HAL
open science

Subjective quality assessment : a study on the grading scales : illustrations for stereoscopic and 2D video content

Bensaied Rania

To cite this version:

Bensaied Rania. Subjective quality assessment : a study on the grading scales : illustrations for stereoscopic and 2D video content. Signal and Image Processing. Institut National des Télécommunications, 2018. English. NNT : 2018TELE0013 . tel-02042947

HAL Id: tel-02042947

https://theses.hal.science/tel-02042947v1

Submitted on 20 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESE DE DOCTORAT CONJOINT TELECOM SUDPARIS
et L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité : Informatique et Télécommunications

Ecole doctorale : Informatique, Télécommunications et Electronique de Paris

Présentée par

Rania Bensaied Ghaly

Pour obtenir le grade de

DOCTEUR DE TELECOM SUDPARIS

Subjective quality assessment: a study on the grading scales.
Illustrations for stereoscopic and 2D video content

Soutenue le 05 juillet 2018

devant le jury composé de :

Pr. Patrick GALLINARI, Université Pierre et Marie Curie	**Président**
Pr. Patrick LE CALLET, Ecole Polytechnique - Université de Nantes	**Rapporteur**
Pr. Azeddine BEGHDADI, Université Paris 13	**Rapporteur**
Pr. Faouzi GHORBEL, ENSI Tunis	**Examinateur**
Pr. Maryline LAURENT, Télécom SudParis	**Examinateur**
Dr. Didier NICHOLSON, Directeur R&D - VITEC	**Examinateur**
MdC HDR. Mihai MITREA, Télécom SudParis	**Directeur de thèse**

Thèse n° 2018TELE0013

To my daughter and my loving husband.

Acknowledgment

I would like to thank all members of the jury. I thank Professor Patrick LE CALLET and Professor Azeddine BEGHDADI for accepting being my thesis reviewers and for their insightful comments and encouragement. I also thank Professor Patrick GALLINARI, Professor Faouzi GHORBEl, Professor Maryline LAURENT and Dr. Didier NICHOLSON for accepting being my thesis examiners.

I would like to express my deepest thanks and gratitude to my supervisor Mihai MITREA for his continuous support, patience, motivation, and immense knowledge. His precious guidance and fruitful ideas helped me in all the time of research and writing of my scientific papers as well as this manuscript. I deeply hope that we can continue our collaboration.

I am also grateful to my dear Mrs Evelyne TARONI for her encouragement, kindness, enthusiasm and constant support during this thesis; her precious advices have been invaluable.

I also want to give special thanks to all the members of ARTEMIS department of Telecom SudParis; particularly, to Afef Chammem, Ismail Boujelben and Marwa Ammar, with who I had the opportunity to collaborate during the development of this thesis.

I am thankful to Ecole Nationale des Sciences de l'Informatique in Tunis, for the sound education I received there during my engineering program.

Last but not least, I would like to write some words in French.

Je porte en moi la mémoire de mon oncle Kais et mon beau-père Hamda qui se sont éteints sans que je puisse leur dire au revoir. Je ne vous oublierai jamais, que le bon DIEU vous accueille dans son paradis.

Je remercie mes grands parents Ali et Emna et mes parents, Sarah et Khomsi, symbole de la bonté et source de tendresse. Je remercie également mes sœurs Syrine et Dorra et mon frère Dhia. Je tiens aussi à remercier ma belle-mère Fatma, mes belles-sœurs Monia, Mouna, Houda et mes beaux-frères Mounir et Hafedh ainsi que mon cher cousin Fredj, qui m'ont toujours encouragée et soutenue avec grand amour. Merci pour vos sacrifices, vos prières et votre amour inconditionnel. Votre bénédiction m'a permis de devenir ce que je suis aujourd'hui. J'espère avoir réussi à vous rendre fiers de moi.

Je remercie aussi à mes meilleurs amis Souha et Wael. Ils ont été toujours là, dans les meilleurs comme dans les pires.

Enfin, je réserve mes derniers remerciements à Aymen, mon cher mari, mon âme sœur et mon meilleur ami, pour son soutien, son grand amour et sa patience dans les moments difficiles qu'il a dû endurer pendant cette période de thèse.

Table of Contents

ABSTRACT **11**

I. INTRODUCTION **25**

I.1. What does image quality mean? **27**

I.2. ITU subjective image quality assessment **29**
- I.2.1. Viewing conditions 29
- I.2.2. Panel composition 30
- I.2.3. Evaluation methodology and grading scales 31
- I.2.4. The content under evaluation 33
- I.2.5. Scoring and post-processing of the scores 34
- I.2.6. Controversial ITU issues 34

I.3. Thesis structure **36**

II. STATE OF THE ART **39**

II.1. Methodological panorama of subjective visual quality assessment **42**

II.2. Insight on the grading scales for subjective evaluation **49**
- II.2.1. Background studies in subjective quality evaluation 49
- II.2.2. Visual quality assessment 52

II.3. Graphical user interface for ITU visual quality assessment **57**

II.4. Thesis objectives **59**

III. TEST-BED **61**

III.1. Overview **63**

III.2. Viewing conditions **64**

III.3. The evaluation methodology and the grading scales **65**

III.4. Content under evaluation **66**
- III.4.1. Stereoscopic video corpus 66
- III.4.2. 2D video corpus 68

III.5. Panel composition and size **69**

III.6. Scoring and post-processing of the scores **70**
- III.6.1. Scoring application 70
- III.6.2. Post-processing of the scores 74

III.7. Conclusion **75**

IV. BRIDGING CONTINUOUS AND DISCRETE UNLABELED QUALITY SCALE EVALUATION 77

IV.1. Continuous *vs*. discrete unlabeled scale evaluation 79

IV.2. Gaussian assumption 81
IV.2.1. *MOS* and confidence limits computation for *q* grading scales 81
IV.2.2. Illustration of continuous to discrete *MOS* computation 81
IV.2.3. Investigation on the accuracy of the results 92

IV.3. Beyond Gaussian assumption 97
IV.3.1. Gaussian mixture estimation 97
IV.3.2. *MOS* and confidence limits computation for *q* grading scales 98
IV.3.3. Illustration of continuous to discrete *MOS* computation 99
IV.3.4. Investigation on the accuracy of the results 109

IV.4. Conclusion 114

V. QUANTIFYING THE IMPACT OF THE SEMANTIC LABELS 115

V.1. Verifying the existence of the semantic impact of the labels 117

V.2. Methodological framework for semantic impact assessment 118
V.2.1. Evaluation principle 118
V.2.2. Methodology presentation 118
V.2.3. Quantitative results 122

V.3. Conclusion 142

VI. CONCLUSION AND FUTURE WORK 143

VII. APPENDIXES 147

VII.1. Appendix A Discrete unlabeled scale simulation (Gaussian distribution) 148

VII.2. Appendix B Discrete unlabeled scale simulation (Gaussian mixture) 179

REFERENCES 211

LIST OF PUBLICATIONS 219

LIST OF ACRONYMS 221

List of Figures

Figure I-1: Visual acuity and color vision charts: Snellen (left) and Ishihara (right).....25
Figure I-2: Continuous rating scale for DSCQS method.....26
Figure I-3: Five-grade impairment scale for DSIS method.....26
Figure I-4: Categorical scale for Comparison method.....27
Figure I-5: Categorical scale for Single Stimulus method.....27
Figure II-1: Grading scales used in [SIM11]; (a) DSIS method scale, (b) DSCQS method scales.....38
Figure II-2: Grading scales used in [McK78].....44
Figure II-3: Grading scales used in [Alb81].....45
Figure II-4: Grading scales used in [JON86].....47
Figure II-5: Graphical user interface for ITU visual quality assessment.....51
Figure II-6: State-of-the-art synopsis.....53
Figure III-1: Left and right views sampled from the 3DLive corpus.....61
Figure III-2: Scoring synopsis.....64
Figure III-3: Login interface.....65
Figure III-4: Video control interface.....65
Figure III-5: Scoring interfaces.....67
Figure IV-1: Continuous to discrete scale mapping for the subjective evaluation scores.....74
Figure IV-2: Subjective evaluations for high-quality stereoscopic video content.....79
Figure IV-3: Subjective evaluations for low quality stereoscopic video content.....81
Figure IV-4: Subjective evaluations of high quality 2D video content.....83
Figure IV-5: Subjective evaluations for low quality 2D video content 85
Figure IV-6: Precision in the evaluation of high quality stereoscopic video content (Image quality).....88
Figure IV-7: Precision in the evaluation of high quality stereoscopic video (Depth perception).....88
Figure IV-8: Precision in the evaluation of high quality stereoscopic video (Visual comfort).....88
Figure IV-9: Precision in the evaluation of low quality stereoscopic video (Image quality).....89
Figure IV-10: Precision in the evaluation of low quality stereoscopic video (Depth Perception).....89
Figure IV-11: Precision in the evaluation of low quality stereoscopic video (Visual Comfort).....90
Figure IV-12: Precision in the evaluation of high quality 2D video content.....90
Figure IV-13: Precision in the evaluation of low quality 2D video content.....90
Figure IV-14: Subjective evaluations for high-quality stereoscopic video content, non-Gaussian case 96
Figure IV-15: Subjective evaluations for low quality stereoscopic video content, non-Gaussian case 98
Figure IV-16: Subjective evaluations for high quality 2D video content, non-Gaussian case 100
Figure IV-17: Subjective evaluations for low quality 2D video content low quality 102

Figure IV-18: Precision in the evaluation of high quality stereoscopic video, non-Gaussian case (Image quality).....104

Figure IV-19: Precision in the evaluation of high quality stereoscopic video, non-Gaussian case (Depth perception).....104

Figure IV-20: Precision in the evaluation of high quality stereoscopic video, non-Gaussian case (Visual comfort).....105

Figure IV-21: Precision in the evaluation of low quality stereoscopic video, non-Gaussian case (Image quality).....105

Figure IV-22: Precision in the evaluation of low quality stereoscopic video, non-Gaussian case (Depth Perception).....105

Figure IV-23: Precision in the evaluation of low quality stereoscopic video, non-Gaussian case (Visual comfort).....106

Figure IV-24: Precision in the evaluation of high quality 2D video content, non-Gaussian case 106

Figure IV-25: Precision in the evaluation of low quality 2D video content, non-Gaussian case 107

Figure V-1: Principle of the semantic impact evaluation.....112

Figure V-2: Methodological framework.....115

Figure V-3: Number of failed binomial tests between the values taken by the Y and Z r.v. when scoring high quality stereoscopic video by the *validation* panel 117

Figure V-4: Number of failed binomial tests between the values taken by the Y and Z r.v. when scoring low quality stereoscopic video by the *validation* panel 118

Figure V-5: Number of failed binomial tests between the values taken by the Y and Z r.v. when scoring high quality 2D video by the *validation* panel.....119

Figure V-6: Number of failed binomial tests between the values taken by the Y and Z r.v. when scoring low quality 2D video by the *validation* panel 119

Figure V-7: Number of failed binomial tests when assessing high quality stereoscopic video by the *validation* panel 121

Figure V-8: Number of failed binomial tests when assessing low quality stereoscopic video by the *validation* panel 122

Figure V-9: The number of failed binomial tests when assessing high quality 2D video by the *validation* panel 123

Figure V-10: The number of failed binomial tests when assessing low quality 2D video by the *validation* panel 124

Figure V-11: The number of failed binomial tests when assessing high quality stereoscopic video by the *cross-checking* panel.128

Figure V-12: The number of failed binomial tests when assessing low quality stereoscopic video by the *cross-checking* panel ..129

Figure V-13: The number of failed binomial tests when assessing high quality 2D video by the *cross-checking* panel 130

Figure V-14: The number of failed binomial tests when assessing low quality 2D video by the *cross-checking* panel 131

List of Tables

Table A-1: Grading scales on the subjective quality assessment: constraints, challenges, current limitations and contributions. 15

Table I-1 General viewing conditions for subjective assessments....22

Table I-2 Preferred Viewing Distance-PVD for video according to ITU-R BT.500-11. 22

TableII-1: Methodological panorama of subjective visual quality assessment 39

Table II-2: Background studies related to grading scales. 44

Table II-3: Grading scales in visual quality assessment. 47

Table III-1 Viewing conditions for test-bed....56

Table III-1: Panel composition and size: a total of 640 observers, split in 4 sub-panels of 160 observers each....61

Table III-3: Outliers' detection results....66

Table V-1: Quantifying the semantic impact for *q=3*....123

Table V-2: Quantifying the semantic impact for *q*=5....124

Table V-3: The generality of the quantization of the semantic impact for *q=3*....130

Table V-4: The generality of the quantization of the semantic impact for *q=5*....131

Table V-5: Reference values for the semantic impact for *q=3*....132

Table V-6: Reference values for the semantic impact for *q=5*....133

Contexte

Les méthodes subjectives d'évaluation de la qualité du contenu visuel servent à quantifier les performances des systèmes multimédias avec des observateurs humains. Afin de garantir la reproductibilité et la comparabilité des résultats, il est indispensable de disposer de conditions d'évaluation bien configurées en se référant, par exemple aux recommandations de l'Union Internationale des Télécommunications (UIT) (UIT-R BT.1788, R BT.500-11, UIT-R BT.500-13, UIT-T P.913,…).

Les spécifications de l'UIT définissent diverses exigences relatives à l'infrastructure, à la conception du groupe d'observateurs (le panel) et à la manière dont les sessions de test doivent être conduites. Tout d'abord, l'infrastructure et les conditions de test sont définies. Par exemple, l'éclairage environnemental, la distance entre l'écran et le spectateur ou l'angle de vue doivent être correctement configurés avant l'évaluation. Ensuite, la taille du panel est dimensionnée en fonction de la sensibilité et de la fiabilité de la procédure de test. Par exemple, il est généralement recommandé d'avoir au moins 15 observateurs, tandis que leur acuité visuelle et leur vision couleur sont testées conformément à des procédures préétablies, comme les diagrammes de Snellen ou Ishihara. Les méthodes d'évaluation et les échelles de notation sont également spécifiées. Les méthodes le plus souvent considérées sont la DSCQS (*Double-Stimulus Continuous Quality-scale*), DSIS (*Double-Stimulus Impairment Scale*) ou bien encore SSCQE (*Single-Stimulus Continuous Quality Evaluation*). Enfin, une fois l'évaluation de l'échelle choisie, les notes attribuées par les observateurs sont enregistrées et ensuite interprétées statistiquement. Par exemple, lorsqu'ils considèrent l'échelle SSCQE à 5 niveaux, les observateurs notent en sélectionnant l'une des étiquettes suivantes: *Excellent*, *Good*, *Fair*, *Poor*, *Bad* (dont la traduction française est Excellent, Bon, Satisfaisant, Médiocre, et Mauvais). Par la suite, ces étiquettes sont converties en valeurs entières (Excellent est mappé à 5 tandis que Bad est mappé à 1); les valeurs aberrantes sont éliminées et le MOS (Mean Opinion Scores, c'est-à-dire la moyenne de ces valeurs) est calculé en même temps que ses limites de confiance.

Notre thèse porte sur divers aspects liés à la spécification des échelles de gradation, de leur fondement théorique à leur utilisation pour les vidéos 2D et stéréoscopiques.

Limitations de l'état de l'art

Depuis 100 ans déjà, divers domaines de recherche (psychologie, psychophysique, sociologie, marketing, médecine, ...) ont envisagé l'utilisation d'échelles d'évaluation dans les évaluations subjectives. Malgré cette longue et fructueuse histoire, aucun consensus n'est encore atteint sur l'utilisation d'une échelle spécifique dans un but spécifique, et plusieurs types d'échelles coexistent encore tout se contredisant: échelles étiquetées graphiques / numériques / sémantiques ou continues / discrètes.… De plus, la dynamique des échelles numériques varie avec l'expérience: par exemple, les échelles continues peuvent aller de 0 à 75, à 100, à 120 ou même à 200, tandis que Les échelles discrètes peuvent comporter entre 2 et 11 classes d'évaluation.

Pour l'évaluation du contenu visuel, les Recommandations de l'UIT ont prouvé leur efficacité et sont déjà utilisées de manière intensive dans plusieurs études de recherche visant une grande variété

d'applications (évaluation / étalonnage d'appareils, compression, reconstruction d'images en 3D, tatouage, etc.). Certaines études utilisent des échelles à 5 niveaux de qualité tandis que d'autres sur 11 niveaux. Pourtant, aucune réponse sur la manière de choisir ce nombre ou les niveaux de qualité eux-mêmes n'est fournie. De plus, la Rec. UIT-T P.913 va au-delà de ces recommandations et fournit une liste de modifications acceptables (telles que l'utilisation d'échelles non étiquetées, l'utilisation d'échelles numérotées non étiquetées, etc.)

La relation entre les échelles continue et discrète est abordée de manière récurrente dans les études de recherche. Par exemple, il a été montré que les évaluations de l'échelle discrète présentent le niveau de stabilité le plus élevé, du moins pour l'expérience considérée (une auto-évaluation des connaissances antérieures en statistique). Cette étude soulève également une interrogation quant à la signification même des termes continus et discrets lors des évaluations subjectives.

L'impact des étiquettes sémantiques est discuté et détaillé dans diverses études de recherche. Selon certaines études, les étiquettes UIT adjacentes sont caractérisées par des distances sémantiques non uniformes ; pourtant, un tel comportement n'est pas quantifié. D'autre part, certaines études affirment le contraire, c'est-à-dire que la sémantique des étiquettes UIT adjacentes n'a pas d'impact sur les résultats. Si certains résultats correspondent à des études subjectives effectuées pour différentes langues (japonais, allemand, anglais, français et italien), la Rec. UIT-T P.913 postule explicitement que le MOS est invariant vis-à-vis de la traduction des étiquettes sémantiques, mais ne fournit aucun motif pour cela.

Au-delà des particularités de la stratégie d'évaluation elle-même, des doutes sur le modèle statistique des résultats d'évolution (les scores) surgissent: l'hypothèse implicite de la distribution gaussienne est aussi contredite dans l'interprétation de certains résultats dans UIT-R BT.500 11/13.

Enfin, bien que l'UIT demande toujours que le nombre de sujets impliqués dans les évaluations soit supérieur à 15 (par exemple, UIT-R BT.500-13), les études expérimentales mettent en évidence une grande variabilité de ce paramètre ; dans ce contexte, exprimer de manière pragmatique l'influence théorique de la taille du panel d'observateurs quant à l'exactitude des résultats reste une question ouverte.

Objectifs

La présente thèse a pour objectif principal d'analyser les recommandations de l'UIT et d'étudier théoriquement certains de leurs aspects essentiels liés aux échelles d'évaluation. Trois axes principaux de recherche sont pris en compte.

D'abord, l'étude sera menée au niveau théorique. À cet égard, la thèse tentera de relier les procédures d'évaluation des échelles continue et discrète et de déterminer si le nombre de classes sur les échelles discrètes est un critère relevant dans les interprétations des résultats ou un simple paramètre. L'étude de l'influence du modèle statistique des scores et de la taille du panel (nombre d'observateurs) dans l'exactitude des résultats entre également dans le cadre de la thèse.

Deuxièmement, au niveau méthodologique, la thèse abordera la question de la quantification du biais induit dans les expériences de qualité vidéo subjective par les étiquettes sémantiques (par exemple, *Excellent*, *Good*, *Fair*, *Poor* et *Bad*) généralement associées aux échelles d'évaluation discrètes.

Enfin, d'un point de vue expérimental, les deux directions mentionnées nécessitent des conditions expérimentales capables d'appuyer leur précision et leur pertinence statistique. En conséquence, spécifier et déployer un tel cadre expérimental devient un objectif en soi. De cette manière, une étude sur la variabilité pratique des résultats expérimentaux et méthodologiques avec le type de contenu (vidéo 2D ou stéréoscopique) ou avec sa qualité (évaluée par des métriques objectives) devient également possible.

Contributions théoriques

Il est théoriquement démontré que les scores attribués par les observateurs sur une échelle de notation continue peuvent être mappés a *posteriori* sur toute échelle de notation discrète, avec une précision statistique contrôlée. À cet égard, en utilisant des transformations de variables aléatoires non-linéaires, la formule reliant les fonctions de densité de probabilité modélisant les échelles de gradation continue et discrète est établie. Les moments du premier et du deuxième ordre (permettant de calculer le MOS et les limites de confiance) sont ensuite déduits.

Ces résultats sont génériques et ne tiennent pas compte de la particularité de la fonction de densité de probabilité d'origine ni du partitionnement sur une échelle discrète (les partitions égales et non-uniformes sont couvertes). Cependant, dans la perspective de transformer ces formules théoriques en outils efficaces d'évaluation de la qualité visuelle, il est nécessaire de disposer d'une description statistique de la densité de probabilité modélisant les scores attribués par l'observateur sur l'échelle continue. Par conséquent, l'étude sous-jacente est également menée. Elle commence par examiner l'hypothèse gaussienne qui est rejetée (sur la base du test de concordance statistique de Chi-2) et propose ensuite des estimations basées sur des mélanges gaussiens. L'impact du nombre d'observateurs humains sur la précision de l'évaluation de la qualité est ensuite évalué

Contributions méthodologiques

Afin de quantifier l'impact psycho-cognitif des étiquettes sémantiques dans l'évaluation de la qualité, la thèse avance une procédure d'investigation interdisciplinaire ; cette procédure combine les formules théoriques mentionnées, l'estimation statistique et les tests statistiques.

D'abord, afin de déterminer si un tel impact sémantique existe, une comparaison basée sur le test de Student est effectuée entre les valeurs moyennes (représentant le MOS) correspondant aux échelles continue (non étiquetée) et discrète, sémantiquement étiquetée. Comme les expériences démontrent que l'impact sémantique existe, une procédure pour sa quantification est définie par la suite. La deuxième étape consiste donc à définir une variable aléatoire discrète auxiliaire, caractérisée par une partition inégale mais par des probabilités a *posteriori* égales. L'impact sémantique est quantifié en définissant un coefficient. Cette variable aléatoire auxiliaire est estimée par des tests binomiaux entre les différences de longueur des classes de partition entre cette variable aléatoire auxiliaire et la variable aléatoire correspondant à l'échelle étiquetée sémantiquement,

Étude expérimentale

L'étude expérimentale est réalisée conformément aux recommandations UIT-R BT.1788, UIT-R BT.500-11, UIT-R BT.500-13 et UIT-T P.913 et prend en compte la méthode SSCQE. Il convient de souligner que la thèse examine la procédure d'évaluation elle-même et n'est pas centrée sur le contenu à évaluer. Par conséquent, le cadre expérimental conçu et déployé dans la thèse devrait adapter / étendre les spécifications UIT de manière à s'adapter à cet objectif.

Une application Android a été développée afin de permettre alternativement l'évaluation sur des échelles continues et discrètes sémantiquement étiquetées. La première échelle se situe entre 0 et 100, avec 10 points équidistants (les scores sont enregistrés avec une précision unitaire). Le deuxième type d'échelle analyse alternativement les évaluations sur 5 et 3 niveaux, avec les libellés suivants: *Excellent, Good, Fair, Poor, Bad* et respectivement *Good, Fair, Bad.*

Le contenu à évaluer est représenté par quatre corpora composés par des vidéos de haute et basse qualité (selon des mesures objective *a priori* de qualité), vidéo 2D et stéréoscopique. Chacun de ces corpus dure environ 20 minutes. Notez que lors de l'étude de la procédure d'évaluation, le contenu lui-même ne fait que garantir des conditions réalistes de notation.

Au total, 640 observateurs humains participent aux expériences. Ils sont regroupés en quatre panels principaux de 160 observateurs chacun, un panel pour chaque type de contenu. Afin de garantir la répétabilité des résultats, chaque panel principal a été divisé en trois sous-panels, appelés panels de référence (60 observateurs), de validation (50 observateurs) et de contrôle croisé (50 observateurs). Les sujets des panels de référence ont attribué des scores sur les échelles continues. Les sujets des panels de validation et de vérification croisée sont finalement chacun divisés en deux sous-ensembles et ont attribué des notes sur des échelles de notation à 5 et 3 niveaux.,.

La relevance statistique des résultats est assurée en calculant des limites de confiance de 95% pour le MOS et en appliquant les différents tests statistiques (Chi-2, Student et binomial) à un niveau de signifiance statistique de 0,05.

Tableau 1 : Les échelles de classement de l'évaluation subjective de la qualité: contraintes, défis, limitations et contributions.

Contraintes	Défis	Limitations	Contributions de la thèse
Etude théorique pour les échelles de notation	• Échelles de notation continues et discrètes • Nombre variable de niveaux de qualité sur l'échelle de notation • Comportement statistique des scores	• Utilisation de différentes échelles de notation avec différents niveaux sans aucune étude sur leur impact sur les résultats • L'hypothèse implicite que les scores ont des distributions gaussiennes	• Etablir une formule théorique pour mapper les notes attribuées sur une échelle de notation continue à toute échelle de notation discrète (partitionnée de manière uniforme ou non), avec une précision statistique contrôlée: ○ Formule de filtrage non linéaire de variable aléatoire ○ Calcul des paramètres (MOS et écart type) • Estimation de la fonction de densité de probabilité pour les scores des observateurs via un mélange gaussien ○ Application de l'algorithme de Expectation Maximization EM) ○ Nombre optimal des paramètres du mélange gaussien • Evaluer l'impact du nombre d'observateurs humains sur la précision de l'évaluation de la qualité
Cadre méthodologique pour l'évaluation de l'impact des étiquettes sémantiques	• Quantification de l'influence psycho-cognitive	• Au cours de la procédure d'évaluation, on suppose que: ○ aucune contrainte psycho-cognitive n'est imposée aux observateurs par les étiquettes sémantiques ○ les scores sont attribués uniquement en fonction du classement de l'étiquette sur les échelles	• Définir un cadre méthodologique pour l'identification et l'évaluation de l'impact des étiquettes sémantiques: ○ Comparaison basée sur les tests appariés de Student pour déterminer si un tel impact existe ○ Définition d'une variable aléatoire auxiliaire: partition inégale mais distances de probabilités *a posteriori* égales ○ Définition et évaluation d'un coefficient d'impact sémantique par des tests binomiaux répétés
Cadre expérimental	• Répétabilité et pertinence statistique pour les résultats • Possibilité d'analyser la dépendance par rapport au contenu 2D et stéréoscopique • Vidéo haute et basse qualité • Logiciel de notation	• Configurations expérimentales orientées vers l'évaluation du contenu plutôt que vers l'évaluation de la procédure • Manque d'approches consensuelles • Pas de confrontation entre vidéo 2D et stéréoscopique • Aucune comparaison entre les échelles de notation à 5 et 3 niveaux	• 640 observateurs humains (160 observateurs pour chaque type de contenu), répartis en 3 types de sous-panels ○ 4 x 60 pour le panneau de référence ○ 4 x 50 pour le panneau de validation ○ 4 x 50 pour le contrôle de contrôle croisé • corpus vidéo 4 x 20 minutes: ○ contenu 2D et stéréoscopique ○ vidéo haute et basse qualité (35dB <PSNR <40dB et 25dB <PSNR <30dB) • Limites de confiance à 95% pour le MOS et niveaux de signification de 0,05 pour chacun des tests considérés (Chi -2, test de Student, binomial) • Discussion sur le type de contenu ○ vidéo 2D vs vidéo stéréoscopique ○ vidéo de basse qualité vs haute qualité • Logiciel de notation Android: ○ Echelle de notation continue ○ Echelle de notation discrète étiquetée sémantiquement (utilisée alternativement à 5 et 3 niveaux)

Abstract

Context

The subjective visual quality assessment methods are essentially used to gauge the performance of multimedia systems with the help of responses obtained from observers who investigate the content displayed by the system under test. Hence, in order to ensure the reproducibility and comparability of the results, well-configured, consensual evaluation conditions are particularly required and the International Telecommunication Union (ITU) Recommendations save as a ground in this respect (ITU-R BT.1788, ITU-R BT.500-11, ITU-R BT.500-13, ITU-T P.913, …).

The ITU specifications define various requirements related to the infrastructure, the design of the panel of observers and the way in which the test sessions should be conducted. First, the test infrastructure and the environment conditions are defined. For instance, the environmental illumination, the distance between the screen and the viewer or the angle of view should be properly set-up prior to the evaluation. Secondly, the panel size is designed depending on the sensitivity and reliability of the test procedure and on the anticipated effect sought. For instance, the number of observers is usually recommended to be at least 15 while their visual acuity and color vision are tested according to pre-established procedures, like the popular Snellen or Ishihara charts. Thirdly, the evaluation methods and the underlying grading scales for subjective quality assessment are specified. Among these, the most popular are the Double-Stimulus Continuous Quality-Scale (DSCQS), the Double-Stimulus Impairment Scale (DSIS) and the Single Stimulus Continuous Quality Evaluation (SSCQE). Finally, once the scale evaluation is chosen, the scores assigned by the observers are recorded and subsequently statistically interpreted. For instance, when considering the 5-levels SSCQE scale, the observers score by selecting one of the following labels: *Excellent*, *Good*, *Fair*, *Poor* and *Bad*. Subsequently, these labels are converted into integer values (*Excellent* is mapped to 5 while *Bad* is mapped to 1); the outliers are eliminated and the *MOS* (mean opinion score, *i.e.* the average of these values) is computed alongside with its confidence limits.

The present thesis focuses on various aspects related to the grading scales specification, from their theoretical basis to their usage for conventional (2D) and stereoscopic video.

State-of-the-art limitations and constraints

For 100 years already, various research fields (psychology, psychophysics, sociology, marketing, medicine, ...) have considered the use of rating scales in subjective evaluations [FRE23], [FRO89]. Despite this long and fruitful history, no consensus is reached yet on a usage of a specific scale for a specific purpose, and several scale typologies still coexist and contradict each-other: graphic *vs*. numerical *vs*. semantic labeled scales or continuous *vs*. discrete scales or … Moreover, the dynamics of numerical scales is varying with the experiment: for instance, continuous scales can range from 0 to 75, to 100, to 120 or even to 200 [AIT69], [BON74], [McG84] while the discrete scales can feature between 2 and 11 evaluation classes [FRE23].

For the visual content evaluation, the ITU Recommendations have proven their effectiveness and are already intensively used in several research studies aiming at a large variety of applications (device evaluation/calibration, compression, 3D image reconstruction, watermarking, *etc*.). Some studies report experiments had carried out on 5 quality levels while other on 11 quality levels. Yet, no answer on how to choose either this number or the quality levels themselves is provided. Moreover, the ITU-T P.913 goes beyond such particular evaluation set-ups and provides a list of some acceptable changes (like the use of unlabeled scales, the use of numbered unlabeled scales, *etc*.) and even of some acceptable yet discouraged changes (like increasing the number of qualitative levels). Here again, all these modifications are left to the experimenter choice and no particular guide is provided.

The relationship between continuous and discrete scales is recurrently addressed in research studies. For instance, [SVE00] shows that assessments on the discrete scale have the highest level of stability, at least for the experiment under consideration (a self-assessment of the previous knowledge in statistics). This study also raises a concern about the very meaning of the *continuous* and *discrete* terms during the subjective evaluations.

The impact of semantic labels is discussed and detailed in various research studies. On the one hand, some studies state that adjacent ITU labels are characterized by non-uniform semantic distances [JON86], [NAR93]; yet, such a behavior is not quantified. On the other hand, some studies [ZIE07] claim the contrary, *i.e.* that the semantic of adjacent ITU labels does not impact the results. While some results correspond to subjective studies carried out for different languages (Japanese, German, English, French and Italian), the ITU-T P.913 explicitly postulates that the *MOS* is invariant with respect to the semantic labels translation, but does not provide any ground for this.

Beyond the peculiarities in the evaluation strategy itself, doubts about the statistical model of the evolution results (*i.e.* the scores) arise: the implicit assumption of Gaussian distribution [SIM09], [SES10], [WIN09] is also implicitly contradicted in some results interpretation in ITU-R BT.500 11/13.

Finally, although the number of subjects involved in evaluations is always requested by ITU to be larger than 15 (e.g. ITU-R BT.500-13), the experimental studies bring to light a large variability of this parameter and pragmatically expressing the theoretical influence of the size of the observers panel in the accuracy of the results remains an open question.

Objectives

The present thesis has as main objective to reconsider the ITU recommendations and to investigate on theoretical basis some of their key aspects related to evaluation scales. Three main research directions are to be considered.

First, the investigation will be carried out at the theoretical level. In this respect, the thesis will try to bridge the continuous and the discrete scale evaluation procedures and to investigate whether the number of the classes on the discrete scales is a criterion meaningful in the results interpretations or just a parameter. Studying the influence of the statistical model of the scores and the size of the panel (number of observers) in the accuracy of the results is also under the scope of the thesis.

Secondly, at the methodological level, the thesis will address the issue of quantifying the bias induced in subjective video quality experiments by the semantic labels (*e.g. Excellent*, *Good*, *Fair*, *Poor* and *Bad*) generally associated to the discrete grading scales.

Finally, from the experimental point of view, these above-mentioned two directions require an experimental test-bed able to support their precision and statistical relevance. Hence, specifying and deploying such an experimental test-bed becomes an objective *per-se*. This way, an investigation on the practical variability of the experimental and methodological results with the type of content (2D or stereoscopic video) or with its quality (as assessed by objective metrics) becomes also possible.

Theoretical contribution

First, it is theoretically demonstrated that the scores assigned by the observers on a continuous grading scale can be a posteriori mapped to any discrete grading scale, with controlled statistical accuracy. In this respect, by using non-linear random variable transformations, the formula connecting the probability density functions modeling the continuous and discrete grading scales is established. The first and second order moments (allowing for the *MOS* and the confidence limits to be computed) are subsequently derived.

These results are generic and do not take into account either the original probability density function peculiarity or the discrete scale partitioning (both even and uneven partitions are covered). However, in the perspective of turning these theoretical formulae into effective quality evaluation tools, a statistical description for the continuous scale probability density function modeling the observer's scores is required. Hence, the underlying study is also conducted. It starts by investigating the Gaussian hypothesis that is rejected (based on the Chi-square goodness-of-fit test) and follows by suggesting approximations based on Gaussian mixtures. The impact of the number of human observers in the precision of the quality evaluation is subsequently assessed.

Methodological contribution

A cross-disciplinary investigation procedure is defined (combining the above-mentioned theoretical formulae, statistical estimation and statistical tests) to quantify the psycho-cognitive impact of the semantic labels in the quality evaluation.

First, in order to bring to light whether such a semantic impact exist, a comparison (based in the Student's paired test) between the average values (representing the *MOS*) corresponding to the continuous (unlabeled) and discrete, semantically labeled scales is carried out. As the experiments demonstrate that the semantic impact exists, a procedure for its quantification it is also defined. Hence, the second step is to define an auxiliary discrete random variable, which is characterized by uneven partition but by equal *a posteriori* probabilities. By comparing the differences in the partition classes length between this auxiliary random variable and the random variable corresponding to the semantically labeled scale, the semantic impact is quantified (by defining an underlying coefficient). This auxiliary random variable is estimated trough repeated binomial tests.

