Objective assessment of stereoscopic video quality of 3DTV

Darya Khaustova

ANNÉE 2015

UNIVERSITÉ DE RENNES 1

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

École doctorale Matisse

présentée par

Darya KHAUSTOVA

préparée au centre Inria Rennes - Bretagne Atlantique
et à Orange Labs

Objective assessment of stereoscopic video quality of 3DTV

Thèse soutenue à Rennes
le 30.01.2015

devant le jury composé de :

Prof. Patrick LE CALLET
IRCCyN, Université de Nantes / *Rapporteur*

Prof. Touradj EBRAHIMI
Ecole Polytechnique Fédérale de Lausanne / *Rapporteur*

Dr. Mohamed-Chaker LARABI
Maître de conférences, l'Université de Poitiers / *Examinateur*

Prof. Luce MORIN
INSA, Rennes / *Examinateur*

Prof. Christine GUILLEMOT
Directeur de Recherche, INRIA Rennes / *Directeur de Thèse*

Dr. Olivier LE MEUR
Maître de conférences, l'Université de Rennes 1 / *Co-directeur de Thèse*

Acknowledgments

The three years of my PhD have passed so fast. Though this period has taken about 5% of my lifetime (if to consider that I have 60 deliberate years). Now it is time to summarize.

First of all, I would like to sincerely thank to my supervisors Jérôme Fournier, Oliver Le Meur and Emmanuel Wyckens. They have managed to support, inspire, encourage and motivate me.

I must express my deepest gratitude to Bernard Letertre, Jean-Charles Gicquel, and Amelie Lachat for their contribution and help with subjective experiments, equipment, discussions and advices.

I particularly thank to my PhD jury members: Patric Le Callet, Touradj Ebrahimi, Chacker Larabi, Luce Morrin, and Cristine Guillemot. I really appreciate all your kind and professional feedbacks about my work and discussion we had during the question session.

Dear colleagues from HEAT team, Orange and SIROCCO, INRIA, I am so grateful for your support, concern and friendship. It would be such a shame to forget somebody's name here so I would not personalize this paragraph.

It would not be possible to record my fully controlled, located in the zone of comfort stereoscopic 3D videos without my film crew (and actors in the same time): Julien Libouban, Remi Rouaud, Didier Gaubil and Jean-Yves Leseure. Without our 3D clips nobody would ever believe that sunny hot days are possible in Rennes!

I thank to my dear friends Anna, Bogdan, Fabien D., Fabien R., Fanny, Ferran, Lucile, Mael (in alphabetical order) for all their concerns, support and friendship! I am so lucky to know you guys! Fanny thank you so much for being my friend, for your never ending optimism and support in the moments when "ça va pas". Also what an amazing "pot" you have organized for me with the help of other friends!

I appreciate the support of my family during the time I was supposed to support them. Mom, dad, I hope to be able to hug you and to celebrate my diploma soon despite of all the obstacles... Vova thank you for always believing in me. I wish you didn't have to leave time after time...

Thank you all!

Contents

Résumé en français : Evaluation objective de la qualité vidéo en TV 3D relief

Introduction

Les mesures de distorsion spatiale et temporelle habituellement utilisées pour tester la qualité des contenus stéréoscopiques étaient jugées insuffisantes car les notions de profondeurs d'image devaient être également considérées. Par conséquent, le terme de qualité d'expérience [Le Callet et al., 2012] (QoE) a été redéfini afin de caractériser l'expérience utilisateur des images stéréoscopiques dans sa globalité. La recommandation ITU-R BT.2021-13 prend en compte trois axes des perceptions primaires influençant la qualité perçue, la qualité intrinsèque, la profondeur et le confort visuelle des images [ITU, 2012a].

Les problèmes liés au confort visuel résulteraient des déficiences visuelles des téléspectateurs. Par exemple Solimini et al. ont conduit une étude sur une large variété de films en stéréoscopie. 953 questionnaires provenant des spectateurs ont été collectés où 60.4% d'entre-eux ont évoqué des problèmes de fatigue visuelle, vision double, vertige, mal de tête, nausée, de palpitation durant le visionnage du film [Solimini et al., 2012]. Par conséquent, dans le cas d'images stéréoscopique 3D, les exigences minimales sur les performances du système seraient de n'avoir aucun inconfort visuel.

Actuellement, le test subjectif est le moyen la plus adaptée pour refléter l'opinion des observateurs ou des clients sur la qualité d'un service proposé. Cependant, les services temps réel nécessitent l'usage de métriques objectives capables de prédire et le surveiller la qualité à la volée. Aussi, ces mesures doivent être capables de garantir un niveau de qualité vidéo suffisante pour les utilisateurs. Par conséquent, les objectifs de cette thèse sont orientés vers ces différents volets précités. Tout d'abord, nous investiguons sur l'attention visuelle en 3D afin de concevoir une nouvelle métrique objective. Ensuite, il est nécessaire d'adapter les métriques selon les besoins d'un service, par exemple, sur le niveau d'acceptabilité.

État de l'art

La disparité binoculaire n'est pas la seule source d'information pour évaluer la profondeur. Bien qu'avec des systèmes stéréoscopiques, l'amélioration de la perception de la profondeur est générée principalement par la disparité binoculaire. La sensation de profondeur apparait quand le cerveau fusionne deux images plates légèrement différentes. Cependant, les limitations du système visuel humain peuvent influencer la perception de

profondeur avec un système 3D en relief. Par exemple, des conflits peuvent apparaitre quand la divergence entre images est trop grande ou quand le système visuel humain est sensible à des vues non naturelles. De tels conflits peuvent créer des inconforts visuels, lesquels pourraient être évités si les scènes reconstruites étaient présentés dans une zone de confort visuel limité à 0.2 dioptrie. De plus, la quantité de profondeur perçue dépend de la taille de l'écran et de la distance de visualisation. Par conséquent, quand nous étudions la perception de profondeur, il est nécessaire d'inclure la distance de visualisation, la disparité entre vue afin de généraliser les résultats.

Les paramètres techniques de la chaine de diffusion de contenus 3D TV peuvent avoir un impact potentiel sur l'expérience utilisateur. Au niveau de la production de contenus, trois paramètres de prise de vue influencent la profondeur perçue, la distance focale, baseline , la distance de convergence; au niveau de la visualisation, les paramètres humain sont la distance binoculaire, la largeur de l'écran, et la distance di visualisation. Ainsi, l'absence de maitrise des paramètres de prise de vue et de l'environnement de visualisation entravent la production de résultats. De plus, la qualité de la perception de la profondeur est basée sur l'absence défauts visuels. Ces imperfections peuvent être produits a chaque niveau de la chaine de diffusion c'est-à dire, à la production des contenus, au codage, à la transmission et à la visualisation. Par conséquent, il est très important de prendre en compte l'impact de la technologie afin d'éviter des coûts de production prohibitifs.

La recommandation ITU-R BT.2021-13 fixe trois dimensions primaires perceptuelles, lesquelles influence la QoE perçue, la qualité de l'image, la qualité de la profondeur, le confort visuel [ITU, 2012a]. Il a été prouvé que ces trois composantes de base de la QoE 3D peuvent avoir un lien direct entre les paramètres de qualité techniques contrairement aux composants perceptuels de plus niveau (naturel, sentiment de présence). Pour la qualité de la profondeur, l'état de l'art montre que ce concept est assez difficile à interpréter par les observateurs, d'où les possibilités de trouver des indicateurs plus représentatifs. Finalement, les images stéréoscopiques non correctement capturées et rendus peuvent induire un inconfort visuel, lesquelles ont un impact sur la QOE globale et indépendamment de la qualité de l'image [Tam et al., 1998, Kaptein et al., 2008]. Les conflits d'accommodation [Yano et al., 2004, Lambooij et al., 2007, Hiruma and Fukuda, 1993] et des asymétries entre les vues sont des sources manifestes d'inconfort visuel pour les systèmes 3D [Kooi and Toet, 2004, Chen, 2012]. Par conséquent, importe quel service 3D devrait minimiser l'inconfort visuel perçu pour ces client regardant ce service. Il serait donc intéressant de caractériser objectivement l'impact des paramètres techniques afin de surveiller la qualité perçue des images stéréoscopiques.

Le panorama des métriques objectives pour l'évaluation de la QoE 3D démontre que la plupart des métriques existantes sont adaptés de l'évaluation de la qualité 2D. De tel métriques ne sont pas en mesure de détecter des problèmes provenant du confort visuel. Néanmoins, les métriques existantes en 3D n'ont pas incluses toutes les sources d'inconfort visuel. Par conséquent, nous concluons que les métriques pour la QoE 3D actuelles sont perfectibles et doivent être amendées a) avec les dimensions perceptuelles primaires, b) avec la technologie des écrans et format, c) l'environnement de visualisation (taille de l'écran, distance de visualisation). Un tel modèle sera donc capable de mesurer la qualité des images stéréoscopiques mais également son rendu. Si l'une de ces composantes est absente, il devient alors difficile de conclure sur la qualité d'expérience pour les images 3D.

La première partie de la thèse étudie l'attention du système visuel humain pour la

création d'une métrique de la qualité objective en 3D. La seconde partie de la thèse est dédié à la conception d'un modèle QoE 3D lequel est basé sur les seuils de perception humain et le niveau d'acceptabilité (voir Figure 1).

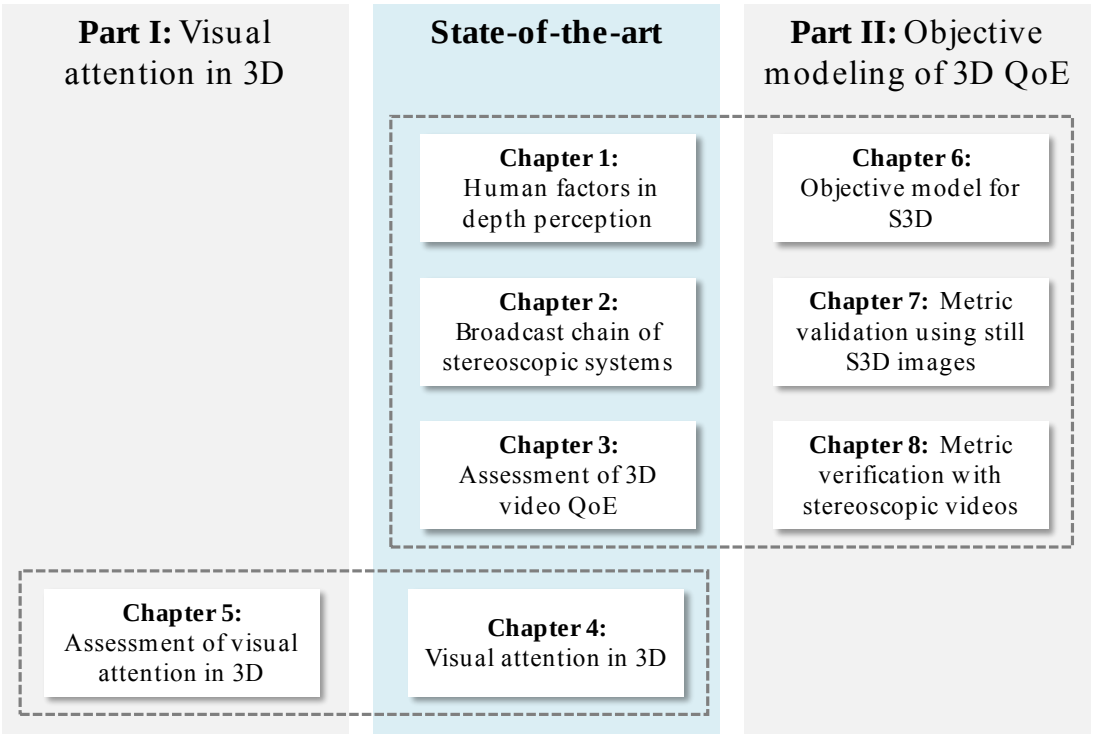


Figure 1: Présentation des chapitres de thèse.

Partie I: attention visuelle dans un contexte 3D

La première partie de cette thèse introduit l'état de l'art relatif à l'attention visuelle. L'analyse d'études récentes permet de comparer l'attention visuelle à la fois dans des conditions stéréoscopiques et dans des conditions de visualisation non-stéréoscopiques. Il est intéressant de noter le manque de consensus d'une part sur les indicateurs utilisés et d'autre part dans les conditions expérimentales. Par ailleurs, la plupart des études ne prennent pas en considération la zone de confort visuelle lorsqu'un contenu est affiché en condition 3D. Par conséquent, l'absence de standard ou de protocole sur la façon d'effectuer des tests de suivi de regard ne facilite pas la généralisation et l'analyse des études actuelles.

A partir ces études, nous présentons trois études expérimentales pour comparer le déploiement de l'attention visuelle dans un contexte de visualisation 2D et 3D stéréoscopique sur des images fixes. Les stimuli sont parfaitement contrôlés à la fois durant leur création et leur affichage. Cela permet d'étudier l'impact sur l'attention visuelle de différents niveaux de disparités binoculaire (incluant des niveaux entraînant un inconfort visuel) ainsi que des niveaux de complexités différents de texture. Les expérimentations sont menées grâce à un oculomètre et un écran 3D. Les fixations visuelles collectées sont utilisées pour construire des cartes de saillance. Ces cartes représentant les zones de saillance en 2D ou en 3D sont analysées afin de déterminer le

degré de similarité entre déploiement oculaire.

L'objectif de la première expérimentation est d'étudier l'influence de la profondeur sur l'attention. 28 observateurs ont été impliqués dans cette expérimentation. Les stimuli sont des scènes simples avec une disparité croisée ou non croisée. Un test univarié de significativité (ANOVA) est utilisé et montre que le facteur texture est plus important que le facteur profondeur pour la sélection des objets. Les objets avec disparité croisée sont plus importants comparativement à une visualisation 2D. Cependant, pour une disparité non croisée, aucune différence significative est observée entre objet visualisé en 3D ou en en 2D. L'analyse des mouvements oculaires ne révèle pas de différences sur les amplitudes de saccades. Les durées de fixations sont cependant significativement plus longues dans des conditions de visualisation stéréoscopiques que pour une visualisation en 2D. Nous pensons que ces résultats peuvent être utilisés pour affiner la conception des modèles de saillance dans un contexte 3D.

L'objectif de la seconde expérimentations est de confirmer les résultats précédents sur des stimuli complexes. 6 scènes présentant différentes structures sont générées avec le logiciel Blender. Pour ces scènes, les paramètres suivants sont modifiés : complexité de la texture et la profondeur de la scène (en modifiant l'écart entre les caméras et la distance de convergence). Les observateurs impliqués dans l'expérimentation de suivi de regard, regardent les scènes de façon libre (expérimentation sans tâche) avec des niveaux de profondeur et de complexités de texture variés. Pour éviter un effet mémoire, chaque observateur ne voit qu'une seule fois une scène donnée. 135 observateurs (106 hommes et 39 femmes âgés de 21 à 60 ans) participent à l'expérimentation. Chaque scène est regardée par 15 observateurs. Les données collectées sont utilisées pour construire des cartes de saillance afin d'analyser l'impact des différentes conditions de visualisation. Les résultats indiquent l'introduction de la disparité tend à réduire l'amplitude de saccades. Cependant, la durée de fixation n'est pas affectée. L'analyse des cartes de saillance ne révèle pas de différence entre les conditions 2D et 3D pour une durée de visualisation de 20 secondes. Nous n'avons pas constaté que le confort (inconfort) visuel engendrait une modification de l'attention visuelle.

L'objectif de la troisième expérimentations est de compléter l'étude précédente avec des images qi seront affichées en disparités croisées. 51 observateurs participent à l'expérimentation. Il a été observé que plus la disparité croisée était importantes plus l'observateur focalise son attention sur l'objet (affiché avec une disparité croisée). Cette observation reste valable lorsque l'objet apparaît dans la zone d'inconfort visuel (engendré par une disparité croisée excessive). En outre, aucune preuve ne permet de dire qu'une disparité excessive influence le déploiement oculaire.

Finalement, une nouvelle métrique utilisant carte de profondeur et carte de saillance est proposée afin de comparer les stratégies visuelles dans des conditions visuelles différentes. La métrique permet de comparer notamment l'attention visuelle en 2D et en 3D ainsi que pour des paramètres d'affichage et de texture différents. Les résultats obtenus confirment les conclusions des tests de suivi de regard.

Partie II: un modèle objectif de la QoE vidéo 3D

Cette partie de la thèse définit un modèle permettant de prédire la QoE 3D et propose des critères d'évaluation associés à trois axes perceptuels de la QoE. Tenant compte de l'importance, pour tout système 3D, de garantir un confort visuel à ses téléspectateurs le modèle est réduit à une métrique objective de la prédiction du confort visuel. Lorsqu'un

niveau de distorsion provoquant un inconfort est détecté, il peut être classé automatiquement dans une catégorie représentée par une couleur : (1) Vert – pas de gêne visuelle, (2) Orange - acceptable, mais induit une gêne visuelle, (3) Rouge - niveau de gêne inacceptable. La frontière entre les catégories “Vert” et “Orange” définit le seuil de gêne visuelle, tandis que la frontière entre le “Orange” et “Rouge” définit le seuil d'acceptabilité. Potentiellement, les seuils de visibilité pourraient également être introduits dans la métrique en créant une catégorie supplémentaire dite “Jaune”, où un défaut est visible mais non gênant, donc acceptable.

De plus, la métrique proposée utilise les seuils perceptuels pour définir l'impact des paramètres techniques sur l'axe du confort visuel dans la QoE de la vidéo stéréoscopique. Après une mesure objective des paramètres techniques et la comparaison avec les seuils perceptuels, les catégories de couleurs reflètent le jugement des spectateurs basé sur le niveau d'acceptabilité et la gêne visuelle induite. De plus, il a été proposé de créer une échelle subjective se basant sur les catégories de couleur du modèle objectif. Cela permet d'établir un lien direct entre les expériences subjectives et les prédictions objectives. Les avantages de l'approche proposée sont la possibilité d'omettre la prédiction des notes MOS et d'ajuster la métrique en fonction du niveau d‘acceptabilité et de gêne visuelle de l'usager. De plus, la méthode ne dépend pas d'une technologie 3D précise et aucune référence n'est requise pour prédire l'inconfort visuel.

La métrique proposée est ensuite validée par des tests subjectifs avec des images stéréoscopiques fixes et animées présentant différents niveaux d'asymétrie. L'utilisation du modèle proposé comme échelle de couleur subjective a démontré qu'il était possible d'obtenir directement et en même temps les seuils d'acceptabilité et de gêne visuelle grâce à l'échelle de couleur. Toutefois, ces seuils ne sont pas les mêmes lorsqu'ils sont évalués avec des échelles standards. Les différences de valeur des seuils d'acceptabilité ont été expliquées par des concepts d'évaluation différents : le premier reflète l'acceptabilité globale, le second évalue l'acceptabilité en fonction de l'acceptation du niveau de gêne causé par la distorsion. Fondamentalement, l'acceptabilité est pondérée par la gêne visuelle. De même, elle peut être pondérée par d'autres critères d'évaluation (distorsion géométrique, flou, bruit, etc.) Nous croyons que la métrique de couleur proposée pourrait être transférée à l'évaluation d‘autres technologies où les dégradations peuvent être mesurées et associées à des seuils perceptuels.

Finalement, la métrique proposée est également utilisée avec des séquences vidéo stéréoscopiques contrôlées. La performance de la métrique est évaluée en comparant les prédictions objectives avec des notes subjectives pour différents niveaux d'asymétrie pouvant provoquer un inconfort visuel. La comparaison des seuils d'acceptabilité pour des images fixes et animées a montré des résultats significativement équivalents. Par conséquent, les niveaux d'acceptabilité obtenus pour des images 3D fixes peuvent être réutilisés pour des vidéos S3D. De plus, nous cherchons à évaluer l'impact sur les prédictions objectives de deux asymétries agrégées sur un même stimulus. Il a été démontré que les jugements sur l'agrégation d'asymétries géométriques sont majoritairement basés sur un décalage vertical global : 82% pour un décalage vertical et 18% pour l'agrandissement. Tandis que l'agrégation d'une asymétrie verte et d'un décalage vertical montre un impact quasi équivalent sur le jugement des sujets : 45% pour l'asymétrie verte et 55% pour le décalage vertical.

Conclusions

La première partie de cette thèse explore l'importance de l'attention visuelle dans la conception d'une métrique de qualité objective de la 3D. La stratégie d'observation des sujets pour des images stéréoscopiques fixes localisées à l'arrière de l'écran est similaire à celle utilisée pour des images 2D. Le regard est plutôt guidé par la saillance des objets que par la quantité de disparités décroisées. Par conséquent, dans la seconde partie de cette recherche l'effet de l'attention visuelle en 3D n'est pas considéré sachant que la plupart des contenus produits pour le cinéma ou la télévision sont des contenus avec des disparités décroisées.

Il a été conclu que les objets avec des disparités croisées attirent le maximum d'attention: plus il y a de disparités croisées, plus l'attention visuelle est dirigée sur cette partie de l'image. De plus, aucun résultat n'a montré que l'inconfort visuel généré par des disparités excessives influence la façon dont nous observons les images.

Finalement, une nouvelle métrique utilisant carte de profondeur et carte de saillance est proposée afin de comparer les stratégies visuelles dans des conditions visuelles différentes. La métrique permet de comparer notamment l'attention visuelle en 2D et en 3D ainsi que pour des paramètres d'affichage et de texture différents. Les résultats obtenus confirment les conclusions des tests de suivi de regard.

La seconde partie de la thèse a été dédiée à l'élaboration d'un modèle objectif de QoE pour la vidéo S3D, basé sur les seuils perceptuels humains et les niveaux d'acceptabilité. Les résultats des expériences subjectives avec des images stéréoscopiques fixes et animées ont montré de fortes corrélations entre les notes subjectives et les prédictions objectives faites en utilisant les seuils perceptuels obtenus pour toutes les asymétries testées (avec au minimum r=0,87). Cela implique qu'il est possible de classifier le paramètre technique mesuré dans une des catégories objective, en utilisant les niveaux d'acceptabilité et de gêne visuelle correspondant. Il a aussi été établit que la mesure objective peut être utilisée comme une échelle subjective pour évaluer les niveaux d'acceptabilité et de gêne visuelle en même temps.

L'avantage de cette métrique est qu'elle peut être ajustée en accord avec les attentes marketing, techniques ou d'autres domaines en changeant le pourcentage d'acceptabilité ou le niveau de gêne visuelle, e.g. adaptation de la largeur des catégories objectives. L'acceptabilité dépend de l'acceptation du niveau de gêne causé par la distorsion. Fondamentalement, l'acceptabilité est pondérée par la gêne visuelle. De façon similaire, elle peut être pondérée par d'autres critères (distorsion géométrique, flou, bruit, etc.) Nous croyons que la métrique de couleur proposée pourrait être transférée à l'évaluation d'autres technologies où les dégradations peuvent être mesurées et associées à des seuils perceptifs. Cependant, des tests complémentaires doivent être effectués pour vérification.

Introduction

Motivations and objectives

The major objective of this thesis is to consider an industrial need for an objective video quality characterization on the fly. The main idea is to use three color labels for a categorization of video sequences: green, orange and red. Each color category is linked to perceived video quality by perceptual thresholds. For example, acceptability, visibility and visual annoyance levels can serve as perceptual thresholds. Adjustable acceptability level makes this idea useful for marketing or service applications, which set sometimes requirements to video quality in terms of acceptability.

The proof of described concept is done using stereoscopic 3D video sequences. The reason of such choice is the added perceptual depth dimension. Thus, the measure of spatial and temporal distortions usually used in 2D video to assess the video quality of stereoscopic content became incomplete.

The term Quality of Experience [Le Callet et al., 2012] (QoE) was proposed to characterize the overall viewing experience of stereoscopic images. Recommendation ITU-R BT.2021-13 determines three primary perceptual dimensions, which influence the perceived QoE: picture quality, depth quality, and visual comfort [ITU, 2012a].

The problems with visual comfort dimension resulted in raised concern about possible side effects on spectators' health. For instance, Solimini et al. has conducted survey during various stereoscopic movies. 953 questionnaires of spectators have been collected and 60.4% of individuals reported at least one symptom related to tired eyes, double vision, headache, dizziness, nausea and palpitation while watching a movie [Solimini et al., 2012]. Hence, in case of stereoscopic 3D (S3D), the minimum requirement for the stable system performance should be absence of visual discomfort.

Currently a subjective assessment is a convenient way to reflect opinion of the viewers or customers about the quality of proposed service. However, the real-time services require objective metrics that are able to predict and monitor the video quality on the fly. Also it should be able to guaranty certain quality level of provided video to end users.

Therefore, the objectives of this thesis are:

- To investigate the interest of visual attention in 3D for designing of a new objective metric.
- To propose a new objective model, which links tangible aspects of viewing experience with the 3D technological parameters.
- To consider the possibility to tune the designed metrics based on different service requirements, for example, acceptability level.

Thesis outline

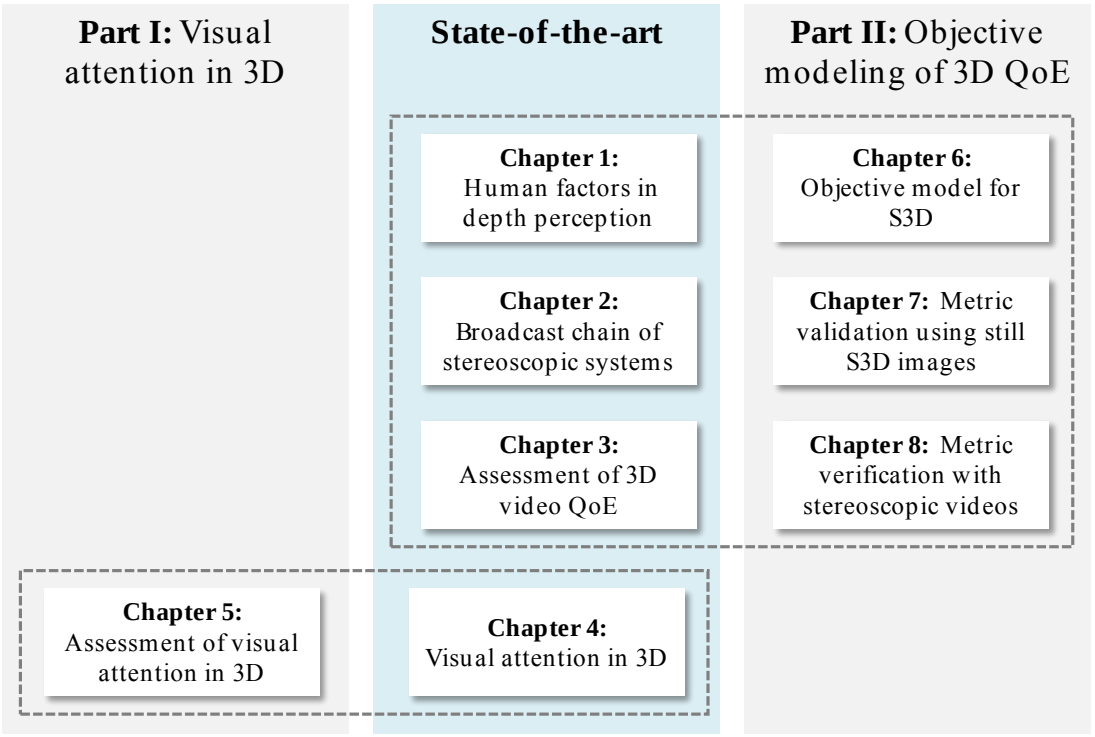


Figure 2: Overview of the thesis chapters.

Chapter 1 explains the mechanisms of human depth perception followed by a description of the functional principles of simple stereoscopic stereoscopic systems (S3D systems). The major advantage of such systems is an enhanced depth perception. Hence, depth perception is reviewed considering viewing distances, screen disparities and possible visualization artifacts. In artificial visualization environment, the limitations of the HVS can result in visual discomfort and visual fatigue. Therefore, the concept of comfortable viewing zone is described.

Chapter 2 discusses the impact of different technologies at every stage of the broadcast chain on final user viewing experience in detail. The basic principles of stereoscopic content acquisition and generation, 3D shooting rules, data representation, coding, and transmission and display technologies are presented.

Chapter 3 introduces the definition of 3D QoE and discusses what the quality means for 3D images and video. Then the existing models of 3D QoE are reviewed. Considering that a minimum system requirement for any stereoscopic system is to guarantee visual comfort for its viewers, we review subjective and objective ways of 3D QoE assessment with an emphasis on this issue. Moreover, the inherent properties of a comprehensive objective metric for the objective 3D QoE are discussed.

The first part of this thesis consists of two chapters (Chapter 4 and Chapter 5) and investigates whether visual attention of the viewers should be considered when designing an objective 3D quality metrics.

Chapter 4 introduces the state-of-the-art studies about visual attention and the methods to compare visual attention qualitatively and quantitatively. It also provides a review of recent studies comparing visual attention for S3D and 2D conditions.

In Chapter 5 presents three subjective experiments which aim to investigate an impact of depth on visual attention. First, the visual attention in 2D and 3D is compared using simple test patterns. The conclusions of this first experiment are validated using complex stimuli with crossed and uncrossed disparities. In addition, we explore the impact of visual discomfort caused by excessive disparities on visual attention. Lastly, the new objective depth metric is proposed in accordance with the subjective test results. This metrics allows comparing visual attention between 2D and 3D conditions as well as 3D conditions with different amount of depth.

The second part of the thesis is composed of three chapters (Chapter 6, Chapter 7 and Chapter 8) and dedicated to the design of an objective model of 3D video QoE.

The Chapter 6 presents a new objective model that uses perceptual thresholds to define the impact of technical parameters on the 3D video QoE. After objective measurement of 3D technical parameters and comparison with perceptual thresholds, it should be possible to predict evoked perceptual state, which reflects the viewers' categorical judgment based on stimulus acceptability and induced visual annoyance. The proposed model can be used as subjective color scale, where color category reflects the subjective judgments about a perceived stimulus.

Taking into account that the most important task for any 3D system is to guarantee visual comfort to its viewers, the model was tested for prediction of visual comfort. The goal of Chapter 7 is to verify proposed model comparing predicted categories with the votes from subjective test. Additionally, the possibility to use the proposed model as a new subjective scale is explored. For the validation of proposed model, subjective experiments with fully controlled still stereoscopic images with different types of view asymmetries are conducted. Finally, perceptual thresholds obtained with the proposed subjective scale are compared to the thresholds obtained with conventional assessment methods using impairment and acceptability scales.

Chapter 8 validates the proposed model with stereoscopic video sequences as well as compares perceptual thresholds for still and moving images. The metric performance is evaluated by comparing objective predictions with subjective scores for various levels of view discrepancies, which might provoke visual discomfort. Furthermore, this chapter explores how objective predictions should be affected if two view asymmetries were aggregated in a single stereoscopic stimulus.

The conclusions and perspectives of this thesis are given in Conclusions.

Chapter 1

Human factors in depth perception

1.1 Introduction

Currently there is nothing surprising about watching 3D movies, sporting events, or advertisements in cinemas or at home. Enhanced depth perception, sense of presence, and naturalness are some of the many reasons why 3D has become so widespread. But the most important reason is probably the improved entertainment experience. Besides entertainment, 3D technology can be applied in areas such as video games, medicine, telecommunications, robotics, and engineering.

All existing 3D systems are based on the principles of human depth perception. Thus, the first chapter of this thesis explains its basic mechanisms in relation with the *Human Visual System (HVS)* followed by a description of the functional principles of simple stereoscopic systems.

1.2 Principles of depth perception

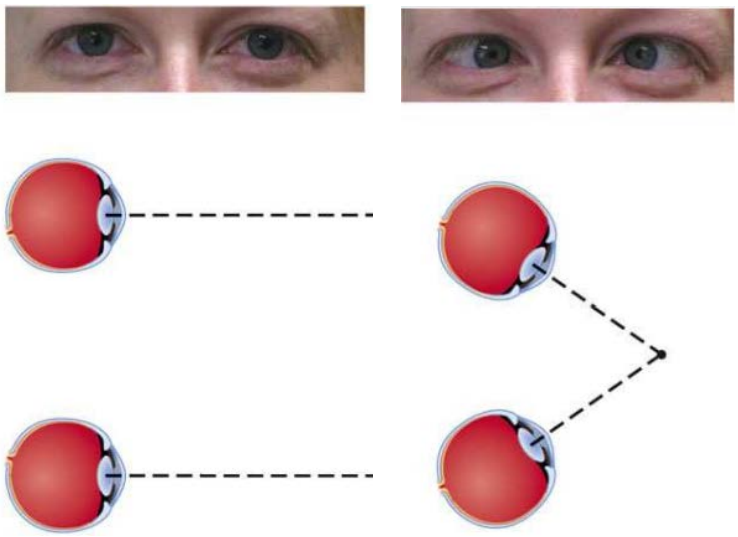
The HVS is endowed with the ability to reconstruct the three-dimensional world from two-dimensional images projected onto the retinas. The perceived depth is established through a variety of depth cues. These cues can be divided into three major groups: (1) *Oculomotor* – the cues based on the physical abilities of our eye muscles and lenses; (2) *Monocular* – the cues that require information from a single two-dimensional view; and (3) *Binocular* – the cues that extract information from both eyes [Goldstein, 2013]. The following sections give a detailed overview of the cues from each group.

1.2.1 Oculomotor cues

Oculomotor cues require feedback from muscles in the eye to provide information to the brain about the locations of objects in space. This muscular reaction causes *vergence* and *accommodation* of the eye [Holliman, 2003].

Vergence occurs when the eyes are moving inward or outward to insure that an object being fixated upon can be projected into the central part of the retinas. If an object of interest is located nearby, the eyes converge to fixate on it. However, when the object of interest moves away, the eyes diverge. This principle is demonstrated in Figure 1.1.

Furthermore, the distance between objects can be estimated by measuring the angle between the optical axes of two eyes.



(a) Man looks at the horizon(the optical axes of his eyes are in optical infinity)

(b) Man looks at the object 5 cm away (the optical axes of his eyes converge at the distance of 5 cm)

Figure 1.1: Eyes convergence principle from [Goldstein, 2013].

Accommodation occurs when the eyes are focusing on an object of interest to perceive it in a sharpened manner. During this process, depending on the distance, the shape of the eye lens is changes due to the ciliary muscles, while the focused object remains projected onto the retina. Points situated outside of the accommodated area cannot be properly projected onto the retina and are perceived as blurred within the limit of ± 0.2 diopters [Yano et al., 2004]. This limit is known as the *Depth of Focus (DoF)* and it is located around the accommodation distance. The size of the DoF diminishes when the pupil diameter increases and expands when the pupil diameter decreases.

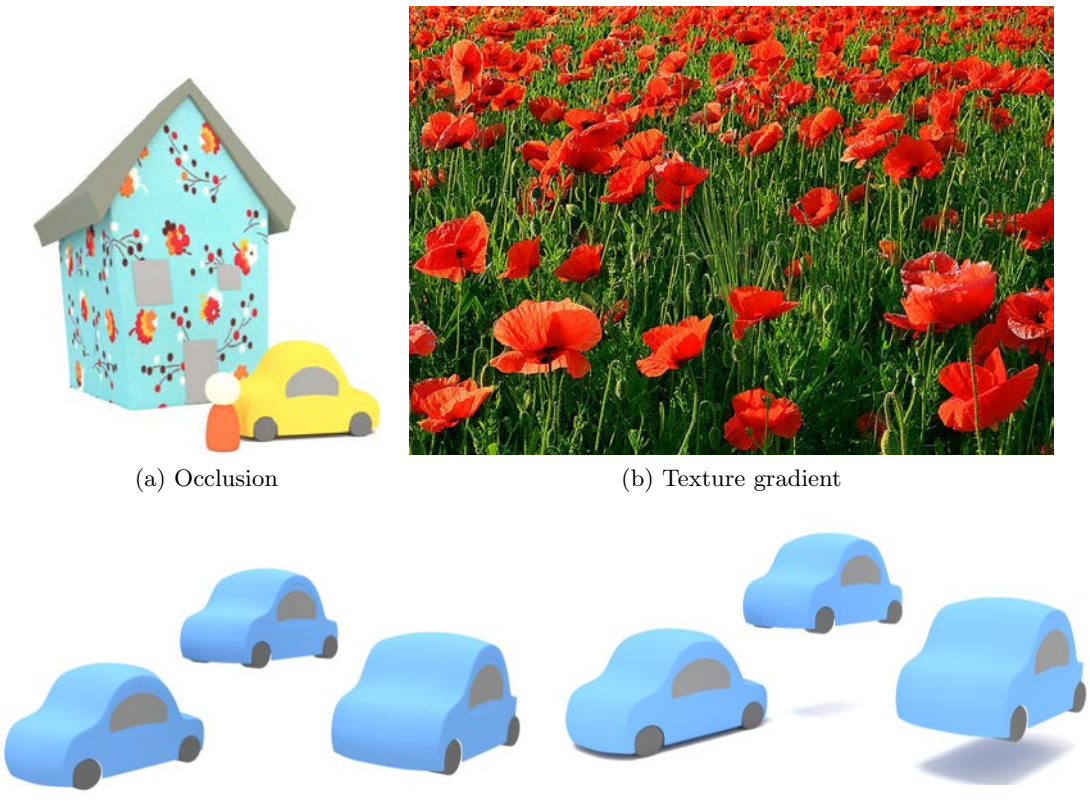
When the eyes fixate on an object of interest, the lens focuses to keep the perceived object sharp. Thus, accommodation and vergence intrinsically interact with each other [Suryakumar et al., 2007, Polak and Jones, 1990]. Basically, the amount of accommodation needed to focus an object is proportional to the amount of vergence appropriate to fixate the same object in the center of each retina. However, vergence is mostly driven by disparity [Stark et al., 1980] and it is more effective than accommodation at close distance estimation [Cutting and Vishton, 1995, Tresilian et al., 1999], whereas accommodation is guided by retinal blur [Phillips and Stark, 1977].

1.2.2 Monocular depth cues

Monocular cues estimate depth relying only on information from one eye. There are two categories of monocular cues: *pictorial cues*, which provide depth information from static 2D images and *movement-based cues*, which extract depth information from changes in retinal images over time, i.e. from movement. Accommodation, as described in the previous section, is also considered a monocular cue.

The most important pictorial cues [Goldstein, 2013, Holliman, 2003, Cutting and Vishton, 1995, Mendiburu, 2009] are listed below:

- Occlusion: occurs when one object partially covers another object. It suggests that the occluded object is farther away (see Fig. 1.2.a).
- Relative size: the sizes of similar objects on the retina are estimated. Closer objects result in larger retinal images.
- Relative height: the vertical position of the point in the visual field is estimated. Thus, objects which are closer to the horizon are seen as being further away.
- Texture gradient: repetitive patterns look smaller and denser as the distance increases (see Fig. 1.2.b).
- Aerial perspective: due to the presence of particles in the air (moisture, pollution), distant objects look less saturated and sharp. Moisture in the air can cause a color shift towards blue and pollution towards gray or brown colors [Mendiburu, 2009].
- Linear perspective: parallel lines seem to converge and recede to the horizon.
- Shadows: reflections or shadows casted by an object can give an idea about its location [Coren et al., 1994]. For example, without shadows, the blue cars seem to be located on the ground in Figure 1.2.c. However, when shadows are added, some of the blue cars appear to float in the air (see Fig. 1.2.d).



(a) Occlusion (b) Texture gradient

(c) No-shadows (d) Casted shadows

Figure 1.2: Monocular depth cues.

All pictorial cues take place when the observer's position and scene are stationary. When motion is added, other mechanisms complement depth perception. Goldstein defines two motion-produced cues: motion parallax and deletion/acceleration [Goldstein, 2013].

Motion parallax occurs when the movement of an observer or a scene induces change in the relative position of near and far objects on the retina. Relative retinal positions of near objects change faster than farther ones. This information is used by the HVS to extrapolate depth from the scene. Absolute depth information can be estimated when the speed and the direction of the movement are known [Ferris, 1972]. Finally, *deletion* and *acceleration* occurs when the observer moves through the environment and a near object covers or uncovers a more remote object [Kaplan, 1969].

1.2.3 Binocular depth cues

Binocular vision seems to be natural to us, but in fact our brain has to solve the complex problem of interpreting 2D retinal projections of the scene into a 3D perception. For example, in Figure 1.3 the yellow car is projected differently to each eye. It happens because each eye has a slightly different *Field of View (FoV)* due to their horizontal separation. This separation is called the *Interpupillary Distance (IPD)* and its average value is 63 mm [Dodgson, 2004] for adults. Furthermore, the discrepancy between the projection of the car onto the retina of one eye and the other is called *retinal disparity* or *binocular disparity*. The brain processes these binocular disparities and creates the impression of perceived depth, which is called *stereopsis* (see Fig. 1.3). Stereopsis helps to discriminate the difference in depth, to order the objects, to judge slant or curvature, to obtain relief and shape, to judge motion speed and direction in depth, to recover surface properties, or to obtain precise measurements of depth between objects

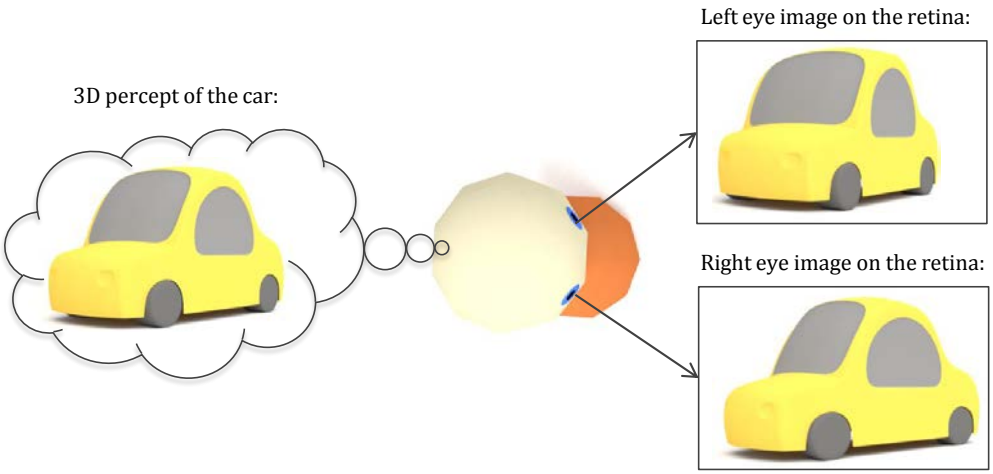


Figure 1.3: The figure sees the car from different angles because of the horizontal separation of the eyes. Using this difference, the brain fuses the left and right retinal image into a 3D perception.

Figure 1.4 illustrates binocular vision. The two eyes verge and their optical axes intersect on the fixation point F. Once the eyes have accommodated and converged, the fixation point is projected to the same position on both retinas. Hence, the fixation point acquires *zero retinal disparity*. All points that have zero disparity fall on the

horopter and are perceived at the same depth. The points that are situated in front of the horopter have negative or *crossed disparities* (point B in Fig. 1.4), whereas the points located behind the horopter have positive or *uncrossed disparities* (point A in Fig. 1.4).

The shape of horopter is approximately a circle, which is called the *Vieth-Müller circle*. The radius of this circle is determined by accommodation power and vergence [Schreiber et al., 2006]. This gives the correct representation of zero-disparity points only because the real shape of the horopter is non-linear [Blakemore, 1970] but in practice this difference can be neglected [Ijsselsteijn, 2004]. The region around the horopter, where the brain can fuse two retinal images is called *Panum's fusional area* (see Fig. 1.4)

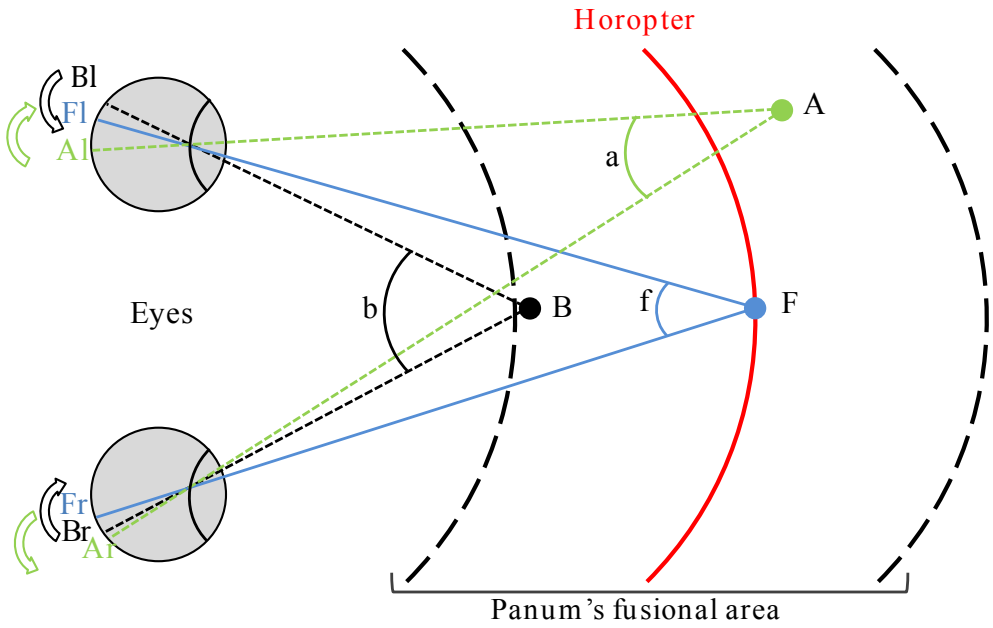


Figure 1.4: The horopter and Panum's fusional area. All points that fall on the horopter have zero retinal disparity. The points located around the horopter fall on Panum's fusional area, where binocular fusion can take place.

In practice, one of the ways to estimate binocular disparity is to calculate the difference between the fixation point and the vergence angle [Holliman, 2003]. In the case of crossed disparity, the difference will be negative: $\alpha_b = f - b$. In the case of uncrossed disparity, the difference will be positive: $\alpha_a = f - a$, where a, b, f are vergence angles as illustrated in Figure 1.4.

As was explained in the previous section, depth is perceived not only because of binocular disparity but also due to various pictorial cues. So, in real life it is difficult to find such situations where stereopsis is created only by binocular disparity. However, this is possible with the help of synthetic stimuli called *random-dot stereograms*; an example of such a stimulus is presented in Figure 1.5. To prove that stereopsis can only be created by binocular disparity, Julesz designed stereoscopic images of random-dot patterns that did not contain any pictorial cues [Julesz, 1971]. The idea is to take two identical random dot patterns and then to shift a square section of one pattern to the side. When the generated images are presented to the corresponding eyes, the created

disparity produces the feeling that the square with the random dots floats above the flat background.

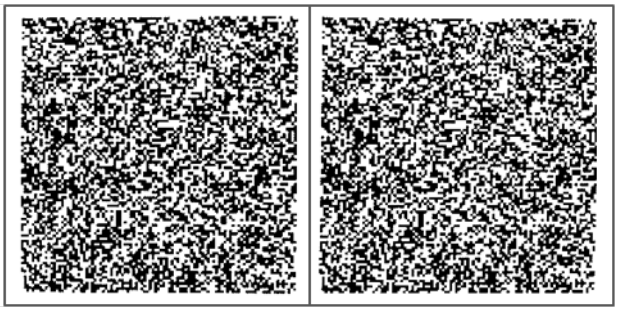


Figure 1.5: Left and right image of a random-dot stereogram. It is very difficult to discriminate which group of dots is shifted in the right image in comparison with the left, but the HVS can easily detect differences when corresponding views are presented to the left and right eyes.

Therefore, the results of Julesz's experiment have demonstrated the existence of neurons which react to different amounts of disparity. The *disparity-selective cells* or *binocular depth cells* were revealed in the visual cortex of the brain [Barlow et al., 1967, Hubel and Wiesel, 1970]. Furthermore, Poggio and Fisher classified binocular cells of the visual cortex in monkeys and found that about 67% of cells are sensitive to zero disparities, 19% to uncrossed disparities, and 9% to crossed disparities [Poggio and Fischer, 1977]. Uka et al. reported that binocular cells respond best to a certain amount of absolute disparity and confirmed that they are mostly tuned for either crossed or uncrossed disparity [Uka et al., 2000].

1.2.4 Depth cue interactions

The previous sections describe oculomotor, monocular, and binocular cues that contribute to depth perception. In everyday life we rarely find situations when only one cue is present. Redundant information occurs due to the importance of correct perception for humans. The brain must combine sensory information from multiple cues to create a single perception of depth.

Cutting and Vishton investigated a synergy between some of the cues and found out that the efficiency of some of them depends on distance [Cutting and Vishton, 1995]. The results of their research are presented in Figure 1.6. Depth is the mean distance of two objects from an observer (horizontal axis). Depth contrast is defined as the ratio of the just-noticeable distance between these objects over depth (vertical axis). Small depth contrast leads to high depth quantization, e.g. the observer can perceive very small differences in the distance between objects. Independent of distance, occlusion always contributes to depth perception and dominates over all other cues as shown in Figure 1.6. Since it is very easy for the HVS to define when an object covers the view of another one, occlusion is considered as one the least ambiguous depth cues. However, it gives no information about the amount of depth between two objects [Banks et al., 2012]. Binocular and oculomotor cues provide such information but they are more effective for the distance of closer objects (the closer the distance is, the smaller the depth contrast is). Binocular disparity decreases when an object moves at least 10 meters away from an observer, which makes these depth cues ineffective.

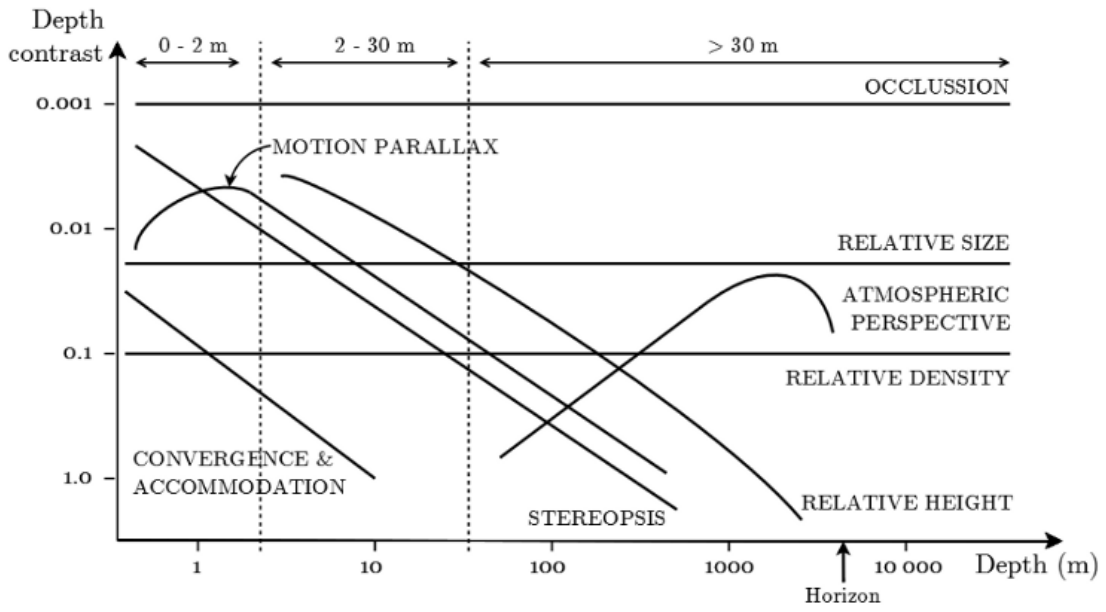


Figure 1.6: Depth thresholds for different depth cues as a function of distance from an observer; from [Cutting and Vishton, 1995].

As can be seen in Figure 1.6, the HVS produces redundant information taken from different cues at various distances. These cues can be perceptually combined in many ways [Howard, 2012]. Among the cues are:

- *Dominance.* Information provided from the occlusion cue dominates over all the other cues.
- *Summation.* For example, when monocular visual acuity is enhanced by binocular vision, [Banton and Levi, 1991] e.g. a person sees worse with one eye than with two open eyes.
- *Averaging.* These are additive interactions between cues.
- *Disambiguation.* The information from one cue disambiguates the interpretation of another cue. For example, blur resulting from focusing can disambiguate stereopsis [Banks et al., 2012].
- *Calibration* and *adaptation* happens when one cue helps to interpret information of another cue.
- *Dissociation.* The interaction of connected cues applied to a common feature of an object, e.g. the cues are applied differently depending on an object's feature or location.

A combined model for the integration of the cues or a generalized theory on how the brain extracts depth does not exist yet [Banks et al., 2012].

1.2.5 Individual differences

Humans perceive depth using depth cues. However, stereopsis is created only by binocular depth cues, so this sensation is susceptible by non stereo blind people. In other words, stereoblindness is the inability to perceive stereoscopic depth cues [Shibata et al., 2011].

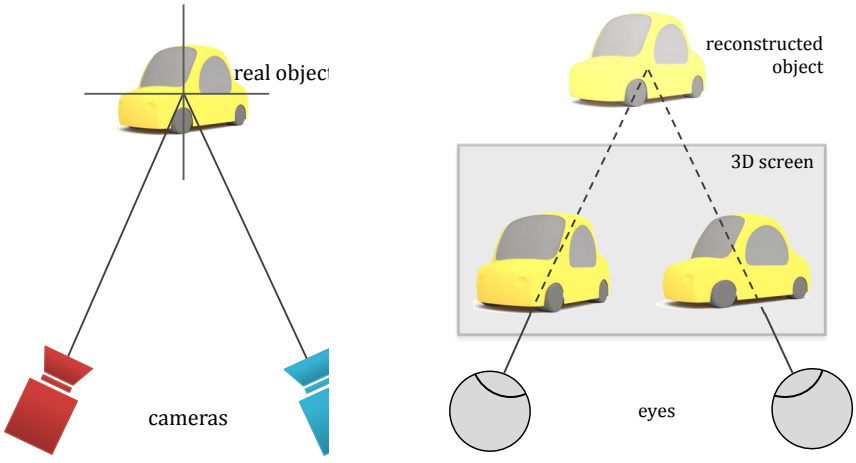
The studies of this phenomenon were conducted by Richards, who tested 150 participants for stereoblindness with random-dot stereograms [Richards, 1970]. He found that 4% were not able to perceive depth and 10% had difficulties detecting its direction relative to a background. Strabismus and amblyopia are some common reasons for stereoblindness.

There are several characteristics of the HVS that differ individually and affect stereopsis. For example, IPD varies from person to person and changes with the age. The range of IPD is from 40 to 80 mm for extreme cases and children [Dodgson, 2004]. Thus, for a given viewing distance and screen disparity, people with smaller IPD would perceive more depth (see more details in the next Section 1.3). Another characteristic is the pupil diameter, which influences the amount of depth e.g. DoF, when an image is perceived sharply [Lambooij et al., 2007].

The properties of the HVS change with age due to structural changes in the eyes. For example, the ability to accommodate decreases with age because the eye lens loses its elasticity; usually, almost no accommodation remains by the age of 55 years. However, the accommodation system of children develops until 7 years old and it is not known whether stereoscopic content influences their visual system [Lambooij et al., 2007].

1.3 The simple stereoscopic imaging system

Shooting and displaying three-dimensional images is an attempt to imitate what we see with our two eyes. Hence, a basic idea to imitate the HVS is to replace the left and right eyes with two horizontally separated video cameras. Next, to use a screen that is able to directly show the recorded left and right views to the corresponding eyes. Then the brain fuses these images, which results in depth perception. An example of such a simple stereoscopic system is illustra



(a) Image acquisition (b) Image visualization

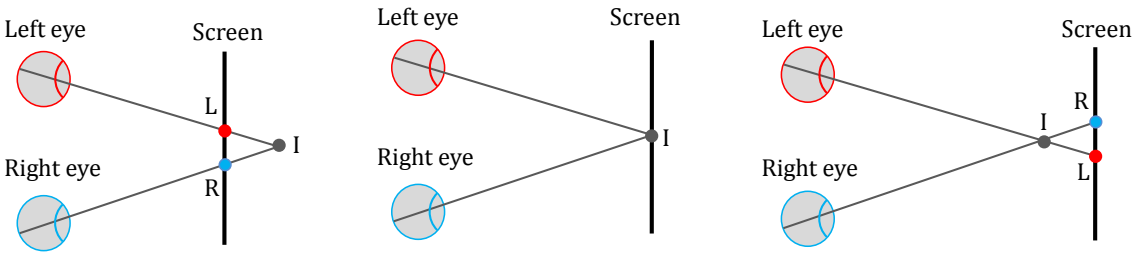
Figure 1.7: A simplest stereoscopic imaging system.

The horizontal separation between the left and right cameras is called *interaxial* or *baseline distance*. It creates the desired differences between the left and right views and, consequently, produces binocular disparity. Then, the binocular disparity drives accommodation and vergence, which are combined with the monocular depth cues of

a scene, including the occlusion, relative size of objects, shadows and relative motion. Finally, all this information is processed by the brain to create depth perception.

Typically, the recorded views are displayed on the same planar screen. The separation created between the position of corresponding points in the left and right images is called *parallax* or *screen disparity*. There are three types of parallax depending on the position of the reconstructed object relative to the screen plane, which is illustrated in Figure 1.8:

- Positive parallax. The object is perceived behind the screen plane. The displayed object is shifted to the left for the left eye and to the right for the right eye.
- Zero parallax. The object is perceived on the screen plane. The left eye and right eye image are in the same position on the screen.
- Negative parallax. The object is perceived in front of the screen. The displayed object is shifted to the right for the left eye and to the left for the right eye.



(c) Negative parallax

Figure 1.8: Types of the screen parallax.

Assuming an ideal planar stereoscopic display, Figure 1.9 shows the geometry of perceived depth, which depends on the screen parallax P, the viewing distance V and the separation between the observer's eyes (e). Z_m is the distance from the perceived object to the screen; Z_i is the distance from the perceived object to the viewer. From similar triangles, the equation 1.1 derives the perceived depth as the function of the parameters mentioned above; similarly, the equation 1.2 deduces the distance from a virtual object to the viewer.

$$Z_m = \frac{P \times V}{e - P} \tag{1.1}$$

$$Z_i = \frac{V \times e}{e - P} \tag{1.2}$$

From the equation 1.1, the perceived depth depends on the viewing distance and the screen disparity. The next section presents the consequences of such dependence.

1.3.1 Perceived depth as function of viewing distance

From the equation 1.2, the perceived depth depends on the viewing distance as illustrated in Figure 1.10. Therefore, a viewer that moves closer to the screen increases retinal disparities and perceives the reconstructed objects as being closer.

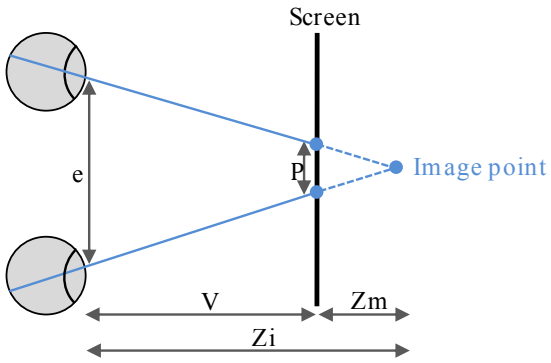


Figure 1.9: Perceived depth as a function of screen disparity.

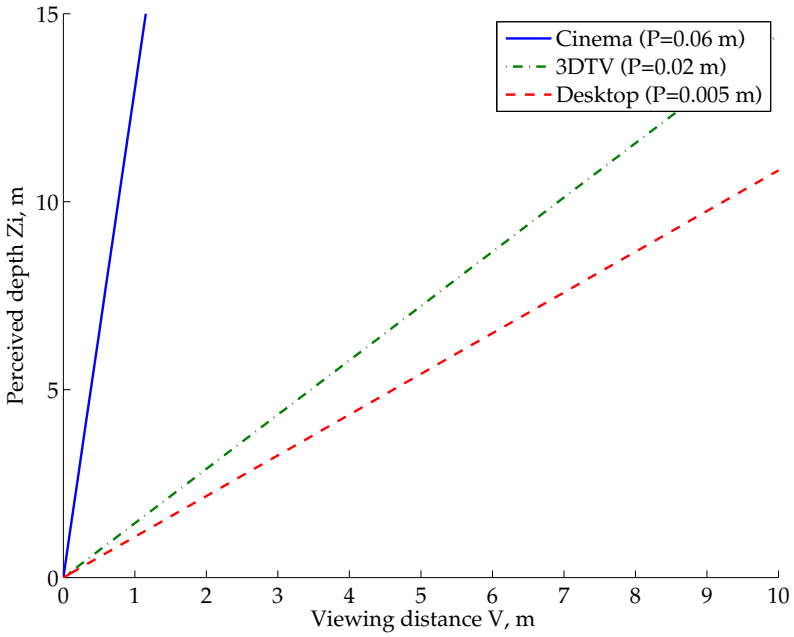


Figure 1.10: Perceived depth as a function of viewing distance and screen disparities.

1.3.2 Perceived depth as function of screen disparities

Also, in accordance with the equation 1.2, the perceived depth depends on the screen disparity. This means that the larger the parallax is, the larger the reconstructed virtual space will be. This dependency is illustrated in Figure 1.11. A more detailed example for cinema, TV, and desktop screens is given in Table 1.1, where the viewing distance is fixed at 4 meters. Screen parallax will increase when a content designed for a 3DTV is displayed on a cinema screen. As result, the objects are reconstructed as being farther in depth. Conversely, there is a reduction in screen parallax when the same content designed for the cinema is displayed on a desktop monitor. Hence, objects seems to be closer in depth.

Thus, it is very important to take into account that the screen parallax of the same 3D content varies depending on the screen size and viewing distance. This change in the disparity magnitude leads to a different perception of the depth of the objects. There are two ways to deal with this problem. The first way is to change the screen parallax

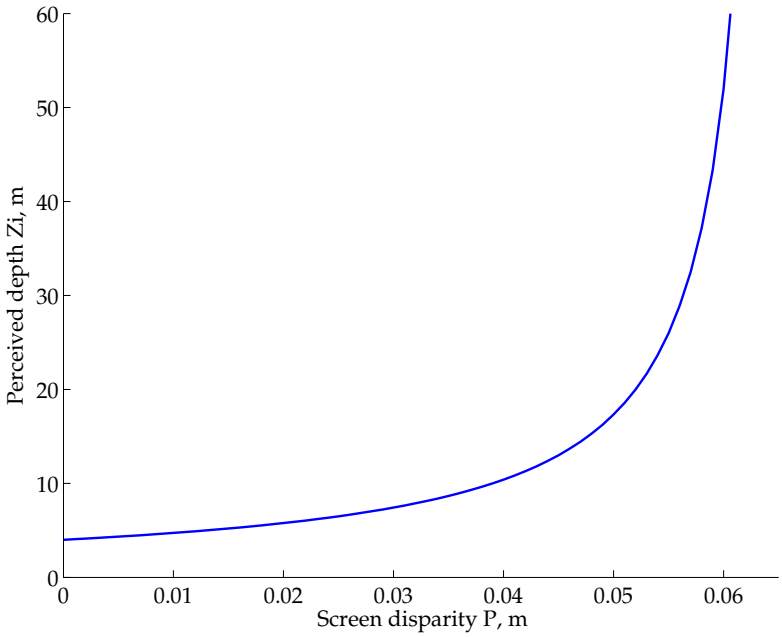


Figure 1.11: Perceived depth as a function of screen disparities.

Table 1.1: Perception of the same content on different screen sizes.

Display	P, m	V, m	IPD, m	Z_m, m
Cinema	0.06	4	0.065	52
3DTV	0.02	4	0.065	5.78
Desktop	0.005	4	0.065	4.3

to match the screen size in postproduction. For example, the screen parallax should be adapted for displaying content made for the cinema screen on a 3DTV. The second way is to change the viewing distance depending on the screen size. However, in this case, immersion, perception, and visual experience will differ due to changes in field of view [Sandrew, 2012].

1.3.3 Artifacts related to S3D visualization

Reconstructed stereoscopic content might be very different from a real world scene in terms of depth perception. The viewing distance can be modified while viewing but IPD and shooting parameters are unchangeable during the visualization stage. Thus, there is a possibility that any introduced disparities are unnatural to the HVS and that the reconstructed scene looks distorted.

In cinematography the *roundness factor* indicates the depth realism of produced stereoscopic content. A reconstructed object that has the same shape proportions as in real life has a roundness factor equal to 1. To facilitate content creation by filmmakers and stereographers, Smith et al. proposed a mathematical analysis [Smith and Collar, 2012]. It uses the selected camera and the scene parameters as the input and then predicts whether the intended artistic effect can be created. Also, Mendiburu provided the guidelines on how to refine the depth effect depending on the roundness factor [Mendiburu, 2009]. However, any introduced intervals of the roundness

factor are rather empirical and are not supported by any subjective experiments. For example, it is not possible to conclude whether the viewer will perceive any shape improvement if the roundness factor changes from 1.2 to 1.1 or if 0.7 is really the perceptual limit when viewers start to perceive a flattening of the objects. Such studies have not yet been conducted.

The most common artifacts related to depth perception are reviewed below.

1.3.3.1 Cardboard effect

The *cardboard effect* is the phenomenon when the depth planes of an image look as they are made from cardboard, while the objects in depth look flattened. This is the result of a perception mismatch between the binocular and the perspective depth cues, which is created by insufficient depth or disparity information [Boev et al., 2008]. This effect can be avoided by an accurate reproduction of binocular disparity during filming [Yamanoue et al., 2000].

1.3.3.2 Size distortion artifacts

Similar to the cardboard effect, the *puppet-theater effect* is also the result of flaws in the recreation of binocular disparities. Disproportion in reconstructed space leads to undesired geometrical distortions in depth perception because in real life distance and angular size are strongly linked. The puppet-theater effect makes 3D reconstructed objects appear unnaturally smaller in comparison with real objects. For example, people in stereoscopic pictures may look as small as the puppets. Usually effects are more noticeable for objects with familiar sizes. Distortion can be avoided by the correctly representing binocular disparity, which also applies in the case of the cardboard effect [Yamanoue et al., 2000].

Gigantism and *miniaturization* are more general cases of size distortion artifacts. Binocular vision provides an idea about the structure of a 3D scene. When this information is incorrect, the whole scene is exposed to large stereoscopic distortion and can appear gigantic or miniaturized. These distortions are produced while shooting with improper parameters [Devernay and Beardsley, 2010].

1.3.3.3 Window violation

The *window violation* effect is the result of a depth cue conflict. It occurs when objects with crossed disparity are cut off by the border of the screen [Mendiburu, 2009, Devernay and Beardsley, 2010]. The conflict is created because the brain does not know how to interpret a screen border located behind an object that occludes it at the same time. This violation can destroy the 3D effect or induce visual discomfort. Window violations on the top and bottom of the screen are considered to be less annoying than ones on the left and right borders [Collins et al., 2011].

Window violations can be avoided during the shooting stage or corrected by post processing technique, which is called *floating window*. Floating window is basically a crop mask, which hides the part of the object that causes the violation.

Subtitles in stereoscopic movies can also create the window violation effect. They are often rendered at the display plane and hence can produce depth discontinuities when the *region of interest (ROI)* of a scene is rendered in another depth plane. Therefore, subtitles reduce visual comfort [Lambooij et al., 2013]. Such conflict can be avoided if the subtitles are placed at the depth of a scene.

1.3.3.4 Stereoanomaly

Stereanomaly is a phenomenon when stereopair is perceived as inverted and objects with negative disparities acquire positive disparities and vice versa. The image pair perceived due to stereoanomaly is called *pseudoscopic* because of the inversed order of presentation. For example, in the case of S3D football game content, when the image intended for the left eye is seen by the right eye, which can result in perceiving the field to be in front of the players. This can be avoided by displaying the same image for the left and right eyes several times to insure the correct presentation order. Also this anomaly can be created by the brain if there is degraded stimuli (such as degraded images with low luminance, contrast, or resolution) and may occur for 20-30% of people [Patterson, 2007].

1.4 Limits of the HVS in binocular depth perception

Despite the numerous advantages that our visual system has, there are several limitations on how stereoscopic content should be displayed, which are discussed below.

1.4.1 Fusion range limits

As explained in Section 1.2, the principle of any stereoscopic system is to deliver two slightly different images to the corresponding eyes. The brain processes the created binocular disparity and fuses the two images into a single 3D view. However, there is a certain disparity value beyond which the images will not fuse and the objects are seen in different positions. The range over which fusion occurs is known as *Panum's Fusional Area*. The shape limits of Panum's fusional area are not constant over the retina. Qin et al. have investigated limits in 16 equidistant different directions from 0 to 360 degrees and concluded that the shape of Panum's fusional area has "the shape of an ellipse off-centered toward the nasal side on the horizontal meridian of the retina" [Qin et al., 2006].

Most studies have concentrated not on the shape but on the size of Panum's fusional area, i.e. the fusion limit of binocular disparities. Julesz and Schumer have performed experiments with random-dot stereograms and found that the limit of fusion is equal to 50 minutes of an arc [Julesz and Schumer, 1981]. By stabilizing images on the retina, this threshold can be increased [Fender and Julesz, 1967]. By using line stereograms, Mitchell found a lower threshold of 10 minutes of an arc for crossed and uncrossed disparities [Mitchell, 1966]. Yeh and Silverstein defined 24 minutes of arc as the limit for an uncrossed disparity and 27 minutes of arc for crossed for the short duration stimuli that did not require any vergence movements. But the limits increased to 1.57 degrees for uncrossed and 4.93 degrees for crossed [Yeh and Silverstein, 1990] for longer durations with allowed eye movements. Similar temporal dependency has been found by Schor and Tyler; when the disparities were changing slowly (0.2 Hz), the limits of binocular fusion were higher than for fast changes (5 Hz). They also found that stimuli with low spatial frequency increases the size of Panum's fusional area in comparison with high frequency [Schor and Tyler, 1981].

Such diversity in the results can be explained by the influence of various factors on the size of Panum's fusional area. Among those factors are stimulus size, exposure duration, continuous features, temporal effects, amount of luminance, and individual differences [Lambooij et al., 2007]. For now, there is still no common consensus on how large the disparity can be in order to avoid diplopia (double vision).

Naturally, the HVS system can fuse horizontal disparities as well as vertical disparities. The situation when binocular disparities are constant but eyes are converging results in a perspective distortion of the retinal image, which leads to an enlargement of vertical disparities. Moreover, all the points which do not fall on the horopter are projected to the retina with either a vertical disparity or both a vertical and a horizontal disparity (see Fig. 1.12)[Tyler, 2006]. Then the visual system uses these vertical disparities to scale the information from the binocular disparities [Banks et al., 2012]. This has been demonstrated in an experiment by Rogers and Bradshaw [Rogers and Bradshaw, 1993]. They fixed binocular disparity at 10 minutes of arc and placed the patterns at a distance of 57 cm. Then they found that the eyes of the subjects converged at different distances, changing the amount of vertical disparities of the proposed pattern. A convergence distance of 28 cm
perception of d

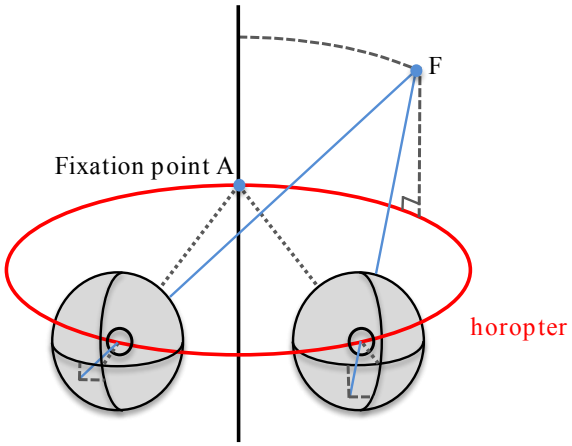


Figure 1.12: Fixation point A lies on the horopter and thus, produces zero-retinal disparity; point F is projected to the retinas with horizontal and vertical disparities (adapted from [Tyler, 2006]).

Nielsen and Poggio have examined the fusion limits of the HVS when vertical disparities are introduced into random-dot stereograms. The stimuli were presented for 117 ms. A fusion limit of 3.5 minutes of arc of vertical disparity for the central region of the stereogram was reported [Nielsen and Poggio, 1984]. However, an experiment by Stevenson and Schor extended this value to 45 minutes of arc [Stevenson and Schor, 1997], where observers could still have stereopsis and discriminate in depth. Such a difference in results was explained by the influence of the size of random-dot stereograms on the field of view.

The limit of the smallest angle of binocular disparity that can be detected by the HVS is called *stereoscopic acuity* or *stereoacuity*. Stereoacuity is lower for an uncrossed disparity than for a crossed disparity [Lam et al., 2002] and it reduces away from the horopter. The studies of Coutant et al. demonstrated that 97.3% of participants out of 188 had a stereoacuity of 2.3 minutes of arc or better [Coutant and Westheimer, 1993].

1.4.2 Accommodation and vergence limits

The eyes must converge and accommodate quite accurately to perceive objects properly. As explained in Section 1.2.1, the accommodation and vergence systems are connected

and can not operate entirely independently. For example, for a certain amount of vergence, accommodation remains within the limits of the eye's DoF. When the amount of vergence exceeds the limit of the eye's DoF, the image appears blurred and unclear. In the same manner, vergence should remain inside Panum's fusional area (0.25°-0.5°) [Watt and MacKenzie, 2013] or images will not be fused and double vision will occur. Thus, to avoid side effects it is important to understand the degree of freedom of both systems, when their responses are decoupled.

For the first time this question was addressed as the concept of the *zone of clear single binocular vision (ZCSBV)* in ophthalmological studies. The task was to find out the conditions when the patient can clearly see the set of vergence and focal stimuli using binocular vision. The various focal stimuli were designed to determine maximum convergence and divergence, when a patient focuses on an object. So the ZCSBV represents the mapped vergence ranges for the several focal distances. Inside this zone, the accommodation and vergence responses can be decoupled without any conflicts and mismatches. However, it does not mean that visual discomfort or fatigue can be avoided. This observation was made by Percival in 1982, who was prescribing spectacles with optical corrections. He proposed to use only the middle third of ZCSBV for comfortable vision [Percival, 1892]. This area was named after him as *Percival's Zone of Comfort (ZoC)*.

Shiabata et al. estimated the ZCSBV from the literature and plotted accommodation versus vergence distances in diopters [Shibata et al., 2011]. The ZoC was evaluated from the questionnaires collected during their experiment. Their results are presented in Figure 1.13 and adapted from [Watt and MacKenzie, 2013]. The ZCSBV is illustrated in light gray and ZoC in dark gray colors. accommodation and vergence are plotted in diopters (D) ($1/d$

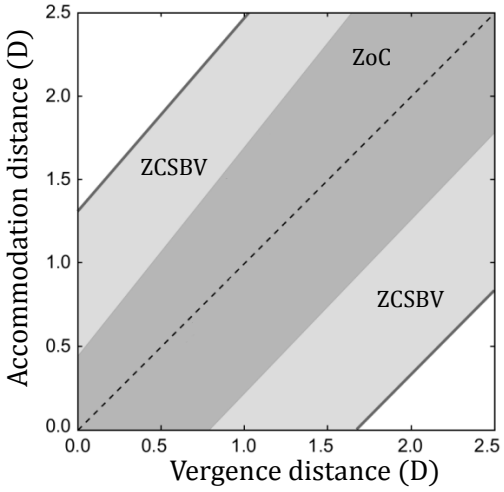
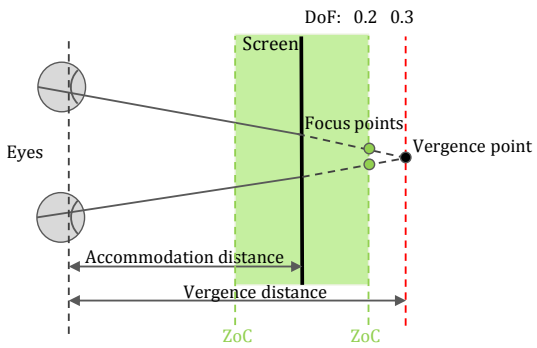


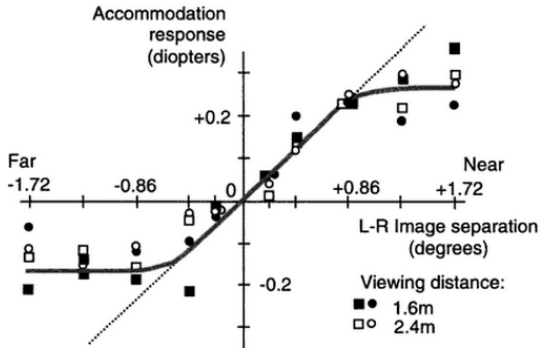
Figure 1.13: ZCSBV and ZoC estimated by Shiabata et al.(adapted from [Watt and MacKenzie, 2013]).

In stereoscopic systems, oculomotor cues do not always act in S3D systems as in real life. The difference between real-world viewing and the viewing of stereoscopic images is illustrated in Figure 1.14.a. *Depth of Field* (DoF) in this case means the amount of reconstructed depth around the screen plane [Yano et al., 2002]. The *accommodation-vergence conflict* is the mismatch between real life and an artificial environment. It

happens beyond the limits of DoF=±0.2 [Yano et al., 2004]. When a reconstructed scene is farther than this limit, the focus remains at a depth corresponding to DoF=±0.2 (see Fig 1.14.b). Hiruma et al. have demonstrated that the accommodation response is driven by the vergence and the HVS acts naturally within the zone of comfort for stereoscopic viewing [Hiruma and Fukuda, 1993].



(a) Accommodation-vergence conflict when a scene exceeds ZoC.



(b) Relationship between image disparity and accommodation response; from [Hiruma and Fukuda, 1993].

Figure 1.14: Accommodation-vergence conflict.

The vergence-accommodation in stereoscopic displays leads to visual discomfort [Lambooij et al., 2007, Yano et al., 2004, Okada et al., 2006, Eadie et al., 2000]. By using the concept of ZoC explained earlier, visual discomfort can be avoided. In extreme cases, vergence-accommodation conflict leads to diplopia, blur, or both simultaneously.

Shiabata et al. evaluated discomfort related to the vergence-accommodation conflict and approximated the ZoC from subjective scores that were obtained [Shibata et al., 2011]. The reported discomfort was affected by viewing distance and type of disparity (crossed or uncrossed). A vergence-accommodation conflict produced slightly higher discomfort for far viewing distances compared to near distances. Discomfort was higher at near viewing distances for objects with crossed disparities, while it was higher at far viewing distances for uncrossed disparities. The produced ZoC expands when viewing distance increases (see Fig. 1.15). At a close viewing distance, it is very easy to produce the accommodation-vergence conflict because the ZoC is quite narrow. However, this does not mean that accommodation-vergence conflict can be completely avoided at far viewing distances. For example, the conflict can be achieved by showing an object with crossed disparity too close to the audience at a cinema.

