MAXIMUM DISPARITY ESTIMATION FOR DEPTH DISCOMFORT DETECTION

by

Ömer Can Gürol

B.S., Electronics Engineering, Işık University, 2010

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Electrical and Electronics Engineering

Boğaziçi University

2013

ACKNOWLEDGEMENTS

Foremost I would like to thank my supervisor Prof. Bülent Sankur for his amazingly useful comments and suggestions during the learning process of this thesis. I would also like to thank Assoc. Prof. Burak Acar for his encouraging guidance and reviews.

Furthermore I would like to thank Mehmet Güney from Digitürk A.Ş. Software Development Team for introducing me to this very interesting topic and the support he provided on the way. Also, I like to thank Bernhard Moser for fruitful discussion about Herman Weyl's Discrepancy Measure as well as sharing their MATLAB codes.

I would like to thank the 3D shooting crew at Digitürk A.Ş and all participants of the subjective tests we performed for sharing their valuable time during data collection and evaluation.

Finally I would like to thank my beloved family and friends, who supported me with their encouragement and endless patience through the entire process. I will be grateful forever for your love and support.

ABSTRACT

MAXIMUM DISPARITY ESTIMATION FOR DEPTH DISCOMFORT DETECTION

This thesis documents a work in which prediction of depth related discomfort levels, while watching stereoscopic 3D videos is studied. In commercial 3D videos, excessive depth levels can cause discomfort to the viewer and hence decreases the users quality experience of the video. Therefore detecting excessive levels of depth is important to maintain a better visual quality in 3D videos. In this work a scheme is presented for detection of depth discomforts, resulting from excessive depth levels. An exaggerated depth corresponds to high disparity levels between stereo image pairs. In order to detect depth discomforts, we developed and tested algorithms to detect and track maximum disparities. The maximum disparities are extracted from sparse disparity maps, where the disparities are obtained for only certain edge locations. The sparse disparity maps are obtained using five varieties of block matching, which are: Sum of Absolute Differences (SAD), Herman Weyl's Discrepancy Measure (HWDM), Adaptive Support Windows (ASW), Sum of Absolute Differences of Scale Invariant Feature Transforms (SADSIFT) and Correlation of Gradient Orientations (CGO). A comparative study of these five methods is performed in terms of their performances in estimation of maximum disparities. Also subjective tests are run by collecting viewer discomfort data and using maximum disparity statistics as a predictor of user experience. By examining our results, we observed CGO performs better in maximum disparity estimation. Also it is shown, that the maximum disparity statistics obtained through CGO can be used to predict the number of viewers, which experience depth discomforts.

ÖZET

EN BÜYÜK AYRIKLIKLARIN KESTİRİMİ İLE DERİNLİK RAHATSIZLIĞI TESBİTİ

Bu tezde stereoskopik 3B (üç boyutlu) videolar izlenirken ortaya çıkabilen derinlik kaynaklı rahatsızlıkların öngörülmesi üzerine yapılan bir çalışma belgelenmektedir. Ticari amaçlı 3B videolarda ortaya çıkabilen aşırı derinlik seviyeleri izleyicileri rahatsız edebilir ve bundan dolayı da kullanıcının video izleme deneyiminin kalitesi düşer. Bu nedenle görsel kalitesi yüksek 3B videolar elde edilebilmesi açısından, videolardaki aşırı derinlik seviyelerinin tesbiti önemlidir. Bu çalışmada aşırı derinlik seviyelerinden kaynaklanan derinlik rahatsızlıklarının tesbiti için bir plan ortaya konulmaktadır. Aşırı derinlik, stereo imge çiftleri arasında yüksek ayrıklık (disparity) seviyelerine işaret etmektedir. Bu nedenle, derinlik rahatsızlıklarının tesbiti için en yüksek ayrıklık seviyelerini tesbit ve takip eden algoritmalar denenmiştir. En yüksek ayrıklıklar, sadece ayrıtlar üzerindeki ayrıklıkları içeren, seyrek ayrıklık haritalarından elde edilirler. Seyrek ayrıklık haritalarını elde etmek için beş farklı blok eşleme yöntemi denenmiştir: Mutlak Farkların Toplamı (SAD), Herman Weyl'in Uyuşmazlık Ölçütü (HWDM), Uyarlanabilir Destek Pencereleri (ASW), Ölçekten Bağımsız Öznitelik Dönüşümlerinin Mutlak Farklarının Toplamı (SADSIFT) ve Gradyant Yönelimlerinin Korrelasyonu (CGO). Bu beş yöntem en yüksek ayrıklıkların kestirimindeki başarımlarına göre karşılıklı olarak incelenmişlerdir. Ayrıca izleyici rahatsızlığı ile ilgili verilerin toplanması ve en yüksek ayrıklık istatistiklerinin kullanıcı deneyimini ön görmede kullanılması ile öznel testler de yapılmıştır. Sonuçlarımızı incelediğimizde CGO yönteminin en yüksek ayrıklıkların kestiriminde daha başarılı olduğu gözlenmiştir. Ayrıca CGO ile elde edilen en yüksek ayrıklık istatistiklerinin derinlik rahatsızlığından şikayatçi olabilecek izleyici sayısının ön görülmesinde kullanılabilecekleri gösterilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

b	Horizontal length of the search block
C	Total number of reliable disparity estimates
D_f	Binocular disparity of a far object point
D_n	Binocular disparity of a near object point
d	Candidate disparity value
d^C	Cross-check disparity estimate
d^E	Estimated disparity
d_v^E	Estimated disparity value at threshold v
d^G	Ground truth disparity
d_v^G	Ground truth disparity value at threshold v
f_R	Reference Image
f_T	Target Image
G_i	Complex gradient map of image i
h	Vertical length of the search block
I_q	Integral image along direction q
k	Disparity search range parameter
N	Search block size
O_L	Gradient orientation map of the left image
O_R	Gradient orientation map of the right image
p_s	Start frame
p_e	End frame
q	Direction of integration
R	Rounding operator
$Ratio95\%$	95 percentile ratios
s	Dilation width
$SR\{p_s, p_e\}$	Slew rate of d_v^E values of a scene starting at frame p_s and ending at frame p_e
T	Disparity estimation error test
v	Threshold at 95% of sorted reliable disparities

W	Size parameter of rank filter window
x	Vertical coordinate of an edge point
y	Horizontal coordinate of an edge point
Z	Gradient magnitude threshold
α	Angular disparity at left eye
β	Angular disparity at right eye
$\mu 95$	Mean of the largest 5% of true disparities
μd	Mean of d_v^E values of a scene starting at frame p_s and ending at frame p_e
$\sigma\{p_s, p_e\}$	Standard deviation of d_v^E values of a scene starting at frame p_s and ending at frame p_e

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
ASW	Adaptive Support Windows
CGO	Correlation of Gradient Orientations
CPU	Central Processing Unit
$Diff95\%$	95 percentile absolute error
$Erroneous\%$	Percentage of erroneous disparities
HVS	Human Visual System
HWDM	Herman Weyl's Discrepancy Measure
$LRBT$	Left to right and bottom to top
$LRTB$	Left to right and top to bottom
MAE	Mean Absolute Error
$RLBT$	Right to left and bottom to top
$RLTB$	Right to left and top to bottom
SAD	Sum of Absolute Differences
SADSIFT	Sum of Absolute Differences of Scale Invariant Feature Transforms
SIFT	Scale Invariant Feature Transforms
TV	Television

1. INTRODUCTION

In recent decade many developments have been established in the field of stereoscopy and 3D technologies. Although the foundations for stereoscopic movie production was known since the beginning of the twentieth century, the amount of commercial stereoscopic content had been scarce until the last decade. The emergence of commercial 3D TV's and developments in digital 3D video production and delivery increased the demand for stereoscopic content. This demand required a wider understanding of stereoscopy in order to provide high quality visual experience for the viewers. Hence the number of researches in stereoscopy and 3D technologies also increased significantly in recent decade.

Basically stereoscopy can be described as a technique to create an illusion of depth for the viewer, when the scene is observed with both eyes. In normal conditions we perceive our surroundings with both of our eyes and each eye gets a slightly shifted image of the observed scene in reference to the other eye. These shifted images are merged by Human Visual System (HVS) in order to create an impression of depth of the observed scene. This property of HVS enables the stereoscopy, such that by presenting to each eye slightly shifted images, an illusion of depth is created and a more realistic visual experience for the viewer is aimed.

Of course in order to achieve a realistic depth illusion, the image pairs presented to each eye should meet certain quality requirements individually and also they should be compatible with each other. To satisfy these requirements certain measures should be taken during capturing, production, transmission and viewing stages of the stereo content. For 2D visual content many methods are available today to enhance or preserve the quality during any of these stages. However for stereo 3D content, improving some of the quality degrading effects originating from previous stages is not always an easy task. Therefore, quality assessment methods are actually required in every step of 3D movie broadcasting and stereo visual quality assessment methods are expected to become a very important part of 3D content delivery pipeline in a near future.

1.1. Motivation

Creating a satisfying experience for the viewer is the most crucial goal in commercial 3D technologies. In this sense, creating a satisfying depth illusion while assuring the eye comfort of the viewer is the most important aspect. This requires some 3D specific quality evaluation and improvement methods to be considered, in addition to the quality enhancement measures taken in 2D content production and delivery. In 3D videos besides having visually good images for each stereo channel, it is also important that the channels match with each other. Stereo quality control methods have identified a number of factors affecting viewing experience, such as parallax (offset between image pairs) irregularities, focus mismatch, color mismatch, geometry mismatch, vertical parallax, object edge tearing, cardboard effect, pincushion distortion etc [1]. In this work, we focus solely on the assessment of parallax related errors that can lead to severe viewer dissatisfaction.

In order to measure the parallax between stereo image pairs, pixel wise disparities must be obtained. The disparity value associated with a pixel is the amount of offset between that pixel and its shifted version in the other image, such that the disparity values reflect the local parallax and hence relative amount of local depth. Therefore it is possible to observe the disparities as an indicator of depth and by extracting the necessary statistics from them, the viewer satisfaction related to parallax irregularities can be predicted.

In this thesis, a scheme is presented for automatic assessment of stereo disparity quality from a viewer point of view, with the purpose of estimating the out-of-range disparity occurrences in order to detect scenes with depth discomfort. The motivation for this work is to develop algorithms that will estimate disparity fields and spot scenes where the extent of disparity exceeds comfort level. Thus a video document, e.g., a cinema film or a TV film can be marked with shots where disparity is out of range, and these can constitute an overall figure of disparity merit for the video work. As a concrete example, with this algorithm a broadcasting company can perform quality analysis for the material provided by various 3D video content vendors and decide whether the

movie can be broadcasted or not. Also with real-time application of this algorithm 3D movie shooting operators can keep track of the depth levels while shooting and hence they can take higher quality stereoscopic shots by keeping the depth in appropriate range for comfortable viewing.

1.2. Outline

The thesis is organized as follows: In this chapter an introduction is done by describing stereoscopy and the motivation is given to explain the underlying idea of the thesis. Chapter 2 inspects the literature about depth discomforts and stereo video quality assessment, also depth perception of HVS is briefly explained. Chapter 3 includes a brief literature review about disparity estimation and elaborates our sparse disparity map estimation methodology. In Chapter 4 the experimental setup is described together with the data sets we used. Also results of our experiments are given in Chapter 4. Chapter 5 concludes the thesis with a summary of the described system and results we obtained.

2. DEPTH PERCEPTION AND DISCOMFORTS

2.1. Perception of Depth in Stereoscopic Visual Content

The mechanism of depth perception in the human visual system (HVS) is fairly well understood. It is known that depth perception uses both psychological and physiological cues. On the psychological aspect, HVS uses separate families of depth cues, which include binocular parallax, motion parallax, accommodation, and perspective, while the physiological aspect consists of separate visual mechanisms and neural paths [1]. These depth cues depend on different physiological events and they are effective in different visual ranges.

For example accommodation cue is most effective in determining the depth of objects close to the eye and it depends on the amount of contractions and expansions of the eye lens, while focusing on objects at various distances. Motion parallax is used when the changing posture of a moving object is used to estimate its depth and is effective in larger distances. Perspective cue is mostly effective for determining the depth of the objects far away from the eye and depends on physiological events such as overlap, shadow, apparent size and texture, while estimating the depth of the objects. Binocular parallax refers to the difference in images between the two eyes caused by their different location and it is effective in estimation of depth of the objects, which are in short to middle distance from the eye. This binocular cue of HVS is the main factor that enables the stereoscopy.

Binocular cues are mainly the consequence of having two eyes horizontally positioned in the head. HVS receives two views of a scene, which slightly differ because they are from two different perspectives due to horizontal separation of eyes (typically 50 to 75 mm). These two images are fused by HVS to produce stereoscopic depth. When our eyes fixate at a point, the images of that point fall in both eyes on the same relative coordinates on the *fovea*, which is the back part of the eye. The fixation point falls on the *horopter*, which is a curved line or surface that contains all points at the

same geometrical or perceived (empirical) distance of the fixation point [2,3]. In accordance with the horopter, points located in front or behind of it are imaged in different locations of the eye, and these differences are described as *binocular disparity*. In other words, it is the relative spatial distance between similar points, which share the same physical origin, in left and right stereo image pairs. Binocular disparity is defined as *negative* if an object is before the horopter curve and *positive* if the object is behind the horopter curve. The brain uses binocular disparity to infer depth information from the two-dimensional retinal images resulting in 3D perception, that is, stereopsis. In Figure 2.1, the geometrical and empirical horopter curves can be observed together with the *Panum's fusional area*, which is the are where HVS can successfully perform the binocular fusion of the images in both eyes. The objects outside this area result in double vision and the size of this area depends on many spatial and temporal properties of the fixed object [3].