Experimental study

The experimental study is carried out according to the ITU-R BT.1788, ITU-R BT.500-11, ITU-R BT.500-13, ITU-T P.913 recommendations and considers the SSCQE method. It should be emphasize that the thesis investigates the evaluation procedure itself and is not focused on the content to be evaluated. Hence, the test-bed designed and deployed in the thesis should consider the ITU specifications as a backbone and should adapt/extend them so as to fit to the evaluation procedure investigation.

In order to score the content, a versatile Android application has been developed so as to alternatively allow the evaluation on continuous and discrete, semantically labeled scales. The former ranges between 0 and 100, with 10 even marks (the scores are recorded with unit precision). The later alternatively considers 5 levels and 3 levels evaluations, with the following labels: *Excellent*, *Good*, *Fair*, *Poor*, *Bad* and *Good*, *Fair* and *Poor*, respectively.

The content to be evaluated is represented by four corpora, representing high and low-quality video (*e.g.* 35dB < PSNR < 40dB and 25dB<PSNR<30dB), both 2D and stereoscopic video. Each of these corpora is about 20 minutes long and is presented to the observers into downgraded versions obtained through watermarking or compression methods (16 downgraded versions for the stereoscopic contents, 16 downgraded versions for high quality 2D content and 28 downgraded versions for the low-quality 2D content). Note that as the evaluation procedure is investigated, the content itself just ensures realistic conditions for scoring.

A total of 640 human observers are involved in the experiments. They are grouped in four main panels of 160 observers each, on panel for each type of content. In order to grant result repeatability, each main panel was split into three sub-panels, referred to as the *reference* (60 observers), *validation* (50 observers) and *cross-checking* (50 observers) panels. The subjects in the reference panels scores on the continuous scales. The subjects in both the *validation* and *cross-checking* panels are finally partitioned into two sets, scoring on 5 level grading scales and on 3 level grading scales, respectively.

The statistical relevance of the results is ensured by computing 95% confidence limits for *MOS* and by applying the various the statistical tests (Chi-square, Student's and binomial) at 0.05 significance level.

Table A-1: The grading scales on the subjective quality assessment: constraints, challenges, current limitations and contributions.

Constraints	Challenges	Current limitations	Thesis contributions
Theoretical basis for grading scales	• Continuous vs. discrete grading scales • Variable number of quality levels on the grading scale • Observers scores statistical behavior	• Use of different grading scale with different quality levels, without any investigation about their impact in the meaning of results • Observers' scores are implicitly assumed to be Gaussian distributed	• Establishing theoretical formula for mapping the scores assigned on a continuous grading scale to any discrete grading scale (be it evenly partitioned or not), with controlled statistical accuracy: ○ Non-linear random variable filtering formula ○ Parameter computation (*MOS* and standard deviation) • Estimating the probability density function for the observers scores through a Gaussian mixture ○ Expectation maximization (EM) algorithm application ○ Optimal number of components in the Gaussian mixture investigation • Assessing the impact of the number of human observers in the precision of the quality evaluation
Methodological framework for assessing the semantic label impact	• Psycho-cognitive influence quantification	• During the evaluation procedure, it is assumed that: ○ no psycho-cognitive constraint is imposed to the observers by the semantic labels ○ the scores are assigned solely based on the label rank on scales	• Specifying a methodological framework for identifying and assessing the impact of semantic labels: ○ Student's paired test based comparison for identifying whether the semantic impact does exist ○ auxiliary discrete random variable definition: un-even partition but *a posteriori* equal measurements distances ○ auxiliary random variable estimation ○ definition and evaluation of a semantic impact coefficient trough repeated binomial tests ○
Experimental test-bed	• Repeatability and statistical relevance for the results • Possibility of investigating the dependency on the content • 2D and stereoscopic content • High and low-quality video • Soring tool	• Experimental set-ups oriented towards content evaluation rather then procedure evaluation • Lack of consensual approaches • No confrontation between 2D and stereoscopic video • No comparison between 5 and 3 levels grading scales	• 640 human observers (160 observers for each type of content), partitioned in 3 types of sub-panels ○ 4 x 60 for the *reference* panel ○ 4 x 50 for *validation* panel ○ 4 x 50 for *cross-checking* panel • 4 x 20 minutes video corpora: ○ 2D and stereoscopic content ○ high and low-quality video (35dB < PSNR < 40dB and 25dB<PSNR<30dB) • 95% confidence limits for *MOS* and 0.05 significance levels for each and every considered test (Chi-square, Student's paired, Binomial) • Discussion on the content type ○ 2D video vs. stereoscopic content ○ low *vs*. high quality content • Android scoring application: ○ continuous grading scale evaluation ○ discrete, semantically labeled, evaluation (alternatively used with 5 and 3 levels)

I. Introduction

The present thesis is developed under the framework of subjective visual quality evaluation, a research field for which the International Telecommunication Union (ITU) Recommendations are expected to offer a ground for ensuring the reproducibility and comparability of the results.

Hence, this Introduction chapter succinctly presents the main aspects related to an ITU test session, from the viewing conditions and the panel composition to the post-processing of the scores, passing from the evaluation methodology and the grading scales.

At least three open and surprising issues are thus identified; they relate to: (1) the continuous vs. discrete evaluation scales, (2) the statistical distribution of the scores assigned by the observers and (3) the usage of semantic labels on the grading scales.

I.1. What does image quality mean?

Quality is an ever evolving, yet ever fascinating concept whose meaning goes far beyond the objective of the present thesis.

According to the Oxford English dictionary [WEB01], the first meaning of quality is "*The standard of something as measured against other things of a similar kind; the degree of excellence of something.*" This linguistic definition is very broad, and encompasses various aspects of our daily, professional and private lives. Hence, some 40 years ago, a more pragmatic approach steamed from business and stated that *"quality is conformance to requirements, not goodness"* [CRO82].

It can be considered that a milestone in the societal perception of the quality has been reached some 30 years ago when the International Organization for Standardization (ISO) elaborated a family of standards specifically addressing the quality issues [WEB02]: *"ISO 9001:2015 sets out the criteria for a quality management system*". This way, quality becomes an abstract concept, rather related to the creation/production process than to the results of such process.

One surprisingly aspect is common to the above-mentioned definitions: the quality is always implicitly assumed to be an *a posteriori* measurable notion on which we always have some *a priori* expectations/knowledge. Yet, the way of measuring the quality and expressing with rigor these expectations (where are they come from? how relevant are them? are they individual or shared by various persons?) remain controversial issues.

The earliest mentioning of image quality is credited to date back to the invention of optical instruments, at the beginning of the XVIIst century [ENG99]. The term became more and more common with the introduction of photography and television, and several definitions coexist today [KUN16], [MAR11], [AVI01], [PRE15], [CHA13].

Intuitively, image quality is defined as the difference between a processed image and its original representation, from both visual and cognitive points of views. According to Janssen [JAN99], in order to bring together these approaches, the image quality should be defined in the context of visuo-cognitive systems, and should relates to the *"observer's interaction with his environment"*; the notions of *usefulness* and *naturalness* are subsequently defined as being the image capacity of being *close* to both the original representation and to knowledge about the reality of the observer.

The simplest taxonomy in image/video quality evaluation [JAH97] relates to two pragmatic situations.

First, according to *what* it is to be evaluated, *no-reference* and *reference* quality assessments are encountered. A *no-reference* image quality assessment assumes that solely the content under evaluation is available and that its quality is to be evaluated with respect to some pre-established, commonly (implicitly or explicitly) agreed criteria, derived from some *a priori* knowledge/expectations. On the contrarily, in *reference* quality assessment, the content under evaluation is compared to other content whose quality is supposed to serve as reference; actually, such an approach rather evaluates the difference in quality then the quality itself.

Secondly, according to *how* the evaluation takes place, *objective* and *subjective* evaluation procedures are deployed. An *objective* image quality evaluation solely relays on visual content (the one to be

evaluated and, eventually, some reference content) but does not directly and explicitly considers any human observer. This class covers a large variety of image quality metrics, from the popular PSNR (peak signal to noise ratio) to the sophisticated biological-inspired measures, [BEG13]. On the contrarily, *subjective* evaluation procedures consist in inquiring a panel of human observers about the way they perceive the quality of the content under evaluation (be it presented by itself or in conjunction to some reference content).

Of course, such taxonomy is very broad and, according to the targeted application, one several types of evaluation quality procedures can be deployed. For instance, when considering an image acquisition system (a camera), the image quality intuitively relates to the difference between the digital, captured, image and the pristine (natural) representation of the scene. As the natural representation is not available during the evaluation, a no-reference (subjective or objective) evaluation procedure is likely to be set. However, should we be interested in comparing the performances of two cameras, we may considered the images captured by one as reference and to evaluate the second camera images with respect to this reference. When considering now a compression or watermarking application, the image quality relates to the visual differences between the original and the processed image. In such a case, both original (reference) and processed images can be made available for evaluation and the difference between them can be evaluated either by some objective or by subjective evaluations. Of course, this is not compulsory: watermarking applications can also consider no-reference quality evaluation [COX02].

The present thesis focuses on the subjective, no-reference evaluation of visual content. In order to ensure the reproducibility and comparability of the results, well-configured, consensual evaluation conditions are particularly required and the International Telecommunication Union (ITU) Recommendations offers a ground in this respect (ITU-R BT.1788, ITU-R BT.500-11, ITU-R BT.500-13, ITU-T P.913, …).

I.2. ITU subjective image quality assessment

The ITU Recommendations provide methodologies for assessing the overall image quality and the overall image impairment of distorted still images and video. The main aspects related to any test session are thus specified:

- *the viewing conditions*: the luminance and chrominance conditions for the content under evaluation and for the ambient lighting, as well as the relative position of the observer with respect to the display;
- *the panel composition and size*: how many observers in the panel, their typology, the way of testing their visual acuity;
- *the evaluation methodology and the grading scales*: the evaluation/reference content availability, the way to presenting them to the observer, the way the scores are assigned, ….
- *the content under evaluation:* the type of content, its duration, *etc*.
- *the scoring and the post-processing of the scores*: score gathering, outlier detection, *MOS* (mean opinion score) computation, …

In the following sections, these aspects will be detailed according to ITU-R BT.1788, ITU-R BT.500-11, ITU-R BT.500-13. These three references are considered as being the widest used and, at the same time, representative for the two types of the targeted content namely the 2D video and the stereoscopic video. Note that ITU also considers possible modifications of the testing session set-up, as described in ITU-T P.913.

I.2.1. Viewing conditions

The ITU provides detailed specifications for the luminance and the chrominance conditions during the tests sessions. Note that such conditions apply both to the content displaying and to ambient lighting. Additionally, the *viewing conditions* also refer to the position of the observers: the angle of view with respect to the center of the display and the recommended viewing distance between the observer and the display. The latter is computed as the product between the *Preferred Viewing Distance* – PVD (a standard ratio value for a fixed picture high) and the picture height. Tables I-1 and I-2 present these values according to the ITU-R BT.500-11.

Table I-1: General viewing conditions for subjective assessments

Rec. ITU-R BT.500-11
Ratio of luminance of inactive screen to peak luminance $\leq$ 0.02
Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white: $\approx$ 0.01
Display brightness and contrast: set up via PLUGE software
The viewing distance and the screen sizes are to be selected in order to satisfy the Preferred Viewing Distance PVD, see Table I-2.
Maximum observation angle relative to the normal (this number applies to CRT displays, whereas the appropriate numbers for other displays are under study): 30
Ratio of luminance of background behind picture monitor to peak luminance of picture: $\approx$ 0.15
Chromaticity of background: D65
Other room illumination: low

Table I-2: Preferred Viewing Distance-PVD for video according to ITU-R BT.500-11.

Screen diagonal (inch)		Screen height (m)	PVD
4/3 ratio	16/9 ratio		
12	15	0.18	9
15	8	0.23	8
20	24	0.30	7
29	36	0.45	6
60	73	0.91	5
> 100	> 120	> 1.53	3-4

I.2.2. Panel composition

In order to set-up a proper panel, ITU recommends at least 15 observers to compose the panel. They should be non-expert, in the sense that they are neither directly concerned with image quality as part of their normal professional activity, nor experienced assessors.

Prior to the session, the observers should be screened for normal (or corrected to normal) visual acuity, color vision and dynamic stereopsis. The visual acuity can be screen by the Snellen or Landolt charts while color vision can be tested by the Ishihara chart, see Figure I.1. The dynamic stereopsis can be tested by 8 main vision tests (referred to as VT-01 to VT-08): the tests VT-04 and VT-07 are compulsory while the remaining six tests are for more detailed characterization.

No specific mention about the gender involvement is made.

Figure I-1: Visual acuity and color vision charts: Snellen (left) and Ishihara (right).

I.2.3. Evaluation methodology and grading scales

The evaluation methodologies described in ITU-R BT.1788 and ITU-R BT.500-11/13 can be structured into three categories, referred to by ITU as *double stimulus*, *comparison* and *single stimulus* methods, as detailed here after. Note that from the conceptual point of view, the *double stimulus* and *comparison methods* relate to the reference quality evaluation methods while the *single stimulus* method to no-reference quality evaluation methods.

I.2.3.1. Double stimulus methods

In its broadest usage, this type of methods is meant to provide a measure of the quality of a processed content assuming its original (pristine) content is also available and investigated during the assessment session.

Two main types of methods belong to this class, namely the *Double Stimulus Continuous Quality Scale* (DSCQS) and the *Double Stimulus Impairment Scale Method* (DSIS).

According to a *DSCQS* method, the observers assess the overall content quality of both reference and processed content. The results thus obtained can result in different scores between reference and test content, thus indicating a lost in quality. According to ITU-R BT.500-11/13, this evaluation takes place on scales that *"provide a continuous rating system to avoid quantizing errors, but they are divided into five equal lengths which correspond to the normal ITU-R five-point quality scale"*, as illustrated in Figure I-2.

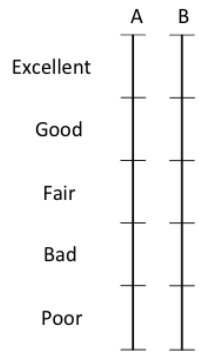


Figure I-2: Continuous rating scale for DSCQS method (source: ITU-R BT.500-13, page 15).

In *DSIS* case, the same reference and processed content are presented to the observer for scoring as in the DSCQS case, but only the impairments (the visual differences) are judged during the evaluation. This time, a " *five-grade impairment scale should be used*", as illustrated in Figure I-3.

5	imperceptible
4	perceptible but non annoying
3	slightly annoying
2	annoying
1	very annoying

Figure I-3: Five-grade impairment scale for DSIS method (*c.f* ITU-R BT.500-13 page 21).

To conclude with, *DSCQS* provides absolute quality evaluation for both the content under evaluation and its reference while the *DSIS* evaluates the impairments between these two types of content.

I.2.3.2. Comparison methods

The stimulus comparison methods consider two visual contents and provide a relative measure of the quality difference between the two of them. While the content is presented to the observer the same way as in double stimulus method, the observer is asked to evaluate the visual impact of the difference between the two types of content. Note that during comparison method, the two content can be of different types (*i.e.* different natural scenes captured by two different cameras). Yet, in the case in which one the content correspond to a reference and the other to a processed version of that reference, the comparison method becomes equivalent to DSIS. The grading scale is this time a "*categorical scale*", as illustrated in Figure I-4.

-3	Much worse
-2	Worse
-1	Slightly worse
0	The same
+1	Slightly better
+2	Better
+3	Much better

Figure I-4: Categorical scale for Comparison method (*c.f* ITU-R BT.500-13, page 20).

I.2.3.3. Single stimulus methods

In single stimulus methods, the subject assesses each content individually, independent with resect to its reference.

The most popular single-stimulus method is *Single Stimulus Continuous Quality Evaluation* (SSCQE). This method is introduced by ITU in order to emulate the most common situation in practice, namely when the final user watches video on a private, professional uncontrolled set-up (*e.g.* 2D or stereoscopic TV set). The years following this, the practical relevance of such an approach steadily increased, as the double stimulus method of laboratory testing does not replicate the single stimulus home viewing conditions.

SSCQE can be applied on both quality scaling and impairment scaling using corresponding rating scales, as illustrated in Figure II-5. Of course, the interpretation of the results is different. In the former case, an absolute quality evaluation of the content is obtained, based on its visual appealing. The latter implicitly assumes that the scoring observer has some background knowledge about would-be reference content and consequently, he/she scores such would be impairments.

Quality		Impairment	
5	Excellent	5	Imperceptible
4	Good	4	Perceptible but non annoying
3	Fair	3	Slightly annoying
2	Poor	2	Annoying
1	Bad	1	Very annoying

Figure I-5: Categorical scale for Single Stimulus method (*c.f* ITU-R BT.500-13, page 18).

I.2.4. The content under evaluation

The choices related to the content under evaluation are implicitly or explicitly addressed at various levels of the ITU Recommendations.

First, different ITU specifications relate to different type of content (*e.g.* 2D video or stereoscopic video). Moreover, upper limits are set for the total duration of the testing sessions (30 min) and for the individual sequence duration (30 s); yet, continuous time evaluation sessions are also included in the ITU specifications.

No clear specifications are set either to the total number of excerpts (still image or video sequence) to be evaluated or to the number of excerpts an individual viewer is recommended to evaluated. However, the total number of evaluations (the sum of the number of excerpts each evaluator scores) should meet (*via* the confidence limits and the underlying relative error) the precision required by the application.

I.2.5. Scoring and post-processing of the scores

The scoring follows the constraints and procedure inner to the evaluation procedure and the grading scale. In practice both convention (*pencil and paper*) and software scoring tools are encountered [MSU13].

The first step in the scores post-processing is the observer outliers' detections and ITU recommends in this respect a kurtosis criterion.

Once the outliers eliminated, the mean opinion score *MOS* is computed:

$$MOS = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{I-1}$$

where $x_1, x_2, \dots, x_N$ are the N scores assigned by the N observers (after the outliers elimination).

The *MOS* is expected to be presented alongside with its 95% confidence lower and upper limits $MOS - \delta$ and $MOS + \delta$ respectively, where δ is the 95% estimation error:

$$\delta = 1.96\frac{S}{\sqrt{N}} \tag{I-2}$$

where *S* is the unbiased estimator for the standard deviation:

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - MOS)^2 \tag{I-3}$$

I.2.6. Controversial ITU issues

This walk-through some ITU specifications brings to light that rather than imposing consensual subjective quality evaluation test conditions, the ITU Recommendations let all doors open for an experimenter and that several surprising, controversial aspects are still open to questions and interpretations:

- the terms *continuous* and *discrete* can be jointly used for a same grading scale (e.g. DSCQS) and their common, mathematical meanings seem no longer to hold;

- the *MOS* and its confidence limits can be computed on 15 and less data (assuming outliers are discarded) with traditional formulae which hold (for such few data) only assuming that the scores are Gaussian distributed; moreover, for 15 and less data, the relative estimations error is expected to be somewhat large;
- the grading scales are presented to the observers across with some semantic labels (*Excellent*, *Good*, *Fair*, *Poor* and *Bad*) which are subsequently mapped to values between 5 and 1; such an even mapping implicitly assumes that the 5 lables have an equal semantic impact in the evaluation.

The next Chapter presents a state-of-the-art study bringing to light how the research community currently addresses these open issues.

I.3. Thesis structure

This succinct introduction brings to light that rather than imposing consensual subjective quality evaluation test conditions, the ITU Recommendations let all doors open for an experimenter.

Such a situation becomes even more challenging these days when the home video consumption becomes more and more demanding in quality. For instance, [WEB03], since 2005, YouTube continuously increases the quality of its video archives. Initially, YouTube video is displayed at a 320×240 resolution, using a version of H.263 codec. Two years latter, an option to watch videos in 3GP format on mobile phones is added. In March 2008, a high-quality mode is introduced, the aspect ration is changed to 16:9 and the MPEG-4 AVC/H.264 encoder is provided by default. The 1080p HD support appears in 2009 while starting by 2010, the 4K format (up to 4096×3072 pixels) becomes available. In June 2015, support for 8K (7680×4320 pixels) is added and HDR video is made available one year latter. Beyond Internet video, note that at the time of writing (the fourth quarter of 2017) Netflix reached 117.58 million streaming subscribers worldwide [WEB04].

Hence, a particular attention should be paid on each and every detail involved in the quality assessment and a particular focus should be made on the single stimulus evaluation that better fits the home evaluation conditions, as stated by ITU-R BT.1788, ITU-R BT.500-11, ITU-R BT.500-13, ITU-T P.913, …

In order to encompass issues from ITU recommendations to reliable and precise subjective quality evaluation, the present thesis has the following structure.

Chapter II presents a concise state-of-the-art study, bringing to light both a methodological panorama of subjective visual quality assessment (Section II.1) and some insights on the grading scales for subjective evaluation (Section II.2). Section II.3 presents some current-day interfaces for ITU-based quality evaluations while Section II.4 identifies the main state-of-the-art limitations and précises the challenges taken by the thesis.

Chapter III describes the test-bed designed and deployed in the thesis. It observes to the ITU-R BT.1788, ITU-R BT.500-11, ITU-R BT.500-13, ITU-T P.913 specifications and the SSCQE method. The chapter is structured according to the main aspects in any ITU evaluation session, namely: the viewing conditions, the panel size and composition, the evaluation methodologies and grading scale, and, finally, scoring and post-processing of the scores.

Chapter IV theoretically bridges the continuous and the discrete scale evaluation procedures and investigates whether the number of the classes on the discrete scales is a criterion meaningful in the results interpretations or just a parameter. The instantiations for Gaussian and non-Gaussian models for the scores assigned by the observers on continuous scale are presented in Chapters IV.2 and IV.3, respectively. Both Chapters IV.2 and IV.3 include the theoretical ground for *MOS* and its confidence limits computation as well as experimental illustrations; studies on the accuracy of the results according to the number of observers in the panel are also presented.

Chapter V starts by demonstrating the existence of the semantic impact (Chapter V.1) and follows by defining a methodological framework for quantifying this semantic impact (Chapter V.2). The methodological approach is presented alongside with the underlying quantitative results.

Conclusions are drawn and perspectives are open in Section VI.

Appendixes A and B complete the quantitative results presented in Chapter IV.

II. State of the art

The state-of-the-art study is structured into two main parts, related to the methodological panorama of subjective visual quality assessment and to a presentation of the insights on the grading scales for subjective evaluation. Specifically, 37 studies, spanning about 100 years of research carried-out in various fields (visual quality, psychology, psychophysics, sociology, marketing, medicine, ...) are investigated. By regrouping their key aspects into three tables and one synoptic representation, it is noticed that the controversial ITU issues identified in the Introduction section (relating to the continuous vs. discrete evaluation scales, to the statistical distribution of the scores assigned by the observers and to the usage of semantic labels on the grading scales) are not yet solved. This way, three objectives are identified for the thesis:

- bridging at the theoretical level the continuous and the discrete scale evaluation procedures and investigating whether the number of the classes on the discrete scales is a criterion meaningful in the results interpretations or just a parameter; studying the theoretical influence of the statistical model of the evolution results and of the size of the panel (number of observers) in the accuracy of the results are also here targeted;

- quantifying the bias induced in subjective video quality experiments by the semantic labels (e.g. Excellent, Good, Fair, Poor, and Bad) generally associated to the discrete grading scales;

- designing and deploying and experimental test-bed able to support the precision and statistical relevance for the targeted results.

The Introduction chapter identifies several surprising, controversial ITU aspects, related to the use of *continuous* and *discrete* terms in conjunction with a grading scale, to the statistical model of the scores assigned by the observers as well as to the use of semantic labels for the grading scales. Actually, these aspects echo a longer research effort, carried out even outside the visual quality evaluation.

For 100 years already, various research fields (psychology, psychophysics, sociology, marketing, medicine, ...) consider the use of rating scales in subjective evaluations. For instance, back to 1923, a study published in the *Journal of Educational Psychology* [FRE23] raises the need for "*constructive effort … towards improving the means whereby ratings are obtained".* According to this study, a discrete grading scale has as advantages its simplicity and its self-explanatory character. On the contrarily, on continuous scales, the observer is somehow freed from the constraints imposed by the "*quantitative terms*". The study also discusses various labeled scales: the number of grading points varies from 2 to 11 while the labels assigned to these scales can be as different as *Very impressive*, *Brilliant*, *etc*. (for a psychology application).

One year before I was born, the study in [FRO89] discusses the choice of the methodology to measure the health state (labels of the scale, data interpretation). It is brought to light that the choice of the labels depends on the investigator's purpose and recommend limiting the number of the labels to "*9 or even fewer attributes*".

As a comprehensive state-of-the-art survey in such a broad field is practically impossible to be achieved, the present thesis will illustrate the variety of approaches through 37 studies, structured at two levels:

- *methodological panorama of subjective visual quality assessment*, see Section II.1 and Table II.1, discusses 21 studies, namely [MUE02], [NIN06], [CAM07], [BEN08], [HEW08], [NIN09], [SIM09], [Gol10-01], [GOL10-02], [SIM11], [BOS11], [CHE12-01], [CHE12-02], [URV12], [CHA13], [BOS13], [PIE14], [HAN15], [FAN15], [SU15] and [GAO13];
- *insight on the grading scales for subjective evaluation* (Section II.2), discusses 16 studies, divided into two sub-sections:
    - *background studies in subjective quality evaluation*, see Section II.2.1 and Table II.2.1: [MAT71], [MCK78], [COX80], [ALB81], [SVE00], [PRE00];
    - *visual quality assessment focused studies*, see Section II.2.2 and Table II.2.2: [JON86], [NAG93], [TEU96], [WAS98], [WIN03], [ZIE07], [WIN09], [PEC08], [HUY07, [HUY11].

Additionally, the issues connected to the software support for scoring are presented in Section II.3.

II.1. Methodological panorama of subjective visual quality assessment

This section regroups 21 studies related to the subjective quality evaluation for image/video and/or audio content, presented in chronological order (between 2002 and 2015).

For each investigated state-of-the-art study, the main ITU aspects related to a subjective evaluation (as introduced in Section I.2) will be addressed: the viewing conditions, the panel composition, the evaluation methodology and the grading scales, the evaluated content and the scoring and the post-processing of the scores.

The study [MUE02] considers a video-conferencing application and aims at investigating the variation in subjective quality evaluation introduced by the interaction between the audio and video modalities. Two tests are conducted: a passive test (in which the subject is simply listening to/viewing the content) and an interactive communication (person to person audio/video data exchange). The viewing conditions (in the ITU sense) are not presented. A panel of 20 observers is composed and asked to rate using a 5 point rating scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad*. The results of the study show a strong interaction dependency between audio and video content and discuss a unique benefit on multimedia quality for its psychological effects.

The study reported in [NIN06] is centered on the role that the visual attention can play in image quality assessment. The viewing conditions are set according to ITU-R BT.500-10, with subjects positioned at a viewing distance 4 times larger than the height of the picture. The panel is composed by 20 observers. A corpus of 20 images is evaluated: 10 images are considered as undistorted while the rest of 10 have various artifacts. A 5 levels discrete impairment grading scale is considered.

A methodology for subjective assessment of stereo images is presented in [CAM07]. The tests take place in a laboratory environment set according to ITU-R BT.1438. 17 observers are asked to score according to a *SAMVIQ* (Subjective Assessment of Multimedia Video Quality) method, with a 5-levels discrete labeled scale: *Excellent*, *Good*, *Fair*, *Poor* and *Bad*. While the size of the evaluated corpus is not précised, six different distorted contents (JPEG and JPEG2000 compressions blur) are considered.

The aim of [BEN08] is to introduce an objective quality metric for stereo images quality assessment which relies on both the use of 2D metrics and depth information. The testing environment complies with ITU-R BT.500-11. 17 observers scores 66 video sequences (6 reference and 60 downgraded sequences) according to a *SAMVIQ* methodology, assigned on a continuous scale ranging from 0 to 100.

The study in [HEW08] investigates the correlation between subjective and objective assessment of color plus depth map 3D video. The testing conditions are not explicitly stated. 32 observers participated on the experiments using a DSCQS method, with a scale ranging from 1 to 5, where 1 represents *Bad* image quality/depth perception and 5 represents *Excellent* image quality/depth perception. The stimulus set contains 13 coded video sequences and the original, uncompressed version of each scene is used as the reference in the evaluation test.

The study in [NIN09] advances a perceptual full reference video quality assessment metric based on the temporal evolutions of the spatial distortions. This metric is compared to a subjective rating in order to

investigate its efficiency. The tests are conducted into an environment observing to ITU-R BT.500-10. 36 observers participated on subjective assessment using an impairment discrete scale to rate the impairments induced by the MPEG-4 AVC (H264) codec (used at 5 different encoding configurations) on 10 reference video sequences.

A procedure for subjective evaluation of the JPEG XR codec for compression of still pictures is described in [SIM09]. The viewing conditions are set according to ITU-R BT.500-11 and are described in detail. A total of 16 subjects took part in experiments and each subject attended four test sessions, scoring a total of 208 test images. A DSCQS method is considered, with a grading scale featuring both continuous markers – form 0 to 100 and the 5 generally considered discrete labels *Excellent*, *Good*, *Fair*, *Poor* and *Bad*.

The studies in [GOL10-01] and [GOL10-02] are devoted to the definition of objective stereoscopic video quality metrics. The stereoscopic video database that contains a large variety of scenes captured by using a stereoscopic camera setup consisting of two HD camcorders with different capture parameters. The testing conditions are set according to ITU-R BT.500-11. 20 subjects (6 females, 14 males) are scoring the content; among them, 3 are discarded as outliers. An SSCQS method and a 5 levels discrete grading scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad* are considered. The corpus considers 6 different scenes and, for each scene, 5 different capturing parameters. This study brings to light a particular point of view related to stereoscopic video evaluation: "*for non-expert viewers it is quite difficult to distinguish 5 quality levels (excellent, good, fair, poor, bad)"* and *"3 quality levels (good, fair, bad) may be more appropriate"*.

This study in [SIM11] describes a subjective evaluation framework which is meant to serve as a test-bed for the joint collaborative team video coding (JCT-VC) efforts towards the definition of the HEVC standard. The testing conditions are set according to ITU-R BT.500-11 and are described in detail. A total of 494 naive observers are inquired, among which about 30% are females. Two different evaluation methods are considered:

- the DSIS method, with a joint numbering and semantic labels marks (*Imperceptible*, *Perceptible but non annoying*, *Slightly annoying*, *Annoying* and *Very annoying*), as depicted in Figure II-1.a.
- the DSCQS method, with a grading scale featuring both continuous markers – form 0 to 100 and the 5 generally considered discrete labels *Excellent*, *Good*, *Fair*, *Poor* and *Bad*, as depicted in Figure II-1.b.

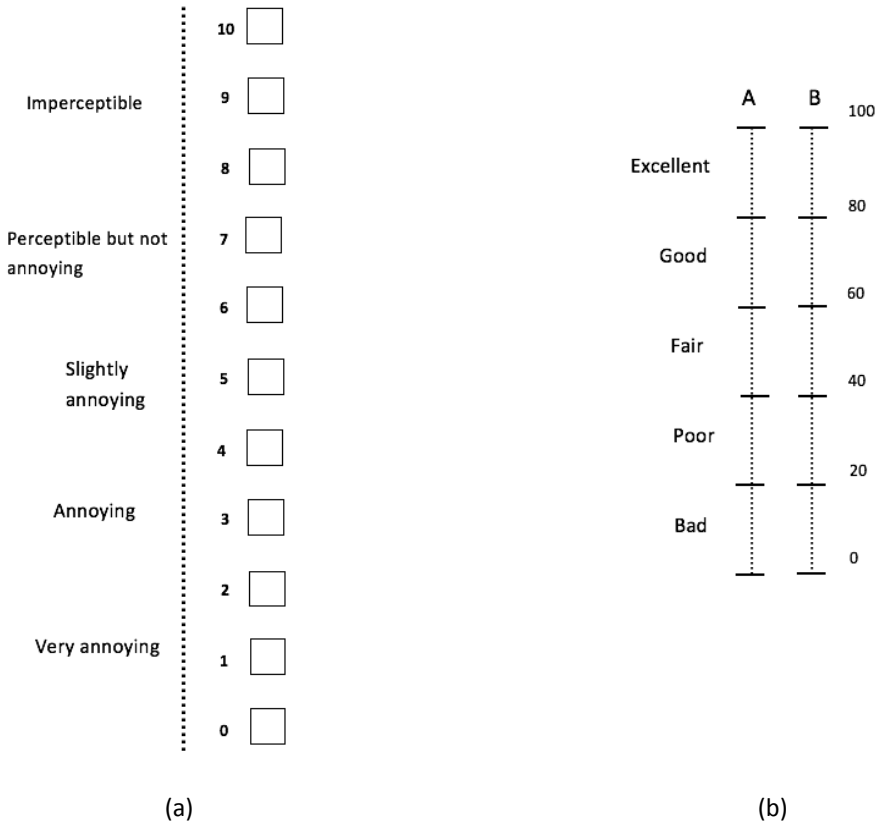


Figure II-4: Grading scales used in [SIM11]; (a) DSIS method scale, (b) DSCQS method scale.

In [BOS11], a DIBR (Depth Image-Based Rendering) synthesized view evaluation problem is considered, namely the reliability of some objective quality metrics in 3DTV. The testing conditions are set according to ITU-T P.910. A panel of 43 naïve observers participated in the experiment. Three test sequences are used to generate four different viewpoints, corresponding to 12 synthesized sequences for each tested algorithm (84 synthesized sequences in total). Two evaluation methods are considered:

- *Absolute Categorical Rating* (ACR), with *Excellent*, *Good*, *Fair*, *Poor* and *Bad*;
- *Paired Comparisons* (PC), according to which the observer selects the content that best matches the targeted criterion; the conversion between the results thus obtained and a virtually continuous scale is discussed.

The study [URV12] provides the scientific community with a database of high quality Full-HD stereoscopic sequences, shot with a semiprofessional 3D camera. The performances of this database are hinted to by an example of usage for evaluating encoding properties. The underlying subjective video quality evaluation follows ITU-T P.910. 29 observers (12 females, 17 males) performed the ACR-HR (ACR – Hidden Reference) evaluation on a 5-level quality scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad*. The total evaluated content is composed of 99 sequences of 16s and 11 sequences of 13s.

The study [CHE12-01] aims at enhancing the understanding of the perception of MPEG-4 AVC compressed stereoscopic 3D videos, in particular spatial video quality, depth quality, visual comfort and overall 3D video quality. Actually, a subjective assessment test was conducted in order to identify

whether some of the subjective quality scores can be incorporated into a single stereo 3D quality database or whether separate databases are needed to study different aspects of 3D QoS (Quality of Experience). A SSCQS method, with continuous grading scale between 0 and 100 was considered. 11 video sequences (a 3D pristine video, a 2D pristine video (right view), and nine distorted videos) were shown to the subjects. The number of observers is not presented.