As it was already mentioned, the limits of the ZoC are often related to the amount of reconstructed depth range. Perceived depth falls in the range of the ZoC if the eyes accommodate within DoF=±0.2 [Yano et al., 2004, Yano et al., 2002] or less rigorously within DoF=±0.3 [Lambooij et al., 2007]. Otherwise, decoupling of accommodation and vergence happens outside of the DoF. Another approach to minimize the accommodation-vergence conflict is to control the amount of horizontal disparity presented on the screen or screen parallax, which is expressed in percentage of the screen width: 2% for uncrossed disparity and 1% for crossed disparity. This rule is often applied in practice by filmmakers [Mendiburu, 2009]. However, it does not take into account the viewing distance and the screen size. But these factors influence the produced retinal disparity and the perceived ZoC as demonstrated by Shi-

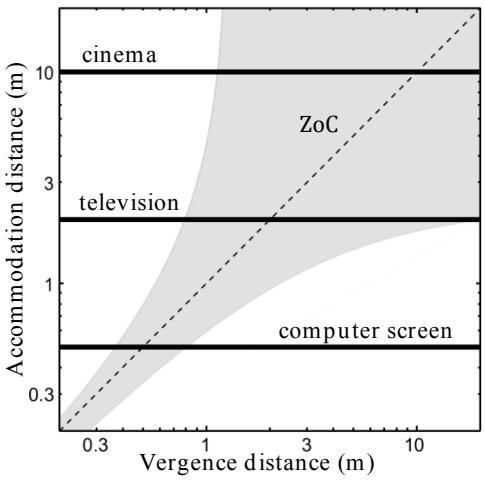


Figure 1.15: ZoC as the function of viewing distance estimated by Shiabata et al. (adapted from [Watt and MacKenzie, 2013]).

bata et.al (see Fig. 1.15). Hence this rule does not guarantee the absence of the accommodation-vergence conflict. This drawback can be avoided by using the limits in terms of retinal disparity. Limit $\pm 1^\circ$ of visual angle was reported by several studies [Yano et al., 2002, Speranza et al., 2006, Kuze and Ukai, 2008] and is considered a generalization of all the previously described approaches by the recommendation of ITU-R BT.2021 [ITU, 2012a].

Chen summarized most of the proposed limits of the comfortable viewing zone and plotted them as a function of the viewing distance as illustrated in Figure 1.16 [Chen et al., 2011]. As can be seen from the figure, the threshold of DoF=$\pm$0.2D is the most rigorous limit when compared with the others. Hence, it was recommended to use this value to be absolutely sure that visual comfort is guaranteed. This threshold was confirmed with a subjective experiment, which demonstrated that visual comfort decreased beyond this limit.

1.4.3 Extreme convergence and divergence

In real life when we look at far away objects, optical axes of our eyes are parallel and the perceived distance is optical infinity. Hence, objects with uncrossed disparity appear at an infinite distance when the positive screen parallax is equal to IPD in stereoscopic viewing systems. From the equation 1.2, the perceived depth increases with positive screen disparities for a fixed viewing distance until the disparity reaches its maximum limit compelled by IPD (see also Fig. 1.11).

However, the limit of maximum screen disparity equal to IPD may be transgressed during the shooting process with a parallel camera configuration by using a large camera baseline. The resulting disparity presented on the screen is larger than the IPD of a viewer. Therefore, the eyes have to move outward (see Fig. 1.17.a) to fuse the images. This type of unnatural situation can cause eye strain or the inability to fuse a stereopair. Typically, humans can achieve about 1° of divergence [Watt and MacKenzie, 2013]. Similarly, extreme values of negative parallax force the eyes to move excessively inward as illustrated by Figure 1.17.b leading to the same consequences [Reeve and Flock, 2010].

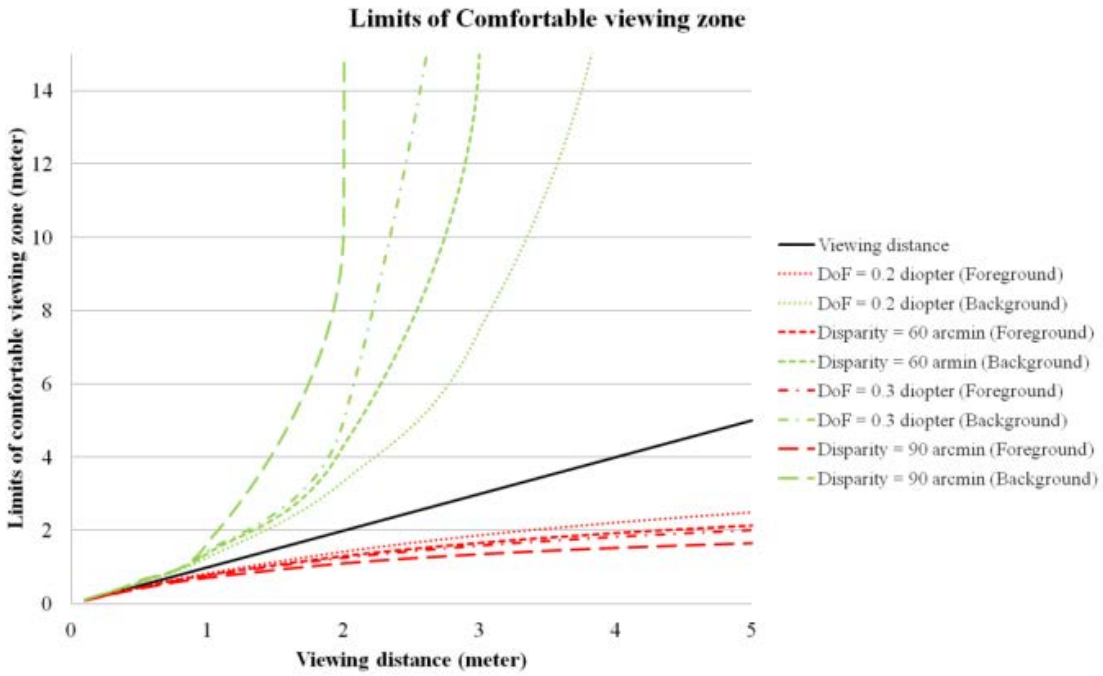
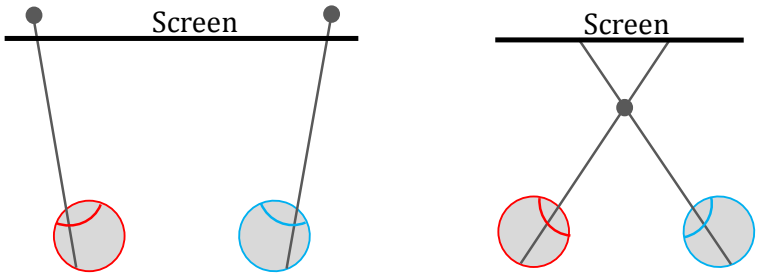


Figure 1.16: Limits of the comfortable viewing zone collected from literature by [Chen, 2012]



(b) Extreme convergence

Figure 1.17: Extreme divergence and convergence.

1.5 Conclusions

This chapter describes the principles of human depth perception and its limitations. The simplest stereoscopic system was described with its potential impact on human depth perception. To sum-up the key points of Chapter 1:

- Binocular disparity is not the only source of depth information for humans. However, in stereoscopic systems the HVS reconstructs depth mostly owing to binocular disparities.
- Sensation of depth occurs when the brain fuses two slightly different flat images. Some conflicts can happen when the discrepancies between images are too large or when the HVS is susceptible to some degree of unnatural viewing.
- The amount of perceived depth depends on screen size and viewing distance.

Therefore, when studying depth perception the viewing distance and screen disparities must be considered for the generalization of the results.

- Even simple stereoscopic systems can cause visual discomfort due to the vergence-accommodation conflict. Therefore, the reconstructed scene should remain within the comfortable viewing zone limited by DoF=$\pm$0.2 diopter.

So, all stimuli produced in this thesis will consider these guidelines.

Chapter 2

Broadcast chain of 3D systems

Contents

2.1 Introduction

Starting with Avatar in 2009, the financial success of 3D movies has promoted the development of new 3D digital technologies like *three-dimensional television (3DTV)* and digital cinema. This boom has led to advances in 3D display technologies and the emergence of stereoscopic broadcasting services, which aim to enhance user experience. However, the quality of such services can be influenced or sometimes limited by human

Figure 2.1: Stereoscopic broadcast chain (video processing chain).

This chapter discusses the impact of different technologies at every stage of the broadcast chain on final user perception in detail. The basic principles of stereoscopic content acquisition and generation, data representation and coding, and transmission and display technologies are considered below.

2.2 3D content production

There are many ways to generate stereoscopic content but the basic principle remains the same: separate the left and right points of view. There are three types of systems based on the number of cameras involved in the process of acquisition: monoscopic systems, dual-camera systems, and multi-view systems.

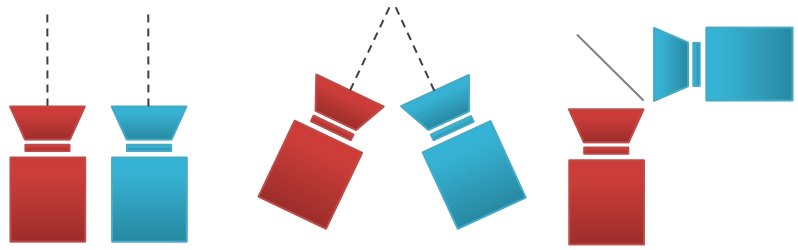
The main advantage of *monoscopic systems* is that only one traditional camera is required. Computer vision algorithms are used to reconstruct depth from a set of two-dimensional color images. Some of the developed algorithms extract depth from various monocular depth cues [Tam and Zhang, 2006]. The main interest in such systems is that the automatic or semi-automatic 2D to 3D conversion is cheaper in comparison with fully manual processing by an expert [Zhang et al., 2011]. However, the remaining challenge is the correct reconstruction of occluded objects or regions.

Another type of monoscopic system requires additional equipment for *a depth map* acquisition, which contains the values related to the depth of each pixel of a captured image [McCarthy, 2010]. Often laser or infra-red sensors are used as additional equipment. Usually, the data captured by a sensor is a monochrome 8-bit image, which is simple to store and compress [Fehn, 2003]. Unfortunately, such sensors have limited depth accuracy, a quite narrow sensing range, and a lower spatial resolution (and sometimes lower temporal resolution) than color sensors. All these problems make outdoor shooting quite problematic. Nevertheless, the challenge of occluded objects remains. Another approach computes the depth map by disparity estimation [Kauff et al., 2007, Scharstein and Szeliski, 2002].

The *stereoscopic dual-camera system* consists of two 2D cameras (or two sensors embedded in one camera body), which capture the scene from slightly different points of view [Dumbreck, 1993]. The cameras can be rigged parallel to one another, converged (*toed-in*), or even perpendicular (*mirror rig*) (see Fig. 2.2). Due to the physical size of the cameras and lenses, the minimum baseline distance with a side-by-side configuration is limited. This baseline distance limitation can be overcome by a mirror rig, where cameras are separated with a semi-transparent mirror. The light passes through the mirror and reaches horizontally placed camera, while the light reflected from the mirror hits the vertical camera. The baseline distance on a mirror rig can be reduced almost to zero, making such a configuration very suitable for close shooting. However, the mirror reduces light and a color correction of one view relative to another may be required. Independent of the configuration, the stereoscopic dual-camera system is able to provide control over the space with greater plasticity than any other system. Thus, in the next section shooting with parallel and toed-in cameras and its influence on depth perception will be discussed in detail.

Finally, a *multi-view system* is an array of more than two monoscopic cameras. Such systems result in more precise control over the space. However, calibration in optics, position, color, and luminance is required for the whole array. The systems are quite bulky due to the amount of equipment required.

Besides using camera systems, stereoscopic content can also be generated with 3D



(a) Parallel rig (b) Toed-in rig (c) Mirror rig

Figure 2.2: Stereoscopic dual-camera systems.

computer graphics techniques [Laszlo Szirmay-Kalos, 1996, Watt, 2000]. Software such as Blender, 3ds Max, Maya, and many others provide possibilities to create synthetic scenes. Different views are rendered using virtual cameras. Very accurate depth maps can also be easily created. The main convenience of such an approach is that virtual cameras have no physical restrictions on camera parameters (focal length, sensor size, etc.) and location within synthetic scenes. In addition, they are free from lenses with optical flaws. Synthetic scenes are an alternative to real shooting. On the other hand, rendering large projects is very time consuming and requires powerful equipment.

2.2.1 Controlling the perceived depth with a dual-camera configuration

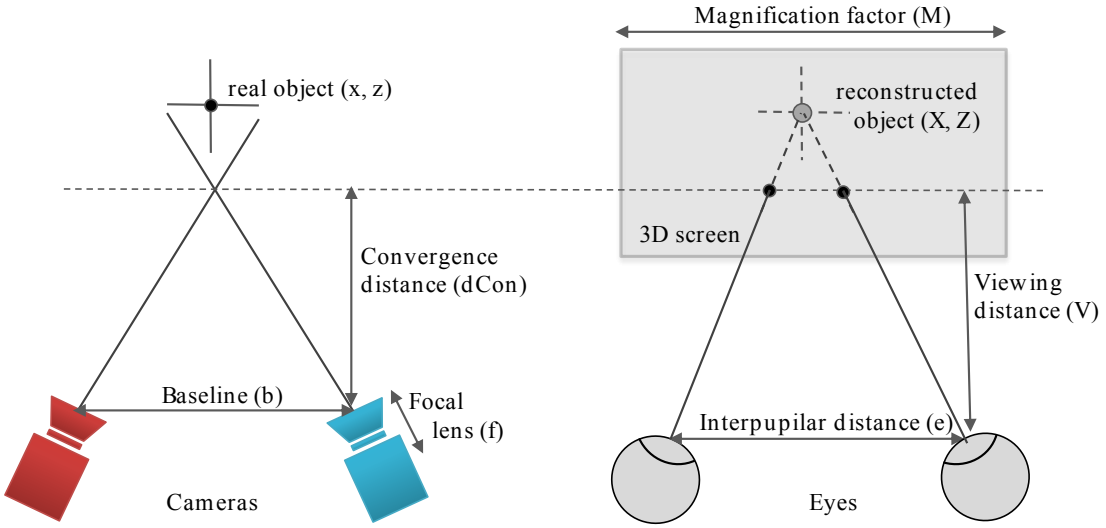


Figure 2.3: Camera and visualization spaces.

For example, it is possible to dynamically increase and decrease the amount of depth in a scene by modifying the baseline distance between cameras. The larger the baseline is, the larger the separation between a pair of displayed stereoscopic images is. Hence,

more depth will be perceived and the objects will be more distant from each other. Another reason to regulate the baseline is to manage the field of view of the scene space, which might change due to the visualization environment (display size and viewing distance)[Collins et al., 2011].

The convergence point of two cameras defines the position of the objects relative to the screen but it does not change the amount of total depth in the scene. Objects located in front of the convergence point appear in front of the display plane and have negative parallax, while objects farther than the convergence point appear behind the

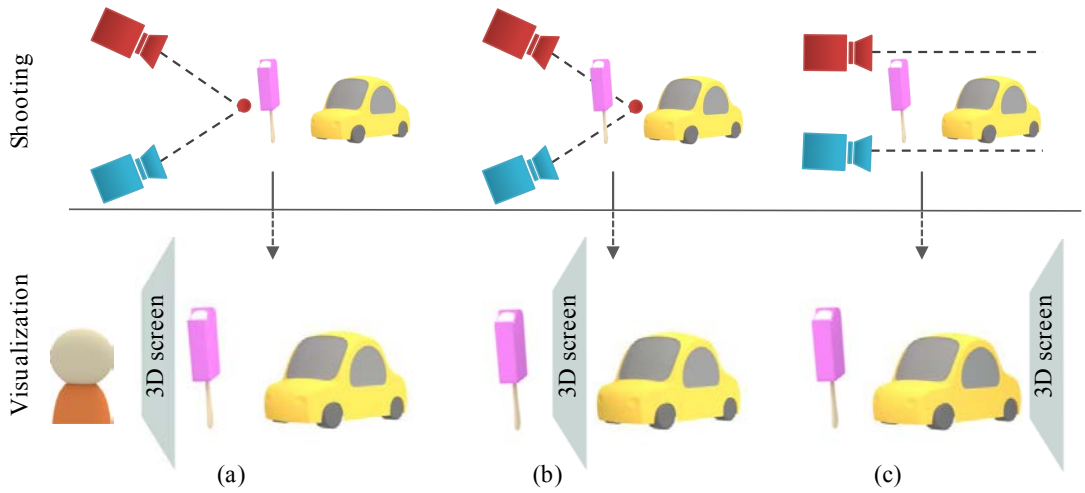


Figure 2.4: The effect of convergence distance on perceived depth. (a) The cameras are converging in front of the car – the scene is viewed behind the display plane. (b) The cameras are converging behind the ice-cream – the ice-cream pops-out of the display, while the car remains behind the display plane. (c) The cameras are parallel. The convergence point is infinity. The scene is displayed in front of the display plane.

The convergence point can be easily controlled by using a toed-in camera configuration, where two cameras converge and diverge similarly to the HVS. However, noticeable geometrical distortions can be produced. When shooting with parallel cameras (without toe-in), the convergence point is infinity and all objects obtain negative parallax. Post-production is required to adapt convergence point using *Horizontal Image Translation (HIT)*.

In addition, perceived depth may be affected by the choice of the camera's focal lens. A long focal lens increases the size of objects and decreases the field of view of a scene [Ijsselsteijn et al., 2000]. Consequently, higher disparities are captured and the background is perceived as being further away. The baseline might be reduced if the created depth is outside the comfort zone. However, this scenario may cause the cardboard effect (see Section 1.3.3.1). The opposite is true for short lenses.

In the visualization space, the perceived depth depends on the IPD, the viewing distance, and the display size, as illustrated in Figure 2.3 (for more details see Section 1.3.1 and Section 1.3.2). The geometrical relationship between the parallel camera space and the final visualization space was derived by Woods et. al. [Woods et al., 1993] as following:

$$Z = \frac{Vez}{ez + Mfb(1 - \frac{z}{dCon})} \tag{2.1}$$

$$X = \frac{Mefx}{ez + Mfb(1 - \frac{z}{dCon})} \tag{2.2}$$

$$Y = \frac{Mefy}{ez + Mfb(1 - \frac{z}{dCon})} \tag{2.3}$$

where (X, Y, Z) is the location of the point in the visualization space, as seen by an observer when displayed on a screen, (x, y, z) is the location of the same point in the camera space (in front of the camera), e – IPD, f– camera focal length, b – camera baseline distance, $dCon$ – convergence distance, and M – magnification factor.

The magnification factor is defined as:

$$M = \frac{W_{screen}}{w_{sensor}} \tag{2.4}$$

where W_{screen} is the screen width, and w_{sensor} – camera sensor width.

Chen et al. in [Chen et al., 2011] defined the local depth variation around the depth plane as the derivative of Z in the visualization space in respect to z of the camera space from the equation 2.1:

$$D_z = \frac{dZ}{dz} = \frac{VeMfb}{ez + Mfb(1 - \frac{z}{dCon})^2} \tag{2.5}$$

D_z identifies the depth distortion around the depth plane z as a function of shooting and visualization parameters. Viewers perceive objects being at the same distance in depth as real life for stereoscopic images when D_z is equal to 1.

In the same way, the derivatives of X and Y in the visualization space relative to x and y are defined from the equations 2.2 and 2.3:

$$D_x = \frac{dX}{dx} = \frac{Mef}{ez + Mfb(1 - \frac{z}{dCon})^2} \tag{2.6}$$

$$D_y = \frac{dY}{dy} = \frac{Mfb}{ez + Mfb(1 - \frac{z}{dCon})^2} \tag{2.7}$$

These equations describe the change of image size on the x and y axes respectively without considering the depth component.

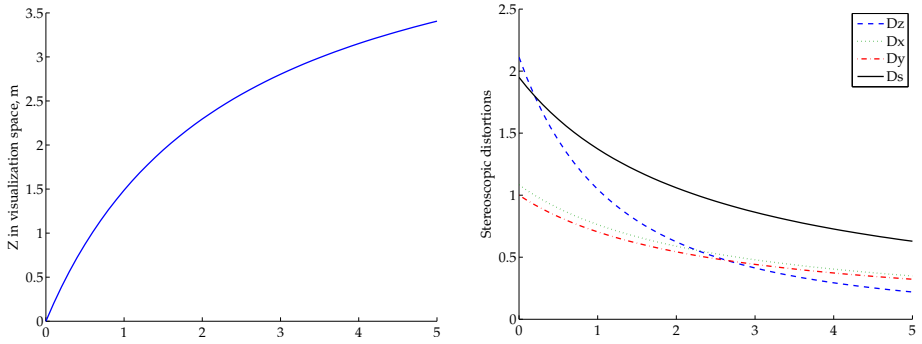
The factor for characterizing the shape distortions of an object in the visualization space is defined as follows:

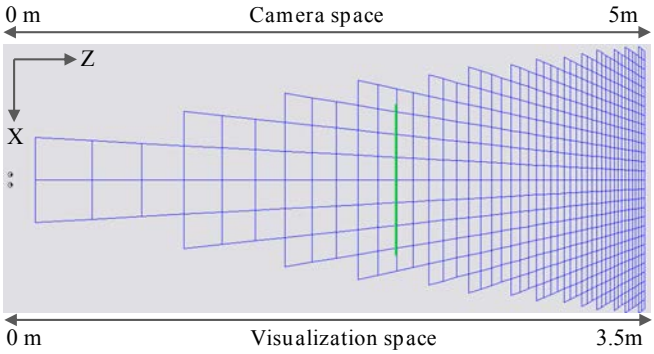
$$D_s = \frac{D_z}{D_x} = \frac{Vb}{ex + Mfb(1 - \frac{z}{dCon})} \tag{2.8}$$

where V is visualization distance. If this ratio is equal to 1, the object's shape in the visualization space is proportional to the camera space, otherwise a stereoscopic shape distortion appears. If $D_s > 1$, then the shape is stretched along the depth axis in the visualization space; if $D_s < 1$, the shape is compressed in depth in comparison with the camera space. Therefore, as follows from this equation, it is possible to influence the shape of objects in the visualization space by controlling the shooting parameters. This influence was demonstrated by numerous studies

[Ijsselsteijn et al., 2000, Woods et al., 1993, Goldmann et al., 2010a]. However, this information does not provide any guidance on how to choose the camera parameters to avoid distortions in the stereoscopic content acquisition. This issue will be addressed in the section 2.2.3 "3D Shooting Rules".

An example of the relationship between the camera and the visualization spaces is presented in Figure 2.5. The curves are plotted for the stereoscopic camcorder Panasonic AG-3DA1, which uses a static interaxial distance of $b = 60mm$, a focal length distance of $f = 4.2mm$, $w_{sensor} = 3.2mm$, and a convergence distance of $dCon = 2.14m$. This type of camera is pre-calibrated during the manufacturing stage, which allows view asymmetries to be avoided. The relationship is computed for a scene, which is situated between 0-5 m in camera space, and visualized with a display width of $W_{display} = 93cm$ and a viewing distance of $V = 2.38m$. Figure 2.5.a allows the percieved distance in the visualization space in comparison with the camera space to be estimated. For example, 1 meter in camera space appears to be 1.5 meters away from the viewer. Figure 2.5.b illustrates the depth distortions. At a distance of $z < 2.3$ meters D_s is bigger than 1 so the shape of visualized objects is stretched in depth, whereas for a distance of $z > 2.3$ meters, the shape of the visualized objects is compressed.





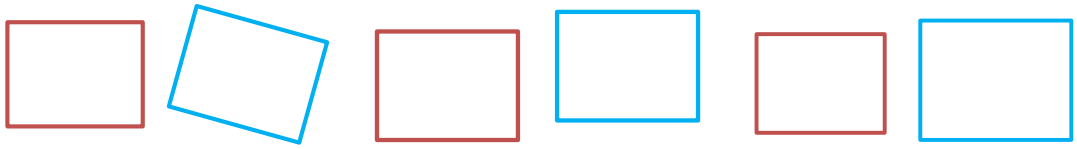
(c) Illustration of the shape distortion in the visualization space (the dimension of each rectangle is 0.25 × 0.25 m; solid green line - display plane)

Figure 2.5: Relationship between camera space and visualization space: f=4.2mm, CD=2.14m, b=60mm, B=65mm, $w_{sensor} = 3.2mm$,$W_{display} = 93cm$, V=2.38m.

2.2.2 View alignment problems

Stereoscopic shooting is quite a delicate process and requires a high degree of accuracy because numerous artifacts and view impairments may be produced during capture. These view alignment issues can be separated into several different categories:

- *Optical asymmetries* occur when the focal setting of one camera is different than the other. As result, the left or right view may be magnified (see Fig. 2.6.c).
- *Geometrical asymmetries.* Typically there are three possible geometrical misalignments of cameras during 3D acquisition. A vertical misalignment is created when one camera is shifted vertically in comparison to the other camera (Fig. 2.6.b). A *rotation* misalignment is produced when one camera is tilted more than the other one (see Fig. 2.6.a). *Keystone distortion* is generated by toed-in camera rigs, which have different orientations in space (Fig. 2.7). As a result, each image of the stereo pair is projected as a trapezoid to a sensor and then vertical disparities are generated. Unlike keystone distortion, depth plane curvature is a result of the wrong horizontal disparities, which are created because the inner parts of the sensors are closer to the camera baseline and thus the scene is captured from a farther distance than by the outer parts. Both distortions are strongest in the corners of the stereopair.



(a) Rotation (b) Vertical shift (c) Magnification (difference of focal lens)

Figure 2.6: View asymmetries.

- *Color* and *luminance asymmetries.* These variations are inherent to the lenses, mirrors, and cameras. The white and black levels and color gamut must be adapted when a two cameras configuration is used. Otherwise, transparent mirrors in the mirror rig can cause discrepancies in colors and luminance due to the difference in reflectance and the transmission of beam-splitters.

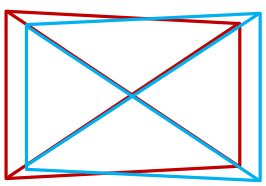


Figure 2.7: Keystone distortion.

The advantages and drawbacks of camera systems and different rigs with a predisposition to view asymmetries are summarized in Table 2.1.

Table 2.1: Advantages and disadvantages of camera systems regarding view asymmetries

Rigs	Advantages	Drawbacks
2D + sensor	Free of geometrical, luminance and color distortions.	Requires a sensor to capture depth map. The sensors often have a lower spatial resolution than color sensors, limited depth accuracy, a quite narrow sensing range, and, sometimes, a lower temporal resolution than a 2D camera sensor. The occluded areas should be reconstructed.
Toed-in	Due to the imitation the HVS convergence of the HVS. Hence it is easier to avoid the window violation effect. Does not require the HIT with post processing and due to this is able to keep a landscape ratio.	Two cameras can produce various geometrical, color and luminance asymmetries.
Parallel	No keystone distortion or depth plane curvature.	Two cameras can produce various geometrical, color and luminance asymmetries. The minimum interaxial distance can not be less than the camera's width. Hence, the zoom is required to make close and macro shots. Requires HIT with post-production; hence a wide-aspect ratio is lost.
Beam-splitter rig (half-mirorr)	The interaxial distance can be reduced to 0.	This system is quite bulky, heavy, and fragile. Mirrors are sensitive to dust and fast accelerations. Keystone distortion can be produced due to inaccurate mirror placement. Color and luminance asymmetries can be created due to the difference in reflectance and transmission of beam splitters. The images should be flipped to the original orientation before visualization.
Static interaxial	Imitates the HVS by having a baseline distance equal to the IPD. The cameras are usually compact and simple to manage. Cameras are pre-calibrated by the manufacturer. So it is supposed that there are no geometrical, luminance, or color asymmetries.	These cameras are mostly suitable for indoor scenes due to the fixed interocular distance.
Multi camera	More precise control over the space.	The systems are quite bulky and expensive due to the amount of equipment required. The luminance, color, and geometrical alignment are required for the whole optical array.

2.2.3 3D shooting rules

In this section the rules for stereoscopic video acquisition are divided into two major categories: the cinematographic and scientific rules. The reason for such a division is that movie creation is considered an artistic process in cinematography. That is why directors and cinematographers are graphic artists that "do not want to create visual mediums by blindly relying on trigonometric formulas" as explained by Mendiburu in [Mendiburu, 2009]. On the other hand, scientists have derived formulas to predict the perceived depth, compute the zone of comfort, and avoid space distortions. Anyhow, both groups aim to produce comfortable and pleasant to watch stereoscopic content.

Numerous books about 3D cinema provide tutorials on stereoscopy [Mendiburu, 2009, Lipton, 1997, Kaminsky, 2011, Zone, 2013]. The stereoscopic rules that are usually described come from experience and are quite simple to follow during the acquisition process. For example, *1/30 rule* states that the camera baseline should be no more than 1/30 of the distance to the foreground (nearest object in the scene). The main goal of this rule is to keep the scene within the comfort zone by reducing excessive binocular disparity. For cinema shooting, the camera baseline should be changed to 1/100 and for very short lenses to 1/10 of the distance to the foreground. Another recommendation is to select a focal length under 30 mm, because, according to Mendiburu, "long lenses make poor 3D; short lenses make great 3D" [Mendiburu, 2009]. Basically, by using short lenses, it is easier to preserve the original proportions of the objects in space. In addition, it is recommended to check the produced stereoscopic video on the target display size, which might be simply impossible for small budgets. Besides, all the rules listed above only provide rough empirical estimation of camera parameters.

The fundamental work about the effect of stereoscopic camera shooting parameters and display systems on the plane curvature, depth non-linearity, depth and size magnification, shearing distortion, and keystone distortion was presented by Woods et al. [Woods et al., 1993]. It has been demonstrated that the correct choice of system parameters helps to get rid of some distortions. For instance, to avoid geometrical distortions like the puppet-theater effect (see Section 1.3.3.2), keystone distortion, and depth plane curvature, it is recommended to avoid the use of a toed-in camera rig. A parallel camera rig was recommended as the solution against geometrical distortions produced by a toed-in rig. This was confirmed by Yamanoue et al. who demonstrated that the parallel rig preserves linearity during the conversion from real space to stereoscopic images [Yamanoue, 2006]. However, no quantitative guidance was given by Woods et al. on how to avoid discomfort caused by distortions.

Jones et al. presented the method to calculate the camera baseline using the relationship between the camera and scene parameters and visualization environment (display size and position of viewer) [Jones et al., 2001]. Therefore, the space captured by the camera is mapped to a perceived depth range for a given stereoscopic display. This approach suggests controlling stereoscopic distortions by changing the position of the camera or viewer. Likewise, the algorithm to reduce stereoscopic distortions was proposed by Holliman et al. [Holliman, 2004]. This algorithm optimizes perceived depth in the region of interest by considering the DoF around the display plane, e.g. comfortable viewing. Moreover, Chen et al. considered stereoscopic distortions and visual comfort and proposed a new stereoscopic video shooting rule [Chen et al., 2011], which states that in order to optimize camera parameters, keep the perceived depth range within a comfortable viewing zone prior to the optimization of stereoscopic distortions, if the two

conditions can not be fulfilled simultaneously.

Finally, Gunnewiek and Vandewalle have presented mathematical methods to display stereoscopic content realistically, which take into account the camera parameters of the original content and viewing conditions [Gunnewiek and Vandewalle, 2010]. It was recommended to use the same ratio between the focal length of the camera and the width of the sensor as between the viewing distance and the width of the display. Basically, the field of view during scene acquisition should match the field of view during rendering.

2.3 3D visualization

It is important to keep in mind that the technology behind 3D displays has strengths and weaknesses in producing high quality 3D content. This should be taken into account in order to understand how different displays influence the viewer's experience. However, this thesis only considers stereoscopic systems that take advantage of human binocular vision by presenting stereopairs of 2D plane images. There are three main groups of such 3D displays (see Fig. 2.8 for a taxonomy): (1) direct view stereoscopic displays, which require eyewear and may be classified based on the multiplexing scheme: color multiplexed, polarization multiplexed, and time multiplexed; (2) binocular head-mounted displays, which are built into the eyewear itself; and (3) autostereoscopic direct-view

Figure 2.8: Classification of planar stereoscopic displays.

2.3.1 Displays with color multiplexed approach

The color multiplexed approach is based on the color filtering of the left and right views. The stereopair is superimposed simultaneously and then the content is displayed on any standard 2D screen. To perceive 3D, the viewer should wear glasses with color filters so that the left and right eyes receive only the corresponding images. There are three possible options for color filters: red–cyan, yellow–blue and green–magenta as illustrated in Figure 2.9. Historically, the first anaglyph glasses were red-cyan glasses. Yellow-blue glasses transmit light unevenly; hence, a neutral density filter is required on top of the yellow filter. Lastly, green-magenta glasses transmit light more uniformly than the other two options.

Compatibility with any existing color display and a low price for glasses are the main advantages of the color multiplexing technique. This technique, however, has poor color

(a) Red–cyan (b) Green–magenta (c) Yellow–blue

Figure 2.9: 3D glasses with color filters.

performance and a low separation power between the views. So it is recommended to use it only when 3D display is not available.

2.3.2 Displays with polarization multiplexed approach

In this approach, light polarizations of the left and right views are mutually orthogonal. There are two types of content polarization: linear or circular (see Fig. 2.10). For the visualization, glasses and views with the same direction of polarization are required. The unintended view is blocked when its polarization direction is orthogonal to the polarization of the glasses. Unfortunately, both linear and circular polarizations are sensitive to the position of the viewer's head. However, circular polarization provides a larger degree of freedom in tilting the head to the left or the right. Similar to anaglyph glasses, the production of polarized glasses is quite cheap but the polarized approach preserves a full color image.

(a) Linear polarization (b) Linear polarization (c) Circular polarization

Figure 2.10: Polarized 3D glasses.

In cinema, 3D images can be displayed using two-projector set-ups with polarization filters, where the content preserves the full spatial and temporal resolutions. But the cost and visualization imperfections increase due to the difficulty in aligning the two projectors. Instead, a single projector with a polarization switch placed in front of the projection lens can be used. In stereoscopic 3DTV solutions, a polarization filter is placed in front of the panel. Hence, the left and right views are interleaved and polarized on a line or column basis for multiplexing. As a result, half of the vertical or horizontal resolution per view is lost. The display filter should be accurately mounted on the panel. Otherwise, undesired light leakage may happen when the filter is not fully orthogonal to the glasses filters. Apart from this, brightness is reduced due to the polarization filter even in 2D mode [Borel and Doyen, 2013].

2.3.3 Displays with time multiplexed approach

The time multiplexed approach alternates the left and right views of S3D content at high frame rates to separate the views. The frame rate is twice the normal frame rate while displaying, therefore the HVS is incapable of tracking such a high frequency change from

one view to another and the stereoscopic video is perceived smoothly. For 3D visualization, the viewer has to wear active glasses, which are synchronized to the display using either infrared or radio commands. The glasses are composed of thin LCD films that block one view from reaching the other during a certain period of time. In comparison with polarized glasses or color glasses, this solution is more expensive to produce, requires a battery and is heavier to wear. However, it is less sensitive to the tilt of the viewer's head. Moreover, full spatial resolution is preserved in television applications but only half of the temporal resolution is kept in the case of format compatibility. In a cinema, an infrared emitter usually synchronizes the active shutter glasses with the stereoscopic content delivered by the projector.

2.3.4 Head Mounted Displays (HMD)

Head-mounted displays are binocular systems which consist of two separated mini displays with linked relay optics. Since the device is worn on the head, the user is not attached to a specific viewing position and can perceive full immersion from the displayed scene. However, it is not easy to provide a large field of view and high resolution at the same time [Ferrin, 1999]. Another challenge of the HMD design is the precise calibrations between the two displays in terms of luminance, color, and geometry.

2.3.5 Autostereoscopic displays

Another group of stereoscopic displays is autostereoscopic that do not require any eyewear. There are different types of autostereoscopic displays: (1) two-view (binocular) displays, which display a single stereopair, (2) multiview displays that generate multiple stereopairs for viewing by several users, and (3) super multiview displays, where the number of presented images is large enough to provide the user with a sensation of continuous motion parallax. In this section, only the technologies of two-view display systems will be discussed. The information about the other two types can be found in [Urey et al., 2011].

The main idea behind two-view systems is to display a single stereopair in a way that the left and right views are delivered to the corresponding eyes. However, the viewing distance and location of the viewer's eyes are restricted by a particular zone or *viewing window*, where a sensation of stereopsis is possible. Two-view systems with head tracking or eye-tracking are capable of providing more flexibility in regards to the viewer's position or for displaying the same stereopair to multiple viewers. The parallax barrier and lenticular sheet are two state-of-the-art optical elements for the generation of viewing windows. If both barriers are electronically controllable, the system can easily switch from 3D to 2D and vice versa.

The *parallax barrier* is composed of vertical slots separated by strips of black mask, which block the radiation of light in undesired directions. Another more expensive solution is using another LCD panel instead of the black stripes. The black mask is created by controlling the pixels of the top panel. In both cases, only half of the pixels are visible to the right eye and the other half only to the left. The underlying principle of such a barrier is illustrated in Figure 2.11. The optimal viewing distance is directly proportional to the distance between the display and the parallax barrier and inversely proportional to the display pixel size [Urey et al., 2011]. Such barriers can be mounted in the screens of laptops, tablets, or mobile phones. It is just enough to manufacture the plastic films with black stripes. Also, parallax barriers are adaptable to any new

technologies,

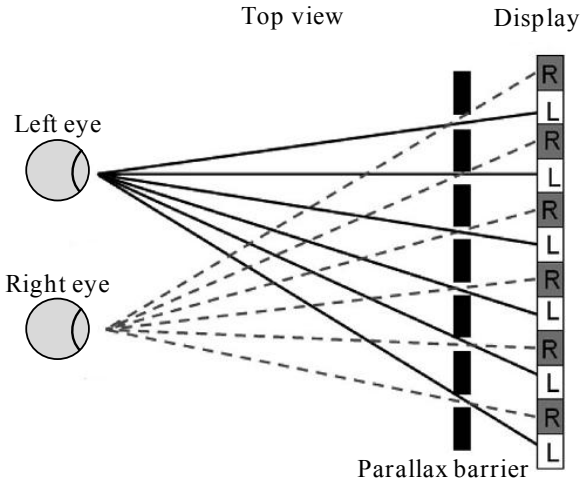


Figure 2.11: Working principle of parallax barrier (adapted from [Urey et al., 2011]).

A *lenticular sheet* can consist of cylindrical lenses that direct radiated light from a pixel in specific directions (see Fig. 2.12). Therefore, the diffused light can only be seen from one particular angle, which dramatically limits the viewing position in front of the displ
pane

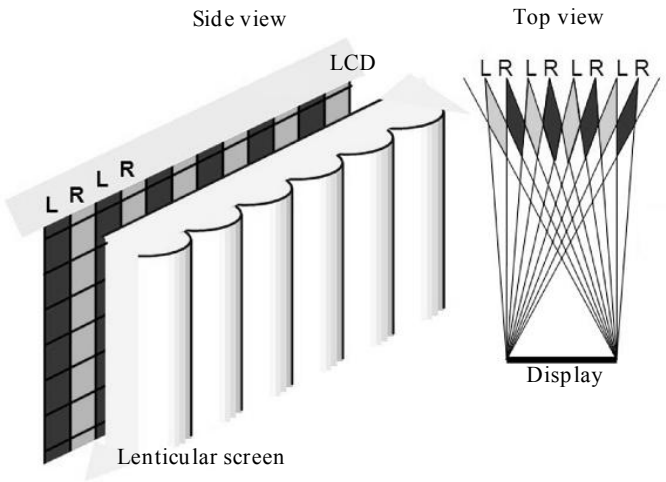


Figure 2.12: Underlying principle of a lenticular sheet (adapted from [Urey et al., 2011]).

Autostereoscopic displays are glasses-free which is quite convenient for end-users, but the main drawback of parallax barrier and lenticular sheet is that a wrong position in front of the display may lead to a pseudoscopic image (see Section 1.3.3.4) or the loss of depth perception.

To summarize, a short overview of 3D display technologies with their advantages and drawbacks is presented in Table 2.2.

Table 2.2: A short overview of advantages and drawbacks of 3D display technologies

3D display technology	Advantages	Drawbacks
Stereoscopic (with glasses)		
Anaglyph	Compatibility with legacy broadcast chain in terms of visualization, representation, and transmission.	Color and luminance asymmetries.
Polarized	Switching between 2D and 3D mode, full temporal resolution, no flicker, no batteries, and light glasses.	Loss of spatial resolution, luminance, sensitivity to viewer's head position, crosstalk.
Active shutters	Switching between 2D and 3D mode, full spatial resolution.	Loss of temporal frequency, loss of luminance, crosstalk, loss of synchronization, flicker, heavy glasses in comparison with polarized solution.
HMD	Full spatial and temporal resolution, full immersion with a scene.	Difficulty to provide large FoV and high resolution. Requires precise calibration in terms of luminance, color, and geometry.
Autostereoscopic (without glasses)		
Parallax barrier and lenticular sheet	No glasses.	Loss of spatial resolution, luminance, limited viewing position.
Multiview autostereoscopic	No glasses and several viewers in the same time.	Limited single view resolution.
Autostereoscopic with eye tracking or motion sensor	No glasses, support of motion parallax.	High complexity to calculate the precise position and generate synthetic views.

2.3.6 Visualization artifacts

Unfortunately none of the technologies listed above are free of flaws, which cause visualization artifacts in addition to other drawbacks inherent to 3D acquisition. All of the view misalignment problems discussed in Section 2.2.2 can potentially occur when two mechanical projectors are used to present an image to each eye. Often low image luminance and contrast are intrinsic to stereoscopic visualization due to light losses in filter-based systems and systems with glasses. In LCD panels, contrast reduction is the result of backlight leakage from pixels that are turned off. Furthermore, it influences stereoacuity and thus the ability to distinguish fine stereoscopic details [Legge and Yuanchao, 1989].

The most common problem of all stereoscopic displays is probably *crosstalk* or *ghosting*, which is the result of an imperfect separation of the left and right views in the stereoscopic system. Hence the part of the signal intended for the left eye leaks into the right eye or vice versa. As a consequence, a ghost image or double contours can be perceived by viewers (see Fig. 2.13). Even more crosstalk is introduced with a higher contrast of the content and disparity values [Boev et al., 2008]. In addition, small levels of crosstalk may reduce the amount of perceived depth, but high levels reduce the viewer's comfort [Watt and MacKenzie, 2013].

Among the sources of crosstalk are display persistence, an imperfect combination of the eye filters, the viewer being incorrectly positioned in front of the display, and tilt of the head in cases of linear polarization. A detailed overview of hardware-related reasons of the appearance of crosstalk is given by Woods in [Woods, 2012].

(a) 3D image without ghosting

(b) 3D image with ghosting

(c) Crosstalk: difference of images a and b

Figure 2.13: The effect of crosstalk.

Sheer distortion, the picket-fence effect, and image flipping are other visualization artifacts linked with technologies and a change in position of the viewer. *Sheer distortion* occurs in planar stereoscopic displays when the viewer moves laterally in front of the display plane. Therefore, wrong head parallax can be introduced due to the fact that the same stereopair is delivered to the eyes along the movement. This leads to an incorrect perception of object distances and a false perception of an object's motion [Woods et al., 1993]. This effect can be minimized in multi-view autostereoscopic systems, where several views can be seen along the movement [Meesters et al., 2004]. However, sheer distortion may still occur when the sampling between views is large.

The *picket-fence effect* is an issue inherent to some autostereoscopic displays with parallax barriers (see Section 2.3.5). The black matrix between the pixels creates visible gaps. Hence, periodically brighter and darker vertical stripes are perceived when the observer moves sideways in front of the screen. Such an effect can also be generated by

optical filters for spatial multiplexing. It is possible to lessen their visibility by tilting the optical filter in respect to the display panel [Berkel and Clarke, 1997].

The introduction of head tracking helps to minimize all artifacts connected with a change of the viewer's position [Woods et al., 1993, Meesters et al., 2004].

Another group of issues is connected to the temporal multiplexing of the signals in 3D. For example, *flicker* is produced by temporal variations in the luminance of an object or a scene [McCarthy, 2010]. In field-sequential stereo displays it becomes visible when the scene presented to the eye is replaced by a dark interval slow enough to be detected by the HVS. The visibility of flicker does not depend on the capture rate but only on the presentation rate [Hoffman et al., 2011]. Thus, to reduce its visibility, double-flash and triple-flash protocols, where the image is presented two or three times before refreshing, are used. However, multi-flash protocols can create blur: when smooth-pursuit eye movements are made, the same image pair is delivered to the eye several times, hence the same object is projected onto different retinal parts [Banks et al., 2012].

Finally, the moving object can be improperly perceived in depth due to a temporal delay of the input to one of the eyes. To begin with, the creation of a stereopair requires the simultaneous matching of the left and right views, but the left and right images are not available at the same time in sequential stereo displays. Thus, periodically, the HVS matches an image from the left eye with the image from the right eye that was captured at different times. Hence, an incorrect disparity would be produced for an object moving in depth, resulting in a pseudoscopic image [Watt and MacKenzie, 2013]. Another possible consequence is that objects moving in the same direction seem to be closer, while objects moving in opposite directions seem to be farther away than they should be [Banks et al., 2012].

2.4 3D representation, coding and transmission

3D video representation has an influence on the broadcast chain because it is inseparable from content acquisition, transmission, coding, displaying, and end-user perception. In this section, the S3D video format, 2D-plus-depth format, multiview video format (MVV), multiview video-plus-depth (MVD), and layered depth video (LDV) with its coding schemes will be described briefly. A detailed review of the existing formats can be found in the literature [Vetro, 2010, Gautier et al., 2010, Muller et al., 2010].

2.4.1 3D representation formats

Conventional S3D video format. A stereoscopic 3D video format is the most popular and simplest format owing to compatibility with existing encoders, transmission channels, and receivers. It consists of two views devoted for each eye. Two times more raw data is produced during shooting in comparison with a regular 2D video. The *frame-compatible stereo formats*is the way to represent S3D content. During this process the left and right views are multiplexed spatially or temporally into a single frame or a sequence of frames [Vetro, 2010]. With spatial multiplexing, the left and right views are sub-sampled and then interleaved into a single frame. Several sub-sampling patterns are available: *side-by-side, top-bottom, checkerboard, and column and raw interleaved.* In the case of temporal multiplexing, a single frame can be extracted by interleaving the left and right views of alternated frames. Unfortunately, both types of view multiplexing lead to the loss of spatial or temporal resolution. Due to the view mixture, frame-

compatible formats may cause color bleeding artifacts [Cagnazzo et al., 2013]. Mostly frame-compatible formats are used for 3DTV broadcasting services.

In order to keep full resolution, the *frame packing* format was designed. The key difference with the previously described formats is that each sub-frame maintains the original full resolution. This format consists of two full resolution sub-frames which are stacked horizontally or vertically. A single packed frame is decoded by splitting it into left and right views and then displaying them in a frame sequential manner.

The advantage of S3D video representation is the compatibility with stereoscopic displays with passive or active systems. However, at the shooting stage of the S3D content, the target display size is predefined. Thus, the viewing distance for S3D video representation is quite constrained. A change in the display size may produce content distortion and/or discomfort.

2D-plus-depth video format consists of only one view with a linked 3D map (see Fig. 2.14). This format is facing the problems generated by monoscopic systems while capturing. The information from occluded areas is missing and it can be reconstructed with the help of inpainting algorithms [Grossauer and Scherzer, 2003, Bertalmio et al., 2000, Guillemot and Le Meur, 2014]. The representation of this format is fully compatible with any type of conventional displays.

(a) 2D video (b) Depth map

Figure 2.14: 2D-plus-depth: 2D video and corresponding depth map.

Multiview video format (MVV) is composed of N views captured from slightly different view points. This format is suitable for autostereoscopic displays, where multiple views are displayed at the same time but the viewer sees only a pair of adjacent views depending on the viewing angle. MVV can be easily converted to 2D or S3D video format by extracting one or a pair of views.

Multiview video-plus-depth (MVD). It is another multiview format that is basically the synthesis of 2D-plus-depth format and MVV: each of the N views is captured from slightly different points with its associated depth (see Fig. 2.15). Thus, the accuracy of reconstruction is higher than for any other format.

Layered depth video (LDV) is illustrated in Figure 2.16. The first concept of a layered depth image (LDI) to describe a 3D scene was introduced by Shade et al [Shade et al., 1998]. The idea is to record for every pixel not only depth and color information, but also information about the occluded pixels in the next layer. Thus, LVD can be considered as a sequence of LDI. The first layer of LDV can serve as 2D-plus-depth format.

Figure 2.15: Multiview video-plus-depth: N views and corresponding depth maps from [Cagnazzo et al., 2013]

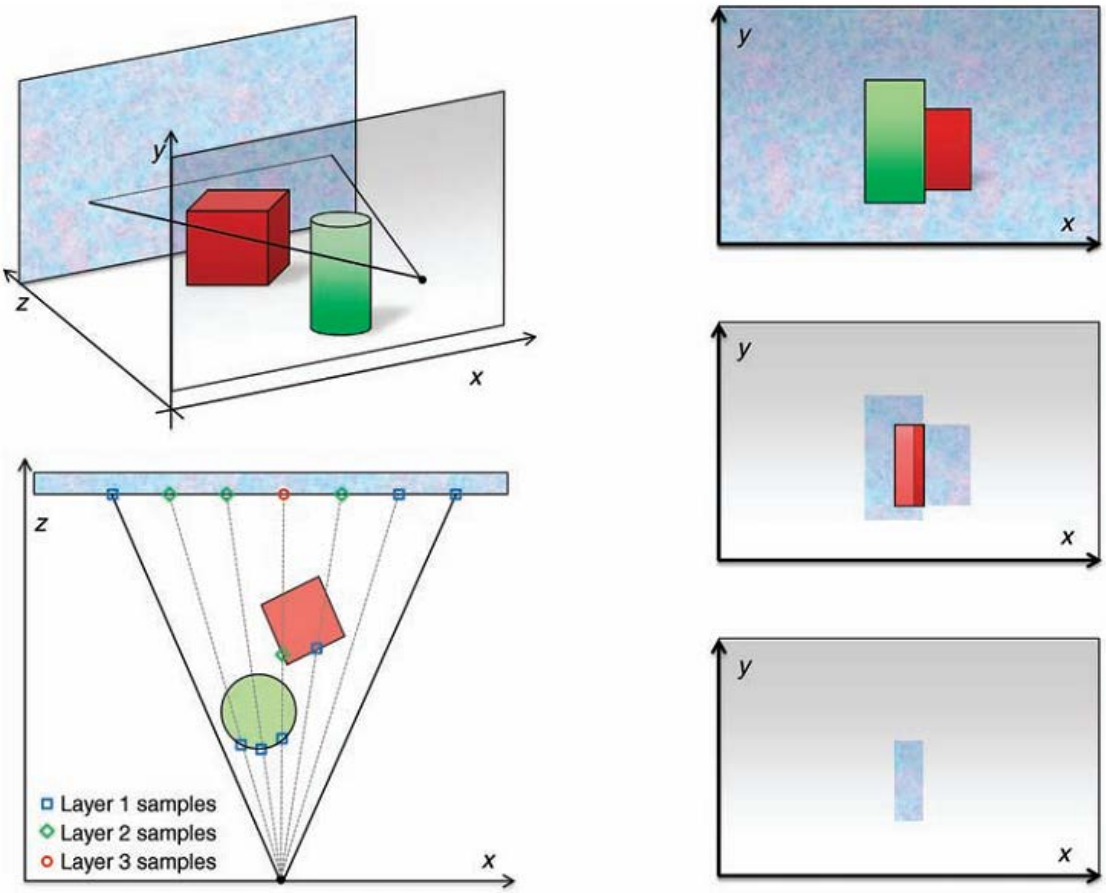


Figure 2.16: The layered depth image representation. On the left: top - 3D scene, bottom - view points. On the right from the top: the first, second, and last layers from [Cagnazzo et al., 2013]

The summary of video formats with their advantages and drawbacks is presented in Table 2.3. Frame compatible formats are the most widely used formats due to them being supported by 3DTVs available on the market.

Table 2.3: 3D video formats with their advantages and drawbacks

Format	Principle	Advantage	Disadvantage
Frame sequential	Left and right views are alternated in consecutive frames	Compatible with legacy displays, transmission and decoders. No loss of spatial resolution	Sometimes double bitrate is required in comparison with frame compatible format
Frame compatible	Spatial multiplexing of the left and right views	Compatible with legacy displays, transmission and decoders	Loss of spatial resolution
Frame packing	Left and right views stacked together in a single frame	Preservers full spatial resolution	Supported only by HDMI 1.4 compliant displays
2D-plus-depth	The depth map is used to reconstruct render two views	The depth map is easy to store, compress, and transmit	Occluded areas are not available during the rendering
MVV	More than two views are captured by a multi-camera system	Support of free-viewpoint displays	Increase of the bitrate requirements.
MVD	More than two views are captured by a multi-camera system with a depth map	Less views are required in comparison with MVD since a depth map is available	Occluded areas are not available during the rendering
LDV	First layer is 2D-plus-depth and other layers provide information about occluded areas	Improves the quality of 2D-plus-depth	Requires additional bandwidth

2.4.2 Coding, transmission and related artifacts

There are many ways to encode S3D video using any of following coding schemes: simulcast, frame-compatible stereo interleaving, and multiview video coding (MVC) (see Fig. 2.17).

For *simulcast*, the views are transmitted independently and no synchronization is performed between two views. Simulcast is a simple method with low computational complexity and delay since the S3D views may be encoded independently from each other. It is compatible with any kind of standard video coders. However, this is not an optimal solution regarding rate distortion because the redundancy of information between views is not taken into account. A significant gain in bit-rate is possible by applying *asymmetrical coding* [Stelmach and Tam, 1998, Stelmach et al., 2000]. This coding takes advantage of the property of binocular vision to average the brightness and contrast of stimuli. Hence, one view is encoded at a spatial lower quality than another view.

For the frame-compatible stereo formats the coding scheme illustrated in Figure 2.17.b might be suitable. However, for the correct interpretation and de-

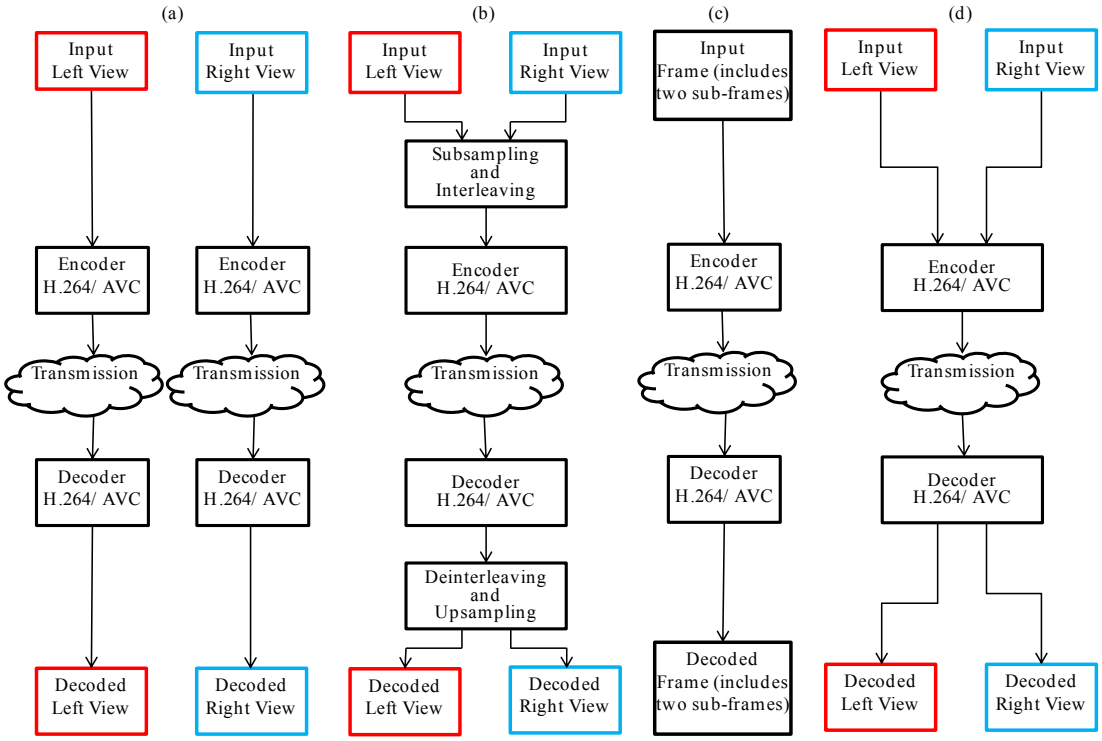


Figure 2.17: Coding schemes. (a) *H.264/AVC* simulcast. (b) Frame-compatible format. (c) Frame packing. (d) H.264/MVC (adapted from [Cagnazzo et al., 2013])

interlacement of content, *supplementary enhancement information* (SEI) is required. It has been standardized within the structure of H.264/MPEG-4 AVC [ITU, 2010, ISO, 2010], but SEI is not the normative part of the decoding process.

Finally, if the bandwidth is critical *multiview video coding (MVC)* illustrated in Figure 2.17.d [ITU, 2010, ISO, 2010] can be used as an extension of the H.264/AVC. This scheme takes advantage of view redundancies and improves compression by a prediction algorithm. The current block of pixels is predicted using different views (or the view) as the reference. Hence, compression is more efficient and the bit stream has a similar structure to the H.264/AVC.

The 2D-plus-depth video format is similar to 2D video data, except that some extra capacity is required for the depth map storage. Therefore, for compression optimization purposes, a standardized format MPEG-C Part 3 is used [ISO, 2007]. Thus, the overall bitstream rate is increased only by 20%-30% of the HD 2D content bitstream rate [Fehn, 2003]. The advantage of MPEG-C Part 3 is compatibility with traditional 2D devices.

The MVV format requires N times more capacity for raw data storage and the bitstream should be increased linearly with the number of encoded views. The MVC is mostly devoted to the encoding of MVV. It merges a traditional intra-view motion-based prediction with an inter-view disparity-based prediction [Merkle et al., 2007]. MVC coding algorithms could be adapted and used for MVD and LDI encoding [Yoon and Ho, 2007] as well. MVD video content and depth can be encoded and transmitted independently or jointly to use view redundancies for more efficient coding performance.

Some artifacts produced during coding are identical to conventional media due to the

similarity of used algorithms. One of such artifacts is *blockiness*, which appears when block-based coding schemes with quantization levels of compression ratios are used. For high compression ratios the quantization gets coarse enough for the HVS to detect it. Especially, blocks can be easily noticed in areas with low spatial frequencies, in areas with high spatial frequencies they are visible as blur. In color channels, blockiness results in *color bleeding*.

The reduction of high frequency components during coding leads to mosaic patterns, staircase effect and ringing [Boev et al., 2008]. *Mosaic patterns* arise when areas with high frequencies lose resolution in horizontal and vertical direction. *Staircase effect* appears at diagonal edges and *ringing* occurs in high contrast areas and generates ripples and shimmering near the edges. Both artifacts are result of coarse quantization and hence often generated together.

In this stage of broadcast chain the cardboard effect (see Section 1.3.3.1) may appear due to encoding of the depth map with high quantization levels. Finally, the asymmetrical stereo coding may generate distortion of depth when the quality mismatch between the left and right views are too big and can not be compensated by the HVS [Boev et al., 2008].

2.5 Conclusions

This chapter discusses the broadcast chain of stereoscopic systems. It is important to understand that final depth perception can be influenced at every stage of this chain. To sum-up the key points of Chapter 2:

- The amount of the perceived depth depends on the six main parameters. Three parameters are from the shooting side: camera focal length, baseline, and convergence distance (See Section 2.2.1); and three parameters are from the visualization side: human interpupillary distance, display width, and viewing distance (see Section 1.3.1 and Section 1.3.2 for examples).

- The optimal quality of the perceived depth is based on an absence of visual artifacts. These imperfections may be produced at every stage of the stereoscopic broadcast chain. Therefore, it is very important to consider from the beginning the impact of the technologies and avoid expensive post production processes.

- The influence of technical parameters on human perception can be studied only when shooting and the visualization environment are controlled. Otherwise, it would not possible to generalize the results.

Chapter 3

Assessment of 3D video QoE

Contents

3.1 Introduction

Chapter 2 was a review of the technical parameters of the broadcast chain that can have a potential impact on the perception of stereoscopic content. This chapter discusses what quality means for S3D images and video and how it can be assessed subjectively and objectively. Traditional 2D concepts of quality evaluation are considered along with how applicable they are to 3D.

3.2 3D Quality of Experience

As explained in Chapter 1, added depth information involves binocular vision. Improperly captured or rendered stereoscopic information (see Chapter 2) can violate this natural physiological mechanism and induce visual discomfort during the visualization

stage. Thus, any 3D video quality evaluation must consider the visual comfort of the viewers.

Furthermore, it has been demonstrated that depth perception can be evaluated independently from image quality. For instance, Tam et al. has shown the low correlation between subjective image quality and perceived depth [Tam et al., 1998]. Later, Seuntiens et al. proved that image quality decreases with increasing compression independent of the depth level [Seuntiëns et al., 2006]. This independency was later confirmed in [Kaptein et al., 2008] by demonstrating that the attribute image quality does not consider the added value of depth. As a consequence, the term *Quality of Experience (QoE)* was proposed to characterize the overall viewing experience of stereoscopic images.

In recommendation ITU-T P.10/G.100 Am.2, QoE is defined as "The overall acceptability of an application or service, as perceived subjectively by the end user" [ITU, 2008]. This definition also includes the acceptability of the service. However, this leads to uncertainty as to when the term *Quality of Service (QoS)* should be used (see [ITU, 2011]). In 2010 Brooks and Hestnes proposed an approach to define and measure QoE relative to QoS. They specified QoE as "a measure of user performance based on objective and subjective psychological measures of using a service or product" [Brooks and Hestnes, 2010]. Then in 2012 the European Network of Excellence "Qualinet" defined QoE as "the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state" [Le Callet et al., 2012].

3.3 Components influencing 3D video QoE

Recommendation ITU-R BT.2021 determines three primary perceptual dimensions, which influence the perceived QoE: picture quality, depth quality, and visual comfort [ITU, 2012a]. These and some other attributes will be discussed in this section.

3.3.1 Picture quality

Picture quality in this case refers only to conventional 2D image quality without the added value of depth. In stereoscopic systems, various artifacts influencing image quality can be introduced during the acquisition (see Section 2.2.2), encoding, and transmission (see Section 2.4.2) stages of the broadcast chain. It is important to understand what exactly picture quality implies in different studies. For example, in some studies picture quality refers to spatial image/video quality [Chen et al., 2012a] or perceived sharpness [Tam et al., 1998].

3.3.2 Depth quality

In recommendation ITU-R BT.2021, depth quality is defined as "the ability of the system to deliver an enhanced sensation of depth" [ITU, 2012a]. As explained in Section 1.2.2, humans can extract depth information using only monocular depth cues. For instance, in [Chen et al., 2012c] depth quantity for 2D stimuli received an MOS score of 25 on a numerical scale from 0 to 100. But depth quality for stereoscopic content considers only the depth gain from binocular vision. Furthermore, some studies replace the concept of depth quality by the amount of depth

[Lambooij et al., 2007], by depth quantity [Chen et al., 2012b, Chen et al., 2012c], or depth rendering [Barkowsky et al., 2009, Chen, 2012, Chen et al., 2012c].

Interestingly, the study of Chen et al. demonstrated that there is almost a linear relationship between the MOS of "depth quantity" and the DoF independent of the content, while the MOS of "depth rendering" mostly depends on the content [Chen et al., 2012c]. Hence, based on the distribution of MOS, it seems that the subjects easily judge the attribute "depth quantity" but have difficulties with "depth rendering".

3.3.3 Visual (dis)comfort and visual fatigue

Stereoscopic systems exploit advantages of binocular vision, sometimes forcing HVS to act inordinately. This happens because the presented stimuli create artificial situations which are unnatural in the real world. So eye strain, discomfort, headaches, and visual fatigue can easily occur. Visual discomfort and fatigue were discriminated by Lambooij et. al in [Lambooij et al., 2007]. *Visual fatigue* was defined as the decrease in the performance of visual functions, which can be measured objectively and subjectively. It is caused by multiple excessive efforts of the visual system. Among the symptoms of visual fatigue are eye tiredness, pain and soreness around the eyes, headaches, blurred and double visions, difficulty focusing, and so on [Ukai and Howarth, 2008]. These symptoms could be divided into four groups: (1) asthenopic related (e.g. eye strain, tired eyes), (2) ocular surface related (e.g. dry eyes, red eyes), (3) vision related (e.g. double vision, reduced visual acuity), (4) the extra-ocular group (e.g. headaches, neck pain) [Balter et al., 2008].

On the other hand, *visual discomfort* is the perceptual reaction of the observer to unnatural visual stimuli or to an unnatural environment, which may combine several sensations and symptoms. If the cause is eliminated and the negative sensation vanishes, visual discomfort disappears immediately, unlike visual fatigue, which accumulates over time and then requires a sufficient recovery period for the visual system [Urvoy et al., 2013].

Furthermore, *comfort* is when there is "satisfaction with the visual environment" (part of the definition of Walter Grondzik [Grondzik,]).

3.3.4 Additional perception dimensions

Besides the primary perceptual dimensions, recommendation ITU-R BT.2021 proposes naturalness and sense of presence as additional dimensions to assess the psychological impact of 3D technologies.

Naturalness is defined as "a truthful representation of reality" or "perceptual realism" [ITU, 2012a]. Such a definition is related to the disproportions of the objects' shapes in space, which are responsible for an unreal or unnatural appearance. The most well-known examples are cardboard and puppet theater effects (see Section 1.3.3).

Contrary to the studies that demonstrated a low correlation between perceived image quality and perceived depth [Seuntiëns et al., 2006], the study of naturalness indicated that this attribute comprises the image quality and added value of stereoscopic depth [Seuntiëns et al., 2008]. In the same study, Seuntiëns et al. investigated *viewing experience* as an alternative evaluation concept. It was defined as "the users' perceptual and cognitive experience of the entire application". The same trend as for naturalness was found: viewing experience decreases with degradation of image quality and integrates depth at the same time.

Another dimension recommended by recommendation ITU-R BT.2021 is a sense of presence or *"being there"*, which is defined as "the subjective experience of being in one place or environment even when one is situated in another". This association was discovered by Freeman and Avons, who used focus groups to study viewer reactions and sensations after watching 3DTV [Freeman and Avons, 2000].

For now there are no standard methodologies for the assessment of additional perceptual dimensions.

3.4 Models of 3D QoE

This section is an overview of the S3D models of QoE that have been proposed in the literature. The main goal of such models is to integrate perceptual attributes related to perceived depth, visual discomfort, and image quality into one concept that allows for an assessment of the quality of the perceived stereoscopic stimuli. Correct modeling of the viewers' experience will help to establish the link between the indicators of QoE and the technology variables of any 3D system.

The first model to characterize 3D viewing experience was proposed by Seuntiëns [Seuntiëns, 2006] (see Fig. 3.1). This model encompasses all primary perceptual dimensions and basically reflected in the idea about multidimensional 3D QoE recommended by ITU-R BT.2021 [ITU, 2012a].

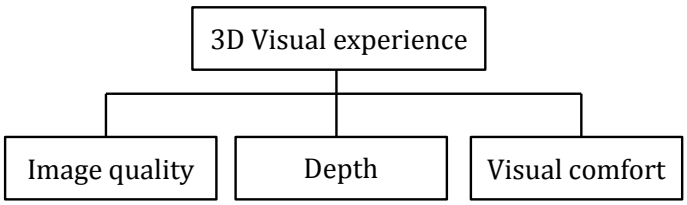


Figure 3.1: The first model of 3D visual experience proposed by Seuntiëns [Seuntiëns, 2006].

Later, he complicated this model by adding naturalness (see Fig. 3.2), which was determined as being capable of taking into account an added depth dimension [Seuntiëns et al., 2005]. Naturalness was considered as a higher level concept in this model. As defined, naturalness and viewing experience (EC) consist of a weighted image quality ($\alpha \cdot IQ$), a perceived depth ($\beta \cdot D$), and a residual term (γ):

$$EC = \alpha \cdot IQ + \beta \cdot D + \gamma \tag{3.1}$$

It was highlighted that the coefficients α and β are only relative contributions and not absolute coefficients. Also, the role of visual comfort was not investigated in his work which is why it is connected with a dashed line to 3D visual experience.

Another model based on the concept of naturalness was proposed by Lambooij et al. [Lambooij et al., 2011] (see Fig. 3.3). Through subjective experiments, the weights of the model components were evaluated. It was reported that naturalness is determined by approximately 76% of perceived image quality and 26% of perceived depth. In addition, naturalness was compared to viewing experience as an alternative evaluation concept. Even though variations in image quality were quite similar, perceived depth was reflected more in naturalness.

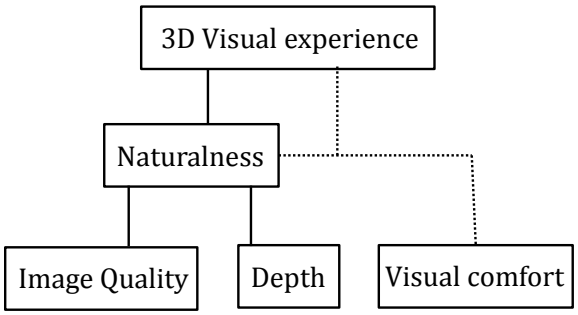


Figure 3.2: Improved
[Seuntiëns, 2006].

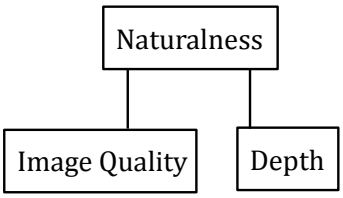


Figure 3.3: 3D quality model proposed by Lambooij et al. [Lambooij et al., 2011]

However, neither Seuntiëns's nor Lambooij's model considers visual comfort. This issue was addressed in a subjective study by Chen et al., who evaluated visual comfort, visual experience, and depth rendering of stereoscopic synthetic images [Chen et al., 2011]. The results indicated that discomfort induced by increasing the DoF decreases the subjective scores of the visual comfort, the visual experience, and even the depth rendering.

Therefore, in their next study, Chen et al. concentrated on searching for the most representative quality indicators to construct a model of QoE [Chen et al., 2012c]. The proposed model of 3D QoE is depicted in Fig. 3.4. This model distinguishes higher (depth rendering, naturalness, and visual experience) and lower (image quality, depth quantity, and visual comfort) levels of 3D QoE attributes. Similar to Lambooij et al., they assumed that high levels of 3D QoE might be represented by a weighted sum of 2D image quality (IQ) and depth quantity (DQ). But they considered visual comfort (VC) as well:

$$QoE = \alpha \cdot IQ + \beta \cdot D + \gamma \cdot VC \tag{3.2}$$

α, β, γ– the weights of 2D image quality, depth quantity, and visual comfort, respectively.

The weights were estimated with a linear regression analysis and their correlation with the high level concepts was computed. The coefficients of linear fitting indicated that visual comfort is a dominant factor for visual experience (58.8%) and naturalness (54.1%). Later, a final model was proposed by Chen in his PhD thesis. He had eliminated higher level concepts of naturalness and depth rendering. The result was that it looks the same as Seuntiëns's first model (illustrated in Figure 3.1); although, the use of the visual comfort axis has been justified by subjective experiments [Chen, 2012].

Finally, Vlad et al. designed a new model (illustrated in Figure 3.5), where the depth axis used in all previous models was replaced by realism [Vlad et al., 2013]. The remarks of the subjects collected during the subjective test revealed the connection between realism and the cardboard effect. The computed correlation coefficients demonstrated

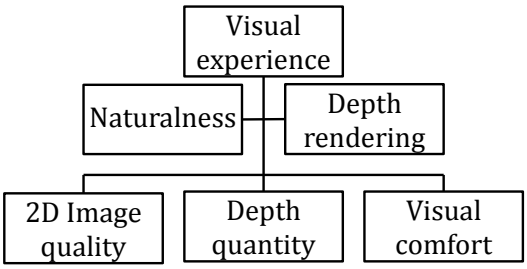


Figure 3.4: 3D QoE model proposed by Chen et al. [Chen et al., 2012c]

an independency between the three proposed attributes of the 3D QoE. However, the weight of ea

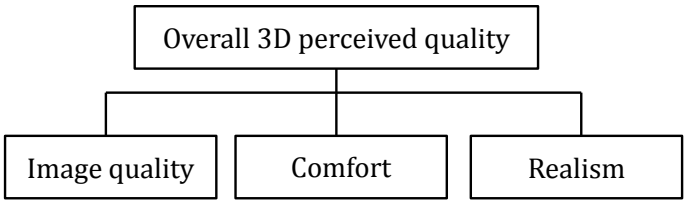


Figure 3.5: 3D QoE model proposed by Vald et al. [Vlad et al., 2013]

Models of the 3D QoE define perceptual attributes, the composition of which should determine the QoE of any stereoscopic stimulus. Chapter 2 discusses various technical parameters and properties of the HVS that influence the 3D video QoE. Therefore, the next question is: how can it be assessed subjectively and objectively?

3.5 Subjective assessment methods of 3D QoE

Subjective studies are the most direct way to evaluate viewers' opinions about a set of images or video sequences. Though it is preferable to follow the established standards of subjective quality assessment to produce compatible and reproducible results across different studies. A big discussion about the new requirements of the subjective quality assessment methodologies for 3D was given by Chen et al. [Chen et al., 2010]. Such requirements include the methods of testing, grading scales, viewing conditions, selection of stimuli, subjects, and even methods for the statistical analysis of the results. For stereoscopic 3DTV systems, all these issues are addressed by the International Telecommunication Union (ITU) in recommendation ITU-R BT.2021 [ITU, 2012a]. However, most of the assessment methods have been adapted from recommendation ITU-R BT.500, which is for the assessment of 2D picture quality [ITU, 2012b]. Therefore, the Video Quality Expert Group (VQEG) started the 3DTV project in 2009 to advance the field of 3D video quality assessment by investigating new subjective and objective assessment methods. The goal of this project is to favor standardization activities for subjective and objective measurements of 3DTV QoE.

Recommendation ITU-R BT.2021 defines several methods for the evaluation of basic perceptual attributes (see Section 3.3). These methods propose suitable evaluation scales, time, order, and the methods of presenting stimuli to each observer. All standardized methods are briefly discussed below. Some examples are given to demonstrate

their usage in the subjective assessment of the perceptual attributes of 3D QoE. The more detailed information about all possible grading scales and test trial structure of the methods as well as the way to analyze and present the results can be found in recommendation ITU-R BT.500-13 Annex2 [ITU, 2012b].

The single stimulus (SS) method. Subjects rate each stimulus in the set without a reference on the selected scale. For picture quality and depth quality assessment the discrete five-grade scale (Fig. 3.6.a) and the standard ITU continuous quality scales (CQS) can be used (Fig. 3.6.b). The quality labels are "Excellent", "Good", "Fair", "Poor" and "Bad". The same scales can be used for the assessment of visual comfort but w
"Uncc

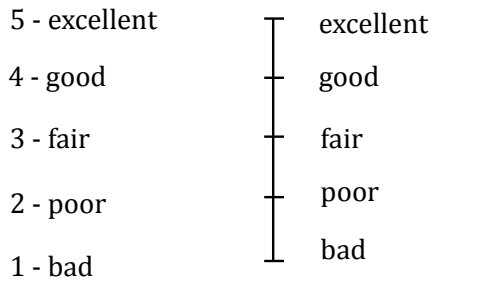


Figure 3.6: The labeled ITU scales a) Discrete five-grade scale and b) Continuous quality scale (CQS) [ITU, 2012a]

At the end of an experiment, the collected individual scores are converted to *mean opinion scores (MOS)*.

For example, the SS method with a discrete scale was used by Lambooij et al. to evaluate the image quality, depth, naturalness, and viewing experience of stereoscopic images with various camera baselines, blur, and noise levels ([Lambooij et al., 2011]). The impact of the acquisition distortions of stereoscopic images on image quality was rated on a continuous quality scale (SSCQS) in the study by Goldmann et al. [Goldmann et al., 2010a]. SSCQS was applied as well to the study subjects' ratings of spatial video quality, depth quality, visual comfort, and overall 3D video quality in [Chen et al., 2012a].

Besides, the discrete five-grade scale and CQS proposed in recommendation ITU-R BT.2021, some studies have adopted the *impairment scale (IS)* from recommendation ITU-R BT.500. This scale is illustrated in Figure 3.7. For instance, Seuntiëns et al. assessed the effect of symmetric and asymmetric JPEG coding and camera separation on perceived sharpness and perceived eye-strain with the IS [Seuntiëns et al., 2006]. SSIS was applied for the subjective evaluation of crosstalk perception in [Wang et al., 2014] and [Xing et al., 2010a].

The single stimulus continuous quality evaluation (SSCQE) method. This method is designed for the assessment of long video sequences. Subjects constantly rate the selected attribute by adjusting a slider in accordance with their perception. Usually the slider is situated on a continuous quality scale with a range from 0 to 100.

Yano et al. assessed visual comfort with the SSCQE using stimuli with a duration of 15 minutes [Yano et al., 2002]. Ijsselsteijn et al. applied the SSCQE to assess the

presence, perceived depth, and naturalness of depth [Ijsselsteijn et al., 1998].

The double stimulus continuous quality scale (DSCQS) method. Subjects rate a series of stimuli pairs. Each pair consists of a reference and a test stimulus with a time duration of 10 seconds. These two images/videos are presented one after the other twice. During the second presentation, the subjects should rate the attribute being assessed using a
structu

5 - imperceptible

4 - perceptible, but not annoying

3 - slightly annoying

2 - annoying

1 - very annoying

Figure 3.7: The five-grade impairment scale (IS) [ITU, 2012b].