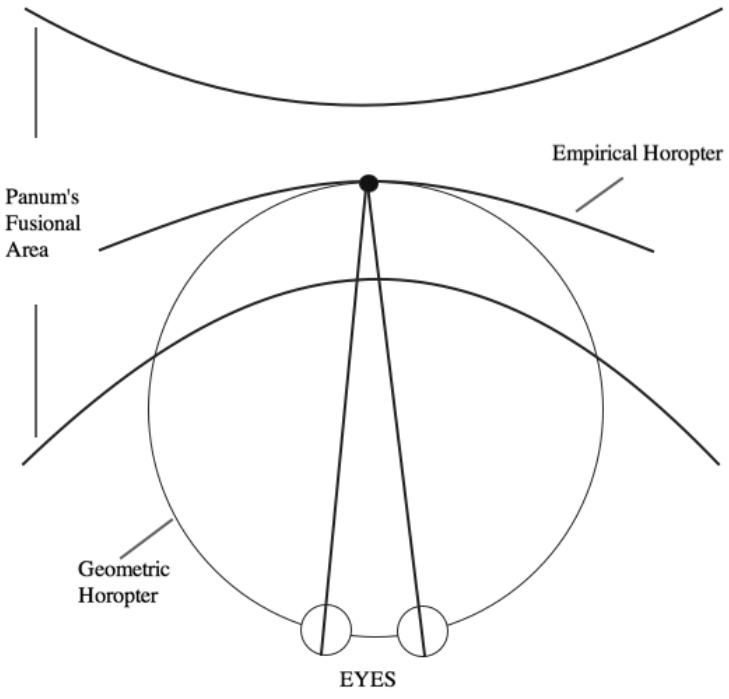


Figure 2.1. Definition of geometrical and empirical horopters and fusional limits [3]. Objects outside of the Panum's fusional are result in double vision, since HVS can not fuse the single 3D object.

In stereoscopic screens a depth illusion is created by presenting the left and right camera views separately to the eyes of the viewers, such that the eye separation requirement for binocular vision is ensured. In commercial 3D TV's the eye separation is mostly enabled with the users wearing active or passive glasses, although there are also some relatively new technologies such as autostereoscopic displays, which do not require wearing any glasses [1, 4, 5]. Active glasses require a power source and they enable the eye separation by sequentially opening and closing the shutters in them in synchronisation with the display rate of the screen, such that the image pairs are temporally separated for both eyes. Passive glasses do not require any power source and they achieve the eye separation by filtering the image polarized by the display with their polarization filters [3, 5].

The creation of depth illusion via binocular disparity in stereoscopic screens is influenced by the size of the 3D display and the viewing distance, given the same relative parallax. Thus, the disparity requirements vary proportionally for cinema viewing (typically 20 m), home TV viewing (typically, 1.5 m) and mobile device viewing (typically 0,2 m). In practice, smaller screens require a larger stereo baseline to provide more disparity as a fraction of the image width and to retain a good impression of depth.

In optics, binocular disparity is usually measured in units of arcmin, such that 1 arcmin is $1/60^{\text{th}}$ of a degree. In this sense binocular disparity for a near object is calculated as $D_n = -\beta - \alpha$ and for a far object as $D_f = \beta - \alpha$, where α and β are angular disparities at left and right eyes as can be seen in Figure 2.2 [6]. Note that $D_f \geq 0 \geq D_n$.

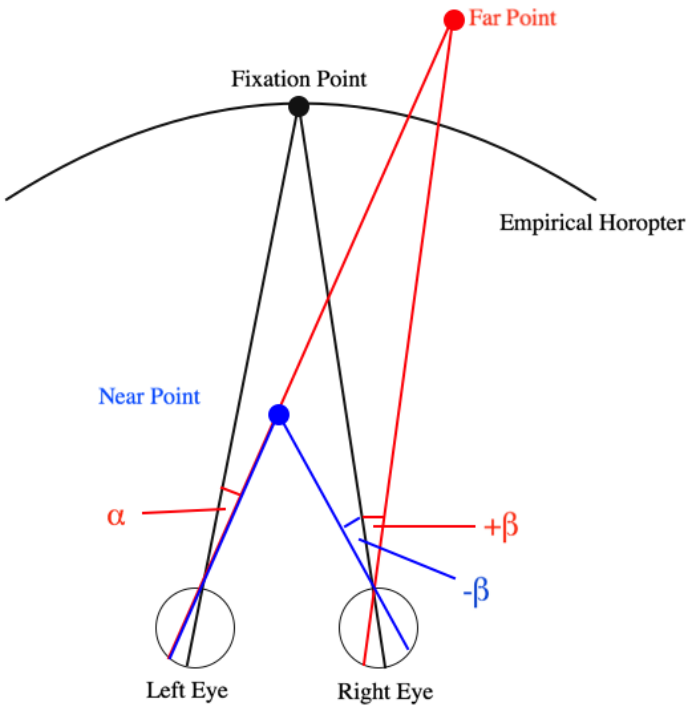


Figure 2.2. Geometry of binocular disparity. Binocular disparity for near object is calculated as $D_n = -\beta - \alpha$ and for a far object as $D_f = \beta - \alpha$, hence binocular disparities of near and far objects are named as negative and positive respectively.

For stereoscopic content it has been recommended that binocular disparity be upper bounded at 60 arcmin to ensure visual comfort for the majority of viewers [2,7,8]. Also it has been suggested that up to 2 degrees (120 arcmin) of disparity is tolerated between two images before the sensation of depth is lost, however such a disparity level is not guaranteed to be comfortable for the viewers [9].

In digital environments the disparity is defined as a percentage of horizontal image resolution or in units of pixels or sub-pixels. In this sense the term disparity mostly refers to the relative distance in units of pixels between points in both images, which are corresponding to the same physical origin. The relation between pixel disparities and binocular disparity depends on conditions like viewing distance, eye separation, screen size and resolution. Therefore defining a pixel disparity range independent of

these conditions for safe binocular vision is not possible. However it can be said that pixel disparities with higher percentages of resolution would mostly cause visual fatigue or a discomfort. As a general rule, having a negative pixel disparity range of 2% and a positive pixel disparity range of 1% of horizontal resolution has been recommended for experiencing less visual fatigue in movie theatres and similar conclusions have been reached for 3D TV's [3].

In order to create a satisfying depth illusion, the pixel disparities between the image pairs should be made compatible with HVS 3D perception. Very high disparity values correspond to an exaggerated depth range and may cause the depth of the scene to be perceived inaccurately by the observer. Such defective image pairs can cause the depth illusion to be weakened in the observer, and it can even result in headache or nausea when exposed for a long time [10].

2.2. Depth Discomforts

Although the research focused on stereoscopy has increased significantly in recent decade, number of research focused on measurement and elimination of stereoscopic discomforts has been limited until recent years. The types of discomforts experienced while watching 3D movies and their causes are well documented in a number of researches [1–3, 7, 9]. In these studies the discomfort types are well classified according to their origins in the 3D video delivery pipeline, related depth perception cues and physiological causes. In general the main causes of stereoscopic discomforts can be classified as:

- Accommodation-vergence conflicts [3]: While watching stereoscopic content lenses in our eyes accommodate according to the distance to the screen, however we follow the 3D scene and therefore the eyes converge on virtual objects, whose disparities are varying. If the disparities of these objects in the scenes are excessive, the conflict between accommodation and vergence increases and the eye starts to accommodate to the vergence point instead of the screen, resulting in blurry vision and visual fatigue.

- Excessive disparities [2, 3, 9]: As we mentioned previously, the disparity of the images should be limited in a comfort zone . If the disparities exceed the Panum's fusional area limits, the stereo pair images can not be fused and the sensation of depth is lost. However even if they are within the limits of the fusional area, the viewer may experience visual fatigue when exposed to excessive disparities for a long time.
- Crosstalk [9]: If the separation of the image pairs between two eyes is not performed properly, some part of the left image can be perceived by the right eye and some of the right image can be perceived by the left eye. This causes the contours in the stereo scene to be perceived in doubles, creating the so called "ghosting" effect.
- Differences between image pairs [11]: Usually the only difference between stereoscopic image pairs should be the perspective, such that one image should be slightly horizontally shifted version of the other. However different effects can lead the two image pairs to have more differences than perspective. Difference in color, blur, vertical disparities, noise and occlusions between image pairs can decrease the perceived stereo quality.
- Absence of motion cues [11]: Normally humans also benefit from the motion cues to perceive depth. The movement of objects or movement of the head gives certain cues about the depth of the objects. However in stereoscopic videos, the change of head position does not result with any depth cue. In videos the object motion keeps helping on depth perception, but if disparity levels of the objects are not properly set, they may conflict with the motion cues.
- Human factors [3]: There are also some human factors related to special conditions of the viewers. Watching the stereoscopic videos at unusual angles or distances to the screen may result in higher binocular disparities and other distortions to be perceived by the eye causing visual fatigue and decreased image quality. In addition to that it should be noted that the binocular vision is not same for every individual. The depth illusion can not occur for some, whose visual systems do not often refer to binocular cues for depth perception. Also the

inter-pupillary distance varies between humans, therefore comfortable disparity levels determined for some people can cause discomforts at others.

The discomfort types encountered in stereoscopic displays are fairly well understood, however there are quite limited work on measurement and prediction of discomforts. A commonly accepted methodology for evaluation of stereo discomforts is not yet available. The proposed approaches in existing studies are investigating very different aspects of visual discomforts and the strategies for quantification of discomforts can be significantly different between these studies. Existing approaches in discomfort measurement can be classified in two groups: objective and subjective methods. Objective methods try to measure physiological responses of HVS in order to measure the discomfort level by tracking the eye response [12] or the brain activity [13], while subjective methods depend on the answers of the test subjects about how well they perceived the stereo scenes. In general subjective methods are more preferable, since they are easier and cheaper to implement. The rest of this section is dedicated to brief explanation of some of the available subjective discomfort measurement and prediction methods.

In [14] a quality analysis method for stereo images is performed by applying some well known 2D quality metrics on stereo images and disparity maps. Different distortions have been applied on images and the disparity map. Through subjective tests the relevancy of applying 2D metrics on stereo has been tested. The 2D metrics are applied separately for each image including the disparity map and then their results are fused together. Of course this study only considers the noise and compression related discomforts, which causes differences between image pairs.

The visual comfort in stereo videos is also related to the planar motion. In [15] relation between visual discomfort and planar motion at different depth levels is investigated. By performing subjective tests with videos containing different depth levels and velocities, they concluded that experienced discomfort increases with increasing velocities. It has been also stated that, small objects with large disparities and large objects with small disparities can be equally disturbing for the viewers.

A method for quality assessment of 3D content has been proposed in [16], which depends on statistical features extracted from the disparity maps, disparity gradient maps and spatial activities in the image pairs. Also for videos motion compensated disparity difference maps have been used. Different statistical features have been extracted from the disparity related maps and the image pairs and the relevancy of these features in discomfort prediction have been investigated. It has been concluded that for images, mean and median of the disparity maps, and for videos spatial activity from the image pairs and the motion compensated disparity statistics are the most important features in stereo quality assessment.

In [17] a visual metric for discomfort prediction has been proposed, which considers human attention. The disparity statistics obtained from the salient regions in the images are used for prediction of subjective test scores. It has been shown that in comparison to global evaluation of disparity maps, extraction of statistical features from the salient regions increases the performance in discomfort prediction.

Also there are studies to predict discomfort levels, which consider the size of the objects together with the disparity magnitudes. In [18] features extracted from disparity maps, disparity gradient maps, spatial frequencies and object width have been used for discomfort prediction and it has been reported that object size related features improve the performance. In [6] the influence of binocular disparity, retinal blur and object size, on perception of depth in stereoscopic content has been investigated through subjective tests. It has been stated that beside of binocular disparities, retinal blur and object size provide very important depth cues and hence improve the depth perception.

The work proposed in this thesis is solely focused on predicting the stereo discomforts, resulting from the excessive disparity magnitudes present in the video scenes. This study contributes to the literature in several aspects:

- First of all, instead of performing the discomfort analysis for each frame in a video, we investigated the subject responses given to video scenes containing many frames. Our approach considers the maximum disparities extracted for each

frame in a video scene and certain maximum disparity statistics are calculated for each scene. By performing the discomfort analysis on video scene level, the scenes, which mostly contain discomfort causing frames, can be differentiated from the scenes with fewer number of faulty frames.

- Secondly, the maximum disparity statistics are used to predict the number of subjects, who reported a discomfort for that video scene. Predicting the number of subjects instead of giving a quality score is more intuitive and has a physical meaning. For example a 3D TV broadcast company can decide whether to air a content or not, based on such an analysis, such that they would have a number about how many of their customers would comfortably enjoy it.
- Finally, we also introduce a point disparity estimation algorithm. In the literature most of the studies have used existing disparity estimation algorithms, which estimate the disparities for each pixel and resulting dense disparity maps. Dense disparity maps can be very helpful in many tasks, however for discomfort prediction in videos, obtaining disparity at each pixel of each frame is redundant and costly. Therefore we propose a point disparity estimation method, which estimates the disparities only at certain image specific locations. For this approach, we also provide comparative analysis of five different cost calculation methods. The details of the point disparity estimation algorithm is given in the next chapter.

3. POINT DISPARITY ESTIMATOR

3.1. Related Work

Disparity estimation has been the subject of much interest in the last two decades and a plethora of algorithms have been developed. With existing benchmark stereo image sets [19, 20], various approaches on dense disparity map estimation has been studied comparatively [21, 22]. Mainly the existing stereo matching algorithms can be categorized in two groups: local and global methods. Local methods treat a region of pixels in the reference image independently and try to find the point in the other image, where this ensemble of pixels best match. Global methods try to minimize an energy function over all pixels by depending on some assumptions about the variability of disparity values over the image. Local methods are also referred as correlation based methods and global methods are referred as pixel based methods.

Disparity estimation can be formulated as a motion estimation problem, where the motion vectors are longer and expected to be in horizontal direction. In such a sense some common motion estimation approaches are applied for stereo matching. Sub-pixel accuracy estimation of disparities [23] and coarse-to-fine estimation using image pyramids at different scales [24] are some of them. These methods are known to yield accurate results in general, but the existence of large displacement vectors requires some other measures to be taken in disparity estimation.

