L}$ on the left scanline in the left image pair the corresponding pixel on scanline in the right image is

$$x_R = m_L(x_L).x_L + d_L(x_L)$$

And for right image pair,

$$x_L = m_R(x_R).x_R + d_R(x_R)$$

Here $\mathrm{m_L}, \mathrm{m_R}$ are the stretch factors and is the amount by which each image pair is resampled before using Birchfield-Tomasi method. The disparities are then computed using the following equations

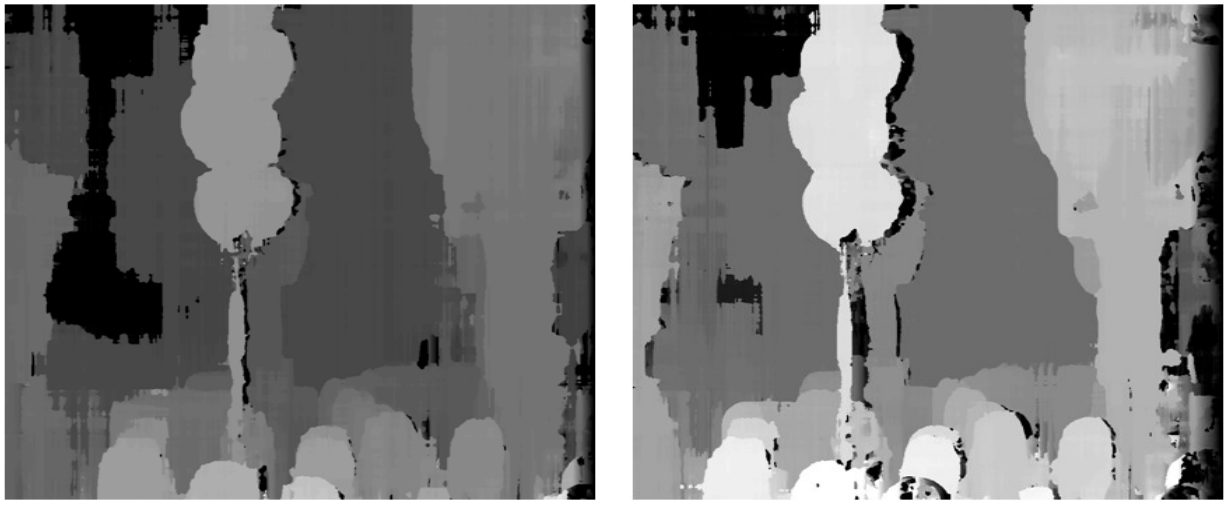
$$\Delta_L(x_L) = x_R - x_L = (m_L(x_L) - 1).x_L + d_L(x_L)$$

$$\Delta_R(x_R) = x_L - x_R = (m_R(x_R) - 1).x_R + d_R(x_R)$$

For finding occlusions a modified uniqueness constraint is also specified for one-to-one correspondence between left and right image scanlines. This uniqueness constraint is used to find

occluded pixels by Left-Right consistency check between the left and right disparity maps. The correspondence between image scanlines as already mentioned can be of N to M pixels where N≠M. N pixels may correspond to M pixels as long as it is unique. This is equivalent to uniqueness in scene space rather than image space. So, the new uniqueness constraint implies that consistency check be applied on regions rather than individual pixels to ensure that occluded pixels are found correctly.

Figure 32: Left: Ground truth Right: Shape and correspondence method disparity maps



Baseline 1 **Baseline 2**

Figure 33: Balloon sequence disparity maps using shape and correspondence stereo algorithm

Figure 32 and Figure 33 show the results of this stereo correspondence algorithm on two different images. The results of disparity maps generated for high resolution video sequences from this method were quite good compared to the other stereo correspondence methods. Further comparison is given in next section.

5.5 Comparison of disparity map methods

The methods tested for disparity map generation in this thesis were compared using the Middlebury stereo correspondence evaluation page [21]. The cost volume filtering method described in [19] is one of the best local methods on stereo evaluation page. A snapshot of the Middlebury stereo evaluation page is shown below.