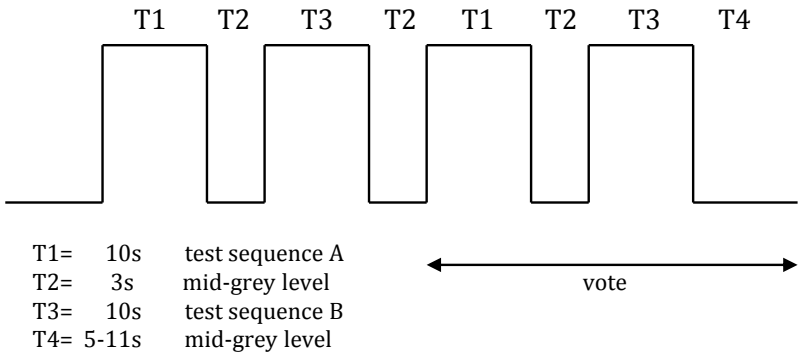


Figure 3.8: The presentation structure of the DSCQS method [ITU, 2012a].

If the reference was included in the test, the *difference mean opinion scores (DMOS)* could be computed as a mean of the difference between the scores of the distorted images and their reference.

DSCQS was applied in the study by Stelmach et al. to rate the overall quality, sharpness, and overall sensation of depth [Stelmach et al., 2000]. Kooi and Toet designed their own labels("equal viewing comfort", "slightly reduced viewing comfort", "reduced viewing comfort", "considerably reduced viewing comfort", and "extremely reduced viewing comfort") and applied this scale to assess the visual discomfort induced by view asymmetries of stereoscopic images [Kooi and Toet, 2004].

The stimulus-comparison (SC) method. All possible pair combinations from the set of stimuli are presented to the subjects. The subjects compare two images/videos in each pair and rate their relationship in terms of preferences using the scale in Figure 3.9.

Barkowsky et al. investigated the influence of depth rendering on the quality of visual experience using a paired comparison [Barkowsky et al., 2009]. Also, this method was used by Li et al. to study visual discomfort induced by motion in stereoscopic displays [Li et al., 2011, Li et al., 2012].

-3 - much worse

-2 - worse

-1 - slightly worse

0 - the same

1 - slightly better

2 - better

3 - much better

Figure 3.9: The labeled ITU scale for the SC method [ITU, 2012a].

The subjective assessment methodology for video quality (SAMVIQ). This is another method that might be suitable for the assessment of 3D stimuli. SAMVIQ was defined in recommendation ITU-R BT.1788 [ITU, 2007]. SAMVIQ was developed on the basis of the DSCQS method for the evaluation of a large range of image quality. It is claimed to provide reliable discrimination at high and low quality levels [Kozamernik et al., 2005, Blin, 2006]. If needed, both hidden and explicit references can be used. All stimuli are presented at the same time on a multi-stimulus button form (see example in Fig. 3.10). Except for the explicit reference, all other stimuli are

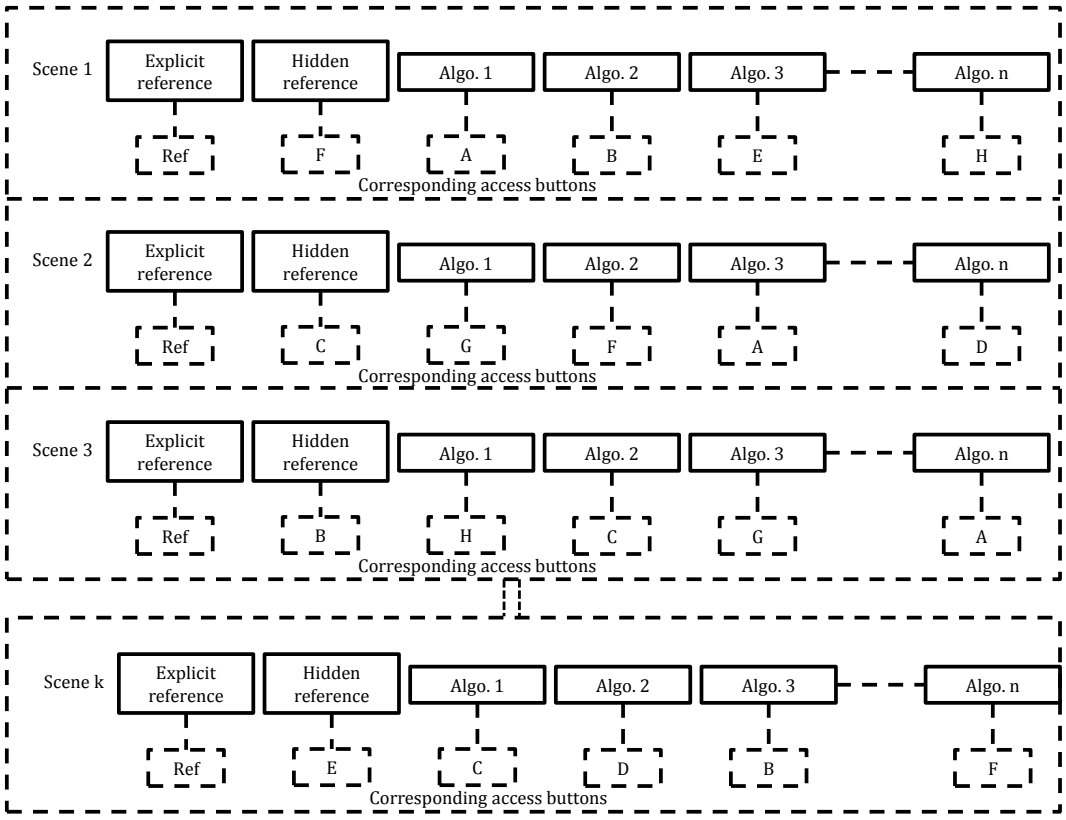


Figure 3.10: SAMVIQ method [ITU, 2007].

SAMVIQ methodology was used by Chen et al. to evaluate change in visual experience, image quality, depth quantity, and visual comfort with a variation of the DoF on the labeled CQS with grades from 0 to 100 [Chen et al., 2012b, Chen et al., 2012c]. In his thesis he also estimated various levels of stereoscopic view asymmetries on CQS and IS [Chen, 2012].

Furthermore, Chen et al. demonstrated through two subjective experiments that visual comfort is the dominant factor of visual experience [Chen et al., 2011, Chen et al., 2012c]. To guarantee the best QoE for viewers, the priority is to make sure that stereoscopic videos do not cause visual discomfort. So, alternative non-standard ways to assess visual discomfort and fatigue will be reviewed in the next section.

3.5.1 Assessment of visual discomfort and fatigue

There are different approaches to assess visual discomfort and fatigue. The two major groups remain the same: subjective and objective. Subjective visual discomfort and fatigue are often evaluated via questionnaires and subjective experiments. However, taking into account that visual discomfort vanishes if its cause is eliminated, objective indicators are inherent to visual fatigue only.

Subjective experiments evaluate the level of visual comfort using standard or adapted assessment scales [Kooi and Toet, 2004, Nojiri et al., 2003, Yano et al., 2002, Wöpking, 1995]. Questionnaires reveal the presence of symptoms relevant to visual discomfort and fatigue [Watanabe and Ujike, 2013, Solimini, 2013, Shibata et al., 2011, Kuze and Ukai, 2008, Ujike et al., 2008, Howarth and Costello, 1997, Kennedy et al., 1993]. A subjective assessment should include the full range of the factors' variation [ITU, 2012b]. Thus, for the evaluation of visual comfort, an extremely uncomfortable case should be included, but might be damaging to the viewers. According to Tam et al., there are two consequences: (1) it complicates the measurement of tolerance limits and long term effects; (2) it raises the question of the ethical requirements for such experiments [Tam et al., 2011].

Objective measurements of visual fatigue assess various indicators such as pupillary diameter, eye blink rate, visual or stereo acuity, accommodative response, tear film breaking time, eye movement velocity, and so on [Lambooij et al., 2009]. Taking into account the variety of possible indicators, many studies are searching for the most representative indicator of visual fatigue. For instance, pupillary diameter was found to be linked with accommodative functions [Uetake et al., 2000] and a reduction in its diameter is correlated with visual fatigue [Murata et al., 2001]. Furthermore, eye-blinking rate was proposed as the indicator of visual fatigue by Stern et al. based on the literature review [Stern et al., 1994]. This indicator has been evaluated in several studies [Kim et al., 2011a, Iatsun et al., 2013]. They demonstrated that the eye-blinking rate increases with visual fatigue. On the other hand, Li et al. argued that eye-blinking rate is not always a correct indicator for stereoscopic motion pictures because of differences in the eye blinking mechanisms for planar and in-depth motions [Li et al., 2013]. Two types of eye-movements were examined regarding visual fatigue: fixation and saccades. However, eye movements were found to be dependent on the content, which interferes with accurately measuring visual fatigue [Iatsun et al., 2013]. Instead Iatsun et. al proposed to use the number of saccades as an evaluation parameter.

Another objective way to assess visual fatigue is using psychophysical devices for taking measurements. For instance, some studies apply electroencephalography (EEG) [Calore et al., 2012, Kim and Lee, 2011, Li et al., 2008, Chen, 2012] and fusional mag-

netic resonance imaging (fMRI) [Kim et al., 2011b] to assess brain reactions to comfortable and uncomfortable stimuli, to identify indicators of visual fatigue, or to analyze the level of emotional involvement during stereoscopic visualization.

The reasons why stereoscopic systems induce visual fatigue were divided by Yano et al. in [Yano et al., 2004] into four groups:

1. Accommodation-vergence conflict;
2. Excessive binocular parallax;
3. Misalignments between the left and right views;
4. Differences between the characteristics of the left and right views.

As it was explained in Chapter 1, accommodation – vergence responses are coupled creating the *oculomotor balance* in real-world viewing. But visual fatigue induced by watching stereoscopic stimuli violates this balance [Inoue and Ohzu, 1997] and influences on the accommodation and vergence functions as well as on its ratios [Schor and Tsuetaki, 1987, Ukai et al., 2000]. Therefore, various studies evaluated different objective indicators of visual fatigue related to accommodation-vergence conflict, such as:

- The amplitude of accommodation [Yano et al., 2002, Yano et al., 2004, Emoto et al., 2004];
- The amplitude of fusion [Emoto et al., 2005, Lambooij et al., 2009];
- AC/C ratio is the amount of accommodative convergence (AC) per unit of accommodative (A) response (accommodation can be still stimulated by covering one eye; the closed eye still converges driven by the coupling of responses) [Ukai et al., 2000];
- CA/C ratio is the amount of convergence accommodation (CA) to convergence (C) response (vergence is stimulated by converging pinhole pupils, accommodation responds as a result) [Fukushima et al., 2009].

Moreover, large screens disparities induce the accommodation-vergence conflict [Ukai and Kato, 2002] and discontinuing changes of parallax contribute to visual fatigue [Emoto et al., 2004] and discomfort [Nojiri et al., 2004].

Unfortunately, a clear relationship between subjective and objective measurements of visual fatigue has not yet been established. It is important to take into account that the HVS is able to avoid visual discomfort by adapting to unnatural conditions or changes of the visual environment and then increase its performance. However, in some cases such adaptations result in visual fatigue [Lambooij et al., 2007].

The four groups defined by Yano et al. as sources of visual fatigue can be equally considered as the sources of visual discomfort. In this case, the difference between visual fatigue and discomfort is the duration of viewing of an improper stereoscopic content. This conclusion can be deduced from the definitions of these concepts (Section 3.6.2). For example, as a result of the accommodation-vergence conflict, visual discomfort appears. It causes multiple excessive efforts of the visual system for a period of time, which results in visual fatigue. Therefore, the studies of visual fatigue might require longer test session duration than studies of visual discomfort.

The next two sections discuss the measurement of visual discomfort. It was decided to merge groups one and two proposed by Yano et al. because excessive disparities

induce the accommodation-vergence conflict [Ukai and Kato, 2002] and combine groups three and four since both of them can be considered view asymmetries.

3.5.1.1 Measurement of the visual discomfort associated with accommodation-vergence conflict and excessive disparities

As discussed in Section 1.4.2, discomfort induced by the vergence-accommodation conflict can be avoided or minimized if the screen parallax remains within the limits of the comfortable viewing zone. To verify whether comfort limits are respected, the following strategy could be used:

1. The comfortable viewing zone should be calculated using equations 3.3 and 3.4 from [Chen et al., 2010] taking into account the parameters of the target screen and the viewing distance. The recommended threshold of DoF=0.2 diopters can be selected [ITU, 2012a]. However, it is important to keep in mind the subjective studies that recommend using a threshold of DoF=0.2 diopters for synthetic scenes and DoF=0.1 diopters for natural scenes [Chen et al., 2012c].

$$Z_f = d - \frac{1}{\frac{1}{d} + DoF} \tag{3.3}$$

$$Z_b = \begin{cases} d - \frac{1}{\frac{1}{d}+DoF} - d, & d < 0 \\ \infty, & d \geq DoF^{-1} \end{cases} \tag{3.4}$$

   where Z_f, Z_b – foreground and background distances respectively in real space, d – viewing distance.

2. Perceptual constraints should be interpreted into the physical parameters: obtained values of Z_f and Z_b should be translated into screen parallax in pixels using equations 3.5 and 3.6.

$$D_f = \frac{Z_f \cdot e}{(d - Z_f) \cdot p_w} \tag{3.5}$$

$$D_b = \frac{Z_b \cdot e}{(Z_b + d) \cdot p_w} \tag{3.6}$$

   where, D_f, D_b – comfortable viewing zone of display in pixels for foreground and background respectively; e – interocular distance; p_w – pixel width.

3. Using the target content, the maximum crossed and uncrossed disparities for the foreground and background should be calculated with any existing algorithm for disparity estimation. For instance, the algorithm can be selected based on a comprehensive quantitative comparison between disparity estimation algorithms provided in [Middlebury, 2014, Scharstein and Szeliski, 2002]. Another possibility for a rough estimation is to measure the maximum crossed and uncrossed disparities directly on the target screen. The obtained values should be converted to a number of pixels or a degree of visual angle.

4. If the obtained maximum crossed and uncrossed disparity values of the target content exceed the computed pixel intervals of the comfortable viewing zone then the tested content will provoke visual discomfort due to an accommodation-vergence

conflict for most of the viewers. A possible solution for this problem is to reduce the maximum disparities or increase the viewing distance.

3.5.1.2 Measurement of the discomfort associated with view asymmetries

The sources of view asymmetries were discussed in Section 2.2.2. But where are the limits of view asymmetries which do not induce visual discomfort? This question has been investigated by multiple studies, which evaluated one of the following thresholds:

1. *Visual discomfort threshold (Tdis).* It characterizes the level of (dis)comfort of a presented stimulus according to the subject's opinion.
2. *Visibility threshold (Tvis).* It reflects the subject's opinion on the visibility of a presented degradation.
3. *Visual annoyance threshold (Tann).* This threshold defines the boundary between annoying and not annoying sensation: 50% of subjects consider a stimulus annoying and 50% as not annoying.
4. *Acceptability threshold (Tacc).* This threshold represents the viewer's expectation level of perceived video quality in a certain context and situation. Tacc(80%) means that 80% of viewers find the video quality of a stimulus acceptable, while 20% do not find it acceptable.

These thresholds are estimated via subjective experiments using different kinds of assessment scales. The visual discomfort threshold is often evaluated using a continuous quality scale (see Fig. 3.6.b). To derive the visibility threshold, ITU BT.500 recommends a categorical impairment scale and defines a grade of 4.5 on this scale as the visibility threshold, which is located between "imperceptible" and "perceptible, but not "annoying" (see Fig. 3.7). Similarly, the visual annoyance threshold can be detected using the same scale with a grade of 3.5 "perceptible but not annoying" and "slightly annoying" as proposed by [Chen, 2012]. The acceptability threshold of visual comfort could also be evaluated using a binary scale with the labels "Acceptable" and "Not acceptable".

Another method for deriving the acceptability threshold was proposed by Chen [Chen, 2012]. He designed a method which links the acceptability threshold with visual comfort that is assessed on a continuous quality scale. According to his method, a score of 49 on the visual comfort scale represents 50% acceptability of the visual comfort and a score of 60 equals 80% acceptability (a score between "good" and "fair" on the visual comfort test). This relationship is illustrated in Figure 3.11. An advantage of using this method is that the acceptability threshold can be computed from the visual comfort scale. Hence, an additional subjective test to define the acceptability threshold is not required.

Multiple studies assess different kinds of view asymmetries using the scales described above or other scales adapted from standards. For example, the study of Kooi and Toet [Kooi and Toet, 2004] investigated the dependence of visual comfort from different types of binocular image asymmetries (geometrical, optical, and crosstalk). It was determined that large view discrepancies reduce visual comfort. Taking into account that view asymmetry negatively influences visual comfort and hence QoE, Chen in [Chen, 2012] created a table with visibility thresholds and visual discomfort thresholds, which was proposed in the literature by [Kooi and Toet, 2004, Fournier, 1995, Seuntiëns et al., 2005, Ikeda and Nakashima, 1980, Ion-Paul and Hanna, 1990]. The table 3.1 is adapted from Chen and complemented by some recent findings of [Chen, 2012, Wang et al., 2014].

"shape-preserving interpolant". The results reveal that around 80 percent of subjects accepts the score 60, i.e., between "good" and "fair" on the visual comfort criteria. Only 50 percent of subjects can accept 50, i.e., "fair". 80 percent are generally used as a rule-of-thumb threshold in many service-oriented applications. Thus, the visual comfort should be maintained as higher than 60. The above finding results in a recommendation for optimized perceived depth: For natural scenes, DOF 0.1 should be targeted and for synthetic scenes, the DOF threshold may remain 0.2.

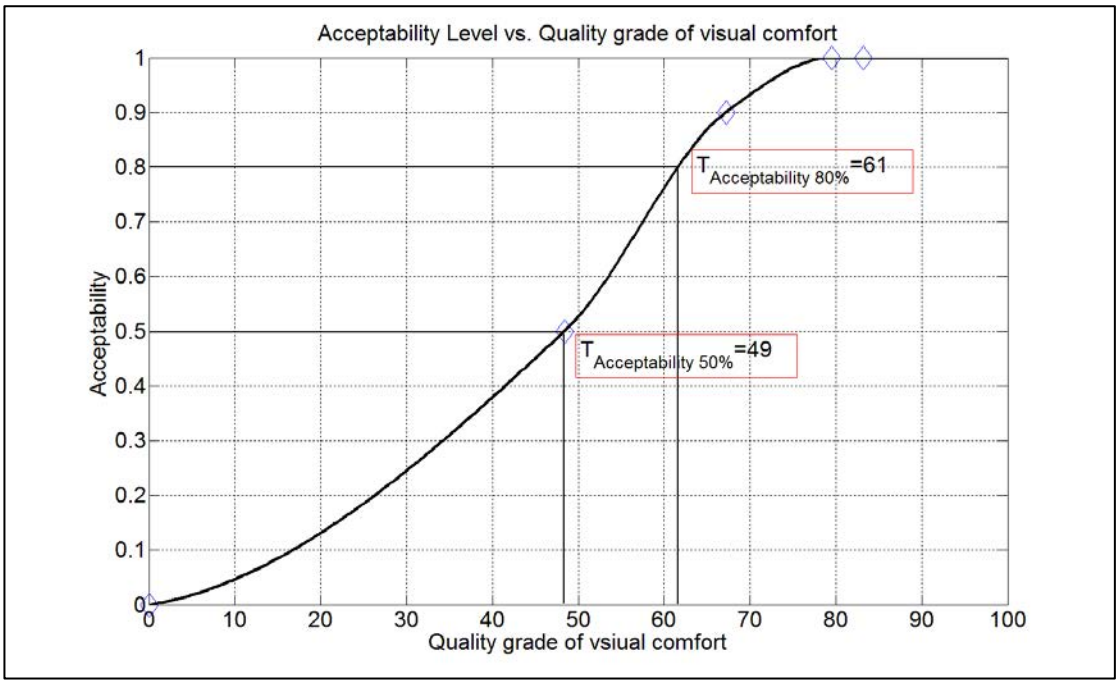


Figure 3.11: Method by Chen to define the acceptability from a visual comfort scale; from [Chen, 2012] p.121

Figure 6-4 : Acceptability vs. Quality grade of visual comfort

6.5 3D QoE modeling

As explained in the previous section, 2D image quality is independent of depth variation while depth quantity and visual comfort shows nearly linear relation with perceived binocular depth. Viewers can judge these three QoE indicators independently so that these three indicators may be categorized as the basic level of

The thresholds were obtained with the following experiment set-ups: (1) Kooi and Toet: 1024×768 resolution, 170×128 cm screen size, 185 cm viewing distance; (2) Chen: HD resolution, 46" line interleaved display, 2.6 m viewing distance; (3) Fournier: SD resolution, 4.5 times image height viewing distance.

As can be seen from the values presented in table 3.1, visual annoying thresholds represent higher levels of degradation than visibility thresholds. This means that, according to the subject's opinion, artifacts or distortions may be visible but not annoying. The thresholds obtained by Chen are more critical than those obtained by Kooi and Toet. Unlike the study of Kooi and Toet, his stimuli were created using a virtual camera which is free of any optic, color, or calibration problems. He explained that his stimuli covered the evaluation scale more uniformly. The stimulus display duration was 8 seconds instead of 3 seconds like in the study by Kooi and Toet, which allowed subjects to judge more critically. Regardless, the authors of the studies reached the same conclusion that a large amount of view asymmetries reduce visual comfort.

3.6 Objective assessment methods of 3D QoE

Subjective testing is the most reliable way of evaluating the perceived quality of stereoscopic still images and videos. However, it is quite costly, time consuming, and should be repeated every time that system parameters are changed. The alternative is new algorithms of objective video quality assessment. The goal of such metrics is to predict image and video quality automatically.

Objective quality metrics can be divided into three groups based on the availability of the original video or image signal. The first group is the *full-reference (FR)* metrics or video similarity and fidelity measurement. Such metrics compare a reference signal (undistorted with perfect quality) with a test signal. The main disadvantage of this

Table 3.1: Summary from literature of view asymmetry thresholds.

Asymmetry	Threshold	Value	Source
Vertical shift	Tdis	34' (<1PD)	[Kooi and Toet, 2004](1)
	Tvis	2.8'	[Chen, 2012](2)
	Tann	7'	[Chen, 2012]
Rotation	Tdis	>1°	[Kooi and Toet, 2004]
	Tvis	0.5°	[Fournier, 1995](3)
	Tvis	0.28°	[Chen, 2012]
	Tann	0.63°	[Chen, 2012]
Keystone distortion	Tdis	0.57° (>1PD)	[Kooi and Toet, 2004]
	Tvis	3'	[Ion-Paul and Hanna, 1990]
Focal length difference	Tdis	2.5%	[Kooi and Toet, 2004]
	Tvis	1%	[Fournier, 1995]
	Tvis	0.55%	[Chen, 2012]
	Tann	1.7%	[Chen, 2012]
Definition difference	Tvis	30%	[Fournier, 1995]
Black level difference	Tvis	1% (0.1 dB)	[Fournier, 1995]
	Tvis	3%	[Chen, 2012]
	Tann	15%	[Chen, 2012]
White level difference	Tvis	15% (1.5 dB)	[Fournier, 1995]
	Tvis	11%	[Chen, 2012]
	Tann	27%	[Chen, 2012]
Crosstalk	Tann	10%	[Wang et al., 2014]
	Tdis	5%	[Kooi and Toet, 2004]
	Tvis	3%	[Wang et al., 2014]
	Tvis	0.2-7%	[Fournier, 1995]
	Tvis	2%	[Seuntiëns et al., 2005]
Color difference	Tvis	15-100 nm	[Ikeda and Nakashima, 1980]
R, G, B level	Tvis	10%	[Chen, 2012]
R, G, B level	Tann	20%	[Chen, 2012]
Temporal sync. diff.	Q. drop	2 frame	[Goldmann et al., 2010b]

group is that in practical applications the reference is often not accessible. The ideal solution in practice would be *no-reference (NR) metrics*, which can be used for any application. A compromise between the two groups can be found by using the *reduced-reference (RR)* metrics, which integrate certain features extracted from the reference signal for comparison.

Perceived 3D video QoE does not only depend on the quality of the signal itself. It can be influenced by "any characteristic of a user, system, service, application, or context" [Le Callet et al., 2012]. Due to all of the influencing factors, it is quite difficult to reach the best case scenario even in the case of 2D: when the objective prediction fully matches the subjective scores. The widely used PSNR and MSE *data metrics* do not consider the video quality as perceived by human observers because the signals are compared without considering the content and the properties of the human visual system. Thus, the efforts of researchers have been directed towards *picture metrics*, which imitate various features of human vision related to image quality perception, such as contrast sensitivity, color perception, and so on.

However, complete success in creating an objective metric considering the properties of the HVS has not been achieved. In 1999, the first attempt (FR-TV Phase I) of the VGEG group to standardize metrics of objective quality assessment failed. It has been reported that none of the seven or eight (out of the nine) tested objective models could outperform the others statistically [VQEG, 2000]. Also, the performance of these models was statistically equivalent to PSNR, which was used as a reference objective model. Hence, VQEG did not validate any models for inclusion in ITU recommendations. Nonetheless, as a result of Phase I, the database of video sequences with associated MOS scores became publicly available. MOS scores were obtained through subjective tests using the DSCQS method (see Section 3.5).

The second phase (FR-TV Phase II) of the VQEG test for FR metrics for SD TV applications defined the four best metrics outperforming PSNR (some correlations reached 94% with MOS, while PSNR about 70%). As in the previous phase, the tests were concentrated on MPEG-2 compression for TV broadcasting. MOS scores were obtained again with the DSCQS method by evaluating the new test sequences. Based on the results reported by VQEG [VQEG, 2003], Rec. ITU-T J.144, and BT.1683 recommend the objective algorithm proposed by the NTIA [Pinson and Wolf, 2004], the Yonsei University [Lee et al., 2006], the Telecommunications Research and Development Center (CPqD) [Lotufo De Alengar et al., 1998], and British Telecom (BFTR).

All later tests by VQEG resulted in ITU recommendations based on the evaluation of metrics for multimedia with smaller frame sizes [VQEG, 2008], NR and RR metrics for standard definition television (625-line and 525-line) [VQEG, 2009], and video quality models that predicted the quality of High Definition Television (HDTV) [VQEG, 2010]. Interestingly, the later trend in standardized metrics is based on a modeling of the human visual system [OPTICOM, 2008] or psycho-visual and cognitive modeling [SwissQual, 2010] (ITU-T Rec. J.247 and ITU-T Rec. J.341 correspondingly).

However, there are still no standardized NR modes, which are necessary for broadcast services in the absence of a reference. Thus, an industry-driven alternative for automatic quality measurement appeared. The idea is to detect simple perceived indicators and then to display an alert when an indicator crosses a threshold associated with the emergency of perceptible degradation of the video or audio. The model avoids a prediction of MOS and uses simple modeling, which is believed to be "potentially more accurate and industrially useful" [Leszczuk et al., 2014]. This project for developing a set of key indicators is in charge of the "Monitoring of Audio-Visual quality by key Indicators"

(MOAVI) subgroup [Wyckens et al., 2012] of the VQEG.

Discussion. From the state of the art studies described above, it seems that a comprehensive objective quality metric for 3D video QoE should take into account:

- The quality of the three primary perceptual dimensions which influence the perceived QoE: picture quality, depth quality, and visual comfort [ITU, 2012a].
- 3D display technology and 3D representation format.
- Visualization environment (display size, viewing distance).

If one of the components listed above is missing, it is impossible to conclude on the overall 3D quality of experience. For now, a metric that encompasses all of these requirements does not exist. Based on the presence of the recommended perceptual attributes for the assessment of QoE, it was decided to divide the overview of the existing metrics into three groups: metrics that assess 2D image quality, metrics that combine the image quality and depth component, and metrics that include visual comfort.

3.6.1 2D image quality

The concept of 3D QoE defines the image as the conventional 2D image quality without the added value of depth (see Section 3.3.1). Hence, all traditional 2D metrics can be used for the evaluation of this attribute of the model. For example, the most widely used classical metrics are mean squared error (MSE), peak signal-to-noise ratio (PSNR), structure similarity index (SSIM) [Wang et al., 2004], and video quality metric (VQM) [ITU, 2004]. You et al. estimated the performance of PSNR, SSIM, and nine other FR metrics using stereoscopic images with different levels of quality degradation [You et al., 2010]. The right view was degraded with four distortion types: Gaussian blurring, JPEG compression, JPEG2000 compression, and white noise. The left view image remained undistorted. All metrics were computed for left and right views and the final quality score was obtained by averaging both values. The highest correlation with subjective scores was found for the SSIM metric. Similarly, may other studies have investigated the adoption of 2D metrics for 3D quality evaluation [Yasakethu et al., 2008, Hewage et al., 2008].

Another open issue that appeared in the context of 3D is how to combine the scores from the left and right views. A possible solution is to consider the properties of binocular vision. For example, Campisi et al. investigated the best way to combine the scores from the right and left views [Campisi et al., 2007]. Three different approaches were compared: averaging the scores, main eye approach (only the score from dominant eye view was considered), and visual acuity approach (the scores from left and right views were weighted in accordance with subject's acuity). The same amount of blurring and JPEG compression were applied to the left and right views. The results did not reveal any improvement from using the main eye or acuity approaches in comparison with averaging.

Only a few 2D metrics were designed intentionally for stereoscopic quality evaluation. For instance, the Stereo Band Limited Contrast algorithm (SBLC) takes into account the monocular properties of the HVS [Gorley and Holliman, 2008]. An input algorithm uses the left and right views of the stereo pair and matches regions of high spatial frequency taking into account sensitivity to contrast and luminance changes. SBLC was found to have a better correlation with subjective scores than PSNR. However, this metric was only devoted to the prediction of a threshold compression level for stereoscopic image

pairs. Similar to SBLC, the proposed Perceptual Quality Metric (PQM) accounts for the luminance and contrast distortions of each pixel [Joveluro et al., 2010]. Its performance was determined to be better than VQM. Another metric designed by Ryu et al. computes luminance, contrast, and structure similarity of each view [Ryu et al., 2012]. The scores are combined based on the binocular suppression principle that implies domination of the high quality view over the degraded one. According to the authors, the metric provides "consistent and outstanding" results in comparison with existing metrics.

All metrics that only consider the monocular properties of the HVS are only suitable for the evaluation of one attribute of 3D QoE: 2D image quality. They can not be accepted directly for the objective evaluation of overall stereoscopic quality assessment because the depth component is not taken into account [You et al., 2010]. This conclusion was validated by Huynh-Thu et al., who specified that 3D objective quality analysis should be applied not to the transmitted signal as in 2D, but rather to its rendered version [Huynh-Thu et al., 2010].

3.6.2 Including depth attribute

To improve traditional 2D metrics and approach human perception, some metrics employ disparity information, which can be estimated from the left and right views or directly available in 2D+depth format.

For instance, Benoit et. al proposed to improve 2D metrics using depth information [Benoit et al., 2008]. Their FR algorithm computes the average signal difference between the left and right views of an original stereopair and its distorted version. This can be done using one of the traditional perceptual quality metrics such as SSIM or C4. The resulting image distortion is combined with disparity distortion (the difference between the original disparity map and distorted one). A significant increase in performance was observed when the SSIM metric was linearly combined with the disparity map distortion. The proposed metric was estimated using degraded stereopairs with JPEG, JPEG2000, or blurring. In the same manner, some other studies have demonstrated that a combination of conventional metrics with disparity information allows a better performance in stereoscopic image quality assessment to be achieved [Yang et al., 2009, You et al., 2010]. Finally, natural scene statistics was considered in NR objective metrics based on a disparity map and a linear rivalry model [Chen et al., 2013]. The model's performance is verified on symmetric and asymmetric distorted stereoscopic images and was found to have higher correlations with DMOS in comparison with Benoit, You, and others' 3D FR metrics.

A new approach in considering the binocular properties of the HVS was applied by Bensalma and Larabi to create a quality metric based on the behavior of simple cells that retrieve disparity information in the visual cortex [Bensalma and Larabi, 2010, Bensalma and Larabi, 2013]. The amplitude, orientation, phase, and size of simple cells were simulated using spatial-frequency transforms. Binocular fusion is implemented by complex cells from the output of simple cells. Finally, a match of the two simple cells from the left and right views is validated if this combination reaches maximal binocular energy. The resulting Binocular Energy Quality Metric (BEQM) is computed as the difference between the binocular energy of the original pair and the degraded pair. Various authors have claimed high performance of the metric in comparison to existing ones. However, only the correlation between the binocular energy and image quality degradation from JPEG symmetrical and asymmetrical compression was evaluated.

The visualization of stereoscopic content was considered in the metric comprising

screen size and disparity by Xing et al. [Xing et al., 2010b]. A prior subjective test defined the viewing location as an insignificant factor of 3D video QoE, while various combinations of the content, baseline, and screen size were found to be significant. The insignificance of the viewing distance can be explained by the content selection. The sequences, which had been chosen for the test, remained within the comfortable viewing zone (DoF=$\pm$0.2), which was calculated taking into account maximum screen disparities and viewing distance. The objective metric encompasses the estimated image disparity plus the weighted screen size. However, it can not detect visual discomfort perceived by the viewers because perceptual thresholds of view asymmetries and ZoC are not taken into account. From perceptual issues, only crosstalk was considered in another work by the same authors [Xing et al., 2010c].

3.6.3 Including comfort attribute

Sohn et. al. proposed two new object-dependent disparity features: relative disparity (the mean disparity difference between neighboring objects) and object thickness (the ratio of mean width relative to the mean absolute disparity of an object) for the evaluation of visual discomfort [Sohn et al., 2013]. Their results demonstrated that the difference in disparities between neighboring objects and the stimulus width should be taken into account in visual discomfort prediction algorithms. Using the new features was able to improve the prediction performance of metrics that use traditional disparity features (those taking into account the disparity magnitude and special frequency of a stereoscopic scene). All these features were collected from state of the art studies. View asymmetry thresholds were not included in the proposed metrics. Nevertheless, the authors made sure that vertical asymmetries did not influence the results of the subjective studies.

Winkler in [Winkler, 2014] presented various metrics which are able to detect some common sources of visual discomfort in stereoscopic content. These metrics detect disparity and view mismatches (based on the work of Takaya [Takaya, 2010]). The disparity range is computed to define the ZoC, to check that maximum disparity does not cause eye divergence, and to verify the disparity transaction between frames to avoid depth discontinuities. Unfortunately, geometrical view asymmetries were not considered. The proposed metrics are computationally efficient and make a step towards the evaluation of 3D QoE by taking into account image quality, depth, and comfort. However, visualization parameters and perceptual thresholds should be integrated to consider the final viewer's perception of rendered stereoscopic content. Another potential issue of this research is how to combine the scores of these and other quality metrics into one score of 3D video QoE.

3.7 Conclusions

This chapter discusses 3D video QoE and reviews subjective and objective methods of its assessment. Therefore, we can draw several important conclusions:

- **3D QoE** has a composite structure. Three primary perceptual dimensions, which influence the final perceived QoE are image quality, depth quality, and visual comfort. Each attribute can be evaluated independently from the others. Unlike the perceptual attributes of higher level such as naturalness and sense of presence,

each primary attribute could be linked directly with the technical parameters of a 3D system.

- **Image quality** can be evaluated separately from the comfort and depth components. In subjective studies it is often evaluated by creating such degradations as JPEG compression, noise, and blur. For its objective evaluation, conventional 2D quality metrics can be applied.

- Opposite to depth quantity, the concept of **depth quality** seems to be quite difficult to judge for viewers. In subjective studies both depth quality and depth quantity are assessed by changing the range of disparities or DoF. Neither of the two concepts take into account the depth component in terms of stereoscopic distortions (e.g magnification/miniaturization of object dimensions and stretching/compression of depth). Objectively, some mathematical methods permit the evaluation of the resulting distortions of rendered content. However, the perceptual limits of geometrical distortions are not known: what level of shape distortion is perceptible, what level of distortion is annoying, what impact various levels of shape distortions have on the overall 3D video QoE, and so on.

- **Visual comfort** is the dominant factor of 3D video QoE. Hence it is important to understand the potential sources of visual discomfort and assess its impact on human perception. Subjectively, discomfort is evaluated by creating stereoscopic stimuli with view asymmetries or outside the zone of comfort. There are a few metrics that exist to assess visual comfort objectively. Unfortunately, none of them takes into account all the possible reasons of visual discomfort.

- The state-of-the-art **objective metrics** in Section 3.6.1 evaluate the quality of the signal without considering the perception of depth involved and resemble 2D metrics concerning spatial distortions. The metrics in Section 3.6.2 consider the depth dimension without taking into account if it remains within the zone of comfort. None of the metrics in Section 3.6.3 examine the potential impact of view asymmetries. So, visual discomfort might not always be predicted correctly. Hence, a comprehensive objective metric of 3D video QoE does not exist at the moment.

- A **comprehensive objective metric** of 3D video QoE should consider all quality aspects of a rendered signal which depends on camera parameters, the visualization environment, display technology, 3D representation format, and the viewer's perception. It seems to be impossible without considering the human perceptual thresholds.

Part I

Visual attention in 3D

Chapter 4

State-of-the-art of visual attention in S3D

Contents

4.1 Introduction

This chapter presents some of the state-of-the-art studies concerning visual attention. It also provides a review of recent studies comparing visual attention for S3D and 2D conditions.

4.2 Visual attention and eye movements

Vision is a continuous process directing our attention toward interesting locations within the environment, while ignoring others. The mechanism that performs the selection is visual attention. The classical definition of attention was given by the pioneering American psychologist and philosopher William James in 1890: *"Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out*

of what seem several simultaneously possible objects or trains of thought. [...] It implies withdrawal from some things in order to deal effectively with others."[James, 1890].

According to Posner [Posner, 1980] visual attention is used:

- to select important areas in our visual field (alerting);
- to search for a target in cluttered scenes (searching).

Natural visual scenes are cluttered and contain many different objects that cannot all be processed simultaneously due to the limited capacity of the HVS [Chun et al., 2011]. So attention is required as a mechanism to avoid overloading the system. *Visual scanning* is a way to select certain objects by looking from one place to another. Scanning is necessary to locate an object of interest in the fovea (the central part of the retina), which is responsible for sharp central vision.

While scanning, several types of the eye movements are possible:

- *Fixation* is the phase when the eyes are almost stationary. The typical duration is around 200-300 ms [Findlay and Gilchrist, 2003]. Generally the duration depends on a number of factors like the depth of processing [Velichkovsky, 2002] and ease or difficulty of perceiving something [Mannan et al., 1995].
- *Saccade* is quick jerky eye movement from one fixation location to another. The length of the saccade is from 4 to 12 degrees of visual angle.
- *Smooth pursuit* is the voluntary tracking of moving stimulus.
- *Vergence* is the coordinated movement of both eyes. Convergence happens when objects move towards the eyes and divergence happens when objects move away from the eyes.

A sequence of eye movements results in a *scanpath*, which represents the pattern of fixations (circles) separated by saccadic eye movements (lines) that occur when a subject viewed the image of a kitchen in Figure 4.1.

The process of scanning involves *overt visual attention*, e.g. attention associated with eye movements. Another type of visual attention is *covert visual attention*, which does not require eye movements. This attention can be voluntarily focused on a peripheral part of the visual field and this is the act of mentally focusing on one of several possible sensory stimuli. Cover attention plays an important role in different sports. For example, a basketball player can look in one direction, while covertly attending his teammate from another [Goldstein, 2013]. Most of the studies deal with overt visual attention, which can be measured using eye-tracking.

4.3 Bottom-up and top-down processes

Yarbus [Yarbus, 1967] demonstrated that eye movements are able to change depending on the question asked to the subject. A reproduction of the painting in the top left corner of Figure 4.2 was presented to an observer and his eye movements were recorded for 3 minutes. The results recorded in Figure 4.2.a- 4.2.g were obtained when the observer was asked to: (a) freely observe the painting, (b) judge the economic status of the family, (c) define the age of each person, (d) figure out what they had been doing before the unexpected visitor arrived, (e) memorize what clothes they were wearing, (f) memorize the positions of the people and the objects in the scene, and (g) estimate how long the unexpected visitor had been away from the family.

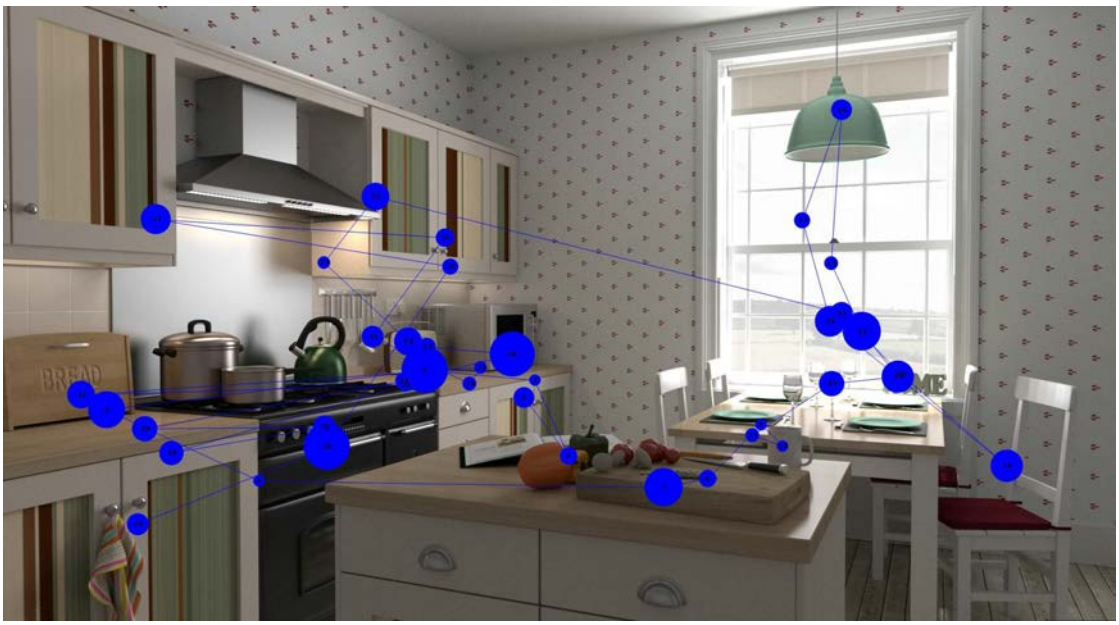


Figure 4.1: Scanpath of an observer looking at the image of a kitchen. Fixations are shown as blue circles and saccadic movements as blue lines.

Yarbus demonstrated that our attention depends on bottom-up features and on top-down information. It has lead to the definition of two types of visual attention:

- *Bottom-up attention* (also exogenous or stimulus-driven attention) is driven by the physical properties of objects (*stimulus salience*) like color, orientation, and intensity. Thus, they draw attention reflexively, in a task-independent way. Bottom-up attention is involuntary, very quick, and unconscious [Borji and Itti, 2012].
- *Top-down attention* (also endogenous or goal-driven attention) is driven by "high level" information, such as current task, knowledge, and expectations. This attention process is voluntary, very slow, and conscious [Desimone and Duncan, 1995].

4.4 Eye-tracking

Eye-tracking is the process of recording eye movements. This technique is used in a variety of disciplines such as psychology, medicine, human factors, marketing, neuroscience, and computer science [Duchowski, 2002]. An *eye-tracker* is a device that is able to provide a quantitative measure of eye position, gaze direction and gaze point, blink, eye movement and scanpaths, pupil size, and pupil dilation.

Eye-tracking experiments are a simple way to conduct human behavioral and psychophysics studies for the desired visual tasks. However, there is no standard for assessment with the eye-tracking technique, so each researcher should carefully prepare the experiment in order to produce results that are reproducible and compatible with existing studies. It is important to consider the quality of visual content, the visual content itself, the presence or absence of a specific task, the age and number of observers, and the viewing duration [Le Meur, 2014].

The size of the stimulus on the retina, e.g. the angular resolution is necessary to segment raw eye-tracking data into fixations and saccades [Le Meur, 2014]. It can be

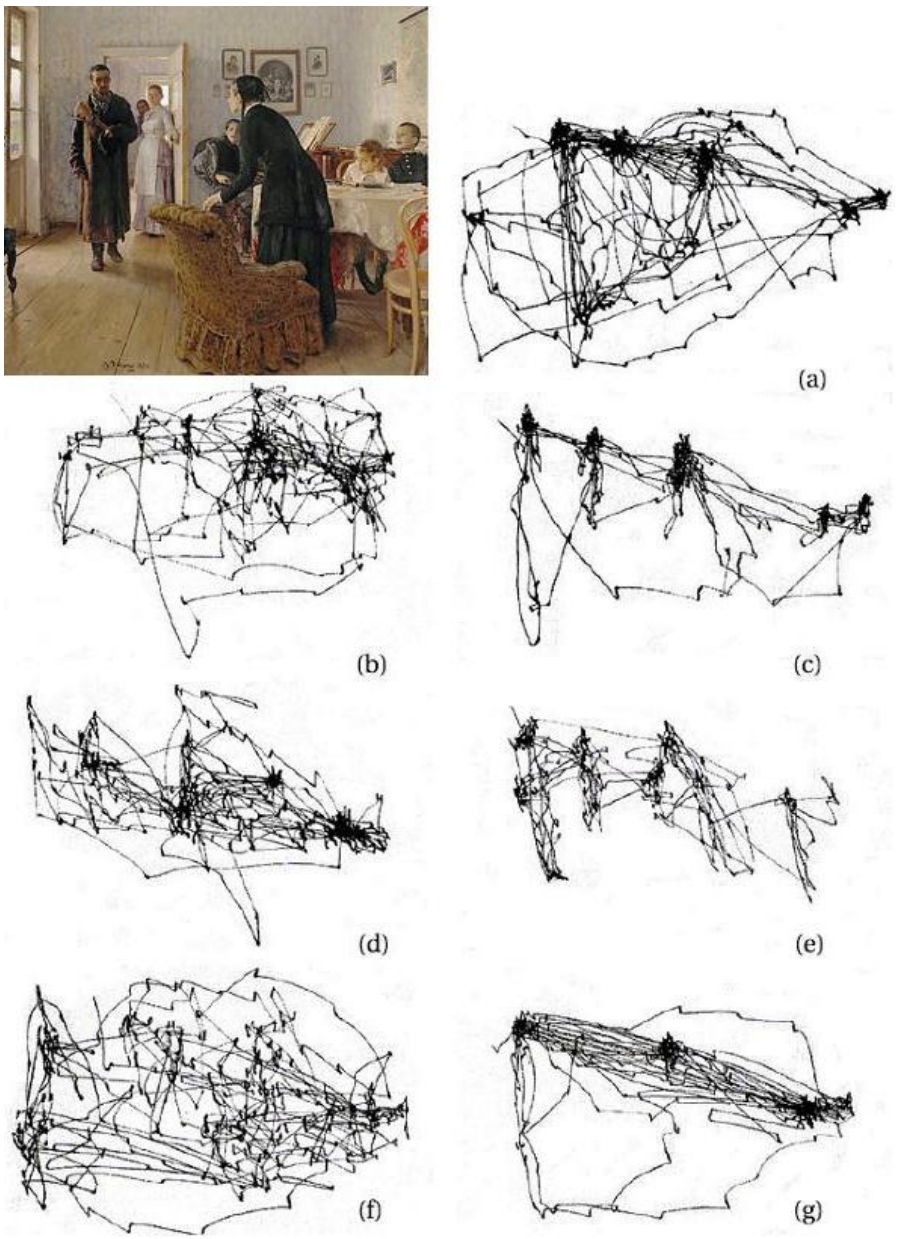


Figure 4.2: The stimulus painting and seven records of eye-movements depending on the task.

computed as the size of the stimulus on the screen and the viewing distance as illustrated in Figure 4.3. The visual angle of the height of the rectangular stimulus can be computed from the equation 4.1 and the width from the equation 4.2:

$$\Theta_H = 2 \cdot arctan(\frac{H}{2d}) \tag{4.1}$$

$$\Theta_W = 2 \cdot arctan(\frac{W}{2d}) \tag{4.2}$$

where H, V - the width and the height of the screen image, d - viewing distance.

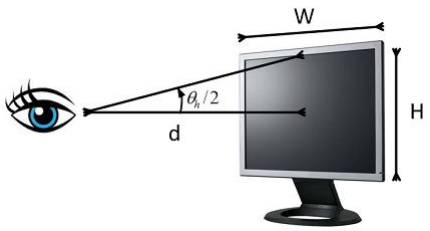


Figure 4.3: Visual angle in pixels per degree Θ_H, d-viewing distance, H, V - the width and the height of the screen image (from [Le Meur, 2014]).

Then it is possible to compute the number of pixels per one degree of visual angle by dividing the horizontal resolution of a screen by Θ_W, obtained with the equation 4.2.

Once the experiment is designed and the raw data has been collected, it is possible to compute scanpaths or *fixation density maps (FDM)*. FDM can be defined by the following equation 4.3:

$$f^i(x) = \sum_{k=1}^{M} \delta(x - x_{f^{(k)}}) \tag{4.3}$$

where x - vector representing spatial coordinates (x, y), $x_{f^{(k)}}$ - the spatial coordinates of k^{th} visual fixation. M is the number of visual fixations for the i^{th} observer. $\delta(.)$ is the Kronecker symbol ($\delta(t) = 1$, if $t = 1$, otherwise $\delta(t) = 0$).

For the N observers, the final fixation map f is described by the equation 4.4:

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} f^i(x) \tag{4.4}$$

Finally, a saliency map S is computed by convolving the fixation map with an isotropic bi-dimensional Gaussian function as follows:

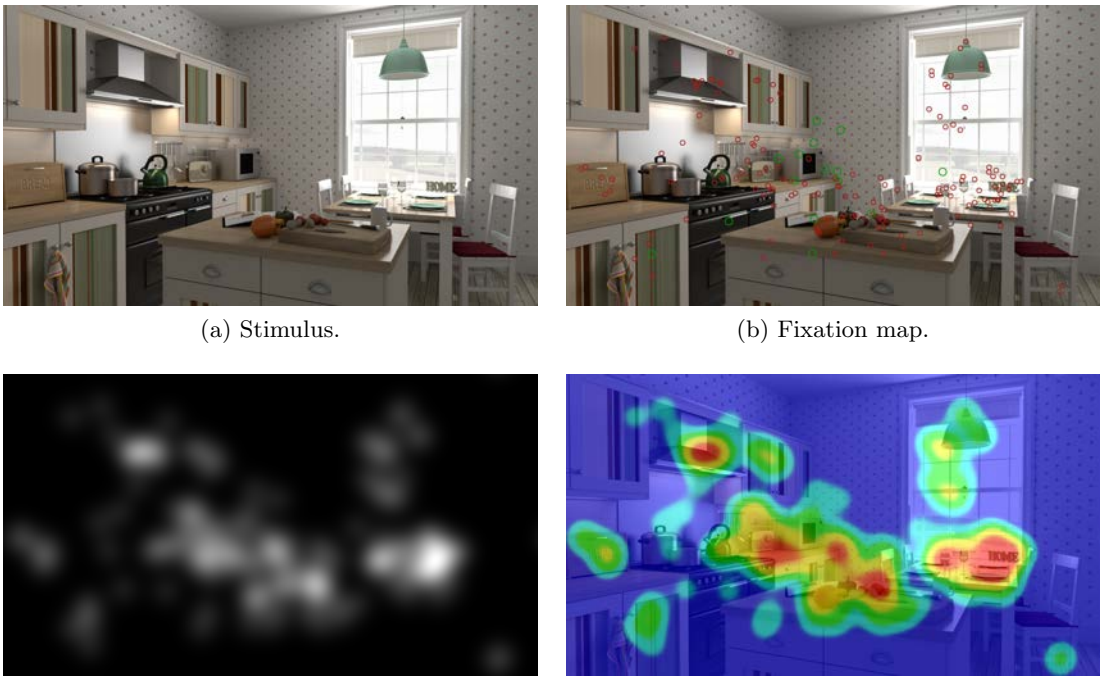
$$S(x) = f(x) * G_\sigma(x) \tag{4.5}$$

where σ is the standard deviation of the Gaussian, which is commonly considered as one degree of visual angle for σ.

Figure 4.4 illustrates an example of fixation and saliency maps. A heat map, illustrated in Figure 4.4.d, consists of the stimulus (Fig. 4.4.a) as a background image and a hotspot mask superimposed on top of it. A hotspot mask is a color map scaled between blue (no fixations) and red (highest number of fixations).

4.5 Studies of visual attention in S3D

Chapter 2 discussed how the production of 3D content can be more complicated than 2D since improper shooting can cause visual discomfort. Thus, the production of visually comfortable stereoscopic content is fundamental to ensure the deployment of 3D cinema, as well as 3DTV at home. To deal with these problems, some studies on the influence of 3D on perception have been performed and reviewed in Chapter 3. This section reviews the studies that explore how stereopsis influences mechanisms of visual attention and whether it causes a change in gaze behavior while watching stereoscopic content.



(a) Stimulus. (b) Fixation map.

(c) Saliemcy map. (d) Heat map.

Figure 4.4: Example of fixation, saliency, and heat maps. The red dots of (b) are fixation points, the green dots are the first fixations.

It was decided to divide the studies into two groups based on the temporal features of the stimuli: the first group is studying visual attention with still images and the second with stereoscopic video content.

4.5.1 Stimuli: still stereoscopic images

Wexler and Ouarti investigated how various aspects of 3D scenes affect visual behavior [Wexler and Ouarti, 2008]. In the experiment three types of inclined surfaces (grid, texture, and dots) were used as stimuli. It was demonstrated that saccades tend to follow surface depth gradients and that vergence is dominated only by binocular disparity.

Jansen et al. [Jansen et al., 2009] studied the influence of disparity on fixations and saccades in the free viewing of 2D and 3D images of natural scenes, pink noise, and white noise. An analysis was performed using data from the left eye. They found that disparity information had an influence on basic eye movements, causing an increase in the number of fixations, a decrease of fixation duration over time (only for pink and white noise), and a shortening of saccade length over time. The saliency of mean luminance, luminance contrast, and texture contrast was compatible across 2D and 3D stimuli. Mean disparity had a time dependent effect for 3D stimuli. The disparity contrast was elevated at fixated regions in 3D noise images but not in 3D natural scenes. They reported that participants fixated closer locations earlier than more distant locations in the image.

Previous works were supplemented by Wismeijer et al. [Wismeijer et al., 2010], who investigated whether saccades are aligned with individual depth cues or with a combination of depth cues. Similar to Wexler and Ouarti [Wexler and Ouarti, 2008], the experiments were conducted using incline in depth surfaces. Such stimuli combined

monocular perspective cues and binocular disparity cues and specified different plane orientations with different degrees of small and large conflict between two sets of cues. It was discovered that the distributions of spontaneous saccade directions followed the same pattern of depth cue combination as perceived surface orientation: a weighted linear combination of cues for small conflicts and cue dominance for large conflicts. By examining the relationship between vergence and depth cues, they reached the same conclusion as Posner [Posner, 1980], that vergence is dominated only by binocular disparity.

Another study by Gautier and Le Meur investigated the influence of disparity on saliency [Gautier and Le Meur, 2012]. Their results claim that visual exploration is affected by the introduction of binocular disparity, i.e. the participants tend to first look at closer areas (in terms of depth) and then direct their gaze to more widespread locations.

Czuni and Kiss analyzed differences in the distribution of fixation points in 3D conditions in comparison with 2D [Czuni and Kiss, 2012]. During the eye-tracking experiment, fixation points were collected in mono and stereo conditions from 66 images. By examining image contours, depth contours, disparity changes between fixation points, and the clustering of fixation points, only slight differences were found in the special distribution of fixation points.

4.5.2 Stimuli: stereoscopic videos

Ramasamy et al. studied the feasibility of using eye tracking for stereoscopic filmmaking in order to identify elements that distract the audience from the flow of the movie [Ramasamy et al., 2009]. The gaze patterns of one scene have been analyzed to identify regions of interest in the frames. It was found that gaze points were concentrated at the far end of a scene showing a long deep hallway in the S3D version while being more spread out in the 2D version.

Other work that used video clips for the exploration of gaze patterns was presented by Häkkinen et al. in [Hakkinen et al., 2010]. In the experiment, observers watched 2D and 3D versions of four short video sequences with durations ranging from 5 to 22 seconds. The task was to compare these two versions and report which version was better. It was reported that in the S3D version eye movements were more widely spread. Therefore, the opposite conclusion was reached in the study by Ramasamy et al. [Ramasamy et al., 2009].

Another study was conducted by Huynh-Thu et al, who discovered no evidence for fixations being more widespread when viewing S3D [Huynh-Thu and Schiatti, 2011]. Although, no strong evidence was found of the opposite either. It was reported that "the spread of fixations depended highly on the content characteristics and narrative flow of the video, and not only on the depth effect provided by the 3D stereoscopic version." During the experiment, observers watched 21 video sequences with various durations ranging from 8 to 143 seconds in 2D and 3D modes. It was reported that the average fixation frequency and average fixation duration were lower when viewing 3D stereoscopic content; while the average saccade velocity was higher.

4.5.3 Analysis of eye movements with state-of-the-art studies

A comparison of eye movements for 2D and S3D conditions from the literature is summarized in Table 4.1, where V is the viewing distance. This table also presents experiment conditions. Interestingly, not all of the studies reported the viewing distance and display

size despite the importance of controlling the visualization parameters for comfortably presenting stereoscopic content (see Chapter 2). No studies have been carried out to investigate whether eye-movements are affected by visual discomfort. However, to our knowledge, just one study by Huynh-Thu et al. [Huynh-Thu and Schiatti, 2011] took into account the limits of the comfortable viewing zone. For example, Jansen et al. [Jansen et al., 2009] displayed stimuli with depth maps restricted to a disparity range between -80 and 80 pixels on an autostereoscopic display in their experiment. However, the comfortable viewing zone for the given conditions should remain within the disparity range -28 and 28 pixels as indicated by the equation 3.5- 3.6.

On the other hand, there is no coherence in the measured indicators between studies. This complicates the comparison of quantitative results. For example, saccade length, saccade duration, or saccade velocity has been reported in Table 4.1. In the case of fixations: the number of fixations and the fixation duration or fixation frequency. Due to the absence of standards in eye-tracking studies, it is quite difficult to know which indicator is the most representative for the comparison of eye movements.

In addition, different studies have found contradictory results for eye movements and gaze distributions. A possible explanation of such results is the absence of control of the visualization space, the rendering of the stimulus, different conditions for watching in 2D and 3D mode (with and without glasses), and the presentation of the same content in 2D and 3D mode twice, which can lead to memorizing and can influence eye-movements.

4.6 Conclusions

This chapter presented state-of-the-art studies about S3D visual attention. In addition, a review of recent studies comparing visual attention for stereoscopic with non-stereoscopic conditions. The conclusions are:

- The absence of a standard or guidelines in protocol for eye-tracking studies leads to a lack of coherence in reported indicators between the studies and the experimental conditions.
- There are some contradictions in the results studying eye-movements in 3D. Such differences can be explained by the absence of control of the visualization space, the rendering of stimulus, different conditions of watching in 2D and 3D mode (with and without glasses), the presentation of the same content in 2D and 3D mode twice, which can lead to memorizing and can influence eye-movements.
- The impact of discomfort on eye movements is not known in S3D condition. However, most of the studies of visual attention have not considered the comfortable viewing zone while displaying 3D content.

In the next chapter, our goal is to design a new subjective experiment with fully controlled technical parameters that takes the limitations mentioned above of the existing studies into consideration.

Table 4.1: Summary of the studies comparing visual attention for 2D and S3D conditions

Fixations: 2D compared to 3D	Saccades: 2D compared to 3D	Spatial distribution of fixation points and other findings
[Jansen et al., 2009], 28 images 20s, 14 observers, 18.1" autostereoscopic display, $V = 2H = 60cm$)		
increase in the number of fixations in 3D, a decrease of fixation duration over time (only for pink and white noise stimuli) in 3D	shortening of saccade length over time in 3D	participants fixated closer locations earlier than more distant locations in the image
[Ramasamy et al., 2009] (1 video clip)		
was not reported	was not reported	fixation points more spread in 2D
[Hakkinen et al., 2010] (4 video clips 5-22 s, 20 observers, 46" display with passive glasses in 3D and 2D mode, $V = 1.5H = 140cm$)		
was not reported	was not reported	fixation points more spread in 3D
[Huynh-Thu and Schiatti, 2011] (21 video clips 8-144 s, 18 observers, 46" display with passive glasses in 3D mode, no glasses in 2D mode, $V = 1.8H = 180cm$)		
average fixation frequency and average fixation duration lower in 3D	average saccade velocity higher in 3D	no difference: fixation distribution depends on content characteristics and narrative flow
[Czuni and Kiss, 2012] (66 images, avg 8 observers per test, display with active glasses in 3D mode, V in not indicated)		
no difference in fixation duration	no difference in saccade duration	fixation points slightly more spread in 2D
[Iatsun et al., 2013] (6 video clips 10 min, 20 observers, 46" display with passive glasses in 3D mode, no glasses in 2D mode, $V = 2H = 114cm$)		
average fixation frequency no difference, average fixation duration lower in 3D	average saccade duration no difference, average saccade number higher in 3D	average blinking number higher in 3D, average blinking duration no difference

Chapter 5

Studies of visual attention in S3D

Contents

5.1 Introduction

This chapter investigates whether the visual attention should be considered when designing an objective 3D quality metric. First, the visual attention in 2D and S3D is compared using simple test patterns. The conclusions of this first experiment are validated using complex stimuli with crossed and uncrossed disparities. In addition, we explore the impact of visual discomfort caused by excessive disparities on visual attention. Finally, a new metrics considering the saliency maps and depth maps is proposed.

5.2 Experiment 1: simple visual stimuli

In Chapter 4 the studies of visual attention in S3D were reviewed. We have noticed that most of them do not consider simuli visualization the comfortable viewing zone. Another issue is that the same stimuli are shown in 2D and 3D modes. This may lead to the memorization of the content, which would then involve the top-down mechanisms of visual attention. Therefore, we decided to use simple stimuli for our experiment instead of complex scenes to avoid the top-down mechanisms of visual attention. In addition, the generated stimuli had to stay within the zone of comfort DoF=$\pm$0.2 diopters.

The goal of the current experiment is to compare mechanisms of visual attention in 2D and S3D using simple controlled visual stimuli to find out whether texture contrast or binocular disparity is a more influential factor in guiding our gaze.

5.2.1 Stimuli generation

Each stimulus contains four spheres equidistant from the center of the screen on a gray background. Any of the four spheres can be in one out of five possible locations in depth:

1. in front of a display: close to a display plane;
2. in front of a display: far from a display plane;
3. behind a display: close to a display plane;
4. behind a display: far from a display plane;
5. in the display plane (in the case of a 2D image).

To study the influence of texture on the selection process, two possibilities were available: a sphere could have the same gray color as the background or a checkerboard texture.

Figure 5.1.a illustrates one of the possible sphere arrangements in depth or *sphere set-up*. By changing the camera baseline and the convergence distance, it is possible to generate three stereoscopic images using the same sphere set-up:

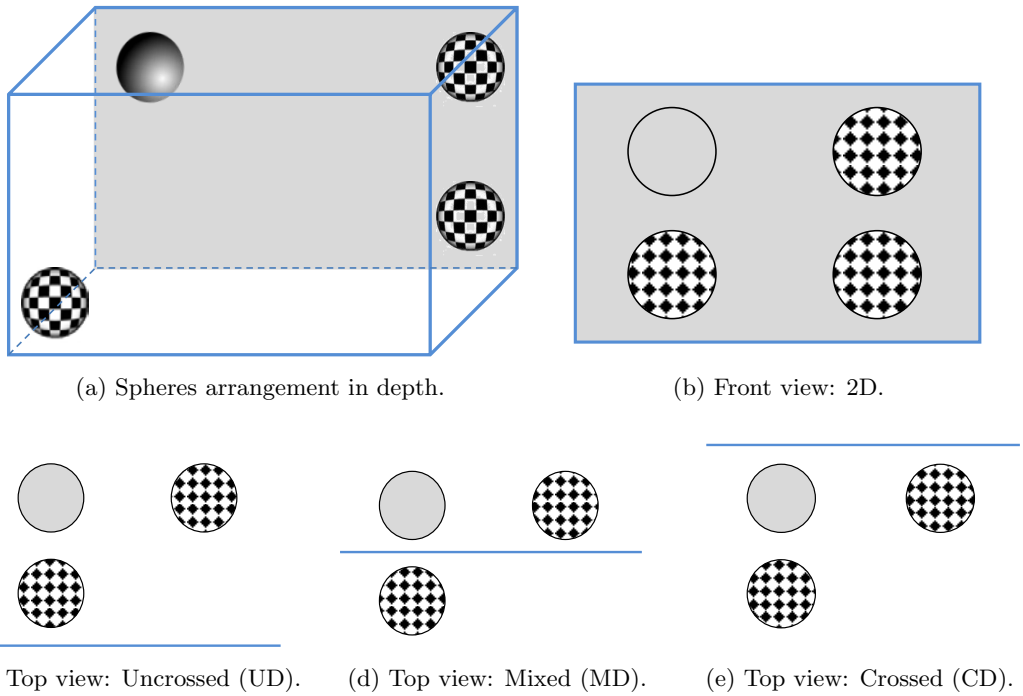
- an image with *uncrossed disparity (UD)* (when all of the spheres are behind the display plane);
- an image *mixed disparities (MD)* (when some spheres are in front of the display plane and the rest are behind);
- an image with *crossed disparity (CD)* (when all of the spheres are in front of the display plane).

The forth option is a 2D image – with a front view of the set-up. Figure 5.1 depicts four stimuli produced from the same sphere set-up. The display plane is the blue solid line. The figures are drawn considering that an observer is located in front of the display plane.

Using Blender software, images were generated with a resolution of 1920×1080 using a virtual camera with a sensor size of $32mm \times 16mm$. Multisampling with 8 sample anti-aliasing was used to smooth the edges. The blur effect was disabled to guarantee the sharpness of the scenes. Shooting was performed with a parallel-rig, using HIT to create the desired disparity. In order to avoid a black border after the post-production shift, extended borders were rendered for every image.

The image parameters are presented in Table 5.1, where b is the baseline distance, $dCon$ - convergence distance, DoF - depth of focus. The disparity range is the depth range of a scene, which consists of maximum crossed and uncrossed disparity on the screen used for the experiment. The amount of depth was defined as DoF=$\pm$0.1 diopters for crossed and mixed disparities and DoF=+0.15 diopters for uncrossed. The total amount of perceived depth in cases of images with mixed disparities reaches DoF=0.2 (0.1 in front of display and 0.1 behind). So, a bigger DoF value was used for uncrossed disparities in order to increase the amount of perceived depth.

In total, 56 images were generated. The 14 different sphere set-ups illustrated in Figure 5.2 were used with 4 variations of disparities, as explained above. Each set-up in Figure 5.2 represents a 2D stimulus and a front view of the 3D representation at the same time. Spheres marked in bold magenta are closer to the observer in depth, independent of disparity type. In 2D, all four spheres are on the display plane. The name of the stimulus consists of the corresponding number for sphere set-up and the designation of



(a) Spheres arrangement in depth. (b) Front view: 2D.

(c) Top view: Uncrossed (UD). (d) Top view: Mixed (MD). (e) Top view: Crossed (CD).

Figure 5.1: Four images with different disparities and the same sphere set-up (a) sphere set-up, e.g. arrangement in depth, (b) 2D image, (c) image with uncrossed disparity, (d) image with mixed disparities, (e) image with crossed disparity. The display plane is the blue solid line.

Table 5.1: Parameters for images with uncrossed (UD), mixed (MD) and crossed (CD) disparities.

Scene	b, mm	dCon, m	Disparity range, mm	DoF, D
UD	315	6	[0;23]	+0.15
MD	500	8	[-15;15]	±0.1
CD	300	9.5	[-15;0]	-0.1

disparities presented in a stimulus. For example, "11_MD" means sphere set-up 11 in Figure 5.2 with mixed disparities in Figure 5.1.d), where the sphere in the bottom left corner comes out of the screen and three other spheres are behind the display plane.

5.2.2 Experimental set-up and methodology

Test set-up: the subjective experiment was performed in the test room in compliance with the recommendation ITU-R BT.500-13. The Tobii x50 eye-tracker was used to track eye movements of observers. An LG 42" 42LW line interleaved stereoscopic display was used for the visualization of the stimuli (see Fig. 5.4.a). Its dimensions are 93×52 cm. It has a resolution of 1920×1080 in 2D and 1920×540 per view in 3D. The viewing distance was 4.5 times the height of the display. An PC was used to record eye-tracking data. The psychophysical test set-up is schematically presented in Figure 5.3.

During tracking, the Tobii x50 eye-tracker uses near infrared diodes to generate reflection patterns on the corneas of the eyes of the observer. These reflection patterns were collected by a camera and analyzed by Clear View software. Finally, it is possible

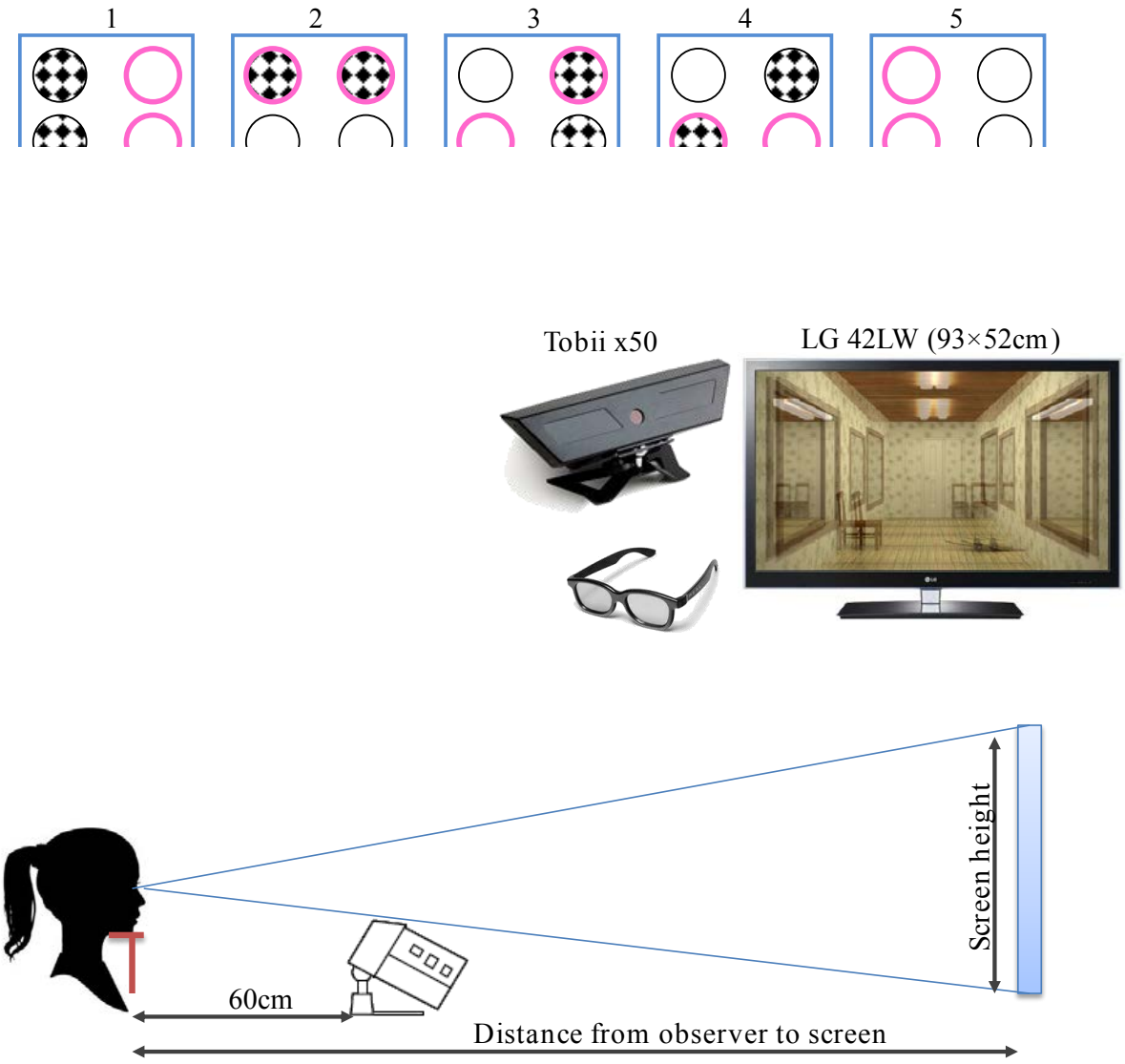
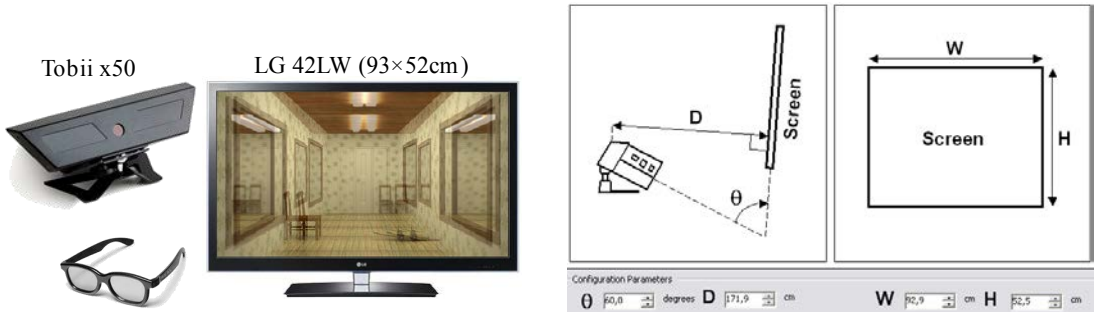


Figure 5.3: General scheme of experimental set-up.

to collect the gaze points on the screen, i.e. the locations where an observer was looking. The Tobii x50 requires geometrical adjustments. The head movements of observers were restricted with the chin rest. The distance from the observer to the eye tracker was 60 cm. There are some restrictions on the placement of the Tobii eye-tracker: first, the distance from the observer to the eye tracker should be around 60 cm; second, the eye tracker should be positioned straight in front of the stimuli and at a particular angle below the user (see Fig. 5.4.b). However, once it has been configured, eye tracking is fully automatic.

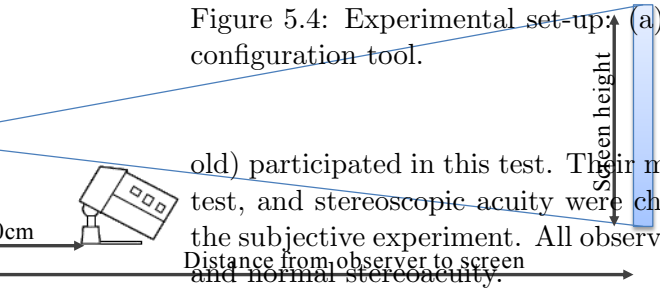
To build saliency maps, it is necessary to calculate the number of pixels per degree of visual angle. Taking into account the width of the screen ($SW = 93cm$) and the distance from the observer to the screen ($SD = 234cm$), the equation 4.2 was applied to calculate the visual angle in degrees $\Theta_W = 22.47$. Therefore, the number of pixels per one degree is $\frac{1920}{22.47} = 85$ pixels per degree.

Obsevers: 28 non-expert observers (19 males and 9 females ranging from 20 to 52 years



(a) Equipment. (b) Tobii x50 configuration tool.

Figure 5.4: Experimental set-up: (a)Equipment used in the experiment, (b) Tobii x50 configuration tool.



old) participated in this test. Their monocular acuity, color vision, far vision test, fusion test, and stereoscopic acuity were checked using Essilor Ergovision equipment prior to the subjective experiment. All observers had normal or corrected to normal visual acuity and normal stereoacuity.

Methodology: for each observer, the experiment consisted of five stages: visual test, reading of the instruction sheet, 2 steps calibration (5-points then 9 points), training, and, finally, the visual attention test. The instruction sheet offered some explanations on how to behave during the calibration stage, the training stage, and during the test itself. Observers were allowed to look at the images freely without any instructions. The instructions were explained by the examination as well to ensure that the observers understood the task.

During the test, all 56 prepared images were displayed to every observer. The duration of the test was 9 minutes 30 seconds. Each image was presented for 5 seconds and separated from the subsequent one by displaying a gray screen for 5 seconds.

Calibration: the eye tracker requires some calibration to learn the characteristics of the eyes of each observer. The observers were asked to put on the passive polarized glasses in order to begin the calibration. During the first step of calibration stage, an observer simply looked at a dot that appeared in different positions of the screen. The calibration procedure was fully automatic and took about 30 seconds. A five-point calibration procedure was used in our experiment.

Even if the software reported that the calibration was done successfully, there were still a few special circumstances in which the system had tracking difficulties, such as for people with bi-focal glasses or people with elements (eye lids, mascara, etc.) that significantly block the eye tracker camera's view of the subject's eyes. Thus, after the first step of automatic calibration, the second step of calibration was performed: a specially designed chart was used in order to check whether the device was able to track the observer's gaze correctly (see Fig. 5.5.a). The calibration image contained nine white dots on top of a picture of an airplane. Observers were instructed to focus on each white point for 3 seconds. Figure 5.5.b presents an example of a successful calibration. If the eye-tracker had difficulties properly recording the data of an observer or the calibration process was unsuccessful, the resulting gaze plot looked similar to Figure 5.5.c. No observers with unsuccessful gaze plots were allowed to participate in the test.

(a) Calibration chart. (b) Successful gaze plot. (c) Unsuccessful gaze plot.

Figure 5.5: Calibration phase: (a) Calibration chart, (b) successful gaze plot, (c) unsuccessful gaze plot.

Training: a training was done in stereoscopic mode using 10 still images with various sphere set-ups and disparity combinations. The training phase was designed to familiarize observers with the test conditions. The duration of the training was 1 min 40 seconds.

5.2.3 Eye-tracking data analysis

The gaze plots of every observer were analyzed in order to find out if the sphere selection preference was based on texture or depth. A table was created which contained the sphere selection priority for each stimulus for each observer. Every sphere had a fixed position number, which was constant within all the images: the sphere in the top left corner is numbered s1, the top right corner – s2, the bottom left – s3, the bottom right – s4 (see Fig. 5.6). Based on the observer's gaze plot for each position number, the selection order number was collected. Figure 5.7 presents the gaze plot of observer 5 for image 11_MD; position number for every sphere presented in Figure 5.6. So observer 5 first selected the sphere with position number s1 (order is 1), then with position number s3 (order is 2), then with position number s4 (order is 3) and the last sphere with position number s2 (order 4). This data was collected for each observer for all the stimuli. See the example of such a table in Table 5.2.

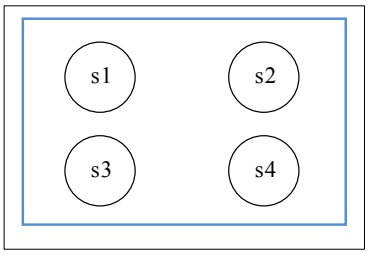


Figure 5.6: Fixed position number of every sphere.