For matching the corresponding points in the images, some matching costs should be calculated at each candidate point and the point with minimum cost should be chosen at the end as the match. Hirschmüller and Scharstein have experimented with different cost aggregation methods using both local and global approaches and they concluded that rank transform performs best for local methods and hierarchical mutual information performs best for global methods [22].

Most of the work on the local methods is focused on smoothing the cost space. By performing this smoothing, obtaining the minimum cost value and hence obtaining the location of the best match becomes easier. This smoothing is performed mostly by using adaptive support windows for calculating correlation values or by smoothing the cost space with an adaptive filter. Kanade and Okutomi have presented a method to chose appropriate windows according to the local variations in intensities and disparities [25]. However performance of this method depends highly on initial disparity estimates presented to the algorithm.

Yoon and Kweon have proposed an adaptive support-weight method, which assigns different support weights to pixels in a support window according to the properties of the pixels in the window [26]. This approach is similar to smoothing the cost space by preserving edges and has yielded accurate results.

Hirschmüller *et al.* have presented another cost volume method, where the cost is calculated by taking four overlapping support windows around the center correlation window and adding two of the minimum cost values from these four windows to the cost values of the center window [27]. It has been claimed that this approach handles the depth discontinuities better and it is faster than other cost volume smoothing methods.

Other approaches have been introduced to smooth the cost space, which are essentially faster approaches than [26]. Richardt *et al.* smoothed the cost volume by using the bilateral grid approach [28] and Hosni *et al.* used a guided filter approach, which is basically an edge preserving filter that smooths the cost volume [29]. Cigla and Alatan performed similar edge preserving cost space smoothing by introducing a filter, which is independent of window size and uses a color similarity based support region [30].

On the other hand global methods minimize an energy function according to the assumption that the disparities should vary smoothly except at the object boundaries. A comparative study is also available, where most popular global energy minimization methods such as graph-cuts and belief propagation are investigated [31].

Graph-cut methods are implemented with high accuracy results [32,33]. However calculation of maximum flow in graph-cuts is an extensive task, which increases the computational time required for these methods.

Felzenszwalb and Huttenlocher have proposed a belief propagation approach, where an iterative strategy is used that approximates Markov Random Fields for finding the best match [34]. The belief that the disparity at a pixel is optimum arises from the belief values of neighboring pixels and at each iteration the belief is propagated to all connected pixels, which enables obtaining smooth disparity maps with precision along the depth discontinuities.

Notice that these more sophisticated global algorithms focused on reconstructing dense disparity field using global optimization [32–34] are not fit for our purposes, due to their time complexity and the fact that they try to estimate the disparity at every pixel, which is highly redundant for our purposes.

To differentiate our method from the dense disparity map methods in the literature, we will call ours the *point disparity estimator* and the outcome as *sparse disparity map*. Our method basically chooses some edge points and estimates the disparities at these points. As illustrated in Figure 3.1, our point disparity estimation algorithm consists of image preprocessing, disparity search guided by the image edge field, and post-processing for error correction.

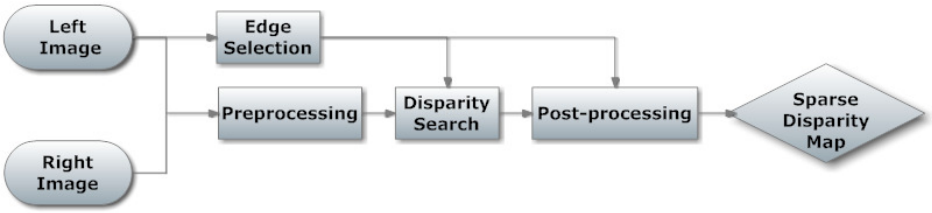


Figure 3.1. General flowchart of the point disparity estimator. Both of the image pairs are subject to preprocessing. Edge selection is only applied to reference image (left image). The resulting edge map guides the disparity search and post-processing steps.

In image preprocessing step the edge maps are obtained and they are refined in order to fit the needs for their guiding roles. Also rank filtering is applied on both of the image pairs prior to disparity search methods based on difference of pixel intensities.

The disparity search step is based on block matching approach. Mainly the block of pixels centered on certain edge locations in a reference image are taken and searched in the target image in horizontal direction in order to determine the point of best match (Figure 3.2). This search is performed according to the similarity of the original block and the blocks of similar size, which are taken around the horizontal search locations in the target image. The block, which is most similar to the original block is called as the *matching block* and the distance between the centers of the original and matching block gives the disparity value. Finally post-processing methods, such as cross-checking and disparities in a band, are applied to the sparse disparity map to eliminate erroneous disparity estimates.

More detail regarding the preprocessing, disparity search and post-processing steps of the point disparity estimator can be found in the following sections together with the details of five different matching cost calculation methods we have investigated.

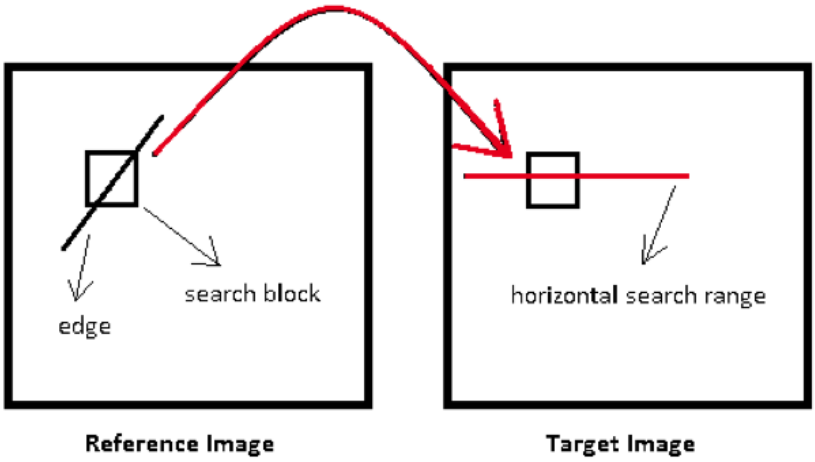


Figure 3.2. Block matching in disparity search step. Search block centering an edge point is taken from the reference image and searched in the target image along the horizontal search range.

3.2. Image Preprocessing For Disparity Estimation

In order to determine the edge locations, where the disparity search will be performed and to improve the performance of the disparity estimation some preprocessing methods are applied. Prior to disparity search step, edge processing and rank filtering steps take place. In the following section more detail about these preprocessing steps can be found.

3.2.1. Extraction and Determination of Edge Points

For point disparity estimation to work, the points, for which the disparities will be estimated, must be given to the algorithm. For our purpose, these points are chosen among the edge points obtained from the reference image. The edge map of the reference image is extracted and then processed further to obtain final edge map, which is used in disparity search and post-processing steps as a guide. The guiding edge map is obtained following these steps:

(i) Detection of edges in the reference image using Canny algorithm
(ii) Elimination of horizontally oriented edges
(iii) Dilation of remaining edges

After obtaining initial edge map of the reference image by using Canny algorithm, the horizontal edges are removed from the map. Remember that we limit our search for stereo correspondences only in the horizontal direction, assuming that the cameras are rectified. If the horizontal edges are not eliminated from the map, they overlap with the horizontal search direction, which yields ambiguous results. This ambiguity originates from the fact, that the image texture does not change significantly along the horizontal edges. As a consequence, the blocks taken along the horizontal direction are very similar to each other and the block matching algorithm gives multiple best matches along the horizontal search line.

In Figure 3.3 an example of this ambiguity can be seen, where a block from the reference image and its corresponding horizontal search range in the target image are shown. It can be observed, that along the horizontal search direction the texture does not change, making it impossible to determine where the best match of the search block is.

To overcome this ambiguity, we chose to avoid obtaining any disparity estimate for horizontally aligned edges, since obtaining few but more accurate estimates is more preferable, compared to obtaining inaccurate estimates. We limit the disparity search only on edges with orientations within the $60^\circ - 120^\circ$ degree cone, eliminating all edge points whose orientation angles remain out of this cone.

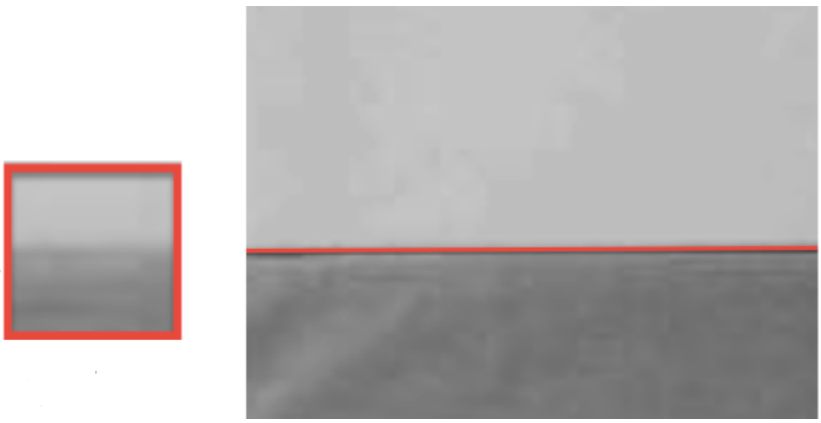


Figure 3.3. Search block and the horizontal search direction. Left: The search block from the reference image; Right: Horizontal search range in the target image (Red line). It can be observed, that the texture does not change significantly along the horizontal search line, resulting ambiguous results in disparity search.

After removing the horizontally oriented edge points from the edge map, the remaining edges are dilated horizontally through morphological operations, to make them s pixels wide. By doing so, the disparities are estimated not only for the original edge points, but also for their neighbors to the left and right. This operation results s disparity estimates for a single original edge point, and then one of these estimates is chosen in the post-processing step (see Section 3.4.2).

An example of these edge processing steps can be seen in Figure 3.4 for the left image of a stereo pair together with the final disparities, which are shown on the left image as blue lines. Note that the disparities are obtained only around the non-horizontal edges.

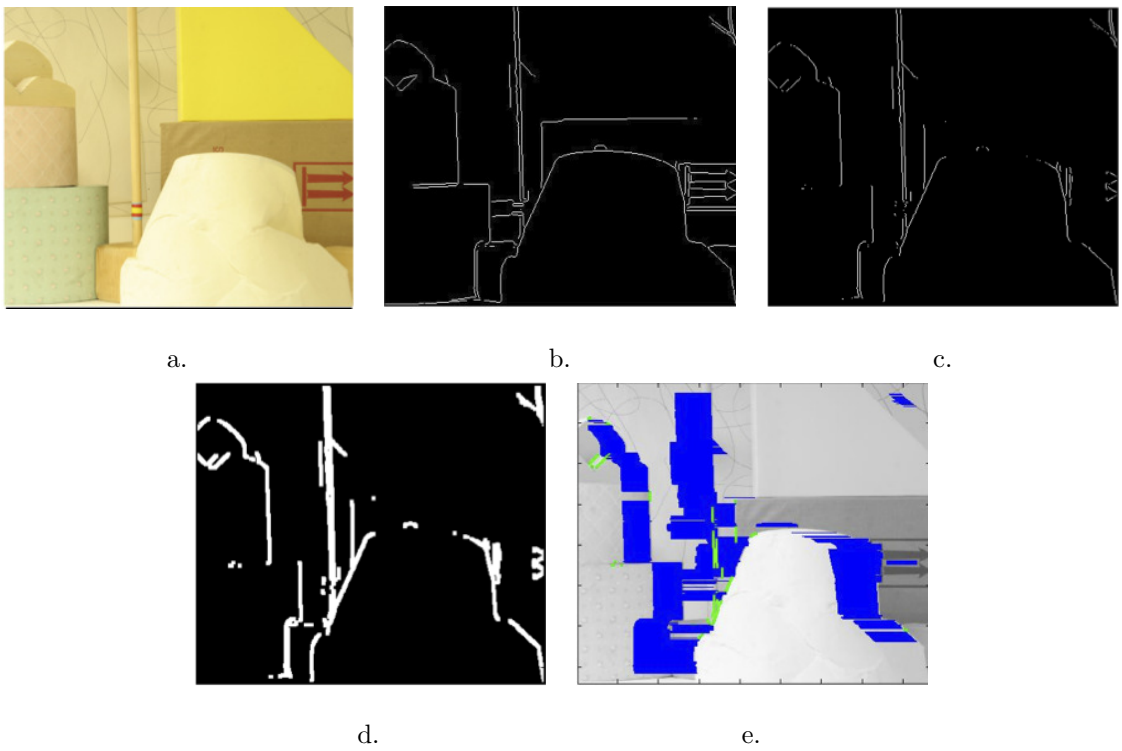


Figure 3.4. An example of edge extraction and determination steps together with the final sparse disparity map. Left image of a stereo pair is given in a and its edge map in b. Horizontally aligned edges are removed and dilated as in c and d, respectively. Estimated sparse disparity map is shown in e with blue lines representing the disparity vectors.

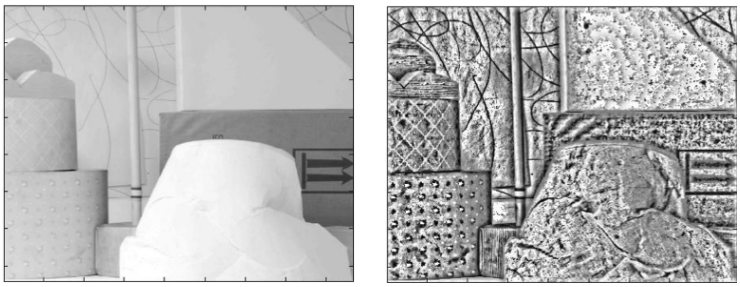
3.2.2. Rank Filtering

To improve the edge disparity field estimation, we preprocessed both (Left and Right) images to mitigate illumination artefacts and enhance the edge structure [22]. We used the rank filtering, proposed in [35], for the image intensity based matching methods. For sum of absolute differences of space invariant feature transforms (SAD-SIFT) and correlation of gradient orientations (CGO) algorithms (see Section 3.3), the rank filtering is not used, since this filtering decreases the dynamic range of the images and hence the extracted SIFT vectors and gradient orientations do not reflect the properties of the original images.