Algorithm	Avg. Rank	Tsukuba ground truth nonocc	Tsukuba ground truth all	Tsukuba ground truth disc	Venus ground truth nonocc	Venus ground truth all	Venus ground truth disc	Teddy ground truth nonocc	Teddy ground truth all	Teddy ground truth disc	Cones ground truth nonocc	Cones ground truth all	Cones ground truth disc	Average percent of bad pixels (explanation)
ADCensus [94]	6.2	1.07 13	1.48 10	5.73 15	0.09 2	0.25 7	1.15 2	4.10 5	6.22 3	10.9 5	2.42 3	7.25 5	6.95 4	3.97
AdaptingBP [17]	7.8	1.11 16	1.37 6	5.79 16	0.10 3	0.21 4	1.44 4	4.22 7	7.06 6	11.8 8	2.48 5	7.92 10	7.32 8	4.23
CoopRegion [41]	7.8	0.87 3	**1.16** 1	4.61 2	0.11 4	0.21 3	1.54 6	5.16 15	8.31 11	13.0 12	2.79 14	7.18 4	8.01 18	4.41
DoubleBP [35]	10.4	0.88 5	1.29 3	4.76 5	0.13 7	0.45 18	1.87 11	3.53 4	8.30 10	9.63 3	2.90 18	8.78 26	7.79 15	4.19
RDP [102]	10.8	0.97 8	1.39 7	5.00 7	0.21 22	0.38 15	1.89 12	4.84 9	9.94 17	12.6 10	2.53 6	7.69 7	7.38 9	4.57
OutlierConf [42]	11.3	0.88 4	1.43 9	4.74 4	0.18 15	0.26 9	2.40 19	5.01 11	9.12 14	12.8 11	2.78 13	8.57 21	6.99 5	4.60
SubPixDoubleBP [30]	15.5	1.24 24	1.76 26	5.98 20	0.12 6	0.46 20	1.74 9	3.45 3	8.38 12	10.0 4	2.93 20	8.73 25	7.91 17	4.39
SurfaceStereo [79]	15.9	1.28 29	1.65 18	6.78 34	0.19 17	0.28 10	2.61 26	3.12 2	**5.10** 1	**8.65** 1	2.89 17	7.95 12	8.26 24	4.06
WarpMat [55]	17.9	1.16 17	1.35 5	6.04 21	0.18 16	0.24 6	2.44 21	5.02 12	9.30 15	13.0 14	3.49 31	8.47 20	9.01 37	4.98
ObjectStereo [98]	18.8	1.22 23	1.62 14	6.36 26	0.59 50	0.69 35	4.61 52	4.13 6	7.59 7	11.2 7	**2.20** 1	6.99 3	**6.36** 1	4.46
PatchMatch [112]	21.8	2.09 62	2.33 48	9.31 59	0.21 21	0.39 16	2.62 27	**2.99** 1	8.16 8	9.62 2	2.47 4	7.80 8	7.11 6	4.59
Undr+OvrSeg [48]	23.5	1.89 54	2.22 47	7.22 43	0.11 5	0.22 5	1.34 3	6.51 27	9.98 18	16.4 32	2.92 19	8.00 13	7.90 16	5.39
GC+SegmBorder [57]	23.9	1.47 42	1.82 28	7.86 51	0.19 18	0.31 11	2.44 21	4.25 8	5.55 2	10.9 6	4.99 69	**5.78** 1	8.66 30	4.52
InfoPermeable [109]	23.9	1.06 12	1.53 11	5.64 12	0.32 31	0.88 45	4.15 46	5.60 17	13.0 40	14.5 16	2.65 10	9.16 33	7.69 14	5.51
CostFilter [95]	24.8	1.51 44	1.85 33	7.61 48	0.20 20	0.39 17	2.42 20	6.16 22	11.8 29	16.0 25	2.71 11	8.24 16	7.66 13	5.55
GlobalGCP [104]	24.9	0.87 2	2.54 51	4.69 3	0.16 13	0.53 23	2.22 17	6.44 24	11.5 26	16.2 28	3.59 34	9.49 42	8.95 36	5.60
AdaptOvrSegBP [33]	26.1	1.69 47	2.04 42	5.64 12	0.14 9	0.20 2	1.47 5	7.04 42	11.1 21	16.4 34	3.60 36	8.96 30	8.84 33	5.59
FeatureGC [107]	26.2	1.08 14	1.59 13	5.82 18	**0.08** 1	**0.16** 1	**1.11** 1	7.17 45	8.25 9	18.5 58	4.33 58	9.40 39	11.1 57	5.72
P-LinearS [99]	27.0	1.10 15	1.67 19	5.92 19	0.53 47	0.89 46	5.71 57	6.69 33	12.0 35	15.9 24	2.60 7	8.44 19	6.71 3	5.68
PlaneFitBP [32]	28.0	0.97 10	1.83 29	5.26 10	0.17 14	0.51 22	1.71 8	6.65 31	12.1 36	14.7 17	4.17 55	10.7 54	10.6 50	5.78
GeoSup [64]	28.1	1.45 40	1.83 31	7.71 50	0.14 10	0.26 8	1.90 13	6.88 38	13.2 44	16.1 27	2.94 21	8.89 29	8.32 26	5.80

Figure 34: Middlebury stereo evaluation results [23]

In the stereo evaluation page, [17] which uses local adaptive support weights is lowest ranked among the methods described apart from the plain window matching and [18] which uses geodesic distances is six places below the cost volume filtering method. Whereas the method finally used in this thesis has not been tested on Middlebury stereo evaluation page. The problem with all the methods described above was that they were tested on only four images on Middlebury stereo evaluation page. The results of the methods on high resolution sequences were not promising compared with the method used in this thesis. For comparison purpose result of the cost volume filtering method on Balloons video sequence frame is shown below.