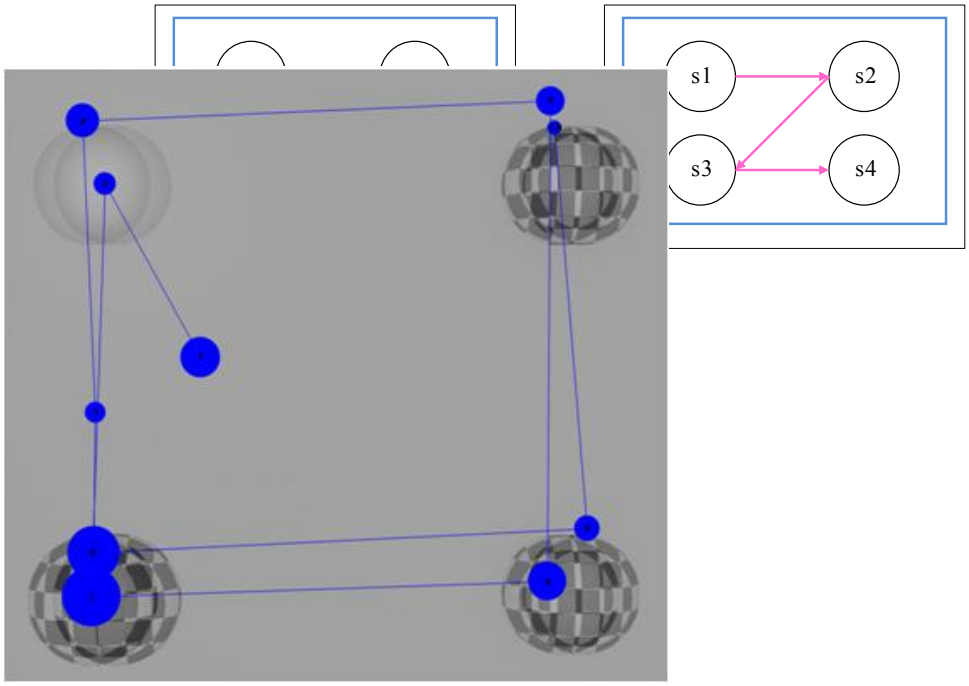


Figure 5.7: Gaze plot of image 11_MD with mixed disparities of observer 5.

Table 5.2: Order of spheres' selection for observer 5 for image 11_MD.

Image	s1	s2	s3	s4
11_MD	1	4	2	3
11_2D	1	2	3	4
11_CD	2	4	1	3
11_UD	1	3	2	4

It is interesting to note that some observers looked at all of the images the same way during the experiment independent of texture or depth variation. For example, all observations were started at the top left corner and then continued clockwise; or another pattern that was observed is from top left to right and then from bottom left to right (see Fig. 5.10). Since the presentation of stimuli was only 5 seconds, supposedly mostly only bottom-up processes should be involved e.g. visual attention should be unconscious. But it seems that some observers were intentionally following the same pattern of observation during the whole test. Since the duration of the test was almost 10 minutes, we found this behavior unnatural.

For 3 observers, we found that for several images all the fixation points were lost. This may have occurred due to some of them closing their eyes during the test or the

signal was lost due to head movement or displacement from the chin rest. Observers whose signals were lost were excluded from the data analysis.

Then, all the gathered information was converted to one data table, where texture: 0 – gray, 1 – checkerboard; depth: 0 – when a sphere has zero disparity, -1 – a sphere with crossed disparity, 1 – a sphere with uncrossed disparity; order is the priority of selection of a given sphere: 1 – selected first, 2 – selected second, etc.; position is a fixed position for spheres for all the images (Fig. 5.6). An example for image 11_MD for observer 5 is presented in Table 5.3.

Table 5.3: Collected data of the image 11_MD for observer 5

Image	Observer	Position	Texture	Depth	Order
11_MD	5	s1	0	1	1
11_MD	5	s2	1	1	4
11_MD	5	s3	1	-1	2
11_MD	5	s4	1	1	3

5.2.3.1 Influence of depth on visual attention

MANOVA univariate tests of significance for order showed that depth significantly influences the order of selection of the spheres $F(2, 4456) = 6.63$, $p < 0.05$, $p = 0.0013$. The analysis was performed for all the data. Then, the data was divided into two separated data sets: (1) all spheres with crossed disparity, (2) all spheres with uncrossed disparity. Next the analysis was performed for each data set separately. The analysis of the crossed disparity set showed that a sphere with a crossed disparity significantly influences the order of selection in 3D compared with 2D: $F(1, 3264) = 13.14$, $p < 0.05$, $p = 0.0003$; whereas the influence of uncrossed disparity on the order of sphere selection was insignificant. The results are presented in Figure 5.8.

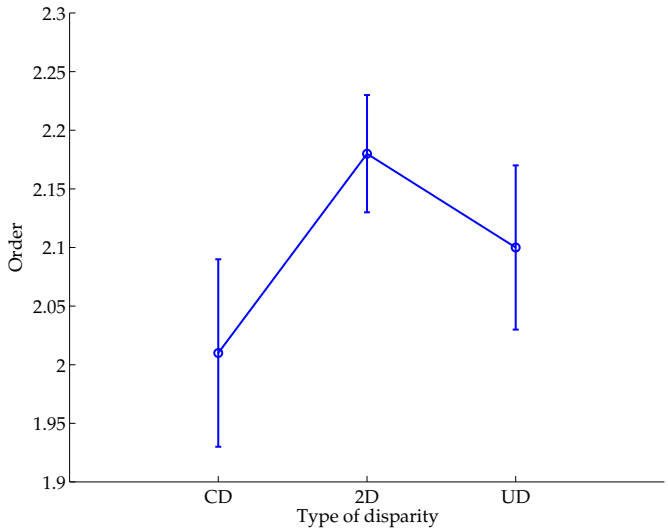


Figure 5.8: Average order of selection of sphere per disparity. Error bars represent 95% confidence interval.

Wang *et al* have reached similar conclusions by demonstrating that objects located closer to spectators attract more visual attention [Wang et al., 2013]. However, their

analysis has not taken into account the position of the screen plane. Hence, no distinction between crossed and uncrossed disparities has been made.

5.2.3.2 Influence of texture on visual attention

Univariate tests of significance for order showed that texture significantly influences sphere selection order $F(1, 4456) = 12.31$, $p < 0.05$, $p = 0.0005$. Also, it is significant for both crossed $F(1, 3264) = 9.95$, $p < 0.05$, $p = 0.0016$ and uncrossed $F(1, 3424) = 8.4$, $p < 0.05$, $p = 0.0038$ disparities. These results are presented in Figure 5.9.

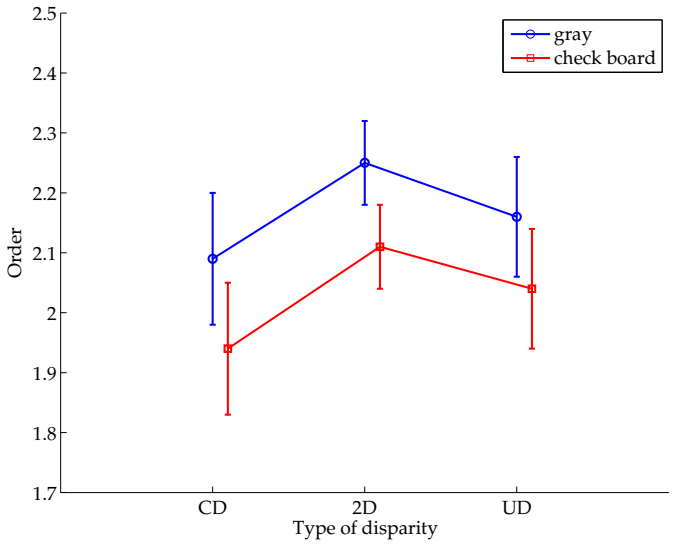


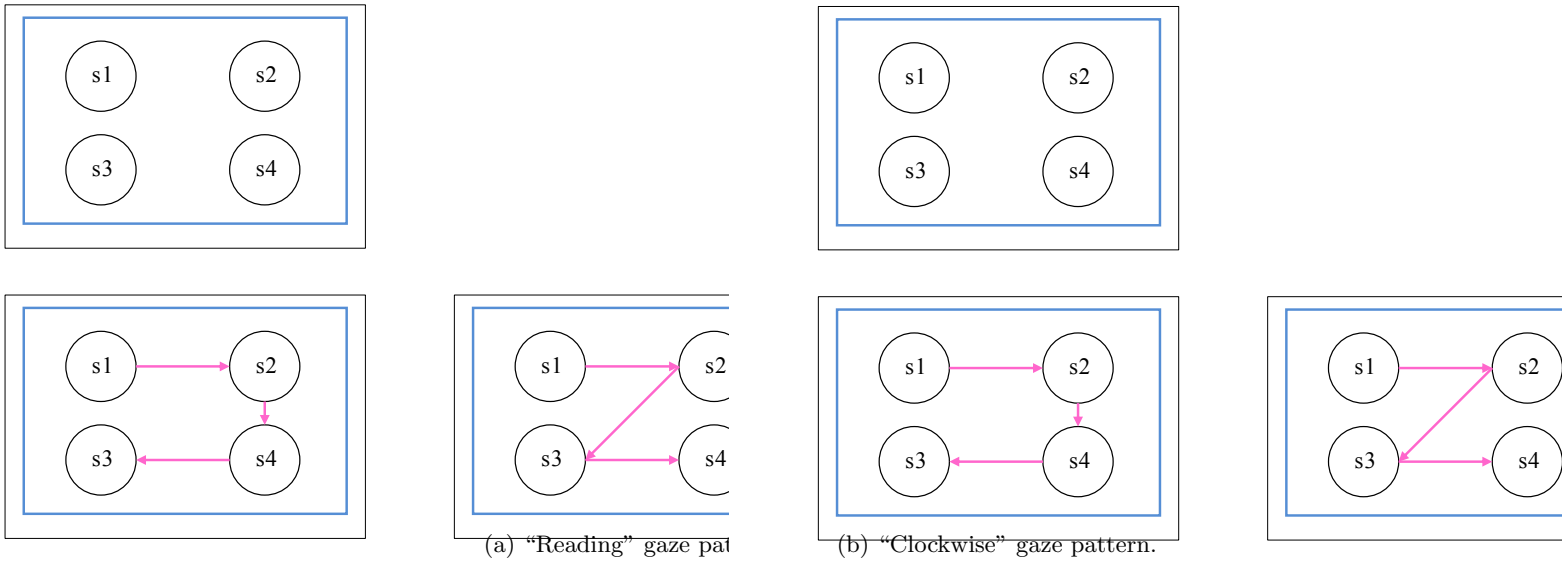
Figure 5.9: Average order of selection of sphere per texture. Error bars represent 95% confidence interval.

Independent of the type of the disparity (zero disparity, crossed, or uncrossed), spheres with the checkerboard texture were selected before spheres with no texture. Spheres coming out of the screen with the checkerboard texture had the highest selection priority.

5.2.3.3 Influence of the position of the spheres on test results

The analysis showed that the position of the sphere significantly influences the selection priority for crossed disparity $F(3, 3264) = 31.14$, $p < 0.05$, $p = 0.0000001$, as well as for uncrossed disparity $F(3, 3424) = 30.17$, $p < 0.05$, $p = 0.0000001$. There could be several explanations for such results: since the spheres were presented in two rows, observers followed the familiar reading pattern of top left to top right, then bottom left to bottom right, as if it was a text (see Fig. 5.10.a). The higher number of sphere position resulted in the lower priority of selection. This tendency can be seen for depth (Fig. 5.11.a), as well as for texture (Fig. 5.11.b). Another possible explanation is that since the time of the presentation of one image was only 5 seconds, it was easier for some observers to apply a scheme of observation, for example, clockwise as it shown in Figure 5.10.b and follow it until the end of the experiment in order to see all the spheres.

The influence of sphere position and depth on selection order is presented in Figure 5.11.a. Figure 5.11.b illustrates that spheres with texture are preferred (have a lower order of selection, e.g. higher priority) to gray spheres. If textured spheres were

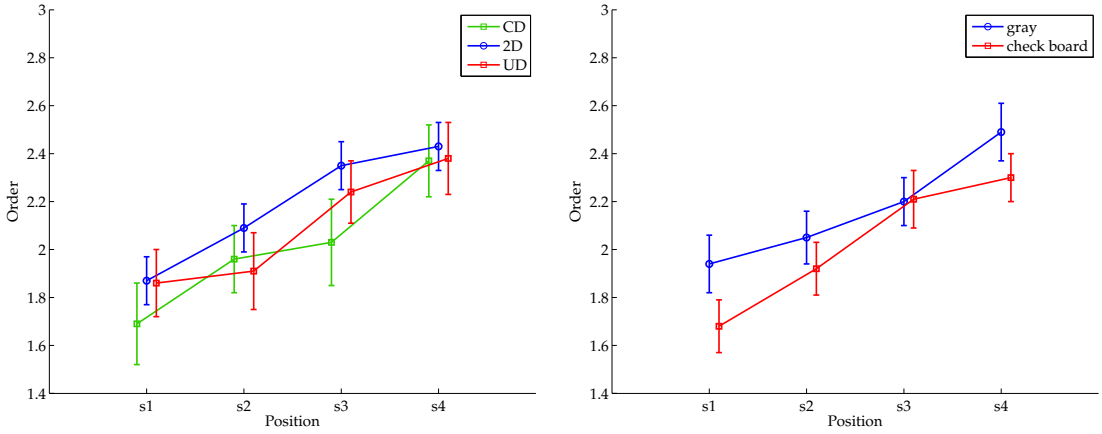


(a) "Reading" gaze pat

(b) "Clockwise" gaze pattern.

Figure 5.10: Stimuli' observation patterns.

situated in the top left corner of the screen, they had a significantly higher priority of selection than a non-textured sphere in the same position of the screen.



(a) Influence of the position of a sphere and depth on the order of selection.

(b) Influence of the position of a sphere and texture on the order of selection.

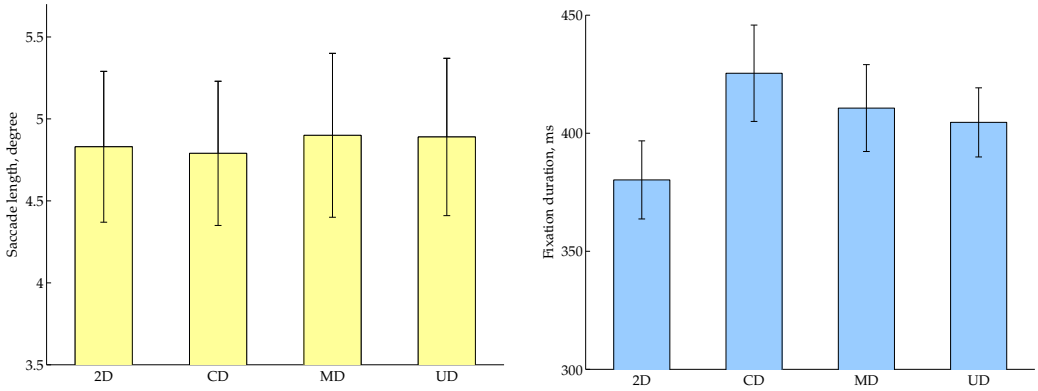
Figure 5.11: (a) Influence of the position of a sphere and depth on the order of selection (b) Influence of the position of a sphere and texture on the order of selection. Error bars represent a 95% confidence interval.

5.2.3.4 Saccade length and fixation duration

The current experiment was designed in a way to involve mostly bottom-up processes of visual attention during eye-tracking with simple visual stimuli. Hence, it was interesting to find out if the depth component influences saccade length and fixation duration.

Figure 5.12.a illustrates that the depth component did not have any influence on the average saccade length considering the confidence interval of 95%. The same results were obtained with a paired samples t-test: there is no significant difference between saccade lengths for zero disparity, mixed, crossed, and uncrossed disparities.

The average fixation duration for every disparity is presented in Figure 5.12.b. The analyses of average fixation duration with a paired samples t-test showed that there is a significant difference for fixation duration $t(13) = -4.06$, $p < 0.05$, $p = 0.0013$ in the scores between zero disparity and mixed disparities, as well as a significant difference between zero disparity and crossed disparity: $t(13) = -2.68$, $p < 0.05$, $p = 0.019$, and between zero disparity and uncrossed disparity: $t(13) = -3.98$, $p < 0.05$, $p =$



(a) Average saccade length. (b) Average fixation duration.

Figure 5.12: (a) Average saccade length. (b) Average fixation duration. Error bars represent a 95% confidence interval.

0.0016. A comparison of the rest of the pairs did not expose any significant results. The introduction of depth (z-axis) increased fixation duration independent of the type of the disparity.

5.2.3.5 Discussion and conclusions

This study was launched to evaluate the features influencing the saliency of the objects in stereoscopic and non-stereoscopic conditions by using content with simple stimuli containing four spheres. The two features of spheres have been evaluated: position in depth and texture. It was discovered that:

- **Eye movements:** There was no significant difference between 2D and 3D conditions for average saccade length. Average fixation durations were higher when viewing stimuli with spheres in 3D.
- **Disparity:** Objects with crossed disparity are significantly important for the selection process as well. However, there was no difference in selection preference for objects with uncrossed disparity in comparison to 2D objects.
- **Texture:** Textured spheres are selected before non-textured independent of their position in depth.

Based on the findings outlined above, it seems that the mechanisms of visual attention for uncrossed disparity images is similar to 2D images. Therefore, saliency maps for stereoscopic scenes remaining behind the display plane can be computed using any of existing 2D visual attention models.

The influence of position on the screen of a sphere had a significant impact on the selection priority. This can be explained by the design of the stimuli: two objects in the top row, two objects in the bottom row, which can be read in the familiar way from top left to top right, then from bottom left to bottom right. Therefore, we decided to design a new experiment using complex visual stimuli to avoid such bias.

5.3 Experiment 2: complex stimuli with only uncrossed disparity objects

The goal of this experiment is to establish the relationship between visual attention, texture complexity and the depth component. In the previous experiment, simple visual stimuli were used for eye-tracking. Each stimulus contained four spheres placed in two rows. However, such an arrangement created a bias based on the sphere position on the screen. To avoid bias, we decided to use controlled scenes with complex stimuli. It was established that texture significantly influenced the priority of selection of spheres, so three levels of texture complexity were applied to every scene in this experiment.

Another open question is whether discomfort sometimes generated by 3D content has an influence on the way we observe images. To this end, this section examines visual attention and investigated potential differences in attention behavior between viewing still images with different levels of texture complexity and disparities. At first, the basic eye movement properties (the length of saccades and the duration of fixations) are analyzed. Then the temporal evolution of these parameters is studied over a viewing duration of 20 s. Then heat maps are explored to investigate if the visual exploration patterns of the observers changed due to the introduced parameters.

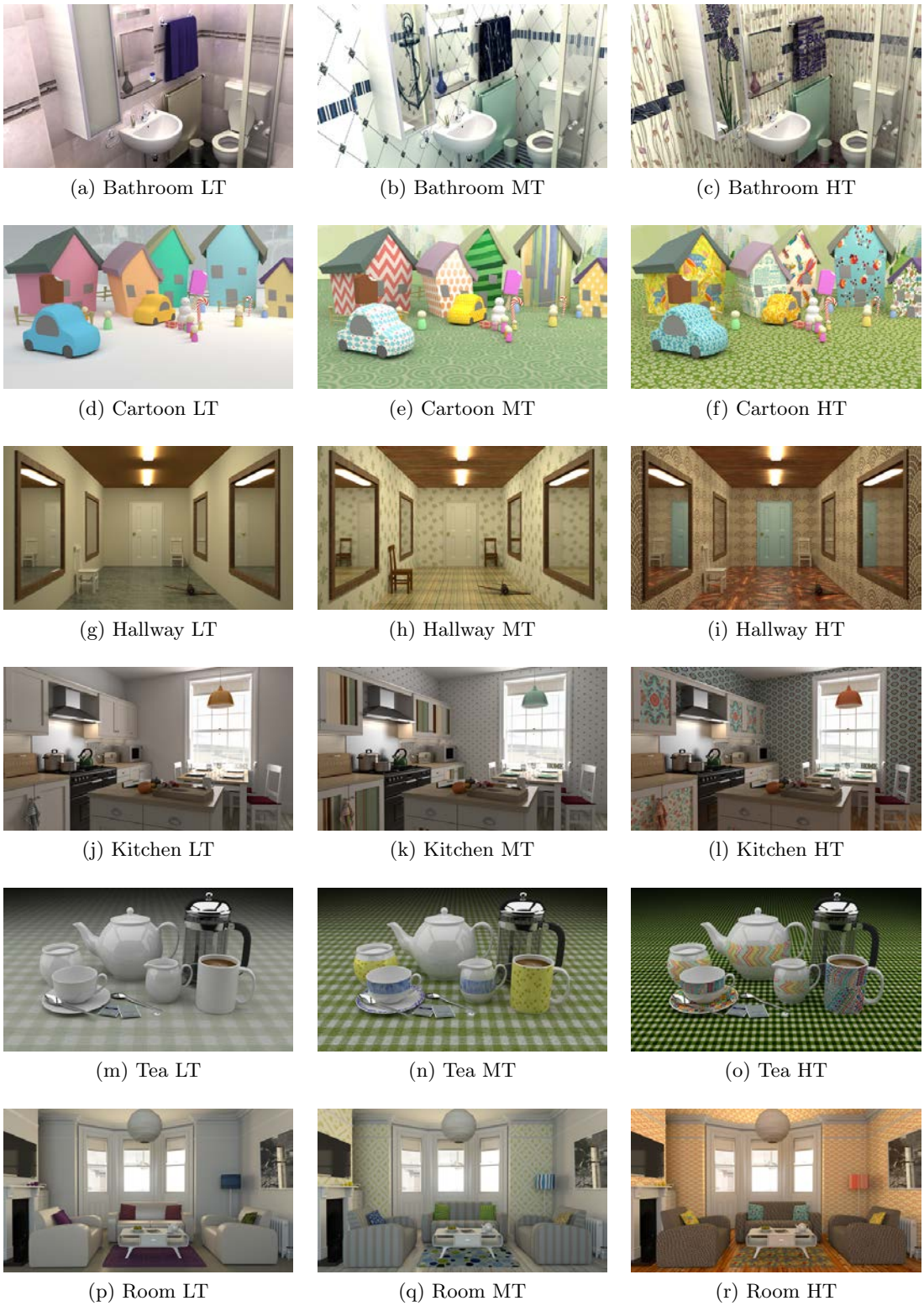
5.3.1 Stimuli generation

All synthetic scenes were generated and rendered using Blender software, which allows for the foreground and background distances of a scene to be measured and for the stereoscopic camera parameters to be controlled accurately. Six different scenes were selected for the experiment: "Bathroom", "Cartoon", "Hallway", "Kitchen", "Tea", and "Room" [Bobal57 et al., 2012]. Each scene has three texture complexities: low, medium, and high. Underscores after the scene name denote its texture complexity: LT – low texture complexity, MT – medium texture complexity, and HT – high texture complexity. In the experiment, low texture is the absence of a pattern on objects and low contrast (if it was possible). Simple geometrical patterns were selected for medium texture complexity and complex non-geometrical patterns were selected for high texture complexity. The contents of the scenes were mainly indoors as it is very difficult to find a suitable substitute for outdoor textures, such as leaves, grass, or sky, when dealing with varying degrees of complexity. Examples of the generated scenes are illustrated in Figure 5.13. The image parameters are presented in Table 5.4.

Table 5.4: Scene parameters.

						DoF=0.1		DoF=0.3	
Scene	f,mm	dCon,m	fg,m	bg,m	roi,m	b,mm	DR	b,mm	DR
Bathroom	32	1.6	1.7	4.4	2.8	41	[0;15]	140	[0;51]
Cartoon	35	5	5.7	22	10.9	96.7	[0;15]	330	[0;52]
Hallway	35	3	5.3	16.8	10	54.6	[0;15]	180	[0;50]
Kitchen	24	1.5	1.6	6.3	2.2	42	[0;15]	145	[0;51]
Tea	35	0.43	0.44	1.2	0.7	10	[0;15]	34	[0;52]
Room	20	2.4	2.7	5.9	3.2	105	[0;15]	317	[0;46]

where f is the camera focal length, b - the baseline distance, $dCon$ - convergence distance, fg – foreground distance, e.g. distance from the camera to the closest object, bg –



(a) Bathroom LT (b) Bathroom MT (c) Bathroom HT

(d) Cartoon LT (e) Cartoon MT (f) Cartoon HT

(g) Hallway LT (h) Hallway MT (i) Hallway HT

(j) Kitchen LT (k) Kitchen MT (l) Kitchen HT

(m) Tea LT (n) Tea MT (o) Tea HT

(p) Room LT (q) Room MT (r) Room HT

Figure 5.13: Stimuli with uncrossed disparities and different texture complexities: LT - low, MT - middle, HT - high.

background distance, e.g. distance from the camera to the farthest object, *roi* - region of interest, and *DoF* - depth of focus. *DR* is the disparity range of a scene in the visualization space, which consists of the maximum crossed and uncrossed disparity in mm on the screen used for the experiment. Camera parameters were selected to correspond to DoF=0.1 diopters for the comfortable condition and DoF=0.3 diopters for the uncomfortable one. Therefore, the reconstructed amount of depth with the same DoF was almost the same for all the stimuli.

Images were rendered with a resolution of 1920×1080 using virtual camera with a sensor size of $32mm \times 16mm$. Multisampling with 8 sample anti-aliasing was used to smooth the edges. The blur effect was disabled to guarantee the sharpness of the scenes. Shooting was performed with a parallel-rig, using HIT to create the desired disparity. In order to avoid a black border after the post-production shift, extended borders were rendered for every image.

A detailed analysis of the relationships between camera space and visualization space and depth distortions for the comfortable condition and the uncomfortable condition is given in Annex A Figures A.7- A.12. The space outside the ZoC is marked in light gray and the region of interest as a magenta line.

In total, 54 images (6 images×3 depth levels×3 textures) were generated. 9 sets containing 6 images with different content were formed in order to prevent the observers from memorizing the images and hence using top-down visual mechanisms. Each set had two 2D images, 2 images with a comfortable depth level, and 2 images with an uncomfortable depth level.

5.3.2 Experimental set-up and methodology

The same experimental set-up and methodology as in Section 5.2.2 was used. The training was done in stereoscopic mode using three images with three levels of depth: 2D, DoF=0.1, and DoF=0.3. Each image was presented for 20 seconds and separated from the subsequent one by displaying a gray screen for 5 seconds. The images were different from those used in the test. The training phase was designed to familiarize observers with the test conditions. The duration of the training was 1 min 20 s.

During the test we displayed one of the nine sets of images. The duration of the test was 2 min 40 s. 135 people (106 males and 39 females from 21 to 60 years old) participated in the test. Thus, each image was observed by 15 subjects.

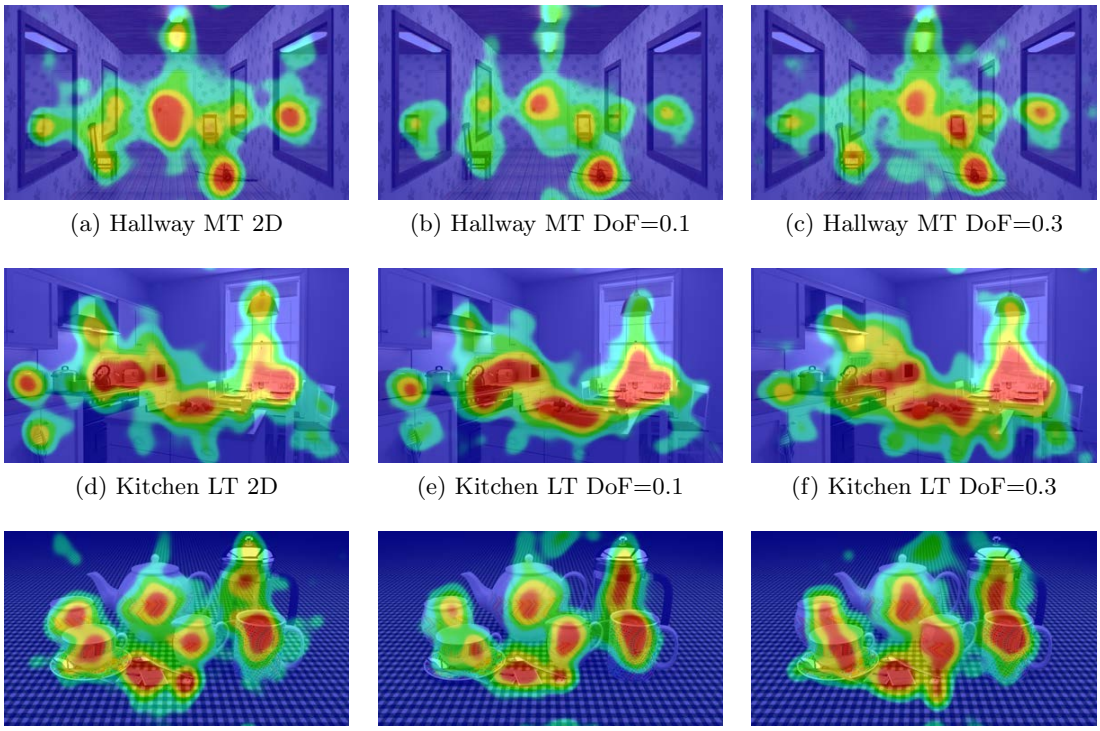
5.3.3 Eye-tracking data analysis

In this section the eye-tracking data is analyzed to study whether the introduction of binocular disparity, discomfort, and texture complexity had an effect on basic eye movement properties. The entire fixation data collected with the Tobii eye-tracker was used for analysis. All observers that could not complete the calibration process using the calibration chart were excluded before the test. In order to analyze the gaze behavior of observers, saliency and heat maps as well as fixation durations and the length of saccades were computed [Le Meur and Baccino, 2012, Le Meur, 2012]. During the experiment, images were separated by a gray slide without a fixation cross in the center, which is why the first fixation of each stimulus was not discarded.

5.3.3.1 Qualitative analysis based on heat maps

The heat maps representing the fixated areas of a stimulus were used to compare the gaze patterns of all observers. This method has been used in various studies [Huynh-Thu and Schiatti, 2011, Hakkinen et al., 2010, Ramasamy et al., 2009] since it allows the gaze behavior of an entire group of observers to be quickly and conveniently visualized. It should be noted that a normalization process is done for each heat map independently. As a consequence, it is difficult to precisely compare heat maps for different scenes.

For example, several heat maps are shown in Figure 5.14 for the scenes "Hallway", "Kitchen", "Tea". Each image on this figure shows a heat map corresponding to a viewing duration of 20 s. Another example is given for the scene "Room" in Figure 5.15: all nine heat maps (3 texture complexities × 3 depth levels) for each scene look qualitatively similar. Only some of the heat maps are presented here since the heat maps for the rest of the scenes did not reveal any particular difference. Visual analysis of the heat maps corresponding to a viewing duration of 20 s did not demonstrate any differences in gaze patterns for conditions with different disparities or for different texture complexities.



(a) Hallway MT 2D (b) Hallway MT DoF=0.1 (c) Hallway MT DoF=0.3

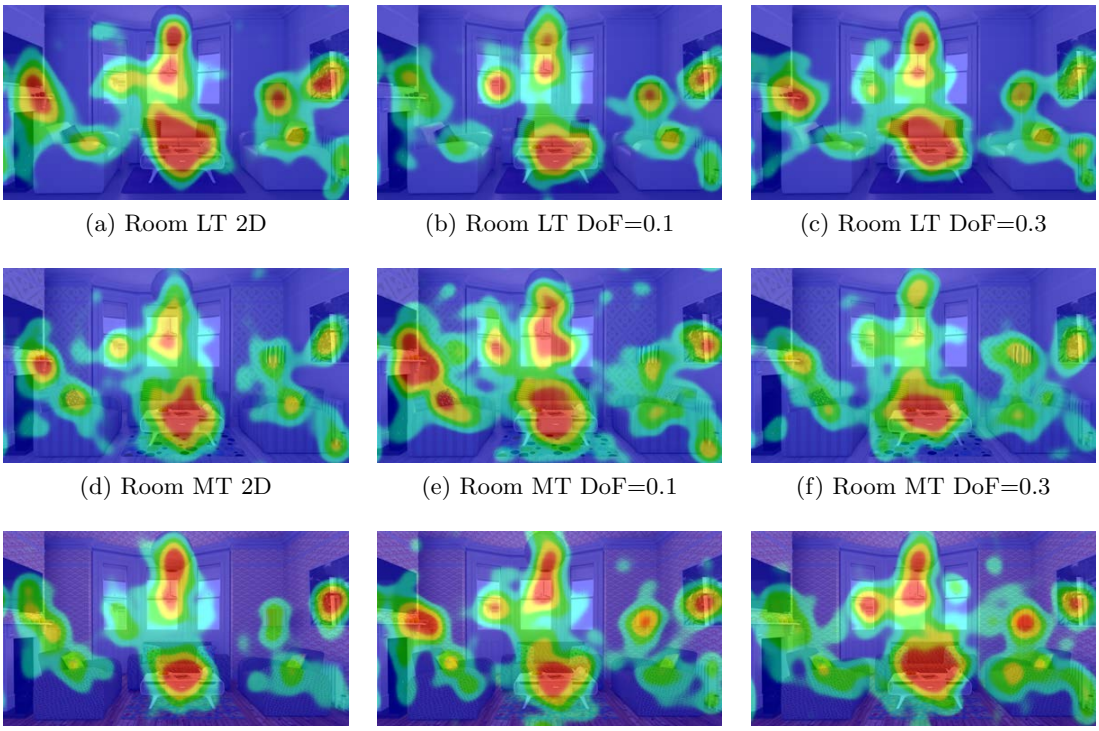
(d) Kitchen LT 2D (e) Kitchen LT DoF=0.1 (f) Kitchen LT DoF=0.3

(g) Tea HT 2D (h) Tea HT DoF=0.1 (i) Tea HT DoF=0.3

Figure 5.14: Heat maps for Hallway, Kitchen and Tea scenes with different depth levels and texture complexities: LT - low, MT - middle, HT - high.

5.3.3.2 Quantitative analysis

The correlation between the pairs of saliency maps for each scene with different depth levels was computed using the Pearson linear correlation coefficient (CC) and Area Under Curve (AUC). The idea is illustrated in Figure 5.16 for the "Bathroom" scene with a low



(a) Room LT 2D (b) Room LT DoF=0.1 (c) Room LT DoF=0.3

(d) Room MT 2D (e) Room MT DoF=0.1 (f) Room MT DoF=0.3

(g) Room HT 2D (h) Room HT DoF=0.1 (i) Room HT DoF=0.3

Figure 5.15: Heat maps for Room scene with different depth levels and texture complexities: LT - low, MT - middle, HT - high.

level of texture complexity. A higher AUC results in a better prediction. A value of 0.5 indicates a random performance while 1.0 denotes a perfect performance. The method used for the computation of the CC metric is presented in Section A.1 and AUC metric

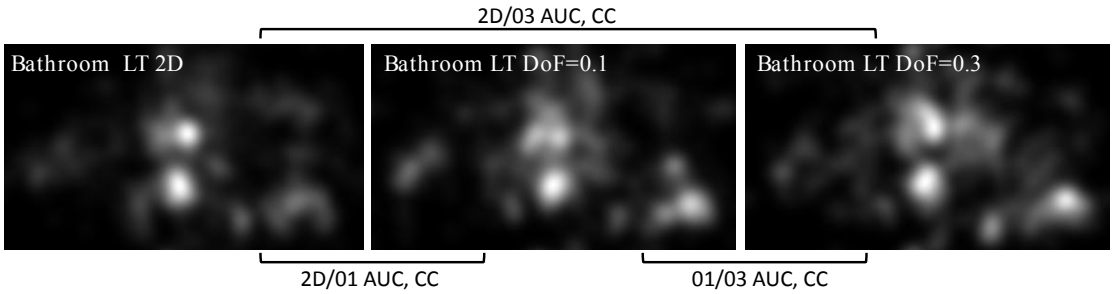


Figure 5.16: AUC and CC coefficients between the pairs of saliency maps for the "Bathroom" scene with different depth levels.

The AUC and CC metrics computed with saliency maps corresponding to the viewing duration of 20s are presented in Figure 5.17 and Figure 5.18 respectively. The average values are presented in Table 5.5, where *min* and *max* are the values corresponding to the minimum and maximum correlation coefficient for all the scenes. The AUC values and the CC values representing the correlation between saliency maps with different disparities are very high. This suggests that there is no strong difference between the

saliency maps for a viewing duration of 20 seconds, which indicates that depth has no obvious influence on visual attention. In spite of this, there is considerable evidence in the literature that disparity has a time dependent saliency effect [Jansen et al., 2009, Huynh-Thu and Schiatti, 2011, Gautier and Le Meur, 2012]. For further analysis, we divided the observation time of 20 seconds into 5 intervals: 1-4 seconds, 5-8 seconds, 9-12 seconds, 13-16 seconds, and 17-20 seconds. Saccade length, fixation duration, disparity impact, and texture complexity impact was analyzed for each time interval.

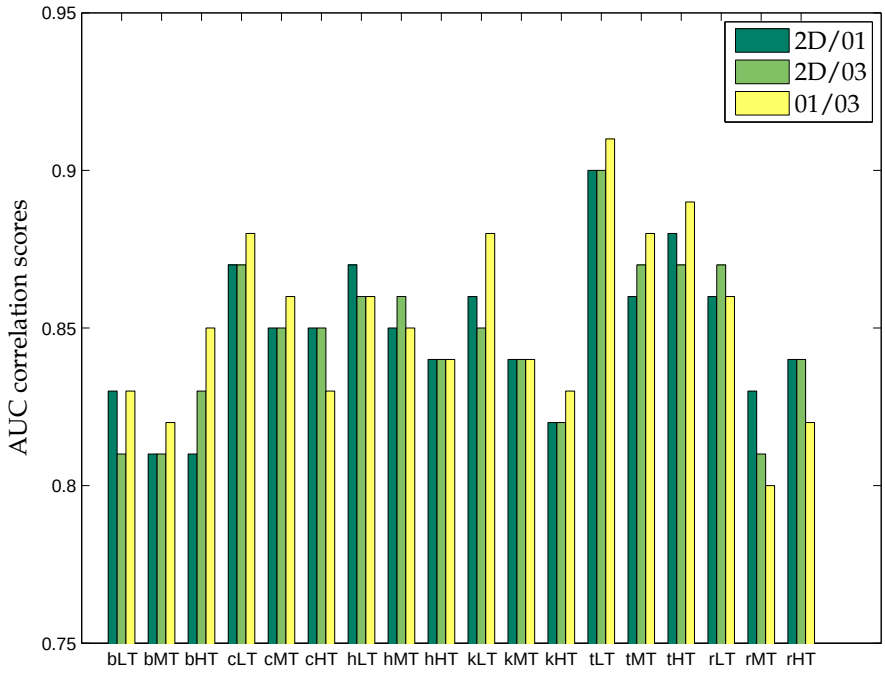


Figure 5.17: AUC correlation values between 2D and 3D DoF=0.1 (2D/01), between 2D and 3D DoF=0.3 (2D/03), and between 3D DoF=0.1 and 3D DoF=0.3 (01/03) saliency maps for a viewing duration of 20s.

Table 5.5: Average AUC and CC values for 20 s between 2D and 3D with DoF=0.1, between 2D and 3D with DoF=0.3, and DoF=0.1 and DoF=0.3

SM	2D/01	2D/03	01/03	min	max
CC	0,87±0,01	0,87±0,02	0,88±0,02	0.76	0.91
AUC	0,85±0,01	0,85±0,01	0,85±0,01	0.8	0.93

5.3.3.3 Saccade length and fixation duration

Each saccade length was measured as the distance between the locations of two fixations in degrees. The results for the average saccade length for each time interval are presented in Figure 5.19. Saccade length has a tendency to shorten over time and with the introduction of disparity. The average decrease of saccade length over time was calculated as the difference between the saccade length of the first time interval and the last one. The difference for 2D: -0.73°, 3D DoF=0.1: -0.49°, 3D DoF=0.3: -0.32°. A paired samples t-test was conducted to compare saccade length for 2D and 3D conditions. There was a significant difference $t(89) = 3.56$, $p < 0.05$, $p = 0.0006$ in the scores

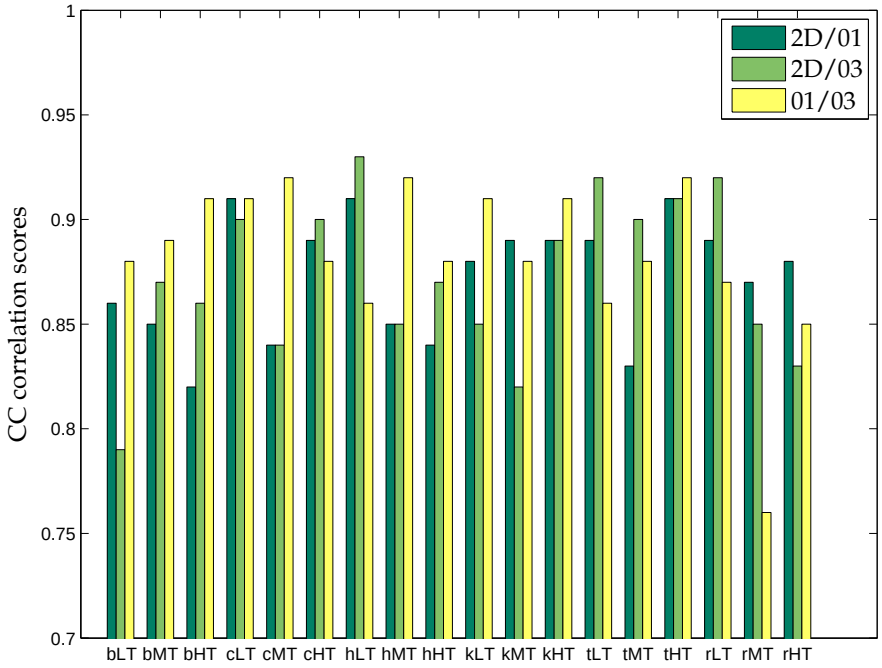


Figure 5.18: CC correlation values between 2D and 3D DoF=0.1 (2D/01), between 2D and 3D DoF=0.3 (2D/03), and between 3D DoF=0.1 and 3D DoF=0.3 (01/03) saliency maps for a viewing duration of 20s.

for saccade length over time for 2D and 3D comfortable conditions. In addition, we found a significant difference of $t(89) = 6.45$, $p < 0.05$, $p = 5.7E - 09$ in the scores for average saccade length over time for 2D and 3D uncomfortable conditions. Finally, a significant difference was detected between 3D comfortable and 3D uncomfortable conditions: $t(89) = 2.24$, $p < 0.05$, $p = 0.027$. Regardless, we did not find any proof from the paired t-tests that texture has an influence on saccade length.

To summarize, the average saccade length decreases constantly over time for all disparities. At the same time, saccade length decreases when bigger disparities are present. These results are in accordance with the study of Jansen et al. [Jansen et al., 2009] that reported that saccade length is reduced in 3D conditions and generally shortens over time.

The introduction of depth into stimuli has an influence on the human visual system since binocular parallax and convergence become involved. Basically, additional time is required in order to verge eyes and fuse delivered images for the perception of depth. Hence it is expected that fixation duration would increase with disparities. However, the statistical analysis of fixation duration showed that there is no relation between the fixation durations and depth levels. The results for the average fixation duration for each time interval are presented in Figure 5.20.

Our results corroborate the findings of neither Huyanh-Thu et al. [Huynh-Thu and Schiatti, 2011] nor Jansen et al. [Jansen et al., 2009] who reported that a disparity cue shortened the median fixation duration. However, we draw attention to the fact that Jansen et al's results were obtained for pink noise and white noise images; there was no effect for natural images. Huyanh-Thu et al. found that for video sequences the average fixation duration was shorter for 3D conditions.

The average increase in fixation duration over time was calculated as the difference

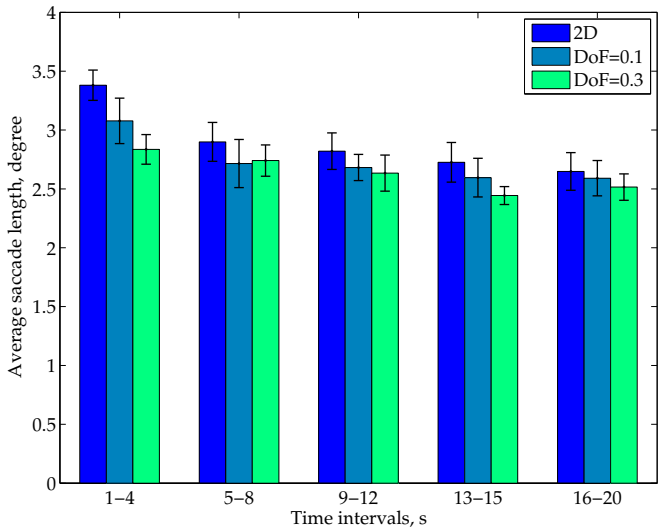


Figure 5.19: Influence of depth on average saccade length over time

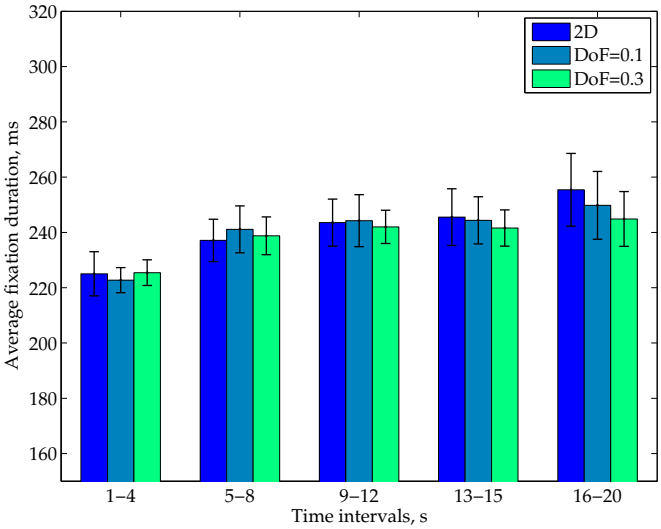


Figure 5.20: Influence of depth on average fixation duration over time.

between the fixation duration of the last time interval and the first one. The difference was +30.4 ms for 2D, +27.01 ms for 3D comfortable, and +19.41 ms for 3D uncomfortable. A paired samples t-test was conducted to compare fixation durations for all corresponding disparities for the first and the last time interval. There was a significant difference for all conditions. 2D: $t(17) = -4.92$, $p < 0.05$, $p = 0.0001$; DoF=0.1: $t(17) = -4.48$, $p < 0.05$, $p = 0.0003$; DoF=0.3: $t(17) = -3.7$, $p < 0.05$, $p = 0.002$. Thus, in Figure 5.20, it can be noted that the fixation duration tended to increase over time. This conclusion is supported by the data of Jansen et al.[Jansen et al., 2009].

With a paired samples t-test, there was no significant influence of texture complexity on fixation duration.

5.3.3.4 Influence of depth on visual attention

In order to assess the effect of disparity on visual attention, the AUC metric was calculated between pairs of saliency maps for 2D, 3D DoF=0.1, and 3D DoF=0.3 over time. The average results are presented in Figure 5.21. All scores are presented in Table A.3.

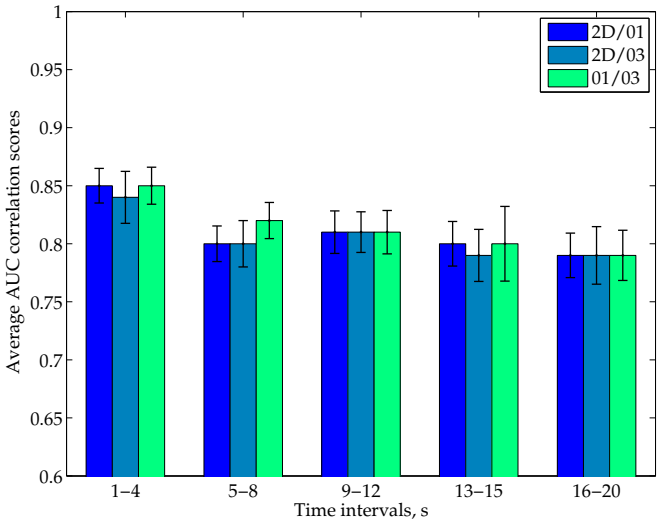


Figure 5.21: Influence of depth on average fixation duration over time: AUC scores.

At each time interval, the "Tea" scene had the maximum AUC value (see Fig. 5.13). This result reflects the particularity of the scene: this scene contained the least number of objects, they were located in the center of the scene at the foreground, and there was no pronounced background (the tablecloth was moving away with a depth to infinity). The depth distortion was not as noticeable for DoF=0.3 as for the other scenes because the objects were at the foreground(see Fig. A.11.d). The "Kitchen" scene (1-4s, 9-12s, 17-20s) and the "Room" scene (5-8s, 13-16s) had the minimum values. This result is also not surprising since these scenes contained a lot of different objects, along with a depth distortion for DoF=0.3 that was very pronounced.

A paired samples t-test was conducted to compare the AUC values for 2D and 3D conditions. There was a significant difference in the scores for 2D/01 and 2D/03 conditions, $t(17) = 2.11$, $p < 0.05$, $p = 0.004$ for the period of time 1-4 seconds. For the rest of the time periods, the differences were insignificant. For the period of time 5-8 seconds, there was a significant difference in the scores for 2D/01 and 01/03 conditions, $t(17) = -3.51$, $p < 0.05$, $p = 0.003$ and in the scores for 2D/03 and 01/03 conditions, $t(17) = -2.48$, $p < 0.05$, $p = 0.024$. The differences were insignificant for the rest of the time periods for both conditions.

The way images are observed during the first time interval 1-4 s is most similar to the observation of videos where the frames change one after the other. Basically, there is no time to explore all the parts of the complex scenes and attention is attracted by salient regions. Besides, at this time period the way images are observed can be influenced by the central bias [Tatler, 2007, Judd et al., 2009, Gautier and Le Meur, 2012]. The demonstrated values of the AUC reflect the fact that the saliency maps are very similar for all the scenes. The minimum value during the first 4 seconds is 0.73 for one of the most complex scenes ("Kitchen") while the average value is 0.85. In our opinion, the high AUC values could indicate that disparity plays a very subtle role in the selection of

salient features of a scene. This hypothesis is supported by the results of our previous experiment in Section 5.2.3.1, but it is not fully supported by the CC values (min 0.47; avg 0.74) which are quite high and are presented in Figure 5.22 and Table A.4 for comparison. Thereby, there is the lack of a standard (a method or a metric), which allows for the evaluation of the depth effect and then a comparison of the results within different studies.

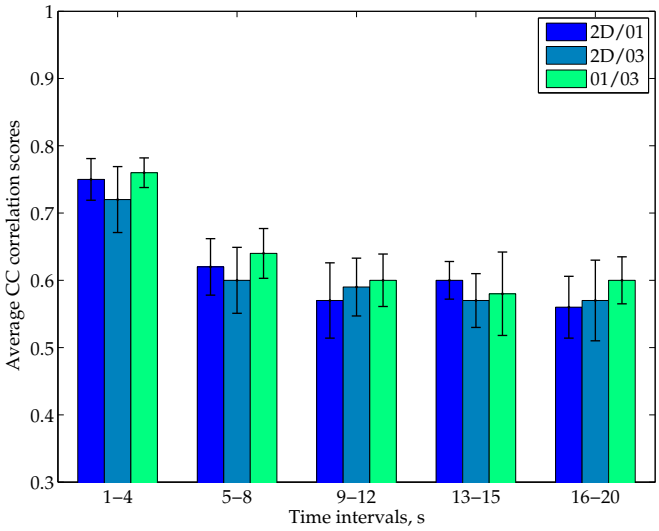


Figure 5.22: Influence of depth on average fixation duration over time: CC scores.

It is likely that during the first time interval all of the observers are attracted by the most salient features of a scene, but in later time periods, the way the image is observed may differ from one observer to another. This hypothesis is supported by the AUC and CC values. The highest correlation values in both cases were obtained for the first time interval. The paired t-test showed that the difference between the first time interval (1-4 s) and the second time interval (5-8 s) is significant for CC values as well as for the AUC. The results of the paired t-test are presented in Table 5.6.

Table 5.6: Paired t-test: difference in AUC and CC scores for the first (1-4 s) and the second (4-8 s) time interval; $t(17)$, $p < 0.05$.

	p, AUC	p, CC
2D/01	0.0001	2.32E-05
2D/03	0.046	0.0017
01/03	0.012	6.7E-06

The results presented in Figure 5.21 and in Figure 5.22 are in accordance with the study of Tatler et al. [Tatler et al., 2005], who found that the consistency between visual fixations of different subjects is high just after the stimulus is displayed but progressively decreases over time. It is likely that just after the stimulus is displayed, our attention is mostly controlled by bottom-up mechanisms, whereas top-down mechanisms become more influential after several seconds of viewing. The second factor is content dependent.

For a viewing duration of 20 s, the average AUC and CC values were presented in Table 5.5 as well as in Figures 5.17- 5.18. These results indicate that for such long durations of time, the depth levels did not have an obvious influence on saliency maps –

the AUC and CC values are very high and similar for every condition. Values for every scene for a viewing duration of 20 seconds are presented in Table A.1.

As in the work of Ramasamy et al., one of our scenes contained a deep hallway [Ramasamy et al., 2009]. For the first 4 seconds, our results correlate with their work: the gaze points were more spread out in 2D conditions and more concentrated at the far end in 3D conditions independent of DoF. But after 4 seconds, the gaze points became more spread out and the heat maps looked similar to non-stereoscopic conditions. For scenes like "Cartoon" or "Kitchen", a similar visual behavior has been noticed but it was less pronounced. A possible reason is the presence of a greater number of objects in the scenes. Nevertheless, clear evidence of this tendency was not revealed for the rest of the scenes. Thus, it is possible that the saliency of the objects in the case of the "Bathroom" and "Room" scenes plays a more important role than the presence of depth. To summarize, we did not observe any particular relation between the depth and the spread of the gaze points. Nevertheless, we believe that an analysis of the heat maps is not fully reliable because a non-normalized color scale between scenes hampers their comparison.

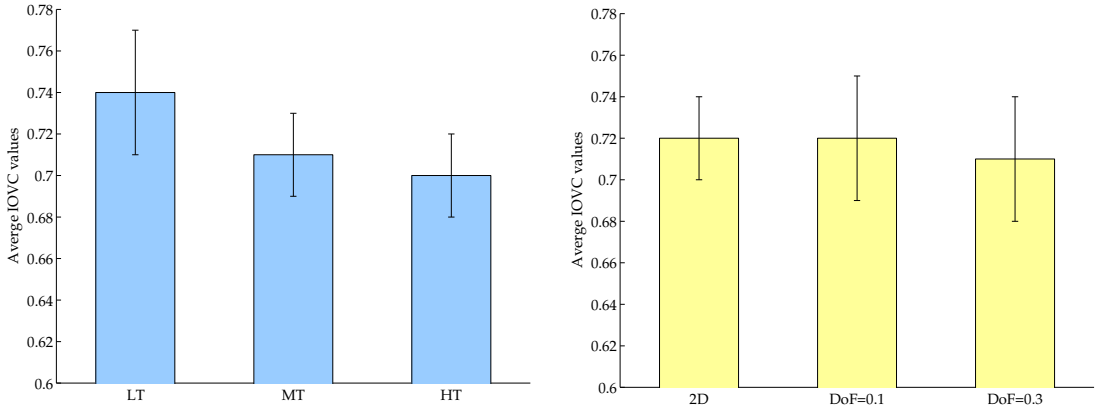
5.3.3.5 Influence of texture on visual attention

This section investigates the impact of texture complexity on visual attention. Thus, the inter-observer visual congruency (IOVC) was calculated, which reflects the visual dispersion between observers or the consistency of overt attention (eye movement) while observers are watching the same visual scene [Le Meur et al., 2011]. The method used for the computation of IOVC values is presented in Section A.3.

As was already mentioned in the previous section, visual attention is controlled by low-level visual features most probably just after each stimulus is displayed. After several seconds top-down processes begin, which is content dependent. As a consequence, a stimulus composed of the salient areas would presumably attract our visual attention, leading to high congruency. The presence of particular features, such as human faces, people, or animals, tends to increase the consistency between observers. On the other hand, congruency tends to decrease with scene complexity.

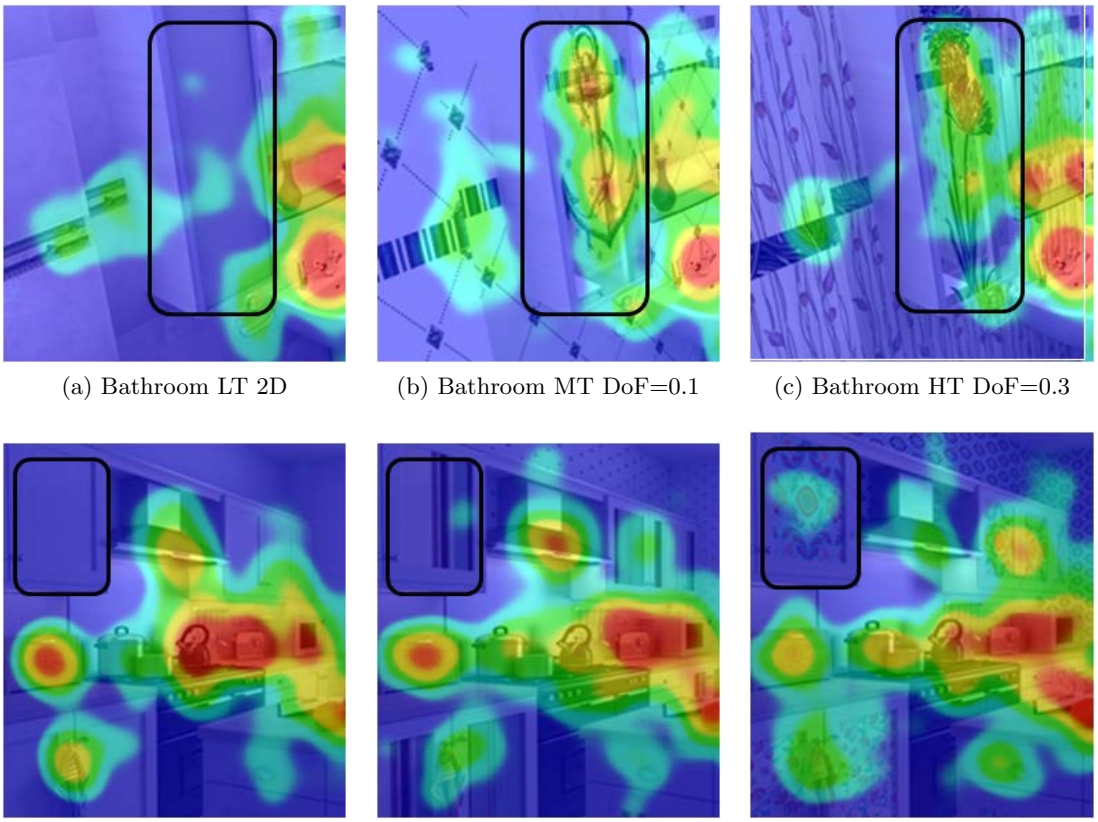
The calculation of IOVC was done with all the fixation data for a viewing duration of 20 s. The computed results are presented in Table A.2 and the average IOVC values for texture complexity are presented in Figure 5.23.a and for depth levels in Figure 5.23.b. A paired samples t-test was conducted to compare inter-observer visual congruency for different texture complexities. There was a significant difference $t(17) = 1.74$, $p < 0.05$, $p = 0.002$ in the scores for congruency for LT and MT complexity. A significant difference was found in the scores for congruency for LT and HT complexity: $t(17) = 1.74$, $p < 0.05$, $p = 0.004$. Nevertheless, no significant difference was detected between MT and HT ($p < 0.05$, $p = 0.08$). For each scene high and medium texture complexities were selected by experts without using any metric. This could be a possible reason that similar results were obtained for both medium and high texture complexities.

In Figure 5.24, fragments of the scene "Kitchen" and "Bathroom" are presented with three texture complexities. It can be seen that the cupboard in the front of the "Kitchen" scene (Fig. 5.24.a) does not attract attention, whereas when there is some pattern on top of the cupboard, its doors become salient (Fig. 5.24.b-c). A similar situation happened with the "Bathroom" scene (Fig. 5.24.d-f). Based on the analysis of the heat maps and the paired t-test of IOVC values, we can deduce that the selected areas of interest depend on the texture. Therefore, the resulting saliency maps might differ.



(a) Influence of texture complexity on IOVC (b) Influence of binocular disparity on IOVC

Figure 5.23: Average IOVC values.



(a) Bathroom LT 2D (b) Bathroom MT DoF=0.1 (c) Bathroom HT DoF=0.3

(d) Kitchen LT 2D (e) Kitchen MT DoF=0.1 (f) Kitchen HT DoF=0.3

Figure 5.24: The first row shows fragments of heat maps for "Bathroom" scene with low, medium, and high texture from left to right. The second row fragments of heat maps are from "Kitchen" with low, medium, and high texture from left to right.

5.3.3.6 Discussion and conclusions

The goal of this study was to estimate the influence of depth, comfort/discomfort, and texture complexity on visual attention. This study was performed using complex stimuli

to validate the conclusions of the previous experiment in Section 5.2. One feature of the study was that all stereoscopic content was with uncrossed disparity, i.e., all objects were behind the display plane. It was discovered that:

- **Eye-movements**: Average saccade length decreased constantly over time for all disparities. At the same time, average saccade length decreased with bigger disparities as well. These results are in accordance with the study of Jansen et al. [Jansen et al., 2009]. The analysis of fixation duration showed that there is no relation between the fixation durations and disparities. Our results are in opposition to former studies, which reported that disparity cues shortened the median fixation duration. For all the scenes, we found that the gaze points were denser and centered in the middle of a scene during the first 4 seconds, but were spread out over the entire scene for the other time intervals. We did not find any strong evidence that depth has an influence on the spread of gaze points. This is in accordance with the conclusion of Huynh-Thu et al.[Huynh-Thu and Schiatti, 2011].
- **Disparity:** No strong evidence was found that indicated the influence of uncrossed disparity. Based on the AUC and CC scores, it can be assumed that the saliency of an object plays a more important role than the depth. The AUC and CC scores remained very high even when an analysis was performed for the first time interval (1-4 s), which implies that only bottom-up mechanisms of visual intention were involved. On the other hand, a visual analysis of the heat maps showed that there is an influence of disparity for some scenes but it is not possible to conclude whether this is significant or not.
- **Discomfort:** All paired t-tests showed that the differences between comfortable (DoF =0.1) and uncomfortable (DoF=0.3) conditions were not significant. Among the reasons that might explain this result is the test methodology. The entire test for each observer lasted 2 min 40 s, and the uncomfortable condition only lasted for 40 seconds. Thus, the visual system was not stressed. After the experiment several observers reported that they experienced discomfort in some cases, but they were still looking at the background because they had never experienced stereoscopic depth distortion (see Ds coefficients for DoF=0.3 Fig. A.7.d- A.12.d) and were curious to observe such content. Therefore, it would be necessary to stress visual systems before the test and then repeat the experiment with stimuli for DoF=0.3. Another possible reason we did not find that the disparities have any pronounced influence is the absence of a method that is sensitive to disparities, which could reveal differences in saliency maps for different DoF. We believe that it is important to define a method or a metric which would allow for a comparison of results between different studies that investigate the effect of disparity on visual attention.
- **Texture:** significantly higher inter-observer visual congruency in cases of stimuli with low texture complexity in comparison with medium and high texture complexities.

5.4 Experiment 3: complex stimuli with crossed disparity objects

The goal of this experiment is to continue the research described in Section 5.3. Visual attention in 3D was studied using complex stimuli with only uncrossed disparity. In order

to complete the previous study a new experiment was designed using 2D, 3D comfortable, and 3D uncomfortable still images containing objects with crossed disparity.

5.4.1 Stimuli generation

The key point of the experiment is to present stimuli with an object(s) in front of a display with a controlled amount of depth to the observer. The depth range is controlled by changing the 3D camera baseline and the convergence distance. All scenes were designed and rendered using Blender software, which allows the measurement of the foreground and background distances of a scene and the accurate control of stereoscopic camera parameters. Four different scenes were selected for the experiment: "Cartoon", "Hall", "Pigs", and "Table" [Monteiro et al., 2013]. Figure 5.25 illustrates examples of the generated scenes. All objects with crossed disparity were selected in a way to avoid the window violation effect. For example, in Figure 5.25.b, the lamp stand was behind the display plane, while the lampshade was coming out of the screen.

(a) Cartoon.

(b) Hall.

(c) Pigs.

(d) Table.

Figure 5.25: Stimuli with crossed disparity objects.

The image parameters are presented in Table 5.7, where f is the camera focal length, b - the baseline distance, $dCon$ - convergence distance, fg – foreground distance, e.g. the distance from the camera to the closest object, bg – background distance, e.g. the distance from the camera to the farthest object, roi - region of interest, and DoF - depth of focus. DR is the disparity range of a scene in the visualization space, which consists of maximum crossed and uncrossed disparity in mm on the screen used for the experiment. The camera parameters were selected to correspond to DoF=±0.1 diopters for the comfortable condition and DoF=±0.3 diopters for the uncomfortable condition. Therefore, the reconstructed amount of depth with the same DoF were the same for all the stimuli. The total amount of perceived depth in the uncomfortable condition reaches

DoF=0.6 (0.3 for the object with crossed disparity and 0.3 for the background).

Table 5.7: Scene parameters.

						DoF=0.1		DoF=0.3	
Scene	f,mm	dCon,m	fg,m	bg,m	roi,m	b,mm	DR	b,mm	DR
Cartoon	35	8	5.2	17.6	13.5	220	[-15;15]	660	[-46;46]
Hall	28	8	5.6	14.2	9	342	[-15;15]	1020	[-46;46]
Pigs	35	2.2	1.4	5.5	4	53	[-15;15]	162	[-47;45]
Table	35	5.45	4.3	7.3	5	325	[-16;15]	1000	[-45;44]

Images were rendered with a resolution of 1920×1080 using a virtual camera with a sensor size of $32mm \times 16mm$. Multisampling with 8 sample anti-aliasing was used to smooth the edges. The blur effect was disabled to guarantee the sharpness of the scenes. Shooting was performed with a parallel-rig, using HIT to create the desired disparity. In order to avoid a black border after the post-production shift, extended borders were rendered for every image.

A detailed analysis of the relationships between the camera space and the visualization space and depth distortions for the comfortable condition and the uncomfortable condition is given in Annex A Figures A.13- A.16. The space outside ZoC is marked in light gray and the region of interest as a magenta line.

The main focus of this study is the influence of depth with crossed disparity on visual attention, so texture complexity was not taken into account. In total, 12 still images were generated (4 scenes $\times$ 3 depth levels). Since it was important to prevent observers from memorizing the stimuli and hence using top-down visual mechanisms, 3 sets were arranged containing 4 images with different contents and different depth levels.

5.4.2 Experimental set-up and methodology

The same experimental set-up and methodology was used as in Section 5.2.2. The training was done in stereoscopic mode using three images with three levels of depth: DoF=0 diopters (2D), DoF=0.1 diopters, DoF=0.3 diopters. Each image was presented for 20 seconds and separated from the subsequent one by displaying a gray screen for 5 seconds. The duration of the training was 1 minute 20 seconds. The images were different from those used in the test. The training phase was designed to familiarize observers with the test conditions.

During the test, only one of the three sets of images was displayed. The duration of the test was 1 minute 40 seconds. 51 people (36 males and 15 females from 22 to 52 years old) participated in the test. So each image was observed by 17 subjects.

5.4.3 Eye-tracking data analysis

In this section we analyze eye-tracking data and study whether the introduction of an object with crossed disparity had an effect on basic eye movement properties. The entire fixation data collected with the Tobii eye-tracker was used for analysis. All observers that could not complete the calibration process using the calibration chart were excluded before the test. In order to analyze the gaze behavior of observers, saliency and heat maps as well as fixation durations and the length of saccades were computed [Le Meur and Baccino, 2012, Le Meur, 2012]. During the experiment, images were sepa-

Table 5.8: AUC and CC correlation values between 2D and 3D DoF=0.1 (2D/01) saliency maps; between 2D and 3D DoF=0.3 (2D/03) saliency maps; between 3D DoF=0.1 and 3D DoF=0.3 (01/03) saliency maps.

	AUC			CC		
	2D/01	2D/03	01/03	2D/01	2D/03	01/03
Cartoon	0.82	0.87	0.84	0.8	**0.94**	0.84
Hall	0.84	0.85	0.83	**0.89**	0.87	0.87
Pigs	0.8	0.84	0.84	0.84	0.87	0.85
Table	**0.88**	**0.9**	**0.88**	0.87	0.87	**0.93**

rated by a gray slide without a fixation cross in the center, which is why the first fixation of each stimulus was not discarded.

5.4.3.1 Qualitative analysis based on heat maps

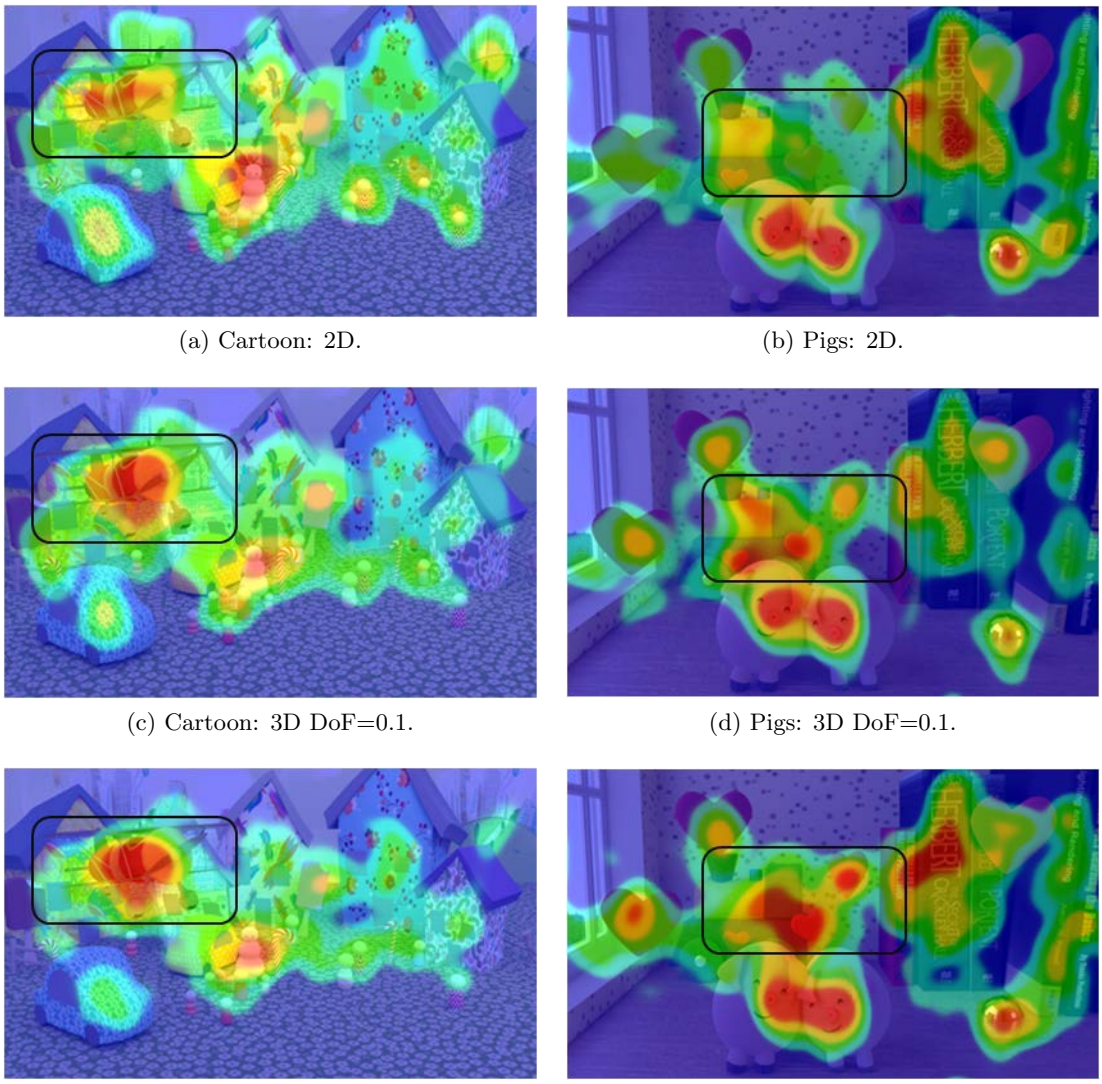
Heat maps representing the fixated areas of a stimulus were used to compare the gaze patterns of all observers. This method has been used in various studies [Huynh-Thu and Schiatti, 2011, Hakkinen et al., 2010, Ramasamy et al., 2009] since it allows the gaze behavior of an entire group of observers to be quickly and conveniently visualized. It should be noted that the normalization process is done for each heat map independently. As a consequence, it is difficult to precisely compare heat maps for different scenes.

For example, several heat maps are shown in Figure 5.26 for the "Cartoon" and "Pigs" scenes. Each image on this figure shows the heat map corresponding to a viewing duration of 20 seconds. In stereoscopic condition, the airplane in Figure 5.26.a attracts attention independent of the cause of discomfort. In the 2D case, the most fixations (the large red spot) are on the snowman, while in 3D, the most fixations are on the airplane. Similar behavior is observed in Figure 5.26.b, where small hearts, which pop-out of the screen, became the main region of interest with depth and received the most fixations. This behavior was exhibited in the other two scenes as well.

5.4.3.2 Quantitative analysis

After computing the saliency maps which represent the density of fixations for an entire image for each scene, their differences were found by calculating the correlations between pairs. As metrics, the Pearson linear correlation coefficient (CC) and Area Under Curve (AUC) were used. In the case of AUC, a higher value means a better correlation: a value of 1.0 indicates a perfect performance, while a value of 0.5 demonstrates a random performance [Le Meur and Baccino, 2012]. The results for the AUC and CC metrics are presented in Table 5.8. The highest result for each column is marked in bold.

All AUC and CC values presented in Table 5.8 are very high. High correlation indicates that there is no strong difference between saliency maps. This implies that visual attention is not affected by different disparities, which is in contradiction with our qualitative results. A viewing duration of 20 seconds is sufficient time for an observer to investigate every object in a still image. Consequently, saliency maps differ mainly in the density of fixations, which cannot be detected by AUC and CC metrics. Hence the difference between 2D and 3D conditions could only be discovered by comparing the number of fixations on the objects.



(a) Cartoon: 2D.

(b) Pigs: 2D.

(c) Cartoon: 3D DoF=0.1.

(d) Pigs: 3D DoF=0.1.

(e) Cartoon: 3D DoF=0.3.

(f) Pigs: 3D DoF=0.3.

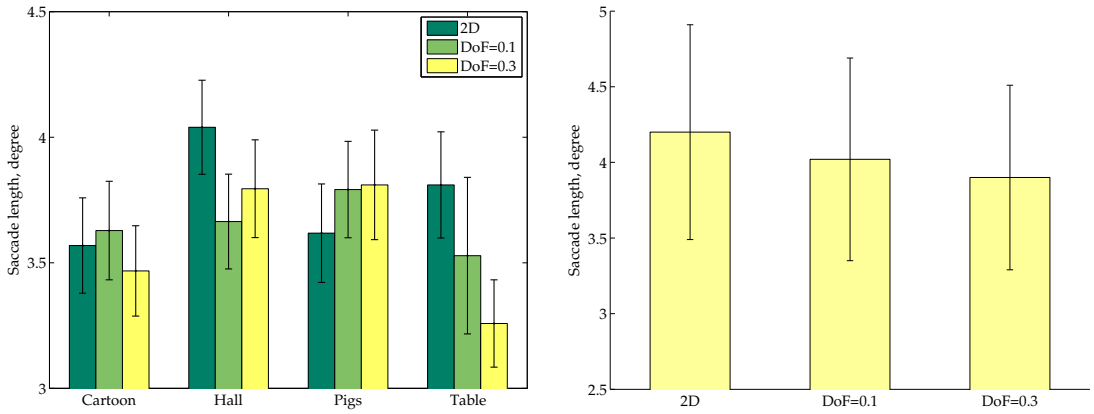
Figure 5.26: Heat maps for the "Cartoon" and "Pigs" scenes with various disparities.

Huynh-Thu et al. performed a quantitative comparison between 2D and 3D using AUC and CC metrics [Huynh-Thu and Schiatti, 2011]. They have found that the differences between saliency maps depended on the content whereas our data presented in Table 5.8 demonstrate a very slight difference between different scenes. One possible reason for such a contradiction is that in the Huynh-Thu et al. experiment, videos of different duration (from 8 to 143 seconds) were used, while the same viewing duration of 20 seconds was used in our experiment. This amount of time may be sufficient to look at all the objects in a still scene, which should not be the case in an experiment with videos.

5.4.3.3 Saccade length and fixation duration

Figure 5.27.a presents the saccade length for every image. There is no clear tendency for saccade length, which seems to be content dependent. With a paired samples t-test

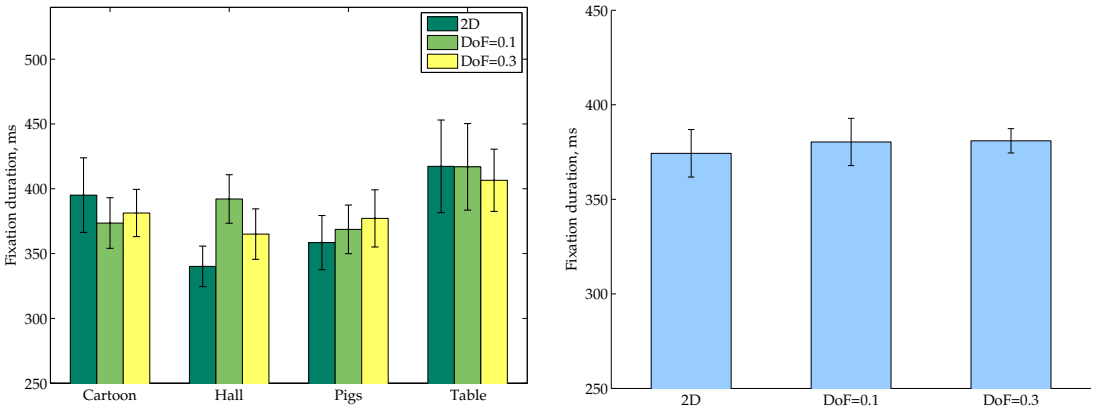
no significant differences were found between saccade length for 2D and 3D conditions. We believe that there are not enough observations to prove a statistical significance in our case. Figure 5.27.b presents the average saccade length for all the scenes. It does not corroborate the results from Jansen et al. [Jansen et al., 2009] or from Huynh-Thu et al. [Huynh-Thu and Schiatti, 2011], who found that saccades in 3D were shorter and faster.