Through rank filtering, local intensity differences in images are emphasized and illumination differences between image pairs are minimized to some degree. It has been shown that rank filtering is an efficient method, which helps to improve the performance of local disparity estimation algorithms [22].

The rank filtering simply considers a window of size $W \times W$ around each pixel, rank orders the pixel values in the range 1 to W^2, and the rank of the pixel in that ordering is used as the new pixel value. This filtering is applied to both of the stereo images and the choice of $W = 15$ was found to be adequate as suggested in [22], so that original gray values are mapped to the range $[1, 225]$.

By applying this filtering to both of the image pairs, their rank filtered versions are obtained and in the disparity search stage they are used instead of intensity images. A gray scale image and its rank filtered version can be seen in Figure 3.5. It can be observed, that rank filtering exaggerates the image texture and this more detailed texture enables more accurate block matching.



a. b.

Figure 3.5. An example of rank filtering with a window size of 15×15. The gray scale image can be seen in a and its rank filtered version in b. Note that rank filtered image has more detailed texture.

3.3. Disparity Search

We will describe the disparity search using block matching and five different methods we assessed. Let us call f_R the reference image and f_T the target image. Obviously either left image or right image can be labeled as f_R.

In disparity search stage a block $f_R(x - h/2 : x + h/2, y - b/2 : y + b/2)$ is designated around an edge location with coordinates (x, y), $f_R(x, y)$ of the reference image f_R. A companion block is searched for in the target image f_T by placing the center of the block on positions $f_T(x, y - k : y + k)$ horizontally shifted. The search block is slid through the target image in a $2k+1$ pixel range symmetrically towards left and right with respect to the current position in f_R and a best matching block position is determined in the target image. When the left image is taken as the reference image, the disparities to the right are defined as *positive*, and the ones to the left as *negative*. The matching costs calculated for all relative displacements in the image plane, that is candidate disparities (d) , form a *cost profile* graph. The location of the minimum cost in the graph is taken as the disparity estimate.

There are two parameters to be set: the search block size, $N = h \times b$, determines the size of the search block, and $-k \leq d \leq k$ determines the length of the horizontal search range . The various methods for point disparity estimation described in the literature differ in the way matching similarity is computed and its aggregation over the search block. We considered five different methods for matching cost calculation and cost aggregation, which are Sum of Absolute Differences (SAD), Herman Weyl's Discrepancy Measure (HWDM), Adaptive Support Windows (ASW), Sum of Absolute Differences of Scale Invariant Feature Transform Vectors (SADSIFT) and Correlation of Gradient Orientations (CGO).

3.3.1. Sum of Absolute Differences (SAD)

SAD is a well studied matching cost calculation method for stereo matching and motion estimation tasks. SAD cost at shift d can be defined as:

$$SAD(d) = \frac{1}{N} \sum_{n=-h/2}^{h/2} \sum_{m=-b/2}^{b/2} |f_R(x+n, y+m) - f_H(x+n, y+m+d)| \; ; \\ -k \leq d \leq k \qquad (3.1)$$

where $N = h \times b$. The SAD value at candidate disparity d is obtained as the sum of the absolute difference of the two blocks from f_R and f_T, at d units horizontal shift from each other.

3.3.2. Herman Weyl's Discrepancy Measure (HWDM) [36]

HWDM is a similarity measure, which recently gained popularity for its usage in texture analysis tasks. HWDM uses the integral image concept. The pixel differences (not the absolute difference) of the two blocks are considered, these differences are integrated along four directions, namely, left to right and top to bottom, left to right and bottom to top, right to left and top to bottom and right to left and bottom to top (LRTB, LRBT, RLTB, RLBT). For example, in the left to right and top to bottom case, the integral image is obtained by summing the pixels first from left to right and then from top to bottom, in other word propagating first in horizontal direction and then in vertical.

Once the four integral images are obtained (each of size $h \times b$), the difference between maximum and minimum values in each integral image is calculated $\max_{x,y}(I_q) - \min_{x,y}(I_q)$, and finally the maximum among these four difference values from integral images is taken as the cost value. Accordingly HWDM cost at shift d can be defined as:

$$HWDM(d) = \max_{q} \left(\max_{x,y}(I_q) - \min_{x,y}(I_q)\right) \; ; \\ q \in \{LRTB, LRBT, RLTB, RLBT\} \qquad (3.2)$$

Here I_q represents the integral image obtained along the q-th direction. The minimum values of the integral images are subtracted from the maximum ones in order to constrain the final costs to be positive. The coordinate location that yields the minimum HWDM is taken as the disparity estimate. An example of HWDM cost calculation for a difference block of size 5×5 can be seen in Figure 3.6. Obviously, similar to SAD, HWDM cost is also obtained from the difference of two blocks, but HWDM cost reflects the maximum local dissimilarity of the blocks in four different directions.

5	-2	0	6	-1
10	6	-12	0	7
7	4	3	-2	-7
-5	12	10	-6	0
-4	-3	9	-6	0

a.

5	3	3	9	8
15	19	7	13	19
22	30	21	25	24
17	37	38	36	35
13	30	40	32	31

13	30	40	32	31
8	27	37	23	23
-2	11	33	19	12
-9	0	19	7	7
-4	-7	2	-4	-4

8	3	5	5	-1
19	4	0	12	6
24	2	-6	3	-1
35	18	-2	-3	-1
31	18	1	-9	-1

31	18	1	-9	-1
23	15	-4	-14	0
12	14	1	-21	-7
7	16	7	-12	0
-4	0	3	-6	0

b.

Figure 3.6. HWDM cost calculation example for a difference block of size 5×5. The difference block is given in a. Four integral images resulting from the difference block in a are given in b. By applying the formula in 3.2, the difference of maximum and minimum values from the integral images are 37, 49, 44 and 52, resulting 52 as the final cost value.

3.3.3. Adaptive Support Windows (ASW) [27]

In this method the search block is divided into five overlapping sub-blocks as shown in Figure 3.7. The final cost is determined from the individual costs of each sub-block. Inside the search block four overlapping sub-blocks with equal sizes and a smaller sub-block in the center are taken. The central smaller sub-block is one third in

size of the larger corner blocks. Accordingly, the final cost value at shift d, $ASW(d)$, is calculated by adding the two smallest SAD costs of the four corner sub-blocks to the cost value of the center sub-block:

$$ASW(d) = SAD_5(d) + \min_{i=1}^{4} SAD_i(d) + \text{second}\min_{i=1}^{4} SAD_i(d) \tag{3.3}$$

where $SAD_i(d)$ represents the SAD cost of the i^{th} sub-block at shift d.

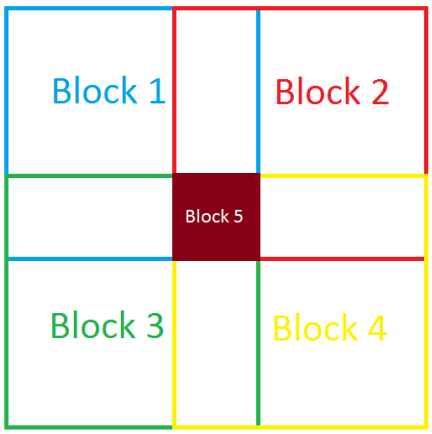


Figure 3.7. Overlapping sub-blocks in ASW. Blocks 1,2,3 and 4 are the support windows of block 5. Size of block 5 is one third of its supporting blocks. Final cost value is obtained by adding the sum of the two minimum costs from the support windows to the cost of block 5.

In ASW the block matching is performed with five overlapping windows instead of a single window. Since the ASW cost includes two of the minimum sub-block costs, the final disparity estimate depends not only to the center blocks cost, but also to its supporting sub-block costs.

3.3.4. Sum of Absolute Differences of SIFT Vectors (SADSIFT)

SIFT (Scale Invariant Feature Transform) vectors are local descriptors of the gradient information in images. In disparity search, we first extract the SIFT vectors at each pixel in both of the stereo image pairs f_T and f_R and obtain SIFT images, whose each pixel is a SIFT vector. Then we perform block matching between these SIFT images.

The SIFT vectors are obtained by dividing the 16×16 neighborhood of each pixel into 4×4 cells and then quantizing the gradient orientation in each cell into 8 bins [37]. Thus each pixel in both of the images is replaced with a SIFT vector of size 128 and the corresponding SIFT images are obtained. The disparity search consists of matching blocks between the reference and target SIFT image blocks using simply the SAD criterion. Note that this algorithm does not need the preprocessing step of rank filtering (Section 3.2.2), since SIFT already uses the image contrast in the neighborhood.

Since each pixel in SIFT images is a SIFT vector, it can be said that each pixel in a SIFT image contains the gradient orientation information around its corresponding pixel in the original image. Therefore block matching on SIFT images can be understood as matching local gradient orientations between image pairs.

3.3.5. Correlation of Gradient Orientations (CGO)

Similar to SADSIFT, CGO cost calculation is also based on matching of gradient orientations. In this method, first the complex gradient fields of the reference and target blocks are calculated. The weakest pixels, that is pixels with gradient magnitude below a given threshold Z are eliminated. The gradient orientation fields $\{O_R, O_T\}$ at the reference and target images, are calculated on the "stronger" pixels. Finally, the CGO cost is computed from SAD scores of the two orientation fields.The CGO cost at shift d is computed as:

(i) Complex gradients of both of the images ($f_i(x,y)$ where $i \in \{R,T\}$) are calculated:

$$G_i(x,y) = \nabla_x f_i(x,y) + j\nabla_y f_i(x,y) \tag{3.4}$$

(ii) Pixels with sum of absolute gradient values less than some threshold Z are eliminated ($G_i(x,y) \rightarrow G_i'(x,y)$):

$$|\nabla_x f_i(x,y)| + |\nabla_y f_i(x,y)| \leq Z \rightarrow 0 \tag{3.5}$$

(iii) Gradient orientation maps ($O_i(x,y)$) are obtained:

$$O_i(x,y) = \frac{G_i'(x,y)}{\|G_i'(x,y)\|} \tag{3.6}$$

(iv) The correlation between gradient orientations at shift d, ($CGO(d)$) is calculated by taking SAD between O_R and O_T:

$$CGO(d) = \frac{1}{N} \sum_{n=-h/2}^{h/2} \sum_{m=-b/2}^{b/2} |O_R(x+n, y+m) - O_T(x+n, y+m+d)| \; ; \quad -k \leq d \leq k \tag{3.7}$$

Note that the threshold Z is determined empirically. In our application we found it adequate to chose it as 5% of the maximum gradient magnitude of an image:

$$Z = \frac{\max_{x,y}(|\nabla_x f_i(x,y)| + |\nabla_y f_i(x,y)|)}{100} \times 5 \tag{3.8}$$

Recall that our disparity search is performed on edges and that CGO considers only strong gradients. Accordingly matching the strong gradient orientations can be understood as matching the edge orientations between image pairs.

It should also be noted that this algorithm also bypasses the rank filtering step due its use of gradients. Rank filtering exaggerates the textures in the images, resulting different gradient structures than the original image. Therefore the rank filtering does not help to improve disparity search performance, when matching of gradient information is under consideration.

3.3.6. Block Sizes

We determined the block sizes empirically. For fairness in the performance comparison of disparity estimation methods, we take the block size that yields the best results for each method. Thus we used the following hxb figures:

- The size of the search block is 25×25 in SAD and ASW.
- In ASW, the four overlapping sub-blocks are of size 15×15 and the middle sub-block is of size 5×5.
- For HWDM the block size is 11×11
- For SADSIFT the block size is 5×5
- For CGO the block size is 15×15.

Note that SADSIFT has smaller block size compared to others, but since each pixel is a 128 dimensional vector, actual block size is $5 \times 5 \times 128$ in SADSIFT. Also recall that each SIFT image pixel contains the gradient orientation information of its 16×16 neighborhood in the original image.

3.4. Post-processing of Edge Disparity Field

In order to refine and correct the sparse disparity map we considered two post-processing approaches, which are *cross-checking* and *disparities in a band*.

3.4.1. Cross-checking

The reliability of the disparity estimates are investigated by performing cross-checking. Using this method, unreliable disparity estimates are eliminated from the sparse disparity map and their locations on the disparity map are labeled as *unreliable*.

In the disparity estimation stage, if a block at some position (x, y) in f_R finds its match in f_T at $(x, y + d^E)$, where d^E is the estimated disparity, then one searches for matching block in f_R this time starting from the reference point in f_T. From this search another estimate d^C is obtained, which is used for checking the reliability of the initial estimate d^E. If the two estimations d^E and d^C are consistent with each other, then the initial estimation is considered as valid. If they are not consistent, then the pixel under investigation is labeled as *unreliable* in the disparity map. The two estimates are accepted as consistent with each other, if the difference between them is equal or less than 2 pixels. This process can be summarized as:

- If $|d^E - d^C| \leq 2$, then d^E is valid.
- If $|d^E - d^C| > 2$, then d^E is *unreliable*

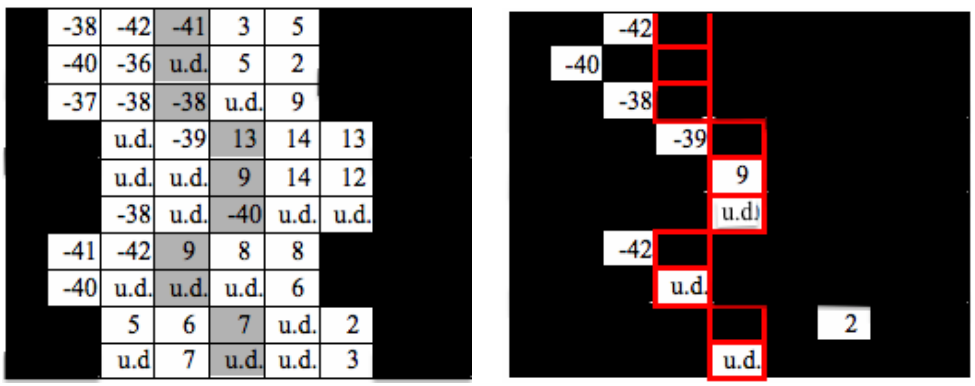
Ideally it would be expected to have both disparity estimates be equal to each other. However, in reality due to sampling and illumination differences between image pairs, expecting both estimates to be equal to each other would result with elimination of most of the disparity estimates from the sparse disparity map. Therefore a difference of 2 pixels is allowed to relax the decision rule.