The study [CHE12-02] aims to explore 3D QoE by constructing the visual experience as a weight sum of image quality, depth quantity and visual comfort. Two experiments in which depth quantity and image quality are varied respectively are designed to validate this model. The test conditions follow ITU-T BT.500-11. In the first experiment, the stimuli consist of three natural scenes and for each scene, four levels of perceived depth variation in terms of depth of focus (0, 0.1, 0.2 and 0.3 diopters) are considered. In the second experiment, five levels of JPEG 2000 compression ratio (0, 50, 100, 175 and 250) are used to represent the image quality variation. 28 observers participated in the test and scored according to a SAMVIQ method labeled with *Excellent*, *Good*, *Fair*, *Poor* and *Bad*.

Robust watermarking techniques for stereoscopic video protection are presented in [CHA13]. The testing conditions are set according to ITU-R BT.500-12. A panel of 15 naïve observers participate in the experiments and is required to evaluate 34 randomly chosen video excerpts of 40 seconds each. The scores are assigned on a 5 levels discrete quality scale with 5 levels with *Excellent, Good, Fair, Poor and Bad* labels according to the SSCQS method.

[BOS13] presents a study on subjective quality assessment for free-viewpoint video sequences generated from decompressed data. The testing conditions are set according to ITU-T BT.500 and are considered to assess 276 FVV (Free Video Viewpoint) sequences. 27 naïve observers participate in the subjective quality evaluation, according to an ACR-HR method and a 5-level labeled quality scale (*Excellent*, *Good*, *Fair*, *Poor* and *Bad*).

The study in [GAO13] investigates the blind image quality assessment (perceptual image quality scores without access to reference images) and organizes a subjective assessment test where the viewing conditions follow the ITU-R BT.500-13. 24 subjects evaluate distorted images with (JPEG2000 compression, JPEG compression, Gaussian Blur, WN and FF) and assign their scores on a continuous scale from 0 to 100.

The effect the chromatic variations in 3D video have on the perceived visual quality is investigated in [PIE14]. The tests conditions follow ITU-R BT.500-12. 25 subjects participate in the test and assigne their scores on an ACR-HR continuous scale with 11 points. A corpus of 8 video sequences (whose duration is not precised) is evaluated.

The study [HAN15] reports on an extensive benchmarking of objective quality metrics for HDR image quality assessment. In total, 35 objective metrics are benchmarked on a database of 20 HDR contents encoded by 3 compression algorithms at 4 bit rates, leading to a total of 240 compressed HDR images. A subjective assessment is conducted to verify the accuracy of the presented metrics. A total of 24 naïve subjects (12 females and 12 males) took part in the experiments. The test conditions follow the ITU-R BT.2022 and the observers assign their score on double-stimulus impairment scale (DSIS) with 5 Labels according to ITU-R BT.500-13.

A method for no-reference quality assessment of contrast distorted images based on the principle of natural scene statistics (NSS) is presented in [FAN15]. In order to validate this method, a subjective assessment test following ITU-R BT.500-12 is conducted. 57 subjects participate in the test. A first group of 22 observers assigned their scores on a discrete scale ranging from 1 to 5 and evaluate a database composed of 415 still images. A second group composed of 35 observers considers a discrete scale ranging from 0 to 9 and evaluate of 200 still images. The type of evaluation (single stimulus or double stimulus) is not précised. Although this study brings to light the usage of two different grading scales, it does not make possible any comparison between the 5 levels and the 9 levels scoring (as the two databases are different).

The study in [SU15] considers oriented correlation models of distorted natural images with application to natural stereopair quality evaluation. A subjective assessment test following the ITU-R BT.500-11 is conducted in order to validate those models. The observers (whose number is not presented) assign their score according to the DSIS method, with a joint numbering and semantic labels marks (*Imperceptible*, *Perceptible but non annoying*, *Slightly annoying*, *Annoying* and *Very annoying*). The viewing conditions are not presented.

The principles of these various research studies are illustrated in Table II-1 which brings to light that the subjective visual quality evaluations are very heterogeneous in their deployment and that a convergence is still far for being reached. While difference are encountered at practically any level (targeted application, experimental conditions, panels, grading scales and processed content), they are particularly controversial for the grading scales; hence, these aspects will be deeper investigated in Chapter II.2.

Table II-1: Methodological panorama of subjective visual quality assessment.

	Context	Testing conditions	Panel	Scale	Evaluated content
[MUE02]	The variation in subjective quality evaluation introduced by the interaction between the audio and video modalities	ITU-T 1984	20	5 levels discrete quality scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad*	-passive test: the subject is simply listening to/viewing the content -interactive communication: person to person audio/video data exchange
[NIN06]	the role of the visual attention in image quality assessment	ITU-R BT.500-10 DSIS	20	5 levels discrete impairment grading scale labeled by *Imperceptible*, *Perceptible but non annoying*, *Slightly annoying*, *Annoying* and *Very annoying*	-10 undistorted images -10 images with various artifacts
[CAM07]	the methodology for subjective assessment of stereo images	ITU-R BT.1438 SAMVIQ	17	continuous scale from 0 to 100 with labels	six different distorted contents (JPEG and JPEG2000 compressions, blur)
[BEN08]	objective quality metrics for stereo images quality assessment relying on both the use of 2D metrics and depth information	ITU-R BT.500-11. SAMVIQ	17	continuous scale ranging from 0 to 100	- 6 original images. - 60 stereo images with different degradation levels (JPEG and JPEG2000)
[HEW08]	the correlation between subjective and objective assessment of color plus depth map 3D video	DSCQS	32	discrete, unlabeled scale ranging from 1 to 5	-13 original uncompressed videos -13 coded video sequences
[NIN09]	full reference video quality assessment metric based on the temporal evolutions of the spatial distortions	ITU-R BT.500-10 DSIS	36	5 levels discrete impairment grading scale labeled by *Imperceptible*, *Perceptible but non annoying*, *Slightly annoying*, *Annoying* and *Very annoying*	10 reference video sequences distorted with the MPEG-4 AVC (H264) codec (used at 5 different encoding configurations)
[SIM09]	a procedure for subjective evaluation of the new JPEG XR codec for compression of still pictures	ITU-R BT.500-11 DSCQS	16	continuous scale from 0 to 100 with the labels *Excellent*, *Good*, *Fair*, *Poor* and *Bad*	still pictures compressed with JPEG XR codec
[GOL10]	objective stereoscopic video quality metrics	ITU-R BT.500-11 (SSCQS)	20	5 levels discrete quality scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad*	6 different stereoscopic videos content; each video is captured with different parameters
[SIM11]	a subjective evaluation framework serving the HEVC standard	ITU-R BT.500-11 DSCQS and DSIS	494	5 levels discrete impairment grading scale labeled by *Imperceptible*, *Perceptible but non annoying*, *Slightly annoying*, *Annoying* and *Very annoying* 5 labels discrete quality scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad*	HD content divided into 5 classes with different spatial and temporal resolutions
[BOS11]	reliability of objective quality metrics in 3DTV considering DIBR (Depth Image-Based Rendering) synthesized view evaluation problem	ACR	43	5 levels discrete quality scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad*	84 synthesized sequences: 3 test sequences generating 4 viewpoints, corresponding to 12 synthesized sequences for each tested algorithm

Table II-1: (continuing) Methodological panorama of subjective visual quality assessment.

[URV12]	a database of high quality Full-HD stereoscopic sequences, shot with a semi-professional 3D camera	ITU-R BT.500 ACR-HR	29	5 levels discrete quality scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad*	99 sequences of 16s and 11 sequences of 13s of full-HD stereoscopic sequences
[CHE12-01]	perception of MPEG-4 AVC compressed stereoscopic 3D videos	SSCQS	NA	continuous scale ranging from 0 to 100	video sequences (a 3D pristine video, a 2D pristine video (right view), and 9 distorted videos with MPEG-4 AVC compressed
[CHE12]-02	3D QoE by constructing the visual experience as a weight sum of image quality, depth quantity and visual comfort	ITU-R BT.500-11 SAMVIQ	28	5 levels discrete quality scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad*	3D distorted content with variation of depth focus and different levels of JPEG2000 compression.
[CHA13]	robust watermarking techniques for stereoscopic video protection	ITU-R BT.500-12 DSCQS	25	5 levels discrete quality scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad*	watermarked stereoscopic video content (34 randomly chosen video excerpts of 40 seconds each)
[BOS13]	subjective quality assessment for free-viewpoint video sequences generated from decompressed data.	ACR-HR	27	5 levels discrete quality scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad*	Free-viewpoint video sequences generated from decompressed data
[GAO13]	blind image quality assessment	ITU-R BT.500-13 Paired comparison	24	continuous scale ranging from 0 to 100	distorted images with (JPEG2000 compression, JPEG compression, Gaussian Blur, WN and FF)
[PIE14]	the effect of the chromatic variations in 3D video on the perceived visual quality	ACR-HR	25	continuous scale with 11 points	8 video sequences of 3D video content
[FAN15]	method for no-reference quality assessment of contrast distorted images based on the principle of natural scene statistics (NSS).	ITU-R BT 500-12	57	discrete scale from 1 to 5 discrete scale from 0 to 9	distorted images based on the principle of natural scene statistics (NSS)
[HAN15]	extensive benchmarking of objective quality metrics for HDR image quality assessment	ITU-R BT.2022 DSIS	24	5 levels discrete impairment grading scale labeled by *Imperceptible*, *Perceptible but non annoying*, *Slightly annoying*, *Annoying* and *Very annoying*	240 compressed HDR images: 20 HDR contents encoded by 3 compression algorithms at 4 bit rates
[SU15]	an oriented correlation models of distorted natural mages with application to natural stereopair quality evaluation	ITU-R BT.500-11 DSIS	NA	5 levels discrete impairment grading scale labeled by *Imperceptible*, *Perceptible but non annoying*, *Slightly annoying*, *Annoying* and *Very annoying*	distorted stereopair images

II.2. Insight on the grading scales for subjective evaluation

The present section investigates the details related to the grading scales. In this respect, both background studies emerged from various research fields and visual quality assessment are discussed.

II.2.1. Background studies in subjective quality evaluation

Conducted under the applied psychology framework, the study in [MAT71] investigates the reliability of 18 different discrete scales, with the number of categories ranging from 2 to 19. The reliability of the scales is discussed with respect to three factors that are a *priori* important in determining the number of alternatives to employ: (a) the proportion of the scale which is effectively considered by the subjects when scoring, (b) the duration required for testing, and (c) whether or not an "uncertain" category is provided. 360 students (20 students for each scale) are considered in the scoring sessions. The scale reliability was assessed by an analysis of the variance (second order moment) of the scores; in this respect, the Fisher's test is used. The conclusion is that, at least from this point of view, the scores are largely independent from the number of rating points on the scale: 16 out of the 18 scales examined did not differ significantly. The two exceptions correspond to the 2 level and 3 level grading scales. The results also show that the testing time increases with the number of levels on the scale while the usage of the "uncertain" category decreases as the number of rating steps increases.

7 years later, in the same research field, the study reported in [MCK78] investigates the reliability and validity of the scores assigned on a continuous scale and on discrete scales with 5, 7 and 11 categories. For both continuous and discrete scales, both labeled and unlabeled versions are presented to 30 subjects involved in the experiments. The conclusions are of different types. First, the continuous scale is "*most pleasing"* to be used. Secondly, the results brought no evidence that the continuous scale would provide either more discrimination or better accuracy than the discrete scales. Concerning the discrete scales, the results brought to light that 5 or 6 categories should be considered for evaluation. It is also stated that even on a continuous scale, the scores assigned by the observers are somewhat clustered into 5 or 6 classes (*"subjects using the continuous scale appear to be operating essentially with 5 or 6 categories"*).

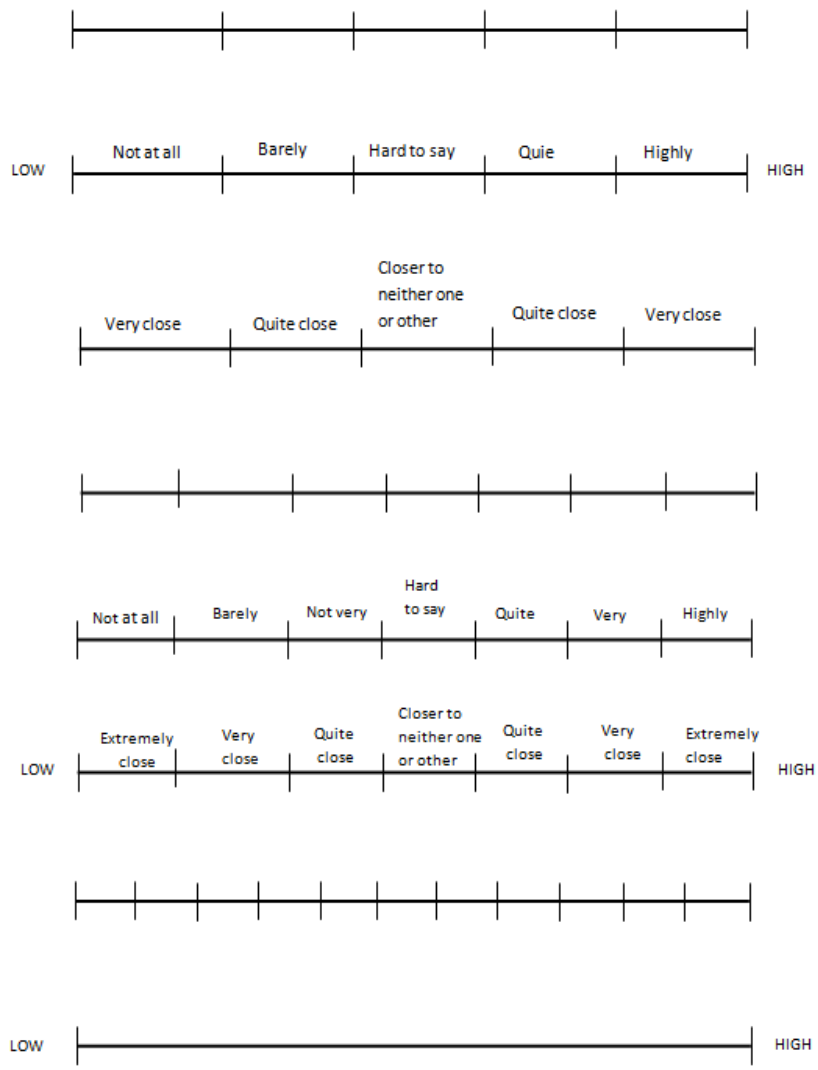


Figure II-5 : Grading scales used in [MCK78].

The problem of the optimal number response alternatives on a discrete scale is also raised under the framework of marketing research, [COX80]. By using information theory tools, it is considered that the optimal number is the one that maximizes the information provided by respondents in a test, while minimizing the response errors or the likelihood of the random responses. In is thud concluded that using a large number of scale categories (higher than 9 to 12) results in no benefit, while a very small number of categories (less than 5 or 6) could produce a loss in accuracy.

In study conducted in academic context, [ALB81] compares a five-category discrete semantic differential scale with the corresponding unlabeled continuous scale. By differential scale it is understood a scale whose extremities are labeled, as for example *Friendly vs. Unfriendly* or *Modern vs. Old fashioned*. The discrete scale is presented to the user with intermediate marks but no semantic labels or numerical values are associated to these marks. 176 university students participated in the assessment test. The results are investigated *via* their mean value and variance, thanks to a paired Student's test. It is thus

demonstrated that no large variability among the discrete scales is encountered: only 4 out of the 30 pairs of scale were significantly different. It is also concluded that similar evaluation result can be obtained from the five-category discrete rating scale and from a continuous scale. Yet, it is explicitly stated that "*there is a considerable advantage of a continuous rating scale in applications for which individual measurement is important …*"

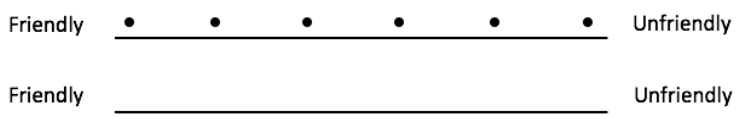


Figure II-6: Grading scales used in [ALB81].

Also in an academic context, [SVE00] compares a visual analog scale (similar to a continuous unlabeled scale), a graphic rating scale (similar to the standard ITU labeled continuous scale) and a five-point verbal descriptor scale (similar to the 5-point discrete category scale). 174 students participated in the subjective assessment test relating to dummy application, namely a self-assessment of the previous knowledge in statistics. As a general conclusion, it is stated that assessments on the discrete scale have the highest level of stability; particularly, it is shown that both the verbal descriptor scale and the graphic scale assessments provided a better order consistency compared to the visual analog scale assessment. The results also show that an increased number of possible responses did not guarantee a higher sensitivity of the assessments.

Coming back to the marketing field, [PRE00] provides to 149 participants a questionnaire concerning the service elements of a recently visited store or restaurant. The questionnaire used scales with a number of judgment category from 2 to 11, and a 101-point scale (from 0 to 100). It is thus shown that the scales that produce the least reliable scores are those with the fewest response categories. However, it is also found that a decrease in reliability is encountered for scales with more than ten response categories. The most reliable scores are found to be those from scales featuring between 7 and 10 response categories.

The principles of these studies are regrouped in Table II-2

Table II-2: Background studies related to grading scales.

	Context	Testing conditions	Panel	Scale	Applicative field
[MAT71]	the optimal number of categories in a discrete scale	N.A	360	18 different labeled discrete scales, with the number of categories ranging from 2 to 19.	psychology
[MCK78]	the optimal number of categories on a graphic rating scale	NA	30	continuous scale discrete scales with 5, 7 and 11 categories	psychology
[COX80]	the optimal number of response alternatives for a scale	NA	225	discrete scales with the number of categories ranging from 2 to 9	marketing
[ALB81]	continuous *vs.* discrete semantic differential rating scales,	NA	176	continuous scale 10 different categorical labeled scales (5 categories)	academic
[SVE00]	comparison of the quality of assessments using continuous and discrete ordinal rating scales	NA	174	continuous unlabeled scale discrete labeled scale	academic
[PRE00]	optimal number of response categories in rating scales	NA	146	continuous 101-point scale (from 0 to 100). discrete scale with number of categories from 2 to 11	marketing

II.2.2. Visual quality assessment

To the best of our knowledge, for visual quality assessment applications, the problem of the semantic impact of the labels in the overall quality evaluation was first raised by [JON86]. Two panels are considered for experiments: 49 persons compose a panel of English speakers (native USA speakers) while 24 persons compose a panel of Italian speakers. The experiments are conducted in parallel for English and Italian, but they will be illustrated here only for English. During the experiments, a continuous scale featuring no intermediate labels but having its two extremities marked with *Best imaginable* and *Worst imaginable* is presented to the panel. The panel members are asked to place, on that continuous scale, according to their own understanding, 15 adjectives (labels): *Superior*, *Ideal, Excellent, Good, Fine, OK, Fair, Passable, Marginal, Not Quite Passable, Poor, Inferior, Bad, Not Usable, Awful*. For both languages, the results showed that the ITU labels were not evenly distributed along the graphic scale suggesting a non-uniform semantic distance between adjacent ITU labels. Specifically, a kind of compression at the end points of the scale was identified and explained as a reluctance of the observers in using the continuous scale extremities. The results also show a clustering tendency, with 9 classes (e.g. the adjectives *Ideal*, *Excellent* and *Superior* are very close each-other).

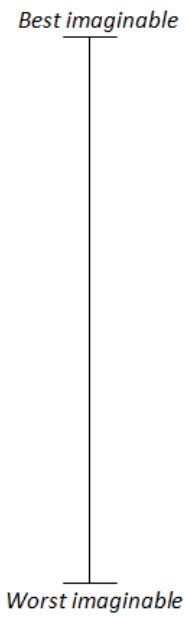


Figure II-4: Grading scales used in [JON86].

These results were later corroborated for other languages, such as Swedish, Dutch and British English, but not for Japanese and German, which exhibited uniformly distributed labels along the scale [NAR93], [TEU96], [ZIE7].

The semantic impact of the ITU Japanese descriptive terms for quality and impairment is investigated in [NAR93]. The two experiments follow the principles in [JON86]. A panel of 40 Japanese speakers is inquired. They are asked to alternatively position on a continuous scale with no intermediate but extremities labels either 13 quality terms or 12 impairment terms. It is thus brought to light that the perceived quality intervals are non-uniform. Yet, in Japanese, they are distributed more evenly than in English, French or Italian, following a similar trend as in the German case. The impairment experiment shows the Japanese terms have lower semantic impact than the corresponding terms in other languages. Note that the underlying experiments for German are presented in [TEU96].

The two ideas of semantic impact and language dependency related to the ITU labels are considered as a starting point for the research study presented in [WAS98]. In order to refine the precision and the stability of the results, the subjective quality evaluation is considered to be a multidimensional process and some means for identifying the different dimensions and the appropriate vocabulary are advanced. In this respect, the use of unlabeled continuous rating scale is considered as a ground for investigation. A panel of 24 subjects is asked to score audio-visual content on a continuous scale whose extremities are labeled by + and – signs. The results show that such a quality rating procedure is "*remarkably consistent*", thanks to the fact that the subjects set their own criteria. The results also show that using an unlabeled scale reduces the tendency of subjects to avoid the end points of the scale.

The SSCQE and DSIS methods are compared on the same test material in [WIN03]. In order to avoid the semantic impact of the *Excellent*, *Good*, *Fair*, *Poor* and *Bad* labels, the SSCQE method is presented to the observer as a vertical slider with only the two labels *Good* and *Bad* at the top and bottom ends of the

slider. On the contrarily, the DSIS scale follows the general ITU recommendations and is presented to the observer on a discrete, 5 levels scale ranging from *Imperceptible* to *Very annoying*. Results obtained from 20 observers show that the two methods are highly correlated and produce comparable quality results.

Three subjective audio quality evaluation tests are presented in [ZIE07]. Each of these three tests considers a different scale: the standard continuous scale with the 5 ITU labels, a 5-point continuous impairment scale, and a label-free continuous scale. The results showed a high similarity between the scores obtained with the ITU labeled scale and the label-free quality scale, with an almost perfect linear regression between them. Hence, this study supports the idea that the ITU quality scale is indeed an equal-interval scale.

The study in [HUY07] investigates the suitability of SAMVIQ assessment methodology; in this respect, two psychovisual experiments are carried out in two different laboratories. The subjective video quality evaluation follows the ITU-T P.910. The observers (whose number is not presented) assign their scores on a continuous 0-100 scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad*. The experimental results indicate that the SAMVIQ methodology provides results comparable to other existing methods, such as the single stimulus ACR methodology.

The ACR and SAMVIQ subjective quality assessment methodologies are compared in [PEC08]. The ACR is presented to the observers with a 5 levels discrete scale associated to the labels *Excellent*, *Good*, *Fair*, *Poor* and *Bad* while the SAMVIQ with a continuous scale ranging from 0 to 100, yet featuring the same labels. The viewing conditions are not précised; the number of observers participating in the test is 43. The results of this study show that the ACR uses 96.3% of the available range while SAMVIQ uses only 82%. It is thus demonstrated that the two assessment methodologies have different behaviors; it is also shown that the relation between their results depends on the evaluated content quality and it is subsequently stated that, for a given number of observers, SAMVIQ is more precise than ACR.

The variability of subjective ratings obtained with different scales (0-100 continuous scale and 5, 9, and 11 discrete scales) is investigated in [WIN09]. The study relays on simulated data instead of real experimental data, since it is considered that the differences among experiments available in the literature are too large for reliable direct comparison. It is concluded that although an increased discretization level of the scale leads in theory to an increase of the standard deviation of the scores (and therefore to a decrease of precision), practical proof of this effect remained inconclusive. He also found that the number of subjects may not need to be as high as generally assumed; in fact, the minimum of 15 recommended by ITU appears to be a very reasonable suggestion.

The study [HUY11] compares 4 different ITU grading scales with labels: two of them are discrete (with 5 and 9 levels) while the other two are continuous, yet with 11-point and 5-point grades. A subjective assessment test following the ITU-R BT.500-11 is conducted. 92 observers assign their scores according to the ACR method using the 4 different scales. The total evaluated content is composed of 128 sequences of 12s each. The results show that no significant statistical difference is found among subjective results obtained with the different four scales.

The examples of these various studies are regrouped in Table II-3.

Table II-3: Grading scales in visual quality assessment.

	Context	Testing conditions	Panel	Scale	Evaluated content
[JON86]	graphic scaling of qualitative terms	NA	73	continuous labeled scale	picture and sound
[NAG93]	graphic scaling and validity of Japanese descriptive terms used in subjective evaluation tests	EBU method (European broadcasting union)	40	continuous labeled scale	HDTV picture
[TEU96]	the validity of CCIR quality indicators along a graphic scale	CCIR 500-4 Recommendations	135	5 levels discrete quality scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad* continuous scale with labeled extremities continuous scale with numbers and labeled extremities	TV content
[WAS98]	perceived quality of speech and video in multimedia conferencing applications	ITU-R	24	continuous unlabeled scale	speech and video in multimedia conferencing applications
[WIN03]	video quality evaluation for internet streaming applications	ITU-R BT.500 DSIS SSCQE	20	5 levels discrete impairment grading scale labeled by *Imperceptible*, *Perceptible but non annoying*, *Slightly annoying*, *Annoying* and *Very annoying* continuous scale with labeled extremities “Bad” and “Good”	Internet streaming content
[ZIE07]	use of graphic scales in modern listening tests	ITU-R BS.1534 MUSHRA	13	5 levels discrete impairment grading scale labeled by *Imperceptible*, *Perceptible but non annoying*, *Slightly annoying*, *Annoying* and *Very annoying* A 5 labels discrete quality scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad*	audio
[HUY07]	examination of the SAMVIQ methodology for the subjective assessment of multimedia quality	SAMVIQ	N.A.	continuous labeled scale	reference sequences at CIF resolution (352 x 288 pixels) and derived from standard- or high-definition original content video content encode at MPEG4, H.264 and WMV9 codec
[PEC08]	suitable methodology in subjective video quality assessment: a resolution dependent paradigm	ACR SAMVIQ	43 observers	5 Labels discrete labeled scale continuous labeled 0-100 scale	HDTV set coded at 8 H.264 bitrates QVGA and VGA contents coded at H.264 bitrates

[WIN09]	the properties of subjective ratings in video quality experiments	ITU-T P.910 DSIS ACR	20 (simulated data)	continuous 0-100 scale discrete 11-point scale (0-10) discrete 9-point scale (1-9) discrete 5-point scale (1-5)	CIF, QCIF, SD,…
[HUY11]	study of rating scales for subjective quality assessment of High-Definition video	ITU-R BT.500 ACR	92 observers	5-point discrete labeled scale 11-point continuous labeled scale 5-point continuous labeled scale 9-point discrete labeled scale.	HDTV content coded at H.264 bitrates

II.3. Graphical user interface for ITU visual quality assessment

In the subjective assessment context, the GUIs are far more appealing versions of the old-fashioned *pencil and paper* evaluation sheets for score gathering. As their practical usage is traversal with respect to the evaluation procedure itself, they are presented into a separate section in the present thesis.

They are specifically designed so as to meet several convergent yet very different requirements. First, they should accurately reproduce the ITU evaluation conditions. Secondly, they should have virtually a zero impact in the evaluation result itself while allowing a speed-up of the evaluation procedure. Thirdly, they should be simple, versatile and user-friendly.

For instance, the MSU Video Group and Media Lab [MSU13] designed a specific GUI for three subjective assessment method, namely the DSIS, DSCQS and a so-called MSU-CQE (Continuous Quality Evaluation), as illustrated in Figure II-5. For the DSIS method, videos are successively presented (first the reference, then the distorted version). After their playback, the observer is asked to score using impairment scale, Figure II-5.a. In the DSCQS case, videos are played simultaneously in the same window. Each pair is repeated a given amount of times (*repetition* parameter in the Task Manager). During playback the observer is free to switch between two videos, named *Sequence A* and *Sequence B* by pressing Tab button. One of videos is the reference and the other is the impaired, but the observer is not informed about it. After playback the observer is asked to give his opinion about each video sequence, see Figure II-5.b. According to the MSU-CQE method, two sequences are played simultaneously; if during the playback, the assessor estimates that one sequence is worse than another, he/she strokes the left or the right arrow keys on his keyboard, depending on a position of the sequence he/she dislikes. In such a case, a red mark appears on top of the video the observer is voting against, see Figure II-5.c.

An interface for evaluating visual quality content according to an ACR method is provided by AccepTV [WEB05]. Here, the assessor asks to play the video, score it and then he/she can ask to watch it again, see Figure II-5.d.

As a current trend, these examples show that the ITU recommendations can be translated into user-friendly, application oriented software implementations. The study developed in the present thesis follows such a principle and develops an interface matched to the peculiarities of our investigation.

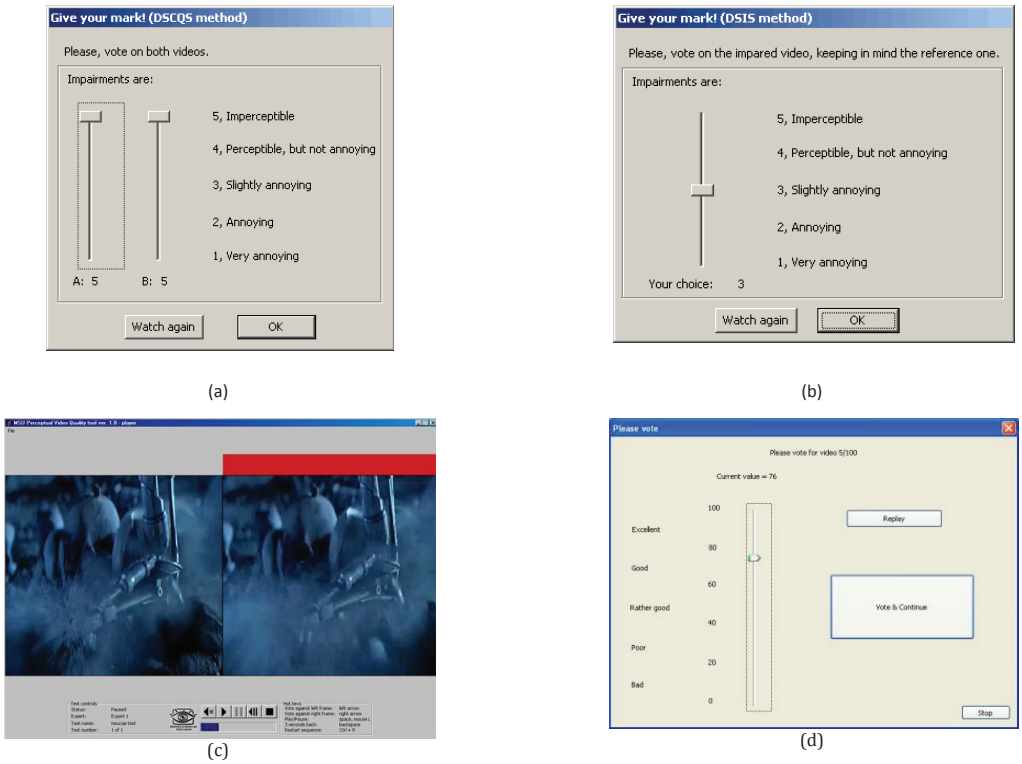


Figure II-5: Graphical user interface for ITU visual quality assessment: (a) A graphical user interface designed for the DSCQS methodology [MSU13], (b) A graphical user interface designed for the DSISS methodology [MSU13], (c) A graphical user interface designed for the MSU-CQE methodology[MSU13], (d) A graphical user interface designed for the ACR method by AccepTV [WEB05].

II.4. Thesis objectives

The present state-of-the-art study spans about 37 studies (*cf*. the synoptic representation in Figure II-6) covering about 100 years of research and, surprisingly, brings to light that controversial, open questions still arise about some key aspects related to the grading scales in subjective quality assessment. Actually, a basic taxonomy would consists in continuous *vs*. discrete and labeled *vs*. unlabeled categories:

- the a *priori* most intuitive category of scale, *i.e.* the continuous unlabeled scales is considered in 3 studies : [BEN08], [URV12], [GAO13];
- continuous scales with labels are also considered in 4 studies: [JON86], [NAR93], [CAM07], [SIM09]; an interesting point common to these 4 studies is the fact that the while the continuous scales range from 0 to 100, their discrete counter-parts are labeled with discrete terms (*Excellent*, *Good*, *Fair*, *Poor* and *Bad*).
- when considering discrete scales, the general tendency is to use the ITU labels, be them for quality (*Excellent*, *Good*, *Fair*, *Poor* and *Bad* in [MAT71], [COX80],[ZIE07],[MUE02], [GOL10-01], [GOL10-02], [SIM11], [BOS11], [URV12], [CHE12-02] [CHA13], [BOS13]) or for impairments (*Imperceptible, Perceptible but non annoying, Slightly annoying, Annoying and Very annoying* in [NIN06], [NIN09], [SIM11], [BOS11], [HAN15], [SU15]);
- some studies utilize discrete scale without labels [HEW08],[PIE14], [HAN15].

Yet, this taxonomy is far for being complete, and various intermediate situations are encounted:

- the studies [TEU96], [PRE00], [WIN03] consider both continuous and discrete labeled scale;
- in its turn, [WIN09] utilize continuous and discrete unlabeled scales;
- [MCK78] utilize discrete labeled and unlabeled scale;
- [ALB81], [SVE00] utilize continuous and discrete, labeled and unlabeled scales.

Consequently, the present thesis has as main objective to reconsider the ITU recommendations and to investigate on theoretical basis some of their key aspects related to evaluation scales. In other words, the main objective is to bridge, from both theoretical and methodological points of view, the continuous unlabeled and the discrete labeled realms (see Figure II-6). In this respect, three main research directions are to be considered:

First, the investigation will be carried out at the theoretical level. In this respect, the thesis will try to bridge the continuous and the discrete scale evaluation procedures and to investigate whether the number of the classes on the discrete scales is a criterion meaningful in the results interpretations or just a parameter. Studying the influence of the statistical model of the evolution results and the size of the panel (number of observers) in the accuracy of the results is also targeted.

Secondly, at the methodological level, the thesis will address the issue of quantifying the bias induced in subjective video quality experiments by the semantic labels (*e.g. Excellent*, *Good*, *Fair*, *Poor*, and *Bad*) generally associated to the discrete grading scales.

Finally, from the experimental point of view, these above-mentioned two directions require an experimental test-bed able to support their precision and statistical relevance. Hence, specifying and deploying such an experimental test-bed becomes an objective per-se. This way, an investigation on the

practical variability of the experimental and methodological results with the type of content (2D or stereoscopic video) or with its quality (as assessed by objective metrics) becomes also possible.