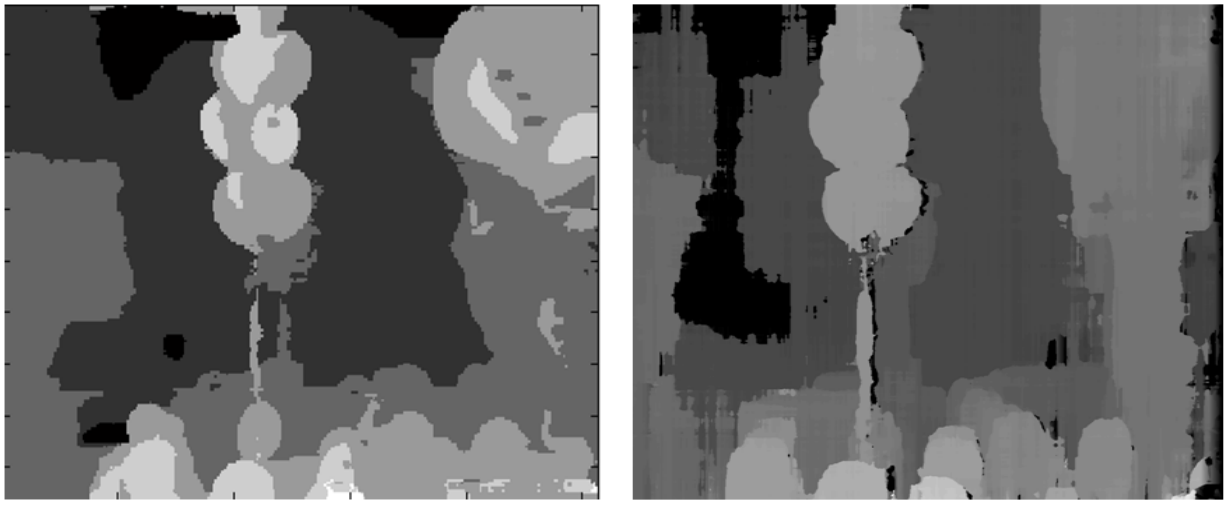


Figure 35: Left: Disparity Map Cost Volume Filtering, Right: Disparity Map Stereo Correspondence

We can see from the result above that stereo correspondence method which is not listed in the Middlebury stereo evaluation page is at least better for sequences used in this thesis than the cost volume filtering method.

Chapter 6

Objective 3D Quality Model

6.1 Introduction

In previous chapters different parts of the objective model have been described. Chapter 4 described the 2D methods, while in chapter 5 the details of 3D parameter estimation algorithms were provided. In this chapter details about the complete model are given. Apart from complete model, the specific parameters that correlate best with the subjective test are explained.

6.2 Block Diagram

Figure 36 below shows the block diagram of the complete model. This model comprises of the 2D objective quality method as well as 3D algorithms. The final output of the model is estimated MOS.

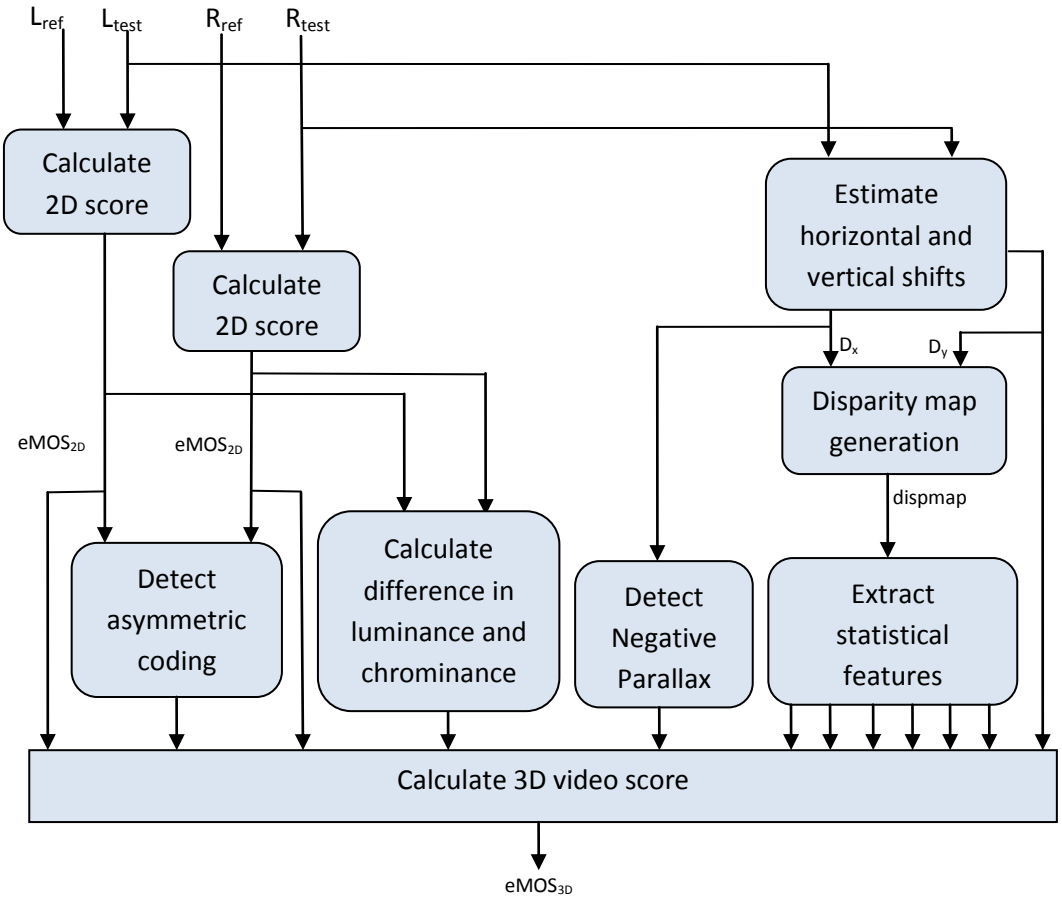


Figure 36: Objective evaluation model for 3D video

This model takes as input left, right stereo frames of test and reference video as it is a full reference model. This model is applied to each frame of video and results are accumulated across all frames. Further details about the blocks are provided in next sections.