Cross-checking is useful in determining the occluded regions and hence helps to reduce wrong disparity estimates. Since *unreliable* estimates are taken out of the dis-

parity map, this approach yields fewer disparity estimates, but since we are already considering the sparse disparity maps, obtaining fewer but correct estimates is preferred.

3.4.2. Disparities In A Band

A disparity estimate at an edge pixel is chosen from among the s estimates in the band around the edge. Recall that edges were dilated horizontally (Section 3.2.1) by s pixels (typically, $s = 5$) via morphological operators resulting in s disparity estimates at and around each edge pixel. This many-to-one mapping helps to eliminate quite a number of *unreliable* (non-validated) disparities by choosing a valid estimate from the disparity band they are in. It also helps to correct the edge disparity field further. As shown in Figure 3.8 there can be *positive*, *negative* and *unreliable* disparities inside the s-wide band surrounding an edge point.



a. b.

Figure 3.8. An example of disparities in a band post-processing method. The disparity values over the edge bands are given in a and the chosen disparities are given in b. The gray squares in a and red squares in b represent the original edge before dilation. *Unreliable* disparities are marked as u.d.

Along the depth of a scene, the disparity values are expected to range from the smallest *negative* values in the foreground to the largest *positive* values in the background, when left image is taken as the reference. Accordingly, choosing the smallest signed disparities across the band means choosing the disparities that belong

mostly to foreground objects. Since large disparities are typically associated with foreground objects, this processing step is consistent with our goal of efficient sparse disparity map estimation along object edges and is helpful in resolving some of the ambiguities. The s-wide band around each edge pixel, is processed according to the following rule (Figure 3.8):

- If #unreliable estimates $< s/2$, then chose minimum disparity
- If #unreliable estimates $> s/2$ and center pixel is *reliable*, then chose minimum disparity
- If #unreliable estimates $> s/2$ and center pixel is *unreliable*, then leave as *unreliable*

According to this, if the number of *unreliable* estimates is more than the half width of the band and the center pixel of the band is *unreliable* , then the final disparity value is left as *unreliable*. In other cases, the final disparity value is determined by choosing the minimum signed disparity inside the band.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

We describe here briefly the stereo image database, video test material and the metrics used in experiments.

4.1.1. Stereo Image Database

For numerical evaluation of our disparity estimation methods, we used 35 image pairs from the Middlebury stereo image database [19, 20]. This standard database contains dense ground-truth disparity maps for each image pair. The image pairs are known to be rectified, such that all points in an image have their corresponding matches in the other image on the same horizontal coordinates.

The image pairs are known to have only unidirectional disparities, such that all disparity vectors are expected to point to the same direction. As we mentioned in previous chapters, we consider the disparity vectors pointing to the left as *negative* and the ones pointing to the right as *positive* disparities. For all image pairs in the database *negative* disparities are expected, when the left image is taken as the reference.

True maximum disparity values occurring in Middlebury database are as follows: For 6 images the absolute maximum disparities are less than 20 pixels ($\approx 5\%$ of the scene) and in the remaining 27 images absolute maximum disparities vary between 40 to 80 pixels ($\approx 10\% - 20\%$ of the scene). Accordingly, in all of our experiments we took the disparity search range $k = 80$ ($\approx 20\%$ of the scene), such that the disparity search range is between $(x, y - 80)$ and $(x, y + 80)$ for any pixel at location (x, y).

4.1.2. Video Test Material

In addition to the automatic method in Figure 3.1, we also considered a subjective evaluation method for our maximum disparity estimation algorithm. For this purpose we compiled a test video set consisting of 12 different stereo scenes from the footages provided by a digital broadcasting company (Digitürk A.Ş.). Eight of these footages are taken in a football stadium by an expert 3D broadcasting crew, one of them is a computer animation and three of them are taken in public locations around the city. Any information regarding the true depth maps of the scenes is not provided. Sample frames from each of these scenes can be seen in Figure 4.1.

The stereo shots have a frame rate of 25 frames per second and have different durations that range from, about 1 minute to 15 seconds. There is a 1 second length black screen between the scenes. The total length of the test video is 9963 frames ($\approx$ 6.5 minutes).

The original resolution of the videos were 1080×1920 pixels, but we down sampled them to 270×480 for processing with our algorithm. The videos were not rectified, but since their vertical disparity was rather slight, the horizontal only search remained still a good option. The stereo shots are taken with a slight angle between the cameras. Therefore they can contain bidirectional disparities, such that when the left image is taken as reference, then the resulting disparities for background objects can be expected to be *positive* and for foreground objects as *negative*.

Some of these video shots contain large disparities, as the offset between the cameras was intentionally and randomly modulated during the shootings. Although these video scenes do not have ground-truth information, with a careful observation one can roughly estimate the offset between the stereo pairs. We determined the disparity search limit $k = 80$ ($\approx$ 17% of the scene), after observing the scenes with maximum camera offset.

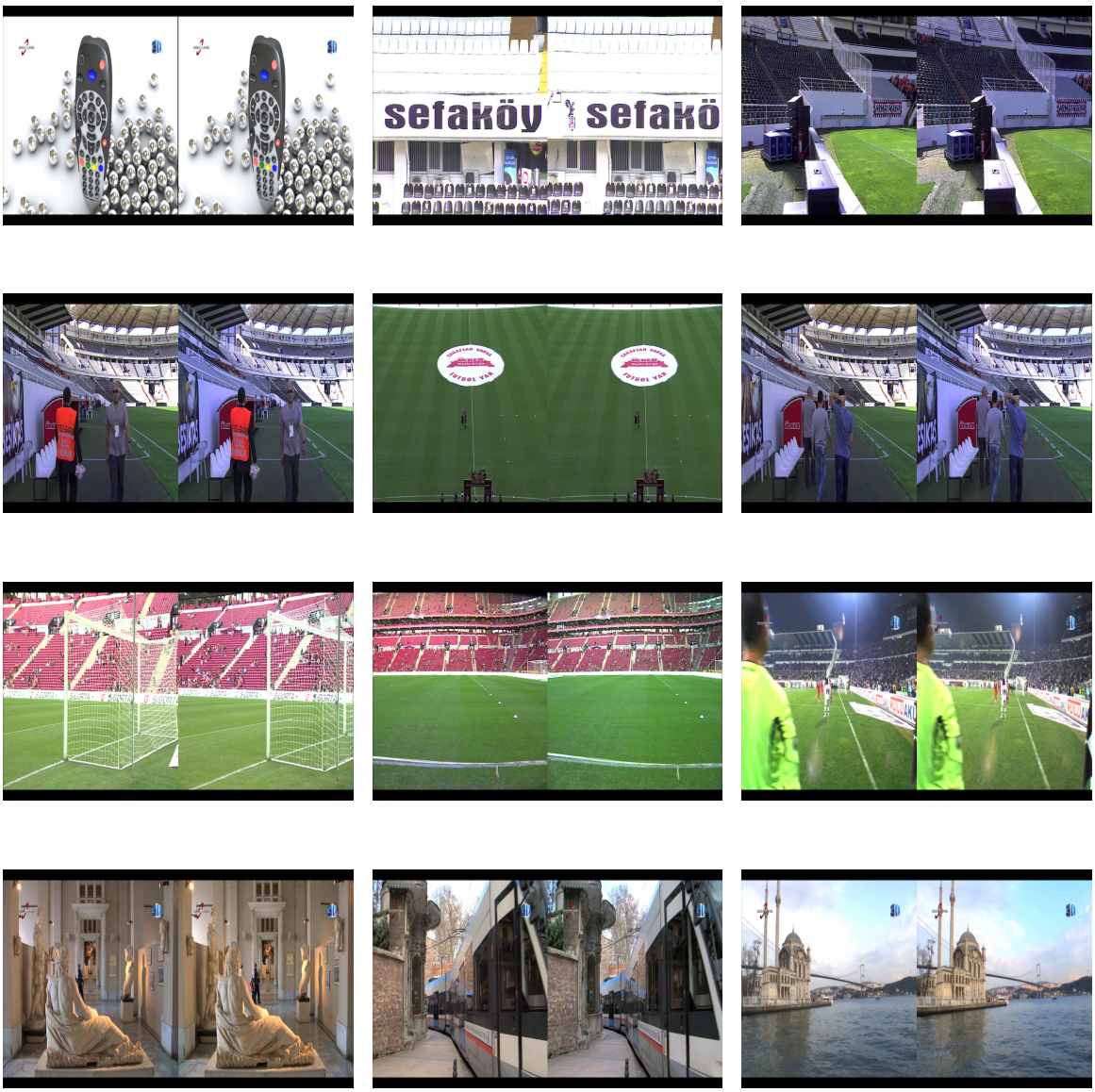


Figure 4.1. Sample frames from each of the 12 scenes in the test video.

4.1.3. Performance Metrics

Our goal in performing point disparity estimation is detecting the largest disparities in the scene and by using them, determining the discomfort level of the audiences watching the scene. Therefore we have developed performance measures to this effect.

We know that disparities, causing discomfort, if any, to the viewer will be among the largest ones. Our experience has also shown that the absolute estimation error prone of the disparities is proportional to the size of the actual disparity, such that as the distribution of disparities gets broader, the more challenging the disparity estimation problem becomes and the likelihood of erroneous disparity estimation increases.

We therefore computed the mean of the largest 5% of the true disparities, the 95-percentile mean, $\mu 95$, for each image and use it as an indicator of expected estimation error. In fact, we rank the Middlebury images according to their $\mu 95$ scores, from small mean maximum disparity to large mean maximum disparity, and plot the disparity estimation performance as a function of $\mu 95$. The $\mu 95$ value for an image pair is calculated by sorting the disparity values in its ground truth disparity map, and then calculating the mean of the upper 5% portion of this sort.

We considered three different criteria for comparing the performances of five different cost calculation methods (see Section 3.3). These criteria are:

- Percentage of Erroneous Disparities ($Erroneous\%$)
- 95 Percentile Absolute Error ($Diff95\%$)
- 95 Percentile Ratios ($Ratio5\%$)

The first criterion (percentage of erroneous disparities), gives us a measure about the accuracy of the disparity estimation, while other two criteria are about the accuracy of the estimation of maximum disparities.

4.1.3.1. Percentage of Erroneous Disparities (*Erroneous*%). In this error criterion, we consider a disparity estimate as erroneous, if it exceeds 10% of its corresponding ground true value. If d^G is the ground truth and d^E is the estimated disparity, then the test $T(d^G, d^E)$ is expressed as:

$$T(d^G, d^E) = \begin{cases} 1 & \text{if } |d^G - d^E| > R(d^G) \\ 0 & \text{if } |d^G - d^E| \leq R(d^G) \end{cases} \tag{4.1}$$

where $R(.)$ is an operator, which takes 10% value and round to the nearest integer value, except that the values smaller than 0.5 are rounded to 1 instead of 0. $R(d)$ can be defined as:

$$R(d^G) = \begin{cases} \lfloor \frac{d}{10} \rceil & \text{if } \frac{d}{10} \geq 0.5 \\ 1 & \text{if } \frac{d}{10} < 0.5 \end{cases} \tag{4.2}$$

Accordingly, $T(d^G, d^E) = 1$ means d^E is an erroneous estimate, and $T(d^G, d^E) = 0$ means the estimate is valid according to this error criterion.

The mapping between the disparity value and tolerable absolute disparity errors is given in Figure 4.2. Notice the staircase behavior due to rounding, effectively quantization.

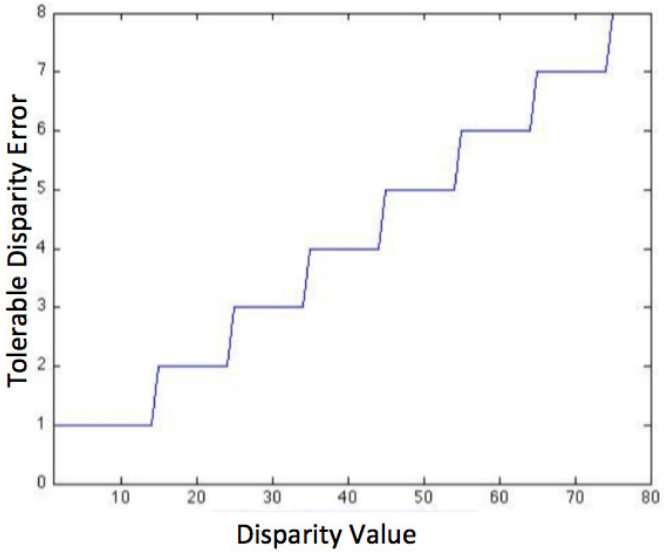


Figure 4.2. 10% criterion. An estimate, whose distance to the true value is below the curve, is taken as erroneous. The staircase behavior is due to round-down.

This criterion tolerates errors in proportion to the actual disparity size; for example, the tolerance for disparity values $d^G < 15$ is 1, for $15 < d^G < 25$ is 2 , while larger disparities allow for larger errors. A disparity estimation figure of merit can be obtained for the whole image in terms of the percentage of erroneous disparity estimates vis-à-vis the total number of reliable disparity pixels, C, at which disparity is estimated.

$$Erroneous\% = \frac{\sum_{i=1}^{C} T(d_i^G, d_i^E)}{C} \times 100 \tag{4.3}$$

While determining the error percentages, we exclude the *unreliable* disparities. Therefore $Erroneous\%$ values are representing how many of determined disparities over an image are erroneous, according to 10% error criterion. Note that this metric gives scores in $[0, 100]$ range, with 0 corresponding to perfect estimation.