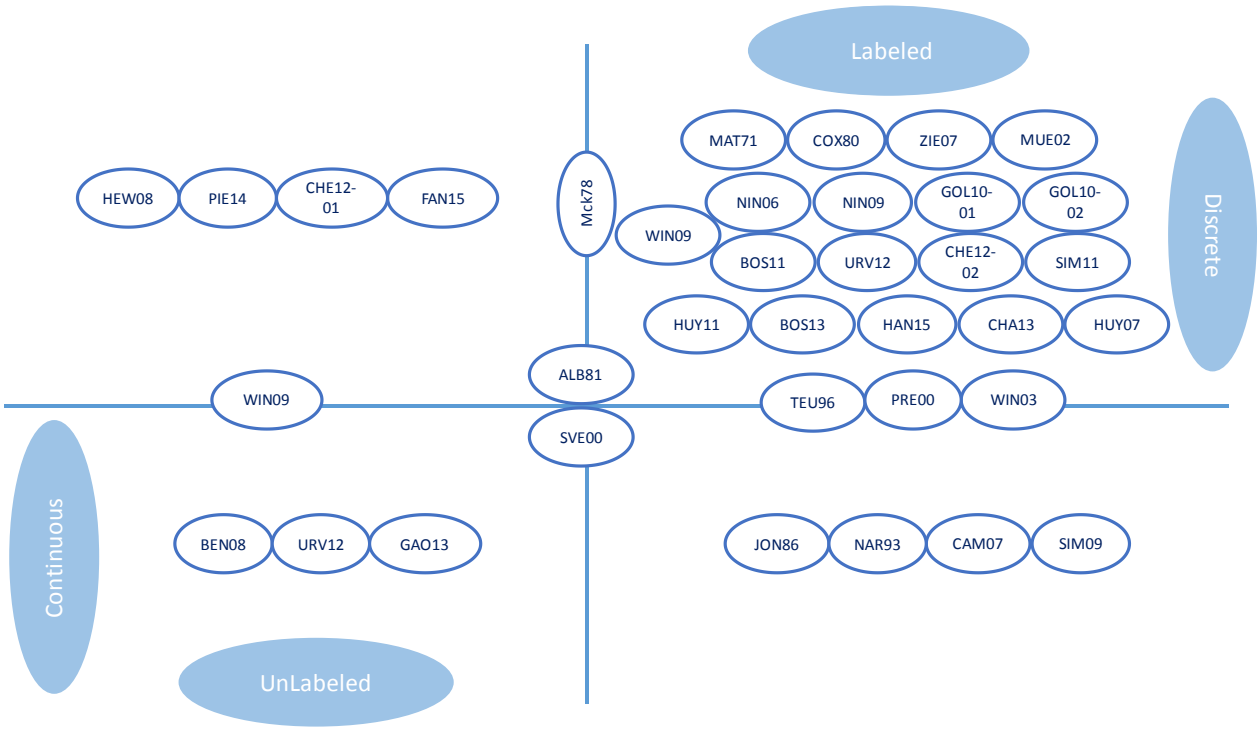


Figure II-6: State-of-the-art synopsis, according to a taxonomy based on the type of scale (continuous *vs.* discrete) and the presence of the labels (unlabeled *vs.* labeled). **The thesis has as objective to theoretically and methodologically bridge the evaluations achieved on continuous unlabeled scales (lower-left quarter) and discrete labeled scales (top-right quarter).**

III. Test-bed

The test-bed designed and deployed in the thesis consider the ITU-R BT.1788, ITU-R BT.500-11, ITU-R BT.500-13, ITU-T P.913 specifications and the SSCQE (Single Stimulus Continuous Quality Evaluation) method as a backbone and should adapt/extend it so as to fit to the evaluation procedure peculiarities.

In order to score the content, a versatile Android application is developed so as to alternatively allow the evaluation on continuous and discrete, semantically labeled scales.

The content to be evaluated is represented by four corpora, representing high and low-quality video (as a priori expressed by the PSNR value), both stereoscopic and 2D video. Each of these corpora is about 20 minutes long and is presented to the observers into downgraded versions obtained through watermarking or compression methods. Note that that as the study investigates the very evaluation procedure, the visual content just ensures realistic conditions for evaluation.

A total of 640 human observers are involved in the experiments. They were grouped in four main panels of 160 observers each, on panel for each type of content. In order to grant result repeatability, each main panel was split into three sub-panels, referred to as the reference (60 observers), validation (50 observers) and cross-checking (50 observers) panels. The subjects in the reference panels scores on the continuous scales. The subjects in both the validation and cross-checking panels are finally partitioned into two sets, scoring on 5 level labeled grading scales and on 3 level labeled grading scales, respectively.

III.1. Overview

By its nature, the present thesis belongs to field of random processes and statistics applied to visual quality assessment; hence, the relevance of the targeted results is essentially determined by the processed experimental data. Consequently, the test-bed designed and deployed for the thesis is presented *prior* to the main theoretical and methodological contributions.

Two criteria are considered in the design of the test-bed:

- the compliance with the most intensively considered ITU specifications;
- the possibility of ensuring reliability and relevance for the objectives targeted by the present thesis.

In the present study, the first criterion mainly applies to the viewing conditions and the evaluation methodology while the second criterion imposes strict constraints on the content under evaluation, thus determining the panel composition and size. The post-processing of the scores derives from the two of them.

It should be emphasizeD that the study investigates the evaluation procedure itself and is not focused on the content to be evaluated. Hence, the two above-mentioned criteria should be considered as complementary: while the ITU specifications are the backbone ensuring compatibility with content evaluation studies, they should be adapted/extended so as to fit to the evaluation procedure peculiarities.

The presentation is structured according to the main aspects related to any test session (Section I.2), namely the viewing conditions, the panel composition and size, the evaluation methodology and the grading scales, the content under evaluation, the scoring and the post-processing of the scores.

III.2. Viewing conditions

The evaluation has been conducted at the Advanced Research &Techniques for Multimedia Imaging Systems (ARTEMIS) Department at Telecom SudParis engineering school in France.

The viewing conditions are set in concordance with ITU-R BT.1788, ITU-R BT.500-11, ITU-R BT.500-13, ITU-T P.913, as detailed in Table III-1

A 47" LG LCD, full HD 3D monitor (1920 x 1080 pixels) and a 400cd/m² maximum brightness are used in the experiments.

The experiments involve maximum 2 subjects per session.

The subjects are seated at a distance D equal to the height of the screen multiplied by factor *F* and defined as the *Preferred Viewing Distance* PVD, see Table I-2.

The observation angle relative to the normal was kept lower than 30 degrees (although this value is not a must for the LCD monitors).

Table III-1 Viewing conditions for test-bed.

Rec. ITU-R BT.500-11	ARTEMIS
Ratio of luminance of inactive screen to peak luminance $\leq$ 0.02	
Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white: $\approx$ 0.01	
Display brightness and contrast: set up *via* PLUGE software	
The viewing distance and the screen sizes are to be selected in order to satisfy the Preferred Viewing Distance PVD, see Table I-2.	
Maximum observation angle relative to the normal (this number applies to CRT displays, whereas the appropriate numbers for other displays are under study): 30	not applicable
Ratio of luminance of background behind picture monitor to peak luminance of picture: $\approx$ 0.15	
Chromaticity of background: D65	
Other room illumination: low	

III.3. The evaluation methodology and the grading scales

III.3.1.1. Test method

After the study and the comparison between different assessment methods seen in the previous chapter, and following the ITU recommendations for evaluation content to be consumed in home environments, a SSCQE (Single Stimulus Continuous Quality Evaluation) method has been adopted.

According to the thesis objectives, this method is considered with two types of scales:

- continuous unlabeled scale, ranging from 0 to 100; such a scale is in concordance with [CAM07], [BEN08], [CHE12-01], [GAO13] and ITU-T P.913;
- discrete labeled scales:
    - with 5 levels: *Excellent*, *Good*, *Fair*, *Poor* and *Bad*; such a scale is in concordance with ITU-R BT.500-11/13 and with the most of the state-of-the-art studies (*cf*. Section II);
    - with 3 levels: *Good*, *Fair* and *Bad*; such a scale follows the ITU principles and the explicit suggestion in [GOL10-01] ITU-T P.913.

III.3.1.2. Assessment session

Training session

At the beginning of the first session, from 2 to 5 training presentations are introduced to stabilize the observers' opinion.

This is a crucial part of method of assessment, since subjects could misunderstand their task; during the training, the observer are explained about what they are going to see, what they have to evaluate and how they will express their opinion. Any question from the subjects is answered in a short and objective way.

If several evaluation sessions are required, only two training presentations are done at the beginning of the next session.

The data issued from these presentations are not taken into account when computing the MOS.

The evaluation session

The evaluation session lasts up to half an hour.

Each observer evaluates video excerpts, extracted from the video corpus (see Chapter III.4 below) and presented in a random order. During the testing session, each sequence is shown once or twice and a break between the presentations is ensured for scoring.

III.4. Content under evaluation

Following the general trend in the state-of-the-art studies, the content under evaluation is selected so as to cater for two particular needs of the thesis, namely to investigate the influence of the type of content and of it's a priori quality in the overall results. Consequently, both stereoscopic and 2D video content are considered, each of which at two quality levels (as a priori expressed by an objective quality metric). Note that in order to ensure realistic conditions, both the stereoscopic and the 2D video content corpora are organized under the framework of French national projects (as subsequently detailed).

III.4.1. Stereoscopic video corpus

The stereoscopic video content processed in the present thesis is produced under the framework of the 3DLive French national project, meant to create expertise in France for shooting and subsequent live TV transmission of 3D stereo contents. The 3DLive corpus sums-up 2 hours, 11 minutes and 24 seconds of stereoscopic video sequences (197000 stereoscopic pairs encoded at 25 frames per second), representing 10 minutes of a rugby match, 10 minutes of a dancing performance, 1 minute of a private gig of rock band "Skip the Use", one hour and 45 minute and 24 seconds of a volley-ball match and 5 minutes of a theater play "Les Fourberies de Scapin" by Molière. These sequences are full HD encoded (1920×1080 pixels), as illustrated in Figure III-1.

High-quality stereoscopic video corpus

From the 3DLive corpus, 16 sequences with individual durations between 40 sec and 80 sec, summing up about 20 minutes are randomly sampled.

Each of these sequences is subsequently watermarked[1] by considering 4 different watermarking methods (namely SS, binary QIM, 5-symbols QIM and IProtect) and, for each of them, 4 different insertion domains (one of the views and 3 different disparity maps, referred to by NTSS, FS-MPEG and 3DV-NTSS). The watermarking parameters are set so as to result into values 35dB < PSNR < 40dB; it is also a posteriori verified that the SSIM (Structural SIMilarity) values are larger than 0.98 while the NCC (Normalized Cross Correlation) values are larger than 0.98.

Thus, a total of 256 (16 original content types x 16 distortion configurations) sequences are obtained. Among these sequences, 1 different type of content for each distortion configuration is randomly selected, thus resulting in 16 sequences summing 20 minutes (each original content type is present once and none of them is distorted the same way). Finally, these 16 sequences are shuffled (randomly reordered) and presented to the observer.

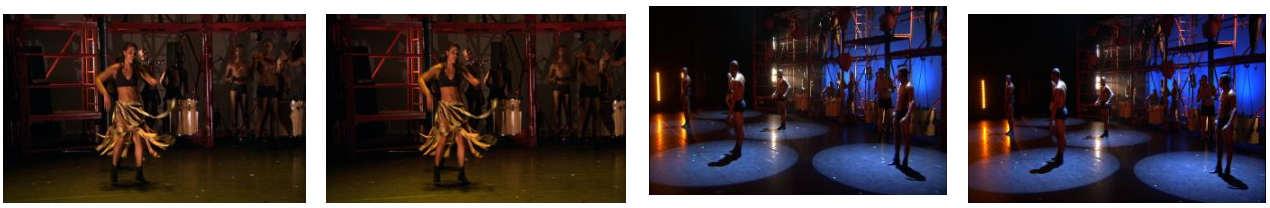
[1] Note that in the present thesis, the watermarking details are irrelevant; yet, more details are available in [HAS17-01], [CHA13-01], [CHA13-02].

All these random sampling and shuffling steps are considered as a possible way to eliminate the influence of the original content and of the distortion types in the final evaluation and to allow the investigation of the grading scales and semantic labels themselves. This principle is kept for the other types of corpora.

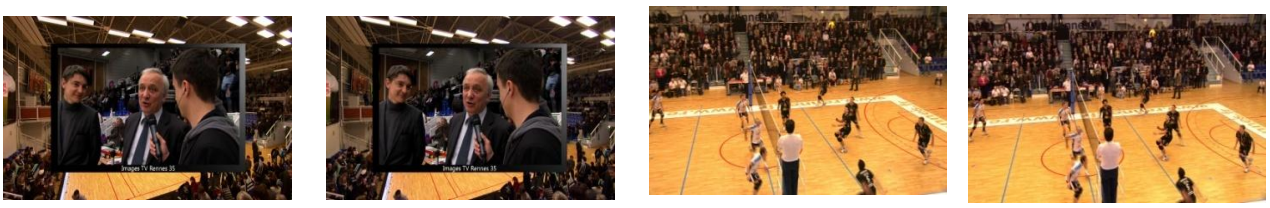
Low-quality stereoscopic video corpus

In order to obtain the low-quality stereoscopic video content, the high quality stereoscopic video corpus is compressed (while keeping the frame resolution and rate constant) so as to obtain 25dB<PSNR<30dB. It should be noticed that the low-quality corpus is downgraded 25db – 30dB with respect to the high quality corpus which is, in its turn, downgraded 35db - 40db with respect to the original content. It is also a posteriori verified that the values corresponding to the SSIM and NCC (with respect to the high quality corpus) range between 0.97 – 0.98 and 0.95-0.97, respectively.

(a) Rugby match sequence (Rugby).



(b) Dancing performance sequence



(c) Volley match sequence

(d) A theater play "les Fourberies de Scapin" sequence

(e) Rock band concert sequence

Figure III-1: Left and right views sampled from the 3DLive corpus.

III.4.2. 2D video corpus

High-quality 2D video corpus

The high quality 2D video corpus corresponds to the left view from the high quality stereoscopic video corpus.

Low-quality 2D video corpus

The low-quality 2D video corpus is organized under the framework of the MEDIVALS (waterMarking et Embrouillage pour la DIffusion et les Echanges Vidéos et Audios Legalisés) French national project. The main MEDIEVALS project is to design and deploy an end-to-end security solution for the audio-visual content, ensuring the tracking of the delivered information and of its property rights.

The video content is encoded at 640x480 pixels, 25 fps. An MPEG-4 AVC encoder is considered, with the baseline profile and a 512 kbps rate.

The corpus has a total duration of 1h30 minutes and is composed of 4 types of professional TV content: news, documentary, movies and talk-shows.

In order to obtain the content to be presented to the observers, from each type of content, a sequence with a duration between 50 and 60 sec is randomly extracted. Then, each of this sequence is downgraded with 7 distortion configurations[2]. These 28 sequences, summing up to 20 minutes, are then shuffled prior to their presentation to the observer.

[2] The distortion modes are irrelevant for the present study; they are derived from a research study related to the compressed-stream watermarking saliency [AMM17-01], [AMM17-02]

III.5. Panel composition and size

A total of 640 human observers are involved in the experiments, as detailed in Table III-3. They are all non-expert viewers with marginal knowledge on the image quality. They speak professional English with a majority of English native speakers. The age distribution ranged from 20 to 37 with an average of 23. All the subjects were screened for fine and dynamic stereopsis, visual acuity using Snellen chart and color vision using the Ishihara test.

They were grouped in four main panels of 160 observers each, on panel for each type of content. In order to grant result repeatability, each main panel was split into three sub-panels, referred to as the *reference* (60 observers), *validation* (50 observers) and *cross-checking* (50 observers) panels. The subjects in both the *validation* and *cross-checking* panels are finally partitioned into two sets, scoring on 5 level grading scales and on 3 level grading scales, respectively.

Table III-2: Panel composition and size: a total of 640 observers, split in 4 sub-panels of 160 observers each

Type of panels	*Panel composition and size*
Panel 1: ***Stereoscopic video content – High Quality***	***Reference:*** 35 males / 25 females aged from 22 to 37 with average of 26. ***Cross-checking:*** ○ *5 Levels grading scale:* 14 males /11 females aged from 21 to 25 with average of 22. ○ *3 Levels grading scale:* 12 males /13 females aged from 20 to 25 with average of 22. ***Validation:*** ○ *5 Levels grading scale:* 14 males /11 females aged from 23 to 30 with average of 24. ○ *3 Levels grading scale:* 12 males /13 females aged from 21 to 24 with average of 22
Panel 2: ***Stereoscopic video content – Low Quality***	***Reference:*** 31 males and 29 females aged from 20 to 31 with average of 25 ***Cross-checking:*** ○ *5 Levels grading scale:* 15males /10 females aged from 21 to 25 with average of 22 ○ *3 Levels grading scale:* 12 males /13 females aged from 22 to 27 with average of 23 ***Validation:*** ○ *5 Levels grading scale:* 13 males /12 females aged from 21 to 25 with average of 22 ○ *3 Levels grading scale:* 15 males /10 females aged from 22 to 28 with average of 24
Panel 3: ***2D video content – High Quality***	***Reference:*** 37 males and 23 females aged from 19 to 34 with average of 27 ***Cross-checking:*** ○ *5 Levels grading scale:* 17 males /8 females aged from 22 to 25 with average of 23 ○ *3 Levels grading scale:* 13 males /12 females aged from 21 to 27 with average of 24 ***Validation:*** ○ *5 Levels grading scale:* 11 males /14 females aged from 23 to 26 with average of 24 ○ *3 Levels grading scale:* 15 males /10 females aged from 21 to 25 with average of 22
Panel 4: ***2D video content – Low Quality***	***Reference:*** 28 males and 32 females aged from 19 to 30 with average of 26 ***Cross-checking:*** ○ *5 Levels grading scale:* 16 males /9 females aged from 21 to 25 with average of 22 ○ *3 Levels grading scale:* 10 males /15 females aged from 21 to 31 with average of 26 ***Validation:*** ○ *5 Levels grading scale:* 14 males /11 females aged from 21 to 25 with average of 23 ○ *3 Levels grading scale:* 11 males /14 females aged from 20 to 25 with average of 22

III.6. Scoring and post-processing of the scores

III.6.1. Scoring application

Following the current state-of-the-art trend for translating the ITU recommendations into user-friendly, application oriented software implementations, an interface matched to the peculiarities of our investigation is developed.

The scoring application is Android-based and runs on compatible tablets and smartphones. It is connected to the server on which the video content to be evaluated is stored and prepared by the test administrator. Of course, the video itself is displayed on the TV screen, see Figure III-2.

The scoring application offers to the observer three successive interfaces, as illustrated in Figure III-3, Figure III-4 and Figure III-5.

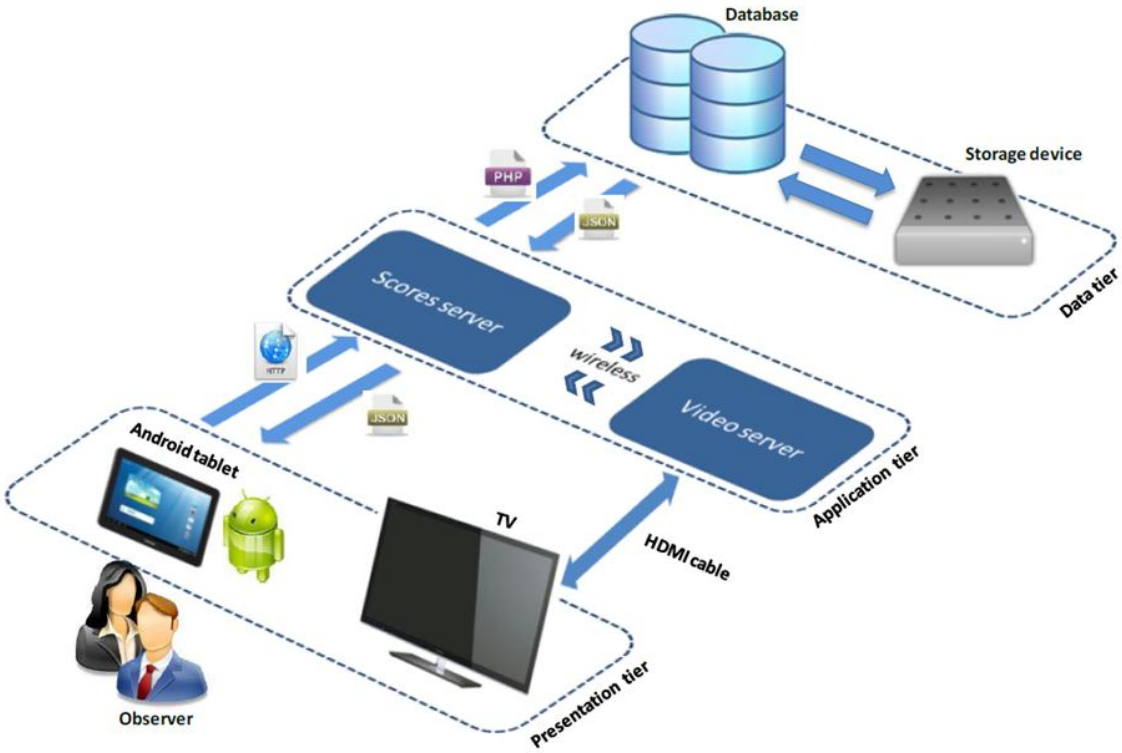


Figure III-2: Scoring synopsis

III.6.1.1. Authentification interface

When the observer launches the application, a login form is displayed, Figure III-3. The observer is required to enter his/her name, age and gender. The gender information is not compulsory (yet, all the

640 subjects in the panel filled-it in). On clicking the *Connection* button, the evaluation process starts and the observer is allowed to load and score videos, see Figure III-4.

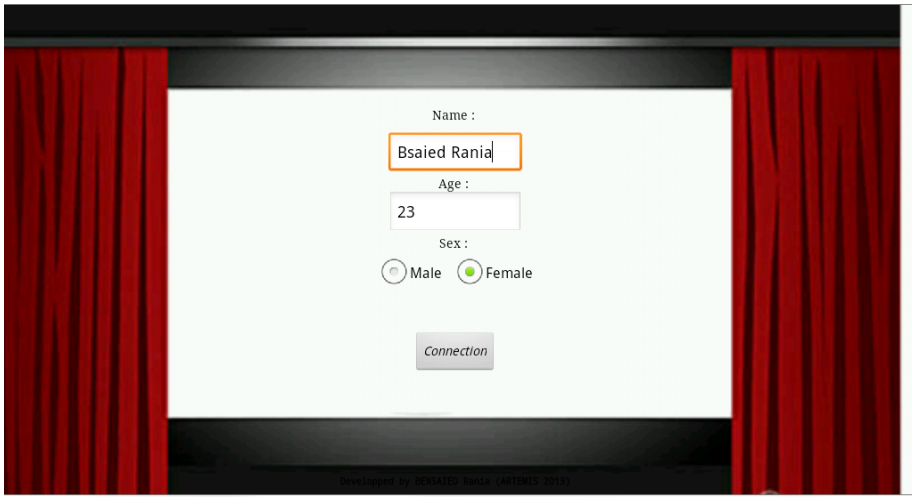


Figure III-3: Login interface.

III.6.1.2. Video control interface

After getting identified, the observer can start loading the videos to be assessed. He/she can ask to play the video and repeat it (the video replay is allowed only twice). Once the video sequence reaches its end, the observer can move to the scoring interface by clicking the "*Scoring*" button, see Figure III-5.

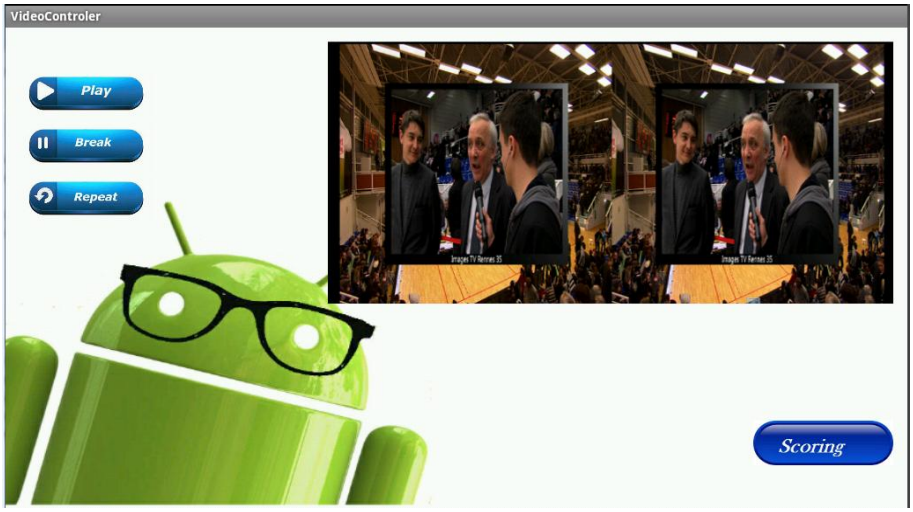


Figure III-4: Video control interface.

III.6.1.3. Scoring interface

When the observer clicks on the "*Scoring*" button from the previous interface, a scoring interface is displayed, Figure III-5.

2D video content scoring

The 2D video content (be it high or low quality) is scored based on the interface presented in Figure III-5.a, Figure III-5.b and Figure III-5.c.

The evaluation concerns the perceived *Visual quality*.

The continuous scale ranges between 0 and 100, with numbered marks (with a 10 precision). Note that the dynamic range is chosen according to both the specifications ITU-T P.913 and to several state-of-the art studies [CAM07], [BEN08], [CHE12-01], [GAO13].

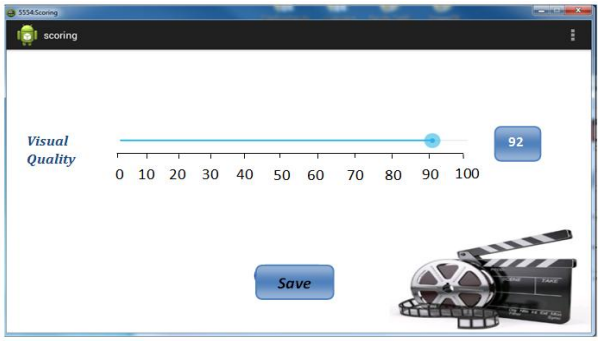
The 5 levels discrete scale is labeled *Excellent*, *Good*, *Fair*, *Poor* and *Bad*; such a scale is in concordance with ITU-R BT.500-11/13 and with the most of the state-of-the-art studies (*cf*. Section II).

Concerning the 3 levels grading scale, the following labels are considered: *Good*, *Fair* and *Bad*; this choice corresponds to the ITU-T P.913 and to the explicit suggestion in [GOL10-01],

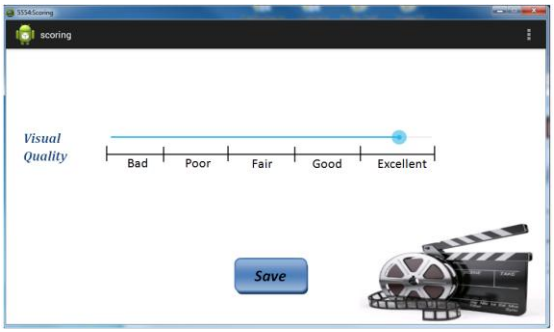
Stereoscopic video content scoring

The stereoscopic video content (both high and low quality) interface is paired design with the 2D evaluation interface Figure III-6.d, Figure III-6.e and Figure III-6.f.

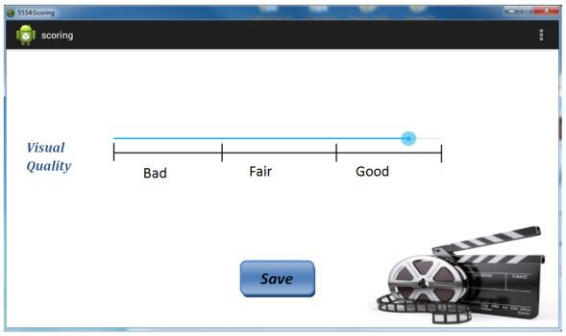
This time, according to ITU-R BT.2021, the evaluation concerns three related yet somewhat complementary criteria, namely *Image quality, Depth perception* and *Visual comfort*; such a choice is also compatible with the state of the art studies [CHE12-01]



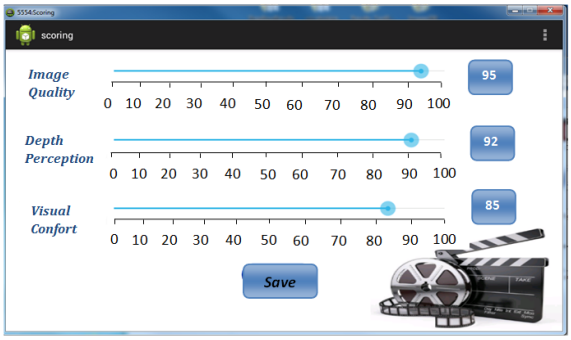
(a)



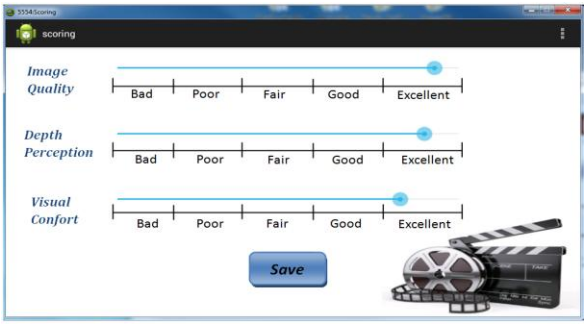
(b)



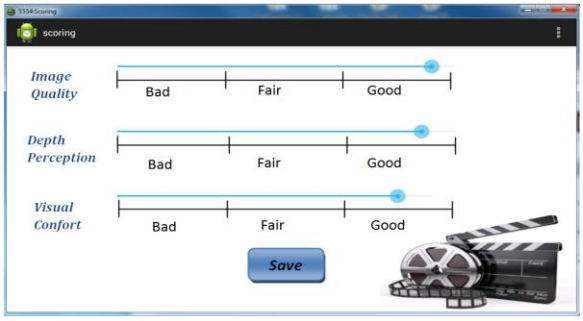
(c)



(d)



(e)



(f)

Figure III-5: Scoring interfaces: (a) Continuous scale scoring interface for 2D video content; (b) 5 levels discrete scale scoring interface for 2D video content; (c) 3 levels discrete scale scoring interface for 2D video content; (d) Continuous scale scoring interface for stereoscopic video content; (e) 5 levels discrete scale scoring interface for stereoscopic video content; (f) 3 levels discrete scale scoring interface for stereoscopic video content.

III.6.2. Post-processing of the scores

Outliers' detection

For each assessed video sequence, the scores distribution across observers was tested by calculating the distribution kurtosis coefficient, according to BT.500-11/13.

Table III-3: Outliers' detection results

Type of panels	*Outliers detection*
Panel 1: ***Stereoscopic video content*** ***High Quality***	***Reference:*** 8 outliers ***Cross-checking:*** 0 outliers ***Validation:*** 1 outliers
Panel 2: ***Stereoscopic video content*** ***Low Quality***	***Reference:*** 3 outliers ***Cross-checking:*** 0 outliers ***Validation:*** 0 outliers
Panel 3: ***2D video content High Quality***	***Reference:*** 3 outliers ***Cross-checking:*** 0 outliers ***Validation:*** 0 outliers
Panel 4: ***2D video content Low Quality***	***Reference:*** 2 outliers ***Cross-checking:*** 0 outliers ***Validation:*** 0 outliers

MOS Computation

Once the outliers eliminated, the mean opinion score *MOS* is computed according to (I-1).

The *MOS* is expected to be presented alongside with its 95% confidence lower and upper limits $MOS - \delta$ and $MOS + \delta$ respectively, where δ is the 95% estimation error computed according to (I-2).

III.7. Conclusion

The test-bed designed and deployed in the thesis follows two complementary criteria: the compliance with the most intensively used ITU specification and the possibility of ensuring reliability and relevance for the targeted objectives. It has the main particularity that it is not oriented towards content quality assessment but towards the investigation of the evaluation procedure itself.

The viewing conditions are set according to ITU-R BT.500-11/13.

The SSCQE evaluation method is considered, with three types of scales:

- a continuous scale ranges between 0 and 100, with numbered marks (with a 10 precision); this dynamic range is chosen according to both the ITU specifications and to several state-of-the art studies ITU-T P.913 and [CAM07], [BEN08], [CHE12-01], [GAO13] ;
- a 5 levels discrete scale labeled *Excellent*, *Good*, *Fair*, *Poor* and *Bad*; such a scale is in concordance with ITU-R BT.500-11/13 and with the most of the state-of-the-art studies (*cf.* Section II);
- a 3 levels grading scale labeled: *Good*, *Fair* and *Bad*; this choice corresponds to the ITU-T P.913 and the explicit suggestion in [GOL10-01].

The content under evaluation is selected so as to ensure realistic evaluation conditions (according to two French national R&D projects) and should allow precision and reliability in the investigation of the quality evaluation procedure. In this respect, 4 corpora of about 20 minutes each are organized. They represent 2D and stereoscopic video content, at both high and low quality (as a priori expressed by a subjective quality metric).

The panel is composed of 640 human observers who were grouped in four main panels of 160 observers each, one panel for each type of content. In order to grant result repeatability, each main panel was split into three sub-panels, referred to as the *reference* (60 observers), *validation* (50 observers) and *cross-checking* (50 observers) panels. The subjects in both the *validation* and *cross-checking* panels are finally partitioned into two sets, scoring on 5 level grading scales and on 3 level grading scales, respectively.

The scoring is achieved by an Android application running on compatible tablets and smartphones.

The post-processing starts by eliminating the outliers then follows by computing the mean opinion score *MOS* alongside with its confidence limits.

IV. Bridging continuous and discrete unlabeled quality scale evaluation

The present chapter deals with the main challenge of the thesis: it theoretically bridges the continuous and the discrete scale evaluation procedures and investigates whether the number of the classes on the discrete scales is a criterion meaningful in the results interpretations or just a parameter.

In this respect, the non-linear relation between the probability density functions modeling the scores assigned by the observers on continuous and discrete unlabeled scales is established. The instantiations of this formula for two cases of practical relevance, namely the case in which the scores according on a continuous scales are Gaussian distributed and the case in which these scores are not Gaussian (but estimated via a Gaussian mixture) are also presented. In the two cases, it is brought to light that the MOS and the related confidence limits solely depend on the average and variances of the continuous scale models and of the q number of quality levels on the discrete scale. The theoretical results are illustrated through experiments corresponding to the scores assigned by the 4 reference panels (one panel for each type of content – see Chapter III).

The impact of the number of human observers in the precision of the quality evaluation is subsequently assessed. It is demonstrated that for three N values ranging from 15 to 50, the relative errors in MOS estimation are constant with both N and q (differences lower than 0.04 and 0.02 being obtained for Gaussian and non-Gaussian cases, respectively).

Hence, in order to converge to a unique answer to the controversial issue related to the number of quality levels on a stereoscopic video grading scale, it is suggested to perform the evaluation on a continuous grading scale, with no semantic labels associated to the scores and to subsequently map these values on the discrete scales.

IV.1. Continuous *vs*. discrete unlabeled scale evaluation

Be there a subjective quality evaluation experiment with X the r.v. (random variable) theoretically modeling the observer's inner appreciation about the content under evaluation. That is, X expresses not only the quality of the content under evaluation but also a large variety of factors related to the observer, but independent with respect to the evaluation procedure itself.