(a) Length of saccade for every image. (b) Average saccade length.

Figure 5.27: Influence of depth on (a) length of saccade for every image, (b) average saccade length. Error bars represent a confidence interval of 95%.

Statistical analysis of fixation durations showed that there is no relation between the fixation durations and the depth levels (see Fig. 5.28.a). With a paired samples t-test, it was found that depth had no significant influence on fixation duration. Figure 5.28.b presents the average fixation duration for all the scenes. Our results do not corroborate the findings of Huyanh-Thu et al. that fixation durations were longer in 2D in videos [Huynh-Thu and Schiatti, 2011] nor Jansen et al. that fixation durations were longer in 3D for still images with uncrossed disparities [Jansen et al., 2009].



(a) Fixation duration for every image. (b) Average fixation duration.

Figure 5.28: Influence of depth on (a) fixation duration for every image, (b) average fixation duration. Error bars represent 95% confidence interval.

5.4.3.4 Discussion and conclusions

To expand on the previous work in Section 5.3, an eye-tracking experiment was designed, where the stimuli contained object(s) with crossed disparity. The difference between visual strategies when observers watch 3D images with crossed disparity (in comfortable and uncomfortable conditions) and 2D images was examined. It was discovered that:

- **Eye movements:** There was no significant difference between 2D and 3D conditions for average saccade length and fixation durations.
- **Disparity and discomfort:** Objects located in front of the display plane are more salient than objects with uncrossed disparity or 2D, even if observers experience discomfort from excessive disparity. Some of subjects reported that visual effort was required to fuse objects in front of the display plane in uncomfortable conditions.

In both experiments with complex stimuli the AUC and CC metrics have demonstrated very high correlations between saliency maps for stereoscopic and non-stereoscopic images. We believe that these metrics are not adjusted to compare gaze in depth for our case since the observation time was sufficient for an observer to investigate every object in a still image. Consequently, saliency maps differed mainly in the density of fixations, which cannot be detected by AUC and CC metrics. Hence the difference between 2D and 3D conditions could only be discovered by comparing the number of fixations on the objects. Therefore, a new metric is proposed to compare gaze points in depth in the next section.

5.5 Weighted Depth Saliency Metric proposal for comparison of visual attention

In the previous sections for the comparison of saliency maps for 2D and 3D conditions, Pearson linear correlation coefficient (CC) and Area Under Curve (AUC) were used. These metrics were designed to assess the degree of similarity between a predicted saliency map computed by a visual attention model and the ground-truth saliency map obtained from the fixation data recorded with an eye-tracker. The degree of similarity is estimated relying on saliency maps, which are two dimensional. Therefore, a comparison of the fixation data for 3D conditions with a different depth level can be quite difficult.

Therefore, being guided by existing metrics with such a principle, it is not possible to conclude whether observers are looking closer or farther in terms of depth. In order to provide a comparison, we propose a new depth metric which takes into account the weighted saliency map and the real depth map. The usage of a weighted saliency map allows a fair comparison between different cases of depth. The depth maps allow the saliency map to be segmented on the basis of the semantic depth information. If needed, visual attention can be computed for different depth layers.

5.5.1 Algorithm

The main aim of our metric is to analyse differently visual attention using the depth information and hence compare saliency when the scene is displayed in different S3D conditions. For example, fixation points were collected for the “Cartoon” scene illustrated in Figure 5.29, which was displayed in 3 different conditions: 2D, 3D with DoF=0.1, and 3D with DoF=0.3.

Figure 5.29: "Cartoon" scene with an airplane in front of the display plane.

Visual attention between these conditions can be compared using the following algorithm, which consists of several stages:

1. Compute saliency map (SM) from the fixation data for all conditions to compare visual attention (see Fig. 5.30).

(a) SM 2D. (b) SM DoF=0.1. (c) SM DoF=0.3.

Figure 5.30: Saliency maps for the "Cartoon" scene displayed with different depth levels.

2. For a fair comparison between saliency maps the weights for every condition k should be computed using the equation 5.1:

$$\omega^k = \frac{min_{p=1..n} \sum_{i=1}^{r} \sum_{j=1}^{c} SM^p(i,j)}{\sum_{i=1}^{r} \sum_{j=1}^{c} SM^k(i,j)} \tag{5.1}$$

where n is the number of conditions to compare, k is n^{th} condition, r, c - number of rows and columns in a SM, i, j - spatial location of pixels e.g. row and column indexes respectively, ω^k - weight coefficient for k^{th} SM.

The computed weights should be applied to corresponding SM using the following equation 5.2. As the result of this operation, all saliency maps will be weighted and have the same sum of all pixels as illustrated in Figure 5.31.

$$WSM^k = \omega^k \cdot SM^k \tag{5.2}$$

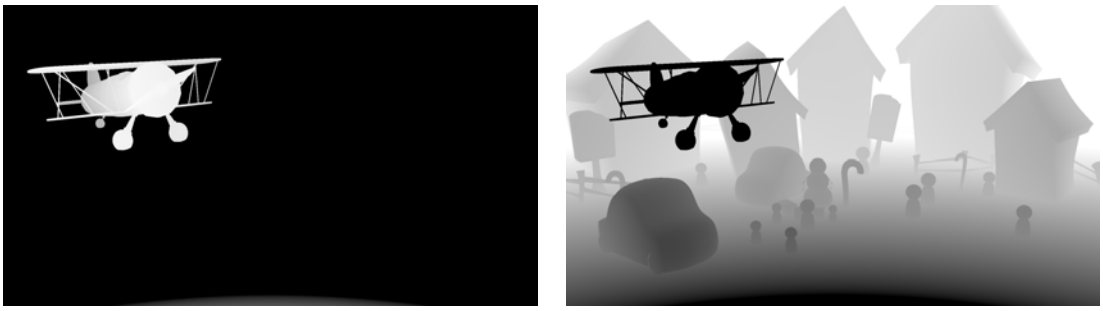
Computed coefficients are required to equate saliency for all the conditions. Therefore, weighted saliency maps (WSM) imply that amount of visual attention is the

same for every condition and only the distribution of gaze may differ. This allows a fair comparison even if a number of observers was not the same for every condition.

(a) WSM 2D ($\omega^1 = 0.43$). (b) WSM DoF=0.1 ($\omega^2 = 0.68$). (c) WSM DoF=0.3 ($\omega^3 = 1$).

Figure 5.31: Weighted saliency maps for the "Cartoon" scene for all the conditions.

3. The depth map of a scene can be used to compare distributions of visual fixations in depth between different conditions. The depth map can be segmented into layers using the semantic information of the scene. Then it would be possible to answer whether the displayed amount of depth has influenced on the distribution of visual fixations by comparing corresponding layers in different conditions. If the scene contains crossed disparity, its depth map can be segmented to two depth layers representing the object in front of the display plane and behind it as demonstrated in Figure 5.32.a and Figure 5.32.b respectively.



(a) Depth map for crossed disparities. (b) Depth map for uncrossed disparities.

Figure 5.32: Result of the "Cartoon" scene segmentation to crossed and uncrossed disparities.

The depth map with uncrossed disparities can be segmented to more layers using segmenting operator g:

$$g(d(i,j)) = \begin{cases} d & \text{if } L_{beg} < d \leq L_{end} \\ 0 & \text{otherwise} \end{cases}$$

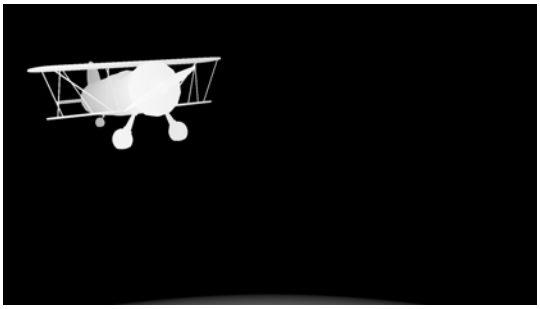
where d - pixel value of the depth map $DM(i,j)$, L_{beg} and L_{end} pixel values at the start and end of the depth layer, respectively.

Therefore, using the segmenting operator $g(d)$, the depth map can be segmented to depth layers following equation 5.3

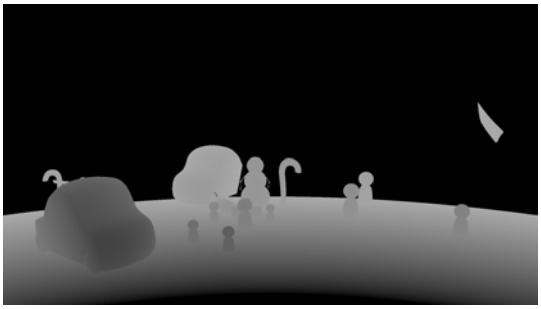
$$DM^L(i,j) = g(DM(i,j)) \tag{5.3}$$

where DM^L is L^{th} depth layer of the depth map DM, g - the segmenting operator, such that $DM = \cup_{L=1..m} DM^L$, m - number of layers, $DM_{Li} \cap DM_{Lj} = \varnothing$.

An example is illustrated in Figure 5.33, where the "Cartoon" scene is segmented into four different depth layers (a) objects in front of the display plane, (b) foreground objects, e.g. objects close to the display plane, and (c) objects in the region of interest, and (d) background objects.



(a) Depth layer 1 (-255:0).



(b) Depth layer 2 (0:170).

(c) Depth layer 3 (170:240).

(d) Depth layer 4 (240:255).

Figure 5.33: Depth layer L (L_{beg}:L_{end}) of the "Cartoon" scene.

4. Finally, our Weighted Saliency Depth Metric ($WSDM$) represents the amount of saliency for every depth layer. It is calculated as an average of weighed saliency map pixels for corresponding non-zero pixels in depth layer following equation 5.4:

$$WSDM^L = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c} \upsilon(DM^L(i,j)) \times WSM(i,j)}{\sum_{i=1}^{r}\sum_{j=1}^{c} \upsilon(DM^L(i,j))} \tag{5.4}$$

where

$$\upsilon(x) = \begin{cases} 1 & \text{if } x \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

For example, the depth metrics was computed for the scene in Figure 5.29 using the weighted saliency maps in Figure 5.31 and depth layers in Figure 5.33. The computed results are presented in Table 5.9 for three visualization conditions.

The obtained values represent the average saliency intensity in each depth layer. Results for different conditions can be compared and an accurate conclusion can be

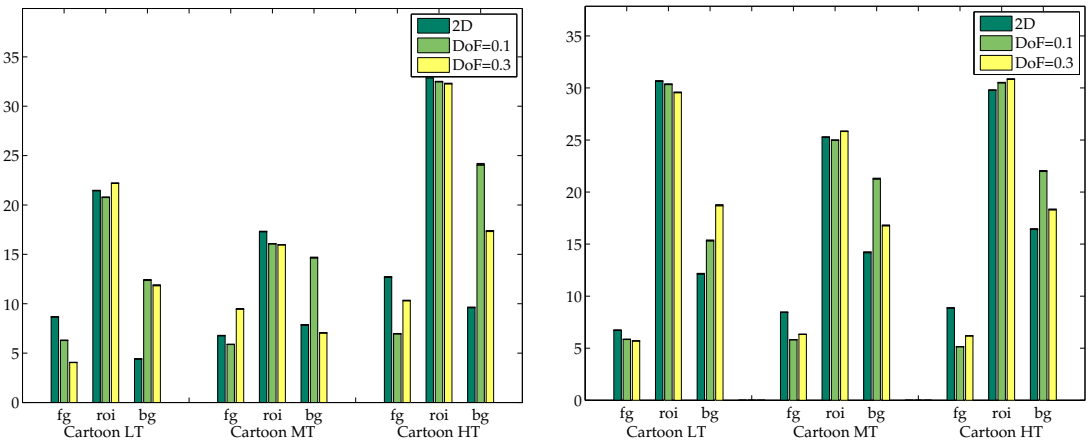
Table 5.9: Depth metric computed for three different visualization conditions and four depth layers.

Cartoon $WSDM^L$	2D	DoF=0.1	DoF=0.3
Depth layer 1 (-255:0)	15.47 ±0.09	48 ±0.18	53.11 ±0.24
Depth layer 2 (0:170)	6.62 ±0.03	3.48 ±0.02	4.29 ±0.02
Depth layer 3 (170:240)	8.72 ±0.02	6.99 ±0.02	5.74 ±0.03
Depth layer 4 (240:255)	2.93 ±0.03	2.79 ±0.03	2.3 ±0.04

made about whether observers looked at a closer location or farther location in terms of depth. Thus, an additional conclusion can be made regarding the influence of binocular features on the saliency of the objects. According to Table 5.9, the interest in an airplane coming out of the screen in 3D increased three times in comparison with 2D. Despite the discomfort caused by excessive disparities, interest in the airplane was higher for DoF=0.3 than DoF=0.1.

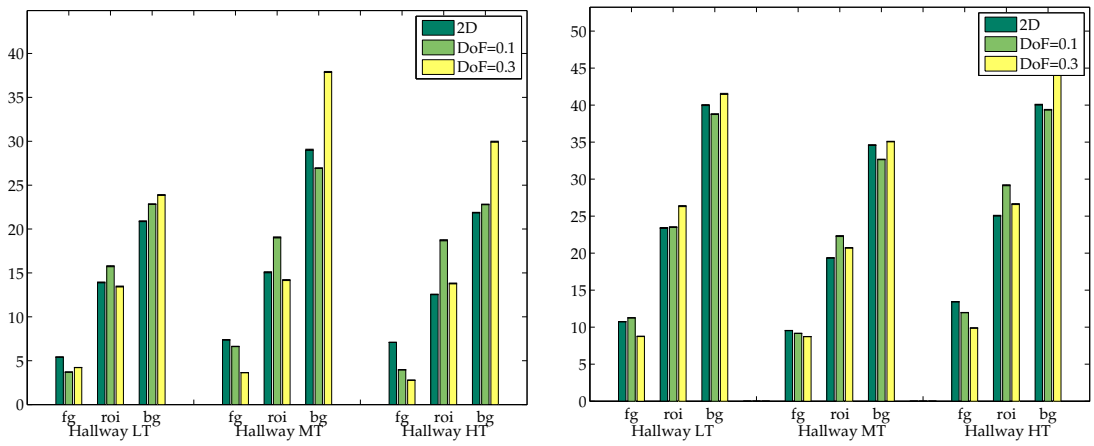
5.5.2 Results

The WSDM was computed using the stimuli from Experiment 2 (Section 5.3) with only uncrossed disparities. The results for the "Cartoon", "Hall", and "Tea" scenes are presented in Figures 5.34- 5.36. Other scenes are presented in Annex A in Figures A.17-A.19. The saliency maps were weighted for three texture complexities (LT, MT, HT) and three depth conditions (2D, DoF=0.1, DoF=0.3). In another words, the same segmented depth layer can be compared for different textures and depth levels.



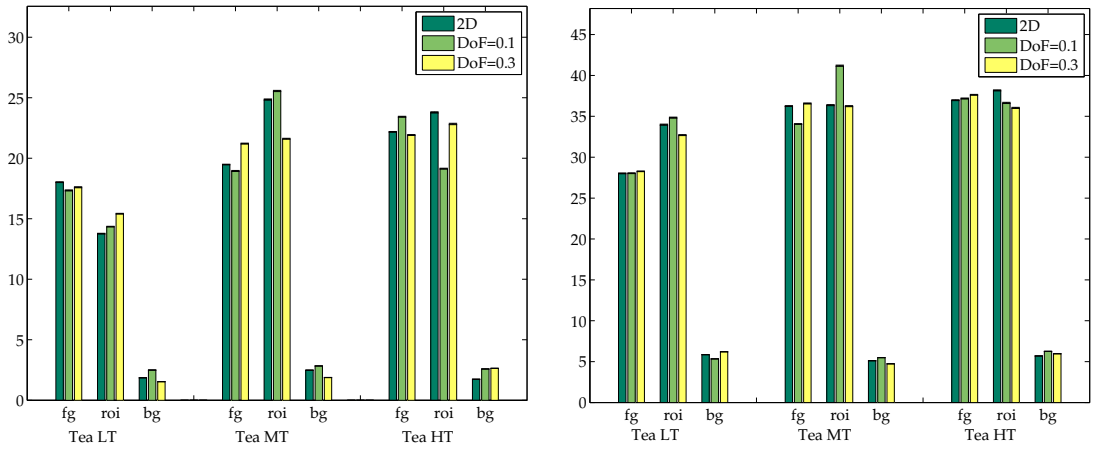
(a) 4s SM (b) 20s SM

Figure 5.34: WSDM for "Cartoon" scene ($fg : [0; 75], roi : (75; 200], bg : (200; 255]$)



(a) 4s SM (b) 20s SM

Figure 5.35: WSDM for "Hallway" scene (fg:$[0;140], roi : (140;200], bg : (200;255]$)



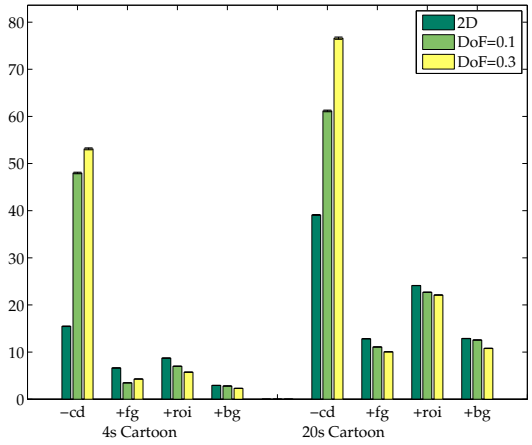
(a) 4s SM (b) 20s SM

Figure 5.36: WSDM for "Tea" scene ($fg : [0;43], roi : (43;65], bg : (65;255]$)

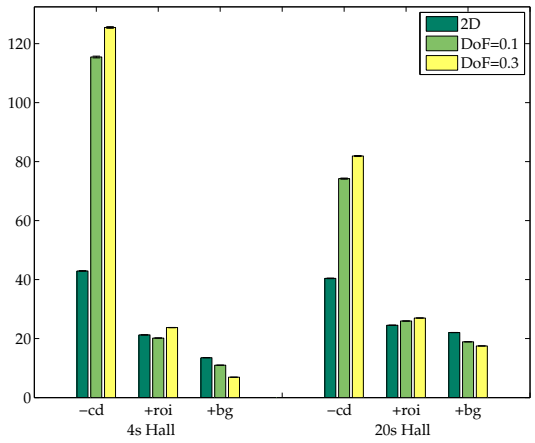
From the presented results it is not possible to conclude that gaze is guided by the amount of presented disparity. The distribution of gaze seems to depend on the saliency of the objects. For example, more attention is paid to the middle depth layer in the "Cartoon" scene, while more is paid to the background in the "Hallway" scene, whereas in the "Tea" scene, the foreground and the region of interest attracted almost an equal level of attention.

Furthermore, it is not possible to generalize where observers look at first and then at later periods of time. There are very few differences between the distribution of attention for the first 4 seconds in comparison with 20 seconds.

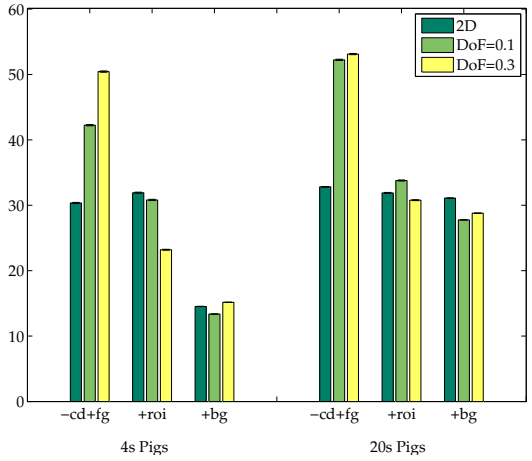
Also, the WSDM metric was computed using the stimuli from Experiment 3 (Section 5.4) with crossed and uncrossed disparities. The results for all the scene are presented in Figure 5.37.



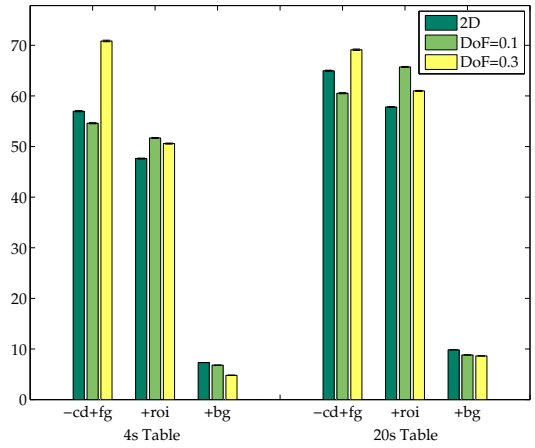
(a) WSDM for the "Cartoon" scene ($cd : [-255; 0], fg : [0; 170], roi : (170; 240], bg : (240; 255]$).



(b) WSDM for "Hall" scene ($cd : [-255; 0], roi : [0; 200], bg : (200; 255]$).



(c) WSDM for "Pigs" scene ($cd + fg : [-255; 85], roi : cd(85; 185], bg : (185; 255]$).



(d) WSDM for "Table" scene ($cd : [-255; -200], roi : cd(-200; 170], bg : (170; 255]$).

Figure 5.37: Depth metric for scenes containing objects with crossed disparities

The same observation pattern was found for all the scenes: objects with crossed disparity attracted the maximum amount of attention. This effect can be seen clearly by comparing the "Cartoon" scene with the crossed disparity airplane in Figure 5.37.a and similar scene in Figure 5.34.b. Though this effect was less pronounced for the "Table" scene (Fig. 5.37.d) since there are no objects with crossed disparity completely outside of the main region of interest.

To summarize, the results obtained with the help of the WSDM metric confirmed the conclusions from the previous experiments. For scenes located behind the display plane, the strategy for observing a still scene is similar to 2D. However, the more crossed disparity that is presented the more visual attention will be attracted by an object even if it causes visual discomfort.

5.6 Conclusions

This chapter describes three subjective experiments, which compare visual attention between 2D and 3D still stereoscopic images. Based on the results of the experiments there are several main conclusions:

- The observation strategy for still stereoscopic images located behind the display plane is similar to the observation of 2D images. Gaze is rather guided by the saliency of objects than by the amount of uncrossed disparities. Therefore, in the next chapters the effect of visual attention in 3D is not considered while taking into account that most of the produced content for cinema or television is content with uncrossed disparities.

- The objects with crossed disparities attract maximum attention: the more crossed disparities that were presented, the more visual attention was directed to that area.

- No evidence has been found that visual discomfort generated by excessive disparities influences the way we observe the images.

- A new weighted saliency depth metric based on the depth map and saliency maps was proposed relying on the results of subjective studies. The metric allowed the comparison of visual attention between 2D and 3D conditions as well as 3D conditions with different amounts of depth owing to weighted saliency maps. The computed results validated the conclusions from the eye-tracking experiments.

Part II

Objective modeling of 3D video QoE

Chapter 6

Objective model for S3D using perceptual thresholds

Contents

6.1 Introduction

In the case of 3D, the minimum requirement for stable system performance should be the absence of visual discomfort (see Section 3.7). Currently a subjective assessment is the best way to reflect the opinion of the viewers or customers about a proposed service. However, real-time services require objective metrics that are able to predict and monitor the video quality on the fly. Also such objective metrics should be able to guarantee a certain quality level of the provided video to end users. This chapter presents a new objective model that meets all the mentioned requirements.

6.2 Background and motivation

As explained in Chapter 3, 3D QoE is a multidimensional concept. Each perceptual attribute has an influence on the final perceived 3D QoE. Therefore, for an objective quality measurement, it is necessary to establish the link between the perceptual attributes and technical parameters of a 3D system. For this purpose, several models for 3D QoE were designed. These models are described in Section 3.4. Most of them consist of the primary perceptual attributes. A possible explanation of this fact is that low level attributes are simpler to evaluate in subjective tests. Besides, presumably they can establish a direct link with the technical parameters of any 3D system; unlike high level concepts, such as naturalness and sense of presence, which are composited attributes themselves. Taking this into account, models based on low level attributes seem to be more practical for application design. Therefore, among all the perceptual attributes, 2D image quality, visual comfort, and depth distortions seem to be more appropriate as basic perceptual attributes from our point of view:

- Image quality refers to 2D image quality in the studies of [Seuntiëns et al., 2006, Kaptein et al., 2008]. Hence, it can be assessed by conventional methods devoted to 2D quality assessment.

- Depth quality as well as depth rendering might be quite complicated to judge in comparison with depth quantity for the subjects [Chen et al., 2012c]. However, depth quantity does not reflect the 3D geometrical distortions (roundness), which is a one of factors considered by 3D producers to create realistic stereoscopic content [Mendiburu, 2009] and one of the concerns of scientists who find that it can be distracting [Smith and Collar, 2012, Smith and Malia, 2013] or uncomfortable for viewers [Doyen et al., 2012]. 3D geometrical distortions are the result of proportion violations between the real world and the visualization space and can occur as magnification/miniaturization of an object's dimensions or stretching/compression of depth, e.g shape distortion. Taking this into account realism linked with 3D image geometry, which was proposed by Vlad et al., might be a suitable perceptual attribute [Vlad et al., 2013].

- Comfort was found to be the most influential attribute on QoE [Chen et al., 2012c, Chen et al., 2012b, Lambooij et al., 2007, Tam et al., 2011]. Thus, the minimum task for any 3D system is to guarantee visual comfort to viewers [Chen et al., 2011].

Taking the discussion into account, Vlad's model [Vlad et al., 2013] of 3D QoE was found to be suitable as the basis of this thesis. But, as mentioned above, the axis "realism" was renamed to "3D geometrical distortion". The composition of the axes of this model should determine the overall perceived 3D QoE of any stereoscopic stimulus. As it has been demonstrated in various subjective experiments, all basic perceptual attributes can be assessed independently from each other. One of the ways to assess each axis is to define its inherent *subjective quality factor*, which viewers can experience subjectively. Then one or several *technical quality parameters* characterizing subjective quality factor should be identified. After the measurement of these technical quality factors with dedicated algorithms, an objective model can be applied. As a result, the stereoscopic video quality should be predicted objectively. The described framework for an objective prediction of 3D QoE is illustrated in Figure 6.1. It can be applied to any basic attribute of 3D QoE.

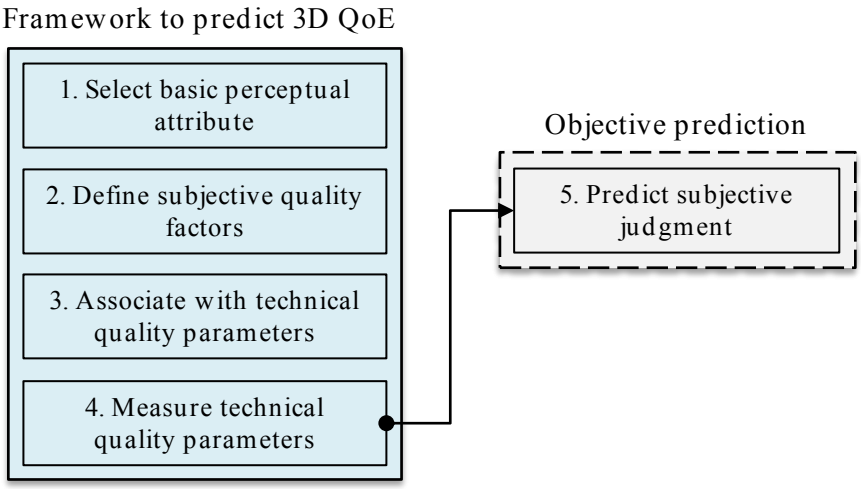


Figure 6.1: The framework to predict 3D video QoE

In this thesis, the following sections focus on the block "Objective prediction". But, first the block "Framework to predict 3D QoE" will be discussed in this section.

Figure 6.2 presents the axes of 3D video QoE with associated subjective quality factors. Selected image quality factors were proposed in [NTT, 2014]. One of the tasks of any selected objective method should be measurement of all artifacts related to 2D image quality axis. This complex issue was investigated a lot by various research groups. Therefore, it will be excluded from the studies of present thesis e.g. all test images must be free from any 2D image quality artifacts (see Chapter 2 for review).

Overall percept	3D video QoE		
Perceptual attributes	Image quality	3D geometrical distortion	Visual comfort
Subjective quality factors	**Spatial distorsion** -Reduced resolution, blurring -Block distortion -False outline **Temporal distortions** -Jerkiness -Flicker -Motion blur -Interruptions and frozen images **Spatio-temporal distortion** -Mosquito noise -"Busy" edges -Disturbance (failure)	**In planar direction** -Magnification -Miniaturization **In Z depth direction** -Compression -Stretching	-Visual annoyance -Visual fatigue

Figure 6.2: Subjective quality factors associated with basic perceptual attributes of 3D video QoE.

Subjectively, 3D geometrical distortions can be perceived as compression or stretching and magnification or miniaturization of objects in depth direction. Such distortions include the cardboard effect, puppet theater effect, gigantism, and miniaturization. Ob-

jectively, shape distortion can be computed when the camera and visualization space parameters are known (see for details Section 2.2.1 equation 2.8).

Even though it is known how to estimate the quantity of 3D geometrical distortions objectively, very few studies have been done to study it subjectively. Mendiburu indicates that a roundness factor between 0.7 and 1 is not discernible under perfect conditions (roundness factor =1) [Mendiburu, 2009]. However, these numbers were obtained empirically and not supported by any subjective tests. Thus, perceptual thresholds of roundness factor have not been studied accurately, especially considering target applications (TV, cinema, etc.). Also, little attention was paid to the impact of depth distortion on visual comfort. Nevertheless, depth distortion by itself supposedly does not violate the physiological mechanism responsible for depth perception like in the case of the vergence-accommodation conflict or severe view asymmetries. Furthermore, in the following experimental work, all test sequences will be free from noticeable stereoscopic distortions.

Visual discomfort as a result of the vergence-accommodation conflict or view asymmetries is a typical problem of 3D systems only. That is why the axis "Visual comfort" in Figure 6.2 gets the top priority in this thesis manuscript. However, visual fatigue will not be taken into account. As explained in Section 3.5.1, it can be induced by multiple excessive efforts of the visual system and requires some time to emerge. But, in the following subjective experiments, stereoscopic videos with a maximum duration of 15 seconds are used, which might be not sufficient to consider visual fatigue. Therefore, further description of the block "Framework to predict 3D QoE" is done for the basic perceptual attribute "Visual Comfort" excluding visual fatigue.

Figure 6.3 characterizes visual annoyance in terms of technical quality parameters (Px), which determine the possible causes of visual discomfort in S3D. Once this subjective quality factor is linked to the technical quality parameters, it can be evaluated subjectively and objectively.

Subjective quality factor	Technical quality parameters **Px**	Objective measurement of Px **Algorithm$_{Px}$**(Left, Right)	Distortion level D of Px, degradation units **Algo$_{Px}$(L,R)= D$_{Px}$**
Visual annoyance	-Vertical shift ($P_{vertical}$)	-ALGO$_{vertical}$(L,R)	- n, ° (lines)
	-View magnification (P_{focal})	-ALGO$_{focal}$(L,R)	- n, °
	-View rotation ($P_{rotation}$)	-ALGO$_{rotation}$(L,R)	- n, °
	-Keystone distortion ($P_{keystone}$)	-ALGO$_{keystone}$(L,R)	- n, °
	-Luminance mismatch (P_{black}, P_{white})	-ALGO$_{black,white}$(L,R)	- n, %
	-Color mismatch (P_{red},P_{green},P_{blue})	-ALGO$_{red,green,blue}$(L,R)	- n, %
	-Maximum crossed and uncrossed disparity (P_{CDmax}, P_{UDmax})	-ALGO$_{CDmax}$(L,R) -ALGO$_{UDmax}$(L,R)	- n, ° (pixels) - n, ° (pixels)

Figure 6.3: Technical quality parameters associated with the basic perceptual attribute "Visual comfort".

For objective evaluation, technical parameters Px should be measured using a dedicated algorithm or formula ($Algorithm_{Px}$). The output of the such algorithm is a distortion value (D_{Px}) in a unit of degradation. For example, vertical shift can be measured in degree of visual angle (n°) or number of lines; mismatch of white luminance level in percentage of mismatch (n%); maximum crossed and uncrossed disparities in degree of visual angle or number of pixels. Though, when it is possible it is recommended to

translate measured values into degrees of visual angle since the display size and visualization distance influence perception of stereoscopic content. But the usage of a degree of visual angle generalizes such dependencies of results.

Several software are available on the market that can accomplish objective measurements (StereoLabs tool, Cel-Scope, Sony MPE-200 etc.). They measure technical parameters but they do not provide reliable information about the impact on human perception.

As discussed in Chapter 1, binocular vision is a physiological mechanism. Hence, without the integration of human perceptual information, it will not possible to predict visual discomfort induced by 3D system and, thus, to conclude about 3D video QoE. Therefore, similar to recently standardized 2D metrics [OPTICOM, 2008, SwissQual, 2010], it seems reasonable to develop a 3D picture metric that considers the properties of human vision, rather than using data metric approaches that only take into account the characteristics of the signal.

Another issue of objective quality measurement is the necessity of establishing the link between the predicted MOS scores and the subjective ones to associate with a certain quality level. For instance, if an objective metric evaluates video quality with score of 23, it is impossible to conclude what it means in terms of quality ("Good?", "Poor?", "Comfortable?"). But when a score of 23 is referenced to a continuous quality scale, it is easy to deduce that the assessed video clip has "poor" quality. Thus, the performance of an objective model can be assessed using the results of subjective tests obtained with exactly the same scale that was used for objective prediction. Also, predicted MOS scores should have a high correlation and reliability with subjective test results.

In the next section a new approach of objective quality assessment is proposed, which characterizes a detected technical quality parameter based on its influence on viewer perception and avoids the direct prediction of MOS with a related quality level to increase reliability and decrease the complexity level.

Based on the above discussion, our motivation is to develop an objective 3D model for the characterization of 3D QoE that fulfills the following characteristics:

- Detected problem is categorized in accordance with human perceptual thresholds.
- Model that predicts category rather than a MOS score.
- Predicted objective scores can be easily validated via a subjective test.

6.3 Objective model proposition

6.3.1 Definition of objective categories

A new objective model that predicts objectively the impact of technical quality parameters relevant to visual discomfort on human perception is proposed. The model consists of three color categories that characterize a detected technical quality parameter in accordance with the evoked perceptual state. In our study *perceptual state* reflects a viewer's categorical judgment based on stimulus acceptability and induced visual annoyance. The proposed objective categories associated with their perceptual states are listed below:

- Green – no annoyance perceived.
- Orange – annoyance is acceptable.

- Red – unacceptable annoyance level.

Therefore, each perceptual state can comply with one or several perceptual thresholds, namely visual annoyance, acceptability. *Acceptability* determines the viewer's expectation level for the perceived video quality in a certain context and situation (inspired by the acceptability for the customer defined as "adequate service" in [Zeithaml et al., 1993]). Still acceptability is a high level concept and also can be considered as "the outcome of a decision which is partially based on the Quality of Experience" [Le Callet et al., 2012].

Figure 6.4 demonstrates how the boundaries between objective color categories are defined. The boundary between the "Green" and "Orange" categories defines the visual annoyance threshold (inspired by the impairment scale), while the boundary between the "Oran" and "Red" categories defines the acceptability threshold. The axis D demonstra
nical quali

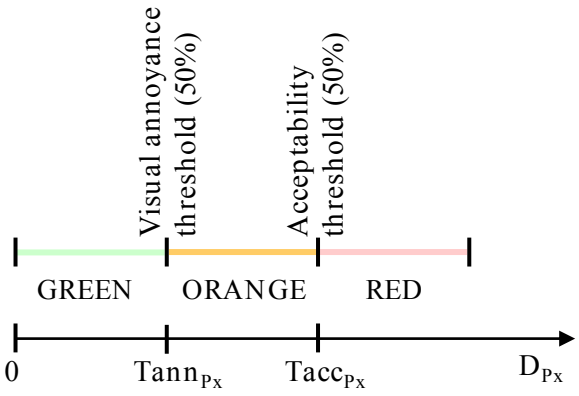


Figure 6.4: Definition of objective categories.

By default, the acceptability threshold level is defined as 50%, e.g. 50% of viewers would rank the subjective quality factor as unacceptable. The acceptability threshold might also be adapted based on service requirements. For example, 80% of acceptability for cinema, 70% for 3DTV service at home, etc. Any selected percentage of acceptability defines the width of "Red" category. The width of "Orange" category is quantified by the both thresholds: the acceptability and visual annoyance perceived by a percentage of viewers.

Summarized information about each category is presented in Table 6.1. Here, a perceptual state is divided into acceptability and visual annoyance subjective reactions to a stimulus. So, each objective color category is a result of two questions in the following order: (1) "Would viewers evaluate a stimulus as acceptable?", and (2) "Would viewers perceive visual annoyance?". However, if the stimulus has been assessed as not acceptable, it falls immediately into the "Red" category and the answer to the second question is not important. A color alert can be displayed during quality monitoring of stereoscopic content provided by a service.

6.3.2 Subjective color scale proposition

A subjective color scale can be constructed from the described above objective color categories as a categorical scale with labels. The "Red" category reflects a judgment

Table 6.1: Detailed description of objective color categories.

Subjective perceptual state		Objective category	Alert	Subjective category label
Acceptable?	Annoying?			
Yes	No	GREEN	Not annoying	Acceptable, not annoying
Yes	Yes	ORANGE	Annoying	Acceptable, but annoying
No	Yes	RED	Not acceptable	Not acceptable

concerning the acceptability of a subjective quality factor. e.g. annoyance, "Orange" if it is visually annoying and acceptable, and "Green" if it is not annoying. The labels for categories are defined in semantic terms in Table 6.1 column "Subjective category label".

The proposed *Color Scale (CS)* is illustrated in Figure 6.5. Color intervals can serve as a better visualization and to facilitate the viewer's choice. In order to assign a category to the viewed stereoscopic stimulus, observers can use the following two-step algorithm: (1) Evaluate if the stimulus is acceptable. If yes, proceed to the second step; if no, choose the "Red" category. (2) Evaluate if the stimulus is visually annoying. If yes, choose "Orange"; if no, choose "Green".

To compute MOS scores, each category should receive a numeric grade. For example, 0 – Not acceptable; 1 – Acceptable, but annoying; 2 – Acceptable, not annoying. Earlier the boundaries of objective categories were defined as 50% acceptability and visual annoyance thresholds. So, similarly for the subjective scale the boundary between the "Orange" and "Green" categories is a score 1.5, e.g. 50% of viewers find a stimulus annoying. The boundary between the "Red" and "Orange" categories is a score 0.5, e.g. 50% of viewers find a stimulus unacceptable.

Figure 6.5: Objective color scale modified into a subjective categorical scale.

Supposedly, the CS can be used in subjective experiments directly to obtain 50% acceptability and visual annoyance thresholds in the same test. For this, following the recommendation ITU-R BT.500-13, the relationship between the MOS and the distortion levels for a technical quality parameter should be approximated. The symmetry logistic function can be used to obtain this continuous relationship following the equation 6.1:

$$MOS_{CS} = \frac{2}{1 + e^{\frac{a+D_{Px}}{b}}} \tag{6.1}$$

where, MOS_{CS} - the MOS score, D - the objective distortion level and a, b – the estimation constants.

Further, acceptability threshold can be estimated as a distortion level corresponding to the score 0.5 and visual annoyance threshold as distortion level corresponding to the score 1.5 from the approximated curve. Schematically, this idea is presented in Figure 6.6.

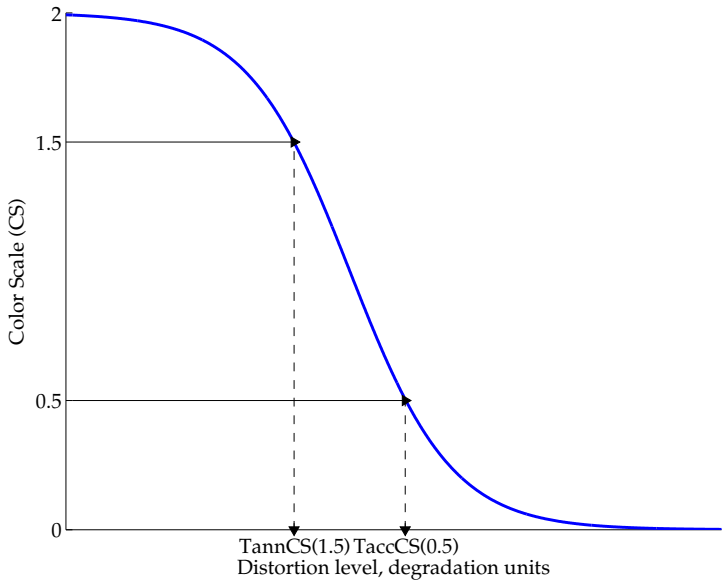


Figure 6.6: Color Scale (CS) curve with 50% acceptability (CS(0.5)) and annoyance (CS(1.5)) thresholds estimated from color scale.

But, it is necessary to verify if the obtained thresholds are the same as those collected with standard test methods.

The next section presents a method how estimate any percentage of acceptability and visual annoyance from the color scale.

6.3.2.1 Color Scale decomposition

The CS in Figure 6.5 is based on two perceptual thresholds. Hence, the constructed scale can be decomposed onto two scales: an acceptability scale and a visual annoyance scale as illustrated in Figure 6.7.

The data collected in subjective experiment with CS can be transformed according to Table 6.2. Such a transformation splits subjective color scale data into annoyance ($CSann$) and acceptability ($CSacc$) data sets imitating the decomposition of color scale to visual annoyance and acceptability scales. For example, votes of a viewer [2, 2, 1, 0, 1] using the categorical CS are converted to acceptability votes as [1, 1, 1, 0, 1] and annoyance votes as [1, 1, 0, 0, 0].

From the two resulting data sets, the acceptability and visual annoyance curves can be approximated to define the distortion level associated with the desired percentage of acceptability and/or visual annoyance. This idea is presented in Figure 6.8 for CSacc

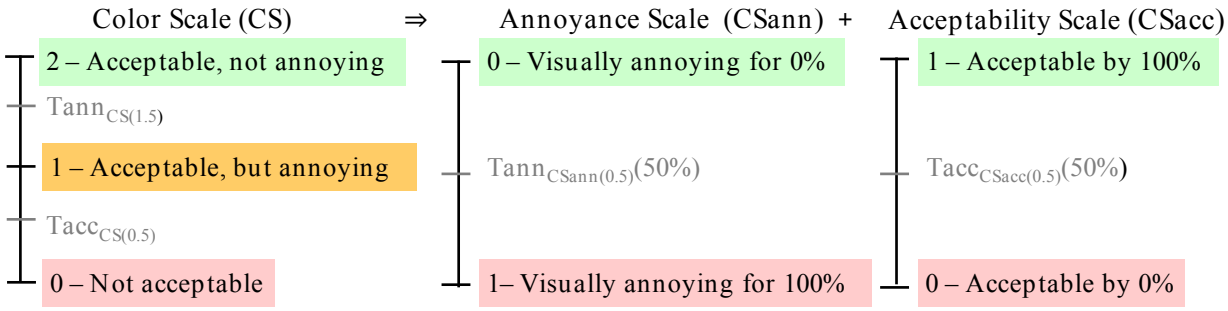


Figure 6.7: Color Scale decomposition.

Table 6.2: Color scale data decomposition to acceptability and visual annoyance data sets

Color Scale data		Acceptability data (CSacc)		Annoyance data (CSann)	
Score	Category	Score	Category	Score	Category
2	Acceptable, not annoying	1	Acceptable	1	Not annoying
1	Acceptable, but annoying	1	Acceptable	0	Annoying
0	Not acceptable	0	Not acceptable	0	Annoying

data set. As a result, the distortion level associated with x% of acceptability is used as the acceptability threshold.

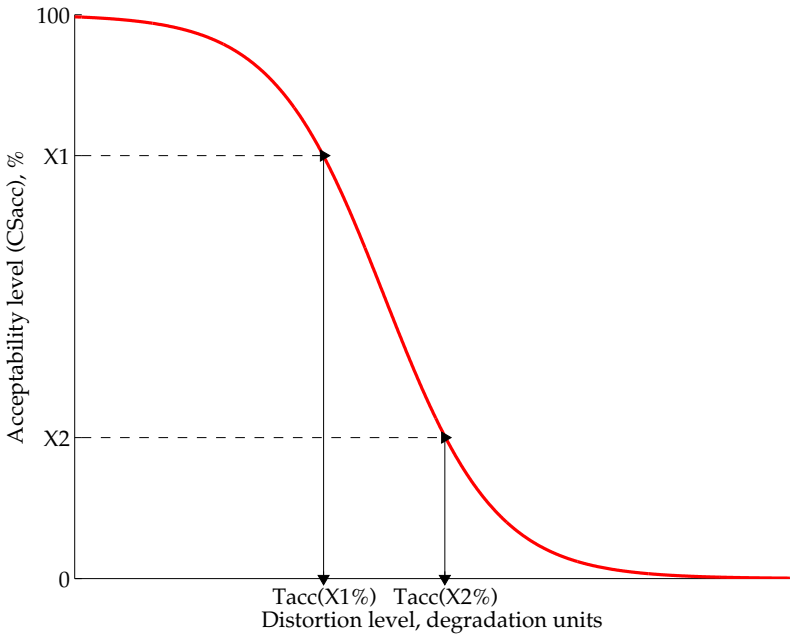


Figure 6.8: Acceptability curve (CSacc) with x1% and x2% acceptability thresholds estimated from CSacc data.

For a curve approximation, MOS scores with binomial probability confidence interval [Clopper and Pearson, 1934, Soper, 2014] should be computed from a transformed data set. Then following the recommendation ITU-R BT.500-13 Annex 2, the relationship between the MOS and the distortion levels should be approximated. This allows the

estimation of a distortion level for desired percentage of acceptability or annoyance. The symmetry logistic function can be used to obtain this continuous relationship following the equation 6.2:

$$MOS = \frac{1}{1 + e^{\frac{a+D_{Px}}{b}}} \tag{6.2}$$

where, D_{Px} - the distortion level of technical quality parameter Px and a, b – the estimation constants.

The tolerance range of each acceptability threshold can be estimated if the confidence interval curves are approximated as well [ITU, 2012b].

6.3.3 Definition of the boundaries of objective categories

Furthermore, to construct an objective model for a 3D system, it is necessary to define the boundaries of objective categories: acceptability and visual annoyance thresholds for all technical quality parameters. Here are several ways to obtain them:

1. The thresholds can be adopted from state-of-the-art studies. However, in this case it is important to make sure that the thresholds were received under the same conditions (screen size, viewing distance, and 3D technology) as the target 3D system. Another solution is to use the generalized thresholds, which do not depend on visualization parameters. For example, thresholds expressed in degree of visual angle, when it is possible.
2. The thresholds can be determined via a subjective test using any standard method [ITU, 2012b, ITU, 2012a].
3. The thresholds can be defined using the proposed subjective color scale. This method allows both perceptual thresholds in the same subjective test to be defined.

Independent of the selected method, all thresholds should be found based on the parameters of the target 3D service (3DTV at home or cinema).

6.3.4 Proposal of Objective Perceptual State Model (OPSM)

Once the limits of the objective categories are identified for the technical quality parameters of all perceptual attributes of 3D QoE, the objective model can be used for monitoring the video quality of a 3D service. The distortion level values of technical quality parameters should be measured using some existing software, tool, or method. Then, the detected level of distortion should be compared with the associated acceptability and visual annoyance thresholds (in degradation units) and placed in the corresponding objective category.

The proposed Objective Perceptual State Model (OPSM) is represented in Figure 6.9. In the Figure, cyan blocks depict the framework created by other researchers, while gray blocks represent the propositions of this thesis, which will be explored and validated in the following chapters. We believe that this model can be applied to any basic perceptual attribute of 3D QoE model and then the obtained scores should be combined. However, this hypothesis will not be explored within the scope of this thesis and the following subjective experiments will only involve the visual comfort axis of 3D QoE. Therefore, in further studies, the proposed model will be referred to as “metric” (OPSM - Objective Perceptual State Metric), taking into account that two other axes of 3D video QoE, i.e depth rendering and 2D image quality, are not considered.

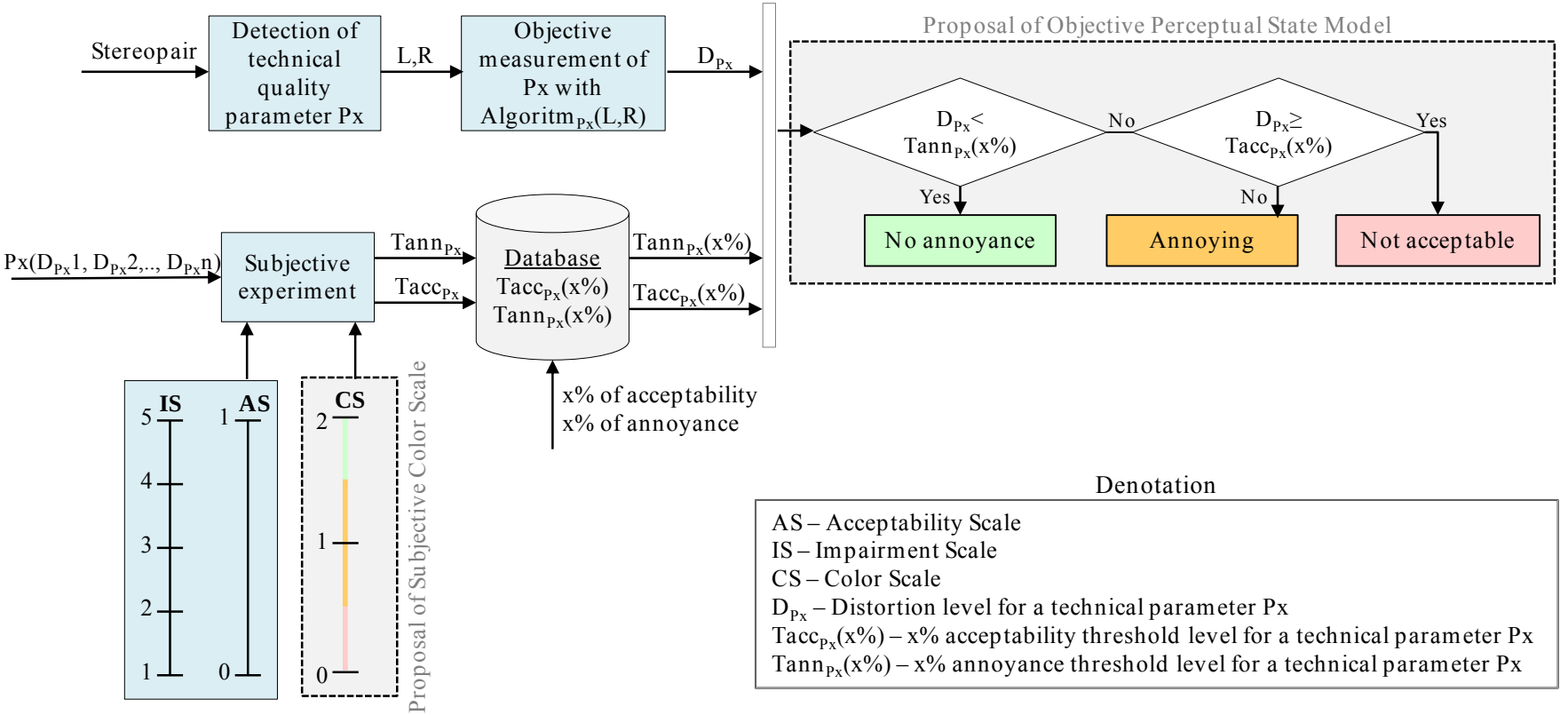


Figure 6.9: Proposal of an Objective Perceptual State Model (OPSM). The inputs required by the metric are the distortion level of a technical quality parameter, an annoyance threshold of the technical quality parameter, and an acceptability threshold of the technical quality parameter: $OPSM(D_{Px}, Tann_{Px}, Tacc_{Px})$.

6.3.5 OPSM validation with subjective experiments

In the following chapters the metric validation will be organized in accordance with Figure 6.9 in four stages as follows:

1. Subjective experiment design. Several distortion levels are introduced for a technical quality parameter to create perceptual variations of stimuli from comfortable to uncomfortable.

2. The subjective evaluation of the generated stimuli using the proposed CS. The MOS scores should be computed from the collected votes of observers. Then subjective color categories can be associated with the obtained MOS.

3. The prediction of objective categories for the generated stimuli using the proposed OPSM metric.

4. A comparison of the OPSM prediction with the subjective experiment results shown graphically and quantitatively (Pearson correlation coefficient).

For example, the technical parameter Px is used for the validation of our metric. The stimuli for the subjective experiment was produced by introducing three different levels of distortions to an undistorted content: $Px(D_{Px}1, D_{Px}2, D_{Px}3)$. Then the MOS scores were computed as presented in Table 6.3.

Table 6.3: Data from a subjective experiment with CS

Stimulus	MOS_{CS}	$Tann_{CS}(50\%)$	$Tacc_{CS}(50\%)$	Subj. category
$D_{Px}1$	1.7			Green (2)
$D_{Px}2$	1.2	1.5	0.5	Orange (1)
$D_{Px}3$	0.3			Red (0)

In the CS, the boundary between the "Orange" and "Green" categories is a score of 1.5 (column $Tann_{CS}$). The boundary between between the "Red" and "Orange" categories is a score of 0.5 (column $Tann_{CS}$). Thus, if MOS_{CS} is $\in [0, 0.5]$ then the subjective category is "Red"; otherwise, if it $\in (0.5, 1.5]$ then the subjective category is "Orange"; otherwise $\in (1.5, 2]$, e.g. "Green". Similarly, the subjective color categories can be read off directly from the graph as illustrated in Figure 6.10.

The prediction of objective color categories for the selected distortion levels are preformed using the proposed OPSM metric (see Fig. 6.9). In our example, distortion levels for the stimuli are already known from the subjective experiment: $Px(D_{Px}1, D_{Px}2, D_{Px}3)$. Thus, only acceptability and visual annoyance thresholds for the technical quality parameter Px are required to accomplish an objective prediction. See the data in Table 6.4:

Table 6.4: Data used for prediction of objective color categories

Stimulus	D_{Px}	$Tann_{Px}(50\%)$	$Tacc_{Px}(50\%)$	Obj. category
$D_{Px}1$	0.5			Green (2)
$D_{Px}2$	1.5	1	2	Orange (1)
$D_{Px}3$	2.5			Red (0)

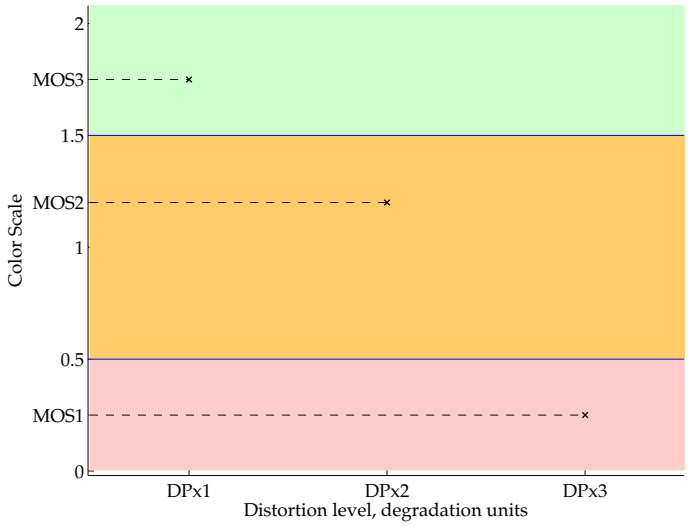


Figure 6.10: Definition of subjective categories from MOS obtained with CS. $MOS1 \in [0, 0.5]$ is in "Red" category, $MOS2 \in (0.5, 1.5]$ is in "Orange" category, $MOS3 \in (1.5, 2]$ is in "Green" category.

If a measured distortion level (D_{Px}) is higher than the associated acceptability threshold ($Tacc_{Px}$), then the objective category is "Red"; otherwise, if it is higher than the associated visual annoyance threshold ($Tann_{Px}$), then the objective category is "Orange"; otherwise it is "Green". This idea is illustrated in Figure 6.11, where perceptual thresholds are presented as solid vertical lines. Such illustrations allow for the association of objective categories with their corresponding ranges of distortion levels.

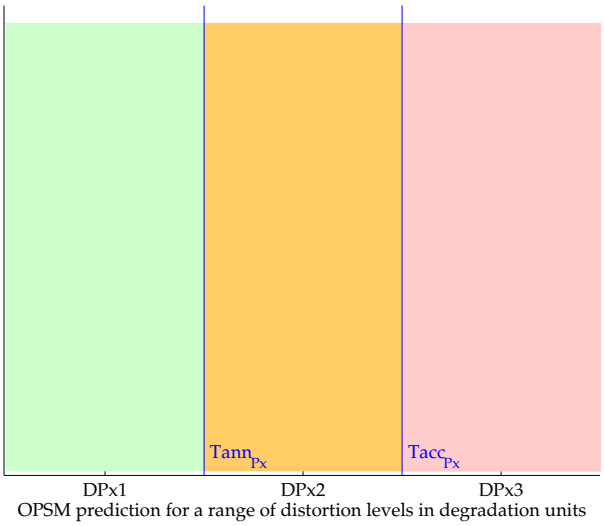


Figure 6.11: The objective prediction of color categories using acceptability and visual annoyance thresholds. $D_{Px}1 < Tann_{Px}$, so the objectively predicted category is "Green"; $Tann_{Px} \leqslant D_{Px}2 < Tacc_{Px}$ - "Orange"; $D_{Px}3 \geqslant Tacc_{Px}$ - "Red".

For a graphical comparison of the subjective categories with the objective prediction, Figures 6.10 and 6.11 can be merged into Figure 6.12. The intersection of the figures creates the color rectangles. The width of the rectangles is defined by perceptual

thresholds. If the subjective prediction matches the objective one for a given distortion level, the corresponding MOS score should be inside the associated color rectangle.

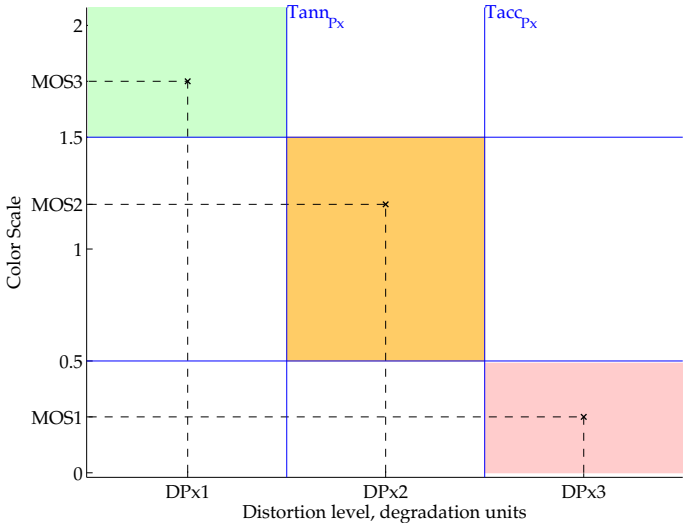


Figure 6.12: Objective categories vs. Subjective categories.

For a quantitative comparison of OPSM prediction with subjective data, the Pearson correlation coefficient (r) can be computed between the subjective and objective categories. In our example, the correlation between the subjective [2,1,0] and the objective [2,1,0] categories is $r = 1$. In the following chapters, the OPSM metric will be tested with various technical quality parameters.

6.3.6 Aggregation of technical quality parameters

It is possible that two or more technical quality parameters can be detected for the same stereopair in Figure 6.9. For example, some percentage of the green channel is mismatched and vertically shifted. So which perceptual state should be predicted?

In the case of visual discomfort, presumably if at least one of the categories is "Red", the overall quality should be in the "Red" category; then, if at least one of the categories is "Orange", the overall quality should be in the "Orange" category; otherwise, it should be "Green". This idea is illustrated in Figure 6.13, where C is the array, which contains the predicted categories for the detected technical quality parameters in the stereopair

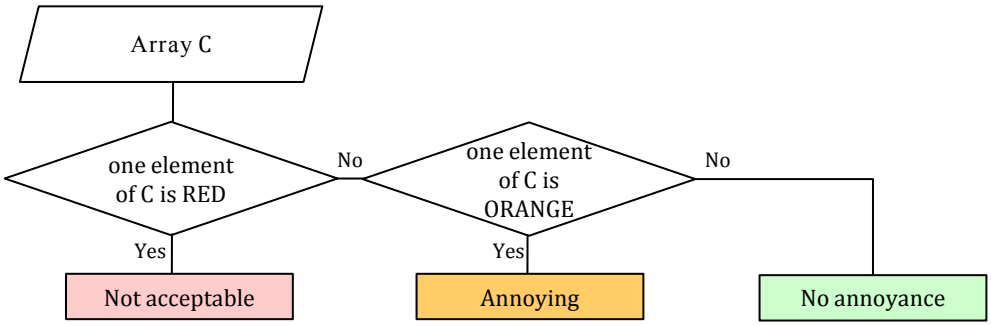


Figure 6.13: Aggregation of technical quality parameters.

For example, the combination of the P_1 and P_2 technical quality parameters can result in various perceptual states depending on their distortion levels as shown in Table 6.5. In this table, the distortion level of any technical parameter D_{P_x} can be in the R - "Red", O - "Orange", or G - "Green" category.

Table 6.5: Aggregation of two technical quality parameters P_1 and P_2

D_{P_1}/D_{P_2}	G	O	R
G	G	O	R
O	O	O	R
R	R	R	R

6.3.7 Acceptability and annoyance thresholds comparison

In the following chapters of this thesis the acceptability and annoyance thresholds for a technical quality parameter P_x obtained in different experiments will be compared. These thresholds can be evaluated using different scales or methods. For this, following the recommendation ITU-R BT.500-13, the relationship between the MOS and the distortion levels for a technical quality parameter P_x should be approximated. Then the perceptual threshold is evaluated from the approximated curve and represents a distortion level D_{P_x}.

To facilitate the understanding, the thresholds will be denoted in the following way:

$$Ttype_{scale(grade)}(level\%)$$

where the descriptors of thresholds' denotation are presented in Table 6.6

Table 6.6: The description of the thresholds' denotation. The descriptors marked in cyan are obligatory.

Descriptor	Definition	Values
Ttype	threshold type	$\{ann, acc, vis\}$
scale	a scale or method used for the definition of a threshold	$\{CS$, $CSann$, $CSacc$, $Chen$, DS, AS, $IS\}$
grade	a grade on the selected scale used for the approximation of a distortion level representing a level of acceptability or visual annoyance	$grade \in scale$
level%	level of acceptability or visual annoyance perceived by a percentage of viewers. The level is associated with the distortion level of a technical quality parameter	$[0, 100]\%$

The descriptor *level* can be omitted in the case of binary scales that provide only two choices to a subject, such as Acceptability Scale (AS). In this case level of acceptability is equal to selected grade on the scale: $Tacc_{AS(0.5)} = Tacc_{AS(0.5)}(50\%)$. It also means that a degradation level was approximated as the grade 0.5 on the AS curve and 50% of viewers find this level of distortion acceptable. However, the annoyance threshold

estimated as the grade 3.5 on Impairment Scale (5-points) is only supposed to have the level of annoyance equal to 50%. Therefore, the level can not be indicated: $Tacc_{IS(3.5)}$ and it just means that the degradation level was estimated from the MOS obtained with the IS as the grade 3.5.

For instance, the annoyance threshold can be described in the following ways:

- $Tann_{CS(1.5)}$ - the threshold is obtained using the Color Scale (CS) as the score 1.5;
- $Tann_{CSann(0.5)}(50\%)$ - the threshold is obtained using the CS annoyance curve (CSann) derived from the CS as the score 0.5 (see Section 6.3.2.1);
- $Tacc_{IS(3.5)}$ - the threshold is obtained using Impairment Scale (IS) as the grade 3.5 (see Section 3.5.1.2) from the approximated curve.

The acceptability threshold can be described in the following ways:

- $Tacc_{CS(0.5)}$ - the threshold is estimated using the Color Scale (CS) as the score 0.5;
- $Tacc_{CSacc(0.5)}(50\%)$ - the threshold is obtained using the CS acceptability curve (CSann) derived from the CS as the score 0.5 (see Section 6.3.2.1);
- $Tacc_{AS(0.5)}(50\%)$ - the threshold is obtained using the Acceptability Scale (AS) as the score 0.5;
- $Tacc_{Chen}(50\%)$ the threshold is obtained using Chen's method (see Figure 3.11).

Basically, the thresholds comparison is the comparison of distortion levels of corresponding technical quality parameters.

6.4 Conclusions

This chapter proposes a new objective model based on perceptual thresholds. Such a model has several advantages:

- The prediction of MOS scores is omitted.
- It can be adapted based on the service requirements of customer acceptability.
- Automatic color warnings with alerts can be displayed during operational monitoring in real time.
- The method does not depend on any precise 3D technology. No reference is required to predict visual discomfort.
- Other perceptual thresholds can be introduced if there is a necessity. For example, a visibility threshold (see Fig. 6.14).
- The possibility of using the proposed model as a subjective scale would allow a direct link between subjective experiments and objective predictions to be established.

With the help of subjective experiments, the following chapters aim to answer several questions regarding different aspects of the proposed OPSM as illustrated in Figure 6.15.

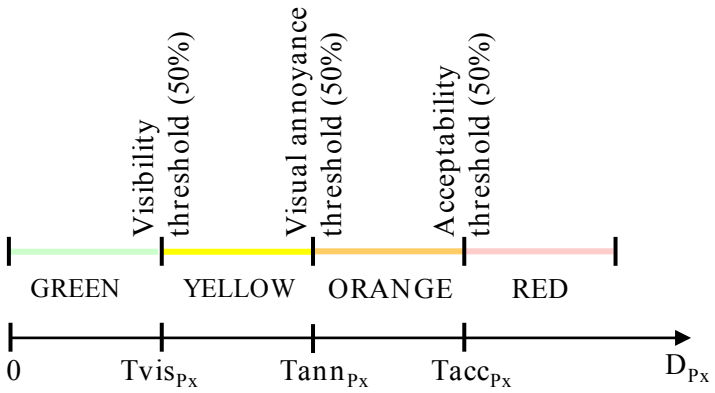


Figure 6.14: Color Scale with visibility threshold.

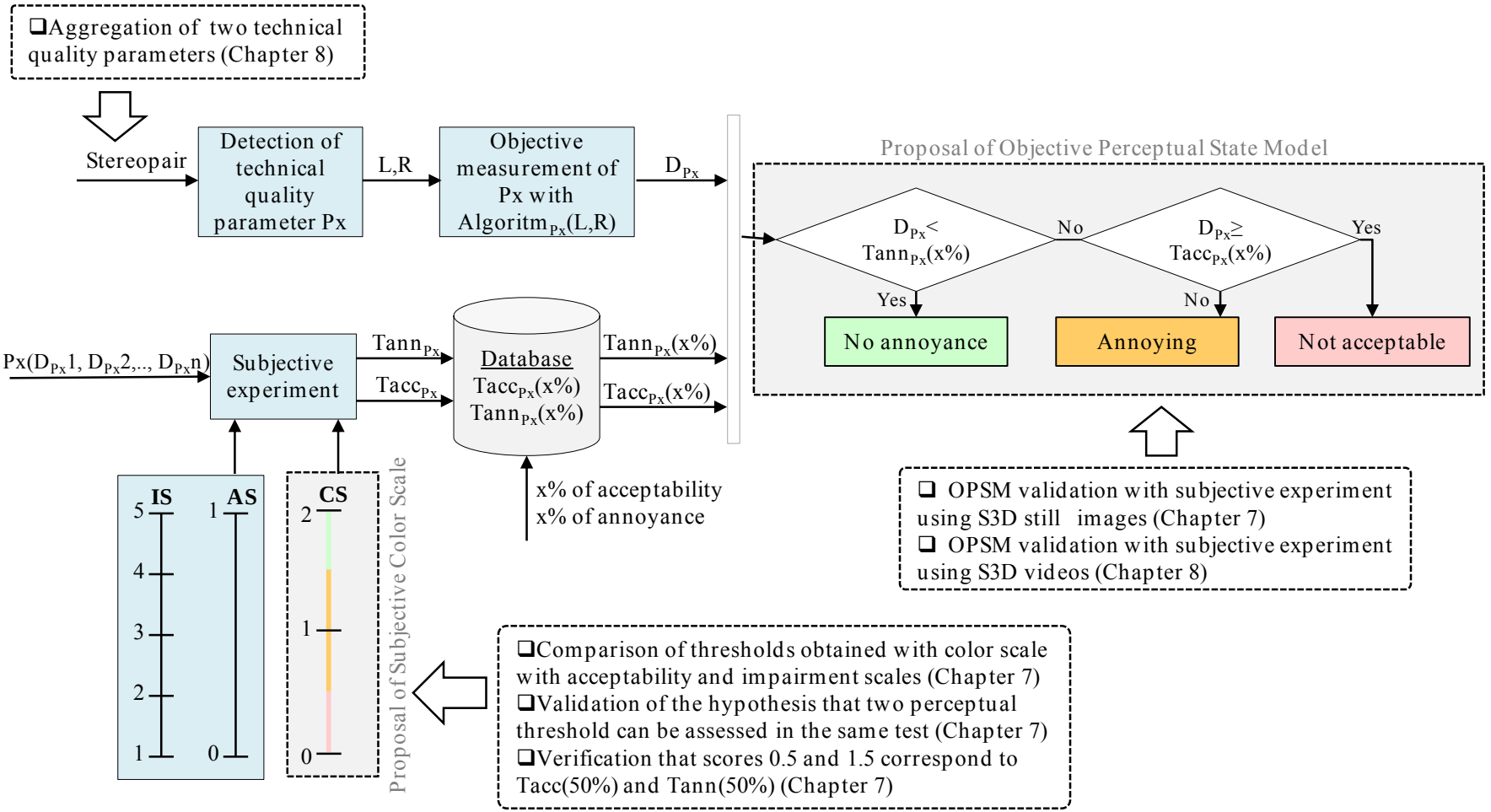


Figure 6.15: Proposal of Objective Perceptual State Model (OPSM)

Chapter 7

Metric validation using still S3D images

Contents

7.1 Introduction

A guarantee of visual comfort for viewers is the minimum requirement for any stereoscopic imaging system. Being able to detect visual discomfort automatically would allow selecting and, if needed, postcorrecting stereoscopic content without any subjective tests. Chapter 6 presented new objective metrics that use perceptual thresholds to define the impact of technical parameters on the visual comfort axis of 3D video QoE. After the objective measurement of 3D technical quality parameters and a comparison with perceptual thresholds, it would be possible to predict the evoked perceptual state, which would reflect a viewer's categorical judgment based on stimulus acceptability and induced visual annoyance. The goal of this chapter is to verify the proposed metric by comparing predicted categories with votes from subjective tests.

7.2 OPSM metric validation. "Color Scale" experiment

View asymmetry is a problem that can be easily introduced in a stereoscopic system. For instance, content creation with a toed-in camera can produce vertical disparity and

keystone distortion. The misalignment of cameras or projectors can lead to vertical shift, rotation, and magnification between views. Color and luminance mismatch of the camera sensors, glasses, or display filters can create other types of view asymmetries. All these view discrepancies can cause visual discomfort [Kooi and Toet, 2004, Chen et al., 2010, Chen, 2012]. It was decided to use a view asymmetry problem for the validation of the proposed objective model. The following subjective experiments will only involve the visual comfort axis of 3D QoE. Therefore, in further studies the OPSM model described in Chapter 6 will be referred as "metric" taking into account that two other axes of 3D video QoE are not considered, i.e depth rendering and 2D image quality.

7.2.1 Stimuli generation

Three stereoscopic images with different levels of complexity were selected for the experiment (see Fig. 7.1). "Forest" with a depth level DoF=0.2 diopters is considered as a high-level texture scene. "Butterfly" with a depth level DoF=0.1 diopters is a mid-level texture scene, and "Basketball" is a 2D scene (DoF=0 diopters) with low-level texture, where the left and right views are identical. "Forest" and "Butterfly" were rendered with a parallel camera configuration using Blender software with a virtual camera sensor set at $32 \times 16mm$. Other camera and scene parameters are presented in Table 7.1, where f – focal length, $dCon$ – convergence distance, b - baseline distance, fg – foreground distance, e.g. the distance from the camera to the closest object, bg – background distance, e.g. the distance from the camera to the farthest object, ROI- region of interest within the scene, DoF – depth of focus, and D_s - the shape distortion factor calculated for the region of interest as explained in Section 2.2.1.

(a) Forest (b) Butterfly (c) Basketball

Figure 7.1: Stimuli used in the experiment: (a) Forest DoF=0.2 D, high texture; (b) Butterfly DoF=0.1 D, middle texture; (c) Basketball DoF=0 D, low texture.

Table 7.1: Scene and camera parameters

Scene	f, mm	dCon, m	b, mm	fg, m	bg, m	RIO, m	DoF, D	D_s
Forest	36	5	5	5	23	7.5	0.2	1.26
Butterfly	70	6.8	118	5.8	12	6.8	0.1	0.69
Basketball	9	5	0	5	10	7	0	-

The advantage of using synthetic scenes is that in virtual space the camera parameters and alignments can be controlled easily to avoid any view asymmetries. The synthetic scenes were based on the open animation project "Big buck bunny" (Blender-Foundation, 2008).

Several different view asymmetries were generated so that all major groups of asymmetries would be represented. These groups are: geometrical (vertical shift, magnifica-

tion, rotation), color (R,G,B), and luminance (black, white) asymmetries.

Geometrical asymmetry. Vertical shift, magnification (focal lens difference), and rotation were selected for image processing. These geometrical asymmetries were intended to imitate the misalignment of the cameras, projector position, or incorrect post-production. The view asymmetries denoted in mathematical formulas can be found in [Chen, 2012] (see pp. 149-150) and an illustration is in Figure 7.2. Vertical shift causes vertical disparity for the entire view of a stereopair. While rotation and magnification produce v

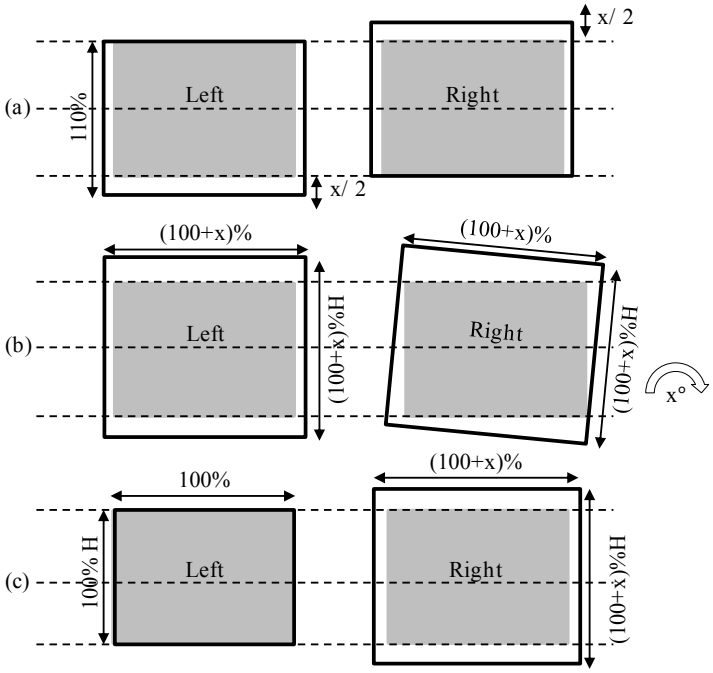


Figure 7.2: Geometrical asymmetries (a) Vertical shift (b) A view rotation (c) A view magnification; from [Chen, 2012].

For the simulation of vertical shift, a 110% resize function with a lanczos3 filter was applied to the original image on both views to avoid a black border. Next, the resized distorted image was cropped from the center. Then, both left and right views were shifted by a $x/2$ percentage of the height of the resized image in order to generate x% of vertical shift. The idea in presented by the equations 7.1- 7.2.

$$I^{dist}_{left} = CROP_{vertical}(RESIZE(I^{origin}_{left}), \frac{x}{2}) \tag{7.1}$$

$$I^{dist}_{right} = CROP_{vertical}(RESIZE(I^{origin}_{right}), -\frac{x}{2}) \tag{7.2}$$

where, I^{dist} – the distorted image, I^{origin} – the original image, and x – the distortion level in percentage of the width of the resized image.

To simulate rotation asymmetry, similar to the manipulation of vertical disparity, both views were resized before the rotation in order to avoid a black border. This is denoted by the equation 7.3:

$$I^{dist} = ROTATE(RESIZE(I^{origin}), x) \tag{7.3}$$

where, I^{dist} – the distorted image, I^{origin} – the original image, and x – the distortion level degree. The rotation was generated with a bicubic interpolation function.

Finally, to create a view magnification, the following equation 7.4 was implemented:

$$I^{dist} = MAGNIFY(I^{origin}, 100 + x) \tag{7.4}$$

where, I^{dist} – the distorted image, I^{origin} – the original image, and x – the distortion level as a percentage of height and width of the original image. The magnification was generated with a lanzcos3 filter.

Color asymmetry. The green channel color asymmetry was selected for image processing (see Figure 7.3). This asymmetry may appear because of the imperfect calibration of a camera's color triangles, color channel multiplex techniques, or a polarized filter in

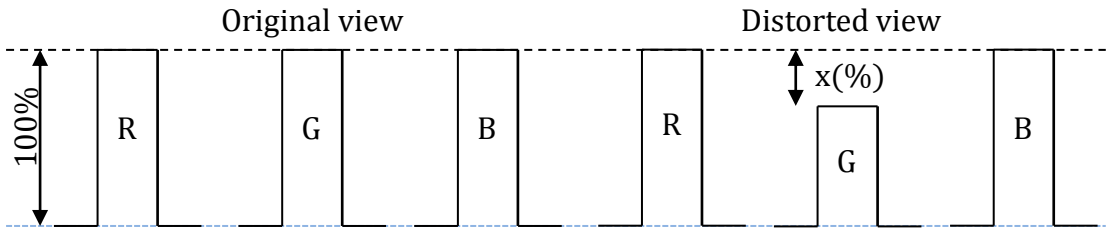


Figure 7.3: Color asymmetry in Green channel; adapted from [Chen, 2012].