6.3 Calculate 2D score

The method used for 2D quality score (PEVQ) is described in Chapter 4. The basic parameter used from the output of PEVQ for 2D quality of video is SCORE. SCORE and DSCORE (differential score) are two outputs of PEVQ that are inversely proportional to each other, so any one of the parameters is suitable for our case. The input to this block is test and reference videos. First left reference and test video are used as input, and then the same is done for right stereo pair sequence. At the end both scores are averaged to get the overall quality of the 2D video.

6.4 Detect asymmetric coding

In asymmetric coding one of the stereo pair is coded with a higher QP which results in higher compression. This phenomenon is useful to compress the video data to save bandwidth as small differences in compression are not noticeable but as the change between stereo pair increases it becomes evident. When stereo pairs are evaluated for 2D quality using PEVQ, the results contain all the PEVQ parameters. In asymmetric case the 2D scores of the video sequence shows evident difference between the two pairs to detect asymmetric coding.

Figure 37 Asymmetric coding of left and right image pairs

Figure 37 shows the left and right frames of a video sequence coded with QP 27 and 38, where higher QP means more compression and thus lower quality. The scores for lovebird sequence with different QP levels between left and right frames are given in table below.

QP	Left Pair Score	Right Pair Score	Score Difference
27-27	4.003	3.972	0.031
27-32	4.003	3.619	0.384
27-37	4.003	3.088	0.915

Table 7: Asymmetric coding scores for Lovebird sequence

The effect of change in QP level for one of the stereo pair becomes evident from the score difference as can be seen in Table 7.

6.5 Difference in Luminance and Contrast

The difference in luminance and contrast between the two left and right stereo pair has considerable impact on the quality of 3D as the subjective test showed. The detection of color changes among the two stereo pairs is done by using three PEVQ parameters contrast, luminance and chrominance. The chrominance of an image contains the color information. It is represented by two components U= Blue-Luma and V=Red-Luma. The luma represents the intensity of light, i.e., the brightness of the image. The three parameters selected from PEVQ best represent the disturbances added for the subjective test and any changes in contrast between the left and right image pair. The difference between the luminance of left and right image pair is used as a parameter for evaluation in objective model. An example for detection of difference in luminance is provided in table below.

Video Sequence	Luminance Difference (Without Luma disturbance)	Luminance Difference (With Luma disturbance)
Balloon	0,011	3,331
GT_FLY	0,003	0,092
Kendo	0,011	2,196
Lovebird	0,073	1,375
Newspaper	0,008	2,098
Poznan	0,098	1,373

Table 8: Luminance disturbance scores

Table 8 shows luminance difference between image pairs for six sequences used when luma is changed for one of the stereo pair. It can be seen that change in luminance is evident from luminance difference parameter.

6.6 3D Parameters

There are two main algorithms for estimation of the 3D parameters used in the model in Figure 36. For initial estimate of vertical and horizontal disparity SIFT was used which has been described in chapter 5. The parameters extracted from SIFT are mean horizontal and vertical disparities. These parameters serve as important benchmark to evaluate the quality of 3D video in itself. Using these parameters vertical and negative or cross-eyed parallax is also detected. If the calculated horizontal disparity is negative then it means that the views have been switched with each other.

The estimate of horizontal shift from SIFT is then used as input to the disparity map algorithm, which serves as the range of disparity for finding dense stereo map. Several parameters were extracted from disparity map to estimate its quality. The parameters include mean, median, standard deviation, skewness, kurtosis and correlation with reference disparity map. Then Laplacian of disparity map was calculated to find regions of rapid changes, this gives the idea of depth changes in the 3D video. All the parameters were calculated per frame and accumulated over the length of the video sequence. The next section describes how the selection of parameters was done among the available parameters.

6.7 Selection of Parameters

There were a lot of parameters that were extracted from the algorithms. PEVQ output contained 21 parameters. SIFT produced results which contained horizontal and vertical disparities. Then several statistical parameters were computed from disparity map. The selection and assessment of which parameters would be optimal to produce the better correlation with subjective test posed some problems. Some parameters calculated handled specific artifacts, e.g., SIFT output parameter of vertical disparity estimate considered only vertical disparity artifacts. So, for each artifact principle component analysis was performed separately to check which parameters were most suitable for that corresponding artifact. Figure 38 below shows the result of principle component analysis on baseline affected sequences. Only parameters which were used finally are shown for easier interpretation of the results.