Let assume that X is continuously distributed in the interval $[0; M]$, according to a probability density function (*pdf*) $p_X(x)$. We denote by *MOS* and σ the X mean value and standard deviation, respectively:

$$\int_0^M p_X(x)dx = 1\,,\ MOS = E(X) = \int_0^M x p_X(x)dx\,,\ \sigma = \sqrt{Var(X)} = \sqrt{\int_0^M x^2 p_X(x)dx - MOS^2} \tag{IV-1}$$

where $E(.)$ denotes the expectation (average) of a *r.v* and *Var()* the variance of a *r.v.*

Assume now the case in which, during the evaluation procedure, a discrete grading scale with q quality levels is imposed to the observers. From the statistical point of view, this means that the observers' scores would be distributed according to a new r.v. Y. Assuming that the constraint of evaluating on q quality levels has no psycho-cognitive impact in the observer's scores, the values taken by Y can be obtained from the values taken by X trough a non-linear mapping function $f(.)$, as illustrated in Figure IV-1:

$$y = f(x) = \begin{cases} 0, & x < 0 \\ i, & (i-1)M/q \le x < iM/q \\ 0, & x \ge M \end{cases}, \text{ where } i \in \{1,2,\ldots,q\} \tag{IV-2}$$

By applying basic non-linear random variable filtering properties, *pdf* of Y denoted by $p_y(y)$ can be expressed as:

$$p_Y(y) = \sum_{i=1}^{q} \delta(y-i) \int_{(i-1)M/q}^{iM/q} p_X(x)dx \tag{IV-3}$$

where $\delta(i)$ denotes the Dirac's Delta function, see Figure IV-1.

The mean value of Y, denoted by MOS_q represents the mean opinion score corresponding to the evaluation on a of q quality level grade scale; it can be computed from (IV-3), by considering the Dirac's Delta function properties:

$$MOS_q = \int_{-\infty}^{\infty} y p_Y(y)dy = \int_{-\infty}^{\infty} y \sum_{i=1}^{q} \int_{(i-1)M/q}^{iM/q} p_X(x)dx\,\delta(y-i)dy = \sum_{i=1}^{q} i \int_{(i-1)M/q}^{iM/q} p_X(x)dx\,. \tag{IV-4}$$

The standard deviation of Y, denoted by σ_q, can also be computed:

$$\sigma_q = \sqrt{\sum_{i=1}^{q} i^2 \int_{(i-1)M/q}^{iM/q} p_X(x)dx - MOS_q^2} \qquad \text{(IV-5)}$$

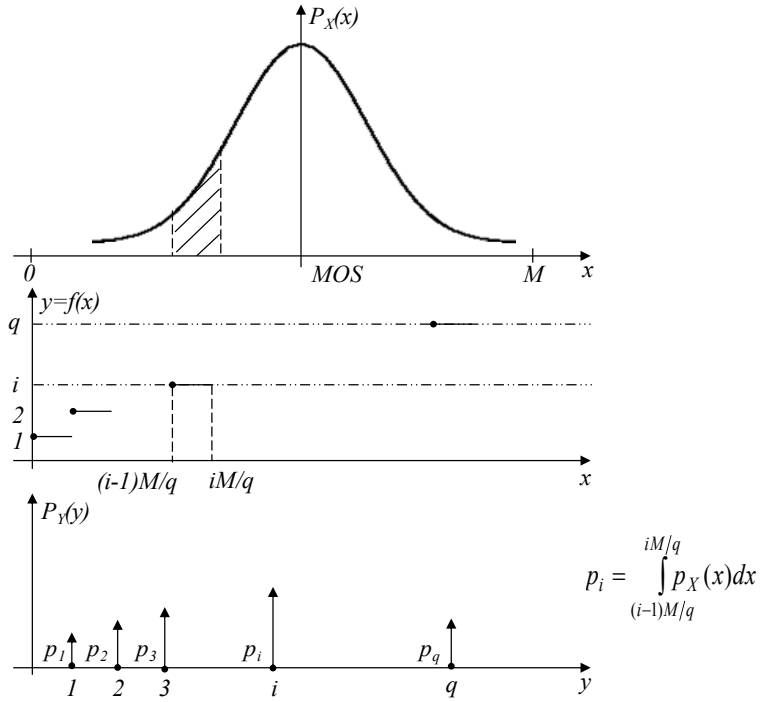


Figure IV-1: Continuous to discrete scale mapping for the subjective evaluation scores.

To conclude with, equation (IV-3) demonstrates that the *r.v. Y* is completely determined by the r.v. *X* and *q*, irrespective from the *pdf* of *X*.

From the quality evaluation point of view, this means that the MOS and σ corresponding to a continuous scale evaluation would allow to a posteriori obtain MOS_q on any q levels grading scale and to also have information about its statistical precision, by computing the 95% confidence limits according to (I-2), where σ is given by (IV-5).

Note that an equation (IV-3) is independent with respect to the law of *X*. Two cases of relevance in quality evaluation are subsequently investigated. First, according to the common trend [SIM09], [SES10], [WIN09], *X* will be considered as a Gaussian law distribution, see Section IV.2. Secondly, as doubts about the possibility of modeling the scores by Gaussian law also arise (ITU-R BT.1788, ITU-R BT.500-11, ITU-R BT.500-13), the *X* law will be represented as a mixture of Gaussian laws, Section IV.3.

IV.2. Gaussian assumption

IV.2.1. *MOS* and confidence limits computation for *q* grading scales

Assuming now a Gaussian behavior for the *r.v.* X, equations (IV-3), (IV-4) and (IV-5) can be expressed in an explicit form:

$$p_Y(y) = \sum_{i=1}^{q} \left[erfc(\frac{(i-1)M/q}{\sqrt{2}\sigma}) - erfc(\frac{iM/q}{\sqrt{2}\sigma}) \right] \delta(y-i) \quad \text{(IV-6)}$$

$$MOS_q = \sum_{i=1}^{q} i \cdot \left[erfc(\frac{(i-1)M/q}{\sqrt{2}\sigma}) - erfc(\frac{iM/q}{\sqrt{2}\sigma}) \right] \quad \text{(IV-7)}$$

$$\sigma_q = \sqrt{\sum_{i=1}^{q} i^2 \cdot \left[erfc(\frac{(i-1)M/q}{\sqrt{2}\sigma}) - erfc(\frac{iM/q}{\sqrt{2}\sigma}) \right] - MOS_q^2} \quad \text{(IV-8)}$$

where $erfc(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} \exp(-t^2) dt$

It can be concluded that if X is Gaussian distributed, the mean value and standard deviation of Y solely depend on the mean value and standard deviation of X, *cf*. (IV-7) and (IV-8) and on the *q* number of levels on the grading scale.

From the quality evaluation point of view, this means that the MOS and σ corresponding to a continuous scale evaluation would allow to a posteriori obtain MOS_q on any q levels grading scale and to also have information about its statistical precision; this property will be investigated in the following section.

IV.2.2. Illustration of continuous to discrete *MOS* computation

In order to illustrate the way in which the continuous scale parameters can be a posteriori mapped for computing the discrete, unlabeled *q* levels grading scales, the experiments are structured at two levels.

IV.2.2.1. Continuous scale evaluation

The following experiment is carried out 4 times, once for each type of content described in Section III.4 (2D and stereoscopic video, high and low quality). For each type of content, only the scores assigned by the 60 observers using a continuous scale (between 1 and *M*) are considered.

For a given type of content, the experiments start by estimating the mean value and the standard deviation (un-biased estimators) of the observers' scores. These estimated values are considered as the theoretical (unknown) parameters MOS and σ of the r.v. X, see (IV-9):

$$MOS = \frac{1}{N}\sum_{j=1}^{N} x_j \ , \ \sigma = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(x_j - MOS)^2} \qquad \text{(IV-9)}$$

We then compute the MOS_q according to (IV-7), for *q*=2, 3,…,9 and their confidence limits according to (I-2) and (IV-8). These values will be further considered as theoretical (reference) values for the mean opinion scores which would have been obtained on a discrete grading scale with *q* semantic levels.

IV.2.2.2. Discrete unlabeled scale simulation

The evaluation on the unlabeled, discrete, grading scales are obtained by mapping the $[x_1, x_2, ..., x_N]$ data to their corresponding values on a would-be *q* levels grading scale, according to (IV-2), thus obtaining values for the r.v. Y: $[y_1 = f(x_1), y_2 = f(x_2), ..., y_N = f(y_{N)}]$.

Consequently, the estimated values $\overline{MOS}_q$ and the 95% confidence limits $(MOS_{q;low}; MOS_{q;up})$ can be computed as follows:

$$\overline{MOS}_q = \frac{1}{N}\sum_{j=1}^{N} y_j \ , \ \overline{\sigma}_q = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(y_j - \overline{MOS}_q)^2}$$
$$MOS_{q;low} = \overline{MOS}_q - 1.96\frac{\overline{\sigma}_q}{\sqrt{N}} \ , \ MOS_{q;up} = \overline{MOS}_q + 1.96\frac{\overline{\sigma}_q}{\sqrt{N}} \qquad \text{(IV-10)}$$

IV.2.2.3. Illustrations

Now, a comparison can be achieved between the theoretical values MOS_q, see Chapter IV.2.2.1 and their confidence limits $(MOS_{q;low}; MOS_{q;up})$ in Chapter IV.2.2.2. The results corresponding to each of the four types of content, to *q*=2, 3,…,9 and to *N*=50, 30 and 15 are detailed in Appendix A. In the sequel of the chapter, only an illustrative selection is considered, as follows:

- Figure IV-2 corresponds to the score assigned for the *Image quality* of the high-quality stereoscopic video content; three subplots, denoted by (a), (b) and (c) correspond to *q* = 3, *q* = 5 and *q* = 9, respectively. For each subplot, three experimental cases are presented (from left to

right), corresponding to *N* = 50, *N* = 30 and *N* = 15. For each *N* value, 16 types of sequences are presented, as detailed in Chapter III.4.1 (they correspond to watermarked content obtained by applying 4 different methods – SS, 2-QIM, 5-QIM and IProtect, in 4 different insertion domains – view-based, NTSS, FS-MPEG, and 3DV-NTSS).

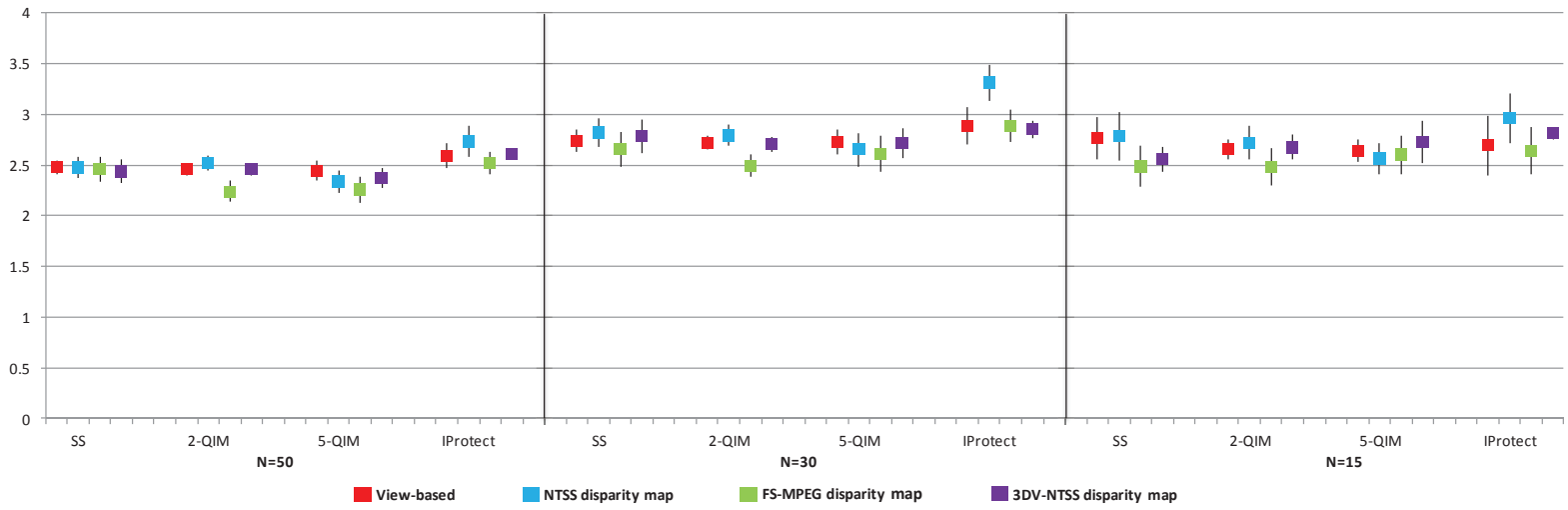
- Figure IV-3 is similar to Figure IV-2 but refers to low quality stereoscopic video content, as detailed in Chapter III.4.2.
- Figure IV-4 is similar to Figure IV-2 but refers to high-quality 2D video, as detailed in Chapter III.4.3; hence, this time, the observers scores the *Visual quality*.
- Figure IV-5 refers to low quality 2D video, as detailed in Chapter III.4.4. Hence, here again the *Visual quality* is scored but this time 28 types of sequences are scored.

IV.2.2.4. Conclusion

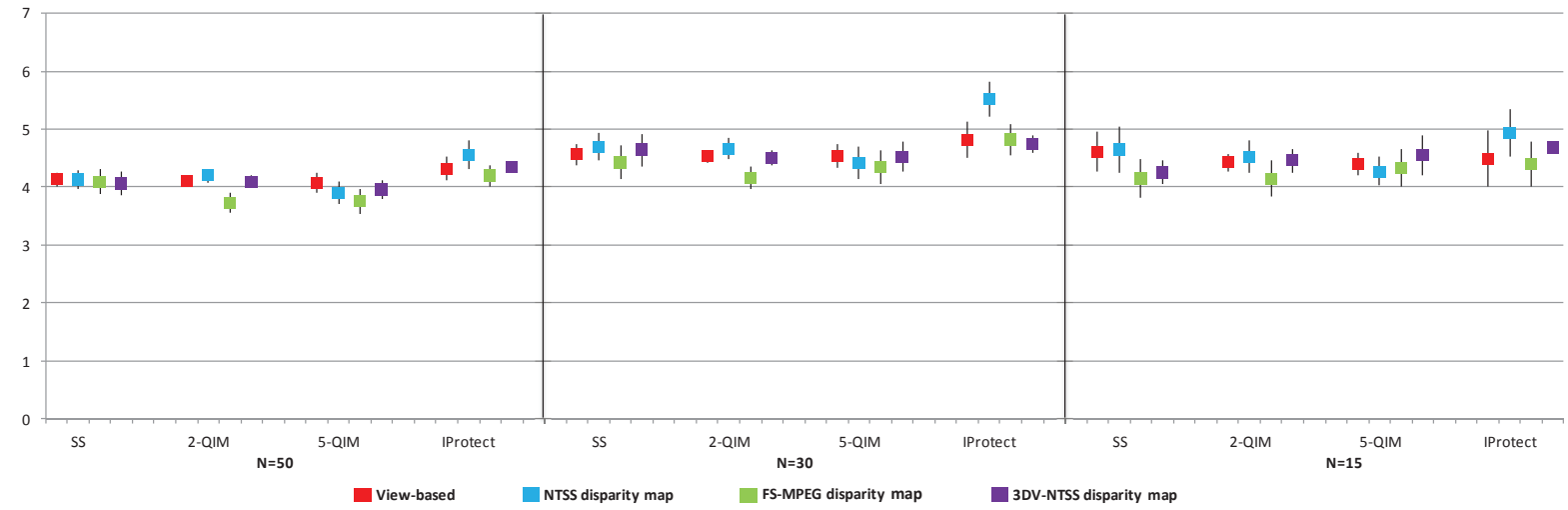
This section instantiates the theoretical formulae (IV-3), (IV-4) and (IV-5) for the case in which the scores assigned by the observers follows a Gaussian law. It is thus demonstrated that the *MOS* and the confidence limits corresponding to a would-be discrete, unlabeled scale with *q* levels can be theoretically computed starting from the *MOS* and the standard deviation of the scores assigned on a continuous grading scale.

The illustrations strengthen these theoretical results: each and every time (*i.e.* for each investigated criterion, each watermarking method, each insertion domain, each q and N values), the theoretical values belong to the experimental confidence limits: $(MOS_{q;low};\ MOS_{q;up})$.

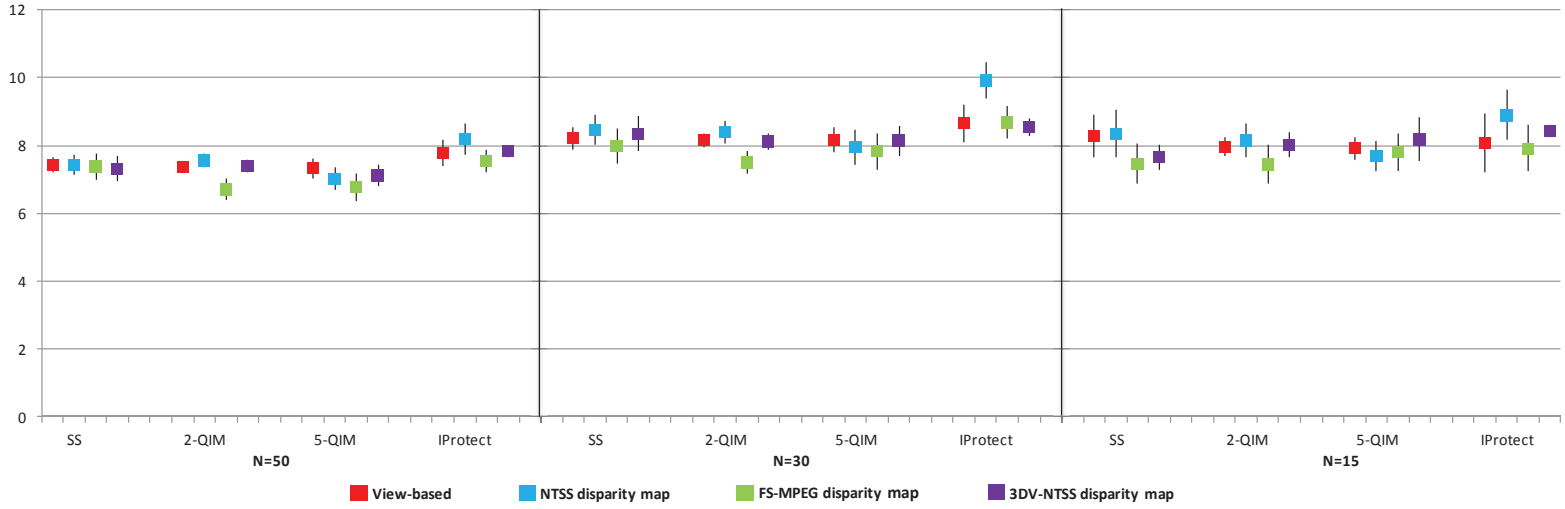
High quality stereoscopic video content



(a)



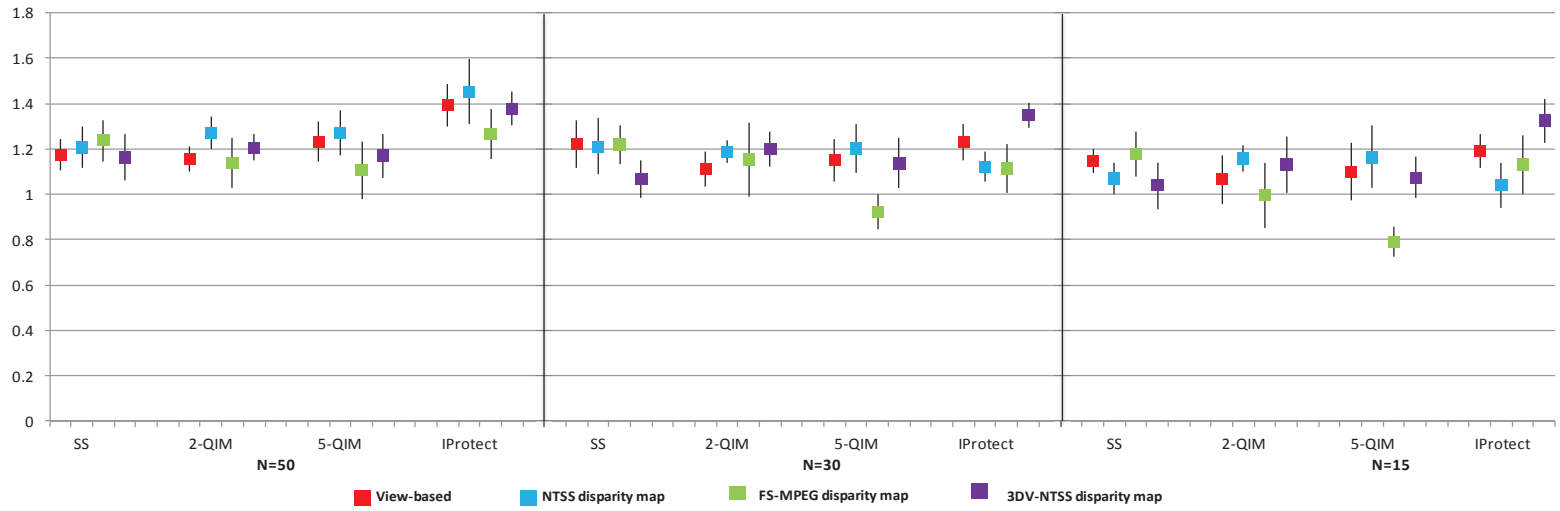
(b)



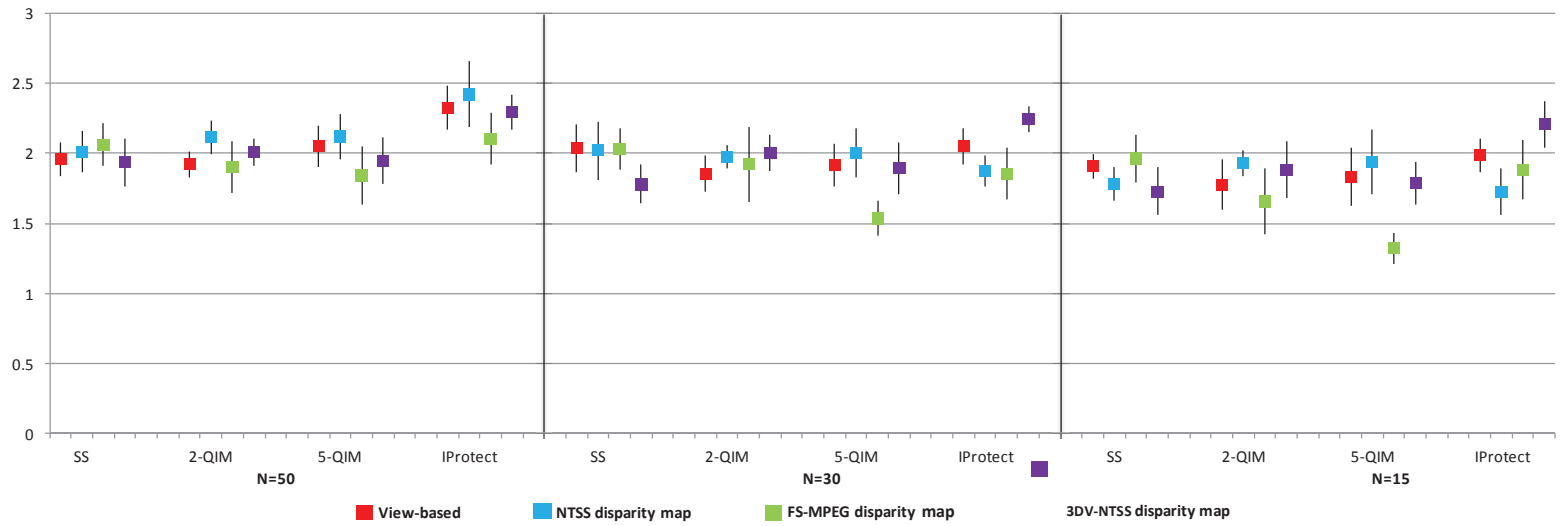
(c)

Figure IV-2: Subjective evaluations for high-quality stereoscopic video content, for grading scales of: (a) $q = 3$ quality levels; (b) $q = 5$ quality levels; (c) $q = 9$ quality levels; and for a number of observers N=50, N=30, N=15.

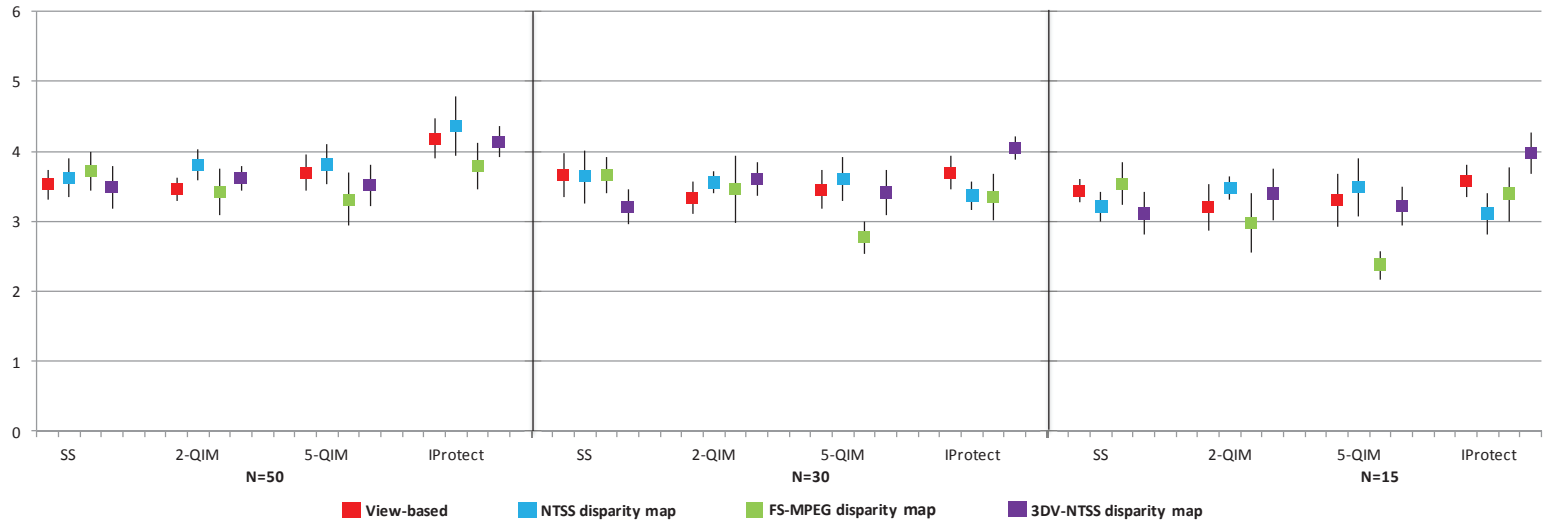
Low quality stereoscopic video content



(a)

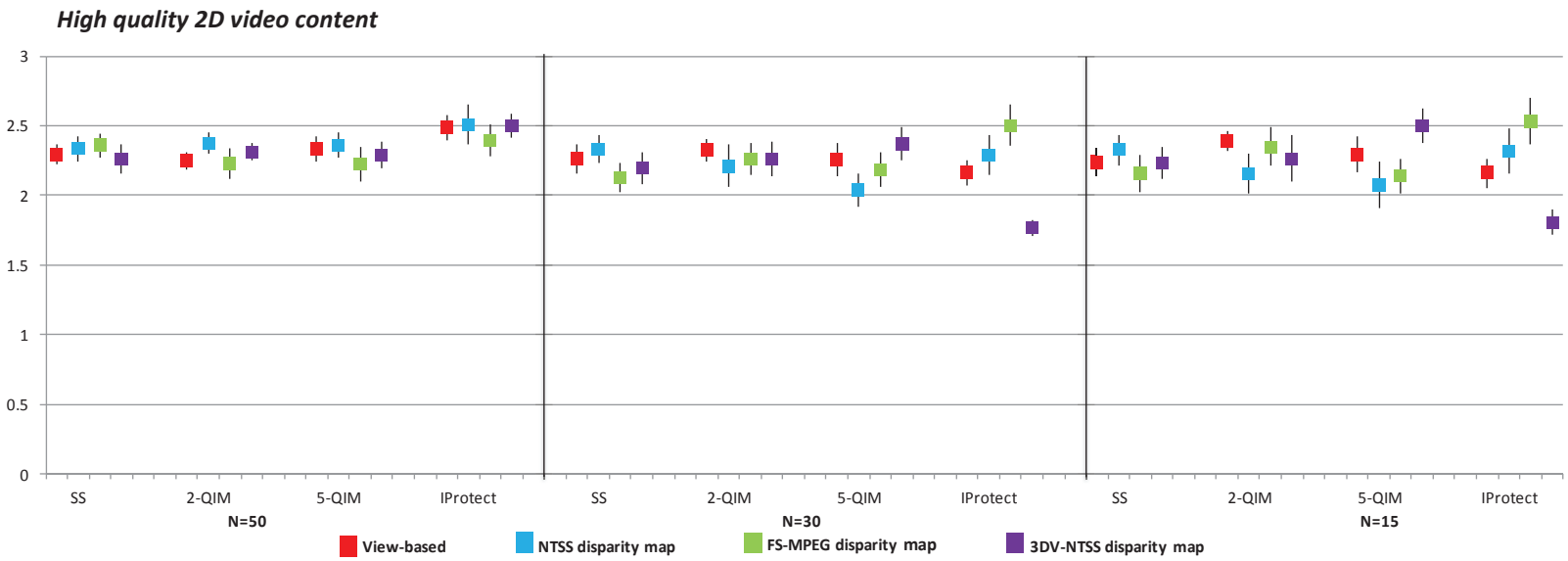


(b)

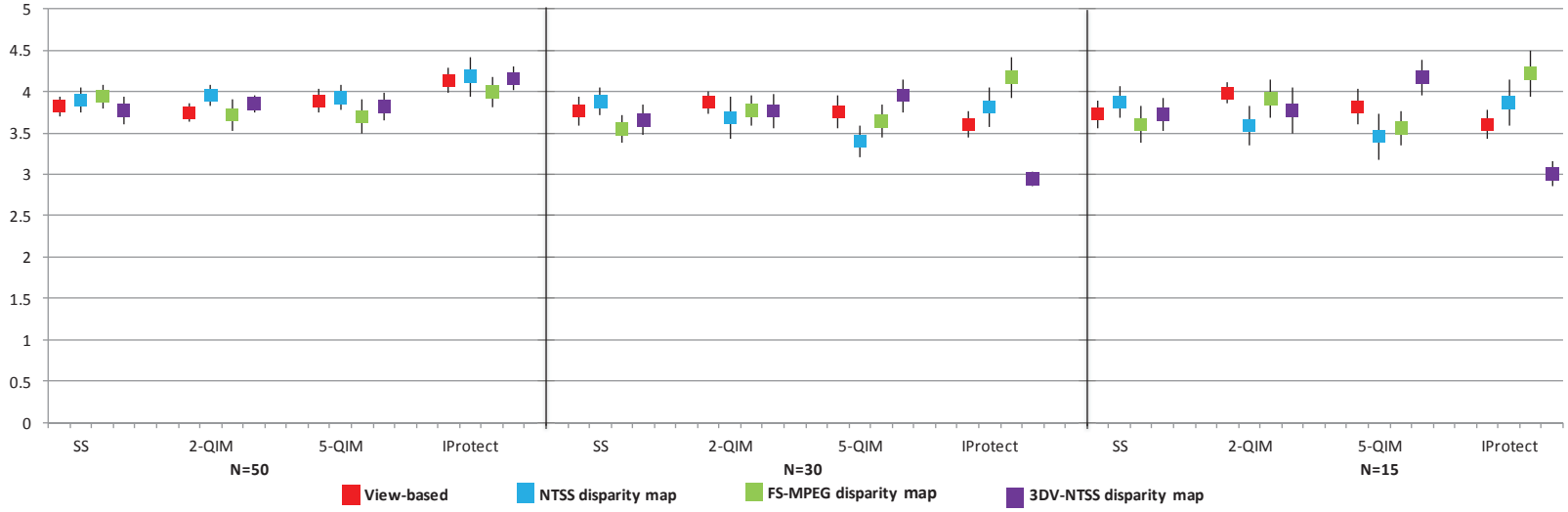


(c)

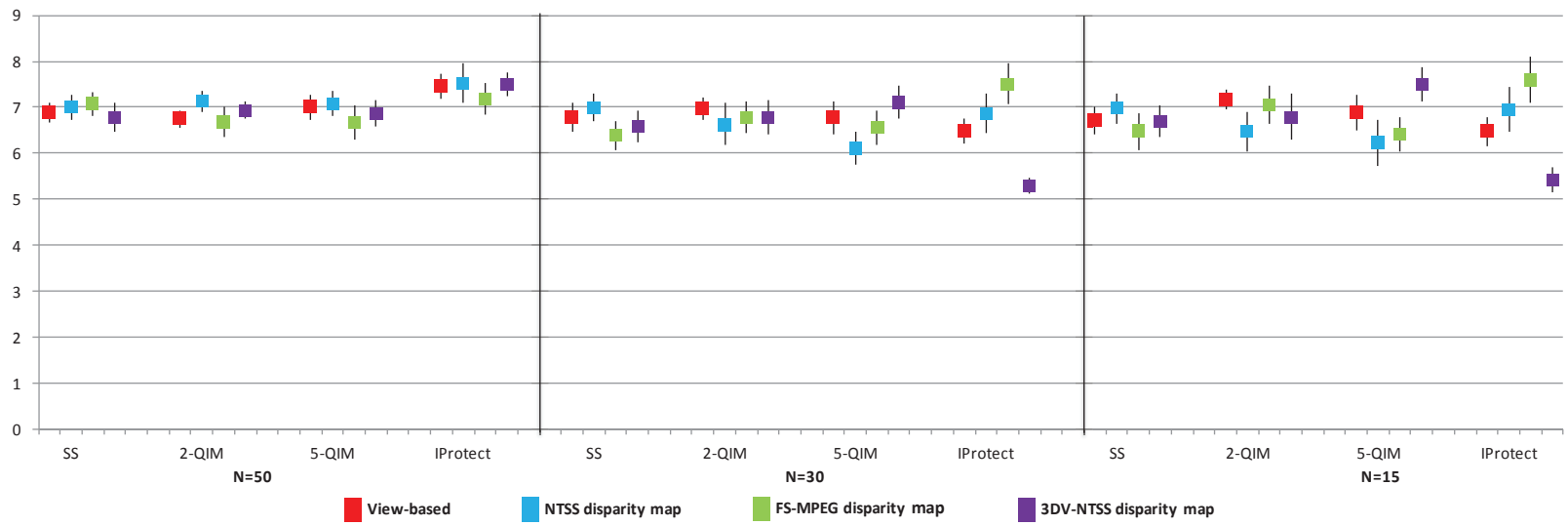
Figure IV-3: Subjective evaluations for low quality stereoscopic video content, for grading scales of: (a) $q = 3$ quality levels; (b) $q = 5$ quality levels; (c) $q = 9$ quality levels; and for a number of observers N=50, N=30, N=15.