4.1.3.2. 95 Percentile Absolute Error ($Diff95\%$). To put into better evidence the disparity performance at larger values, for each image we calculate the 95 percentile of disparities. In other words, we rank the disparities in ascending order and take the pixels corresponding to the highest 5% of disparities, and then calculate the disparity errors at this particular cut point.

More specifically, consider the rank ordered ground-truth disparities $\{d_1^G, d_2^G, .., d_C^G\}$ and rank ordered estimated disparities $\{d_1^E, d_2^E, .., d_C^E\}$, and consider the threshold value v for the largest 95% disparities: $v = \lfloor 0.95 \times C \rfloor$. Then one has:

$$Diff95\% = |d_v^G - d_v^E| \tag{4.4}$$

where the notations d_v^G and d_v^E signify the v^{th} rank ordered true and estimated disparities.

This criterion yields the absolute discrepancy between the 95 percentile values of the ground-truth and estimated disparity sets. Obviously, the range of this metric is between 0 for perfect estimation and k, the largest attainable error.

4.1.3.3. 95 Percentile Ratios ($Ratio5\%$). In this scheme, the disparities are again sorted as in 95 Percentile Absolute Error, but we consider the ground-truth and estimated disparities in the last 5 percentile (between 95 and 100 percentiles) sets, respectively. We then consider the ratio of the means of the disparities d^E and d^G. If this ratio is close to 1, then there is a good agreement; ratio scores above 1 means that the algorithm overestimates the disparities, and the ones below underestimates the disparities. This criterion can be expressed as:

$$Ratio5\% = \frac{\frac{1}{C-v}\sum_{i=v}^{C} |d_i^E|}{\frac{1}{C-v}\sum_{i=v}^{C} |d_i^G|} \tag{4.5}$$

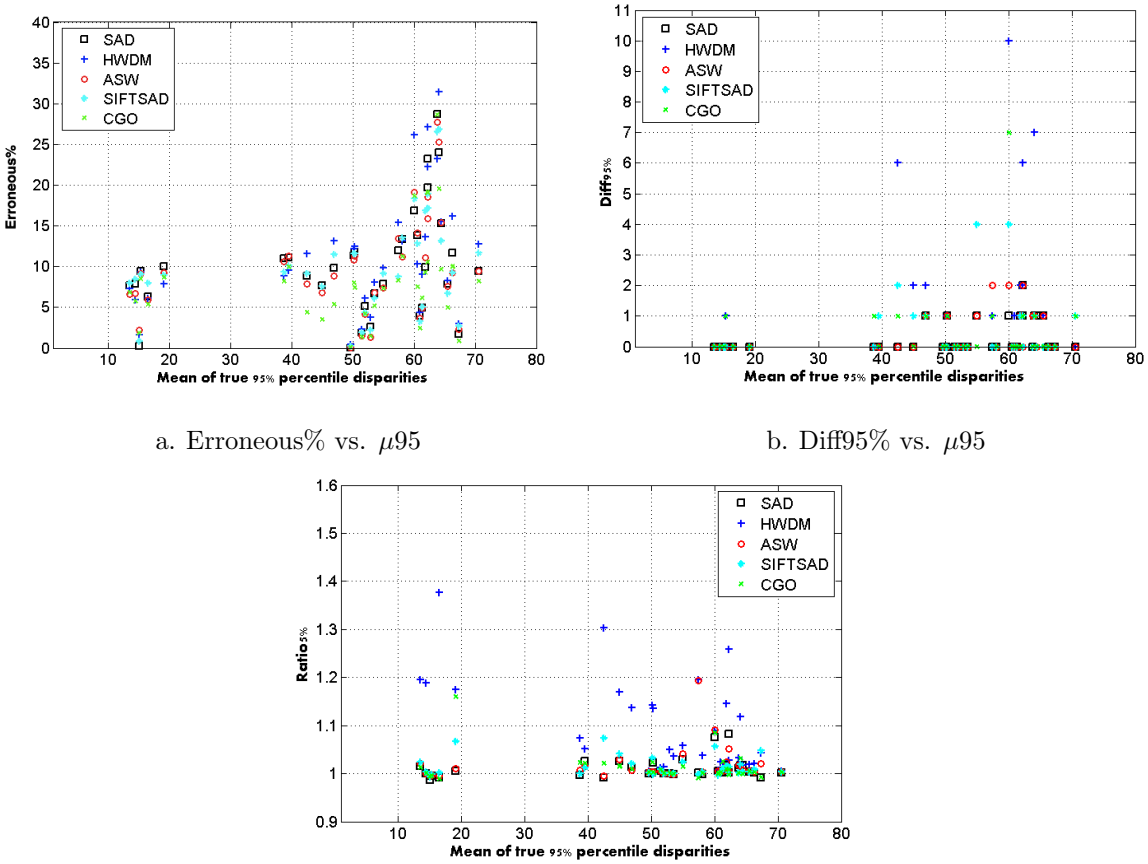
where d_i^G and d_i^E signify the i^{th} rank ordered true and estimated disparities; C is the total number of disparity estimates and $v = \lfloor 0.95 \times C \rfloor$ is the threshold.

4.2. Performance Results on Stereo Images

In this section comparative performance results on the Middlebury images are presented in order to compare the capability of the estimators for large disparity values. For each image pair the sparse disparity maps are obtained, and then the performances of SAD, HWDM, ASW, SADSIFT and CGO cost calculation methods are evaluated. The performance evaluation is performed according to the $Erroneous\%$, $Diff95\%$ and $Ratio5\%$ performance metrics. For each image pair these three values are obtained with five different cost calculation methods.

In Figure 4.3 these results are given in three scatter plots (one for each performance metric), where each dot on the plot represents the result for an image pair. The abscissa is the $\mu 95$ values associated to the image pairs, according to their true maximum disparities, and the ordinates are the $Erroneous\%$, $Diff95\%$ and $Ratio5\%$ values.

Figure 4.3.a shows the scatter plot of disparity estimation errors, where the abscissa is the $\mu 95$ value of the images and the ordinate is the percentage of unsuccessful disparity estimates ($Erroneous\%$). We observe, first, that the disparity estimation error is strongly correlated with the $\mu 95$ coefficient, such that the error is higher for larger $\mu 95$ values. This observation is related to the fact that disparity estimation becomes a harder task when disparity range of the images are larger. Secondly, it can be observed that CGO performs better, since it mostly yields lower $Erroneous\%$ values.



a. Erroneous% vs. $\mu 95$

b. Diff95% vs. $\mu 95$

c. Ratio5% vs. $\mu 95$

Figure 4.3. Comparative performance results of SAD, HWDM, ASW, SADSIFT and CGO for 35 image pairs from the Middlebury dataset. Each dot in the graphs represents the result for an image pair. Image pairs are ordered along the horizontal axes according to their mean of maximum 5% ground truth values ($\mu 95$).

The performance comparison of the different methods according to their *Erroneous*% values is summarized in Table 4.1, where the rank sums for each method is given. For any one method, Rank 1 corresponds to the number of images, where it has performed best, and Rank 5, where it has performed the worst. Again it can be observed, that CGO performs better than others with 22 of 35 image pairs in Rank 1. HWDM performs worst with 22 image pairs in Rank 5.

Table 4.1. Rank sums for each method according to their Erroneous% values.

	Rank 1	**Rank 2**	**Rank 3**	**Rank 4**	**Rank 5**
SAD	4	3	9	14	5
HWDM	3	3	4	3	22
ASW	6	11	9	6	3
SADSIFT	0	12	9	9	5
CGO	22	6	4	3	0

The other two graphs confirm these results: Figure 4.3b and 4.3c show, respectively, absolute maximum disparity estimation error and relative maximum disparity scores. It can be seen that the performance ranking follows the same trend.

With *Diff95%* metric, it is desired to have values close to 0 as much as possible, since this metric is defined as a difference between estimations and true values. In Figure 4.3b it can be observed, that CGO satisfies this criterion well enough, such that *Diff95%* value is 7 pixels for only a single image pair and besides the difference becomes at most 1 pixel.

Similarly *Ratio5%* metric it is expected to have values close to 1 in the ideal case, since this metric is defined as a ratio of estimated and true maximum disparities. Figure 4.3c reflects the results for this metric, and it can be observed that CGO again performs best, since most of its *Ratio5%* values are close to 1. With CGO, *Ratio5%* value is 1.16 for a single image pair and excluding that, the ratio varies between 0.99 and 1.09.

Interestingly enough, all methods overestimate the disparity as $\frac{|d_i^E|}{|d_i^G|} > 1$ in almost all cases (Figure 4.3c). That means the methods we applied do not miss large disparities, but there is a chance to give false alarms by estimating the maximum disparity larger than it actually is. It can be observed, that HWDM is more prone to give false

alarms, since it overshoots more the ideal ratio of 1. CGO has a lower chance of giving false alarms, since its *Ratio5%* values are closer to 1.

Some sample image pairs from the Middlebury database and related sparse disparity maps are given in Figure 4.4. The disparities are represented as disparity arrows on the gray scale left images, such that blue lines are describing the negative disparities and red lines are describing positive disparities with each line being scaled according to its related disparity value. Green dots on the images represent the *unreliable* disparities. Recall, that for the Middlebury images all disparities are expected to be negative valued when left image is taken as reference.

For the image pairs "Bull" and "Venus" , the disparity maps we obtained show the expected disparity characteristics. The disparity vectors are all negative and short as their true disparity maps suggested. Their disparity ranges are less than 20 pixels. As we observed in Figure 4.3a the disparity estimation errors of all of the investigated methods are low for image pairs with small disparity ranges. Therefore the disparity maps obtained with different methods look similar to each other for "Bull" and "Venus" image pairs, which have small disparity range.

However on the visualized sparse disparity maps of the "Lambshade2" image pair the performance of different methods are more accentuated. The broad disparity range of this image pair (about 65 pixels) results higher disparity estimation error rates and therefore enables the errors to be more clearly observable. It can be seen, that there are some false positive disparities found with HWDM and the disparities are not changing smoothly along the edge profiles. Among the results of the five methods, the disparity maps estimated with SADSIFT and CGO seem to be the most accurate ones, since no false positive disparity is present and the disparities seem to be changing smoothly along the edges.

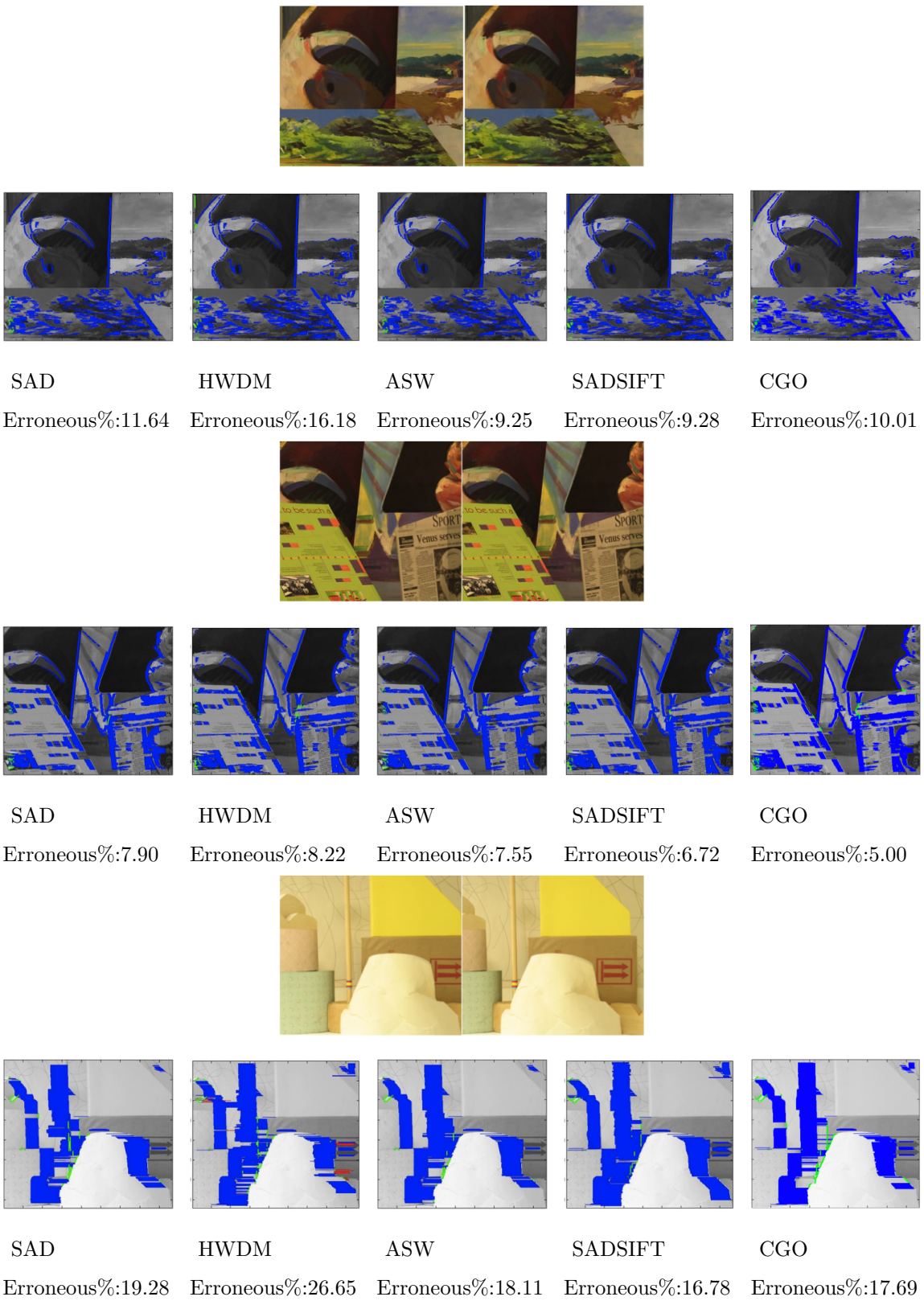


Figure 4.4. "Bull", "Venus" and "Lambshade2" image pairs from the Middlebury dataset and their sparse disparity maps. Blue arrows: negative disparities. Red arrows: positive disparities. Green dots: unreliable disparities.