The stimuli generation followed the equation 7.5:

$$L^{dist}(G) = L^{origin}(G) \times (1 - x) \tag{7.5}$$

where, L^{dist} – the distorted luminance value of the image, L^{origin} – the original luminance value of the image, and x – the distortion level as a percentage of the green color channel.

Luminance asymmetry. The white channel color asymmetry was selected for image processing. TLuminance asymmetry occurs because of an imperfect calibration of the camera's optics in the case of a camera mirror rig, the filters, or the imperfection of the

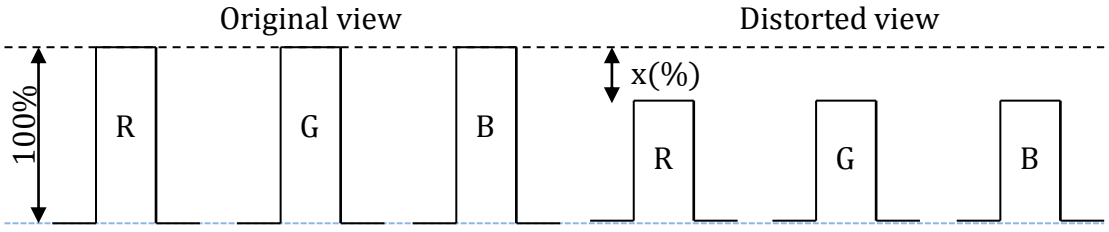


Figure 7.4: White level luminance asymmetry from [Chen, 2012].

The stimuli are generated with the equation 7.6 for white level distortion:

$$L^{dist}(RGB) = L^{origin}(RGB) \times (1 - x) \tag{7.6}$$

where, L^{dist} – the distorted luminance value of the image, L^{origin}– the original luminance value of the image, and x – the distortion level as a percentage of the color channels.

Table 7.2: Five types of view asymmetries with four-levels of distortion.

Distortion level D_{Px}	Level 1	Level 2	Level 3	Level 4
Vertical shift	0.4%	1%	1.4%	1.8%
Rotation	0.2°	0.5°	1°	2°
Magnification	0.4%	1%	1.4%	2%
Green	10%	20%	30%	50%
White	10%	20%	30%	50%

For each asymmetry, four different distortion levels were selected by video experts in the pre-test as shown in Table 7.2. Therefore, in total, 60 sequences (3 scenes * 5 view asymmetries * (4 distortion levels)) were prepared for the subjective experiment.

7.2.2 Experimental set-up and methodology

Test set-up: the subjective experiments were performed in the test room in compliance with the recommendation ITU-R BT.2021. A Hyundai 46" line interleaved stereoscopic display was used for the visualization of the stimuli. The dimensions of the display are 102×56 cm; the resolution in 2D is 1920×1080, and in 3D 1920×540 per view. The luminance, brightness, contrast, and color of the display were adjusted to the normal gamma function (gamma equals 2.2). The display's color triangle is illustrated in Figure B.1. The crosstalk level was less than 3% and the maximum luminance level measured through glasses was $100cd/m^2$. The viewing distance was 4.5 times the height of the display. An additional Dell 22" LCD display was used to present the test interface and store the votes of the observers.

Obsevers: 33 non-expert observers participated in this test. Their monocular acuity, color vision, far vision test, fusion test, and stereoscopic acuity were checked using Essilor ERGOVISION equipment prior to the subjective experiment. All observers had a normal or corrected to normal visual acuity and normal stereoacuity.

Methodology: The instruction sheet presented in Appendix C.1 offered some explanations on how to behave during the experiment and how to rate the sequences. The instructions were also explained by the examination to ensure that the observers understood the task. The SAMVIQ protocol was used to evaluate the sequences on the Color Scale described in Section 6.3.2. The test interface is presented in Figure D.1. The first part of the experiment consisted of 3 tests, where observers assessed 3 types of view asymmetries. To avoid any accumulation of visual discomfort, the observers evaluated the second part of the experiment with the remaining 2 asymmetries on the next day. The asymmetry levels were presented in random order.

In total, every subject had to evaluate 90 stimuli (3 scenes * 5 types of asymmetry * [4 distortion levels + 1 explicit reference + 1 hidden reference]). The visualization time of one stereoscopic pair was 8 seconds. In average it took around 12 minutes for the subject to evaluate one type of asymmetry, e.g. around 1 hour for the whole experiment.

7.2.3 Using the Color Scale for thresholds estimation. "Color Scale" experiment

The MOS scores collected in the "Color Scale" experiment with a 95% confidence interval were computed for all distortion levels of the five view asymmetries. The one way

ANOVA analysis demonstrated that the impact of content was insignificant for a change of MOS scores, while the distortion level had a significant ($p < 0.0001$) impact for all types of asymmetry. The effect of the scene was found to be significant only for rotation asymmetry ($p < 0.02$). This may be explained by the fact that the rotation distortion created false disparities and false depth perception in the case of the 2D scene "Basketball", which was not presented by an explicit reference.

The visual annoyance and acceptability thresholds were calculated from the approximated CS curves for all view asymmetries as the scores 0.5 and 1.5 (for explanations see Section 6.3.2) for all the view asymmetries. The curves were approximated for average scene values considering that the impact of the scenes was found to be insignificant.

The example of such approximation for the focal asymmetry is illustrated in Figure 7.5, where the annoyance threshold was estimated from mean fit as 0.62% of magnification and acceptability thresholds as 1.4%. The tolerance ranges (TR) of these thresholds are estimated from the curves representing 95% confidence interval and indicated in Figure 7.5 as dotted lines. The results of thresholds estimation for all the view asymmetries are presented in Table 7.3.

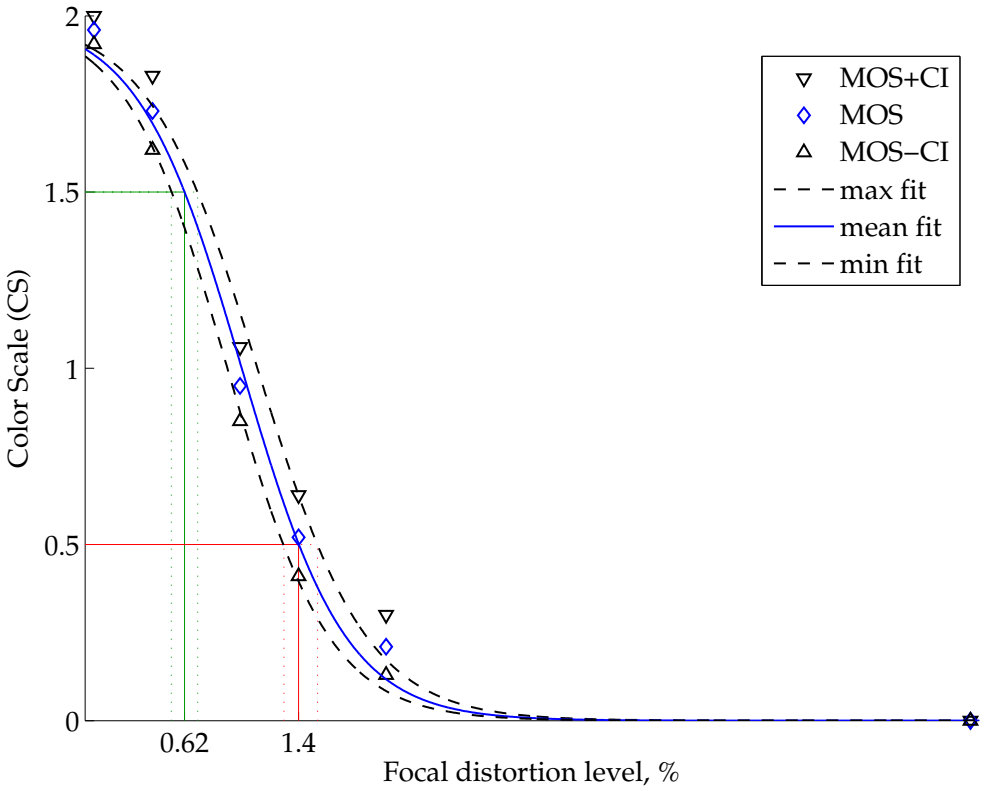


Figure 7.5: Color Scale (CS) curve and two curves representing 95% confidence interval approximated for the focal asymmetry from MOS scores. The acceptability (CS(0.5)) and annoyance (CS(1.5)) thresholds are estimated from the mean fit with the tolerance range estimated from min and max fits.

Table 7.3: Acceptability and visual annoyance thresholds calculated from the "Color Scale" experiment data as distortion levels corresponding to the scores 0.5 and 1.5 on the Color Scale (CS) with tolerance range (TR).

Asymmetry Px	Vertical shift,%	Rotation,°	Focal,%	Green,%	White,%
$Tann_{CS(1.5)}$	**0.7**	**0.5**	**0.62**	**0.233**	**0.244**
TR (upper limit)	0.24	0.11	0.09	0.027	0.003
TR (lower limit)	0.2	0.08	0.09	0.018	0.026
$Tacc_{CS(0.5)}$	**1.64**	**1.15**	**1.4**	**0.405**	**0.434**
TR (upper limit)	0.21	0.12	0.13	0.025	0.031
TR (lower limit)	0.24	0.12	0.1	0.037	0.032

The same thresholds were computed using the Color Scale decomposition method explained in Section 6.3.2.1 to demonstrate that the thresholds computed above as scores 0.5 and 1.5 truly correspond to 50% acceptability and annoyance. The result of this comparison is presented in Figure 7.6.

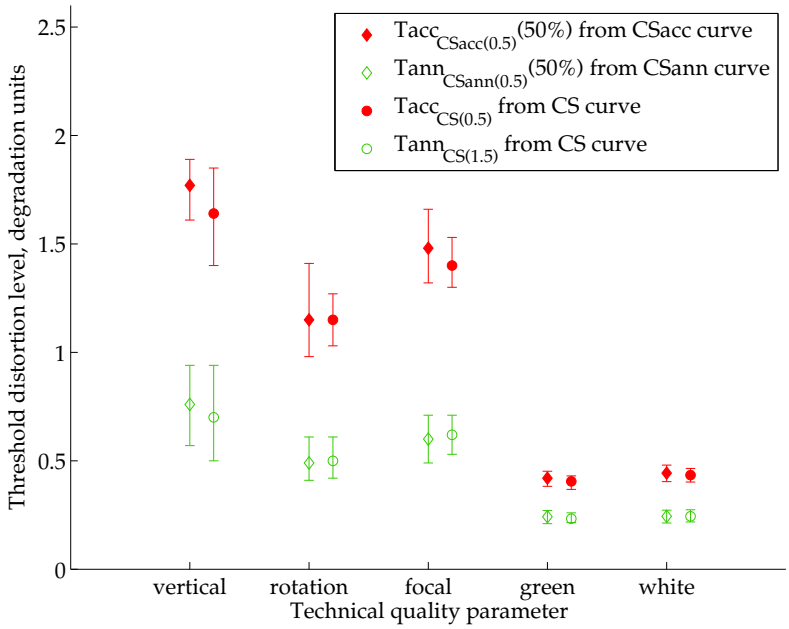
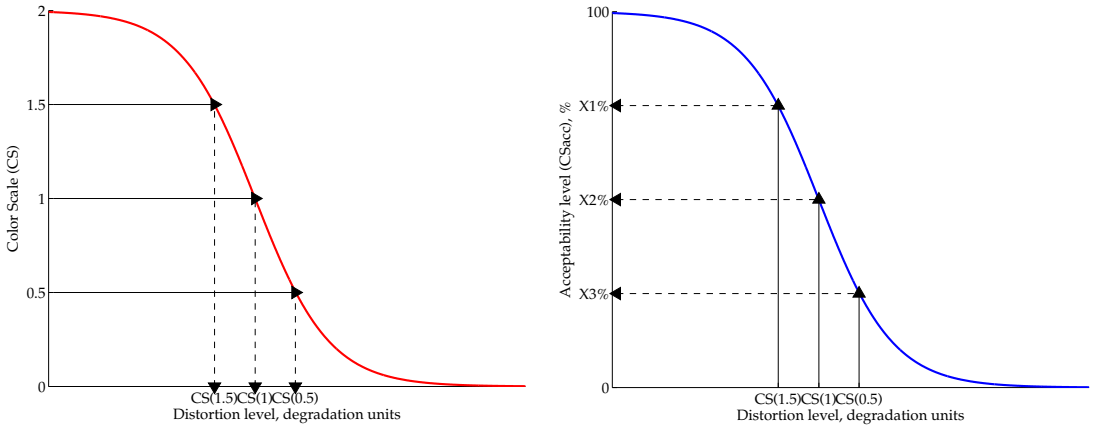


Figure 7.6: Comparison of perceptual thresholds obtained with different methods.

Figure 7.6 demonstrates that thresholds obtained with different methods are similar. Thus, it was confirmed that a 50% annoyance threshold is statistically equivalent to a score of 1.5 on the color scale and 50% acceptability threshold to a score of 0.5. However, how the acceptability and visual annoyance levels change within the CS was not investigated.

Therefore, the degradation levels were estimated with corresponding scores of 1.5, 1, and 0.5 from the CS curve. Then, the corresponding levels (percentages) of acceptability and annoyance on the CSacc and CSann curves were obtained with the Color Scale decomposition method (Section 6.3.2.1). This procedure is illustrated in Figure 7.7.

The results of the CS decomposition are presented in Table 7.4. In the table, the average scores (avg) define the percentage of acceptability and annoyance independent of the asymmetry type. The results for average scores are shown in Figure 7.8.



(a) Estimation of distortion levels from the CS curve

(b) Estimation of x% of acceptability for calculated distortion levels CS(1.5), CS(1), CS(0.5)

Figure 7.7: Data mapping from color scale to approximated acceptability curve.

Table 7.4: Decomposition of the Color Scale (CS) to the acceptability (CSacc) and annoyance (CSann) scales

Asymmetry Px	vertical	rotation	focal	green	white	avg
CS	**CSacc: Acceptable for, %**					
2	100					
1.5	0.97	0.9	0.92	0.99	0.95	0.94
1	0.87	0.75	0.81	0.84	0.83	0.82
0.5	0.6	0.5	0.56	0.56	0.53	0.55
0	0					
CS	**CSann: Visually annoying for, %**					
2	0					
1.5	0.45	0.51	0.52	0.46	0.5	0.49
1	0.78	0.83	0.88	0.75	0.83	0.81
0.5	0.94	0.96	0.98	0.92	0.96	0.95

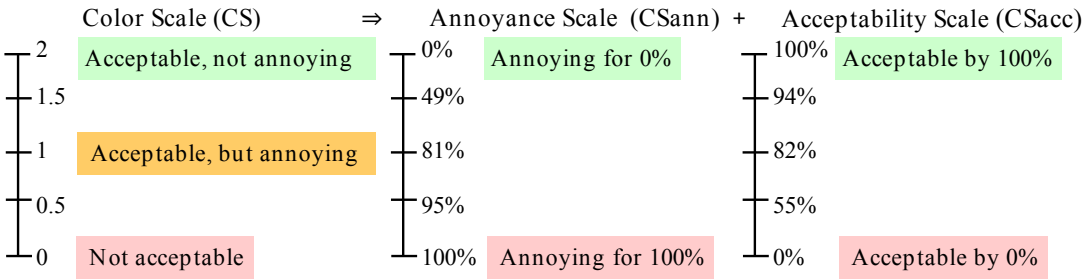


Figure 7.8: Decomposition of the Color Scale to acceptability and visual annoyance.

Furthermore, the average values from Table 7.9 were used to approximate the acceptability and visual annoyance curves to facilitate the definition of the boundaries for any level of annoyance or acceptability. The curves were approximated with the Matlab

Curve fitting toolbox using the shape-preserving interpolant fit. The R-square of the approximation for both curves is more than 0.99.

On the CS, the score 1 (acceptable, but annoying) represents a remarkable point. It can be noticed that when acceptability level is 80%, 80% of observers perceive some visual annoyance.

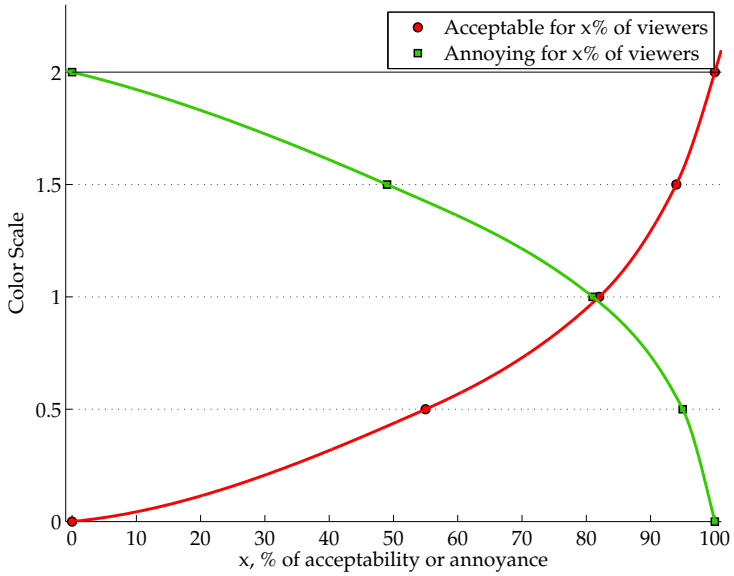


Figure 7.9: Mapping of acceptability and visual annoyance percentage on the Color Scale.

The objective categories are defined by the ranges of distortion, which depend on the selected percentage of acceptability and visual annoyance thresholds. Therefore, by changing the threshold values it is possible to adapt the objective metric to suit any requirement. For example, 80% acceptability is an attractive value for industrial purposes because it can guarantee an optimal solution for the customers. Additionally, the slope of the curve at 50% acceptability or visual annoyance is very steep due to the shape of the logistic function. This produces high variation in acceptability threshold values. For example, the red category of the CS will expand as illustrated in Figure 7.10 if the acceptability threshold is set at 80%.

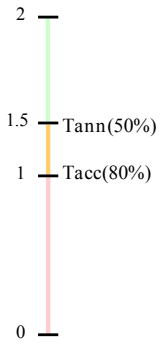


Figure 7.10: Adjustment of the CS by setting a threshold of 80% acceptability on the CS.

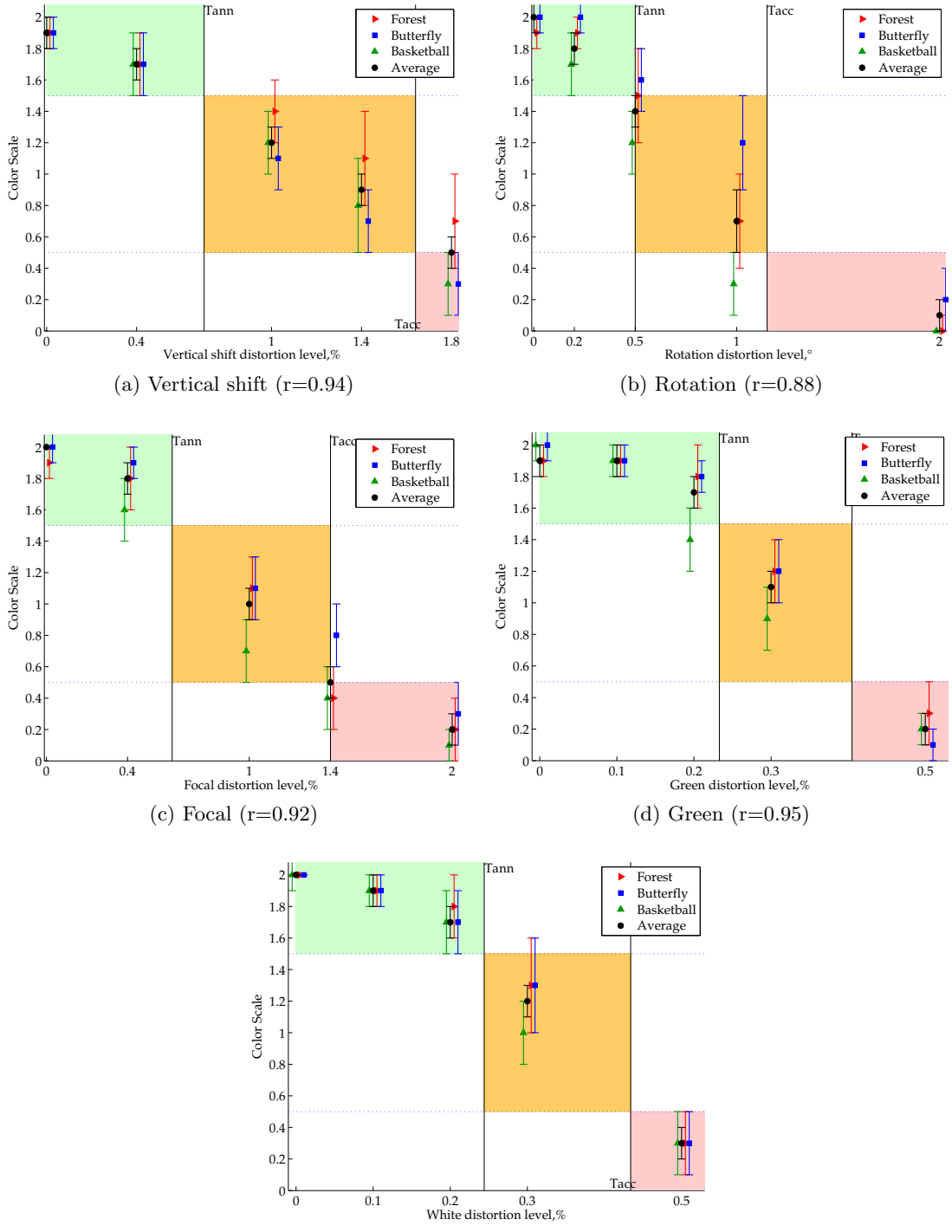
Conclusion: $Tacc_{CS(0.5)} \Leftrightarrow Tacc_{CSacc(0.5)}(50\%)$ and $Tann_{CS(1.5)} \Leftrightarrow Tann_{CSann(0.5)}(50\%)$. It was demonstrated that 50% acceptability and 50% annoyance correspond to scores of 0.5 and 1.5 of the boundaries of the subjective categories. Besides, the decomposition of the CS to acceptability and visual annoyance components permits an adjustment of the boundaries of objective categories in accordance with user requirements.

7.2.4 Result analysis of the "Color Scale" experiment

Figure 7.11 illustrates five plotted graphs representing MOS scores with 95% confidence intervals for five view asymmetries for each scene. These plots allow a direct comparison between subjective results and objective predictions as explained in Section 6.3.5. The boundaries of the objective categories (Tacc and Tann) are plotted from Table 7.3 as a single vertical line representing an estimation from the mean curve. The MOS that do not match the objective predictions are outside the bounds of the color rectangles.

Pearson's correlation coefficients were calculated between the subjective categories for the MOS score and the objective predictions (see Section 6.3.5 for the explanation) for focal (r=0.92), vertical shift (r=0.94), rotation (r=0.88), green (r=0.95), and white (r=1) asymmetries and also for the scenes "Forest" (r=0.9), "Butterfly" (r=1), and "Basketball" (r=0.9).

Conclusion: Generally, for all the asymmetries, high correlations indicate that the OPSM metric performs robustly taking into account that the thresholds obtained with CS were averaged for all the scenes. However, the performance of the metric should be evaluated with the thresholds provided by other studies.



(a) Vertical shift (r=0.94)

(b) Rotation (r=0.88)

(c) Focal (r=0.92)

(d) Green (r=0.95)

(e) White (r=1)

Figure 7.11: Subjective scores of the "Color Scale" experiment versus objective predictions for 5 types of view asymmetries of three scenes "Forest" (r=0.9), "Butterfly" (r=1), and "Basketball" (r=0.9). Tann, Tacc are thresholds estimated from the color scale (Table 7.3).

7.3 Thresholds comparison

In the previous section, the OPSM metric was validated using the thresholds obtained in the same experiment with the CS. However, a comparison of these thresholds with state-of-the-art results is required to validate the CS usage. For this purpose, the thresholds reported by Chen [Chen, 2012](see Chapter 9 p.146) which were obtained in an experiment with the same stimuli as in the "Color Scale" experiment were used. In his experiment, 50% visual annoyance thresholds ($Tann_{IS(3.5)}$) were obtained for various view asymmetries using a score 3.5 on the Impairment Scale (IS). The acceptability thresholds ($Tacc_{Chen}(50\%)$) were derived from the visual comfort scores with the method explained in Section 3.5.1.2. These thresholds are presented in Table 7.5.

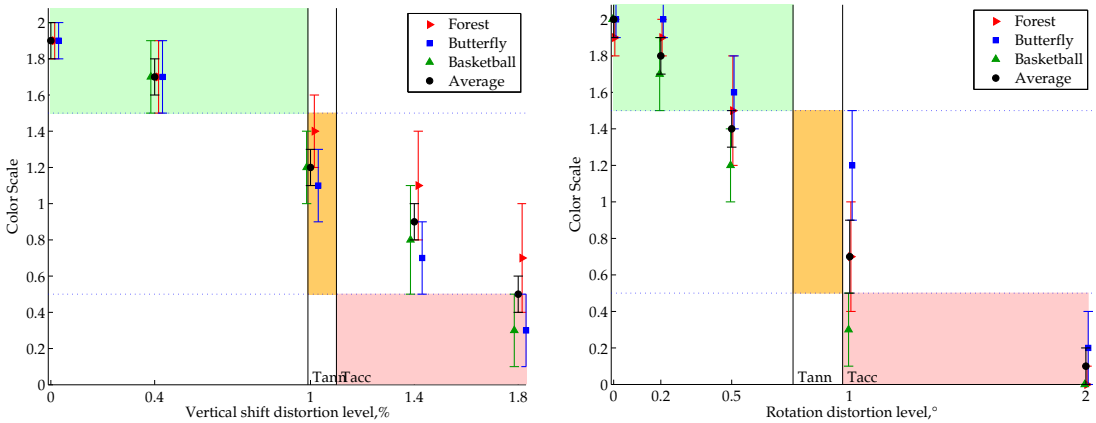
Table 7.5: Visual annoyance and acceptability thresholds from [Chen, 2012] with tolerance range (TR).

Asymmetry Px	Vertical shift,%	Rotation,°	Focal,%	Green,%	White,%
$Tann_{IS(3.5)}$	**0.99**	**0.76**	**0.94**	**0.22**	**0.3**
±TR	0.09	0.1	0.1	0.02	0.002
$Tacc_{Chen}(50\%)$	**1.1**	**0.97**	**1.01**	**0.28**	**0.311**
±TR	0.15	0.14	0.13	0.03	0.032

Figure 7.12 illustrates five plotted graphs representing the MOS scores collected in the "Color Scale" experiment with a 95% confidence intervals for the comparison with the Figure 7.11. The boundaries of the objective categories (Tacc and Tann) are plotted from Table 7.5. These plots allow a direct comparison between subjective results and objective predictions as explained in Section 6.3.5. The MOS that do not match the objective predictions are outside the bounds of the color rectangles.

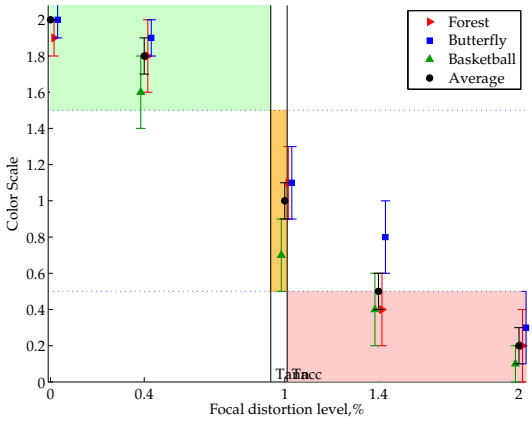
Pearson's correlation coefficients were calculated between the subjective categories for the MOS score and the objective predictions (see Section 6.3.5 for focal (r=0.96), vertical shift (r=0.88), rotation (r=0.85), green (r=0.86), and white (r=1) asymmetries and also for the scenes "Forest" (r=0.88), "Butterfly" (r=0.91), and "Basketball" (r=0.88).

The correlation values between the subjective results and the objective predictions as well as the width of the "Orange" categories for all the plots decreased when using Chen's thresholds. Therefore, a more detailed comparison of the thresholds from the CS (Table 7.3) and from Chen's thesis (Table 7.5) is presented in Figure 7.13.

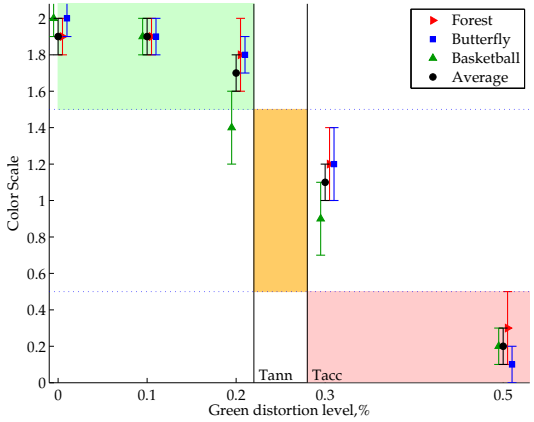


(a) Vertical shift (r=0.88)

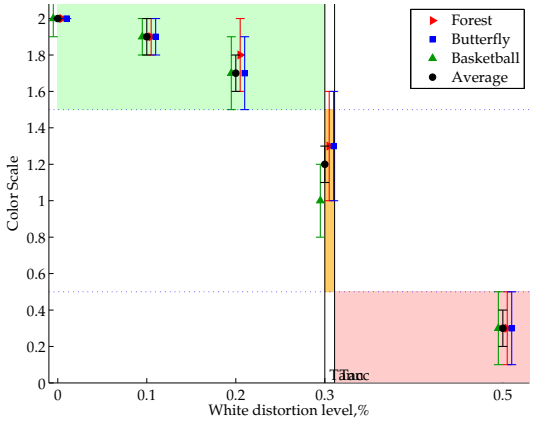
(b) Rotation (r=0.85)



(c) Focal (r=0.96)



(d) Green (r=0.86)



(e) White (r=1)

Figure 7.12: Subjective scores of the "Color Scale" experiment versus objective predictions for 5 types of view asymmetries of three scenes "Forest" (r=0.88), "Butterfly" (r=0.91), and "Basketball" (r=0.88). Tann, Tacc are thresholds from the literature presented in Table 7.3.

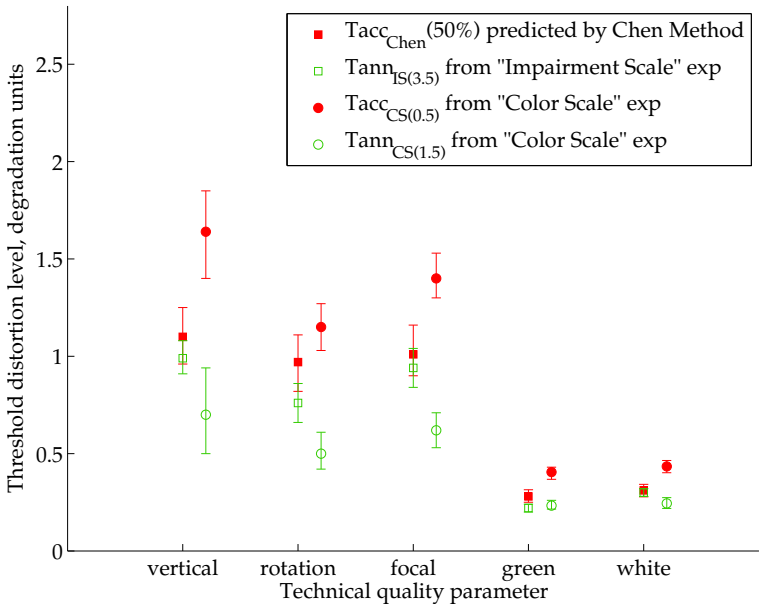


Figure 7.13: Comparison of acceptability and visual annoyance thresholds from the "Color Scale" experiment and state-of-the-art.

Several conclusions can be drawn from Figure 7.13:

- The visual annoyance ($Tann_{IS(3.5)}$) and acceptability ($Tacc_{Chen}(50\%)$) thresholds obtained by Chen are intersecting for all the asymmetries considering the confidence intervals. It means that the "Orange" category overlaps with both the "Red" and "Green" categories at the same time. This explains why the "Orange" categories in Figure 7.12 are so narrow for all of the asymmetry types.
- The visual annoyance ($Tann_{CS(1.5)}$) and acceptability ($Tacc_{CS(0.5)}$) thresholds obtained on the color scale do not intersect, which allows the "Orange" category to be clearly defined.
- $Tann_{IS(3.5)} > Tann_{CS(1.5)}$: the annoyance thresholds obtained with the impairment scale (IS) represent higher degradation levels than those obtained with the color scale (CS) for all types of asymmetries except green. Presumably, this is the result of a difference in the designation of the categories. In the case of the IS, the annoyance threshold is situated between the labels "Perceptible, but not annoying" and "Slightly annoying", e.g. a grade of 3.5. While for the CS, it is placed between the "Acceptable, not annoying" and "Acceptable, but annoying" labels, e.g. grade 1.5., implying that the upper bound does not allow any visual annoyance. However, the IS category "Perceptible, but not annoying" can be interpreted dubiously by observers allowing some slight degree of visual annoyance.
- $Tacc_{Chen} < Tacc_{CS(0.5)}$: the acceptability thresholds obtained with Chen's method are more rigorous than those obtained with the color scale (CS) for all types of asymmetries. A possible explanation for such a difference is that $Tacc_{Chen}$ thresholds were extracted from the visual comfort scores and not from a subjective test.

Conclusion: it is necessary to verify if the 50% acceptability thresholds derived from Chen's method will match acceptability thresholds obtained with dedicated subjective experiment on acceptability.

7.3.1 "Acceptability Scale" experiment

The goal of the following subjective experiment is to assess the subjective acceptability thresholds for five types of asymmetries (vertical shift, rotation, magnification, green, and white) with four levels of distortion as described in Table 7.2. Then the results of the test should be compared with the acceptability thresholds obtained using Chen's method (Section 3.5.1.2).

The test set-up, stimuli, and methodology were the same as previously described in Section 7.2.2. 29 observers participated in the experiment. The instruction sheet can be found in Appendix C.2. T
on the *Acceptability Scale* ,

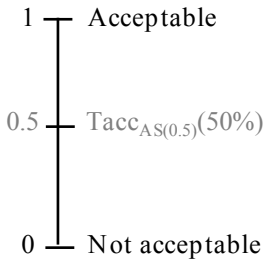


Figure 7.14: The Acceptability Scale (AS) used in the "Acceptability Scale" experiment

MOS scores with a binomial probability confidence interval were computed from collected votes [Clopper and Pearson, 1934, Soper, 2014]. Following recommendation ITU-R BT.500-13, the relationship between the MOS and the distortion levels for each type of view asymmetry was approximated, which allowed the estimation of a distortion level for a desired percentage of acceptability. The symmetry logistic function was used to obtain this continuous relationship following the equation 6.2.

The Matlab Curve fitting toolbox was used to compute the approximation and draw the curves. The R-square of the approximation for each asymmetry was more than 0.98. Thus, based on the R value, all fits can be considered reliable. A 50% acceptability threshold was estimated as a grade of 0.5 from the mean curve. The tolerance range was defined from the minimum and maximum curves obtained as the lower and upper bounds of the MOS with the confidence intervals [ITU, 2012b]. The example of such approximation is presented in Figure 7.15 for the rotation asymmetry. Similarly any other percentage of acceptability can be estimated from the mean curve.

All the obtained 50% acceptability thresholds with tolerance ranges are presented in Figure 7.16. They are compared to the thresholds estimated from the visual comfort test scores by Chen [Chen, 2012]. As can be seen from the plot, the thresholds obtained in the subjective experiment match Chen's thresholds for all type of asymmetries except rotation. Despite the expectations, the rotation acceptability threshold decreased in the subjective test for 0.2°.

Overall, the new acceptability thresholds do not change the objective predictions analyzed in Section 7.2.4 since their variation is insignificant considering the tolerance ranges.

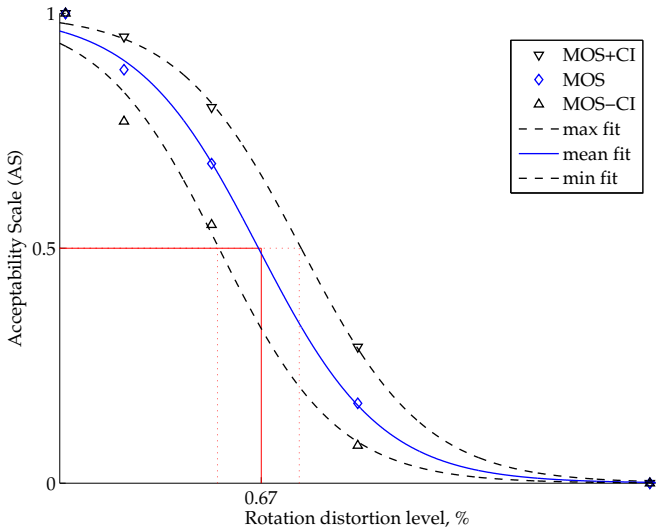


Figure 7.15: Acceptability Scale (AS) curve and two curves representing 95% confidence interval approximated for the rotation asymmetry from MOS scores

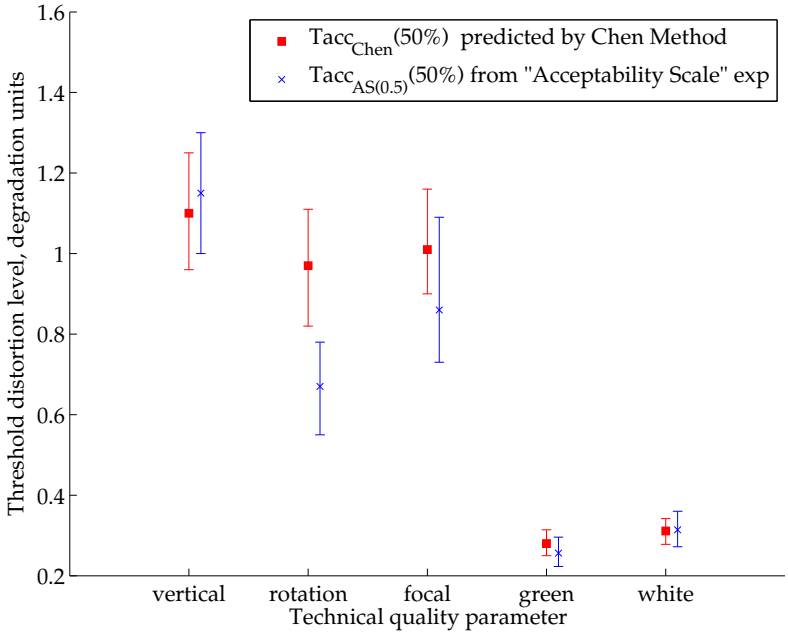


Figure 7.16: Comparison of acceptability thresholds.

Conclusion: $Tacc_{Chen}(50\%) \Leftrightarrow Tacc_{AS(0.5)}(50\%)$. The new acceptability thresholds evaluated with the acceptability scale are equivalent to those derived with Chen's method. However, such results do not clarify why the acceptability thresholds assessed on the CS ($Tacc_{CS(0.5)}(50\%)$) are different from those assessed on the AS ($Tacc_{AS(0.5)}(50\%)$). A possible explanation is that it may be quite difficult for observers to judge two different perceptual criteria on the same scale at the same time. In order to verify this hypothesis, a new subjective test is designed in the next section.

7.4 Methodology development. "Double Scale" experiment

The results of the previous section demonstrated that the acceptability thresholds obtained with the color scale and the acceptability scale are not the same. It is presumed that it may be too complicated to judge two different perceptual criteria at the same time. Therefore, the goal of the "Double scale" experiment is to verify whether acceptability can be evaluated at the same test as visual annoyance. It is suspected that observers accept higher degradation levels with the CS than with the AS, e.g. "Acceptable, but annoying" was preferred to "Not acceptable". So, a new subjective test was designed that resembles the conventional acceptability test with the AS but at the same time keeping all the categories of the CS. For this reason, the two scales were used in this experiment as illustrated in Figure 7.17, a screenshot of the test interface can be found in Figures D.2 and

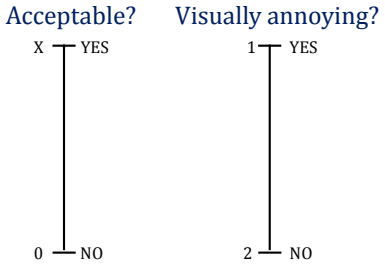


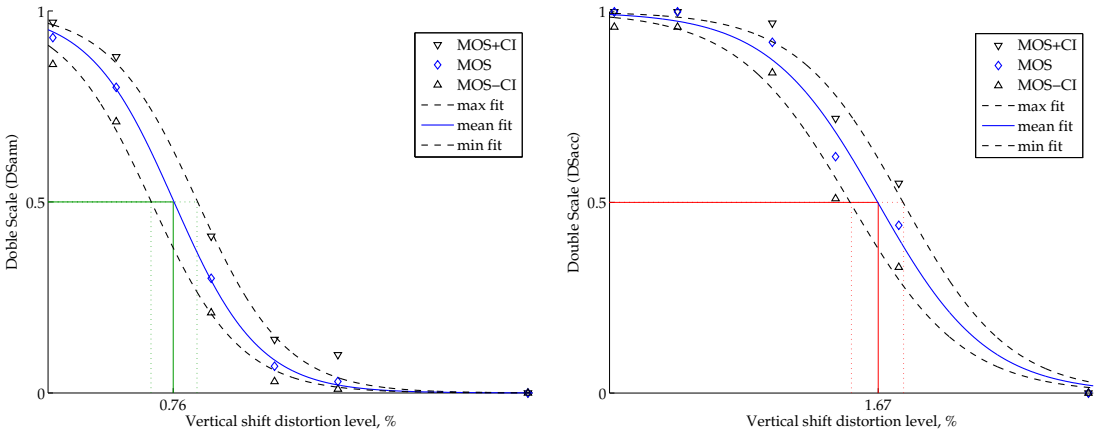
Figure 7.17: Double Scale (DS) experiment.

At the beginning of a sequence evaluation, only the acceptability scale appears on the screen. So firstly, it is proposed to decide on the acceptability of the sequence by simply answering "yes" or "no". If the answer is "no", a score of 0 ("Not acceptable") is recorded and the subject can pass to the evaluation of another sequence. Otherwise, the second scale appears on the screen. Secondly, the subject has to determine if the stimulus is visually annoying or not. If "yes" is chosen, then a score of 1, corresponding to "Acceptable, but annoying", is stored; otherwise, 2 – "Acceptable, not annoying".

The test set-up, stimuli, and methodology are the same as described in Section 7.2.2. 33 observers participated in the experiment. The instruction sheet can be found in Appendix C.3. The SAMVIQ protocol was used to evaluate the sequences with the *Double Scale (DS)* illustrated in Figure 7.17.

7.4.1 Result analysis of "Double Scale" experiment

The MOS scores collected in the "Double Scale" experiment with a 95% confidence interval were computed for all distortion levels of the five view asymmetries. Then, 50% visual annoyance and acceptability thresholds were calculated from the approximated DSann and DSacc curves for all view asymmetries (the same principle was used as for the CSann and CSacc curves in Section 6.3.2.1). The example of such estimation is illustrated in Figure 7.18 for the vertical shift. The thresholds for all the types of asymmetries are in Table 7.6.



(a) 50% annoyance threshold (b) 50% acceptability threshold

Figure 7.18: The example of thresholds estimations for the vertical shift.

Table 7.6: 50% visual annoyance and visual annoyance thresholds calculated from the votes collected in the "Double Scale" experiment with tolerance range (TR).

Asymmetry type	Vertical shift,%	Rotation,°	Focal,%	Green,%	White,%
$Tann_{DSann(0.5)}$	**0.76**	**0.42**	**0.58**	**0.208**	**0.24**
TR (upper limit)	0.15	0.14	0.15	0.032	0.036
TR (lower limit)	0.14	0.16	0.13	0.034	0.038
$Tacc_{DSacc(0.5)}$	**1.67**	**1.02**	**1.47**	**0.41**	**0.447**
TR (upper limit)	0.16	0.18	0.22	0.035	0.03
TR (lower limit)	0.17	0.13	0.23	0.04	0.041

Similar to the CS (see Section 7.2.3), the degradation levels corresponding to scores of 1.5, 1, and 0.5 from the DS curve were estimated. Graphically, the results of the CS decomposition are presented in Figure 7.19. No significant difference was found between

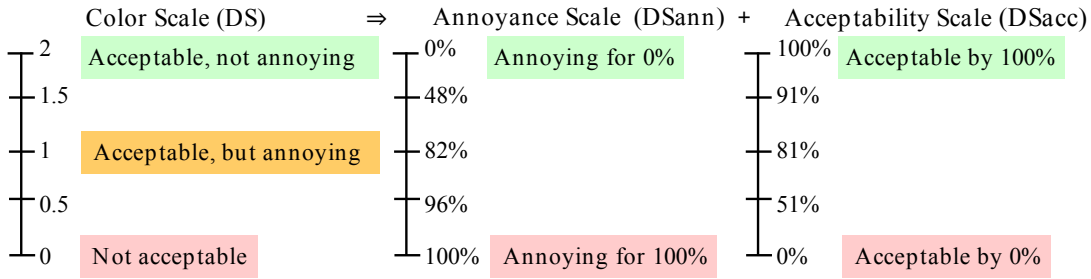


Figure 7.19: Decomposition of the data from the "Color Scale" experiment to acceptability and visual annoyance.

Figure 7.12 illustrates five plotted graphs representing MOS scores with a 95% confidence interval for five view asymmetries for each scene. The boundaries of the objective categories (Tacc and Tann) are plotted from Table 7.6 as a single vertical line representing the estimation from the mean curve. The MOS that do not match the objective

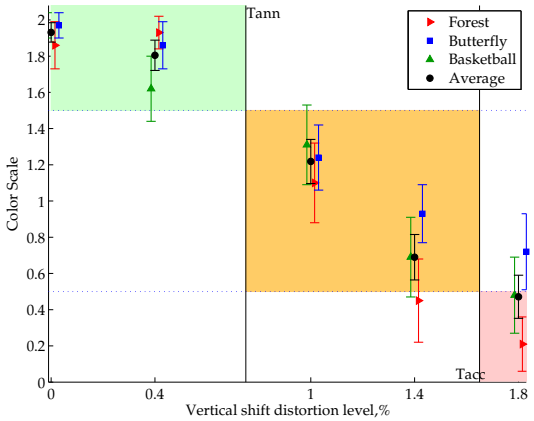
predictions are outside the bounds of the color rectangles. Pearson's correlation coefficients were calculated between the subjective results and the objective predictions for focal (r=0.95), vertical shift (r=0.88), rotation (r=0.92), green (r=0.91), and white (r=1) asymmetries. Also for the scenes "Forest" (r=0.95), "Butterfly" (r=0.9), and "Basketball" (r=0.94).

The one way ANOVA analysis demonstrated that the distortion level had a significant ($p < 0.0001$) impact on the MOS scores of all types of asymmetry. The effect of the scene was found to be significant for focal ($p < 0.03$), vertical shift ($p < 0.03$), and green ($p < 0.02$) asymmetries. For focal asymmetry, the slightly higher votes for the "Butterfly" scene can be explained by the fact that the objects of interest, e.g. the bunny's face and the butterfly, were located in the center of the image where vertical disparities are less pronounced than in the corners of an image. Concerning green asymmetry, the significance can be explained by the color gamut of the scene. The "Butterfly" and "Bunny" scenes are mostly composed of green colors making such impairment less visible. Therefore, for the estimation of perceptual thresholds, it is important to consider content of different complexity, color gamut, and luminance.

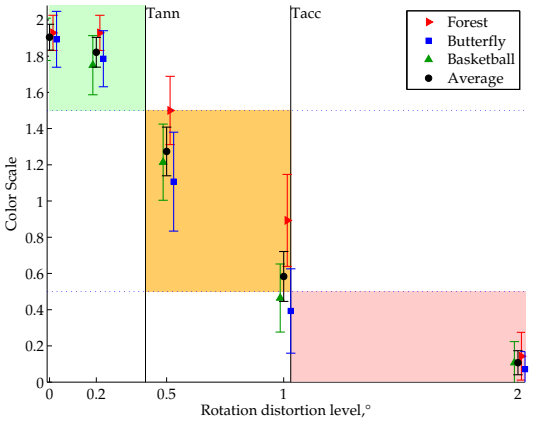
Interestingly, in the case of the "Color Scale" experiment for the same content effect of a scene, only rotation asymmetry was found to be significant. The p-values in both cases are quite close to the threshold of 0.05. A comparison of the two tests is presented in Figure 7.21 to show the difference. The boundaries of the objective categories are plotted from Table 7.6.

The results of the "Double Scale" experiment are the same as the "Color Scale" experiment considering 95% confidence intervals in Figure 7.21. This was confirmed with an F-test (two samples for variances): the difference in MOS scores between the data sets of the "Color Scale" and "Double Scale" experiments is insignificant for all five view asymmetries. Hence, the usage of the two scales in the same test did not improve the methodology of the Color Scale.

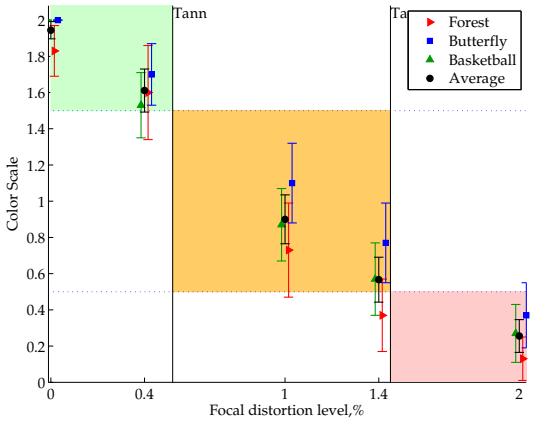
Conclusion: the test interface of the "Double Scale" experiment had no influence on the results, considering that the difference in MOS scores between the data sets of the "Color Scale" and "Double Scale" experiments was found to be insignificant with an F-test (two samples for variance). Thus, this experiment did not clarify why the acceptability thresholds obtained with the CS and AS are different. Therefore, in the next section this issue is investigated by comparing different scales.



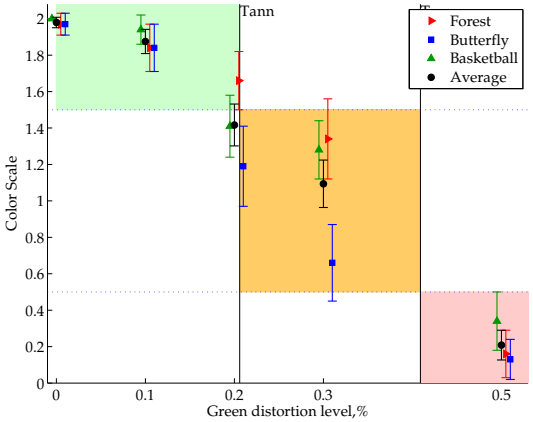
(a) Vertical shift (r=0.88)



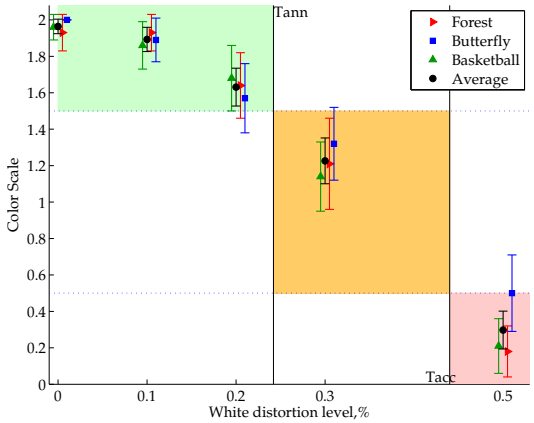
(b) Rotation (r=0.92)



(c) Focal (r=0.95)

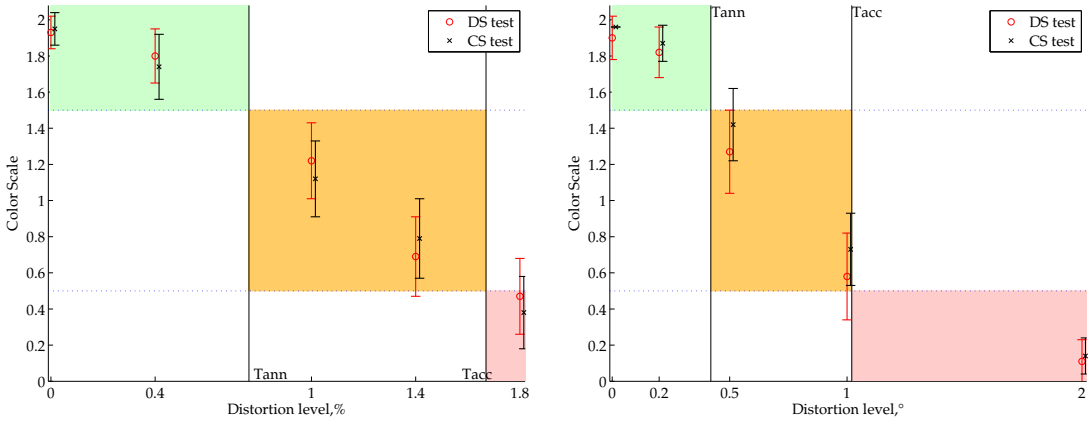


(d) Green (r=0.91)



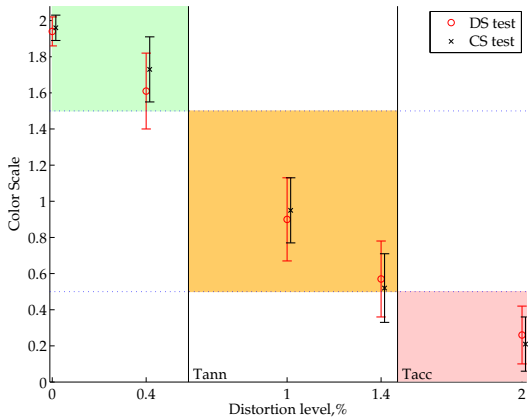
(e) White (r=1)

Figure 7.20: Subjective scores of the "Double Scale" experiment versus objective predictions for 5 types of view asymmetries of three scenes "Forest" (r=0.95), "Butterfly" (r=0.9), and "Basketball" (r=0.94). Tann, Tacc are thresholds from Table 7.6.

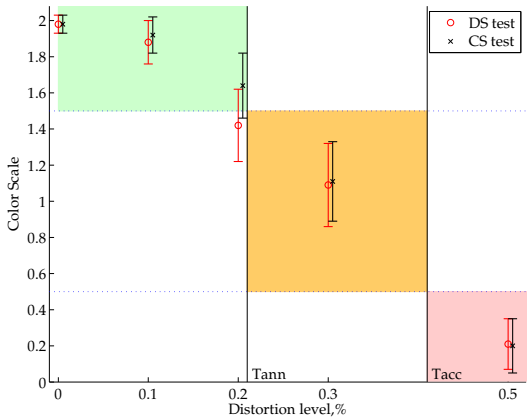


(a) Vertical shift

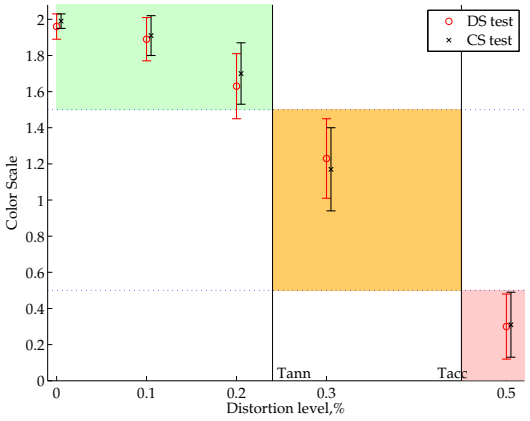
(b) Rotation



(c) Focal



(d) Green



(e) White

Figure 7.21: A comparison of the "Color Scale" and "Double Scale" experiments (average MOS scores for the content). Tann, Tacc are thresholds from Table 7.6.

7.5 Comparison of Color Scale with Acceptability and Impairment Scales

The goal of following analysis is twofold. Firstly, we aim to understand why there is a difference in the assessment methodologies in terms of perceptual thresholds, which was illustrated in Figure 7.13. Secondly, we compare impairment, acceptability, and color scales in terms of visual annoyance and acceptability. Visibility and visual annoyance thresholds as scores of 3.5 and 4.5 on the impairment scale (IS) are estimated from the mean curve of Chen's experimental data [Chen, 2012]. Then obtained distortion values are used to compute the corresponding grades on the color scale curve (CS) and on the visual annoyance (CSann) and acceptability (CSacc) curves (see Section 6.3.2.1) derived from it for five view asymmetries. 20%, 50%, and 80% acceptability thresholds (AS) from the "Acceptability Scale" experiment are estimated in the same manner for five view asymmetries. The result of this mapping is presented in Table 7.7 as average of all view asymmetries. The resulting values for each view asymmetry are presented in Annex B.1.

Table 7.7: Impairment and acceptability thresholds mapped on the color scale, color scale annoyance (CSann), and color scale acceptability (CSacc) with confidence interval (CI) as average for five view asymmetries.

Scale	CS	±CI	CSacc	±CI	CSann	±CI
IS(4.5)	1.86	0.06	0.99	0.01	0.11	0.05
IS(3.5)	**1.24**	0.16	**0.88**	0.06	**0.69**	0.14
AS(0.8)	1.59	0.1	0.95	0.02	0.4	0.08
AS(0.5)	**1.2**	0.11	**0.88**	0.03	**0.71**	0.08
AS(0.2)	0.75	0.07	0.71	0.04	0.9	0.04

The illustration of such mapping for 50% acceptability threshold on the AS (bold

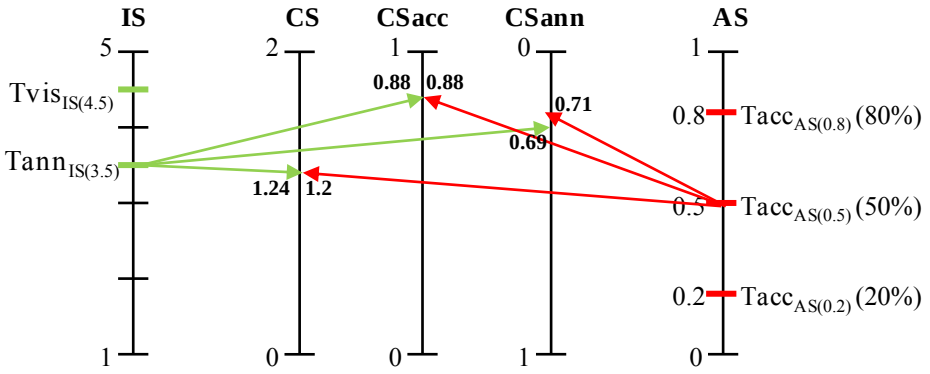


Figure 7.22: A comparison of the impairment and acceptability scales with the color scale and its derivatives.

Several conclusions can be reached from the results:

- IS(3.5)= CS(1.24). This means that 70% of viewers experience visual annoyance at a score of 1.24 on the color scale. So, presumably $Tann_{IS(3.5)}$ does not represent

a threshold of 50% visual annoyance. It confirms the comparison of the thresholds in Figure 7.13, where $Tann_{IS(3.5)} > Tann_{CS(1.5)}(50\%)$.

- IS(3.5) = AS(0.5). The distortion level of these thresholds is similar considering the confidence intervals. Thus, the "Orange" category disappears if these threshold are taken as the boundaries of the objective categories for OPSM.
- AS(0.8) = CS(1.6). This observation implies that when the subjects assess acceptability in our acceptability test, the criteria of acceptance was visual annoyance. In other words, acceptability is a high level concept, where QoE is evaluated. Hence, the threshold AS(0.8) can be used to define the simplified color model, which only consists of the "Green" and "Red" categories as illustrated in Figure 7.23.
- AS(0.2) = CS(0.72). A degradation level representing 20% acceptability on the AS corresponds to 71% acceptability on the CS, while visual annoyance on the CS is 90%. Hence, at this level almost all observers perceive visual annoyance and accept it. Therefo
certain level of visu

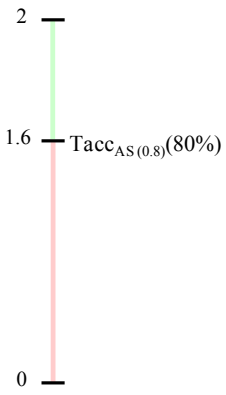


Figure 7.23: Adjustment of CS by setting a threshold $Tacc_{AS(0.8)}$ of 80% acceptability on the CS.

7.6 Conclusions

In this chapter the OPSM metric evaluating visual comfort was validated with three subjective experiments. It was demonstrated that:

- The thresholds of acceptability and visual annoyance can be obtained directly with the color scale. However, these thresholds are not the same when evaluated with standard methodologies.
- Supposedly, the difference in the acceptability thresholds values can be explained by the different evaluation concepts. In the case of the color scale, a subject evaluates the acceptability of the visual annoyance level, while in the case of the acceptability test the quality of the viewing experience is evaluated. In addition, it seems that the main criterion to decide on acceptance in that case is visual annoyance, which can be explained by the instruction sheet of the "Acceptability" experiment (Appendix C.2).

- The threshold of 80% acceptability on the AS can be used to construct a simplified OPSM metric considering visual comfort, which consists only of the "Green" and "Red" categories. This is possible because Tacc(80%) was found to be equal to the 45% visual annoyance threshold on the CS.
- The color metric was validated using the thresholds defined by the color scale with a minimum correlation coefficient of r=0.88 for rotation asymmetry.

In this chapter the OPSM metric was evaluated using still stereoscopic images. In the next chapter the aim is to test the same methodology on video sequences.

Chapter 8

Metric verification with stereoscopic videos

Contents

8.1 Introduction

A new OPSM metric to predict perceptual states of viewers related to stereoscopic content was proposed in Chapter 6. The proposed metric was validated using stereoscopic still images in Chapter 7. The first goal of this chapter is to verify the metric with stereoscopic video sequences as well as compare perceptual thresholds for still and moving images.

8.2 OPSM metric verification with S3D videos

8.2.1 Stimuli generation

Four stereoscopic video sequences with different levels of complexity were selected for the experiment (see Fig. 8.1). The non-compressed scenes were captured with the Panasonic

3D AG-3DA1 camcorder with two sensors and 60 mm as the fixed baseline distance. This camera was selected because both sensors were aligned at the manufacturing stage, so it is expected not to have major problems with geometrical, luminance, or color asymmetries. The restitution of depth is well adapted for indoor scenes (see the depth analysis in Fig. 2.5). Therefore, all scenes except "Alley" feature indoor space (the distance between foreground and background does not exceed 10m). The camera parameters for each scene are presented in Table 8.1, where f – focal length, $dCon$ – convergence distance, fg – foreground distance, e.g. distance from the camera to the closest object, bg – background distance, e.g. distance from the camera to the farthest object, DR - disparity range of the pixels on the screen in mm, and DoF – depth of focus.

(a) Alley (b) Interview

(c) Kitchen (d) Picnic

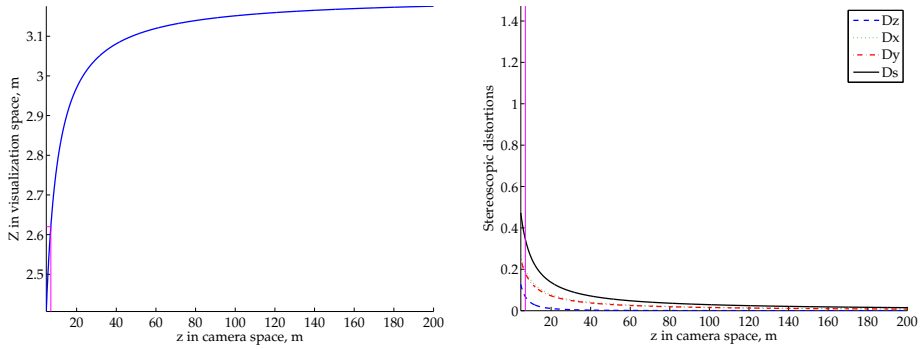
Figure 8.1: Video sequences.

Table 8.1: Scene and camera parameters

Scene	f, mm	dCon, m	fg, m	bg, m	roi, m	DR, mm	DoF, D
Alley	4.3	4.5	4.7	200	6	[0;21]	0.13
Interview	5.5	2.14	2.2	3	2.5	[0;16]	0.1
Kitchen	4.3	2.8	3.3	6.1	4	[0;16]	0.1
Picnic	4.8-8.3	2.96	3.1	10.4-5.5	5	[0;25]	0.15

The relationship between the camera and the visualization spaces for all scenes is presented in Figures 8.2- 8.5. The camera space distance z is limited by the background and foreground of the scene. The location of the object of interest is marked as a single vertical magenta line. The limitation of the camera is clearly demonstrated in the case of the "Alley" scene in Figure 8.2.a: all objects further than 40m are perceived at a distance of 3.1m in the visualization space. Whereas, the scene "Interview" in Figure 8.3 represents an almost ortostereoscopic condition: the distance in camera space

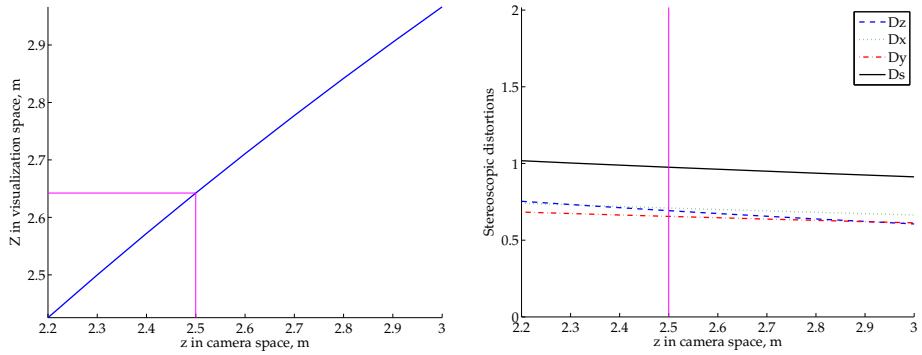
is approximately equal to the distance in the visualization space and the shape distortion coefficient around the object of interest is almost 1. This means that shapes of the faces were preserved in the visualization space. The "Kitchen" scene in Figure 8.4 is compressed in the visualization space in comparison with the camera space. Figure 8.5 illustrates the change in depth restitution before and after zoom in the "Picnic" scene. The shape distortion of the object of interest (Ds) has changed from 0.5 to 0.7, e.g. became less compressed after zoom in the visualization space.



(a) Visualization space Z vs. camera space z

(b) Stereoscopic distortions vs. z in camera space

Figure 8.2: Alley

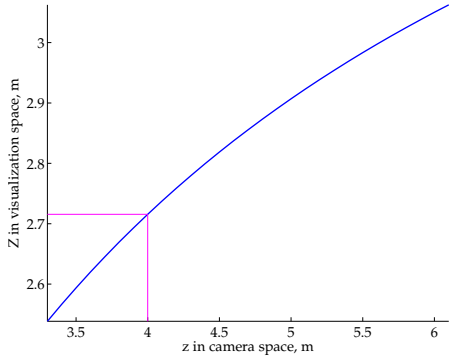


(a) Visualization space Z vs. camera space z

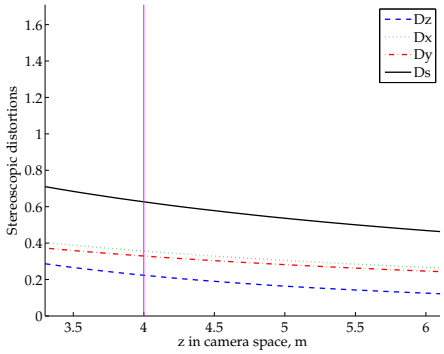
(b) Stereoscopic distortions vs. z in camera space

Figure 8.3: Interview

Vertical shift, magnification, green color, and white luminance view asymmetries with four levels of distortion as shown in Table 7.2 were generated. The rotation asymmetry was not taken into account in this experiment since the view magnification asymmetry produces a similar artifact: vertical disparity is accumulated in the borders of an image. The distortions were created with Virtual Dub software as explained in Section 7.2.1. The exception was the vertical shift: one view was shifted and then black mask was superimposed on top of both views to hide the difference and keep the initial image height.

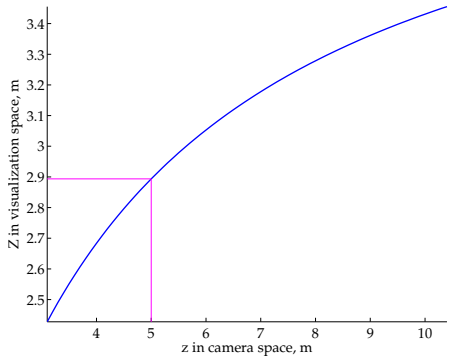


(a) Visualization space Z vs. camera space z

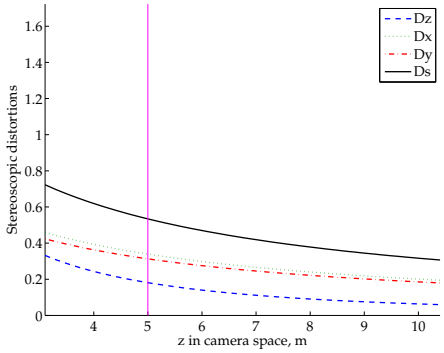


(b) Stereoscopic distortions vs. z in camera space

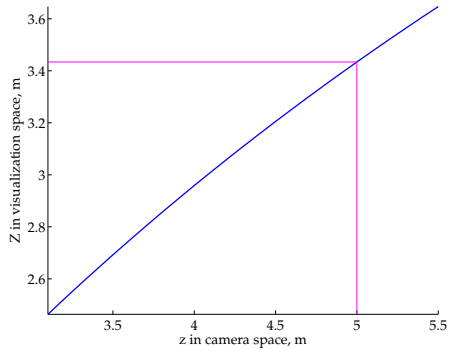
Figure 8.4: Kitchen



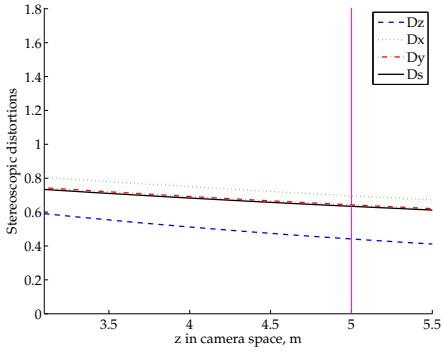
(a) Before zoom: Visualization space Z vs. camera space z



(b) Before zoom: Stereoscopic distortions vs. z in camera space



(c) After zoom: Visualization space Z vs. camera space z



(d) After zoom: Stereoscopic distortions vs. z in camera space

Figure 8.5: Picnic

8.2.2 Experimental set-up and methodology

Test set-up: the subjective experiment was performed in the test room in compliance with recommendation ITU-R BT.2021. The Hyundai 46" line interleaved stereoscopic display was used for the visualization of the stimuli. Its dimension is 102×56 cm; resolution in 2D 1920×1080, in 3D 1920×540 per view. The luminance, brightness, contrast

and color of the display were adjusted to a normal gamma function (gamma equals 2.2). The display's color triangle is illustrated in Figure B.1. The crosstalk level was less than 3% and the maximum luminance level measured through glasses was $100cd/m^2$. The viewing distance was 4.5 times the height of the display. An additional Dell 22" LCD display was used to present the test interface and store the votes of observers.

Obsevers: 30 non-expert observers participated in this test. Their monocular acuity, color vision, far vision test, fusion test, and stereoscopic acuity were checked using Essilor ERGOVISION equipment prior to the subjective experiment. All observers had a normal or corrected to normal visual acuity and normal stereoacuity.

Methodology: The instruction sheet presented in Appendix C.4 offered some explanations on how to behave during the experiment and how to rate the sequences. Also, the instructions were explained by the examiner to ensure that observers understood the task. The SAMVIQ protocol was used to evaluate the sequences on the Color Scale described in Table 8.2. The test interface is presented in Figure D.1. The first part of the experiment consisted of 2 tests, where observers assessed 2 types of view asymmetries. To avoid an accumulation of visual discomfort, observers evaluated the second part of the experiment with the remaining 2 asymmetries after 15 minutes pause. The distortion levels were presented in random order.

Table 8.2: Color Scale (categorical)

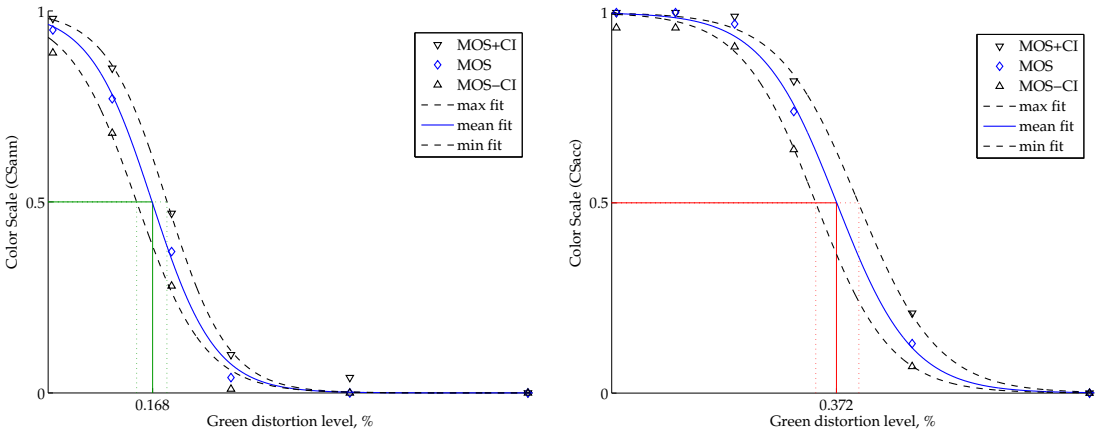
Color Scale (categorical)	
2	Acceptable, NOT annoying
1	Acceptable, BUT annoying
0	NOT Acceptable

In total, every subject had to evaluate 96 stimuli (4 scenes * 4 types of asymmetry * [4 distortion levels + explicit reference + hidden reference]). The visualization time of one stereoscopic pair was 15 seconds. In average it took around 15 minutes for the subject to evaluate one type of asymmetry, e.g. around 1 hour for the whole experiment.

8.2.3 Result analysis

MOS scores with a 95% confidence interval were computed for all distortion levels of the four view asymmetries. The one way ANOVA analysis demonstrated that video content factor was insignificant for the change of MOS scores, while the distortion level had a significant ($p < 0.0001$) impact for all types of asymmetry. Considering that scene content does not influence results, the 50% acceptability ($Tacc_{CSacc(0.5)}$) and 50% visual annoyance ($Tann_{CSann(0.5)}$) thresholds with tolerance ranges were estimated from CS results as an average of the four scenes. The example of such estimation is illustrated in Figure 8.6. The thresholds for all the types of asymmetries are in Table 8.3.

Figure 8.7 shows the MOS scores with confidence intervals of the color scale test versus the distortion levels for all the asymmetries. The boundaries of the objective categories (Tacc and Tann) are plotted from Table 8.3 as a single vertical line representing an estimation from the mean curve. Then the Pearson's correlation coefficients were calculated between the subjective results and objective predictions for focal (r=0.96), vertical shift (r=1), green (r=1), and white (r=0.96) asymmetries as explained in Section 6.3.5.



(a) 50% annoyance threshold (b) 50% acceptability threshold

Figure 8.6: The example of thresholds estimations for the green view asymmetry.

Table 8.3: 50% acceptability and 50% visual annoyance thresholds with tolerance range (TR).

Asymmetry Px	Vertical shift,%	Focal,%	Green,%	White,%
$Tann_{CSann(0.5)}$	**0.71**	**0.71**	**0.168**	**0.18**
TR (upper limit)	0.12	0.10	0.02	0.02
TR (lower limit)	0.10	0.11	0.03	0.02
$Tacc_{CSacc(0.5)}$	**1.57**	**1.61**	**0.372**	**0.35**
TR (upper limit)	0.15	0.15	0.038	0.032
TR (lower limit)	0.16	0.14	0.035	0.04

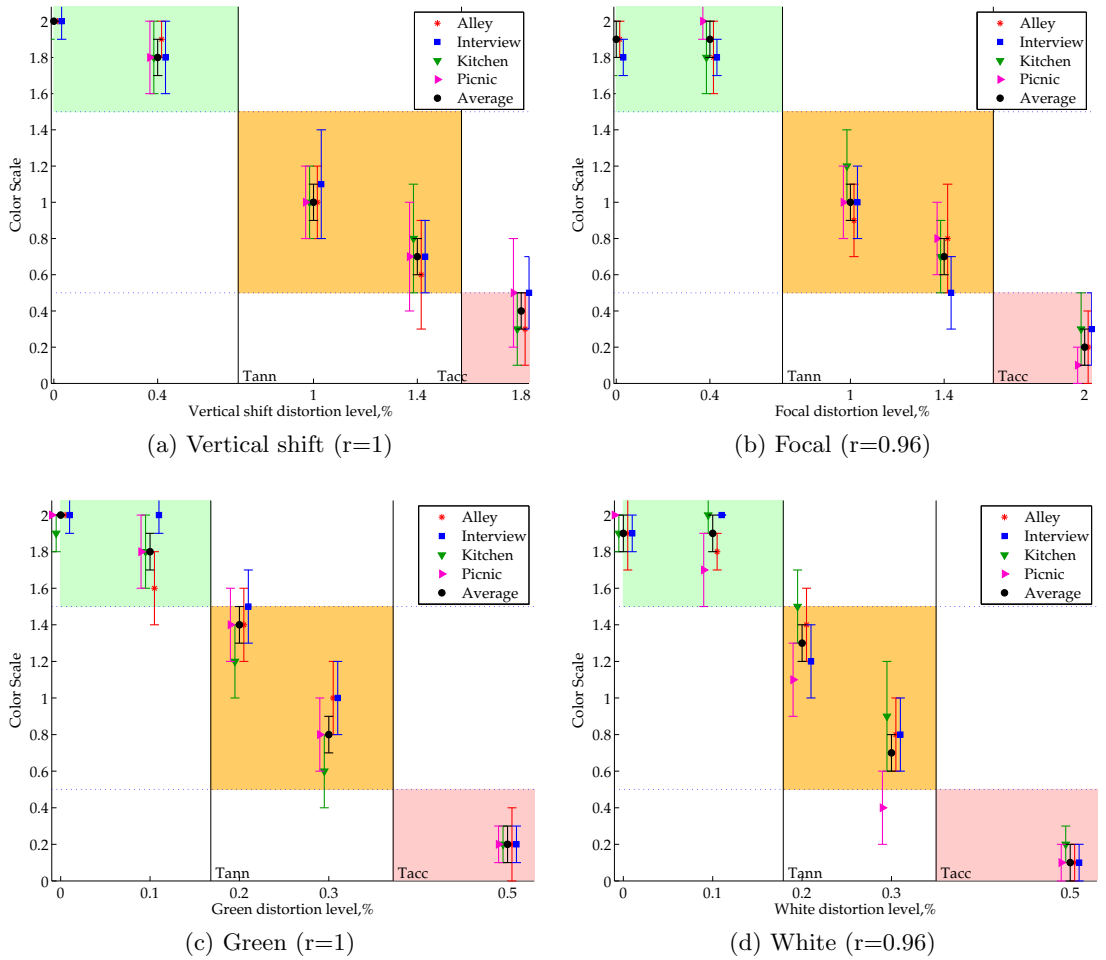
Conclusion: For all the asymmetries, high correlations may indicate that the perceptual objective metric performs robustly taking into account that the thresholds were averaged for all the scenes.