High quality 2D video content
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(a)

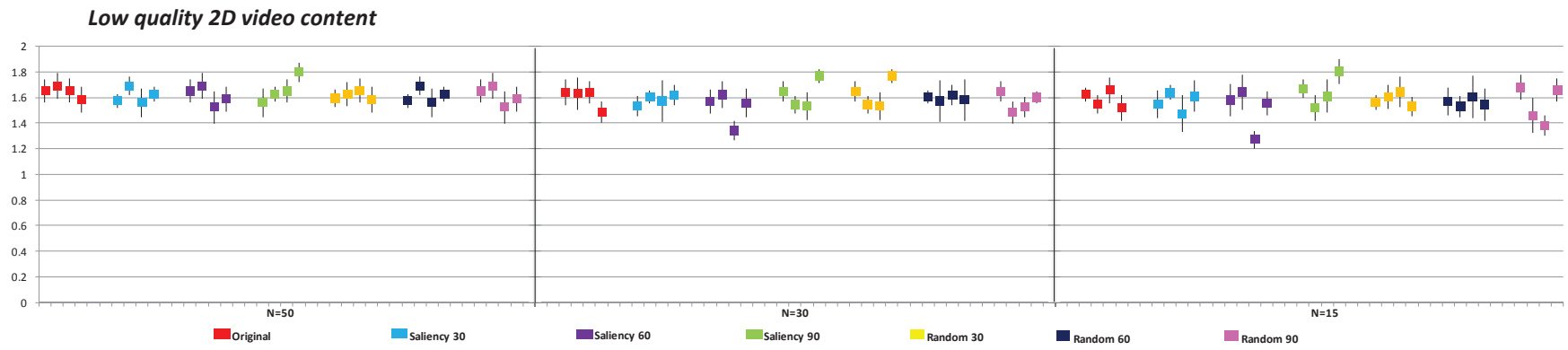

5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(b)

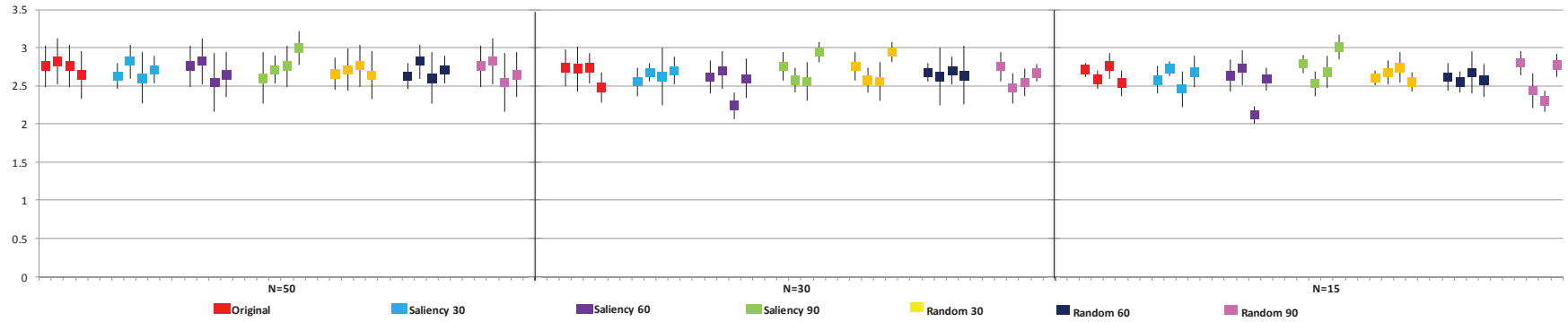


(c)

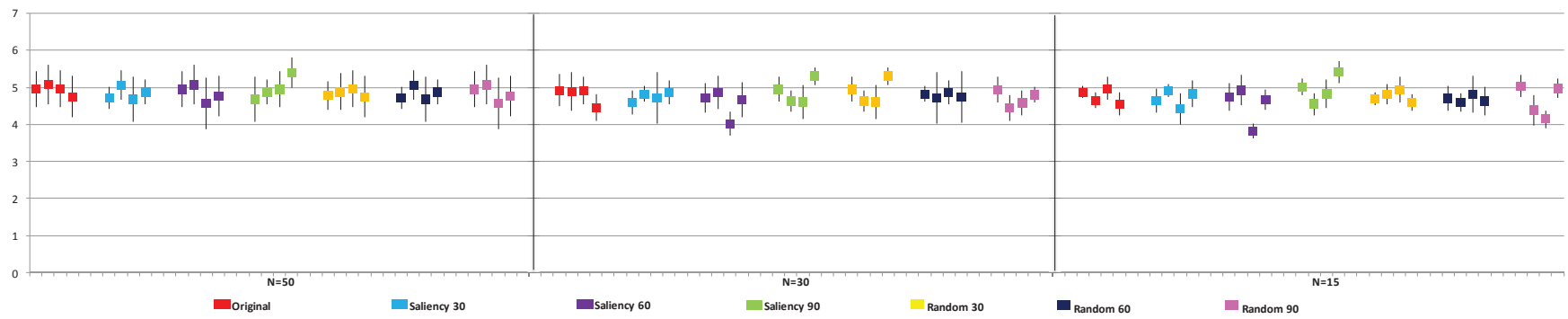
Figure IV-4: Subjective evaluations of high quality 2D video content, for grading scales of: (a) $q = 3$ quality levels; (b) $q = 5$ quality levels; (c) $q = 9$ quality levels; and for a number of observers N=50, N=30, N=15.


Low quality 2D video content
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
N=50
N=30
N=15
Original
Saliency 30
Saliency 60
Saliency 90
Random 30
Random 60
Random 90

(a)


3.5
3
2.5
2
1.5
1
0.5
0
N=50
N=30
N=15
Original
Saliency 30
Saliency 60
Saliency 90
Random 30
Random 60
Random 90

(b)



(c)

Figure IV-5: Subjective evaluations for low quality 2D video content for grading scales of: (a) $q = 3$ quality levels; (b) $q = 5$ quality levels; (c) $q = 9$ quality levels; and for a number of observers N=50, N=30, N=15.

IV.2.3. Investigation on the accuracy of the results

In order to assess the accuracy (precision) of the results, both the absolute and the relative errors in MOS_q computation will be investigated as functions of *N* and *q*. That is, we shall first investigate the influence of *q* for a fixed *N* value.; then, for a given *q* value, we shall investigate the influence of *N*.

For a given number of quality levels on the grading scale *q*, the absolute and relative errors in MOS_q computation with a 95% confidence level, denoted by ε_q and ε_q^r can be approximated by

$$\varepsilon_q = 1.96\frac{\sigma_q}{\sqrt{N}};\ \varepsilon_q^r = \frac{\varepsilon_q}{MOS_q} = 1.96\frac{\sigma_q}{\sqrt{N}}\frac{1}{MOS_q} \tag{IV-11}$$

where MOS_q and σ_q can be computed according to (IV-7) and (IV-8) from the MOS and σ (*i.e.* from the values estimated on the $M = 100$ levels grading scale).

The numerical results, corresponding to the experiments described in Chapter IV.2.2 are presented in Figures IV-6, IV-7, IV-8, IV-9, IV-10, IV-11, IV-12 and IV-13. This time, rather then presenting individual results corresponding to each individual content, average values for a type of content are computed:

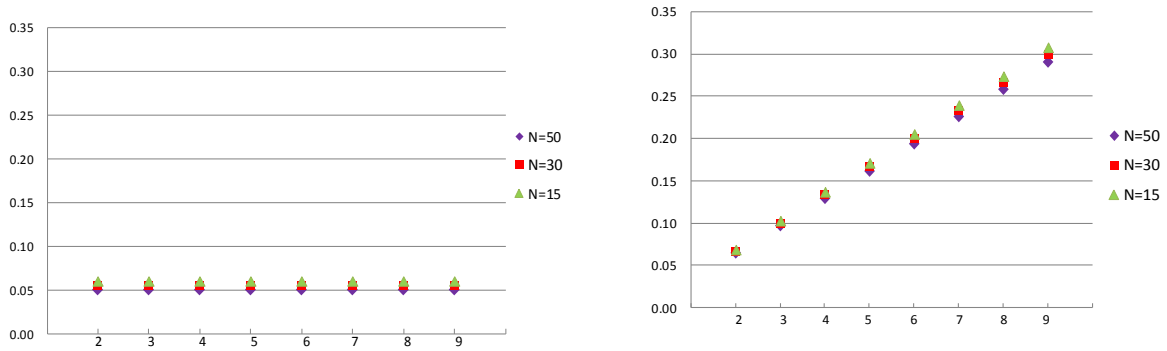
- Figure IV-6 illustrates the overall impact of *N* in the absolute and relative errors of high quality stereoscopic video when assessing the *Image quality* for *N*=50, 30 and 15 and *q*=2, 3,.., 9.
- Figure IV-7 illustrates the overall impact of *N* in the absolute and relative errors of high quality stereoscopic video when assessing the *Depth perception* for *N*=50, 30 and 15 and *q*=2, 3,.., 9.
- Figure IV-8 illustrates the overall impact of *N* in the absolute and relative errors of high quality stereoscopic video when assessing the *Visual comfort* for *N*=50, 30 and 15 and *q*=2, 3,.., 9.
- Figure IV-9 is similar to Figure IV-6 but refers to low quality stereoscopic video content, as detailed in Chapter III.4.2.
- Figure IV-10 is similar to Figure IV-7 but refers to low quality stereoscopic video content, as detailed in Chapter III.4.2.
- Figure IV-11 is similar to Figure IV-8 but refers to low quality stereoscopic video content, as detailed in Chapter III.4.2.
- Figure IV-12 is similar to Figure IV-6 but refers to high-quality 2D video, as detailed in Chapter III.4.3; hence, this time, the observers scores the *Visual quality*.
- Figure IV-13 is similar to Figure IV-12 but refers to low quality 2D video, as detailed in Chapter III.4.4.

As expected, for a given *N* value, the absolute error is an increasing function of *q* (see the left-side plots in Figures IV-6, IV-7, IV-8, IV-9, IV-10, IV-11, IV-12 and IV-13). However, it can be noticed that the relative errors are quite constant with *q* (see the right-side plots in the same 4 figures).
Actually, the experiments bring to light differences in the relative error lower than 0.04, for all the types of content, for *N*=50, 30 and 15 and *q*=2,3,..,9.

$$\max_{q_i,q_j}\left(\left|\varepsilon^r_{q_i,N} - \varepsilon^r_{q_j,N}\right|\right) \leq 0.04 \qquad \text{(IV-12)}$$

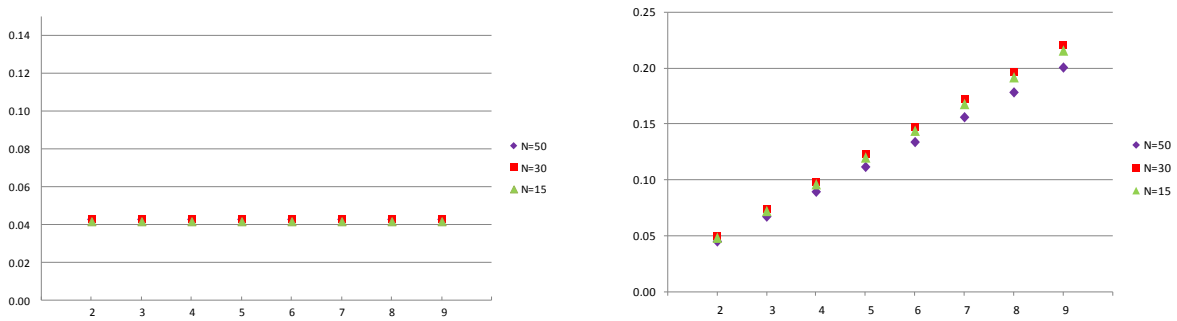
where $q_i, q_j = 1, 2, \ldots, 9$.

High quality stereoscopic video content



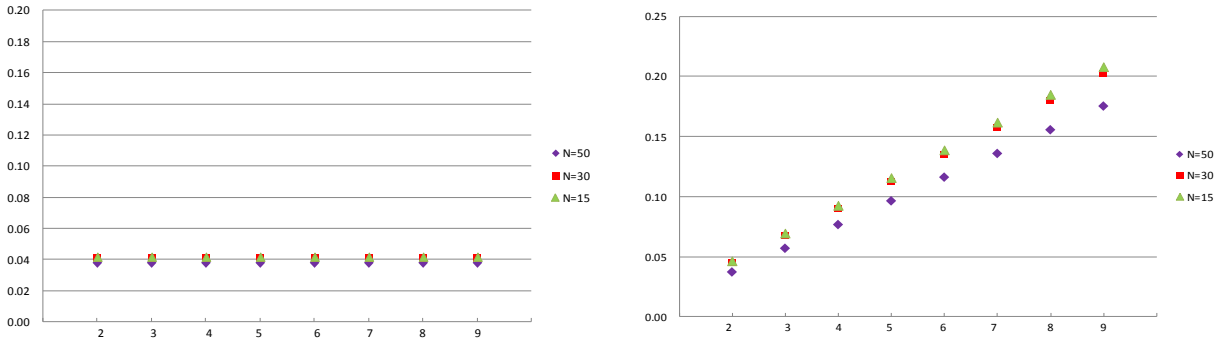
(a) (b)

Figure IV-6: Precision in the evaluation of high quality stereoscopic video content (Image quality); on the abscissa – the number of the quality levels q = 2,3, … ,9; on the ordinate - the 95% error value: (a) Relative error; (b) Absolute error.



(a) (b)

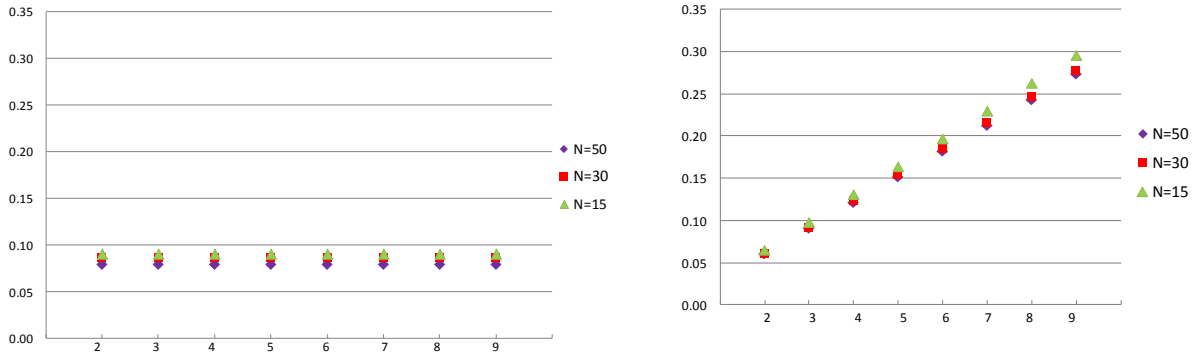
Figure IV-7: Precision in the evaluation of high quality stereoscopic video content (Depth perception); on the abscissa – the number of the quality levels $q = 2,3, \dots ,9$; on the ordinate - the 95% error value:(a) Relative error; (b) Absolute error.



(a) (b)

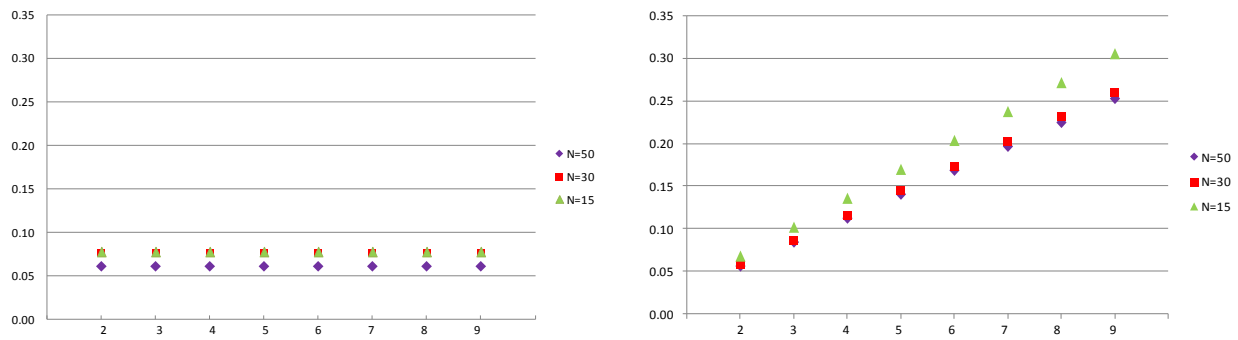
Figure IV-8: Precision in the evaluation of high quality stereoscopic video content (Visual comfort); on the abscissa – the number of the quality levels $q = 2,3, \dots ,9$; on the ordinate - the 95% error value: (a) Relative error; (b) Absolute error.

Low quality stereoscopic video content



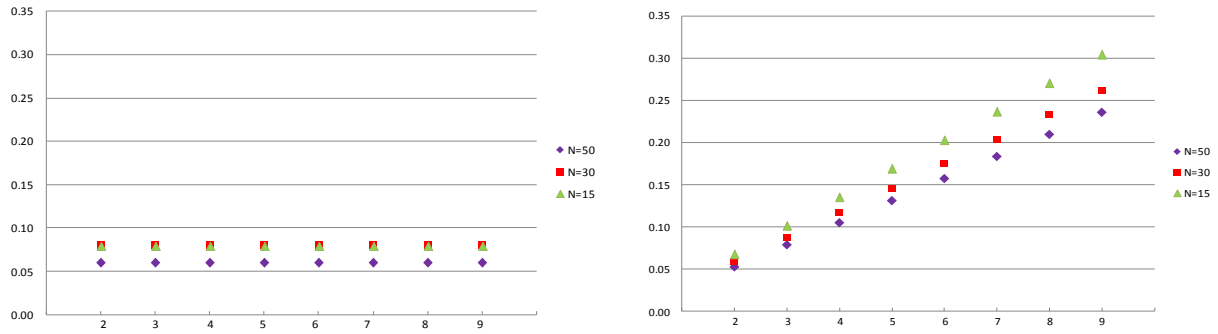
(a) (b)

Figure IV-9: Precision in the evaluation of low quality stereoscopic video content (Image quality); on the abscissa – the number of the quality levels $q = 2,3,\dots,9$; on the ordinate - the 95% error value: (a) Relative error; (b) Absolute error.



(a) (b)

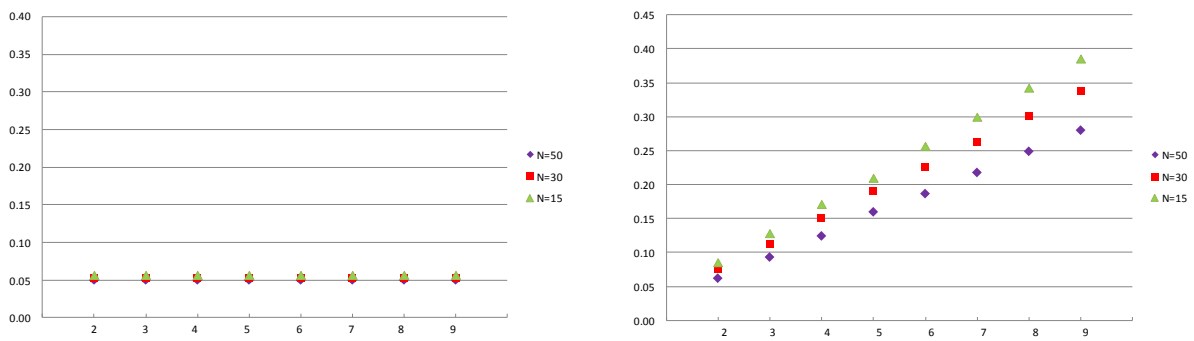
Figure IV-10: Precision in the evaluation of low quality stereoscopic video content (Depth Perception); on the abscissa – the number of the quality levels $q = 2,3,\dots,9$; on the ordinate - the 95% error value; (a) Relative error; (b) Absolute error



(a) (b)

Figure IV-11: Precision in the evaluation of low quality stereoscopic video content (Visual comfort); on the abscissa – the number of the quality levels $q = 2,3,\ldots,9$; on the ordinate - the 95% error value: (a) Relative error; (b) Absolute error.

High quality 2D video content



(a) (b)

Figure IV-12: Precision in the evaluation of high quality 2D video content on the abscissa – the number of the quality levels $q = 2,3,\ldots,9$; on the ordinate - the 95% error value; (a): Relative error; (b) Absolute error.

Low quality 2Dvideo content

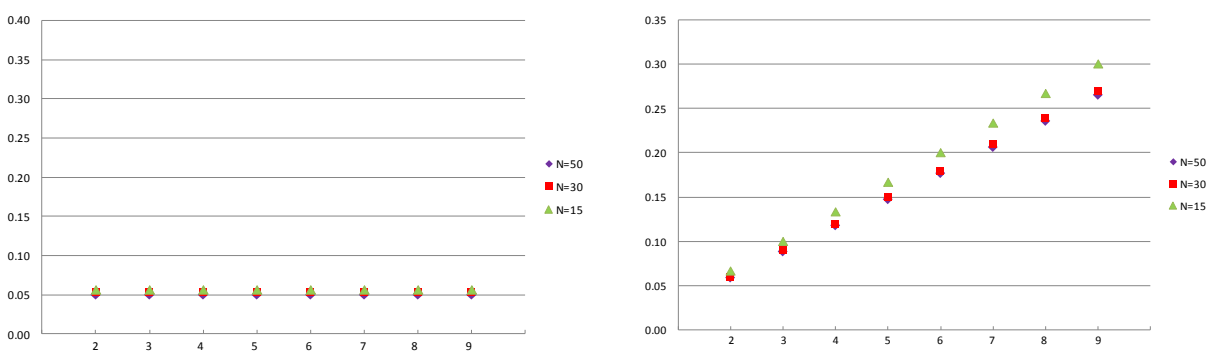


Figure IV-13: Precision in the evaluation of low quality 2D video content on the abscissa – the number of the quality levels $q = 2,3,\ldots,9$; on the ordinate - the 95% error value.: (a) Relative error; (b) Absolute error.

IV.3. Beyond Gaussian assumption

The present Chapter IV.3 goes on step further with respect to Chapter IV.2: it first investigates the validity of the Gaussian hypothesis for the scores assigned by the observers on continuous scale then instantiate the theoretical equations (IV-3), (IV-4) and (IV-5) for an arbitrarily (non-Gaussian) case.

IV.3.1. Gaussian mixture estimation

It is commonly assumed that X *r.v.* scores assigned on a continuous scale are Gaussian distributed [SIM09], [SES10], [WIN09]. However, no firm support is available in this respect and the ITU specification implicitly let the door open for scores that are not Gaussian distributed.

Consequently, the Gaussian behavior of the continuous scale scores is verified by applying the unilateral Chi-square goodness-on-fit tests (with 10 classes, estimated law parameters, and $\alpha = 0.05$).

Such tests are applied for each and every 140 investigated score sets (*cf*. details in Chapter III.4):

- 48 score sets for the high quality stereoscopic content, corresponding to the 16 sequences under test and to the 3 evaluation criteria (Image quality, Depth perception and Visual comfort);
- 48 score sets for low quality stereoscopic content, corresponding to the same cases as above;
- 16 score sets for the high quality 2D video content;
- 28 score sets for low quality 2D video content.

The results of the Chi-square tests demonstrate that the Gaussian hypothesis can be accepted only 15 times while the alternative (non-Gaussian) hypothesis is accepted 125 times. Note that these 15 exceptions are provided by all the four types of content: 3 exceptions for the high quality stereoscopic content, 4 exceptions for low quality stereoscopic content, 5 exceptions for high quality 2D video content and, finally, 3 exceptions for low quality 2D video content. These quantitative values show that the largest relative number of exceptions corresponds to high quality 2D video content, namely 5 out of 16.

Consequently, the study presented in Chapter IV.2 should also be extended with solution for the non-Gaussian models for the scores assigned by the observers. Specifically, the continuous scale scores probability density function $p_X(x)$ is to be estimated. In this respect, a mixture of $K=5$ Gaussian laws combined to an Expectation-Maximization (EM) algorithm [DEM77] is considered and result in a model of the type:

$$p_x(x) = \sum_{j=1}^{K=5} w_j \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x-MOS_j)^2}{2\sigma_j^2}} \qquad \text{(IV-13)}$$

where MOS_j and σ_j^2 are the mean values and the variances of the 5 laws composing the mixture, w_j stands for the underlying weights in the mixture while $j = 1,2, \dots K = 5$.

Of course, such a procedure was carried-out 140 times, once for each of the 140 investigated data sets.

IV.3.2. *MOS* and confidence limits computation for *q* grading scales

Assuming now the Gaussian mixture model (IV-13) for the *r.v.* X, equations (IV-3), (IV-4) and (IV-5) can be expressed in an explicit form:

$$MOS_q = \sum_{i=1}^{q} i \int_{\frac{(i-1)M}{q}}^{\frac{iM}{q}} p_x(x)dx \qquad \text{(IV-14)}$$

$$= \sum_{i=1}^{q} i \int_{(i-1)M/q}^{iM/q} \sum_{j=1}^{K=5} w_j \frac{1}{\sqrt{2\pi\sigma_j{}^2}} e^{-\frac{(x-MOS_j)^2}{2\sigma_j{}^2}} dx$$

$$= \sum_{i=1}^{q} i \sum_{j=1}^{K=5} w_j \int_{\frac{(i-1)M}{q}}^{\frac{iM}{q}} \frac{1}{\sqrt{2\pi\sigma_j{}^2}} e^{-\frac{(x-MOS_j)^2}{2\sigma_j{}^2}} dx$$

$$= \frac{1}{2} \sum_{i=1}^{q} i \sum_{j=1}^{K=5} w_j \left(erfc\left(\frac{1}{\sqrt{2\pi\sigma_j}} \left(\frac{(i-1)M}{q} - MOS_j\right)\right) - erfc\left(\frac{1}{\sqrt{2\pi\sigma_j}} \left(\frac{(i)M}{q} - MOS_j\right)\right)\right)$$

$$= \sum_{i=1}^{q} i \int_{(i-1)M/q}^{iM/q} \sum_{j=1}^{K=5} w_j \frac{1}{\sqrt{2\pi\sigma_j{}^2}} e^{-\frac{(x-MOS_j)^2}{2\sigma_j{}^2}} dx$$

$$\sigma_q{}^2 = \sum_{i=1}^{q} i^2 \int_{\frac{(i-1)M}{q}}^{\frac{iM}{q}} p_x(x)dx - MOS^2{}_q \tag{IV-15}$$

$$= \sum_{i=1}^{q} i^2 \int_{(i-1)M/q}^{iM/q} \sum_{j=1}^{K=5} w_j \frac{1}{\sqrt{2\pi\sigma_j{}^2}} e^{-\frac{(x-MOS_j)^2}{2\sigma_j{}^2}} dx - MOS^2{}_q$$

$$= \frac{1}{2} \sum_{i=1}^{q} i^2 \sum_{j=1}^{K=5} w_j (erfc(\frac{1}{\sqrt{2\pi\sigma_j}} (\frac{(i-1)M}{q} - MOS_j)) -$$

$$- erfc(\frac{1}{\sqrt{2\pi\sigma_j}} (\frac{(i)M}{q} - MOS_j))) - MOS^2{}_q$$

It can be concluded that even if X is not Gaussian distributed, the mean value and standard deviation of Y solely depend on the mean value and standard deviation of the Gaussian *pdf* composing the mixture, *cf*. (IV-14) and (IV-15) and on the q number of levels on the grading scale.

Hence, a similar behavior as in the Gaussian case is expected in quality evaluation: the MOS and σ corresponding to a continuous scale evaluation would allow to a posteriori obtain MOS_q on any q levels grading scale and to also have information about its statistical precision.

IV.3.3. Illustration of continuous to discrete *MOS* computation

The illustrations in this Chapter are structured the same way as Chapter IV.2.1.

The following experiment is carried out 4 times, once for each type of content described in Section III.4 (2D and stereoscopic video, high and low quality). For each type of content, the scores assigned by the 60 observers using a continuous scale (between 1 and *M*) are considered. Subsequently, the underlying Gaussian mixture are estimated and, by applying (IV-14) and (IV-15), the MOS_q can be computed. These values will be further considered as theoretical (reference) values for the mean opinion scores which would have been obtained on a discrete grading scale with q semantic levels.

As in Chapter IV.2, the evaluation on the unlabeled, discrete, grading scales are obtained by mapping the $[x_1, x_2 \dots x_3]$ data to their corresponding values on a would-be q levels grading scale, according to (IV-2), thus obtaining values for the *r.v.* Y: $[y_1 = f(x_1), y_2 = f(x_2), \dots, y_N = f(x_N)]$. Consequently, the

estimated values $\overline{MOS}_{\mathrm{q}}$ and the 95% confidence limits $(MOS_{q;low};\ MOS_{q;up})$can be computed according to (IV-10).

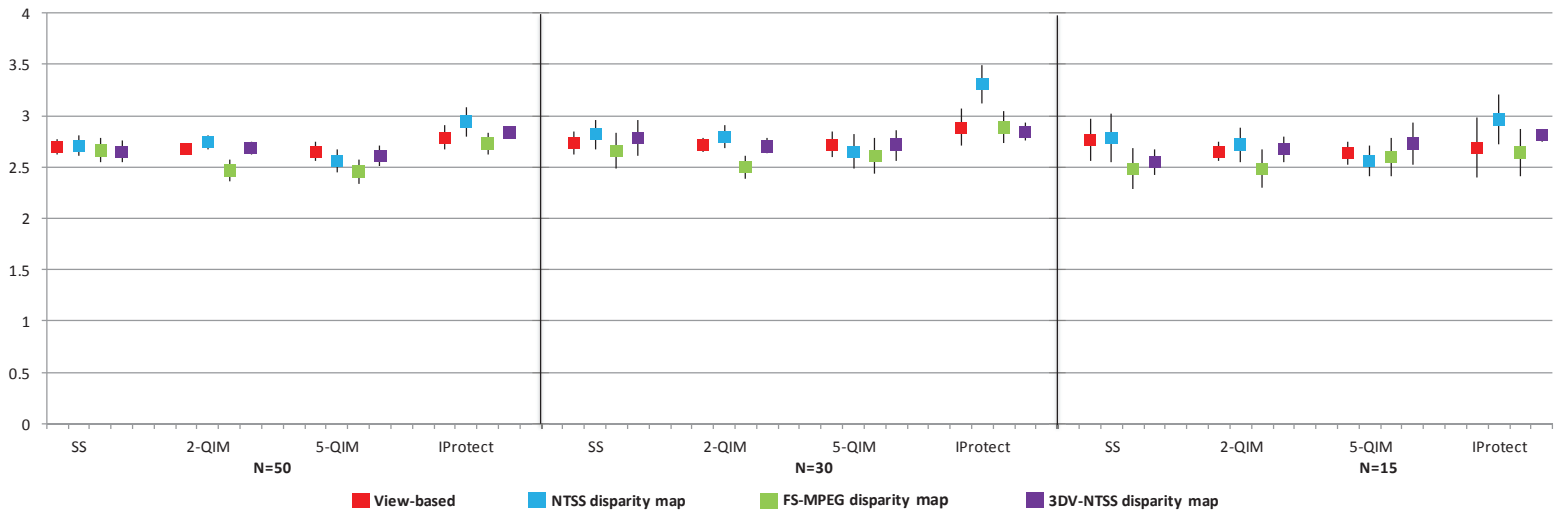
A comparison can be achieved now between the theoretical values MOS_q and their confidence limits $\left(MOS_{q;low}; MOS_{q;up}\right)$.

The results corresponding to each of the four types of content, to *q*=2, 3,…,9 and to *N*=50, 30 and 15 are detailed in Appendix B. In the sequel of the chapter, the same illustrative selection as in Chapter IV.2 is considered, as follows:

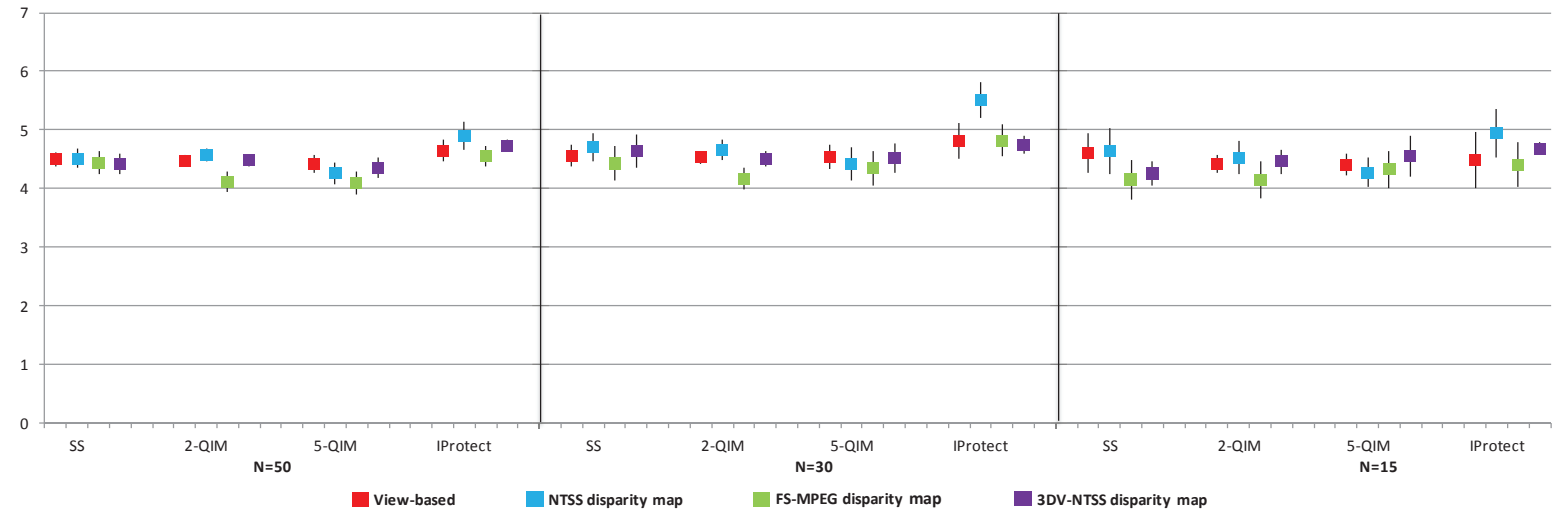
- Figure IV-14 corresponds to the score assigned for the *Image quality* of the high-quality stereoscopic video content; three subplots, denoted by (a), (b) and (c) correspond to *q* = 3, *q* = 5 and *q* = 9, respectively. For each subplot, three experimental cases are presented (from let to right), corresponding to *N* = 50, *N* = 30 and *N* = 15. For each *N* value, 16 types of sequences are presented, as detailed in Chapter III.4.1 (they correspond to watermarked content obtained by applying 4 different methods –SS, 2-QIM, 5-QIM and IProtect, in 4 different insertion domains – view-based, NTSS, FS-MPEG, and 3DV-NTSS.
- Figure IV-15 is similar to Figure IV-14 but refers to low quality stereoscopic video content, as detailed in Chapter III.4.2.
- Figure IV-16 is similar to Figure IV-14 but refers to high-quality 2D video, as detailed in Chapter III.4.3; hence, this time, the observers scores the *Visual quality*.
- Figure IV-17 refers to low quality 2D video, as detailed in Chapter III.4.4. Hence, here again the *Visual quality* is scored but this time 28 types of sequences are scored.