4.3. Disparity Estimation Complexity

The algorithms we considered are implemented in MATLAB and tested on a computer with 2.66 GHz Intel Core Duo CPU. Table 4.2 shows the average processing times for Middlebury database images. The speed of the various stages of the algorithms are tested with same parameter settings given in previous chapters.

Table 4.2. Average processing times for various stages of the algorithms.

Process	**Time (ms)**
Rank filter (only with SAD, HWDM, ASW)	6250
SIFT vector extraction (only with SADSIFT)	71842
Canny Edge Detector	460
Edge elimination and dilation	405
Disparities in a band post-processing	147
Disparity search and cross-checking with SAD	$\approx 92.7 \times 10^3$
Disparity search and cross-checking with HWDM	$\approx 528.6 \times 10^3$
Disparity search and cross-checking with ASW	$\approx 366.8 \times 10^3$
Disparity search and cross-checking with SADSIFT	$\approx 443.0 \times 10^3$
Disparity search and cross-checking with CGO	$\approx 77.8 \times 10^3$
Overall for SAD	$\approx 100.0 \times 10^3$
Overall for HWDM	$\approx 535.9 \times 10^3$
Overall for ASW	$\approx 374.1 \times 10^3$
Overall for SADSIFT	$\approx 515.9 \times 10^3$
Overall for CGO	$\approx 78.8 \times 10^3$

As expected, the highest processing times are observed in disparity search stage. It should be noted that the processing speed in disparity search stage highly depends on the number of detected edge points in the images, such that it takes longer to process images with high number of edge points. Note that the values in Table 4.2 reflect the

average processing times, therefore for image pairs with more edges, processing times can be longer and for image pairs with less edges they can be shorter than given.

Excluding the disparity search stage, the process with highest complexity is the SIFT vector extraction. In order to obtain the SIFT vectors, local gradients around each pixel must be calculated and quantized into orientation bins, therefore a high computational time is observed in this stage. Other pre- and post-processing steps require simpler computations and hence they have low complexity.

Computational times for disparity search stages are given together with the times required for cross-checking, since cross-checking stage requires similar computational measures as in disparity search. Accordingly it can be observed that CGO has lowest complexity in disparity estimation stage and in overall. For CGO the preprocessing stages such as rank filtering or SIFT vector extraction are not necessary. Also the size of the search block used in CGO is smaller compared to the size of the search block in SAD, which yields shorter computational times for CGO in contrast to SAD.

HWDM and SADSIFT are the methods which have highest computational complexity. Disparity search stage in SADSIFT takes shorter computation time compared to HWDM, but in overall both of these methods have the longest computation times, higher than 500×10^3 ms. Computation of integral images in HWDM require many additions, which may be the main reason for high computational times. In fact for a difference image of size $h \times b$, computation of four integral images takes $4 \times 2 \times (h-1) \times (b-1)$ additions, where in SAD the number of additions required is simply $h \times b - 1$. In SADSIFT, the cost calculations are performed with SAD, but the block sizes are $h \times b \times 128$, since each pixel is represented with a SIFT vector of size 128. This requires $h \times b \times 128 - 1$ additions and hence a long computational time.

4.4. Subjective Assessment of Stereo Scenes

In order to determine how many of the viewers experience visual discomfort, while watching stereoscopic videos, we ran a subjective evaluation test, using the stereo video data described in Section 4.1.2. In order to compare the subjective test results with the results of our algorithm, we considered the d_v^E (95% of the sorted disparities) scores of the CGO algorithm for each frame. This metric is expected to reflect the maximum disparity characteristics of the scenes, and hence we are investigating if d_v^E scores and viewer depth discomfort are correlated. The depth discomfort refers to the visual discomfort caused by excessive depth levels (disparities) in the scenes. The depth discomfort levels of the viewers can be predicted using d_v^E scores, if there is a correlation between these scores.

The d_v^E scores of the CGO algorithm on the whole set of video data (12 scenes resulting in 9963 frames) are plotted in Figure 4.5. The ordinate corresponds to the 95% sorted disparity score (see Section 4.1.3) for each frame, and the abscissa is the frame index. Also some example frames from the test video can be observed in Figure 4.6 together with their sparse disparity map visualizations.

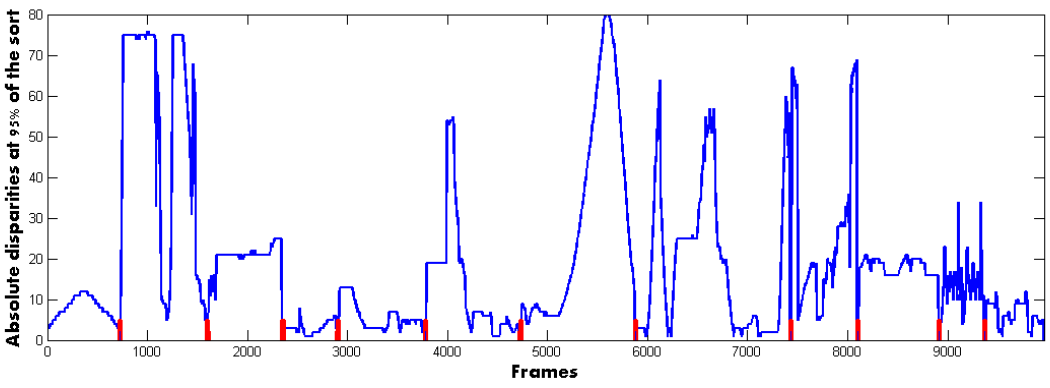


Figure 4.5. Results of CGO algorithm on video test data. Abscissa is the frame number and the ordinate is the disparity value at the 95% of the sorted disparities (d_v^E). Red markers indicate the transitions between scenes.

Figure 4.6. Sample frames from the test video and their related sparse disparity maps obtained with CGO. Sample frame numbers from top to bottom: 1617, 2526, 7712, 8110, 9178. Blue arrows: negative disparities. Red arrows: positive disparities. Green dots: unreliable disparities. Note that some frames contain bidirectional disparities.

In Figure 4.5 it can be observed that some scenes contain very high disparities and they are most likely to cause visual discomfort. Also it can be observed that transitions from low disparities to high disparities can be sharper or gradual depending on the modulations of angle and distance between cameras during shooting.

The subjective tests are performed by having 15 subjects watch the test movie on a commercial 3D TV. The subjects wore appropriate 3D shutter glasses, while watching the scenes. They watched the videos at a distance of 2 meters in a low lit room. After watching each scene two questions were asked to the subjects:

(i) A question on the experienced eye strain: "Did you feel any strain on your eyes at any part of the scene? "
(ii) A question on the overall 3D perception quality: "Did you experience the scene as a single 3D view or were they parts, where the 3D illusion was broken?"

The tally of subjects, who reported eye strain or 3D perception defects for each scene is given in Table 4.3, where the estimated maximum 95% scene disparities appear in the top two rows. Maximum 95% scene disparities are the maximum d_v^E values in each scene. They are also given as a percentage of frame resolution so as to take into account any down sampling effect. Following the maximum 95% scene disparities, the number of subjects who reported eye strain or complaints about 3D perception of the scene are given for each scene. The number of subjects, who reported at least one of the both visual discomfort types is shown in the last column.

It can be observed that for scenes with the largest disparity values, almost all of the subjects have reported eye strain, 3D perception defect or at least one of the both visual discomfort types.

Table 4.3. Maximum 95% scene disparities and subjective test scores for the 12 scenes on 15 subjects.

Scene	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Maximum 95% scene disparities (in % of the resolution)	2.5	15.8	5.2	1.7	2.7	11.5	16.7	13.3	14.4	4.4	7.1	2.5
Maximum 95% scene disparities (in pixels)	12	76	25	8	13	55	80	64	69	21	34	12
#Subjects with eye strain	1	13	2	0	1	10	14	9	8	1	6	2
#Subjects with 3D perception complaints	3	15	4	1	4	14	10	8	7	0	2	0
#Subjects with at least one of both	4	15	4	1	5	15	15	12	11	1	8	2

In order to observe the correlation between maximum disparities and the subject scores for each scene, we performed linear regression analysis, such that a high correlation would suggest that the number of subjects with visual discomforts can be predicted from the disparity statistics in Figure 4.5. Accordingly we considered three different predictors:

- Maximum 95% scene disparities
- Standard deviation of 95% scene disparities
- Slew rate (maximum slope) of 95% scene disparity changes

Standard deviation of 95% scene disparities are calculated for each of the 12 scenes. Standard deviation within a scene gives us a measure about how the 95% disparity values are distributed within the scene. If the disparities are changing much, then they are more likely to cause depth discomforts. Standard deviation of 95% disparities for a single scene ($\sigma\{ps, pe\}$), which starts at frame ps and ends at frame pe, is calculated as:

$$\sigma\{ps, pe\} = \sqrt{\frac{1}{pe - ps} \sum_{i=ps}^{pe} (d_v^E(i) - \mu_d)^2} \tag{4.6}$$

where $d_v^E(i)$ is the 95% disparity of the i^{th} frame and μ_d is the mean of the 95% disparity values between ps and pe:

$$\mu_d = \frac{1}{pe - ps} \sum_{i=ps}^{pe} d_v^E(i) \tag{4.7}$$

We also considered the slew rate of the 95% disparities in Figure 4.5 in order to investigate the correlation between sudden disparity changes between frames and the subject scores. A sudden change of disparities may distort the perception of depth more compared to gradual disparity changes between frames, since the human visual system can not adjust to sudden depth changes. Slew rate of 95% disparity changes are computed again for each of the 12 scenes. Slew rate for a single scene ($SR\{ps, pe\}$), starts at frame ps and ends at frame pe, is calculated as:

$$SR\{ps, pe\} = \max_{ps,\ldots,pe} (|d_v^E(pn + 1) - d_v^E(pn)|) \tag{4.8}$$

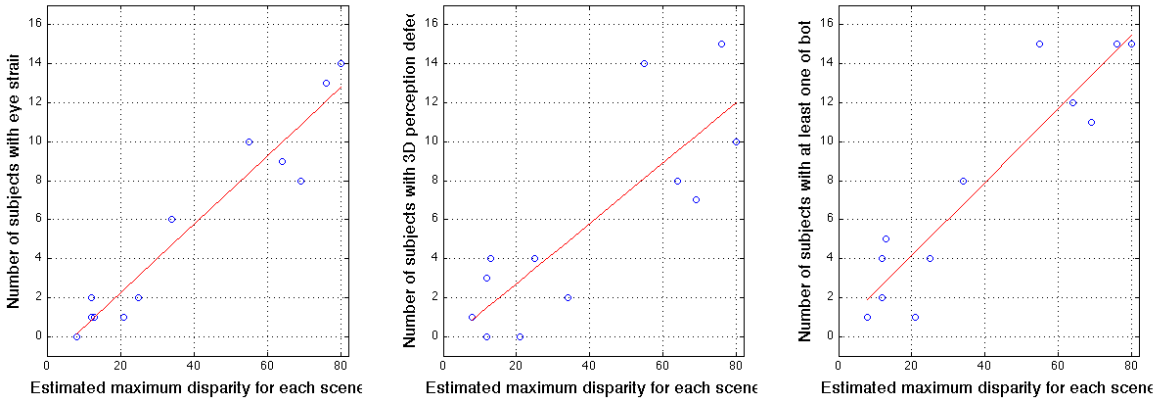
where $pn \in \{ps, ps + 1, \ldots, pe - 1, pe\}$ and $d_v^E(pn)$ represents the 95% disparity value at frame pn.

For each of these three predictors, we considered single variable linear regression. We also performed multivariable linear regression, by using these predictors together such as:

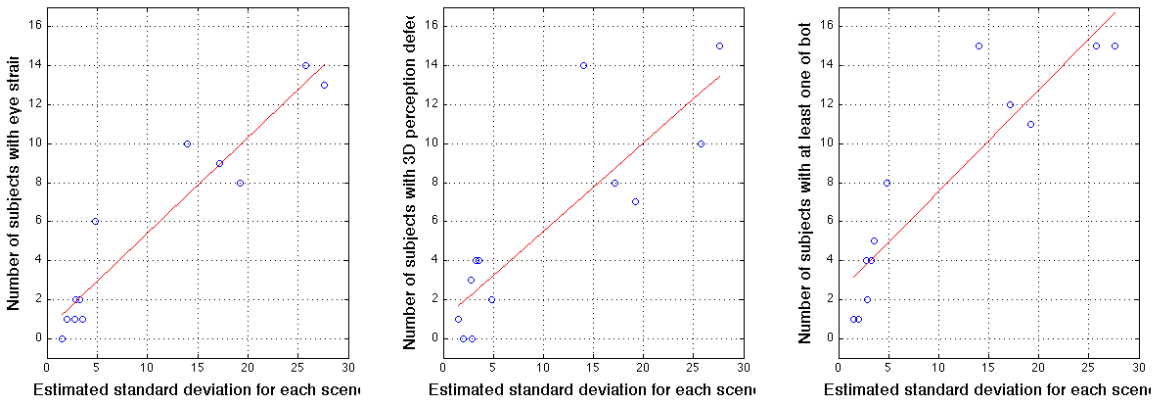
- Maximum disparity and standard deviation of 95% scene disparities as predictors
- Maximum disparity and slew rate of 95% scene disparities as predictors
- Maximum disparity, standard deviation and slew rate of 95% scene disparities as predictors

The single variable regression plots of predicted number of subjects experiencing discomfort as as function of these three predictors are given in Figure 4.7. Note that in single variable regression the regression plots are two dimensional and hence easier to visualize. Visualization of multivariable regression plots is hard if two predictors are available or impossible if more than two predictors are present. Therefore only single variable regression plots are given here.