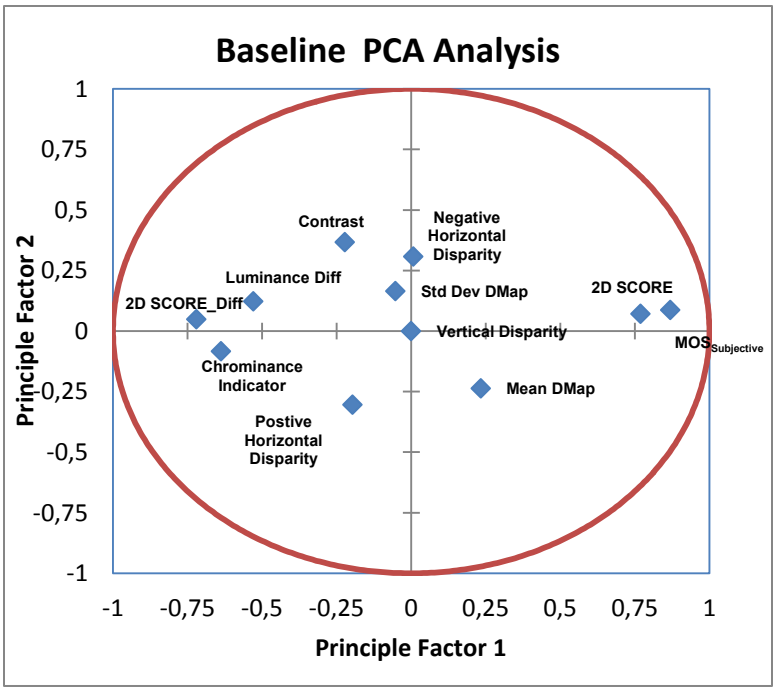


Figure 38: PCA analysis for baseline changed sequences

Principal component analysis (PCA) is shown for principal factors which represent a high variance of the data. As expected 2D quality score, mean disparity maps are best correlated with the subjective test MOS. Another important thing to note is that horizontal disparity is negatively correlated with the MOS. This is due to the reason that as horizontal disparity increases it becomes more and more annoying for viewer and thus the subjective score decreases.

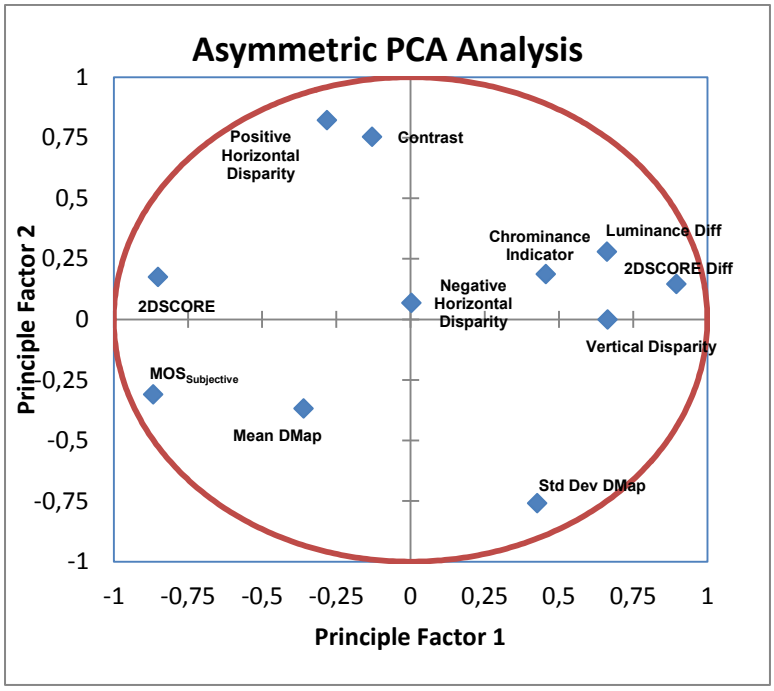


Figure 39: PCA analysis for asymmetric sequences

Figure 39 shows PCA plot for asymmetric sequences. Once again in this case MOS is very closely related to the 2D score. Another important thing to note here is that difference in left and right image pair 2D score '2D SCORE DIFF' is negatively correlated with subjective scores. As the difference in quality between the image pair increases, the subjective score for that sequence declines and alteration in 2D score between pair increases. This results in the negative correlation between the MOS and difference in 2D score of image pair. 10 parameters finally selected which covered all the artifacts introduced in the subjective test are shown in Table 9.

2D Score	2D Score difference	Contrast Indicator	Chrominance Indicator	Luminance difference	Mean DMap	Std DMap	Positive Horizontal Disparity	Negative Horizontal Disparity	Vertical Disparity

Table 9: List of Parameters used for objective evaluation

Chapter 7

Results

7.1 Introduction

This chapter presents the results of the objective evaluation model. The parameters that were used have been explained in the previous chapter. Next section briefly describes the model and then results are presented.

7.2 Method and Results

There was only one subjective test database available that was conducted at Ericsson as part of this thesis. A model was needed to predict the coefficients that best fits the database. To get a somewhat fair interpretation of the results data training and validation was performed by dividing the database into two separate parts. One part which was 70% of the whole data set was used for training and the other part was used for validation. The process of division of database was done randomly. This routine was performed ten times to ensure that overtraining of database did not occur due to too many dependent parameters. This process of training and validation on different parts of database is called cross-validation. When a model is used to predict the coefficients (training) for a dataset, the model will be optimized using the coefficients for the dataset on which it is being trained. So, when validation is performed the model does not fit the validation dataset generally as well as it fits the training data.