The overall results point to a similar conclusion as in Chapter IV.2.2.4. Even when the scores assigned by the observers on a continuous scale do not follow a Gaussian law and are estimated by a Gaussian mixture, the *MOS* and the confidence limits corresponding to a would-be discrete, unlabeled scale with *q* levels can be theoretically computed starting from the *MOS* and the standard deviations of the Gaussian laws composing the mixture. Beyond its theoretical demonstration, the conclusion is also experimentally supported.

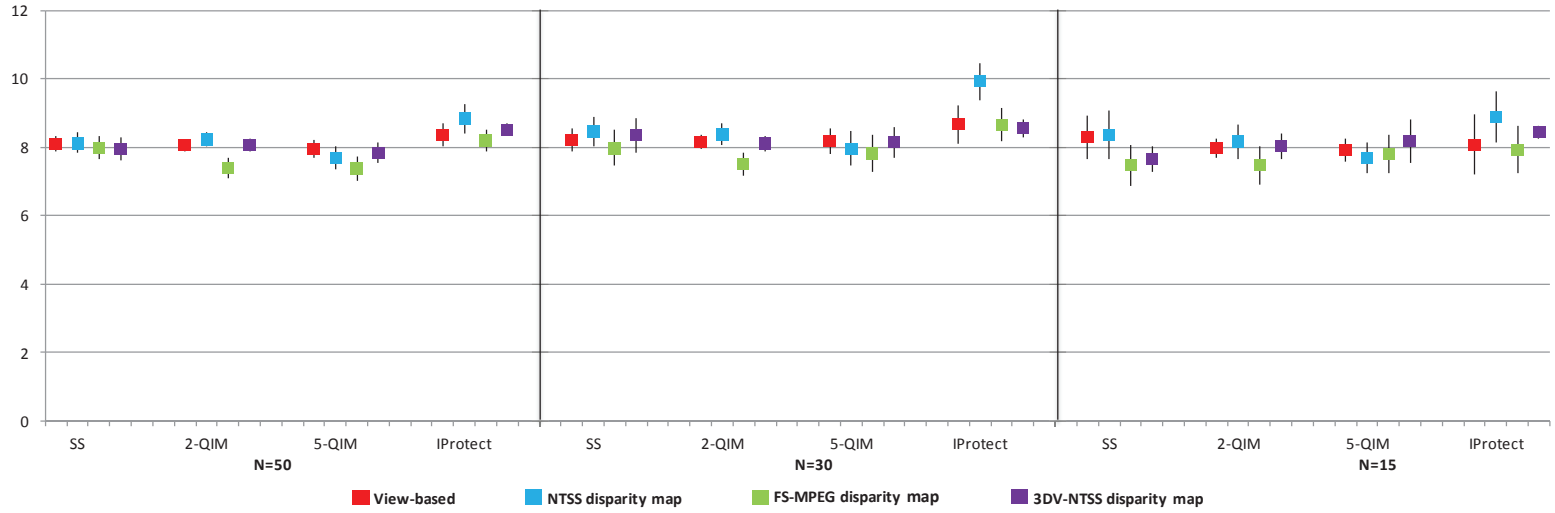
High quality stereoscopic video content



(a)



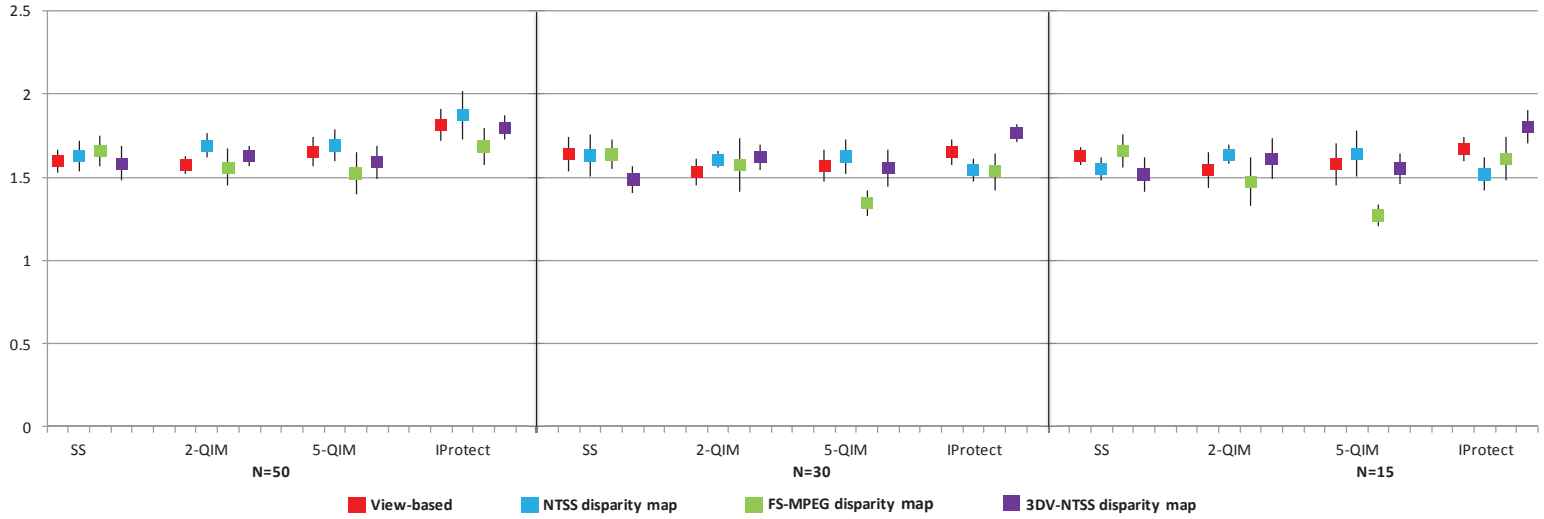
(b)



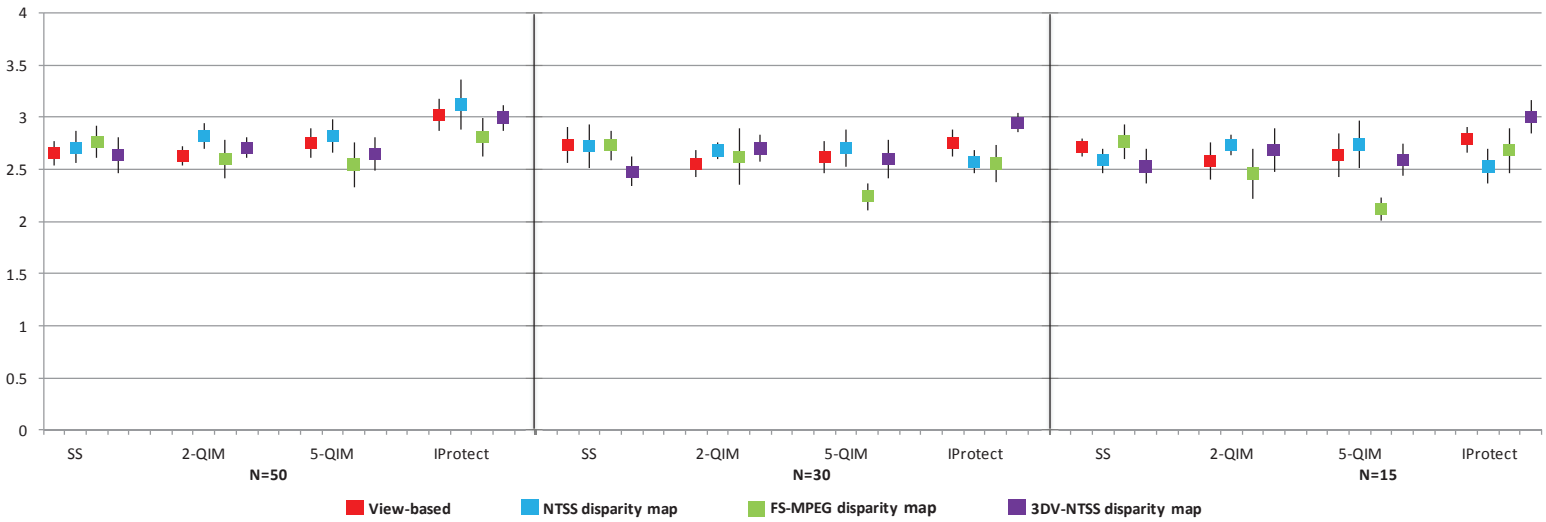
(c)

Figure IV-14: Subjective evaluations for high-quality stereoscopic video content, non-Gaussian case, for grading scales of: (a) $q = 3$ quality levels; (b) $q = 5$ quality levels; (c) $q = 9$ quality levels; and for a number of observers N=50, N=30, N=15.

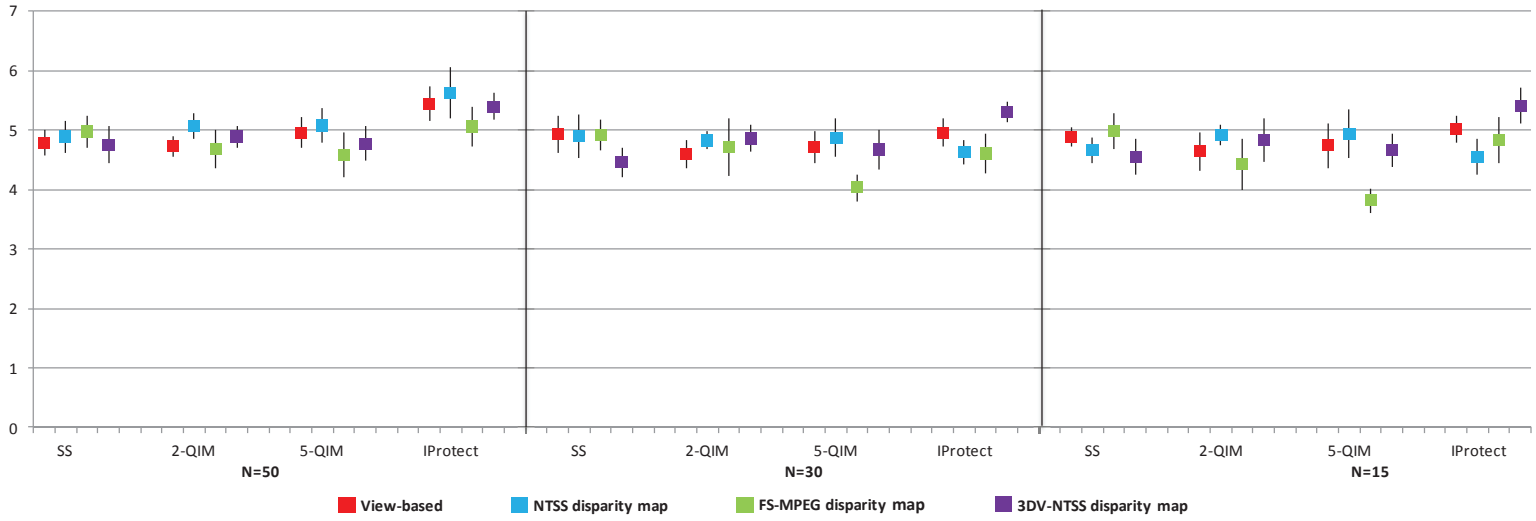
Low quality stereoscopic video content



(a)

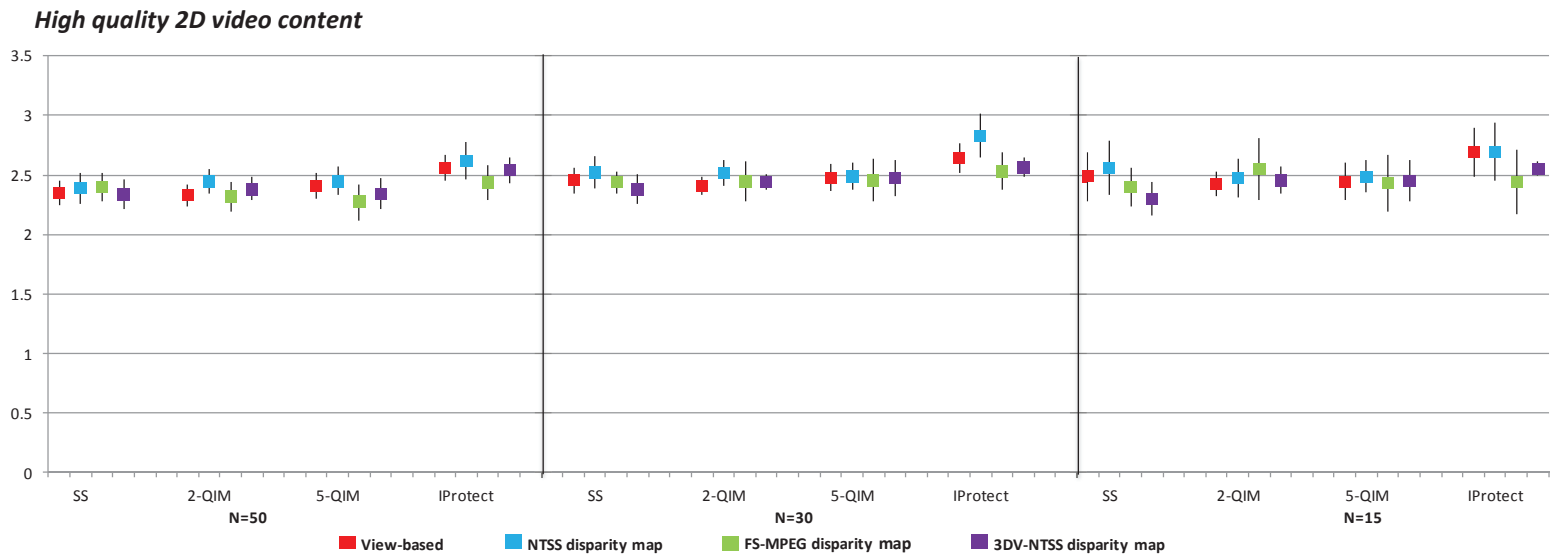


(b)

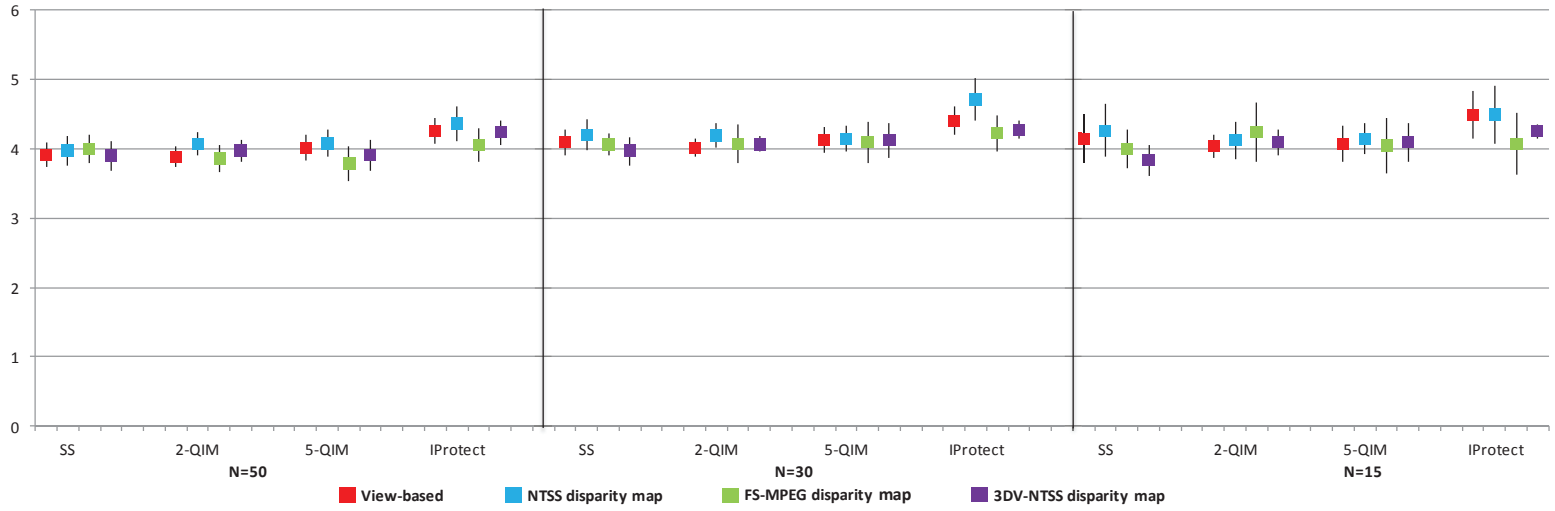


(c)

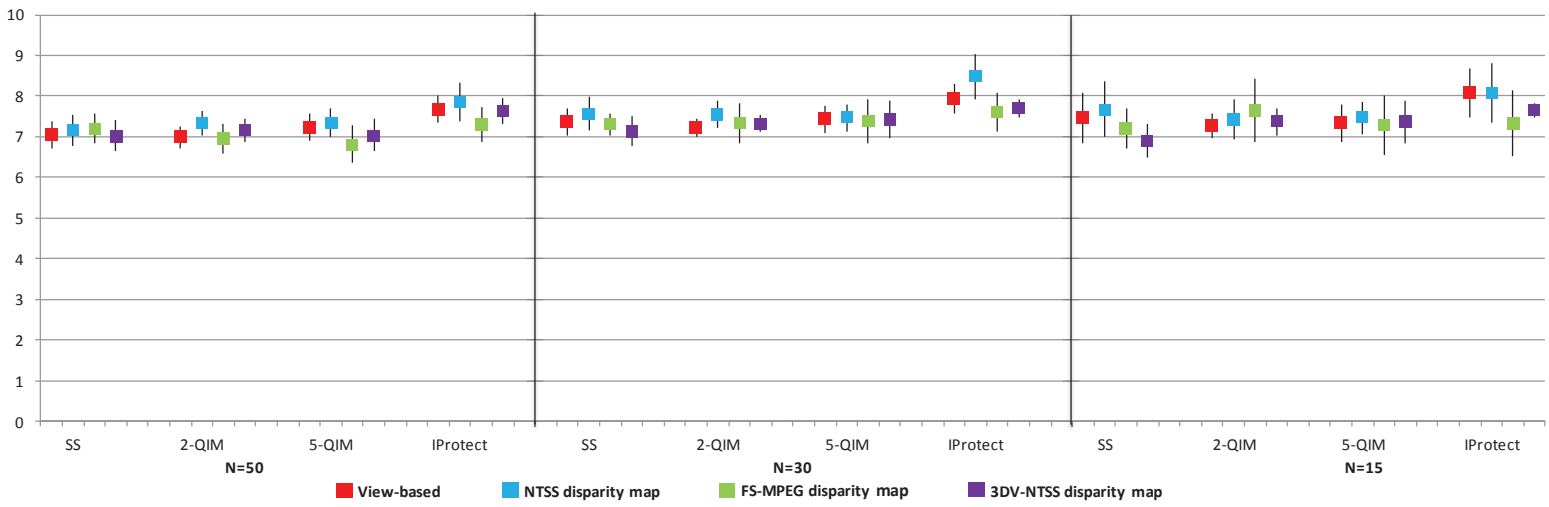
Figure IV-15: Subjective evaluations for low quality stereoscopic video content, non-Gaussian case, for grading scales of: (a) $q = 3$ quality levels; (b) $q = 5$ quality levels; (c) $q = 9$ quality levels; and for a number of observers N=50, N=30, N=15.


High quality 2D video content
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(a)

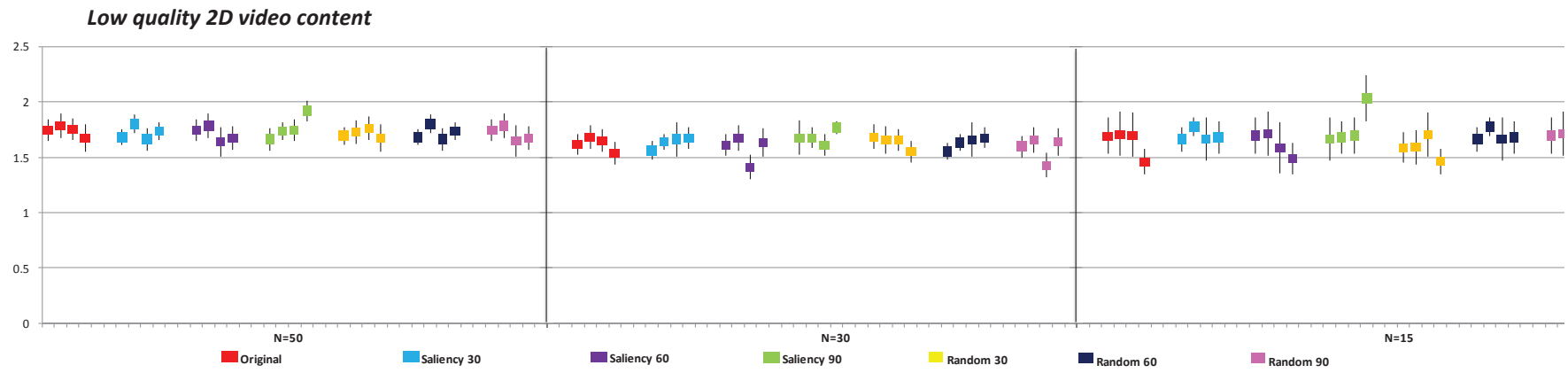

6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(b)

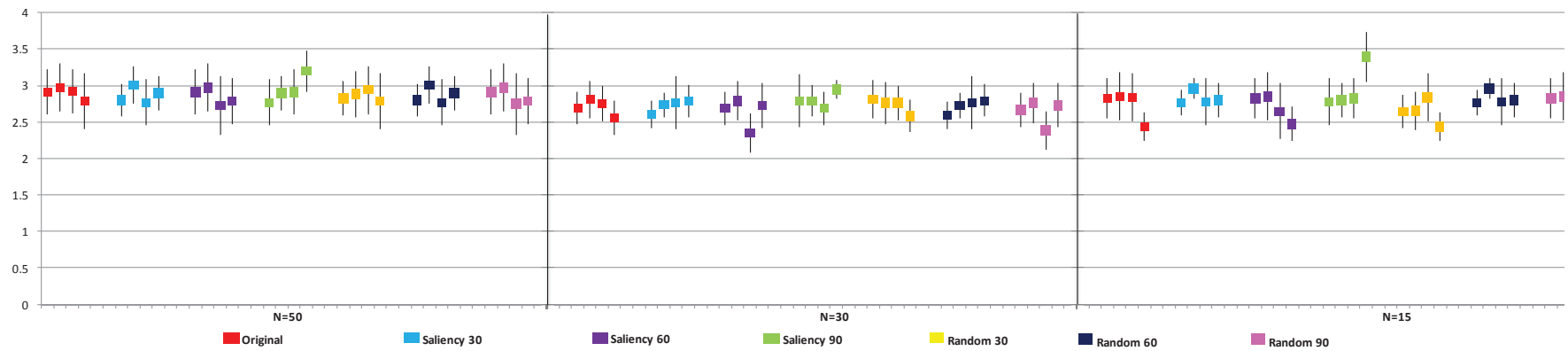


(c)

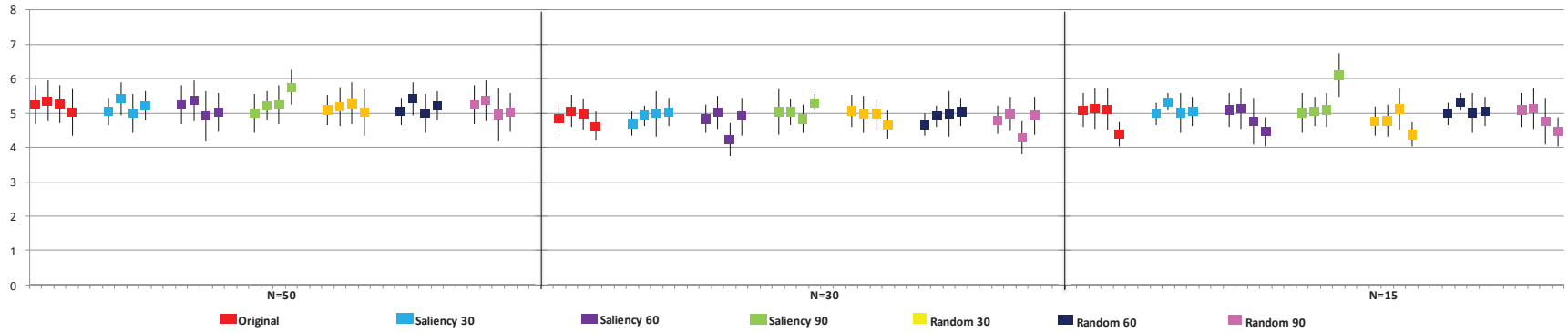
Figure IV-16: Subjective evaluations for high quality 2D video content, non-Gaussian case, for grading scales of: (a) $q = 3$ quality levels; (b) $q = 5$ quality levels; (c) $q = 9$ quality levels; and for a number of observers N=50, N=30, N=15.


Low quality 2D video content
2.5
2
1.5
1
0.5
0
N=50
N=30
N=15
Original
Saliency 30
Saliency 60
Saliency 90
Random 30
Random 60
Random 90

(a)


4
3.5
3
2.5
2
1.5
1
0.5
0
N=50
N=30
N=15
Original
Saliency 30
Saliency 60
Saliency 90
Random 30
Random 60
Random 90

(b)



(c)

Figure IV-17: Subjective evaluations for low quality 2D video content, non-Gaussian case grading scales of: (a) $q = 3$ quality levels; (b) $q = 5$ quality levels; (c) $q = 9$ quality levels; and for a number of observers N=50, N=30, N=15.

IV.3.4. Investigation on the accuracy of the results

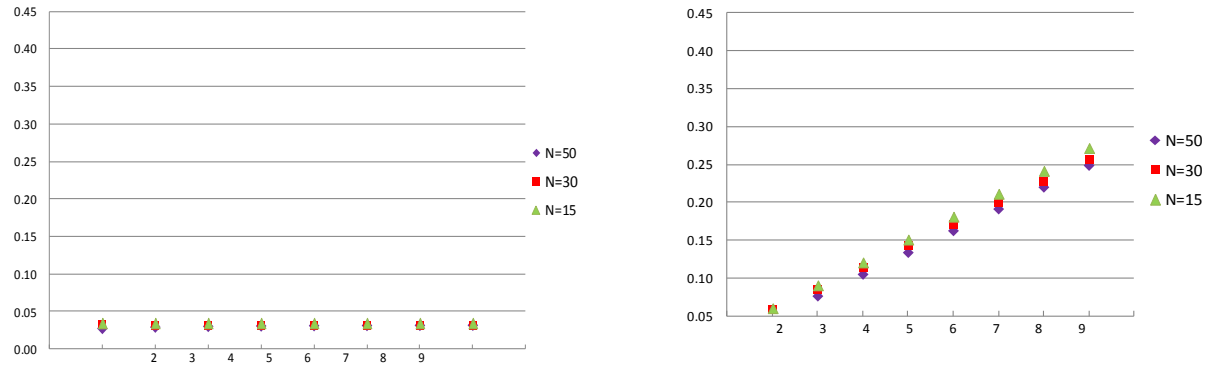
The experiments are conducted the same way as in the Gaussian case, Chapter IV.2.3.

For a given number of quality levels on the grading scale *q*, the absolute and relative errors in MOS_q computation with a 95% confidence level, denoted by ε_q and ε_q^r can be approximated by (IV-11) where MOS_q and σ_q can be computed according to (IV-14) and (IV-15) from the MOS and σ (*i.e.* from the values estimated on the $M = 100$ levels grading scale).

- Figure IV-18 illustrates the overall impact of *N* in the absolute and relative errors of high quality stereoscopic video when assessing the image quality for *N*=50, 30 and 15 and *q*=2,3,..,9;
- Figure IV-19 illustrates the overall impact of N in the absolute and relative errors of high quality stereoscopic video when assessing the depth perception for *N*=50, 30 and 15 and *q*=2,3,..,9;
- Figure IV-20 illustrates the overall impact of N in the absolute and relative errors of high quality stereoscopic video when assessing the visual comfort for *N*=50, 30 and 15 and *q*=2,3,..,9;
- Figure IV-21 is similar to Figure IV-18 but refers to low quality stereoscopic video content, as detailed in Chapter III.4.2;
- Figure IV-22 is similar to Figure IV-19 but refers to low quality stereoscopic video content, as detailed in Chapter III.4.2,
- Figure IV-23 is similar to Figure IV-20 but refers to low quality stereoscopic video content, as detailed in Chapter III.4.2;
- Figure IV-24 is similar to Figure IV-18 but refers to high-quality 2D video, as detailed in Chapter III.4.3; hence, this time, the observers scores the *Visual quality*;
- Figure IV-25 refers to low quality 2D video, as detailed in Chapter III.4.4.

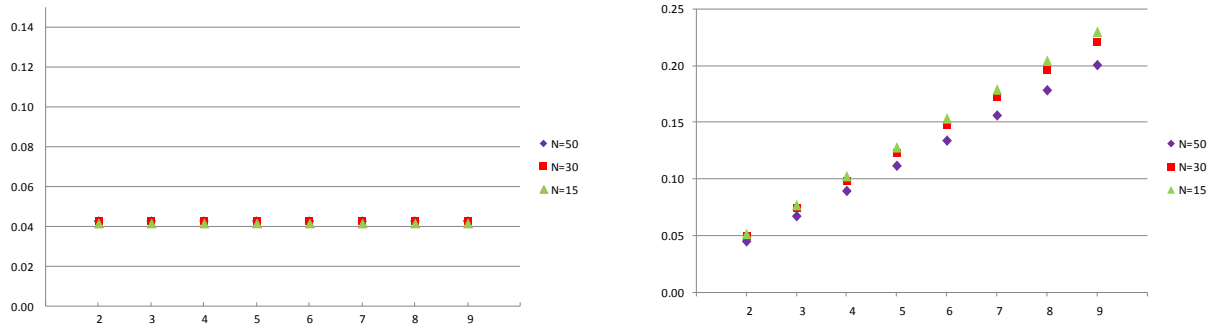
The same general conclusion as in the Gaussian case can be drawn (see Chapter IV.2.3): for a given *N* value, the absolute error is an increasing function of *q* (see the left-side plots in Figures IV-18, IV-19, IV-20, IV-21, IV-22, IV-23, IV-24 and IV-25) while the relative errors are quite constant with q (see the right-side plots Figures IV-18, IV-19, IV-20, IV-21, IV-22, IV-23, IV-24 and IV-25). Yet, the upper limit for the differences between the relative errors corresponding to the various investigated configurations (4 types of content, *N*=50, 30 and 15 and *q*=2,3,..,9) is now reduced from 0.04 (see Chapter IV.2.3) to 0.02.

High quality stereoscopic video content



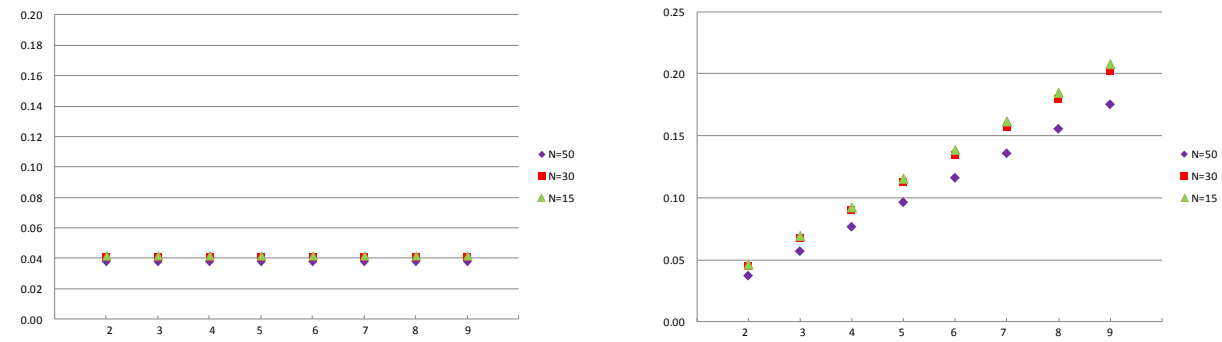
(a) (b)

Figure IV-18: Precision in the evaluation of high quality stereoscopic video, non-Gaussian case (Image quality) on the abscissa – the number of the quality levels $q = 2,3,\ldots,9$; on the ordinate - the 95% error value: (a) Relative error; (b) Absolute error.



(a) (b)

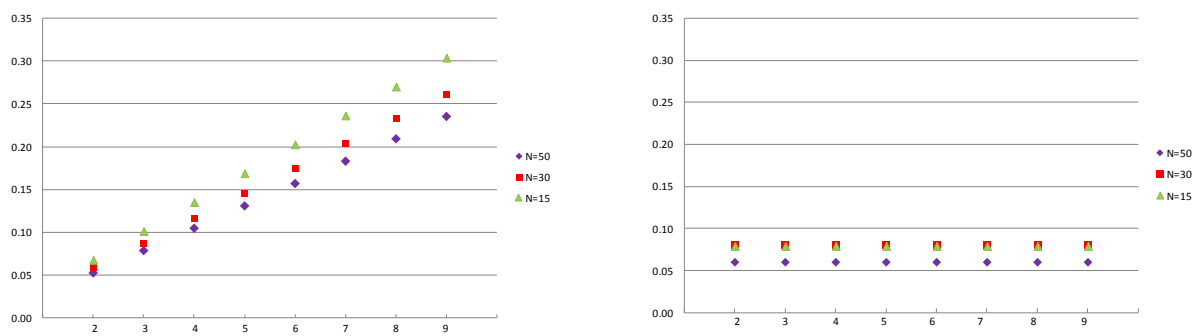
Figure IV-19: Precision in the evaluation of high quality stereoscopic video, non-Gaussian case (Depth perception) on the abscissa – the number of the quality levels $q = 2,3,\ldots,9$; on the ordinate - the 95% error value: (a) Relative error; (b) Absolute error.