8.2.3.1 Stereoscopic video versus images: thresholds comparison

The color scale metric has shown that there is a high correlation between subjective results and objective predictions. However, all objective predictions used the objective boundaries estimated from the same test. Nevertheless, we believe that perceptual thresholds should be independent of the content and observers. Therefore, firstly it was decided to compare acceptability thresholds obtained from the conventional acceptability test on the acceptability scale (AS). Secondly, we compared acceptability and visual annoyance thresholds for stereoscopic videos and images obtained with the color scale (CS).

The same video sequences were used to evaluate the acceptability on the AS. 30 observers took part in the experiment. The MOS with binomial probability confidence interval were computed from the collected votes and compared to the "Acceptability Scale experiment" data in Section 7.3.1. The result of the comparison is presented in Figure 8.8 for 50% acceptability.

The comparison of thresholds indicate that the scene content (images or videos) did not influence the acceptance of view asymmetries. This was confirmed by a t-test (paired



(a) Vertical shift (r=1)

(b) Focal (r=0.96)

(c) Green (r=1)

(d) White (r=0.96)

Figure 8.7: Subjective scores per scene versus objective predictions for four types of view asymmetry (the Pearson's correlation coefficient r is noted in brackets).

two samples for means), which was conducted to compare difference in MOS_{AS} scores between the images and videos for all the view asymmetries. No significant difference has been found. It demonstrates that acceptability thresholds obtained with the AS for still stereoscopic images are valid for stereoscopic videos as well.

Further comparison of acceptability and visual annoyance thresholds for S3D videos and images is shown in Figure 8.9. Here, the results obtained with the CS are compared. Considering the tolerance ranges, the differences for acceptability and visual annoyance thresholds for vertical, focal, and green asymmetry are insignificant between still images and videos. However, there is a significant difference in thresholds for white asymmetry between two different tests. Particularly, the annoyance threshold of white asymmetry reduced by 6% in video test, whereas the acceptability reduced by 9% in comparison with S3D images.

These results are supported by a t-test (paired two samples for means), which was conducted to compare difference in MOS_{CS} scores between the images and videos for all the view asymmetries. No significant difference has been found for focal and vertical asymmetries. However, there was a significant difference for white ($t(5) = 2.01$, $p < 0.05$, $p = 0.029$) and green ($t(5) = 2.26$, $p < 0.05$, $p = 0.032$) asymmetries.

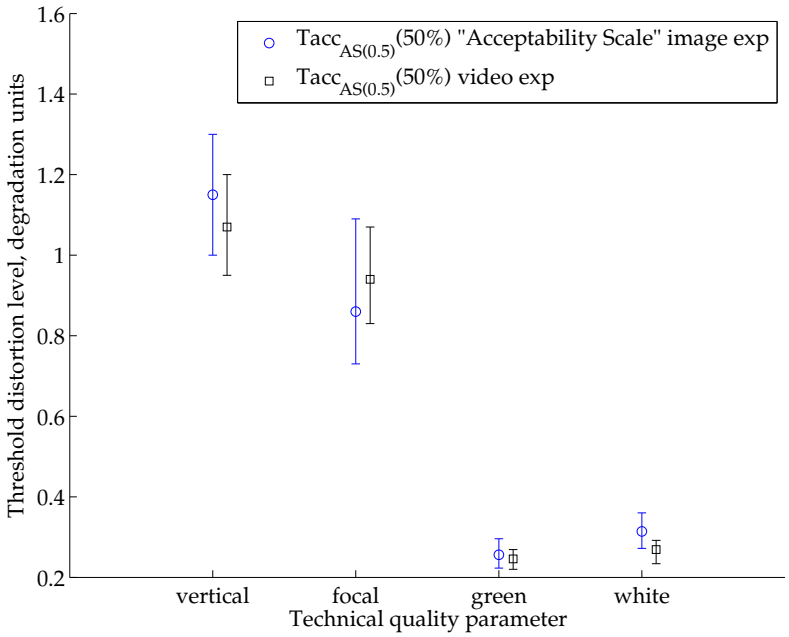


Figure 8.8: Comparison of acceptability and visual annoyance thresholds obtained with the Acceptability Scale (AS) for still and moving images.

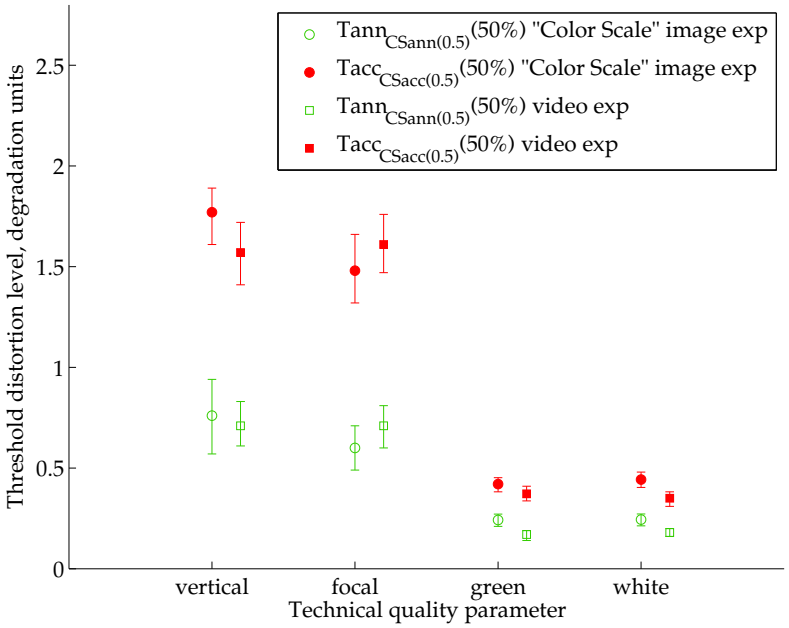


Figure 8.9: Comparison of acceptability and visual annoyance thresholds obtained with the Color Scale (CS) for still and moving images.

It seems that the threshold values depend on the luminance of the stimuli: dark images become darker faster than the light ones as in the case of the "Kitchen" and "Picnic" scenes in Figure 8.7.d. To verify this hypothesis the average luminance was computed for each video and image scenes. The luminance was computed as $L = 0.3R + 0.59G + 0.11B$, where R, G, B are image red, green and blue channels. The

results are presented in Table 8.4, where STD is standard deviation of luminance pixel values.

Table 8.4: Luminance composition of stimuli

Scene	Luminance	STD
Alley	107.8	70.3
Interview	72.4	41.7
Kitchen	120.7	71.7
Picnic	40.6	22.3
avg	**85.41**	**36.1**
Forest	80.5	43.8
Butterfly	132.4	44.1
Basketball	157.5	62.44
avg	**123.3**	**39.2**

Table 8.4 demonstrates that average luminance for S3D images were higher than for videos. There is no scene significance in our data. However, supposedly the variations of perceptual thresholds for luminance asymmetry may be explained by the luminance of a scene: when the luminance of dark stimulus decreases, such degradation is more perceptible than for light stimulus.

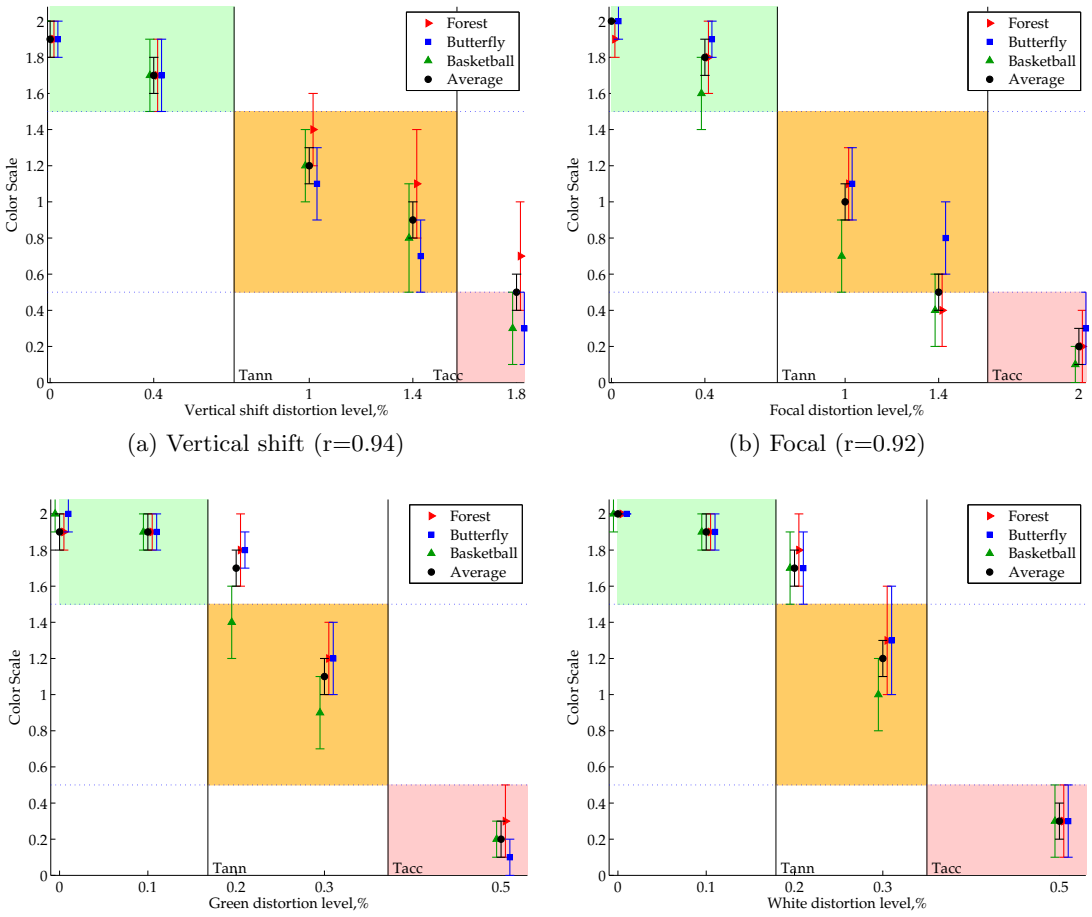
Conlusion: For the estimation of perceptual thresholds, it is preferable to consider the content of different color gamut and luminance. Presumably, perceptual thresholds for luminance and color asymmetries can be refined as a function of color and luminance. However, additional experiments should be carried out to confirm such a necessity.

8.2.3.2 Stereoscopic video versus images: data comparison

The aim of this section is to verify the performance OPSM metric with three still stereoscopic scenes used in Chapter 7 (see Fig. 7.1) when using the perceptual thresholds obtained for the video scenes.

Figure 8.10 illustrates five plotted graphs representing MOS scores with 95% confidence intervals for five view asymmetries for each scene. These plots allow direct comparison between subjective results of the "Color Scale" experiment (see Section 7.2.4) and objective prediction as explained in Section 6.3.5. The boundaries of objective categories (Tacc and Tann) are were extracted from video experiment for the different content than in the "Color Scale" experiment. The thresholds are plotted from Table 8.3 as the single vertical line representing the estimation from the mean curve. The MOS that do not match objective predictions are out of the color rectangles bounds.

The Pearson's correlation coefficients were calculated between the subjective results and objective predictions for focal (r=0.92), vertical shift (r=0.94), green (r=0.9), and white (r=0.87) asymmetries, and also for the scenes "Forest" (r=0.83), "Butterfly" (r=0.92), and "Basketball" (r=0.92).



(a) Vertical shift (r=0.94)

(b) Focal (r=0.92)

(c) Green (r=0.9)

(d) White (r=0.87)

Figure 8.10: Subjective scores for S3D images versus objective predictions made with perceptual thresholds obtained with 3D videos for four types of view asymmetry (the Pearson's correlation coefficient r is noted in brackets).

Conclusion: The results of the subjective experiments of this section with moving and still stereoscopic images have demonstrated high correlations between subjective scores and objective predictions for all tested view asymmetries (with minimum r=0.87). It implies that it is possible to classify detected technical quality parameter to one of the objective categories using the corresponding acceptability and visual annoyance thresholds.

8.3 Aggregation of technical quality parameters

In the previous section, it was established that our OPSM metric can be extended to 3D video content. However, it is not clear how to predict visual discomfort in a case when two view asymmetries are combined. The initial hypothesis on the evoked perceptual states resulted from the combination of two technical quality parameters was discussed in Section 6.3.6. Therefore, the goal of this section is to verify the initial assumption with a subjective experiment.

8.3.1 Stimuli generation

A new subjective experiment was designed in a way to aggregate three distortion levels corresponding to the middle of the "Red" (R), "Orange" (O), and "Green" (G) categories of two view asymmetries as illustrated in Table 8.5. The hypothetical result from asymmetry combination is presented in the table on the intersection of asymmetry categories.

Table 8.5: Aggregation of two technical quality parameters P_1 and P_2

D_{P_x1}/D_{P_x2}	G	O	R
G	G	O	R
O	O	O	R
R	R	R	R

Then, the MOS scores from the video test performed with the CS in the previous section were used to calculate distortion levels corresponding to the middle of each color category of vertical shift (see Figure 8.11), focal, and green level asymmetry. Following the recommendation ITU-R BT.500-13, the relationship between the MOS and the distortion levels for each type of view asymmetry was approximated. The symmetrical logistic function was used to obtain the continuous relationship between the MOS and the distortion level following the equation 6.1.

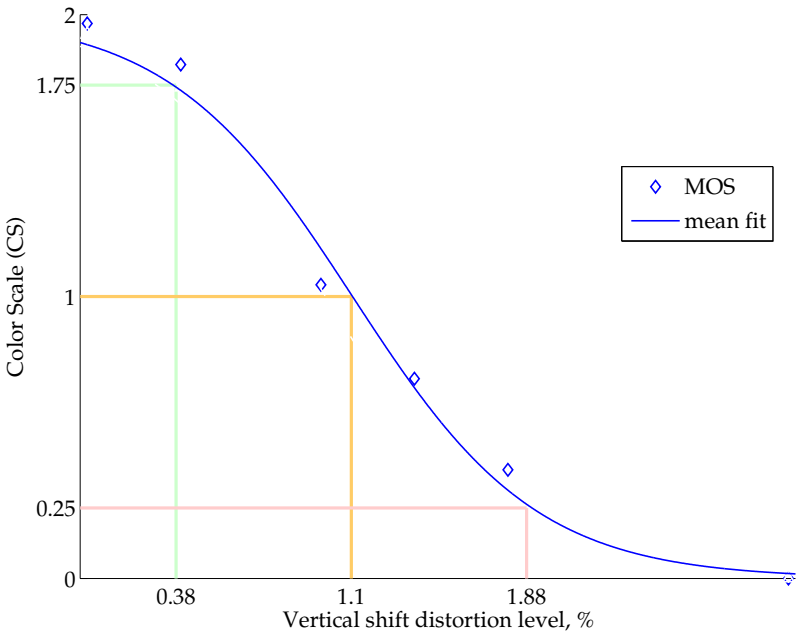


Figure 8.11: Distortion levels corresponding to the middle of each color category of vertical shift.

The distortion levels at the middle of each color category for view asymmetries are presented in the Table 8.6 below.

In the experiment the two following combinations of view asymmetries were used:

1. **Focal with vertical shift asymmetry**. In this case, first, one view was magnified according to the equation 7.4 and then cropped to the original resolution. Next, it

Table 8.6: The distortion levels of view asymmetries for middle of objective color categories

CS grade:	1.75 (G)	1 (O)	0.25 (R)
vertical shift, %	0.38	1.1	1.88
focal, %	0.4	1.1	1.86
green, %	0.108	0.35	0.42
white, %	0.114	0.338	0.4

was shifted and cropped again. Finally, a black mask was superimposed on both views to keep the original resolution. In total, nine distorted video sequences were generated: RR, RO, RG, OR, OO, OG, GR, GO, GG, and the anchor XO. Where, R, O, and G are distortion levels for the focal and vertical shift asymmetries from Table 8.6. The sequence XO is an anchor for a comparison of the results with the previous test. Only the vertical shift distortion corresponding to the middle of the "Orange" category was applied to create this impairment.

2. **Green level with vertical shift asymmetry**. First, the green level of one view was reduced according to the equation 7.5. Next, it was shifted and cropped. Finally, a black mask was superimposed on both views to keep the original resolution. In total, nine distorted video sequences were generated: RR, RO, RG, OR, OO, OG, GR, GO, GG, and the anchor OX. Where, R, O, and G are distortion levels for the green level and vertical shift asymmetries from Table 8.6. The sequence OX is an anchor for a comparison of the results with the previous test. Only the green level distortion corresponding to the middle of the "Orange" category was applied to create this impairment.

The experimental set-up, stimuli, and methodology were the same as described in Section 8.2.2. 33 observers participated in the experiment. The SAMVIQ protocol was used to evaluate the sequences on the Color Scale described in Section 6.3.2. In total, every subject had to evaluate 96 stimuli (4 scenes * 2 asymmetry aggregation * [9 combination of distortion levels + explicit reference + hidden reference + anchor]). The visualization time for one stereoscopic pair was 15 seconds. The experiment was divided into two 25 minutes sessions with a 30 minute break in between. Hence, around one and a half hours for the whole experiment. In each session, the subjects evaluated one type of view asymmetry aggregation.

8.3.2 Result analysis

8.3.2.1 Agregation of green and vertical shift asymmetries

The MOS scores with a 95% confidence interval were computed for nine distortion levels and two anchors for green and vertical shift asymmetry aggregation. The ANOVA analysis demonstrated that the video content was insignificant for changes of MOS scores, while the distortion level had a significant ($p < 0.0001$) impact for all types of asymmetry. The obtained MOS scores were sorted in descending order and the result is presented in Figure 8.12.

The average values for all scenes (see Table 8.7) were translated to the subjective color categories as presented in Table 8.8 to facilitate a comparison with the initial hypothesis in Table 8.5. Two combinations of asymmetries did not match the initial

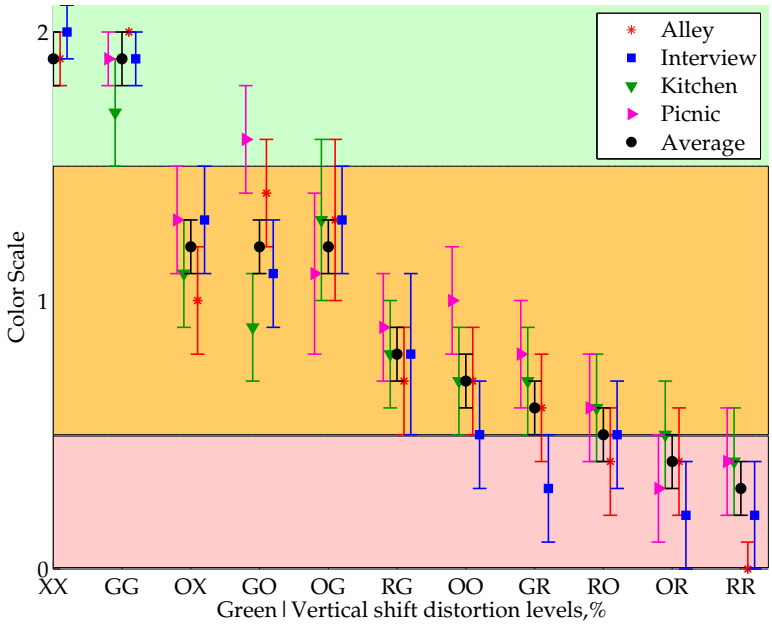


Figure 8.12: Agregation of green and vertical shift asymmetries.

hypothesis: RG and GR. Such votes could be a result of the experiment design, where only 3 stimuli out of 12 were not annoying, which might force observers to evaluate the sequences less rigorously than in a case when only one degradation from the "Red" category is present (like in all the previous subjective experiments).

Table 8.7: Result of aggregation of green and vertical shift asymmetries for all scenes

GrShift:	XX	GG	GO	OX	OG	RG	OO	GR	RO	OR	RR
green,%	0	0.108	0.108	0.35	0.35	0.42	0.35	0.108	0.42	0.35	0.42
shift,%	0	0.38	1.1	0	0.38	0.38	1.2	1.88	1.1	1.88	1.88
avg	1.9	1.9	1.2	1.2	1.2	0.8	0.7	0.6	0.5	0.4	0.3
CI	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Table 8.8: Result of aggregation of green (g) in row and vertical shift (v) in column asymmetries expressed in color categories for all scenes

g/v	G	O	R
G	G	O	**O**
O	O	O	R
R	**O**	R	R

In order to investigate the relationship between the votes and the combination of green and vertical shift asymmetry, it is assumed that the MOS score can be represented as a weighted sum of the green (g) and vertical shift (v) asymmetries expressed as a percentage. Furthermore, to find the weights, a multiple linear regression analysis was performed using the MOS data from this experiment for each scene and the degradation levels in Table 8.6 as values of the independent variables. The weights of the predicted

scores were normalized to a sum of one for both green and vertical shift asymmetries. The following equation 8.1 was obtained:

$$MOS = 2.03 - 0.45g - 0.55v \tag{8.1}$$

The R-square is 0.87 (> 0.7). Therefore there is a linear relationship between the magnification and the vertical shift. More complex models (e.g exponential, logarithmic) did not explain more valiance of the data resulting in the same R-square.

Conclusion: The fitted coefficients demonstrate that both asymmetries contribute almost equally to the scores of the aggregation: 55% of vertical shift and 45% green asymmetry.

8.3.2.2 Agregation of focal and vertical shift asymmetries

The MOS scores with a 95% confidence interval were computed for nine distortion levels and two anchors for green and vertical shift asymmetry aggregation. The ANOVA analysis demonstrated that video content was insignificant for a change of MOS scores, while the distortion level had a significant ($p < 0.0001$) impact for all types of asymmetry. The obtained MOS scores were sorted in descending order and the result is presented in Figure 8.13.

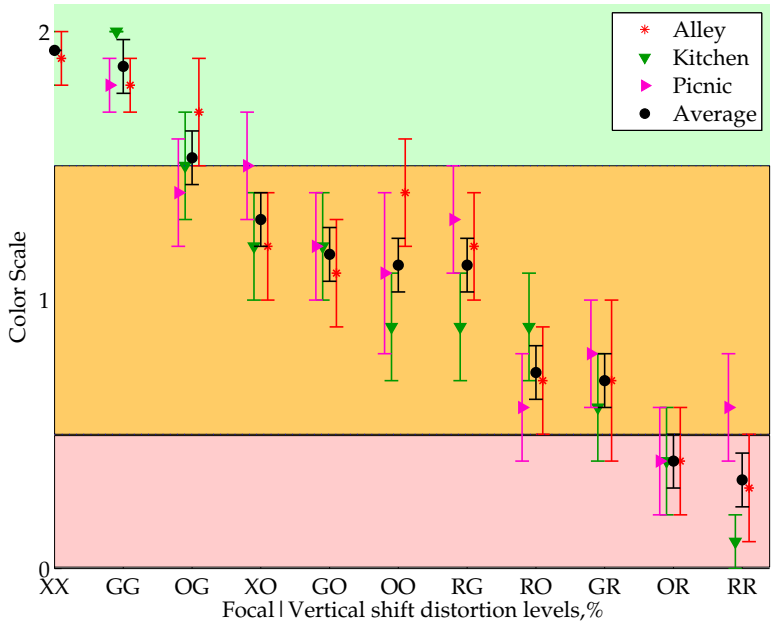


Figure 8.13: Aggregation of focal and vertical shift asymmetries.

The average values for all scenes were translated to color categories and presented in Table 8.9 to facilitate comparison with the initial hypothesis in Table 8.5. Three combinations of asymmetries did not match the initial hypothesis: OG, RG, RO, and GR. For a better understanding of the votes, the stereoscopic scenes were analyzed with the StereoLabs tool, which is able to measure the global vertical shift and zoom after a combination of magnification and vertical shift asymmetries. The MOS scores and global vertical shift are presented in Table 8.10.

Table 8.9: Result of aggregation of focal (f) and vertical shift (v) asymmetries expressed in color categories for all scenes

f/v	G	O	R
G	G	O	**O**
O	**G**	O	R
R	**O**	**O**	R

Table 8.10: Result of aggregation of focal and vertical shift asymmetries for all scenes measured using the StereoLabs tool.

FocVert:	XX	GG	OG	XO	GO	OO	RG	RO	GR	OR	RR
zoom, %	0.1	0.37	0	0.42	0.4	1.13	1.07	1.07	1.77	1.83	1.77
global shift, lines	0	3	6	10	11	12	8	15	18	21	21
avg	1.93	1.87	1.53	1.3	1.17	1.13	1.13	0.73	0.7	0.4	0.33
CI	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

In all cases the decision made by observers about color category is based on global vertical disparity. According to Table 8.3, the "Orange" category is situated between 8 and 17 lines of the vertical shift. Therefore, all the MOS votes in the table are in accordance with this hypothesis. But, for the RG combination, the global shift is lower than in the cases of GO and OO. The reason is the influence of magnification asymmetry, which is in the "Red" category.

In order to investigate the relationship between the votes and the combination of magnification and vertical disparity, it is assumed that the MOS score can be represented as a weighted sum of magnification (f) and vertical shift (v) expressed as a percentage. Furthermore, to find the weights, a multiple linear regression analysis was performed using the MOS data from this experiment for each scene with the measured degradation levels in Table 8.10 as independent variables. The weights of the predicted scores were normalized to a sum of one for both green and vertical shift asymmetries. The following equation 8.2 was obtained:

$$MOS = 2.19 - 0.18f - 0.82v \tag{8.2}$$

The R-square is 0.88 (> 0.7). Therefore, there is a linear relationship between the magnification and vertical shift. More complex models (e.g exponential, logarithmic) did not explain more valiance of the data resulting in the same R-square.

Conclusion: The fitted coefficients demonstrate that the decision about color category is mostly based on vertical shift(82%) rather than on the magnification asymmetry (18%) of an image.

8.4 Conclusions

In this chapter the color metric was validated with stereoscopic videos. In addition, perceptual thresholds obtained for stereoscopic images and videos were compared. It was demonstrated that:

- The perceptual thresholds for stereoscopic images and videos were similar for all the view asymmetries except for white. Such results were explained by the global scene

luminance. Thus, the perceptual thresholds for color and luminance asymmetries should be refined as a function of color and luminance.

- The perceptual thresholds obtained in this thesis can be re-used in another studies. However, the distortion levels were given as a percentage of height and width of the original images. Thus, it is preferable to translate the thresholds to degrees of visual angle in order to consider the influence of visualization parameters.

- The results of the subjective experiments with moving and still stereoscopic images have demonstrated high correlations between subjective scores and objective predictions made using the perceptual thresholds obtained for videos for all tested view asymmetries (with a minimum r=0.87). It implies that it is possible to classify detected image quality parameter to one of the objective categories using the corresponding acceptability and visual annoyance thresholds.

- The combination of green and vertical shift asymmetries mostly followed the algorithm proposed in Figure 6.13. However, in some cases observers selected the "Orange" category despite a choice of "Red" being expected when the asymmetries from the "Red" and "Green" categories were combined. Such results could be explained by the test design, where only 3 stimuli out of 12 were not annoying, which might force observers to evaluate less rigorously. The normalized fitted coefficients showed that vertical shift and green view asymmetries almost have an equal influence on the scores: 55% of vertical shift and 45% of green asymmetry.

- The judgments of the aggregations of geometrical asymmetries are mostly based on vertical shift: 82% of vertical shift and 18% of magnification in our experiment.

Conclusions

Main conclusions

This thesis proposed an objective model for video quality characterization. The main goal was to take advantage of human visual perception by predicting perceptual categories rather than MOS scores. Hence, the model integrated visual annoyance threshold and adjustable acceptability level. The thesis contains two research axes. The first axis (Part I) explores the potential of mechanisms of visual attention in 3D for prediction of visual discomfort and compares visual attention in stereoscopic and non-stereoscopic conditions. The second axis (Part II) investigates the model performance and robustness for prediction of visual discomfort of still and moving stereoscopic images.

In particular, **Chapter 1** investigated the influence of the binocular vision limitations on depth perception in S3D systems. Binocular disparity is not the only source of depth information for humans. Though in stereoscopic systems, the enhanced depth perception is generated mostly owing to binocular disparities. Sensation of depth occurs when the brain fuses two slightly different flat images. Some conflicts can happen when the discrepancies between images are too large or when the HVS is susceptible to some degree of unnatural viewing. Such conflicts can result in visual discomfort that can be avoided if the reconstructed scene remains within the comfortable viewing zone limited by DoF=0.2 diopter. In addition, the amount of perceived depth depends on screen size and viewing distance. Therefore, when studying depth perception the viewing distance and screen disparities must be considered for the generalization of the results.

Chapter 2 reviewed the potential impact of the technical parameters of broadcast chain stages on the final viewing experience. In the stage of the content production three shooting parameters influence on the amount of perceived depth: camera focal length, baseline, and convergence distance; in the visualization stage these parameters are human interpupillary distance, display width, and viewing distance. Thus, the lack of the control of the shooting and the visualization environments hinder the generalization of the results. In addition, the quality of the perceived depth is based on the absence of visual artifacts. These imperfections may be produced at every stage of the stereoscopic broadcast chain. Therefore, it is very important to consider from the beginning the impact of the technologies to avoid expensive postproduction processes.

Chapter 3 presented the existing definition of 3D QoE and analyzed the various models of 3D QoE. It was determined that three primary composites of 3D QoE (image quality, depth quality and visual comfort) can be linked directly to technical quality parameters unlike the higher level perceptual attributes (naturalness, sense of presence, etc.). Concerning depth quality, from the literature overview it was revealed that this concept is quite difficult to judge for the subjects, hence the possibilities to find a more representa-

tive indicator was discussed. The review of the objective metrics for 3D QoE assessment demonstrated that the most of existing metrics were adopted from 2D quality evaluation. Such metrics are not capable to detect problems related to visual comfort. Nevertheless, existing 3D metrics do not consider all possible sources of visual discomfort. Hence, it was concluded that currently a comprehensive objective metric of 3D video QoE does not exist at the moment. The demands inherent to an objective quality model for 3D QoE were formulated. It was stated that it should take into account a) the quality of the three primary perceptual dimensions, b) 3D display technology and representation format, and c) the visualization environment (display size, viewing distance). Such a model should assess not just the quality of the stereoscopic signal but its rendered version. If one of the components is missing, it would be quite difficult to make conclusions about the overall 3D quality of experience.

Chapter 4 introduced the state-of-the-art studies about visual attention. The analysis of recent studies comparing visual attention in stereoscopic with non-stereoscopic conditions has been given. The lack of coherence in reported indicators between the studies and the experimental conditions was demonstrated. In addition, most of the studies examining visual attention in 3D did not consider comfortable viewing zone while displaying 3D content. Therefore, the absence of a standard or guidelines for eye-tracking experiments has lead to difficulties in the generalization of existing studies.

Chapter 5 presented the three experimental studies, which compared visual attention in 2D and 3D conditions. The stimuli were fully controlled in the shooting stage as well as during the visualization. This allowed studying the impact of different level of binocular disparities (including visual discomfort) and texture complexities on visual attention. It was discovered that the observation strategy of still stereoscopic images located behind the display plane is similar to observation of 2D images. A gaze was rather guided by the saliency of objects than by the amount of uncrossed disparities. However, the more crossed disparities were presented the more visual attention were directed to such area, even in the case of visual discomfort caused by the excessive disparities. No evidences have been found that visual discomfort generated by excessive disparities influences the way we observe the images. The new depth metric was proposed based on the results of subjective studies. The metric allowed comparison of visual attention in 2D and 3D conditions as well as 3D conditions with different amount of depth owing to weighted saliency maps and segmentation based on a depth map. The computed results validated the conclusions from the eye-tracking experiments.

Chapter 6 defined the framework to predict 3D QoE and proposed the assessment criteria associated with the basic perceptual attributes of 3D video QoE. Taking into account that the most important task for any 3D system is to guarantee visual comfort to its viewers the model was tested for the objective prediction of visual discomfort. Therefore, the proposed model used perceptual thresholds to define the impact of technical parameters on the visual comfort axis of 3D video QoE. After objective measurement of 3D technical parameters and comparison with perceptual thresholds, it was possible to predict evoked perceptual state, which reflected the viewers' categorical judgment based on stimulus acceptability and induced visual annoyance. In addition, it was suggested to use objective model as subjective scale with color categories. This allowed establishing the direct link between subjective votes and objective predictions. The advantages of proposed approach are the possibility to omit prediction of MOS scores and tune the metric according to the service requirements of customers' acceptability and visual an-

noyance levels. Besides, the method does not depend on any precise 3D technology and no reference is required to predict visual discomfort.

Chapter 7 validated the proposed model in subjective experiments with fully controlled still stereoscopic images with different types of view asymmetries. The exploration to use proposed model as subjective color scale has demonstrated that it was possible to obtain acceptability and visual annoyance thresholds in the same time directly with the color scale. However, these thresholds were not the same when evaluated with standard impairment and acceptability scales. The difference of the acceptability thresholds values was explained by the different evaluation concepts.

Chapter 8 justified the proposed model with fully controlled stereoscopic video sequences. Also perceptual thresholds for still and moving images were compared. The model performance was evaluated by comparing objective predictions with subjective scores for various levels of view discrepancies, which might provoke visual discomfort. Furthermore, this chapter has explored how objective predictions should be affected if two view asymmetries were aggregated in a single stereoscopic stimulus. It was discovered that the judgments of the aggregations of geometrical asymmetries were mostly based on vertical shift.

Future work and perspectives

1. The review of the eye-tracking studies investigating the impact of depth on visual attention has revealed the lack of coherence in reported indicators between the studies and the experimental conditions. We believe that recommendations or guidelines in a protocol for eye-tracking studies, representative indicators of eye-movements and effective methods for qualitative analysis would help the generalization and reproducibility of the results.

2. The huge effort on modeling of the 3D QoE has been done by researches. It has permitted to define primary indicators, which can be estimated independently and then combined as a weighted sum to a single score characterizing the overall viewing experience. Nevertheless, we believe that not enough attention was drawn to the issue of stereoscopic depth distortion, which is important factor (roundness) in cinematography. However, the perceptual limits of geometrical distortions are not known: what level of shape distortion is perceptible, what level of distortion is annoying, what impact various levels of shape distortions have on the overall 3D video QoE and so on.

3. Within the scope of this thesis a new objective perceptual model was proposed. The advantage of this model is that it can be designed according to marketing, technical, or other requirements by changing the percentage of acceptability or visual annoyance thresholds, e.g. adapting objective categories based on requrements. Acceptability depends on the acceptance of the annoyance level caused by the distortion. Basically, acceptability is weighted by visual annoyance. Similarly, it can be weighted by any other assessment criteria (geometric distortion, blur, noise, etc). We believe that the proposed color metric could be transferred to any other technology where degradations can be measured and associated with perceptual thresholds. However, additional tests should be carried out for verification.

4. A long duration subjective experiment could be conducted to investigate the impact on QoE acceptability of recurrences, durations and levels (color categories) of views asymmetries.

Appendix A

Supplementary information for Chapter 5

A.1 Pearson correlation coefficient (CC)

The Pearson correlation coefficient CC between two saliency maps H and P can be computed following the equation A.1:

$$CC_{H,P} = \frac{cov(H,P)}{\sigma_H \sigma_P} \tag{A.1}$$

where $cov(H,P)$ is the covariance between maps H and P and σ_H, σ_P are standard deviations of H and P.

The range of CC values is between 0 and 1. A value of 0 indicates that there is no linear correlation between two maps and 1 - perfect correlation. This coefficient is very simple to compute. It is often used to evaluate performance of computational models of visual attention.

A.2 Area Under Curve (AUC)

Area under the curve (AUC) or ROC area is the indicator of the classification using the ROC analysis - the method to assess the degree of similarity of two saliency maps [Green and Swets, 1966]. One saliency map represents the ground truth, whereas another is the prediction.

The binary classification is applied to every pixel of the both continuous saliency maps. On the threshold basis, pixels are classified onto two groups: fixated (or salient) or as not fixated (not salient). Two different thresholds are applied depending on whether the ground truth or the prediction is considered:

- **Ground truth:** The continuous saliency map is thresholded with a constant threshold T_G^x to keep a given percentage of image pixels, where G is denotation of the ground truth and x of the percentage of an image considered as being fixated. For example, Figure A.1 illustrates the different percentage of pixels that have been fixated.

- **Prediction:** The threshold T_P^x is systematically moved between the minimum and the maximum values of the map. In this case, P denotes the prediction and x

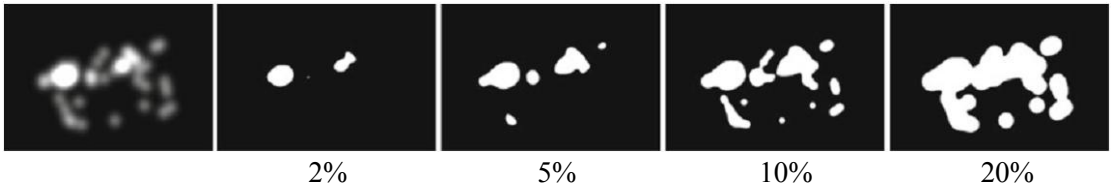


Figure A.1: Thresholded saliency maps to keep the top percentage of salient areas. From left to right: 2%, 5%, 10%, and 20%.

indicates ith threshold. A high-threshold value means an overdetection, whereas a smaller threshold affects the most salient areas of the map.

For each pair of thresholds, the quality of the classification is computed. Four scores represent the true positives (TPs), the false positives (FPs), the false negatives (FNs), and the true negatives (TNs). The true positive score is the number of fixated pixels in the ground truth that are also labeled as fixated in the prediction.

The thresholding operation is graphically demonstrated in Figure A.1. The first continuous saliency map (Fig. A.2.b) is thresholded to keep 20% of the image T_G^{20}. Further, it is compared with the second continuous saliency map (Fig. A.2.d). The classification result is illustrated in Figure A.3. The red and uncolored areas represent pixels having the same label e.g. a good classification (TP). The green areas represent the pixels that are fixated but are labeled as nonfixated locations (FN). The blue areas represent the pixels that are nonfixated but are labeled as fixated locations (FP). A confusion matrix is often used to visualize the algorithm's performance (see Fig. A.4.c).

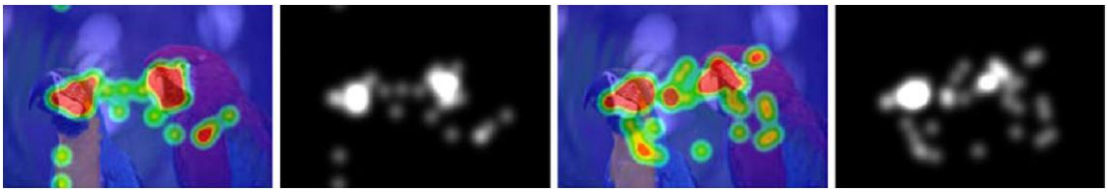


Figure A.2: Heat maps and continuous saliency maps obtained from fixations of two groups of 3 observers.

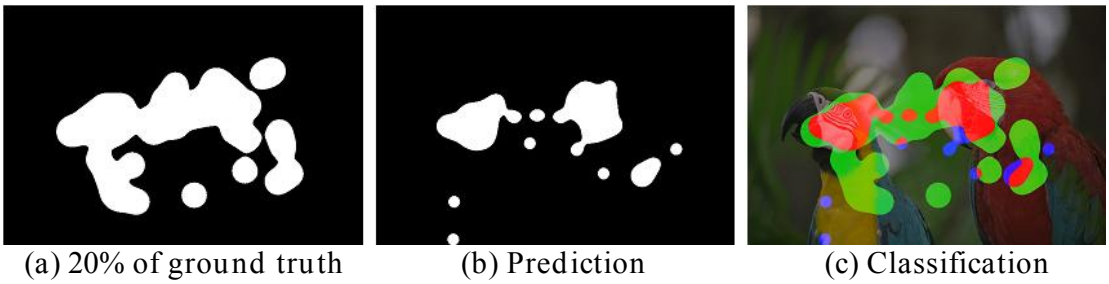


Figure A.3: Classification result (on the right) when a 20% thresholded ground truth (left picture) and a prediction (middle picture) are considered. Red areas are true positives, green areas are false negatives, and blue areas are false positives. Other areas are true negatives.

An ROC curve that plots the FP rate (FPR) as a function of the TP rate (TPR) is used to display the classification result for the set of thresholds used. The TPR, also called sensitivity or recall, is defined as TPR=TP/(TP+FN), whereas the FPR is given by FPR=FP/(TP+FN).

Finally the area under curve(AUC) or ROC area, provides a measure indicating the overall performance of the classification. A value of 1 indicates a perfect classification. The chance level is .5. The AUC curve of Figure A.3 is given in Figure A.4.

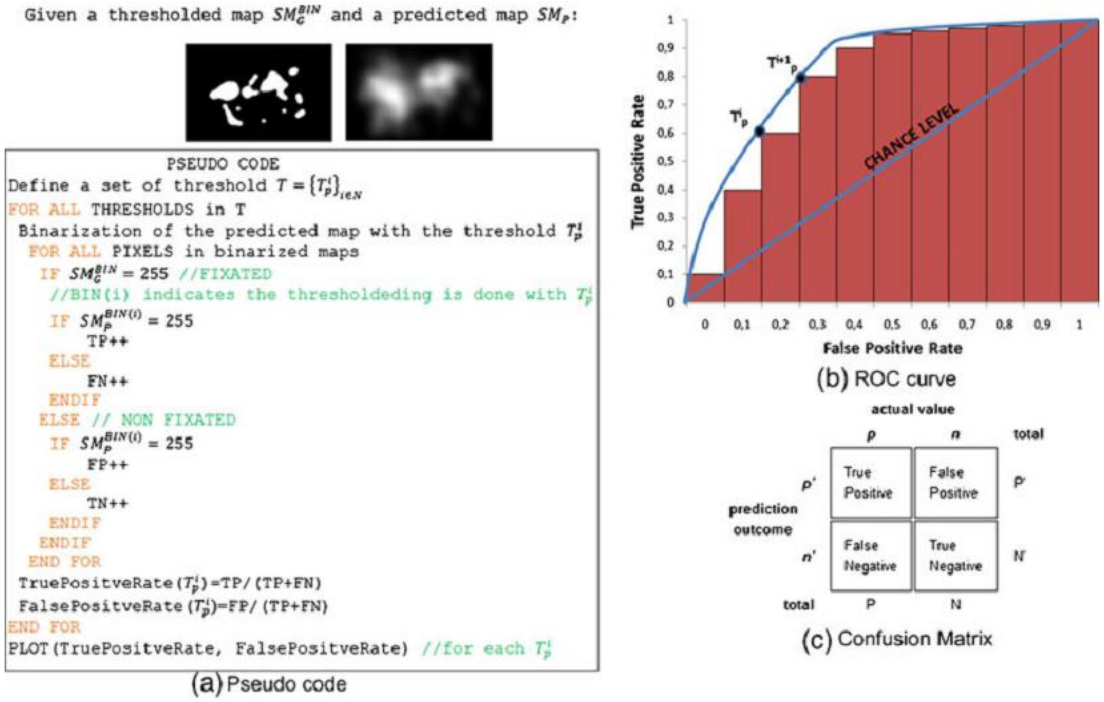


Figure A.4: (a) Pseudo-code to perform the ROC analysis between two maps (b) ROC curve and (c) the confusion matrix.

There are different methods to compute the AUC. The simplest ones are based on the left and right Riemann sums. The left Riemann sum is illustrated Figure A.4. A more efficient approximation can be obtained by a trapezoid approximation: rather than computing the area of rectangles, the AUC is given by summing the area of trapezoids. In Figure A.4, the AUC value is 0.83.

A.3 Measuring the inter-observer congruency (IOVC)

The inter-observer congruency (IOVC) can be assessed using a one-against-all approach [Torralba et al., 2006]. Firstly, a 2D fixation distribution from the fixation data of all observers except one for a given picture should be computed. Then the fixation distributions are convolved with a two-dimensional Gaussian. Each pixel of this map represents the probability to be fixated. The standard deviation of the Gaussian kernel is one degree of visual angle, which represents the estimate of foveal size. The obtained map should be thresholded in a way to select an image area with the highest fixation probability. The threshold can be adapted in order to keep 25% of the image. Secondly, the percentage of the visual fixations of the remaining observers that fall within salient

parts of the threshold saliency map should be computed. This process is iterated for all observers. For a given picture, the variability between observers is the average of the aforementioned percentage over all subjects. Usually the most of the dispersion values are in the range of 0.5 to 1. So this range was has been scaled from 0 to 1. A value of 1 indicates that observers fixate the same areas, whereas a low value suggests that the scan patterns are uncorrelated meaning a strong variability between subjects. Figure A.5 illustrates the method for the ith observer and Figure A.6 the examples of the computed IOVC values.

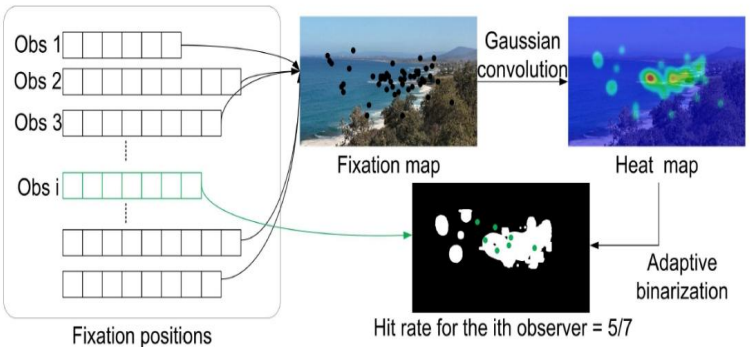


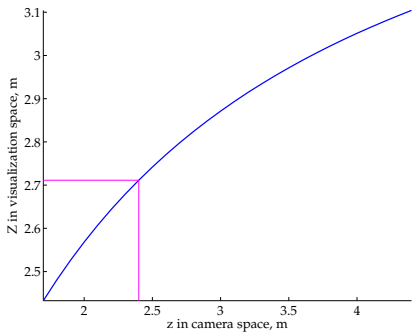
Figure A.5: Measure of the inter-observer congruency. On the left, the spatial coordinates of visual fixations for each observer are presented. On the right, a heat map is computed (on the right) for all fixations except for the ith observer. Finally, the number of fixations of the ith observer that fall into salient regions (white region on the bottom) are counted after an adaptive binarization.

IOVC=0.23 IOVC=0.72 IOVC=0.85

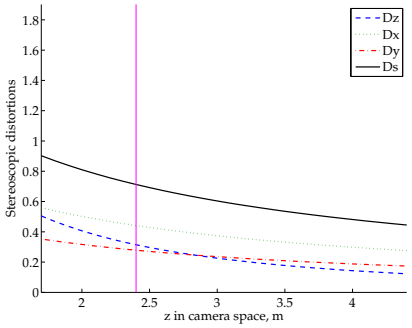
Figure A.6: Examples of pictures associated with their corresponding inter-observer congruency. IOVC is in the range of 0 (strongest) to 1 (lowest).

A.4 Camera space z versus visualization space Z. Stereoscopic distortions in visualization space Z

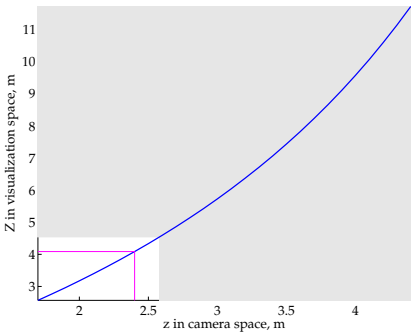
Figures A.7- A.16 present the detailed analysis of relationships between camera space and visualization space and depth distortions for comfortable and uncomfortable condition for all the complex stimuli used in Chapter 5. The space outside ZoC is marked in light gray color and region of interest as magenta line.



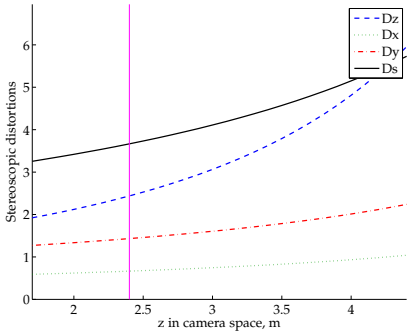
(a) DoF=0.1 Z vs. z



(b) DoF=0.1 S-distortions

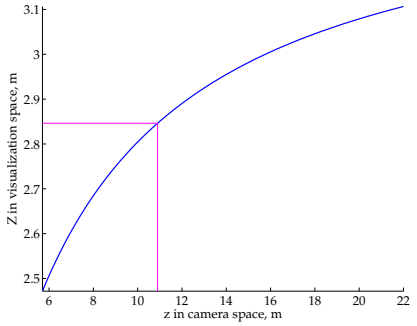


(c) DoF=0.3 Z vs. z

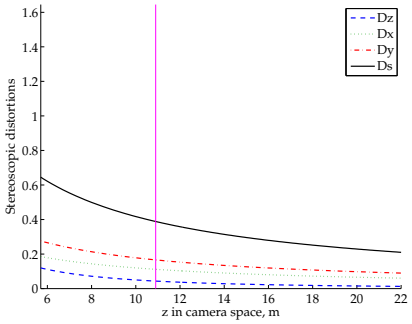


(d) DoF=0.3 S-distortions

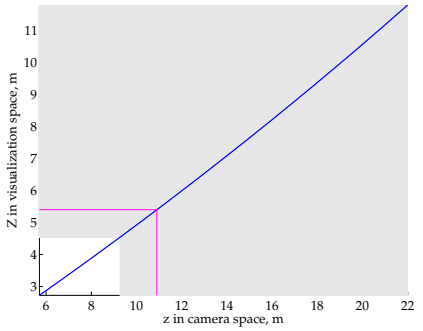
Figure A.7: Bathroom



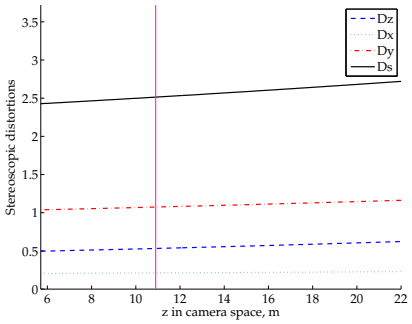
(a) DoF=0.1 Z vs. z



(b) DoF=0.1 S-distortions

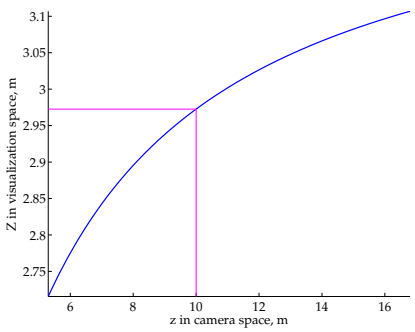


(c) DoF=0.3 Z vs. z

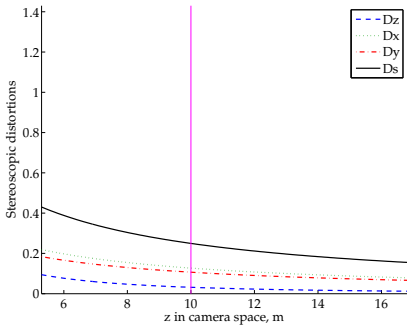


(d) DoF=0.3 S-distortions

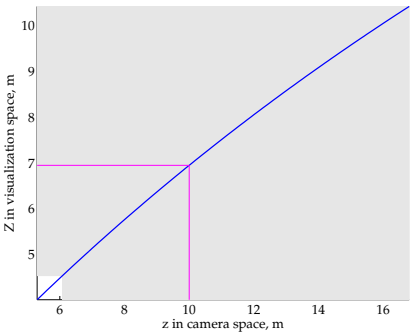
Figure A.8: Cartoon



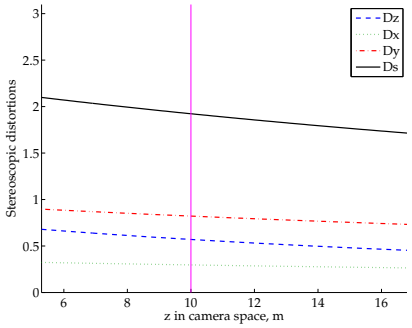
(a) DoF=0.1 Z vs. z



(b) DoF=0.1 S-distortions

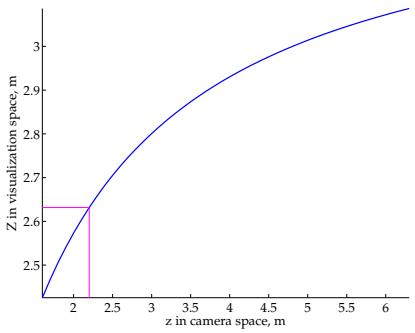


(c) DoF=0.3 Z vs. z

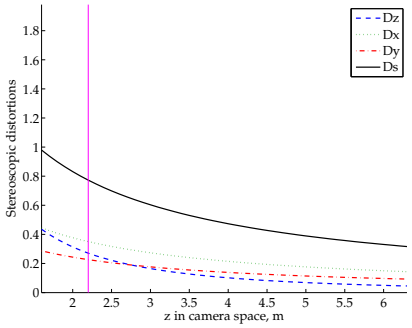


(d) DoF=0.3 S-distortions

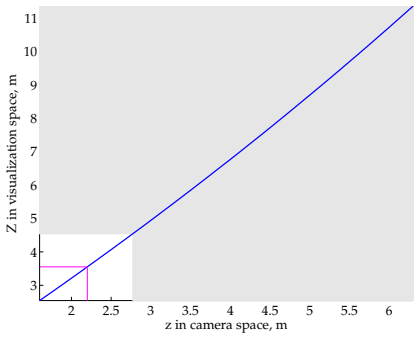
Figure A.9: Hallway



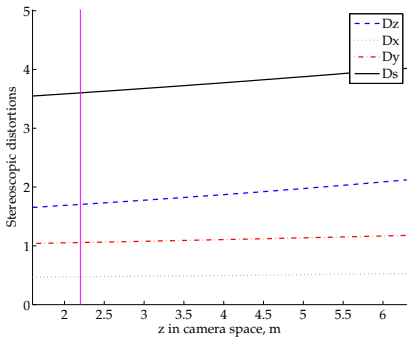
(a) DoF=0.1 Z vs. z



(b) DoF=0.1 S-distortions

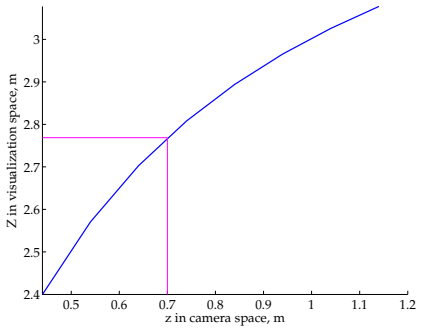


(c) DoF=0.3 Z vs. z

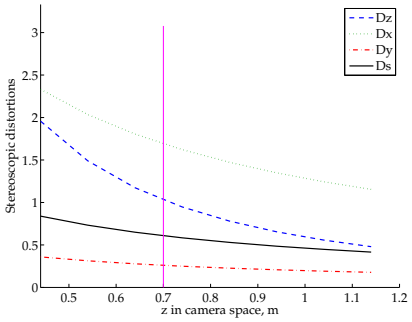


(d) DoF=0.3 S-distortions

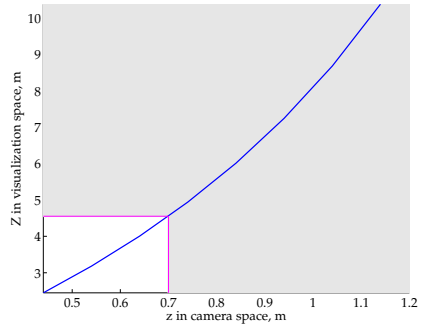
Figure A.10: Kitchen



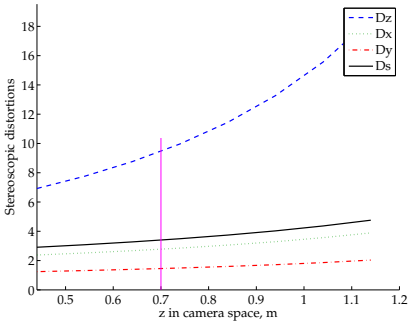
(a) DoF=0.1 Z vs. z



(b) DoF=0.1 S-distortions

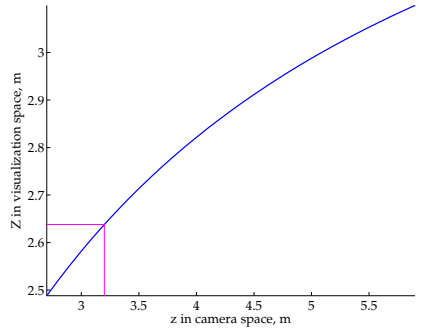


(c) DoF=0.3 Z vs. z

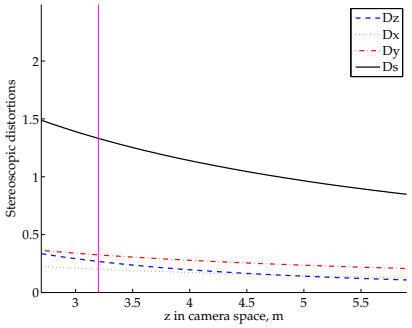


(d) DoF=0.3 S-distortions

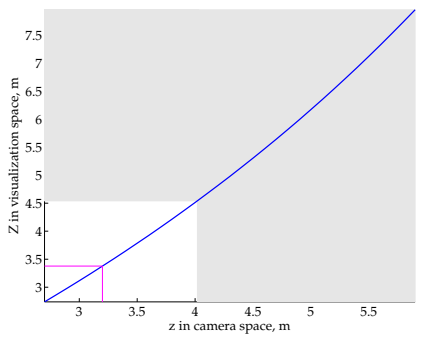
Figure A.11: Tea



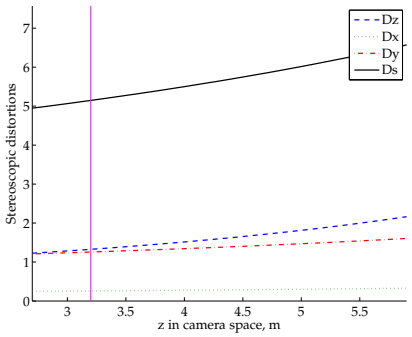
(a) DoF=0.1 Z vs. z



(b) DoF=0.1 S-distortions

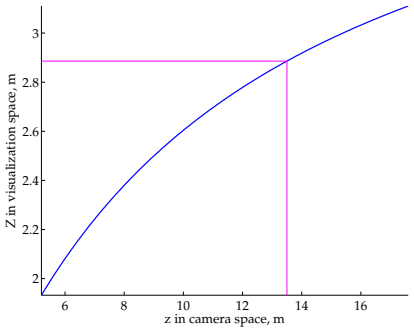


(c) DoF=0.3 Z vs. z

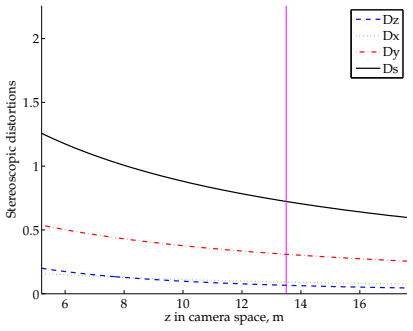


(d) DoF=0.3 S-distortions

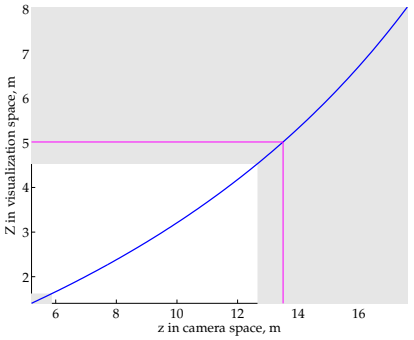
Figure A.12: Room



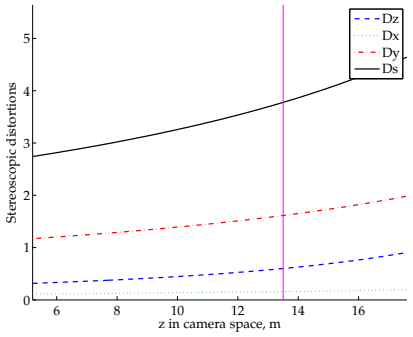
(a) DoF=0.1 Z vs. z



(b) DoF=0.1 S-distortions

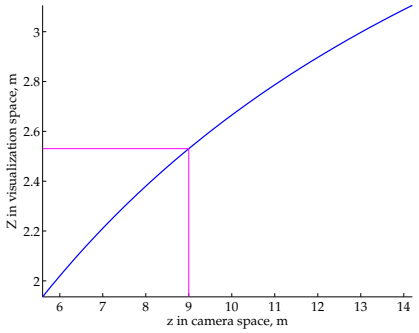


(c) DoF=0.3 Z vs. z

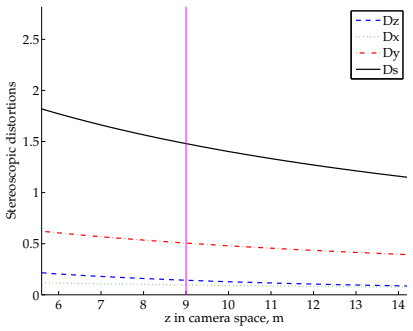


(d) DoF=0.3 S-distortions

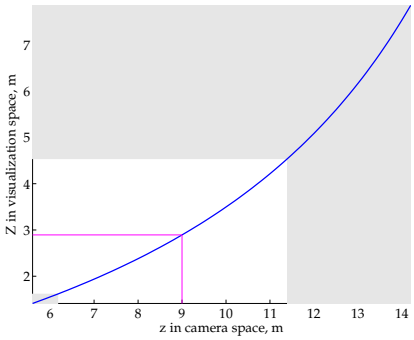
Figure A.13: Cartoon (Crossed disparity)



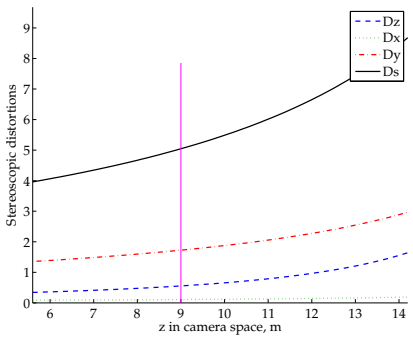
(a) DoF=0.1 Z vs. z



(b) DoF=0.1 S-distortions

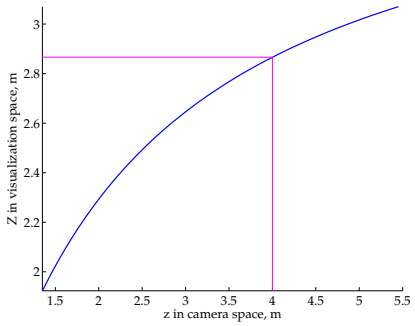


(c) DoF=0.3 Z vs. z

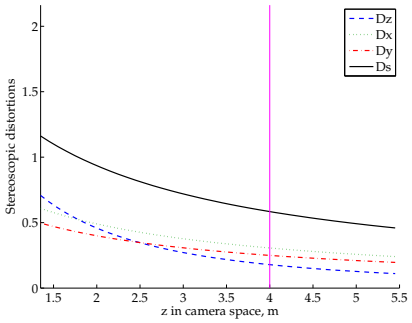


(d) DoF=0.3 S-distortions

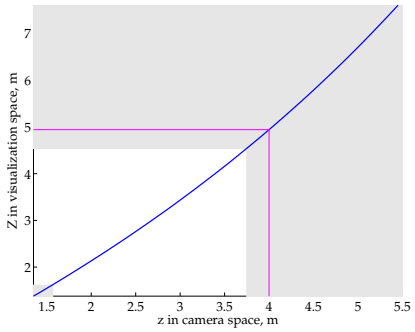
Figure A.14: Hall



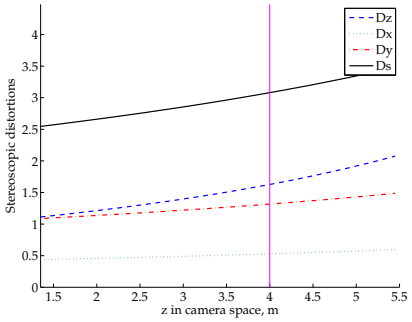
(a) DoF=0.1 Z vs. z



(b) DoF=0.1 S-distortions

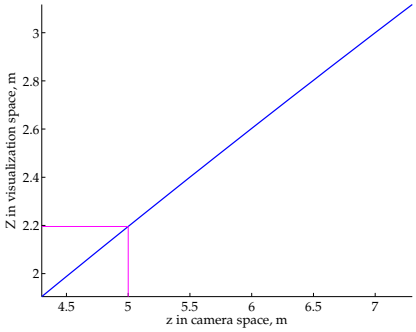


(c) DoF=0.3 Z vs. z

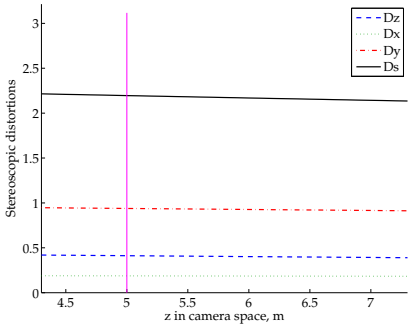


(d) DoF=0.3 S-distortions

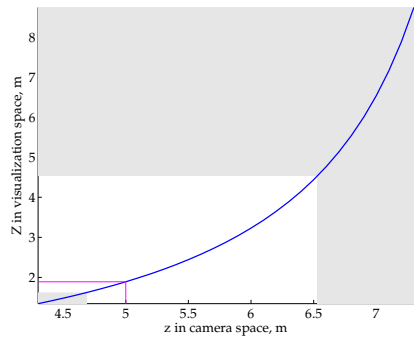
Figure A.15: Pigs



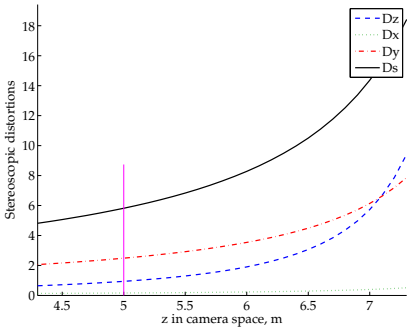
(a) DoF=0.1 Z vs. z



(b) DoF=0.1 S-distortions



(c) DoF=0.3 Z vs. z



(d) DoF=0.3 S-distortions

Figure A.16: Table

A.5 CC, AUC, IOVC data

Table A.1: AUC, CC correlation values between 2D and 3D DoF=0.1 (2D/01) SMs; between 2D and 3D DoF=0.3 (2D/03) SMs; between 3D DoF=0.1 and 3D DoF=0.3 (01/03) saliency maps.

20 s	AUC			CC		
Scene	2D/01	2D/03	01/03	2D/01	2D/03	01/03
Bathroom LT	0.83	0.81	0.83	0.86	0.79	0.88
Bathroom MT	0.81	0.81	0.82	0.85	0.87	0.89
Bathroom HT	0.81	0.83	0.85	0.82	0.86	0.91
Cartoon LT	0.87	0.87	0.88	0.91	0.9	0.91
Cartoon MT	0.85	0.85	0.86	0.84	0.84	0.92
Cartoon HT	0.85	0.85	0.83	0.89	0.9	0.88
Hallway LT	0.87	0.86	0.86	0.91	0.93	0.86
Hallway MT	0.85	0.86	0.85	0.85	0.85	0.92
Hallway HT	0.84	0.84	0.84	0.84	0.87	0.88
Kitchen LT	0.86	0.85	0.88	0.88	0.85	0.91
Kitchen MT	0.84	0.84	0.84	0.89	0.82	0.88
Kitchen HT	0.82	0.82	0.83	0.89	0.89	0.91
Tea LT	0.9	0.9	0.91	0.89	0.92	0.86
Tea MT	0.86	0.87	0.88	0.83	0.9	0.88
Tea HT	0.88	0.87	0.89	0.91	0.91	0.92
Room LT	0.86	0.87	0.86	0.89	0.92	0.87
Room MT	0.83	0.81	0.8	0.87	0.85	0.76
Room HT	0.84	0.84	0.82	0.88	0.83	0.85
avg	0.85	0.85	0.85	0.87	0.87	0.88
CI, ±	0.01	0.01	0.01	0.01	0.02	0.02

Table A.2: IOVC for low (LT), medium (MT), high (HT) texture complexities

	2D			3D DoF=0.1			3D DoF=0.3		
Scene	LT	MT	HT	LT	MT	HT	LT	MT	pHT
Bathroom	0.63	0.63	0.69	0.7	0.66	0.67	0.63	0.68	0.64
Cartoon	0.76	0.74	0.68	0.76	0.72	0.65	0.7	0.74	0.73
Hallway	0.73	0.73	0.69	0.74	0.73	0.72	0.73	0.67	0.73
Kitchen	0.76	0.7	0.69	0.78	0.69	0.7	0.72	0.66	0.71
Tea	0.83	0.76	0.78	0.84	0.79	0.8	0.8	0.82	0.79
Room	0.75	0.69	0.7	0.71	0.65	0.67	0.74	0.66	0.63

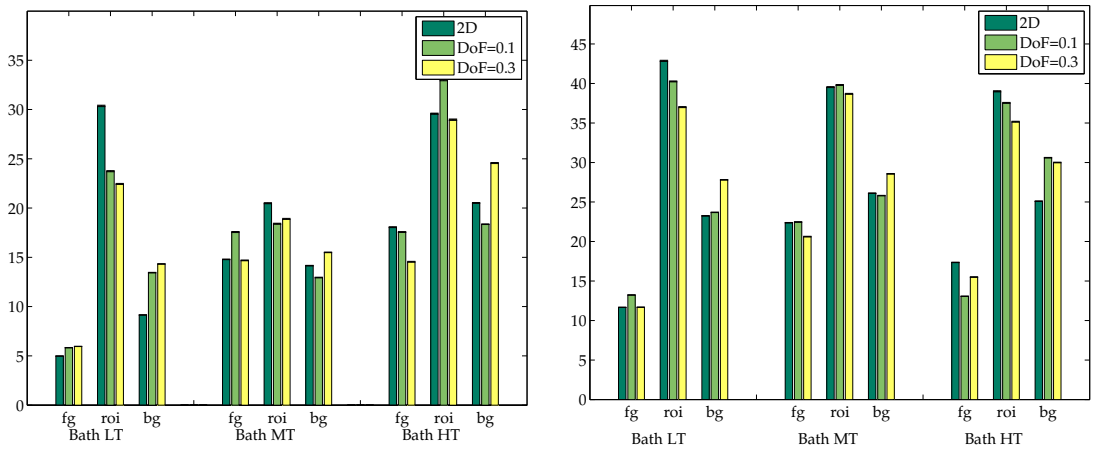
Table A.3: AUC values between 2D and 3D DoF=0.1 (2D/01); between 2D and 3D DoF=0.3 (2D/03); between 3D DoF=0.1 and 3D DoF=0.3 (01/03).

AUC	1-4 s			5-8 s			9-12 s			13-16 s			17-20 s		
Scene	2D/01	2D/03	01/03	2D/01	2D/03	01/03	2D/01	2D/03	01/03	2D/01	2D/03	01/03	2D/01	2D/03	01/03
Bathroom LT	0.9	0.88	0.88	0.75	0.74	0.79	0.79	0.78	0.82	0.76	0.73	0.81	0.75	0.73	0.75
Bathroom MT	0.84	0.82	0.87	0.79	0.8	0.77	0.78	0.73	0.77	0.71	0.76	0.84	0.72	0.76	0.75
Bathroom HT	0.86	0.85	0.86	0.77	0.74	0.81	0.77	0.79	0.81	0.77	0.76	0.83	0.8	0.7	0.8
Cartoon LT	0.86	0.87	0.85	0.85	0.87	0.84	0.86	0.84	0.85	84	0.84	0.77	0.82	0.82	0.85
Cartoon MT	0.88	0.87	0.78	0.8	0.79	0.81	0.76	0.82	0.81	0.79	0.78	0.89	0.79	0.81	0.82
Cartoon HT	0.86	0.85	0.82	0.81	0.82	0.82	0.82	0.83	0.79	0.79	0.81	0.7	0.81	0.76	0.79
Hallway LT	0.88	0.88	0.89	0.85	0.87	0.87	0.81	0.78	0.79	0.83	0.81	0.82	0.8	0.83	0.83
Hallway MT	0.86	0.85	0.83	0.79	0.8	0.83	0.79	0.76	0.83	0.78	0.78	0.7	0.81	0.86	0.72
Hallway HT	0.81	0.77	0.85	0.81	0.8	0.83	0.79	0.83	0.78	0.79	0.77	0.76	0.75	0.78	0.78
Kitchen LT	0.8	0.73	0.86	0.85	0.81	0.86	0.85	0.83	0.85	0.84	0.79	0.74	0.79	0.8	0.81
Kitchen MT	0.85	0.81	0.87	0.8	0.81	0.81	0.76	0.8	0.73	0.81	0.8	0.7	0.81	0.79	0.78
Kitchen HT	0.79	0.75	0.81	0.77	0.82	0.82	0.75	0.82	0.77	0.81	0.78	0.74	0.75	0.65	0.71
Tea LT	0.9	0.91	0.91	0.87	0.83	0.85	0.88	0.87	0.86	0.87	0.91	0.89	0.86	0.87	0.87
Tea MT	0.84	0.84	0.88	0.79	0.76	0.84	0.85	0.87	0.86	0.86	0.84	0.81	0.87	0.84	0.82
Tea HT	0.89	0.88	0.88	0.82	0.83	0.91	0.87	0.86	0.86	0.86	0.83	0.8	0.79	0.82	0.83
Room LT	0.84	0.81	0.84	0.83	0.87	0.84	0.82	0.81	0.81	0.84	0.82	0.82	0.8	0.8	0.79
Room MT	0.85	0.84	0.79	0.76	0.73	0.77	0.79	0.83	0.75	0.75	0.7	0.69	0.79	0.84	0.78
Room HT	0.88	0.86	0.83	0.79	0.79	0.7	0.83	0.77	0.76	0.79	0.74	0.65	0.71	0.72	0.72
avg	0.85	0.84	0.85	0.8	0.8	0.82	0.81	0.81	0.81	0.8	0.79	0.8	0.79	0.79	0.79
CI,±	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.02

Table A.4: CC values between 2D and 3D DoF=0.1(2D/01); between 2D and 3D DoF=0.3(2D/03); between 3D DoF=0.1 and 3D DoF=0.3(01/03).