A linear regression model was used in this thesis although results of a non-linear 2nd order polynomial regression were better on the training data. However, in case of non-linear regression cross validation showed that the variation in results for the training and validation was inconsistent. Also for non-linear regression there is chance of overtraining due to the small size of database. The results of non-linear 2nd order polynomial regression are also provided below. For a dataset with dependent variable y_i, x_i the explanatory variable and $\beta_i$being the regression coefficients the equation of the linear model is of the form

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ \ldots\ldots\ldots\ldots\ldots + \beta_n x_{in}$$

In our case $y_i$are the subjective test scores, x_i are the objective parameters and β_i are the predicted coefficients. The coefficients were predicted using the least squares criterion.

Figure 40 and 41 below show the results of linear regression and non-linear regression. Residuals are also shown for them as well. The non-linear model used here is second order polynomial function. We can see from the graph that in case of non-linear regression the results are better. Table 10 also confirms this by providing the correlation between the subjective MOS and predicted MOS.

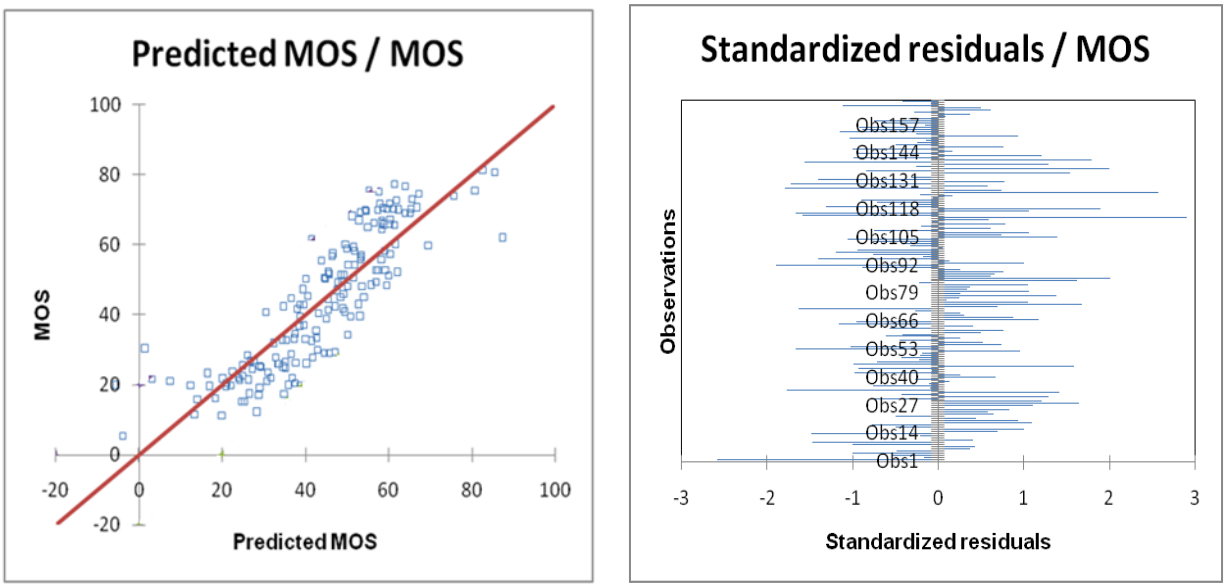


Figure 40: Predicted MOS vs. MOS and Residuals for linear regression

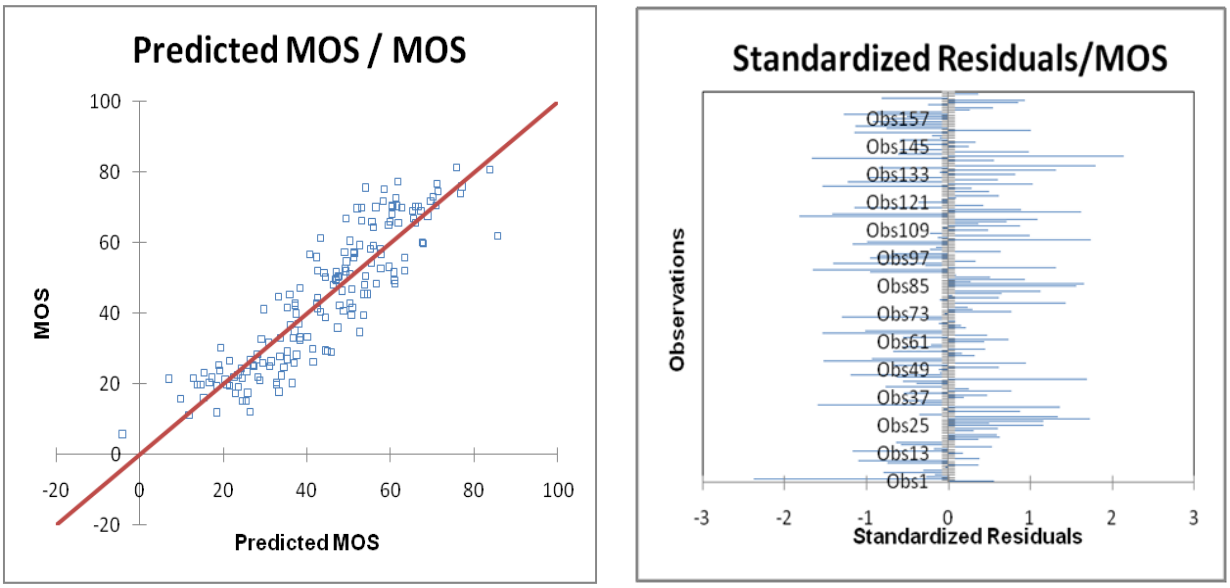


Figure 41: Predicted MOS vs. MOS and Residuals non-linear regression

Regression	Correlation
Linear	0,862104
Non-Linear	0,890285

Table 10: Correlation between Subjective MOS and Predicted MOS

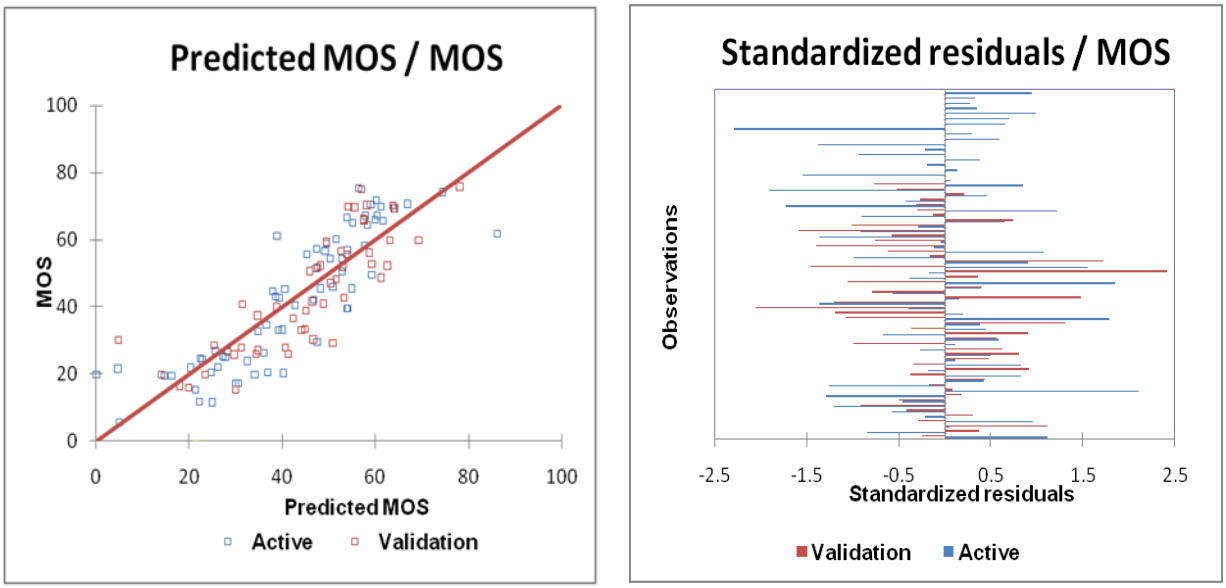


Figure 42: Training and validation results using linear regression

A set of training and validation results are also provided in Figure 42 for evaluation of the objective model. The results of training and validation show that the model performed similarly for validation case as well. From Figure 40 and 41 it can be seen that there is an outlier present with negative value in the predicted MOS. This outlier corresponds to baseline 4 Newspaper sequence with QP 42. Newspaper sequence in general has very high disparity due to the content, so at such a high disparity and QP the MOS was very low. In this case subjective score for Newspaper sequence is extremely low compared with all the other sequences of same disparity and QP. For the GT_FLY sequence with same parameters MOS was very high. This is the reason that linear and non-linear regression model are not able to fit this specific outlier.