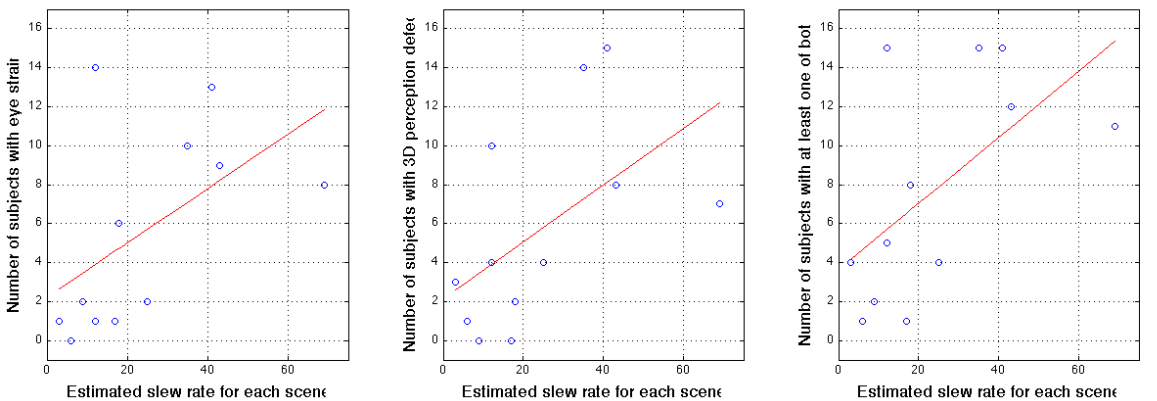
The single variable regression results in Figure 4.7 show that maximum disparity and standard deviation are better statistics compared to slew rate for prediction of number of subjects in all three types of subject scores. In Figure 4.7a and Figure 4.7b it can be observed that the distance between regression plots and the data points are smaller compared to regression plots in Figure 4.7c, meaning that the correlation between slew rate statistic and the number of subjects with discomfort is lower compared to maximum disparity and standard deviation statistics.



a.



b.



c.

Figure 4.7. Single variable linear regression results for predicting number of subjects with discomfort from: maximum 95% scene disparities in a; standard deviation of 95% scene disparities in b; slew rate of 95% scene disparity changes in c. Number of subjects with eye strain, 3D perception complaints and at least one of the both have been predicted separately with each predictor.

In order to compare the single and multiple variable regression results more precisely, we considered to measure the prediction error as mean absolute error (MAE) between original and predicted subject scores. MAE is computed by taking the mean of absolute differences of original and predicted scores over 12 scenes. The MAE scores can be seen in Table 4.4 for each type of subject score and different regression variable setups.

Table 4.4. Single and multiple variable regression results in terms of mean absolute error (MAE).

	Eye Strain	**3D Perception**	**At least one of both**
Maximum disparity	1.0725	2.3406	1.5450
Standard deviation	1.1415	1.9975	1.7070
Slew rate	3.3606	3.3786	3.5799
Maximum disparity + Standard deviation	0.9710	2.1827	1.5623
Maximum disparity + Slew rate	0.8032	2.3386	1.4940
Maximum disparity + Standard deviation + Slew rate	0.8092	2.1633	1.5178

The regression results in Table 4.4 shows that maximum disparity is the best single statistic of viewing discomfort for predicting both the number of subjects with eye strain and the number of subjects experiencing at least one of the both discomfort types. For prediction of number of subjects with 3D perception complaints, standard deviation is the best single statistic. Slew rate is the worst single statistic for prediction of any of the discomfort types.

Among multivariable regression cases, using maximum disparity and the slew rate together give the best predictions of the number of subjects with eye strain and the number of subjects with at least one of the both discomfort types. However multivariable regression does not yield any significantly better results in general, compared to single variable cases. The MAE does not show any significant decrease when multiple statistics are used together, and in some cases MAE scores of multivariable regression results are even worse than the single variable regressions. The multivariable regressions do not show significantly better results, possibly because of limited amount of test data.

In any case, it is encouraging to observe that the number of subjects with eye strain can be predicted with an average error rate of 1 person among 15 subjects (6.7%). Similarly the number of subjects with 3D perception discomfort can be predicted with an approximate average error rate of 2 out of 15 subjects (13.3%). These results show that, using the simple statistics we derived from the sparse disparity maps, the statistics for depth related discomforts of the viewers can be predicted precisely.

5. CONCLUSION

In this study, we presented a method for prediction of depth related discomforts, which result from excessive depth levels in stereo videos. Our approach depends on maximum disparity statistics, which are extracted from the estimated sparse disparity maps. Through comparative performance evaluation of the five block matching based methods we considered, we have shown that maximum disparities can be extracted from sparse disparity maps accurately. Furthermore, we have shown that by tracking the maximum disparities in the video scenes, certain statistics can be derived and they can be used to predict the the number of viewers experiencing depth discomforts for different video scenes.

We compared the five different matching cost calculation methods according to their overall disparity estimation performances and also according to their performances in capturing the maximum disparities correctly. To be fair in our comparison, we chose the parameters, which yield the most successful results, for each method. It has been observed, that CGO method we presented performs best in estimation of overall and maximum disparities. Since this method is based on correlation of edge orientations instead of correlation of pixel intensities, the disparities around the edges can be more accurately estimated. CGO algorithm can be further enhanced by using time correlated information between consecutive frames, to improve estimation reliability and to obtain a smoother disparity time sequence.

The subjective tests we have performed, show that the eye strain of the subjects can be best predicted by using the maximum disparities and the slew rates of maximum disparities in the scenes together. The 3D perception quality can be best predicted from the standard deviation of maximum disparities in the scenes. Using the maximum disparity and the slew rate together for prediction gives slightly better results compared to using only the maximum disparity. However the improvement provided from multivariable prediction is not significant. We believe this is because

of the limited data we had. A future research can focus on collecting more video test scenes and performing the subjective tests for a larger number of subjects.

The proposed maximum disparity to discomfort prediction method can be used together with other video quality measures, such as similarity of color, brightness and focus between image pairs, to obtain a more comprehensive stereo video quality measure. Using such a quality control system would help significantly in every stage of commercial 3D content production and hence increase the 3D viewing comfort of the end user.

REFERENCES

1. Boev, A., D. Hollosi and A. Gotchev, *Classification of Stereoscopic Artefacts*, Tech. Rep. 216503, MOBILE3DTV Project, `http://sp.cs.tut.fi/mobile3dtv/results/#technical-reports` , accessed at May 2013.

2. Lambooij, M. T. M., W. A. IJsselsteijn and I. Heynderickx, "Visual Discomfort in Stereoscopic Displays: A Review", *Proceedings of SPIE, Stereoscopic Displays and Virtual Reality Systems XIV*, Vol. 64900I, 2007.

3. Tam, W. J., F. Speranza, S. Yano, K. Shimono and H. Ono, "Stereoscopic 3D-TV: Visual Comfort", *Broadcasting, IEEE Transactions on*, Vol. 57, No. 2, pp. 335–346, 2011.

4. Onural, L., T. Sikora, J. Ostermann, A. Smolic, M. R. Civanlar and J. Watson, "An Assessment of 3DTV Technologies", *Proceedings of the NAB Broadcast Engineering Conference*, pp. 456–467, 2006.

5. Meesters, L. M., W. A. IJsselsteijn and P. J. Seuntiëns, "A Survey of Perceptual Evaluations and Requirements of Three-dimensional TV", *Circuits and Systems for Video Technology, IEEE Transactions on*, Vol. 14, No. 3, pp. 381–391, 2004.

6. De Silva, V., A. Fernando, S. Worrall, H. K. Arachchi and A. Kondoz, "Sensitivity Analysis of The Human Visual System for Depth Cues in Stereoscopic 3-D Displays", *Multimedia, IEEE Transactions on*, Vol. 13, No. 3, pp. 498–506, 2011.

7. Chen, W., J. Fournier, M. Barkowsky and P. Le Callet, "New Requirements of Subjective Video Quality Assessment Methodologies for 3DTV", *Video Processing and Quality Metrics (VPQM)*, 2010.

8. Jolly, S., J. Zubrzycki, O. Grau, V. Vinayagamoorthy, R. Koch, B. Bartczak,

J. Fournier, J. Gicquel, R. Tanger, B. Barenbrug, M. Murdoch and J. Kluger, *3D Content Requirements & Initial Acquisition Work*, Public Document 215075, 3D4YOU Project, 2009.

9. Ijsselsteijn, W. A., P. J. H. Seuntiëns and L. M. J. Meesters, *Human Factors of 3D Displays*, pp. 217–233, John Wiley & Sons, Ltd, Chichester,UK, 2006.

10. Ukai, K. and P. A. Howarth, "Visual Fatigue Caused by Viewing Stereoscopic Motion Images: Background, Theories, and Observations", *Displays*, Vol. 29, No. 2, pp. 106–116, 2008.

11. Richardt, C., L. Świrski, I. P. Davies and N. A. Dodgson, "Predicting Stereoscopic Viewing Comfort Using a Coherence-based Computational Model", *Proceedings of the International Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging*, CAe '11, pp. 97–104, New York, USA, 2011.

12. Cho, S.-H. and H.-B. Kang, "The Measurement of Eyestrain Caused from Diverse Binocular Disparities, Viewing Time and Display Sizes in Watching Stereoscopic 3D Content", *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 23–28, 2012.

13. Kim, D., Y. J. Jung, E. Kim, Y.-M. Ro and H. Park, "Human Brain Response to Visual Fatigue Caused by Stereoscopic Depth Perception", *Digital Signal Processing (DSP), 2011 17th International Conference on*, pp. 1–5, 2011.

14. Benoit, A., P. Le Callet, P. Campisi and R. Cousseau, "Using Disparity for Quality Assessment of Stereoscopic Images", *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pp. 389–392, 2008.

15. Li, J., M. Barkowsky, J. Wang and P. Le Callet, "Study on Visual Discomfort Induced by Stimulus Movement at Fixed Depth on Stereoscopic Displays Using Shutter Glasses", *Digital Signal Processing (DSP), 2011 17th International Conference on*, pp. 1–8, 2011.

16. Mittal, A., A. Moorthy, J. Ghosh and A. Bovik, "Algorithmic Assessment of 3D Quality of Experience for Images and Videos", *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE), 2011 IEEE*, pp. 338–343, 2011.

17. Sohn, H., Y. J. Jung, S. il Lee, H. Park and Y.-M. Ro, "Attention Model-based Visual Comfort Assessment for Stereoscopic Depth Perception", *Digital Signal Processing (DSP), 2011 17th International Conference on*, pp. 1–6, 2011.

18. Sohn, H., Y. J. Jung, S. il Lee and Y. M. Ro, "Predicting Visual Discomfort Using Object Size and Disparity Information in Stereoscopic Images", *Broadcasting, IEEE Transactions on*, Vol. 59, No. 1, pp. 28–37, 2013.

19. Scharstein, D. and R. Szeliski, "High-accuracy Stereo Depth Maps Using Structured Light", *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, Vol. 1, pp. 195–202, 2003.

20. Scharstein, D., *Middlebury Stereo Datasets*, `http://vision.middlebury.edu/stereo/data/` , accessed at May 2013.

21. Scharstein, D. and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms", *International Journal of Computer Vision*, Vol. 47, pp. 7–42, 2002.

22. Hirschmüller, H. and D. Scharstein, "Evaluation of Cost Functions for Stereo Matching", *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1 –8, 2007.

23. Birchfield, S. and C. Tomasi, "A Pixel Dissimilarity Measure that is Insensitive to Image Sampling", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 20, No. 4, pp. 401–406, 1998.

24. Sun, C., "Rectangular Subregioning and 3-D Maximum-surface Techniques for

Fast Stereo Matching", *Stereo and Multi-Baseline Vision, 2001. (SMBV 2001). Proceedings. IEEE Workshop on*, pp. 44–53, 2001.

25. Kanade, T. and M. Okutomi, "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 16, No. 9, pp. 920 –932, 1994.

26. Yoon, K.-J. and I.-S. Kweon, "Locally Adaptive Support-weight Approach for Visual Correspondence Search", *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 2, pp. 924–931, 2005.

27. Hirschmüller, H., P. Innocent and J. Garibaldi, "Real-Time Correlation-Based Stereo Vision with Reduced Border Errors", *International Journal of Computer Vision*, Vol. 47, pp. 229–246, 2002.

28. Richardt, C., D. Orr, I. Davies, A. Criminisi and N. Dodgson, *Real-Time Spatiotemporal Stereo Matching Using the Dual-cross-bilateral Grid*, Vol. 6313, pp. 510–523, Springer Berlin Heidelberg, 2010.

29. Hosni, A., C. Rhemann, M. Bleyer, C. Rother and M. Gelautz, "Fast Cost-Volume Filtering for Visual Correspondence and Beyond", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 35, No. 2, pp. 504–511, 2013.

30. Cigla, C. and A. Alatan, "Efficient Edge-preserving Stereo Matching", *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 696–699, 2011.

31. Szeliski, R., R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen and C. Rother, "A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 30, No. 6, pp. 1068–1080, 2008.

32. Boykov, Y., O. Veksler and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 23, No. 11, pp. 1222 –1239, 2001.

33. Kolmogorov, V. and R. Zabih, "Computing Visual Correspondence with Occlusions Using Graph Cuts", *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, Vol. 2, pp. 508–515, 2001.

34. Felzenszwalb, P. and D. Huttenlocher, "Efficient Belief Propagation for Early Vision", *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, Vol. 1, pp. 261–268, 2004.

35. Zabih, R. and J. Woodfill, "Non-parametric Local Transforms for Computing Visual Correspondence", *Computer Vision - ECCV '94*, Vol. 801 of *Lecture Notes in Computer Science*, pp. 151–158, Springer Berlin Heidelberg, 1994.

36. Moser, B., "A Similarity Measure for Image and Volumetric Data Based on Hermann Weyl's Discrepancy", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 33, No. 11, pp. 2321 –2329, 2011.

37. Liu, C., J. Yuen and A. Torralba, "SIFT Flow: Dense Correspondence across Scenes and Its Applications", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 33, No. 5, pp. 978–994, 2011.