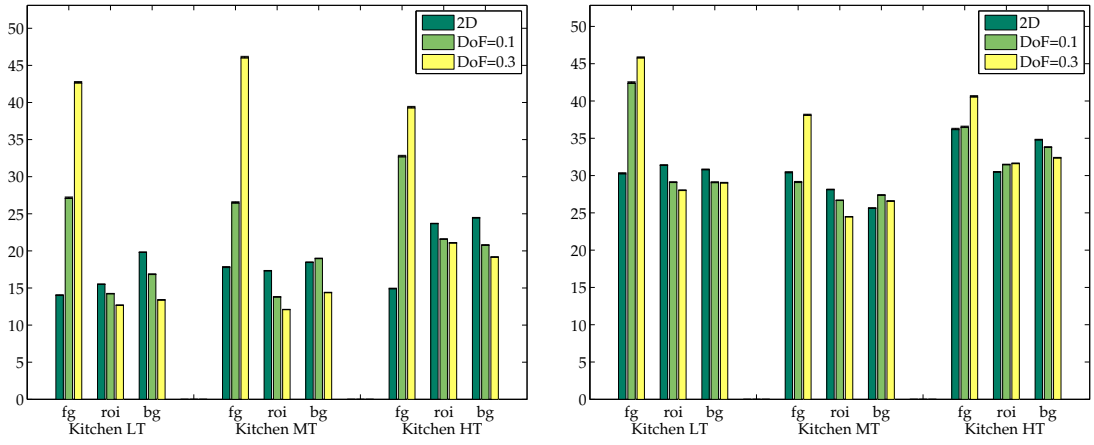
CC	1-4 s			5-8 s			9-12 s			13-16 s			17-20 s		
Scene	2D/01	2D/03	01/03	2D/01	2D/03	01/03	2D/01	2D/03	01/03	2D/01	2D/03	01/03	2D/01	2D/03	01/03
Bathroom LT	0.79	0.77	0.78	0.54	0.43	0.67	0.56	0.57	0.63	0.58	0.44	0.37	0.59	0.48	0.61
Bathroom MT	0.76	0.78	0.85	0.51	0.54	0.64	0.53	0.5	0.58	0.52	0.58	0.59	0.38	0.59	0.53
Bathroom HT	0.78	0.79	0.79	0.55	0.56	0.6	0.64	0.61	0.65	0.53	0.6	0.54	0.6	0.46	0.66
Cartoon LT	0.77	0.79	0.76	0.81	0.76	0.77	0.72	0.64	0.72	0.71	0.7	0.78	0.75	0.77	0.72
Cartoon MT	0.75	0.76	0.72	0.65	0.56	0.65	0.54	0.68	0.69	0.52	0.5	0.81	0.58	0.56	0.64
Cartoon HT	0.7	0.81	0.69	0.48	0.44	0.59	0.61	0.65	0.61	0.64	0.67	0.73	0.57	0.47	0.57
Hallway LT	0.86	0.83	0.86	0.72	0.77	0.73	0.57	0.61	0.64	0.6	0.62	0.58	0.57	0.72	0.6
Hallway MT	0.78	0.81	0.75	0.65	0.55	0.63	0.43	0.34	0.67	0.61	0.55	0.59	0.61	0.76	0.62
Hallway HT	0.71	0.73	0.79	0.6	0.46	0.54	0.52	0.58	0.67	0.61	0.54	0.65	0.42	0.57	0.51
Kitchen LT	0.65	0.51	0.73	0.77	0.5	0.71	0.66	0.58	0.67	0.64	0.48	0.63	0.5	0.5	0.61
Kitchen MT	0.79	0.65	0.75	0.63	0.62	0.72	0.36	0.53	0.41	0.61	0.52	0.66	0.54	0.38	0.64
Kitchen HT	0.61	0.47	0.66	0.53	0.58	0.55	0.34	0.71	0.48	0.58	0.54	0.41	0.61	0.29	0.46
Tea LT	0.83	0.8	0.75	0.66	0.65	0.46	0.71	0.72	0.6	0.54	0.77	0.54	0.72	0.67	0.73
Tea MT	0.65	0.64	0.74	0.55	0.65	0.69	0.58	0.58	0.61	0.59	0.54	0.62	0.5	0.6	0.57
Tea HT	0.78	0.73	0.78	0.74	0.77	0.7	0.7	0.5	0.65	0.66	0.58	0.63	0.62	0.63	0.69
Room LT	0.76	0.69	0.76	0.56	0.7	0.63	0.64	0.58	0.56	0.73	0.6	0.62	0.57	0.64	0.63
Room MT	0.78	0.61	0.75	0.65	0.56	0.63	0.39	0.67	0.44	0.56	0.51	0.41	0.43	0.65	0.51
Room HT	0.83	0.8	0.76	0.62	0.63	0.53	0.69	0.46	0.54	0.52	0.45	0.33	0.44	0.48	0.44
avg	0.75	0.72	0.76	0.62	0.6	0.64	0.57	0.59	0.6	0.6	0.57	0.58	0.56	0.57	0.6
CI, ±	0.03	0.05	0.02	0.04	0.05	0.04	0.06	0.04	0.04	0.03	0.04	0.05	0.05	0.06	0.05

A.6 Results: depth metric



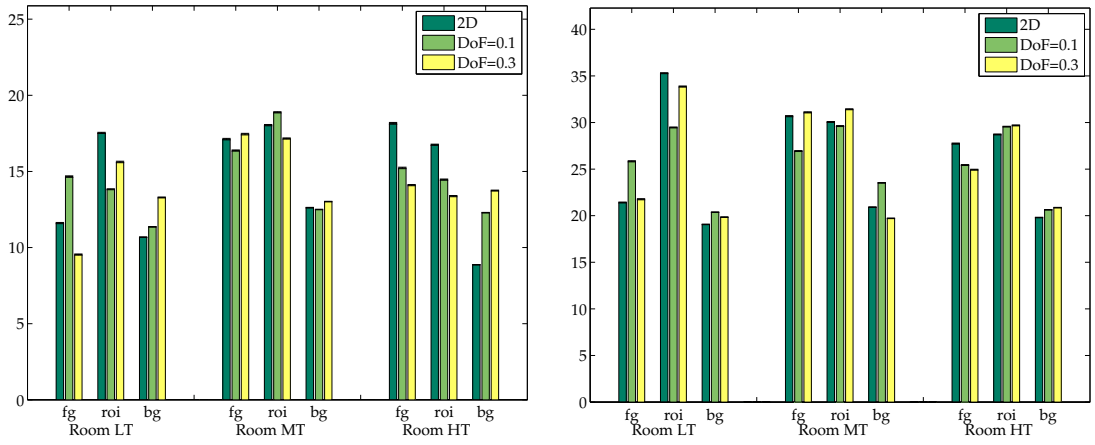
(a) 4s SM (b) 20s SM

Figure A.17: Depth metric for "Bathroom" scene (fg : $[0;60], roi$: $(60;110], bg$: $(110;255]$)



(a) 4s SM (b) 20s SM

Figure A.18: Depth metric for "Kitchen" scene (fg : $[0;110], roi$: $(110;210], bg$: $(210;255]$)



(a) 4s SM (b) 20s SM

Figure A.19: Depth metric for "Room" scene ($fg : [0; 90], roi : (90; 160], bg : (160; 255]$)

Appendix B

Supplementary information for Chapter 7

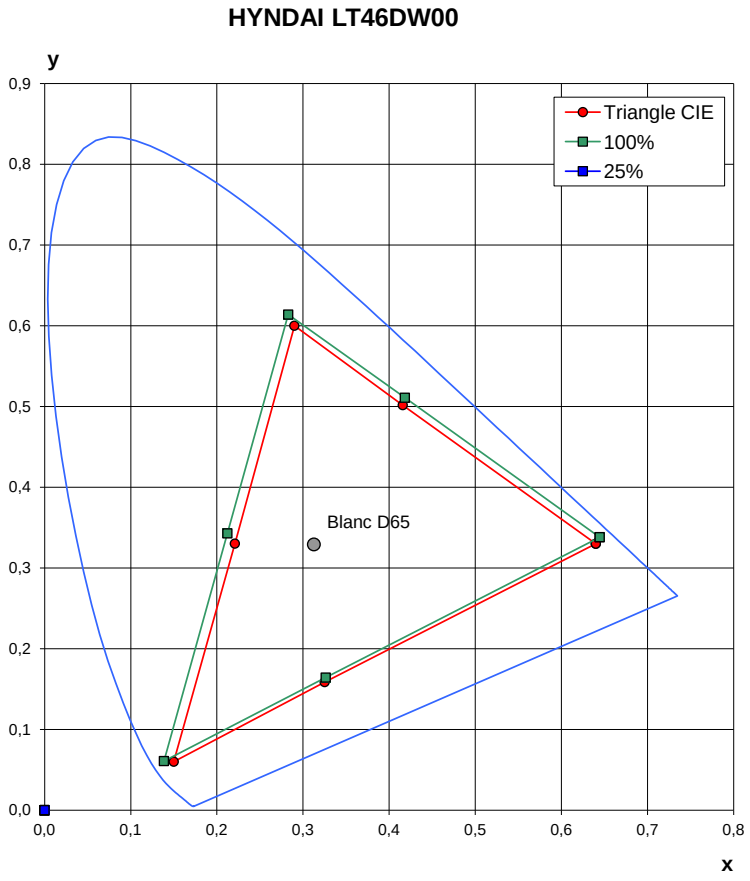


Figure B.1: Color triangle of the Hyundai display. Triangle CIE - Rec.709 color triangle

B.1 Supplementary information for Section 7.5

Table B.1: Mapping of distortion values from Impairment Scale (IS) and Acceptability Scale (AS) on Color Scale (CS)

CS	focal	shift	rotation	green	white	avg	±CI
IS(4.5)	1.86	1.77	1.87	1.95	1.85	1.86	0.06
IS(3.5)	1.1	1.21	1.12	1.56	1.22	1.24	0.16
IS(4)	1.52	1.49	1.58	1.81	1.56	1.59	0.11
AS(0.8)	1.62	1.42	1.65	1.71	1.55	1.59	0.1
AS(0.5)	1.2	1.03	1.28	1.35	1.18	1.2	0.11
AS(0.2)	0.72	0.63	0.8	0.84	0.77	0.75	0.07

Table B.2: Mapping of distortion values from Impairment Scale (IS) and Acceptability Scale (AS) on Color Scale acceptability (CSacc)

CSacc	focal	shift	rotation	green	white	avg	±CI
IS(4.5)	0.98	0.99	0.98	1	0.99	0.99	0.01
IS(3.5)	0.83	0.92	0.79	0.96	0.9	0.88	0.06
IS(4)	0.93	0.96	0.92	0.99	0.96	0.95	0.03
AS(0.8)	0.95	0.95	0.93	0.98	0.96	0.95	0.02
AS(0.5)	0.86	0.87	0.84	0.93	0.88	0.88	0.03
AS(0.2)	0.68	0.7	0.66	0.77	0.72	0.71	0.04

Table B.3: Mapping of distortion values from Impairment Scale (IS) and Acceptability Scale (AS) on Color Scale annoyance (CSann)

CSann	focal	shift	rotation	green	white	avg	±CI
IS(4.5)	0.92	0.81	0.9	0.96	0.89	0.89	0.05
IS(3.5)	0.16	0.33	0.22	0.58	0.28	0.31	0.14
IS(4)	0.5	0.53	0.56	0.81	0.56	0.59	0.11
AS(0.8)	0.62	0.48	0.64	0.72	0.55	0.6	0.08
AS(0.5)	0.21	0.23	0.31	0.43	0.26	0.29	0.08
AS(0.2)	0.04	0.09	0.1	0.18	0.1	0.1	0.04

Appendix C

Instruction sheets used in subjective experiments

C.1 Images: Color Scale experiment

Bonjour,

Au cours de ce test subjectif, vous allez évaluer la qualité d'images stéréoscopiques (3D avec plus ou moins de relief) en utilisant une échelle (port de lunettes spéciales indispensable). Lors de votre évaluation, des critères tels que le confort visuel ressenti ou encore l'acceptabilité des séquences visualisées devront être pris en considération. L'échelle de notation utilisée est une échelle à trois niveaux.

Pour évaluer le confort visuel, la sensation d'une gêne visuelle plus ou moins importante mais aussi l'impression d'un effort visuel pour passer d'une zone de l'image à une autre peuvent être prise en compte.

Concernant l'évaluation de l'acceptabilité vidéo, une séquence peut être considérée comme inacceptable par la présence d'une gêne visuelle fortement désagréable ou insupportable comme par exemple, l'envie de fermer les yeux ou d'éviter de regarder l'écran.

Au cours du test, vous aurez à évaluer successivement 3 scènes comportant chacune 5 séquences et 1 référence : Ref, A, B, C, D et E

Pour cela, vous devez reporter votre opinion sur une échelle comportant 3 niveaux repérés par les termes suivants au tableau ci-dessous.

La durée de présentation de chaque image est de 8 Sec. Vous avez la possibilité de rejouer chaque séquence et d'ajuster votre note.

Pas de gêne visuelle – acceptable	2
Gêne visuelle mais acceptable	1
Inacceptable	0

Important ! Pour vous aidez à évaluer les séquences, nous vous proposons d'utiliser la méthode suivante :

1. Evaluez l'acceptabilité de la séquence vidéo. Pour cela, vous pouvez comparez la séquence à noter avec la séquence de référence. La séquence est-elle acceptable ? Si oui, passez au point suivant. Si non, choisissez la note 0 (Gêne visuelle mais **inacceptable**).

2. Evaluez maintenant la gêne visuelle de la séquence vidéo. S'il vous ressentez ou

percevez de la gêne visuelle – mettez 1 (Gêne visuelle mais acceptable). Sinon, choisissez 2.

Merci de votre participation.

C.2 Acceptability experiment

Bonjour,

Au cours de ce test subjectif, vous allez évaluer le niveau d'acceptabilité d'images 3D relief (port de lunettes spéciales indispensable). Pour vous aider dans votre évaluation, vous pouvez prendre à compte les critères suivants:

- la qualité d'image (présence de dégradations visuelles)
- le confort visuel (sensation d'une gêne visuelle)

Après évaluation des critères ci-dessus, vous devrez prendre une décision sur le niveau d'acceptabilité de l'image présentée : **Acceptable** ou **Non acceptable**.

La durée de présentation de chaque image est de 8 s. Vous avez la possibilité de rejouer chaque séquence et de réajuster vos notes.

Merci de votre participation.

C.3 Images: Doubles Scale experiment

Bonjour,

Au cours de ce test subjectif, vous allez évaluer la qualité d'images stéréoscopiques (3D avec plus ou moins de relief) en utilisant deux échelles (port de lunettes spéciales indispensable). Lors de votre évaluation, des critères tels que le confort visuel ressenti ou encore l'acceptabilité des séquences visualisées devront être pris en considération.

La première échelle est une échelle d'**acceptabilité**. Vous devez évaluer si la séquence est acceptable. ≪**Acceptable? Oui ou Non**≫. Concernant l'évaluation de l'acceptabilité vidéo, une séquence peut être considérée comme inacceptable par la présence d'une gêne visuelle fortement désagréable ou insupportable comme par exemple l'envie de fermer les yeux ou d'éviter de regarder l'écran.

La deuxième échelle est une échelle de **gêne visuelle**. Elle est affichée si vous avez trouvé que la séquence à évaluer était acceptable. Vous devez évaluer si vous avez ressenti une gêne visuelle. ≪**Gêne visuelle? Oui ou Non**≫ Concernant l'évaluation, la sensation d'une gêne visuelle plus ou moins importante mais aussi l'impression d'un effort visuel pour passer d'une zone de l'image à une autre peuvent être prise en compte.

Au cours du test, vous aurez à évaluer successivement 3 scènes comportant chacune 5 séquences et 1 référence : Ref, A, B, C, D et E

Pour cela, vous devez reporter votre opinion sur une ou deux échelles, comme expliquer précédemment. Un code numérique (0 ou 1 ou 2) apparaitra sous chaque séquence d'évaluation une fois votre vote terminé. La signification de ce code est présentée dans le tableau ci-dessous.

Pas de gêne visuelle	2
Gêne visuelle	1
Inacceptable	0

La durée de présentation de chaque image est de 8 Sec. Vous avez la possibilité de rejouer chaque séquence et d'ajuster votre note.

Important! Pour vous aidez à évaluer les séquences, nous vous proposons d'utiliser la méthode suivante :

1. Evaluez l'acceptabilité de la séquence vidéo. Pour cela, vous pouvez comparez la séquence à noter avec la séquence de référence. La séquence est-elle acceptable ? Si ≪Non≫, notez et passez à la séquence suivante. Si ≪Oui≫, notez et passez à l'échelle suivante.

2. Evaluez maintenant la gêne visuelle de la séquence vidéo. Si vous ressentez ou percevez de la gêne visuelle notez ≪Oui≫. Sinon, choisissez ≪Non≫.

Merci de votre participation.

C.4 Videos: Color Scale experiment

Bonjour,

Au cours de ce test subjectif, vous allez évaluer la qualité de vidéos stéréoscopiques en utilisant une échelle (port de lunettes spéciales indispensable). Lors de votre évaluation, des critères tels que le confort visuel ressenti ou encore l'acceptabilité des séquences visualisées devront être pris en considération. L'échelle de notation utilisée est une échelle à trois niveaux.

Pour évaluer le confort visuel, la sensation d'une gêne visuelle plus ou moins importante mais aussi l'impression d'un effort visuel pour passer d'une zone de vidéos à une autre peuvent être prise en compte.

Concernant l'évaluation de l'acceptabilité vidéo, une séquence peut être considérée comme inacceptable par la présence d'une gêne visuelle fortement désagréable ou insupportable comme par exemple, l'envie de fermer les yeux ou d'éviter de regarder l'écran.

Au cours du test, vous aurez à évaluer successivement 4 scènes comportant chacune 5 séquences et 1 référence : Ref, A, B, C, D et E

Pour cela, vous devez reporter votre opinion sur une échelle de couleurs comportant 3 niveaux repérés par les termes suivants au tableau ci-dessous.

La durée de présentation de chaque vidéo est de 15 sec. Vous avez la possibilité de rejouer chaque séquence et d'ajuster votre note.

Pas de gêne visuelle – acceptable	2
Gêne visuelle mais acceptable	1
Inacceptable	0

Important ! Pour vous aidez à évaluer les séquences, nous vous proposons d'utiliser la méthode suivante :

1. Evaluez l'acceptabilité de la séquence vidéo. Pour cela, vous pouvez comparez la séquence à noter avec la séquence de référence. La séquence est-elle acceptable ? Si oui, passez au point suivant. Si non, choisissez la note 0 (Gêne visuelle mais **inacceptable**).

2. Evaluez maintenant la gêne visuelle de la séquence vidéo. S'il vous ressentez ou percevez de la gêne visuelle – mettez 1 (Gêne visuelle mais acceptable). Sinon, choisissez 2.

Merci de votre participation.

Appendix D

Graphical interfaces of subjective experiments

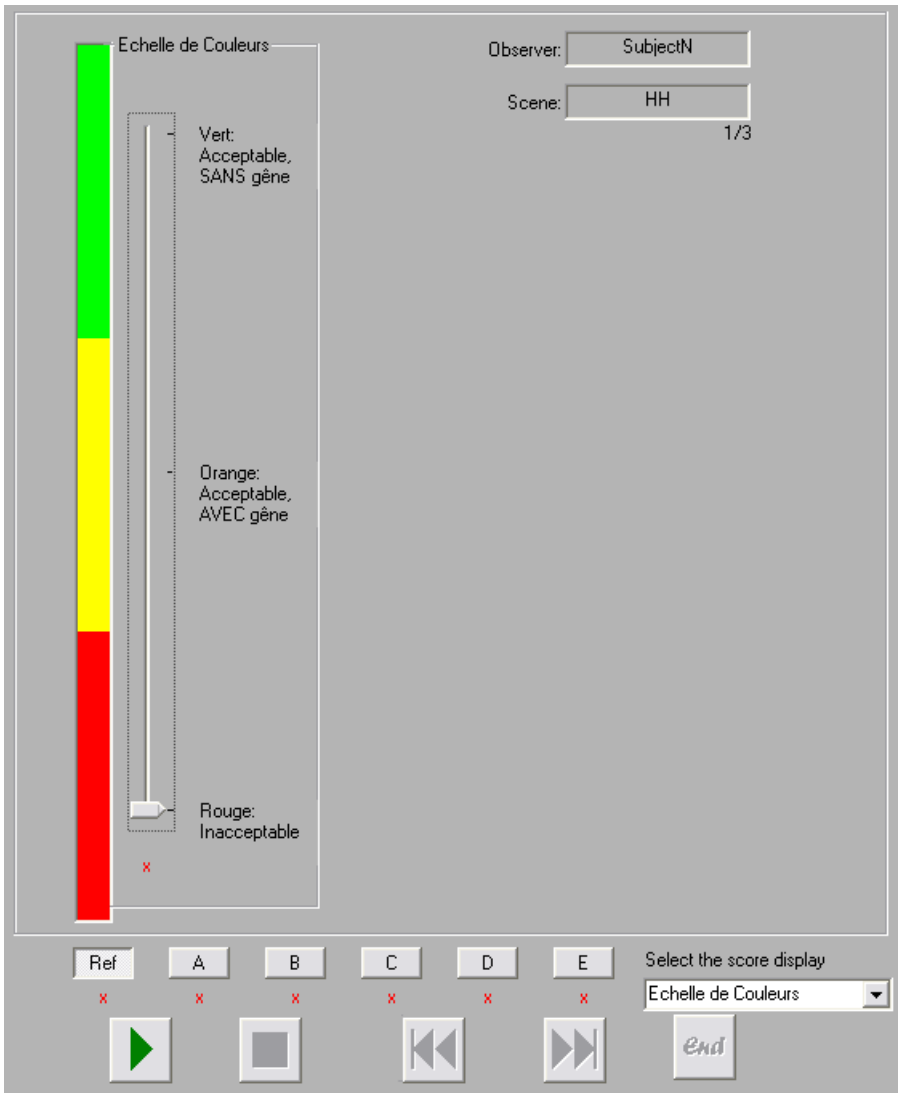


Figure D.1: The interface of the experiment with Color Scale.

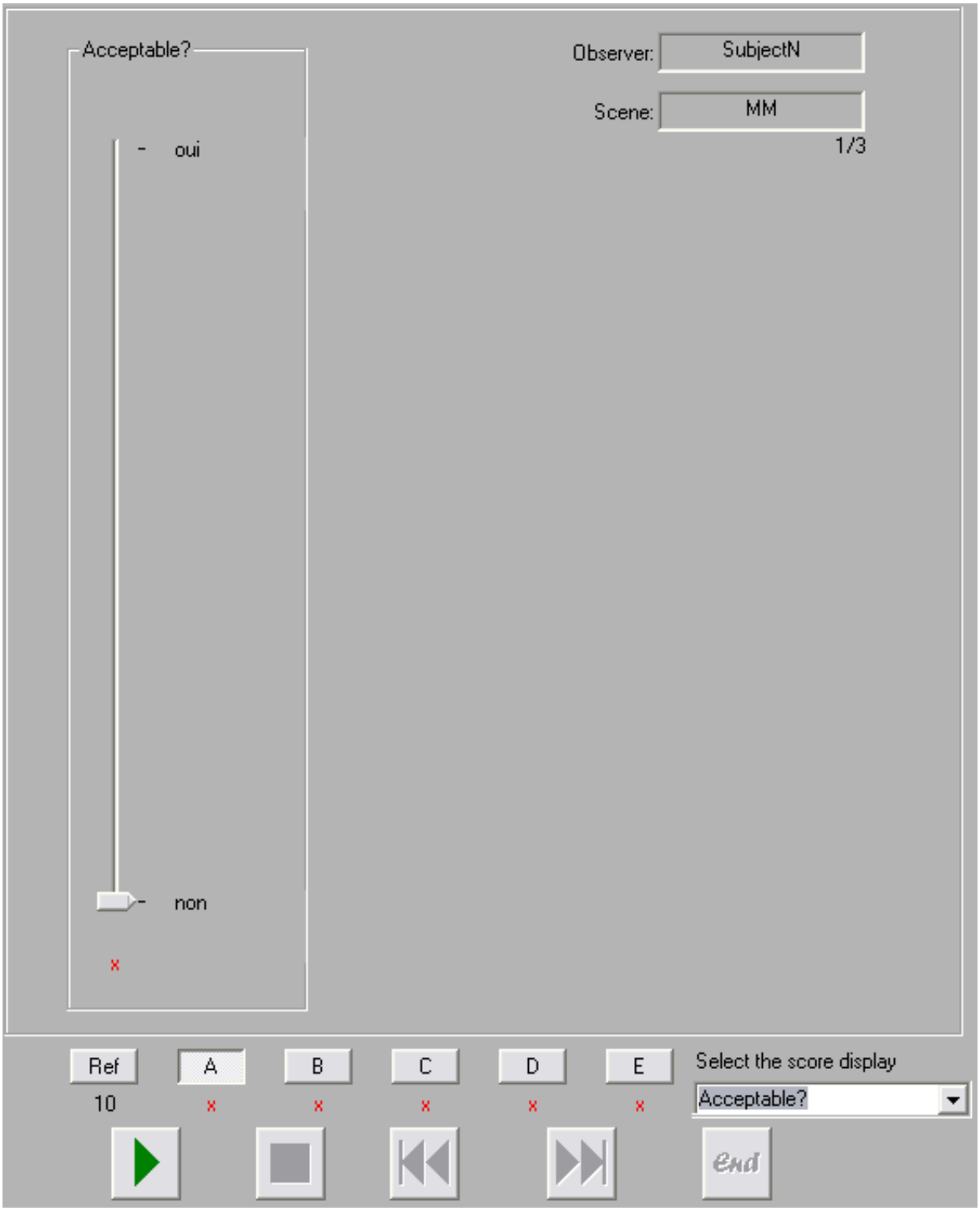


Figure D.2: The interface of the experiment with Double Scale. Visual annoyance scale is invisible.

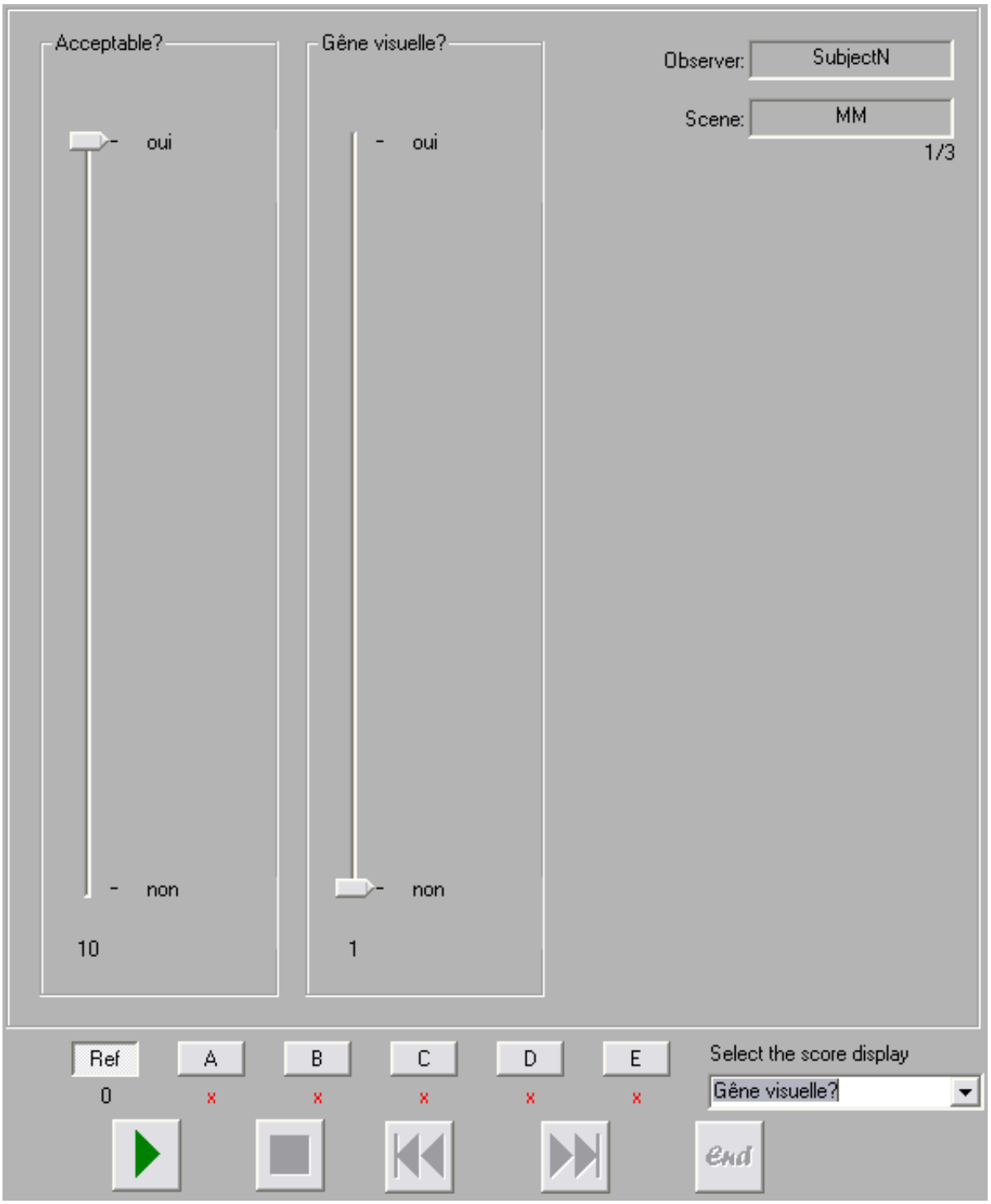


Figure D.3: The interface of the experiment with Double Scale. Subject had selected Acceptable=yes, then visual annoyance scale appeared.

Abbreviations and acronyms

3DTV - three-dimentional thelevision

AS - acceptability scale

AUC - area under the curve

CC - Pearson linear correlation coeffcient

CD - crossed disparity

CQS - continuous quality scale

CS - color scale

CSacc - acceptability scale obtained with decomposition of CS

CSann - annoyance scale obtained with decomposition of CS

DM - depth metric

DMOS - difference mean opinion scores

DoF – depth of focus

DSCQS - double stimulus continuous quality scale

FoV - field of view

FR - full-reference

G - green

HD - high definition

HDM - head mounted display

HIT - horizontal image translation

HT - high texture

HVS - human visual system

IPD - interpupillary distance

IS - impairment scale

ITU - International Telecommunication Union

LCD - liquid-crystal display

LDV - layered depth video

LT - low texture

MD - mixed disparities

MOAVI - Monitoring of Audio-Visual quality by key Indicators

MOS - mean opinion score

MSE - mean squared error

MT - middle texture

MVD - multiview video-plus-depth

MVV - multiview video

NR - no-reference

O - orange

OPSM - Objective Perceptual State Model (Metric)

PQM - Perceptual Quality Metric

PSNR - peak signal-to-noise ratio

QoE - quality of experinece

QoS - Quality of Service

R - red

ROI - region of interest

RR - reduced-reference

S3D - stereosopic 3D systems

SAMVIQ - subjective assessment methodology for video quality

SC - stimulus-comparison

SM - saliency map

SSCQE - single stimulus continuous quality evaluation

SSCQS - single stimulus continuous quality scale

SSIM - structure similarity index

SSIS - single stimulus impairment scale

Tacc - Acceptability threshold

Tann - Visual annoyance threshold

Tdis - Visual discomfort threshold

Tvis - Visibility threshold

UD - uncrossed disparity

VQEG - Video Quality Expert Group

VQM - video quality metric

WDSM- Weighted Depth Saliency Metric

ZCSBV - zone of clear binocular vision

ZoC - zone of comfort

Bibliography

[Balter et al., 2008] Balter, R., Fournier, J., Gicquel, J.-C., Kaptein, R., and Vinayagamoorthy, V. (2008). Human factors for 3d-tv. *3D4YOU. WP4 – Deliverable 4.1.*

[Banks et al., 2012] Banks, M. S., Read, J. C. A., Allison, R. S., and Watt, S. J. (2012). Stereoscopy and the human visual system. *SMPTE motion imaging journal*, 121(4)(1545-0279 (Print)):24–43.

[Banton and Levi, 1991] Banton, T. and Levi, D. M. (1991). Binocular summation in vernier acuity. *Journal of the Optical Society of America A*, 8(4):673–680.

[Barkowsky et al., 2009] Barkowsky, M., Cousseau, R., and Le Callet, P. (2009). Influence of depth rendering on the quality of experience for an autostereoscopic display. In *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, pages 192–197.

[Barlow et al., 1967] Barlow, H. B., Blakemore, C., and Pettigrew, J. D. (1967). The neural mechanism of binocular depth discrimination. *The Journal of Physiology*, 193(2):327–342.

[Benoit et al., 2008] Benoit, A., Le Callet, P., Campisi, P., and Cousseau, R. (2008). Using disparity for quality assessment of stereoscopic images. In *15th IEEE International Conference on Image Processing (ICIP).*, pages 389–392.

[Bensalma and Larabi, 2010] Bensalma, R. and Larabi, C. (2010). A stereoscopic quality metric based on binocular perception. In *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*, pages 41–44.

[Bensalma and Larabi, 2013] Bensalma, R. and Larabi, M.-C. (2013). A perceptual metric for stereoscopic image quality assessment based on the binocular energy. *Multidimensional Systems and Signal Processing*, 24(2):281–316.

[Berkel and Clarke, 1997] Berkel, C. V. and Clarke, J. A. (1997). Characterisation and optimisation of 3d-lcd module design.

[Bertalmio et al., 2000] Bertalmio, M., Sapiro, G., Caselles, V., and Ballester, C. (2000). Image inpainting.

[Blakemore, 1970] Blakemore, C. (1970). The range and scope of binocular depth discrimination in man. *The Journal of Physiology*, 211(3):599–622.

[Blin, 2006] Blin, J.-L. (2006). New quality evaluation method suited to multimedia context samviq. *The Second International Workshop on Video Processing and Quality Metrics for Consumer Electronic, Phoenix, Arizona.*

[Bobal57 et al., 2012] Bobal57, Robo3dguy, Jonfreer, and Jay-Artist (2011-2012). Original design of "bathroom", "cartoon", "hallway", "kitchen", "tea", "room". *http://www.blendswap.com*.

[Boev et al., 2008] Boev, A., Hollosi, D., and Gotchev, A. (2008). Classification of stereoscopic artefacts. *Technical report, Mobile3DTV Project No. 216503*, 7237.

[Borel and Doyen, 2013] Borel, T. and Doyen, D. (2013). *3D Display Technologies*. In Frédéric Dufaux, Béatrice Pesquet-Popescu et Marco Cagnazzo, editeurs,Emerging Technologies for 3D Video: Creation, Coding, Transmission and Rendering. Wiley.

[Borji and Itti, 2012] Borji, A. and Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Brooks and Hestnes, 2010] Brooks, P. and Hestnes, B. (2010). User measures of quality of experience: why being objective and quantitative is important. *Network, IEEE*, 24(2):8–13.

[Cagnazzo et al., 2013] Cagnazzo, M., Pesquet-Popescu, B., and Dufaux, F. (2013). *3D video representation and formats*. In Frédéric Dufaux, Béatrice Pesquet-Popescu et Marco Cagnazzo, editeurs,Emerging Technologies for 3D Video: Creation, Coding, Transmission and Rendering. Wiley.

[Calore et al., 2012] Calore, E., Folgieri, R., Gadia, D., and Marini, D. (2012). Analysis of brain activity and response during monoscopic and stereoscopic visualization. volume 8288, pages 82880M–82880M–12.

[Campisi et al., 2007] Campisi, P., Le Callet, P., and Marini, E. (2007). Stereoscopic images quality assessment. In *Proceedings of 15th European Signal Processing Conference (EUSIPCO'07)*.

[Chen et al., 2013] Chen, M., Cormack, L. K., and Bovik, A. C. (2013). No-reference quality assessment of natural stereopairs. *IEEE Transactions on Image Processing*, 22(9):3379–3391.

[Chen et al., 2012a] Chen, M., Kwon, D., and Bovik, A. (2012a). Study of subject agreement on stereoscopic video quality. In *IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, pages 173–176.

[Chen, 2012] Chen, W. (2012). Multidimensional characterization of quality of experience of stereoscopic 3d tv. *PhD Thesis, IRCCyN*.

[Chen et al., 2010] Chen, W., Fournier, J., Barkowsky, M., and Le Callet, P. (2010). New requirements of subjective video quality assessment methodologies for 3dtv.

[Chen et al., 2011] Chen, W., Fournier, J., Barkowsky, M., and Le Callet, P. (2011). New stereoscopic video shooting rule based on stereoscopic distortion parameters and comfortable viewing zone. In Woods, A. J., Holliman, N. S., and Dodgson, N. A., editors, *SPIE 7863, Stereoscopic Displays and Applications XXII*, volume 7863, San Francisco, California, USA. SPIE.

[Chen et al., 2012b] Chen, W., Fournier, J., Barkowsky, M., and Le Callet, P. (2012b). Quality of experience model for 3dtv. In Woods, A. J., Holliman, N. S., and Favalora,

G. E., editors, *SPIE 8288, Stereoscopic Displays and Applications XXIII*, volume 8288, pages 82881P–9, Burlingame, California, USA. SPIE.

[Chen et al., 2012c] Chen, W., Jérôme, F., Barkowsky, M., and Le Callet, P. (2012c). Exploration of quality of experience of stereoscopic images: binocular depth. In *Proceedings of the Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, pages 1–6.

[Chun et al., 2011] Chun, M. M., Golomb, J., and Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, (1545-2085 (Electronic)).

[Clopper and Pearson, 1934] Clopper, C. and Pearson, E. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26, pp. 404-413.*

[Collins et al., 2011] Collins, B., Derby, J., Dobrin, B., Eklund, D., Hays, B., Houston, J., Joblove, G., and Stephens, S. (2011). The 3d production guide. *3net*.

[Coren et al., 1994] Coren, S., Ward, L. M., and Enns, J. T. (1994). *Sensation and perception*. Harcourt Brace College Publishers.

[Coutant and Westheimer, 1993] Coutant, B. E. and Westheimer, G. (1993). Population distribution of stereoscopic ability. *Ophthalmic Physiology*, (0275-5408 (Print)).

[Cutting and Vishton, 1995] Cutting, J. E. and Vishton, P. M. (1995). Perceiving layout and knowing distances: the integration, relative potency and contextual use of different information about depth. In Epstein, W. and Rogers, S., editors, *Handbook of perception and Cognition.*, volume 5: Perception of Space and Motion, pages 69–117.

[Czuni and Kiss, 2012] Czuni, L. and Kiss, P. J. (2012). About the fixation points in stereo images. In *IEEE 3rd International Conference on Cognitive Infocommunications*, pages 143–147.

[Desimone and Duncan, 1995] Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222.

[Devernay and Beardsley, 2010] Devernay, F. and Beardsley, P. (2010). Stereoscopic cinema. In Ronfard, R. and Taubin, G., editors, *Image and Geometry Processing for 3-D Cinematography*, volume 5 of *Geometry and Computing*, pages 11–51. Springer Berlin Heidelberg.

[Dodgson, 2004] Dodgson, N. A. (2004). Variation and extrema of human interpupillary distance. volume 5291, pages 36–46.

[Doyen et al., 2012] Doyen, D., Sacré, J. J., and Blondé, L. (2012). Correlation between a perspective distortion in a s3d content and the visual discomfort perceived. *Stereoscopic Displays and Applications XXIII*, 8288:82881L–82881L–12.

[Duchowski, 2002] Duchowski, A. T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments and Computers*, pages 455–470.

[Dumbreck, 1993] Dumbreck, A. (1993). Depth of vision-3-d tv. *IEE Review*, 39(2):61–64.

[Eadie et al., 2000] Eadie, A. S., Gray, L. S., Carlin, P., and Mon-Williams, M. (2000). Modelling adaptation effects in vergence and accommodation after exposure to a simulated virtual reality stimulus. *Ophthalmic and Physiological Optics*, 20(3):242–251.

[Emoto et al., 2005] Emoto, M., Niida, T., and Okano, F. (2005). Repeated vergence adaptation causes the decline of visual functions in watching stereoscopic television. *Display Technology, Journal of*, 1(2):328–340.

[Emoto et al., 2004] Emoto, M., Nojiri, Y., and Okano, F. (2004). Changes in fusional vergence limit and its hysteresis after viewing stereoscopic tv. *Displays*, 25(2–3):67–76.

[Fehn, 2003] Fehn, C. (2003). A 3d-tv system based on video plus depth information. In *Conference on Signals, Systems and Computers*, volume 2, pages 1529–1533 Vol.2.

[Fender and Julesz, 1967] Fender, D. and Julesz, B. (1967). Extension of panum's fusional area in binocularly stabilized vision. *Journal of the Optical Society of America*, 57(6):819–826.

[Ferrin, 1999] Ferrin, F. J. (1999). Update on optical systems for military head-mounted displays. *Proc. SPIE 3689, Helmet- and Head-Mounted Displays IV*, 3689:178–185.

[Ferris, 1972] Ferris, S. H. (1972). Motion parallax and absolute distance. *Journal of experimental psychology.*

[Findlay and Gilchrist, 2003] Findlay, J. M. and Gilchrist, I. D. (2003). *Active Vision: The Psychology of Looking and Seeing.* Oxford Psychology Series. Oxford University Press, Oxford.

[Fournier, 1995] Fournier, J. (1995). Etude de la qualité des images stéréoscopiques en télévision.

[Freeman and Avons, 2000] Freeman, J. and Avons, S. E. (2000). Focus group exploration of presence through advanced broadcast services. volume 3959, pages 530–539.

[Fukushima et al., 2009] Fukushima, T., Ukai, K., Wolffsohn, J. S., and Gilmartin, B. (2009). The relationship between ca/c ratio and individual differences in dynamic accommodative responses while viewing stereoscopic images. *Journal of Vision*, (1534-7362 (Electronic)).

[Gautier et al., 2010] Gautier, J., Bosc, E., and Morin, L. (2010). Representation and coding of 3d video data. *Project PERSEE - Schémas perceptuels et codage vidéo 2D et 3D*, page 43.

[Gautier and Le Meur, 2012] Gautier, J. and Le Meur, O. (2012). A time-dependent saliency model combining center and depth biases for 2d and 3d viewing conditions. *Cognitive Computation*, pages 141–156.

[Goldmann et al., 2010a] Goldmann, L., De Simone, F., and Ebrahimi, T. (2010a). Impact of acquisition distortion on the quality of stereoscopic images.

[Goldmann et al., 2010b] Goldmann, L., Jong-Seok, L., and Ebrahimi, T. (2010b). Temporal synchronization in stereoscopic video: Influence on quality of experience and automatic asynchrony detection. In *17th IEEE International Conference on Image Processing (ICIP)*, pages 3241–3244.

[Goldstein, 2013] Goldstein, E. (2013). *Sensation and Perception*. Cengage Learning.

[Gorley and Holliman, 2008] Gorley, P. and Holliman, N. (2008). Stereoscopic image quality metrics and compression. volume 6803, pages 680305–680305–12.

[Green and Swets, 1966] Green, D. and Swets, J. (1966). *Signal detection theory and psychophysics*. Wiley, New York.

[Grondzik,] Grondzik, W. Visual comfort. *http://pages.uoregon.edu/hof/s00moa/pages/Visual%20Comfort.html*.

[Grossauer and Scherzer, 2003] Grossauer, H. and Scherzer, O. (2003). Using the complex ginzburg-landau equation for digital inpainting in 2d and 3d. In Griffin, L. and Lillholm, M., editors, *Scale Space Methods in Computer Vision*, volume 2695 of *Lecture Notes in Computer Science*, pages 225–236. Springer Berlin Heidelberg.

[Guillemot and Le Meur, 2014] Guillemot, C. and Le Meur, O. (2014). Image inpainting: Overview and recent advances. *Signal Processing Magazine, IEEE*, 31(1):127–144.

[Gunnewiek and Vandewalle, 2010] Gunnewiek, R. K. and Vandewalle, P. (2010). How to display 3d content realistically.

[Hakkinen et al., 2010] Hakkinen, J., Kawai, T., Takatalo, J., Mitsuya, R., and Nyman, G. (2010). What do people look at when they watch stereoscopic movies? *Stereoscopic Displays and Applications XXI*, pages 75240E–10.

[Hewage et al., 2008] Hewage, C. T. E. R., Worrall, S. T., Dogan, S., and Kondoz, A. M. (2008). Prediction of stereoscopic video quality using objective quality models of 2-d video. *Electronics Letters*, 44(16):963–965.

[Hiruma and Fukuda, 1993] Hiruma, N. and Fukuda, T. (1993). Accommodation response to binocular stereoscopic tv images and their viewing conditions. *SMPTE Journal*, 102(12):1137–1140.

[Hoffman et al., 2011] Hoffman, D. M., Karasev, V. I., and Banks, M. S. (2011). Temporal presentation protocols in stereoscopic displays: Flicker visibility, perceived motion, and perceived depth. *Journal of the Society for Information Display*, 19(3):271–297.

[Holliman, 2003] Holliman, N. (2003). 3d display systems. *Handbook of Optoelectronics, Institute of Physics Press, ISBN 0 7503 0646*, 7.

[Holliman, 2004] Holliman, N. S. (2004). Mapping perceived depth to regions of interest in stereoscopic images. In Woods, A. J., Merritt, J. O., Benton, S. A., and Bolas, M. T., editors, *Stereoscopic displays and virtual reality systems XI.*, Proceedings of SPIE, pages 117–128. SPIE, Bellingham, WA.

[Howard, 2012] Howard, I. P. (2012). *Perceiving in Depth, Volume 3: Other Mechanisms of Depth Perception*. Oxford Psychology Series. OUP USA.

[Howarth and Costello, 1997] Howarth, P. A. and Costello, P. J. (1997). The occurrence of virtual simulation sickness symptoms when an hmd was used as a personal viewing system. *Displays*, 18(2):107–116.

[Hubel and Wiesel, 1970] Hubel, D. H. and Wiesel, T. N. (1970). Stereoscopic vision in macaque monkey: Cells sensitive to binocular depth in area 18 of the macaque monkey cortex. *Nature*, 225(5227):41–42.

[Huynh-Thu et al., 2010] Huynh-Thu, Q., Le Callet, P., and Barkowsky, M. (2010). Video quality assessment: From 2d to 3d. challenges and future trends. In *17th IEEE International Conference on Image Processing (ICIP)*, pages 4025–4028.

[Huynh-Thu and Schiatti, 2011] Huynh-Thu, Q. and Schiatti, L. (2011). Examination of 3d visual attention in stereoscopic video content. *SPIE 7865, Human Vision and Electronic Imaging XVI*, pages 78650J–78650J.

[Iatsun et al., 2013] Iatsun, I., Larabi, M.-C., and Fernandez-Maloigne, C. (2013). Investigation of visual fatigue/discomfort generated by s3d video using eye-tracking data. In *SPIE 8648, Stereoscopic Displays and Applications XXIV*.

[Ijsselsteijn et al., 1998] Ijsselsteijn, W., de Ridder, H., Hamberg, R., Bouwhuis, D., and Freeman, J. (1998). Perceived depth and the feeling of presence in 3dtv. *Displays*, 18(4):207–214.

[Ijsselsteijn, 2004] Ijsselsteijn, W. A. (2004). *Presence in Depth*. PhD thesis, Eindhoven University of Technology.

[Ijsselsteijn et al., 2000] Ijsselsteijn, W. A., Ridder, H., and Vliegen, J. (2000). Subjective evaluation of stereoscopic images: effects of camera parameters and display duration. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(2):225–233.

[Ikeda and Nakashima, 1980] Ikeda, M. and Nakashima, Y. (1980). Wavelength difference limit for binocular color fusion. *Vision Research*, 20(8):693–697.

[Inoue and Ohzu, 1997] Inoue, T. and Ohzu, H. (1997). Accommodative responses to stereoscopic three-dimensional display. *Applied Optics*, 36(19):4509–4515.

[Ion-Paul and Hanna, 1990] Ion-Paul, B. and Hanna, B. (1990). Influence des distorsions géométriques dans les images stéréoscopiques sur le confort visuel de l'observateur. *La Télévision en Relief*, (CCETT Rennes).

[ISO, 2007] ISO (2007). International standard iso/iec 23002-3.

[ISO, 2010] ISO (2010). International standard iso/iec14496-10.

[ITU, 2004] ITU (2004). Recommendation itu-t j.144. objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference. *ITU-T Telecommunication Standardization Bureau*.

[ITU, 2007] ITU (2007). Recommendation itu-r bt.1788.methodology for the subjective assessment of video quality in multimedia applications. *Broadcasting service (television)*.

[ITU, 2008] ITU (2008). Recommendation itu-t p.10/g.100 amendment 2: New definitions for inclusion in recommendation itu-t p.10/g.100. *Telecommunication Standardization Bureau*.

[ITU, 2010] ITU (2010). Recommendation itu-t h.264. advanced video coding for generic audiovisual services. *Telecommunication Standardization Bureau*.

[ITU, 2011] ITU (2011). Recommendation itu-t p.10/g.100 amendment 3: New definitions for inclusion in recommendation itu-t p.10/g.100. *Telecommunication Standardization Bureau*.

[ITU, 2012a] ITU (2012a). Recommendation itu-r bt.2021. subjective methods for the assessment of stereoscopic 3dtv systems. *Broadcasting service (television)*.

[ITU, 2012b] ITU (2012b). Recommendation itu-r bt.500-13. methodology for the subjective assessment of the quality of television pictures. *Broadcasting service (television)*.

[James, 1890] James, W. (1890). *The Principles of Psychology*. Dover Publications.

[Jansen et al., 2009] Jansen, L., Onat, S., and König, P. (2009). Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*.

[Jones et al., 2001] Jones, G. R., Lee, D., Holliman, N. S., and Ezra, D. (2001). Controlling perceived depth in stereoscopic images. volume 4297, pages 42–53, San Jose, CA, USA. SPIE.

[Joveluro et al., 2010] Joveluro, P., Malekmohamadi, H., Fernando, W. A. C., and Kondoz, A. M. (2010). Perceptual video quality metric for 3d video quality assessment. In *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2010*, pages 1–4.

[Judd et al., 2009] Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. In *IEEE International Conference on Computer Vision*, pages 2106–2113.

[Julesz, 1971] Julesz, B., editor (1971). *Foundations of Cyclopean Perception*, volume 406 of *Science*. University of Chicago Press, Chicago.

[Julesz and Schumer, 1981] Julesz, B. and Schumer, R. A. (1981). Early visual perception. *Annual Review of Psychology*, 32(1):575–627.

[Kaminsky, 2011] Kaminsky, M. S. (2011). *Shoot 3D Video Like a Pro: 3D Camcorder Tips, Tricks and Secrets: The 3D Movie Making Guide They Forgot to Include*. Organik Media, Incorporated.

[Kaplan, 1969] Kaplan, G. (1969). Kinetic disruption of optical texture: The perception of depth at an edge. *Perception and Psychophysics*, 6(4):193–198.

[Kaptein et al., 2008] Kaptein, R. G., Kuijsters, A., Lambooij, M. T. M., Ijsselsteijn, W. A., and Heynderickx, I. (2008). Performance evaluation of 3d-tv systems. volume 6808, pages 680819–11, San Jose, CA, USA. SPIE.

[Kauff et al., 2007] Kauff, P., Atzpadin, N., Fehn, C., Muller, M., Schreer, O., Smolic, A., and Tanger, R. (2007). Depth map creation and image-based rendering for advanced 3dtv services providing interoperability and scalability. *Image Commun.*, 22(2):217–234.

[Kennedy et al., 1993] Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, 3(3):203–220.

[Kim et al., 2011a] Kim, D., Choi, S., Park, S., and Sohn, K. (2011a). Stereoscopic visual fatigue measurement based on fusional response curve and eye-blinks. In *17th International Conference on Digital Signal Processing (DSP)*, pages 1–6.

[Kim et al., 2011b] Kim, D., Jung, Y. J., Kim, EunwooRo, Y.-M., and Park, H. (2011b). Human brain response to visual fatigue caused by stereoscopic depth perception. In *17th International Conference on Digital Signal Processing (DSP)*, pages 1–5.

[Kim and Lee, 2011] Kim, Y.-J. and Lee, E. (2011). Eeg based comparative measurement of visual fatigue caused by 2d and 3d displays. In Stephanidis, C., editor, *HCI International 2011 – Posters' Extended Abstracts*, volume 174 of *Communications in Computer and Information Science*, pages 289–292. Springer Berlin Heidelberg.

[Kooi and Toet, 2004] Kooi, F. L. and Toet, A. (2004). Visual comfort of binocular and 3d displays. *Displays*, 25(2):99–108.

[Kozamernik et al., 2005] Kozamernik, F., Steinmann, V., Sunna, P., and Wyckens, E. (2005). Samviq—a new ebu methodology for video quality evaluations in multimedia. *SMPTE motion imaging journal*, 114(4):152–160.

[Kuze and Ukai, 2008] Kuze, J. and Ukai, K. (2008). Subjective evaluation of visual fatigue caused by motion images. *Displays*, 29(2):159–166.

[Lam et al., 2002] Lam, A. K., P., T., Choy, E., and Chung, M. (2002). Crossed and uncrossed stereoacuity at distance and the effect from heterophoria. *Ophthalmic and physiological optics : the journal of the British College of Ophthalmic Opticians (Optometrists)*, 22(0275-5408 (Print)).

[Lambooij et al., 2009] Lambooij, M., Fortuin, M., Ijsselsteijn, W. A., and Heynderickx, I. (2009). Measuring visual discomfort associated with 3d displays. In *SPIE 7237, Stereoscopic Displays and Applications XX*, pages 72370K–72370K.

[Lambooij et al., 2011] Lambooij, M., Ijsselsteijn, W., Bouwhuis, D. G., and Heynderickx, I. (2011). Evaluation of stereoscopic images: Beyond 2d quality. *IEEE Transactions on Broadcasting*, 57(2):432–444.

[Lambooij et al., 2013] Lambooij, M., Murdoch, M. J., Ijsselsteijn, W. A., and Heynderickx, I. (2013). The impact of video characteristics and subtitles on visual comfort of 3d tv. *Displays*, 34(1):8–16.

[Lambooij et al., 2007] Lambooij, M. T. M., Ijsselsteijn, W. A., and Heynderickx, I. (2007). Visual discomfort in stereoscopic displays: a review. *Stereoscopic Displays and Virtual Reality Systems XIV*, 6490:64900I–13.

[Laszlo Szirmay-Kalos, 1996] Laszlo Szirmay-Kalos, Z. R. B. A. S. (1996). *Theory of Three Dimensional Computer Graphics*. Akademiai Kiado, Budapest.

[Le Callet et al., 2012] Le Callet, P., Möller, S., and Perkis, A. (2012). Qualinet white paper on definitions of quality of experience. *European Network on Quality of Experience in Multimedia Systems and Services*.

[Le Meur, 2012] Le Meur, O. (2012). Fixation analysis software. *http://people.irisa.fr/Olivier.Le_Meur/publi/2012_BRM/index2.html#soft*.

[Le Meur, 2014] Le Meur, O. (2014). Visual attention modelling and applications. towards perceptual-based editing methods. *HDR*.

[Le Meur and Baccino, 2012] Le Meur, O. and Baccino, T. (2012). Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, pages 1–16.

[Le Meur et al., 2011] Le Meur, O., Baccino, T., and Roumy, A. (2011). Prediction of the inter-observer visual congruency (iovc) and application to image ranking.

[Lee et al., 2006] Lee, C., Cho, S., Choe, J., Jeong, T., Ahn, W., and Lee, E. (2006). Objective video quality assessment. *Optical Engineering*, 45(1):017004–017004–11.

[Legge and Yuanchao, 1989] Legge, G. E. and Yuanchao, G. (1989). Stereopsis and contrast. *Vision Research*, 29(8):989–1004.

[Leszczuk et al., 2014] Leszczuk, M., Hanusiak, M., Blanco, I., Dziech, A., Derkacz, J., Wyckens, E., and Borer, S. (2014). Key indicators for monitoring of audiovisual quality. In *Signal Processing and Communications Applications Conference (SIU)*, pages 2301–2305.

[Li et al., 2008] Li, H. C. O., Junho, S., Keetaek, K., and Seunghyun, L. (2008). Measurement of 3d visual fatigue using event-related potential (erp): 3d oddball paradigm. In *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2008*, pages 213–216.

[Li et al., 2012] Li, J., Barkowsky, M., and Le Callet, P. (2012). Analysis and improvement of a paired comparison method in the application of 3dtv subjective experiment. In *19th IEEE International Conference on Image Processing (ICIP)*, pages 629–632.

[Li et al., 2013] Li, J., Barkowsky, M., and Le Callet, P. (2013). Visual discomfort is not always proportional to eye blinking rate: exploring some effects of planar and in-depth motion on 3dtv qoe. In *Proceedings of VPQM 2013*, pages pp.1–6.

[Li et al., 2011] Li, J., Barkowsky, M., Wang, J., and Le Callet, P. (2011). Study on visual discomfort induced by stimulus movement at fixed depth on stereoscopic displays using shutter glasses. In *17th International Conference on Digital Signal Processing (DSP)*, pages 1–8.

[Lipton, 1997] Lipton, L. (1997). *StereoGraphics Developers' Handbook*. StereoGraphics Corporation.

[Lotufo De Alengar et al., 1998] Lotufo De Alengar, R., Da Silva, W. D. F., Falcao, A. X., and Pessoa, A. C. F. (1998). Morphological image segmentation applied to video quality assessment. In *International Symposium on Computer Graphics, Image Processing, and Vision.*, pages 468–475.

[Mannan et al., 1995] Mannan, S., Ruddock, K. F., and Wooding, D. S. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-d images. *Spatial Vision*.

[McCarthy, 2010] McCarthy, S. (2010). Glossary for video and perceptual quality of stereoscopic video. *3D@Home Consortium and the MPEG Industry Forum 3DTV Working Group*.

[Meesters et al., 2004] Meesters, L. M. J., Ijsselsteijn, W. A., and Seuntiëns, P. J. H. (2004). A survey of perceptual evaluations and requirements of three-dimensional tv. *IEEE Transactions on Circuits and Systems for Video Technology.*, 14(3):381–391.

[Mendiburu, 2009] Mendiburu, B. (2009). *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*. Focal Press.

[Merkle et al., 2007] Merkle, P., Smolic, A., Muller, K., and Wiegand, T. (2007). Efficient prediction structures for multiview video coding. *IEEE Transactions on Circuits and Systems for Video Technology.*, 17(11):1461–1473.

[Middlebury, 2014] Middlebury (2014). http://vision.middlebury.edu/stereo/eval/. *Middlebury Stereo Evaluation*.

[Mitchell, 1966] Mitchell, D. E. (1966). Retinal disparity and diplopia. *Vision Research*, 6(7–8):441–451.

[Monteiro et al., 2013] Monteiro, Wfg5001, Robo3dguy, and Jay-Artist (2011-2013). Original design of "cartoon", "hall", "pigs", "table". *http://www.blendswap.com*.

[Muller et al., 2010] Muller, K., Merkle, P., Tech, G., and Wiegand, T. (2010). 3d video formats and coding methods. In *17th IEEE International Conference on Image Processing (ICIP).*, pages 2389–2392.

[Murata et al., 2001] Murata, A., Uetake, A., Otsuka, M., and Takasawa, Y. (2001). Proposal of an index to evaluate visual fatigue induced during visual display terminal tasks. *International Journal of Human-Computer Interaction*, 13(3):305–321.

[Nielsen and Poggio, 1984] Nielsen, K. R. K. and Poggio, T. (1984). Vertical image registration in stereopsis. *Vision Research*, 24(10):1133–1140.

[Nojiri et al., 2004] Nojiri, Y., Yamanoue, H., Hanazato, A., Emoto, M., and Okano, F. (2004). Visual comfort/discomfort and visual fatigue caused by stereoscopic hdtv viewing. volume 5291, pages 303–313.

[Nojiri et al., 2003] Nojiri, Y., Yamanoue, H., Hanazato, A., and Okano, F. (2003). Measurement of parallax distribution and its application to the analysis of visual comfort for stereoscopic HDTV. In Woods, A., Bolas, M., Merritt, J., and Benton, S., editors, *Proc. Stereoscopic Displays and Virtual Reality Systems X*, volume 5006, pages 195–205. SPIE.

[NTT, 2014] NTT (2014). The perception of video quality degradation. *NTT Information Network Laboratory Group. Video quality assessment methods*, http://www.ntt.co.jp/qos/qoe/eng/technology/visual/01_2.html.

[Okada et al., 2006] Okada, Y., Ukai, K., Wolffsohn, J. S., Gilmartin, B., Iijima, A., and Bando, T. (2006). Target spatial frequency determines the response to conflicting defocus- and convergence-driven accommodative stimuli. *Vision Research*, 46(4):475–484.

[OPTICOM, 2008] OPTICOM (2008). "pevq - perceptual evaluation of video quality". *http://www.pevq.org/*.

[Patterson, 2007] Patterson, R. (2007). Human factors of 3-d displays. *Journal of the Society for Information Display*, 15(11):861–871.

[Percival, 1892] Percival, A. S. (1892). The relation of convergence to accommodation and its practical bearing. *Ophthalmological Review*, 11:313–328.

[Phillips and Stark, 1977] Phillips, S. and Stark, L. (1977). Blur: a sufficient accommodative stimulus. *Documenta Ophthalmologica*, 43(0012-4486 (Print)):65–89.

[Pinson and Wolf, 2004] Pinson, M. H. and Wolf, S. (2004). A new standardized method for objectively measuring video quality. *IEEE Transactions on Broadcasting*, 50(3):312–322.

[Poggio and Fischer, 1977] Poggio, G. and Fischer, B. (1977). Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 204(0022-3077 (Print)).

[Polak and Jones, 1990] Polak, N. A. and Jones, R. (1990). Dynamic interactions between accommodation and convergence. *IEEE Transactions on Biomedical Engineering.*, 37(10):1011–1014.

[Posner, 1980] Posner, M. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, 32(1):3–25.

[Qin et al., 2006] Qin, D., Takamatsu, M., and Nakashima, Y. (2006). Disparity limit for binocular fusion in fovea. *Optical Review*, 13(1):34–38.

[Ramasamy et al., 2009] Ramasamy, C., House, D. H., Duchowski, A. T., and Daugherty, B. (2009). Using eye tracking to analyze stereoscopic filmmaking. *SIGGRAPH Poster*.

[Reeve and Flock, 2010] Reeve, S. and Flock, J. (2010). Basic principles of stereoscopic 3d.

[Richards, 1970] Richards, W. (1970). Stereopsis and stereoblindness. *Experimental Brain Research*, 10:380–388.

[Rogers and Bradshaw, 1993] Rogers, B. J. and Bradshaw, M. F. (1993). Vertical disparities, differential perspective and binocular stereopsis. *Nature*, 361(6409):253–255.

[Ryu et al., 2012] Ryu, S., Kim, D. H., and Sohn, K. (2012). Stereoscopic image quality metric based on binocular perception model. In *19th IEEE International Conference on mage Processing (ICIP)*, pages 609–612.

[Sandrew, 2012] Sandrew, B. (2012). Part 2 - sitting too close to your 3d tv will make you blind?

[Scharstein and Szeliski, 2002] Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42.

[Schor and Tsuetaki, 1987] Schor, C. and Tsuetaki, T. K. (1987). Fatigue of accommodation and vergence modifies their mutual interactions. *Investigative Ophthalmology and Visual Science*, (0146-0404 (Print)).

[Schor and Tyler, 1981] Schor, C. M. and Tyler, C. W. (1981). Spatio-temporal properties of panum's fusional area. *Vision Research*, 21(5):683–692.

[Schreiber et al., 2006] Schreiber, K. M., Tweed, D. B., and Schor, C. M. (2006). The extended horopter: Quantifying retinal correspondence across changes of 3d eye position. *Journal of Vision*, 6(1).

[Seuntiëns et al., 2008] Seuntiëns, P., Heynderickx, I., and Ijsselsteijn, W. (2008). Capturing the added value of three-dimensional television: Viewing experience and naturalness of stereoscopic images. *Journal of Imaging Science*, 52(2):20504–1–20504–5.

[Seuntiëns et al., 2006] Seuntiëns, P., Meesters, L., and Ijsselsteijn, W. (2006). Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric jpeg coding and camera separation. *ACM Trans. Appl. Percept.*, 3(2):95–109.

[Seuntiëns, 2006] Seuntiëns, P. J. H. (2006). Visual experience of 3d tv. *PhD Thesis, Technische Universiteit Eindhoven.*

[Seuntiëns et al., 2005] Seuntiëns, P. J. H., Heynderickx, I., Ijsselsteijn, W. A., van den Avoort, P. M. J., Berentsen, J., Dalm, I. J., Lambooij, M. T. M., and Oosting, W. (2005). Viewing experience and naturalness of 3d images.

[Shade et al., 1998] Shade, J., Gortler, S., He, L.-w., and Szeliski, R. (1998). Layered depth images.

[Shibata et al., 2011] Shibata, T., Kim, J., Hoffman, D. M., and Banks, M. S. (2011). The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of Vision*, 11(8).

[Smith and Collar, 2012] Smith, M. D. and Collar, B. T. (2012). Perception of size and shape in stereoscopic 3d imagery. volume 8288, pages 82881O–82881O–31.

[Smith and Malia, 2013] Smith, M. D. and Malia, J. (2013). Controlling miniaturization in stereoscopic 3d imagery. *SMPTE Conferences*, 2013(10):1–13.

[Sohn et al., 2013] Sohn, H., Yong Ju, J., Seong-Il, L., and Yong Man, R. (2013). Predicting visual discomfort using object size and disparity information in stereoscopic images. *IEEE Transactions on Broadcasting*, 59(1):28–37.

[Solimini et al., 2012] Solimini, A., Mannocci, A., Di Thiene, D., and La Torre, G. (2012). A survey of visually induced symptoms and associated factors in spectators of three dimensional stereoscopic movies. *BMC Public Health*, 12(1):1–11.

[Solimini, 2013] Solimini, A. G. (2013). Are there side effects to watching 3d movies? a prospective crossover observational study on visually induced motion sickness. *PLoS ONE*, 8(2):e56160.

[Soper, 2014] Soper, D. (2014). Binomial probability confidence interval calculator [software]. *http://www.danielsoper.com/statcalc*.

[Speranza et al., 2006] Speranza, F., Tam, W. J., Renaud, R., Hur, N., Dodgson, N. A., Merritt, J. O., Bolas, M. T., and McDowall, I. (2006). Effect of disparity and motion on visual comfort of stereoscopic images.

[Stark et al., 1980] Stark, L., Kenyon, R. V., Krishnan, V. V., and Ciuffreda, K. J. (1980). Disparity vergence: a proposed name for a dominant component of binocular vergence eye movements. *American Journal of Optometry and Physiological Optics*, (0093-7002 (Print)):3.

[Stelmach and Tam, 1998] Stelmach, L. and Tam, J. W. (1998). Stereoscopic image coding: Effect of disparate image-quality in left- and right-eye views. *Signal Processing: Image Communication*, 14(1–2):111–117.

[Stelmach et al., 2000] Stelmach, L., Tam, W. J., Meegan, D., and Vincent, A. (2000). Stereo image quality: effects of mixed spatio-temporal resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(2):188–193.

[Stern et al., 1994] Stern, J. A., Boyer, D., and Schroeder, D. (1994). Blink rate: a possible measure of fatigue. (0018-7208 (Print)).

[Stevenson and Schor, 1997] Stevenson, S. B. and Schor, C. M. (1997). Human stereo matching is not restricted to epipolar lines. *Vision Research*, 19(22):2717–2723.

[Suryakumar et al., 2007] Suryakumar, R., Meyers, J. P., Irving, E. L., and Bobier, W. R. (2007). Vergence accommodation and monocular closed loop blur accommodation have similar dynamic characteristics. *Vision Research*, 47(3):327–337.

[SwissQual, 2010] SwissQual (2010). Vquad-hd - objective perceptual multimedia video quality measurement of hdtv. *http://www.vquad-hd.info/*.

[Takaya, 2010] Takaya, K. (2010). Algorithm to realize real-time dense disparity map of stereo vision with the embedded video system for distance sensing applications. In *Proceedings of 19th International Conference on Computer Communications and Networks (ICCCN)*, pages 1–4.

[Tam et al., 2011] Tam, W. J., Speranza, F., Yano, S., Shimono, K., and Ono, H. (2011). Stereoscopic 3d-tv: Visual comfort. *IEEE Transactions on Broadcasting*, 57(2):335–346.

[Tam et al., 1998] Tam, W. J., Stelmach, L. B., and Corriveau, P. J. (1998). Psychovisual aspects of viewing stereoscopic video sequences. volume 3295, pages 226–235.

[Tam and Zhang, 2006] Tam, W. J. and Zhang, L. (2006). 3d-tv content generation: 2d-to-3d conversion. In *IEEE International Conference on Multimedia and Expo*, pages 1869–1872.

[Tatler et al., 2005] Tatler, B., Baddeley, R. J., and Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45:643 – 659.

[Tatler, 2007] Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14).

[Torralba et al., 2006] Torralba, A., Oliva, C., Castelhano, M., and Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*.

[Tresilian et al., 1999] Tresilian, J. R., Mon-Williams, M., and Kelly, B. M. (1999). Increasing confidence in vergence as a cue to distance. In *Proceedings of the Royal Society of London B*, pages 39–44.

[Tyler, 2006] Tyler, C. W., editor (2006). *Binocular Vision. Foundation Volume 2. Chapter 24*. Lippincott Williams and Wilkins, Duane's Ophthalmology on CD-ROM.

[Uetake et al., 2000] Uetake, A., Murata, A., Otsuka, M., and Takasawa, Y. (2000). Evaluation of visual fatigue during vdt tasks. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 2, pages 1277–1282 vol.2.

[Ujike et al., 2008] Ujike, H., Ukai, K., and Nihei, K. (2008). Survey on motion sickness-like symptoms provoked by viewing a video movie during junior high school class. *Displays*, 29(2):81–89.

[Uka et al., 2000] Uka, T., Tanaka, H., Yoshiyama, K., Kato, M., and Fujita, I. (2000). Disparity selectivity of neurons in monkey inferior temporal cortex. *Journal of Neurophysiology*, 84(1):120–132.

[Ukai and Howarth, 2008] Ukai, K. and Howarth, P. A. (2008). Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations. *Displays*, 29(2):106–116.

[Ukai and Kato, 2002] Ukai, K. and Kato, Y. (2002). The use of video refraction to measure the dynamic properties of the near triad in observers of a 3-d display. *Ophthalmic and Physiological Optics*, (0275-5408 (Print)).

[Ukai et al., 2000] Ukai, K., Oyamada, H., and Ishikawa, S. (2000). Changes in accommodation and vergence following 2 hours of movie viewing through bi-ocular head-mounted display. In Franzén, O., Richter, H., and Stark, L., editors, *Accommodation and Vergence Mechanisms in the Visual System*, pages 313–325. Birkhäuser Basel.

[Urey et al., 2011] Urey, H., Chellappan, K. V., Erden, E., and Surman, P. (2011). State of the art in stereoscopic and autostereoscopic displays. *Proceedings of the IEEE*, 99(4):540–555.

[Urvoy et al., 2013] Urvoy, M., Barkowsky, M., Li, J., and Le Callet, P., editors (2013). *Visual Comfort and Fatigue in Stereoscopy*. 3D Video. From Capture to Diffusion. Wiley.

[Velichkovsky, 2002] Velichkovsky, B. M. (2002). Heterarchy of cognition: the depths and the highs of a framework for memory research. *Memory*.

[Vetro, 2010] Vetro, A. (2010). Representation and coding formats for stereo and multiview video. In Chen, C., Li, Z., and Lian, S., editors, *Intelligent Multimedia Communication: Techniques and Applications*, volume 280 of *Studies in Computational Intelligence*, pages 51–73. Springer Berlin Heidelberg.

[Vlad et al., 2013] Vlad, R., Ladret, P., and Guérin, A. (2013). Three factors that influence the overall quality of the stereoscopic 3d content: image quality, comfort, and realism. In *SPIE 8653, Image Quality and System Performance X*, pages 865309–865309.

[VQEG, 2000] VQEG (2000). Final report from the video quality experts group on the validation of objective models of video quality assessment. *www.vqeg,org.*

[VQEG, 2003] VQEG (2003). Final report from the video quality experts group on the validation of objective models of video quality assessment - phase ii. *www.vqeg,org.*

[VQEG, 2008] VQEG (2008). Final report from the video quality experts group on the validation of objective models of multimedia quality assessment. *www.vqeg,org.*

[VQEG, 2009] VQEG (2009). Validation of reduced-reference and no-reference objective models for standard definition television, phase i. *www.vqeg,org.*

[VQEG, 2010] VQEG (2010). Report on the validation of video quality models for high definition video content. *www.vqeg,org.*

[Wang et al., 2013] Wang, J., Perreira Da Silva, Le Callet, P., and Ricordel, V. (2013). A computational model of stereoscopic 3d visual saliency. *IEEE Transactions on Image Processing*, 22(6):2151–2165.

[Wang et al., 2014] Wang, K., Andrén, B., Hussain, M., Brunnström, K., and Osterman, J. (2014). Perception and annoyance of crosstalk in stereoscopic 3d projector systems. volume 9011, pages 901125–901125–6.

[Wang et al., 2004] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

[Watanabe and Ujike, 2013] Watanabe, H. and Ujike, H. (2013). Psychological and physiological effects of stereoscopic movies of real-world scenes containing improper three-dimensional settings. *Health (1949-4998)*, 5(7).

[Watt, 2000] Watt, A. H. (2000). *3D computer graphics*. 3D COMPUTER GRAPHICS. Addison-Wesley.

[Watt and MacKenzie, 2013] Watt, S. J. and MacKenzie, K. J. (2013). *3D Media and the Human Visual System.* In Frédéric Dufaux, Béatrice Pesquet-Popescu et Marco Cagnazzo, editeurs,Emerging Technologies for 3D Video: Creation, Coding, Transmission and Rendering. Wiley.

[Wexler and Ouarti, 2008] Wexler, M. and Ouarti, N. (2008). Depth affects where we look. *Current Biology*, pages 1872–1876.

[Winkler, 2014] Winkler, S. (2014). Efficient measurement of stereoscopic 3d video content issues. In *SPIE 9016, Image Quality and System Performance XI*, volume 9016, pages 90160Q–90160Q–7.

[Wismeijer et al., 2010] Wismeijer, D. A., Erkelens, C. J., van Ee, R., and Wexler, M. (2010). Depth cue combination in spontaneous eye movements. *Journal of Vision.*

[Woods, 2012] Woods, A. J. (2012). Crosstalk in stereoscopic displays: a review. *Journal of Electronic Imaging*, 21(4):040902–040902.

[Woods et al., 1993] Woods, A. J., Docherty, T., and Koch, R. (1993). Image distortions in stereoscopic video systems. volume 1915, pages 36–48, San Jose, CA, USA. SPIE.

[Wyckens et al., 2012] Wyckens, E., Borer, S., and Leszczuk, M. (2012). Moavi (monitoring of audio-visual quality by key indicators) project. In *VQEG*.

[Wöpking, 1995] Wöpking, M. (1995). Viewing comfort with stereoscopic pictures: An experimental study on the subjective effects of disparity magnitude and depth of focus. *Journal of the Society for Information Display*, 3(3):101–103.

[Xing et al., 2010a] Xing, L., Ebrahimi, T., and Perkis, A. (2010a). Subjective evaluation of stereoscopic crosstalk perception. In *SPIE 7744, Visual Communications and Image Processing 2010, 77441V*, volume 7744, pages 77441V–77441V–9.

[Xing et al., 2010b] Xing, L., You, J., Ebrahimi, T., and Perkis, A. (2010b). Estimating quality of experience on stereoscopic images. In *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 1–4.

[Xing et al., 2010c] Xing, L., You, J., Ebrahimi, T., and Perkis, A. (2010c). An objective metric for assessing quality of experience on stereoscopic images. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 373–378.

[Yamanoue, 2006] Yamanoue, H. (2006). The differences between toed-in camera configurations and parallel camera configurations in shooting stereoscopic images. In *IEEE International Conference on Multimedia and Expo*, pages 1701–1704.

[Yamanoue et al., 2000] Yamanoue, H., Okui, M., and Yuyama, I. (2000). A study on the relationship between shooting conditions and cardboard effect of stereoscopic images. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(3):411–416.

[Yang et al., 2009] Yang, J., Hou, C., Zhou, Y., Zhang, Z., and Guo, J. (2009). Objective quality assessment method of stereo images. In *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009*, pages 1–4.

[Yano et al., 2004] Yano, S., Emoto, M., and Mitsuhashi, T. (2004). Two factors in visual fatigue caused by stereoscopic hdtv images. *Displays*, 25(4):141–150.

[Yano et al., 2002] Yano, S., Ide, S., Mitsuhashi, T., and Thwaites, H. (2002). A study of visual fatigue and visual comfort for 3d hdtv/hdtv images. *Displays*, 23(4):191–201.

[Yarbus, 1967] Yarbus, A. (1967). Eye movements and vision. *Plenum*.

[Yasakethu et al., 2008] Yasakethu, S. L. P., Hewage, C., Fernando, W., and Kondoz, A. (2008). Quality analysis for 3d video using 2d video quality models. *IEEE Transactions on Consumer Electronics.*, 54(4):1969–1976.

[Yeh and Silverstein, 1990] Yeh, Y. Y. and Silverstein, L. D. (1990). Limits of fusion and depth judgment in stereoscopic color displays. *Human Factors The Journal of the Human Factors and Ergonomics Society*, (0018-7208 (Print)).

[Yoon and Ho, 2007] Yoon, S.-U. and Ho, Y.-S. (2007). Multiple color and depth video coding using a hierarchical representation. *IEEE Transactions on Circuits and Systems for Video Technology.*, 17(11):1450–1460.

[You et al., 2010] You, J., Xing, L., Perkis, A., and Wang, X. (2010). Perceptual quality assessment for stereoscopic images based on 2d image quality metrics and disparity analysis. In *Proc. of International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, USA*.

[Zeithaml et al., 1993] Zeithaml, V., Berry, L., and Parasuraman, A. (1993). The nature and determinants of customer expectations of service. *Journal of the Academy of Marketing Science*, 21(1):1–12.

[Zhang et al., 2011] Zhang, L., Vazquez, C., and Knorr, S. (2011). 3d-tv content creation: Automatic 2d-to-3d video conversion. *IEEE Transactions on Broadcasting*, 57(2):372–383.

[Zone, 2013] Zone, R. (2013). *3DIY: 3D Moviemaking on an Indy Budget*. Taylor and Francis.

List of Figures

List of Tables

Résumé

Le niveau d'exigence minimum pour tout système 3D (images stéréoscopiques) est de garantir le confort visuel des utilisateurs. Le confort visuel est un des trois axes perceptuels de la qualité d'expérience (QoE) 3D qui peut être directement lié aux paramètres techniques du système 3D. Par conséquent, le but de cette thèse est de caractériser objectivement l'impact de ces paramètres sur la perception humaine afin de contrôler la qualité stéréoscopique.

La première partie de la thèse examine l'intérêt de prendre en compte l'attention visuelle des spectateurs dans la conception d'une mesure objective de qualité 3D. Premièrement, l'attention visuelle en 2D et 3D sont comparées en utilisant des stimuli simples. Les conclusions de cette première expérience sont validées en utilisant des scènes complexes avec des disparités croisées et décroisées. De plus, nous explorons l'impact de l'inconfort visuel causé par des disparités excessives sur l'attention visuelle.

La seconde partie de la thèse est dédiée à la conception d'un modèle objectif de QoE pour des vidéos 3D, basé sur les seuils perceptuels humains et le niveau d'acceptabilité. De plus nous explorons la possibilité d'utiliser la modèle proposé comme une nouvelle échelle subjective. Pour la validation de ce modèle, des expériences subjectives sont conduites présentant aux sujets des images stéréoscopiques fixes et animées avec différents niveaux d'asymétrie. La performance est évaluée en comparant des prédictions objectives avec des notes subjectives pour différents niveaux d'asymétrie qui pourraient provoquer un inconfort visuel.

Abstract

The minimum requirement for any 3D (stereoscopic images) system is to guarantee visual comfort of viewers. Visual comfort is one of the three primary perceptual attributes of 3D QoE, which can be linked directly with technical parameters of a 3D system. Therefore, the goal of this thesis is to characterize objectively the impact of these parameters on human perception for stereoscopic quality monitoring.

The first part of the thesis investigates whether visual attention of the viewers should be considered when designing an objective 3D quality metrics. First, the visual attention in 2D and 3D is compared using simple test patterns. The conclusions of this first experiment are validated using complex stimuli with crossed and uncrossed disparities. In addition, we explore the impact of visual discomfort caused by excessive disparities on visual attention.

The second part of the thesis is dedicated to the design of an objective model of 3D video QoE, which is based on human perceptual thresholds and acceptability level. Additionally we explore the possibility to use the proposed model as a new subjective scale. For the validation of proposed model, subjective experiments with fully controlled still and moving stereoscopic images with different types of view asymmetries are conducted. The performance is evaluated by comparing objective predictions with subjective scores for various levels of view discrepancies which might provoke visual discomfort.