For the objective test result root-mean-square-error [22] was calculated using the following formula.

$$RMSE = \sqrt{\frac{1}{N}\sum_{N} Error[i]^2}$$

Where,

$$Error[i] = MOS[i] - PMOS[i]$$

Here $PMOS[i]$ is the predicted objective score for a video sequence. $RMSE$ value turned out to be equal to 9.67 for scores on a scale of 0-100. For a scale between 0-1 $RMSE$ becomes 0.097.

The last step to assess the objective quality metric was to determine the outlier ratio for the predicted MOS. For objective quality method predicted MOS is an outlier if it lies outside

$$[MOS - 2\sigma, MOS + 2\sigma]$$

The lower the outlier ratio is, better the objective quality method is. By this method the outlier ratio for the validation dataset comes out to be 0.04 or 4%. Due to the small set of database the value of outlier ratio is quite small.

7.3 Summary of Results

The results of objective quality model above show that the model achieves a fair quality in evaluating the artifacts introduced in 3D video sequences. But there was only one database available on which the model could be trained and validated against, so to get a better estimate of the objective quality model a number of datasets would be required for training and validation.

Chapter 8

Conclusion and Future Work

The implemented objective quality model in this thesis is a promising start towards future work in the area of objectively evaluating the 3D video quality. The model presented here takes into account the quality of 2D video as well as commonly occurring 3D artifacts. The artifacts considered in this thesis were coding artifacts, baseline changes, asymmetric coding, negative or cross-eyed parallax, color disturbances, disparity compensation affects and vertical disparity. Asymmetric coding and color disturbances between the stereo pairs were detected using the 2D method PEVQ. Whereas SIFT algorithm was implemented to estimate the initial disparity which was used in detecting negative parallax and vertical disparity. Finally, a disparity map was generated to find the complete depth measure of the stereo video. Several parameters were extracted from each of the algorithms and they were merged using a linear regression model to get the final MOS. The results of the objective model were validated against the available database.

The processing time of the presented model is high due to the resolution of the video sequences used and algorithms implemented in this model. The time consumed by 2D method PEVQ is high compared to the other 2D methods but it provides much more accuracy and flexibility. The feature matching algorithm SIFT is also somewhat computationally complex algorithm. For effectively evaluating the quality of 3D video disparity maps are relied upon. The process of stereo matching is computationally time consuming and thus it increases the time taken to produce the results.

In the proposed model in this thesis there are several features which can be improved. The 2D quality method PEVQ is the only full reference method used in the objective evaluation model. So, it can be replaced by a no reference 2D method to make this model a complete no reference model. The stereo matching algorithm used in the model is not the most efficient in generating the disparity map. It could be replaced by a more efficient and faster global stereo matching algorithm, which will allow an increase in the reliability of this model. Another important improvement area in this model is to use a faster feature matching algorithm in place of SIFT such as Speeded Up Robust Feature (SURF) algorithm. It is faster and more efficient than its predecessor SIFT algorithm. For free viewpoint television, view synthesis is needed which poses the problem of occlusion artifacts. Therefore it is important that occlusion artifacts are also accounted for in the objective model for 3D video as well. Finally, the subjective test datasets that are available for 3D video are few. To effectively evaluate the objective quality models for 3D video, there is a need of more databases, this will enable to train and validate the models thoroughly.

References

[1] A. Vetro, S. Yea, M. Zwicker, W. Matusik, H. Pfister, "Overview of multiview video coding and anti-aliasing for 3D displays", 14th IEEE International Conference on Image Processing, 2007

[2] OpenCv 2.3 Documentation http://opencv.itseez.com/

[3] Patrizio Campisi, Patrick Le Callet and Enrico Marini “Stereoscopic images quality assessment”, European Signal Processing Conference (EURASIP), 2007

[4] Junyong You, Liyuan Xing, Andrew Perkis, Xu Wang “Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis”, 17th IEEE International Conference on Image Processing (ICIP), 2010

[5] Alexandre Benoit, Patrick Le Callet, Patrizio Campisi and Romain Cousseau “Quality assessment of stereoscopic images”, EURASIP Journal on Image and Video Processing, special issue on 3D Image and Video Processing, 2008

[6] Atanas Boev, Atanas Gotchev, Karen Egiazarian, Anil Aksay, Gozde Bozdagi Akar “Towards compound stereo-video quality metric: a specific encoder-based framework”, IEEE Southwest Symposium on Image Analysis and Interpretation, 2006

[7] Paul Gorley, Nick Holliman “Stereoscopic Image Quality Metrics and Compression”, Proceedings of Society of Photographic Instrumentation Engineers (SPIE), 2008

[8] Hang Shao, Xun Cao, Guihua Er “Objective quality assessment of depth image based rendering in 3DTV system”, 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009

[9] Anish Mittal, Anush K. Moorthy, Joydeep Ghosh and Alan C. Bovik “Algorithmic assessment of 3D quality of experience for images and videos”, IEEE Signal Processing Education Workshop (DSP/SPE), 2011

[10] “Objective perceptual multimedia video quality measurement in the presence of a full reference” Recommendation ITU-T J.247 Series J: Cable networks and transmission of television, sound programme and other multimedia signals, 2008

[11] B.D. Chaurasia, B.B.L. Mathur “Eyedness” Departments of Anatomy and Physiology, G.R. Medical College, Gwalior, M.P., India, published under *Acta Anatomica,* 2008

[12] ITU-T P.910 “Subjective video quality assessment methods for multimedia applications”, Series P: Telephone Transmission Quality, Telephone Installations, Local Line Networks, 2008

[13] Ho-Sung Han, Dong-O Kim, Rae-Hong Park “Structural information-based image quality assessment using LU factorization”, IEEE Transactions on Consumer Electronics, 2009

[14] David G. Lowe “Distinctive Image Features from Scale-Invariant Keypoints”, International Journal on Computer Vision, Vol. 2, 2004

[15] Ofir Pele, Michael Werman “A Linear Time Histogram Metric for Improved SIFT Matching”, 10th European Conference on Computer Vision, 2008

[16] Michael Bleyer VU Stereo Vision
http://www.ims.tuwien.ac.at/teaching_detail.php?ims_id=188.HQK

[17] Kuk-Jin Yoon and In-So Kweon “Locally adaptive support-weight approach for visual correspondence search”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[18] Hosni, A Bleyer, M Gelautz, M Rhemann “Local stereo matching using geodesic support weights”, 16th IEEE Conference on Image Processing (ICIP), 2009

[19] Rhemann C, Hosni A, Bleyer M, Rother C, Gelautz M, “Fast cost-volume filtering for visual correspondence and beyond”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011

[20] Middlebury Stereo Evaluation – Version 2 http://vision.middlebury.edu/stereo/eval/

[21] Abhijit S. Ogale and Yiannis Aloimonos “Shape and the stereo correspondence problem”, International Journal of Computer Vision, Vol. 65, 2005

[22] Alexander Wörner, Alex Bourret, The Video Quality Experts Group (VQEG) “RRNR-TV group test plan” for evaluation of objective models, Draft version 1.7, 2004

[23] D. Scharstein and R. Szeliski. “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms”, International Journal of Computer Vision, 2002.