(a) (b)

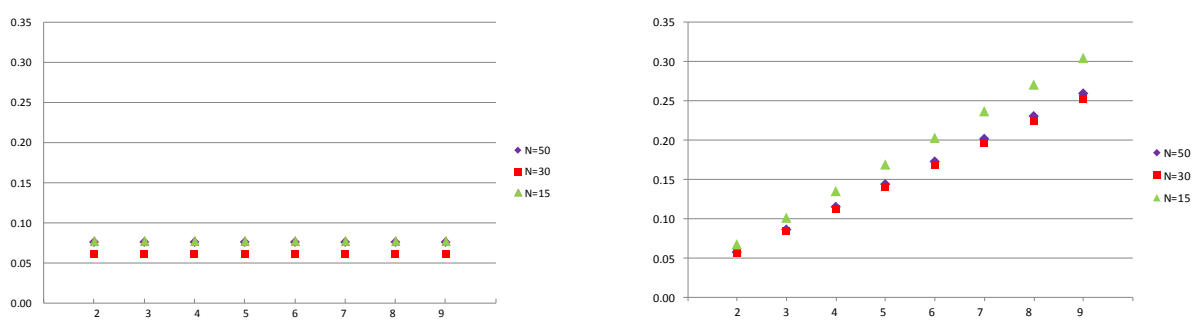
Figure IV-20: Precision in the evaluation of high quality stereoscopic video, non-Gaussian case (Visual comfort) on the abscissa – the number of the quality levels $q = 2,3,\ldots,9$; on the ordinate - the 95% error value (a): Relative error; (b) Absolute error.

Low quality stereoscopic video content



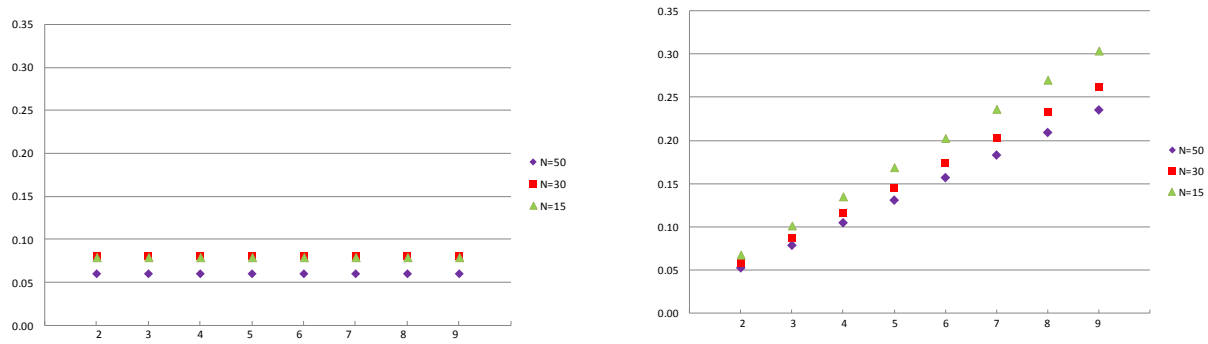
(a) (b)

Figure IV-21: Precision in the evaluation of low quality stereoscopic video, non-Gaussian case (Image quality) on the abscissa – the number of the quality levels $q = 2,3,\ldots,9$; on the ordinate - the 95% error value.: (a) Relative error; (b) Absolute error.



(a) (b)

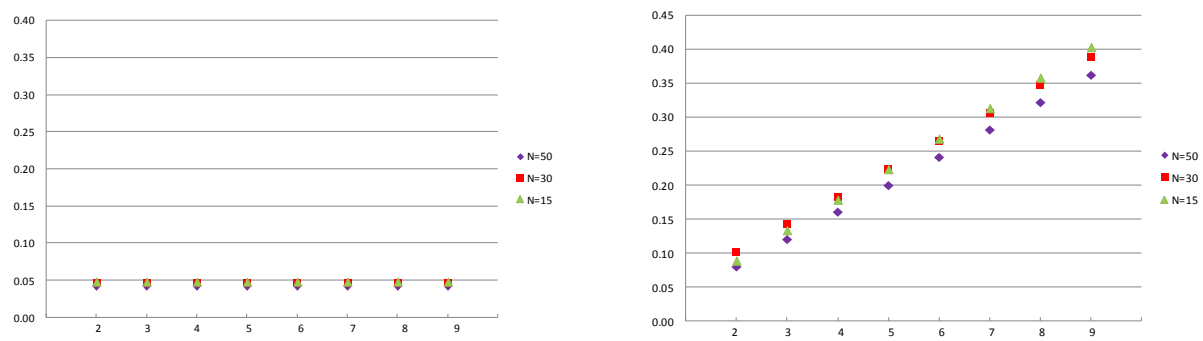
Figure IV-22: Precision in the evaluation of low quality stereoscopic video, non-Gaussian case (Depth Perception) on the abscissa – the number of the quality levels $q = 2,3,\ldots,9$; on the ordinate - the 95% error value.: (a) Relative error; (b) Absolute error.



(a) (b)

Figure IV-23: Precision in the evaluation of low quality stereoscopic video, non-Gaussian case (Visual comfort) on the abscissa – the number of the quality levels $q = 2,3,\ldots,9$; on the ordinate - the 95% error value: (a) Relative error; (b) Absolute error.

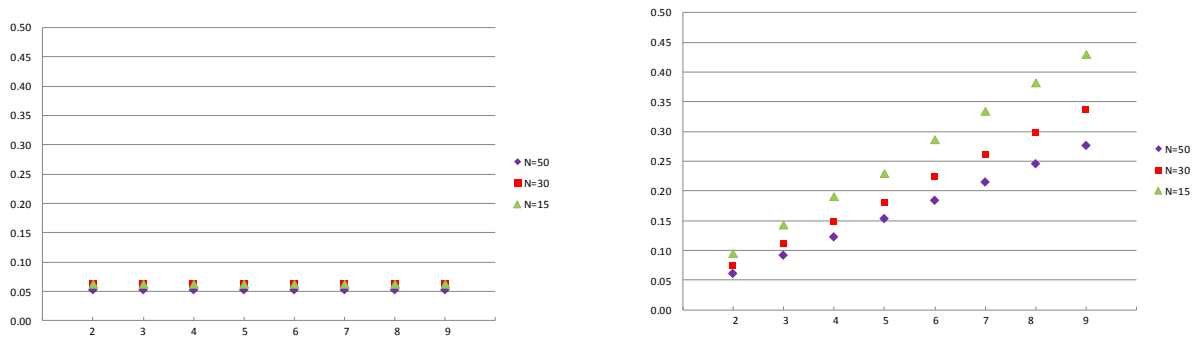
High quality 2D video content



(a) (b)

Figure IV-24: Precision in the evaluation of high quality 2D video, non-Gaussian case on the abscissa – the number of the quality levels $q = 2,3,\ldots,9$; on the ordinate - the 95% error value: (a) Relative error; (b) Absolute error.

Low quality 2D video content



(a) (b)

Figure IV-25: Precision in the evaluation of low quality 2D video, non-Gaussian case on the abscissa – the number of the quality levels $q = 2,3,\dots,9$; on the ordinate - the 95% error value: (a) Relative error; (b) Absolute error.

IV.4. Conclusion

By using non-linear random variable transformations principles, Chapter IV.1 establishes the formula connecting the probability density functions modeling the continuous and discrete grading scales scoring. The first and second order moments (allowing for the *MOS* and its confidence limits to be computed) are subsequently derived. This formula is generic and does not take into account the original probability density function peculiarity. Consequently, Chapters IV.2 and IV.3 consider the instantiations of this formula for two opposite cases of practical relevance, namely the case in which the scores according on a continuous scales are Gaussian distributed and the case in which these scores are not Gaussian (but estimated *via* a Gaussian mixture). In the two cases, it is brought to light that the *MOS* and the related confidence limits solely depend on the average and variances of the continuous scale models and of the q number of quality levels on the discrete scale. The theoretical results are illustrated through experiments corresponding to the scores assigned by the 4 *reference* panels (one panel for each type of content – see Chapter III).

The impact of the number of human observers in the precision of the quality evaluation is subsequently assessed. It is demonstrated that for 3 N values ranging from 15 to 50, the relative errors in *MOS* estimation are constant with both N and q (differences lower than 0.04 and 0.02 being obtained for Gaussian and non-Gaussian cases, respectively).

To conclude with, in order to converge to a unique answer to the controversial issue related to the number of quality levels on a stereoscopic video grading scale, Chapter IV suggests to perform the evaluation on a continuous grading scale, with no semantic labels associated to the scores and to subsequently map these values on the discrete scales.

Such an approach is of benefit to the evaluation procedure, the tests become more user-friendly and the observer is no longer restricted in his/her choice to some pre-established application-dependent options. Moreover, finer and richer information about the evaluated content can be obtained out of a single experiment. Just for illustration, assume the case of a stereoscopic content provider who would like to perform a single experiment in order to assess the quality of a same content which is to be delivered on both TV (hence, strong quality constraints requiring $q=5$ or even more) and on some stereoscopic enabled smart-phone (where a coarser evaluation with $q=3$ is likely to be accepted).

V. Quantifying the impact of the semantic labels

The present chapter addresses the second challenge of the thesis, namely the would-be semantic impact of the ITU labels in the evolution scores: the existence of the semantic impact is first demonstrated, then the possibility of its assessment is addressed.

The existence of the semantic impact is verified by a comparison (based in the Student's paired test) between the average values (representing the MOS) corresponding to the continuous (unlabeled) and discrete, semantically labeled scales.

The main contribution of the chapter is the definition a methodological framework for quantifying this semantic impact. In this respect, an auxiliary discrete random variable is defined: it is characterized by uneven partition but by equal a posteriori probabilities. By comparing (through binomial tests) the differences in the partition class length between this auxiliary random variable and the random variable corresponding to the semantically labeled scale, the semantic impact is quantified (by defining an underlying coefficient).

The experimental results have two-folded outcomes: (1) reference values for the semantic impact (i.e. general values, independent with respect to the observers) are computed for each type of content (stereoscopic and 2D video, high and low quality) and for the two investigated numbers of quality levels (q=3 and q=5); (2) the fact that practically all the labels (and not only Excellent) are involved in the overall semantic impact is demonstrated.

Chapter I.2.6 identified three controversial issues related to the ITU evaluation procedures, related to the use of *continuous* and *discrete* terms in conjunction with a grading scale, to the statistical model of the scores assigned by the observers as well as to the use of semantic labels for the grading scales. While the first two of them are dealt with in Section IV, the present Chapter V addresses the last issue, namely the would-be semantic impact of the ITU labels in the evolution scores.

In this respect, the existence of the semantic impact is first investigated (Chapter V.1), then the possibility of its assessment is addressed (Chapter V.2).

V.1. Verifying the existence of the semantic impact of the labels

The semantic impact of the labels is the fact that the labels attached to the discrete evaluation scale stimulate or make the observer more reluctant in assigning the class corresponding to that level, thus modifying the *MOS*.

Consequently, in order to verify whether such a semantic impact exist, the mean opinion scores corresponding to unlabeled and labeled scales should be compared between them. Of course, should these two types of *MOS* be identical, no semantic impact is encountered. On the contrarily, differences between these two types of *MOS* bring to light the existence of the semantic impact.

The un-labeled scores correspond to the scores assigned by the *reference* panel while the labeled scores correspond to the scores assigned by the *validation* panel (the 5 level grading scales and 3 levels grading scales are alternatively investigated).

140 data sets are processed (see Chapter IV.3.1 and Chapter III.4):

- 48 score sets for the high quality stereoscopic content, corresponding to the 16 sequences under test and to the 3 evaluation criteria (*Image quality*, *Depth perception* and *Visual Comfort*);
- 48 score sets for low quality stereoscopic content;
- 16 score sets for the high quality 2D video content;
- 28 score sets for low quality 2D video content.

For each of the 140 *reference* data sets, the unlabeled *MOS* is computed according to (IV-14) and considered as theoretical values; actually, when applying (IV-14), two q values are alternatively considered, namely q = 5 and q = 3.

The 140 *validation* scores are considered as experimental data which are compared to these theoretical values by applying an unpaired Student's t-test [WAL93]. Such a test is applied with N=25 (the size of *validation* corpus), $\alpha = 0.05$ and estimated variance.

The experimental results bring to light the number of Student tests which are not passed: 23 when q = 5 and 17 when q = 3. As the proportions of test which are not passed are 0.37 and 0.12 (which are larger than $\alpha = 0.05$) we can state that the semantic impact does exist but that it seems to be less important for q = 3 than for q = 5.

V.2. Methodological framework for semantic impact assessment

V.2.1. Evaluation principle

The principle of the semantic impact evaluation is illustrated in Figure V.1.

In Chapter IV, the *r.v. X* and *Y* are introduced so as to model the scores assigned by the observers on a continuous, unlabeled scale and to compute the discrete, unlabeled *q* level grading scale, (IV-14).

Assume now a discrete *r.v. Z* modeling the scores assigned by the observers, on a discrete, labeled scale.

In case no semantic impact exists, the *Y* and *Z r.v.* should be identical (that is, they are expected to take the same *q* values with the same probabilities). On the contrarily, the differences between the *Y* and *Z r.v.* bring to light the existence of the semantic impact and allow for its assessment.

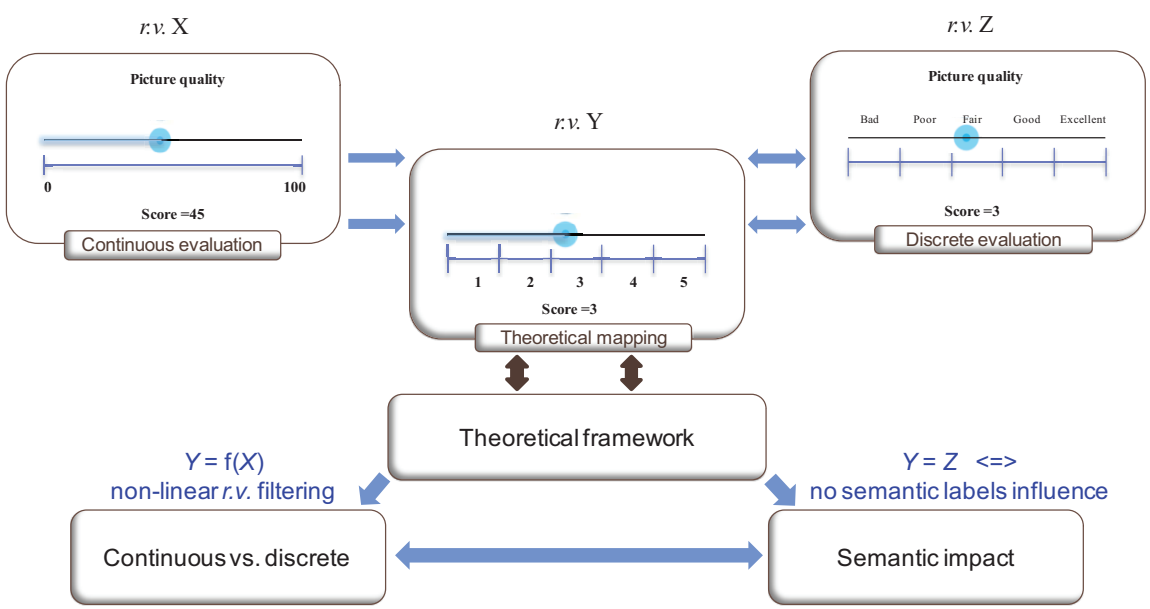


Figure V-1: Principle of the semantic impact evaluation

V.2.2. Methodology presentation

The methodological framework for assessing the semantic impact of the labels is presented in Figure V-2. It is composed of three main steps, related to the computation of the *Y r.v.* by using the results in Chapter IV, to the estimation of the *Z r.v.* from the scores assigned by the observers (the *validation* panels), and to the assessment of the semantic impact as the differences between the *Y* and *Z*.

Step 1: Continuous & discrete unlabeled scales

Step 1.1 Perform the continuous evaluation experiment

According to this step, the human observers are asked to score the content on a continuous scale, *e.g.* between 0 and M; the result of this step is the data set $[x_1, x_2, \dots, x_N]$, corresponding to the *reference* panels.

Step 1.2 Estimate the $p_X(x)$ probability density function (pdf)

This step is performed by a Gaussian mixture, whose parameters are estimated under an EM (expectation-maximization) criterion. The result of this step is the $p_X(x)$ pdf.

Step 1.3 Compute the Y random variable, i.e. the discretization of X according to an even partition $[0 = y_1, y_2, \dots, y_q = M]$ of the $[0, M\ ]$ interval

This step is performed by applying a non-linear random variable filtering operation $Y = f(X)$, according to (IV-2).

Step 2: Discrete, labeled scale evaluation

Step 2.1 Perform the discrete evaluation experiment

According to this step, the human observers are asked to score the content on a q levels semantic-labeled discrete scale; this way, the data sets $[z_1, z_2, \dots, z_N]$ corresponding to the *validation* panels (cf. Chapter III.5) are obtained.

Step 2.2 Estimate the $p_Z(z)$ probability density function

This step can be performed by any discrete *pdf* estimation method, applied to the data set obtained in the previous step. For instance, in the thesis, a frequency based estimation is considered:

$$p_Z(z) = \sum_{i=1}^{q} p_Z(i)\, \delta(z - i) \tag{V-1}$$

where $p_z(z)$ is the relative frequency of the ith quality class.

Step 3: Unlabeled vs. labeled discrete scales

Step 3.1 Find the identity condition

This step searches for the $[0 = y_1, y_2, \dots, y_q = M]$ partition ensuring identity between the Y and Z random variables. In this respect, the $p_Y(y)$ can be considered as theoretical reference and $p_Z(z)$ as an experiment outcome to be validated through a goodness-on-fit test (*e.g.* the binomial test).

Step 3.2 Compute the relative variation of the partition intervals with respect to the uniform partition

This step computes the set of coefficients $\rho_{q-i}, i = 0,1, \dots q - 1$, where:

$$\rho_{q-i} = \frac{y_{q-i} - y_{q-i-1}}{M/q}. \qquad \text{(V-2)}$$

A unitary value for such a coefficient demonstrates that the related semantic label does not modify the evaluation - that is, an even partition $[0 = y_1, y_2, \dots, y_q = M]$ ensures the identity between *Y* and *Z*. A value larger than 1 indicates that the related semantic label makes the observer more likely to score that way while, conversely, a value lower than 1 shows that the related label makes the observers more reluctant in assigning that label when scoring.

Note that this investigation is finer than the one carried out in Chapter V.1: this time the whole *pdf* is considered while in previous case only its mean value.

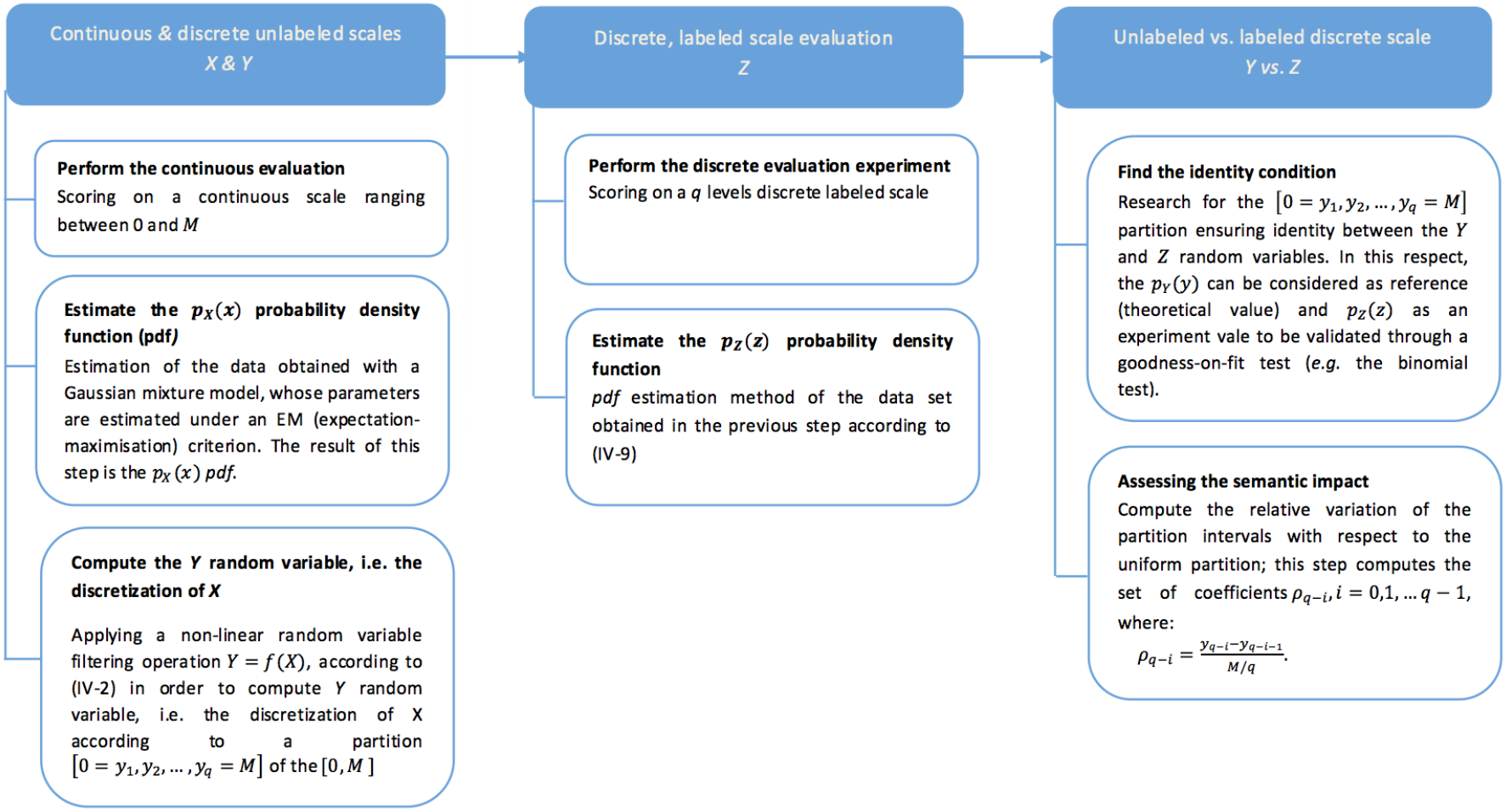


Figure V-2: The methodological framework.

V.2.3. Quantitative results

V.2.3.1. A finer investigation of the semantic impact existence

Chapter V.1 brings to light the existence of the semantic impact in the *MOS* computation; that is, it was demonstrated there that the *r.v*. modeling the scores assigned on unlabeled and labeled scales have different mean values. The methodological solution presented in Chapter V.2.2 makes possible a finer analysis: this time, the very *pdf* modeling the scores assigned on unlabeled and labeled scales can be compared between them.

The same experimental conditions as in Chapter IV and Chapter V.1 are kept.

This time, the experiment is carried out separately for the 4 types of content.

The un-labeled scores correspond to the scores assigned by the *reference* panel while the labeled scores correspond to the scores assigned by the *validation* panel (the 5 level grading scales and 3 levels grading scales are alternatively investigated). Each time, 140 data sets are processed (see Chapter IV.3.1 and Chapter III.4):

- 48 score sets for the high quality stereoscopic content;
- 48 score sets for low quality stereoscopic content;
- 16 score sets for the high quality 2D video content;
- 28 score sets for low quality 2D video content.

Two *q* values are alternatively considered, namely *q* = 5 and *q* = 3.

For each type of content, let us consider the *reference* panel, the same $p_X(x)$ estimated by means of $K=5$ mixture of Gaussian laws, and the same evenly distributed partition of the $[0,M]$ interval. The probabilities of Y taking the values $1,2,3,4,q=5$, denoted by $P(Y=1)$, $P(Y=2)$, … $P(Y=5)$, can be computed according to:

$$p_Y(Y=l)=\sum_{i=1}^{q}\delta(l-i)\int_{y_{i-1}}^{y_i}p_X(x)dx=\int_{y_{l-1}}^{y_l}p_X(x)dx \qquad \text{(V-1)}$$

where $\delta(.)$ denotes the Dirac's Delta function, see Figure IV-1.

These values will be further considered as reference (theoretical) values.

Let us consider now the *validation* corpus. The $p_Z(i)$, $i\in\{1,2,3,4,5\}$ are estimated as the relative frequencies of the scores assigned on the labeled scales, namely *Excellent*, *Good*, *Fair*, *Poor* (for *q* = 5 quality levels evaluation) and *Bad* or *Good*, *Fair* and *Bad* (for *q* = 3 quality levels evaluation).

These values will be further considered as experimental values which are compared to the theoretical values $P(Y=i)$ according to a binomial test (the test on probability), applied at $\alpha=0.05$ [WAL93].

The experimental results will be subsequently detailed for each type of content.

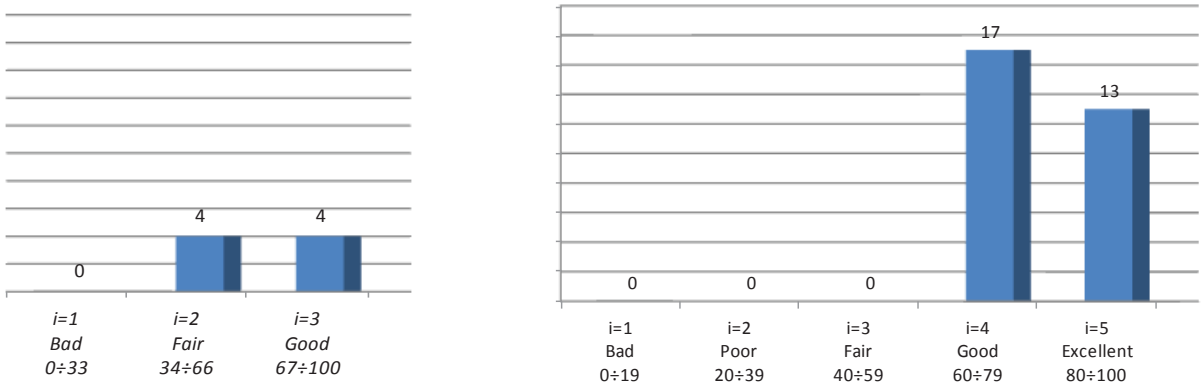
High quality stereoscopic video content

The number of tests which are not passed (*i.e.* the number of cases in which statistical significant differences between the probabilities of Y and Z are encountered) are presented in Figure V-3, as a function of the investigated value *i* (equivalent, of the semantic label): the left side correspond to *q* = 3 while the right side to *q* = 5.

For *q* = 3, Figure V-3.a shows that no statistical differences are encountered for *i=1* (*i.e.* for *Poor*). However, differences are spotted out for $i \in \{2,3\}$, (*i.e.* for *Fair* and *Good*): 4 tests for each class (out of the total of 48) are not passed.

For *q* = 5, Figure V-3.b shows that no statistical differences are encountered for $i \in \{1, 2, 3\}$, (*i.e.* for *Bad, Poor and Fair*). However, differences are spotted out for $i \in \{4,5\}$, *i.e.* for *Good* and *Excellent*): 17 and 13 tests (out of the total of 48) are not passed, respectively.

These results reinforce the preliminary investigation in Chapter V.1: Figure V-3 demonstrates that statistical differences between the *pdf*s of Y and Z *r.v.* exist and that the semantic labels come across with psycho-cognitive side effects in the scores. Moreover, it is thus demonstrated that such semantic impact is not only connected to the *Excellent* label. Chapter V.3.2.2 will quantify these effects.



(a) (b)

Figure V-3: The number of failed binomial tests between the values taken by the Y and Z *r.v.*, when scoring high quality stereoscopic video content by the *validation* panel: (a) for *q=3;* (b) for *q=5*.

Low quality stereoscopic video content

The results corresponding to the evaluation of the low quality stereoscopic video content are presented in Figure V-4, which is organized the same way as Figure V-3. While the semantic impact is also identified, this time it is mainly associated to the *Bad* and *Fair* labels for *q=3* where no statistical differences are encountered for *i=1* (*i.e.* for *Bad*). However, differences are spotted out for $i \in \{2,3\}$, (*i.e.* for *Fair* and *Good*): 1 and 3 tests (out of the total of 48) are not passed, respectively.

For q = 5, Figure V-4.b shows that no statistical differences are encountered for *i=1*, (*i.e.* for *Bad*). However, differences are spotted out for $i \in \{2,3,4,5\}$, (*i.e.* for *Poor, Fair, Good* and *Excellent*): 9, 12, 2 and 1 tests (out of the total of 48) are not passed, respectively.

Figure V-4 strengthen the conclusion from Figure V-3 that statistical differences between Y and Z *r.v.* exist and that the various semantic labels come across with psycho-cognitive side effects in the scores.

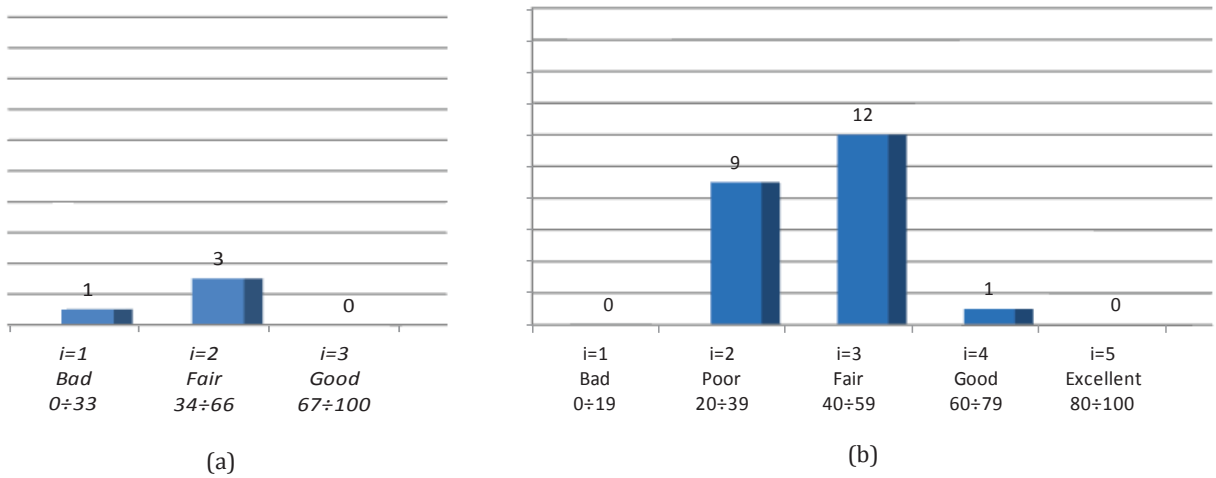


Figure V-4: The number of failed binomial tests between the values taken by the Y and Z *r.v.* when scoring low quality stereoscopic video content by the *validation* panel: (a) for *q=3;* (b) for *q=5*.

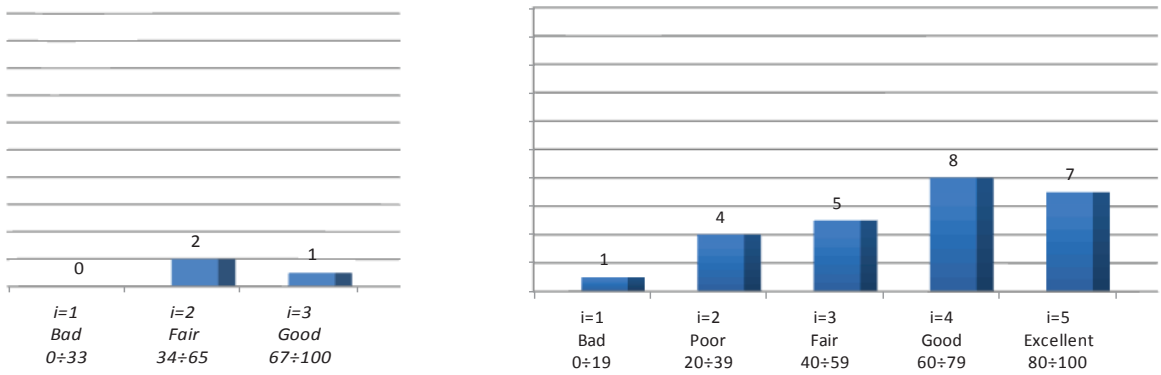
High quality 2D video content

The results corresponding to the evaluation of the high quality 2D video content are presented in Figure V-5, which is organized the same way as Figure V-3.

The semantic impact is now less important, the number of exceptions being under the limit of statistical significance to which the tests are applied: differences are spotted out for $i \in \{2,3\}$, (*i.e.* for *Fair* and *Good*): 2 and 1 tests (out of the total of 16) are not passed, respectively, see Figure V-5.a.

For q = 5, Figure V-5.b shows that differences are spotted out for $i \in \{1,2,3,4,5\}$, (*i.e.* for *Bad*, *Poor, Fair, Good* and *Excellent*): 1, 4, 5, 8 and 7 tests (out of the total of 16) are not passed, respectively.

Here again, Figure V-5 demonstrates that statistical differences between Y and Z *r.v.* exist and reinforce the idea that *Excellent* is not the single label "guilty" for that. Chapter V.3.2.2 will quantify these effects.



(a) (b)

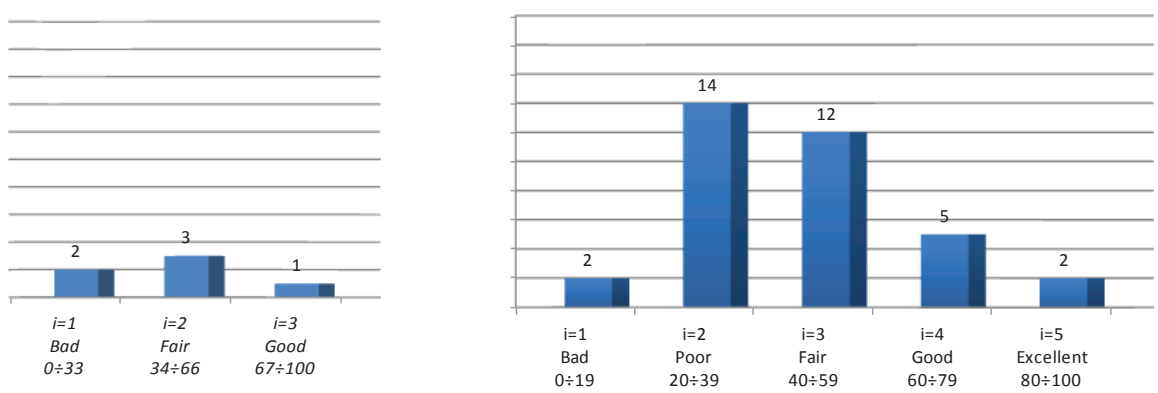
Figure V-5: The number of failed binomial tests between the values taken by the *Y* and *Z r.v.* when scoring high quality 2D video content by the *validation* panel: (a) for *q=3;* (b) for *q=5*.

Low quality 2D video content

The results corresponding to the evaluation of the high quality 2D video content are presented in Figure V-6, which is organized the same way as Figure V-3. This time, the semantic impact it is mainly associated to the *Bad* and *Fair* labels for *q=3*: 2, and 3 tests (out of the total of 16) are not passed, respectively, see Figure V-6.a.

For *q* = 5, Figure V-5.b shows differences for $i \in \{1,2,3,4,5\}$, (*i.e.* for *Bad*, *Poor, Fair, Good* and *Excellent*):2, 14, 12, 5 and 2 tests (out of the total of 28) are not passed, respectively.

Figure V-6 demonstrates the same overall behavior: the existence of the semantic impact, as a consequence of several semantic labels.



(a) (b)

Figure V-6: The number of failed binomial tests between the values taken by the *Y* and *Z r.v.* when scoring high quality 2D video content by the *validation* panel: (a) for *q=3;* (b) for *q=5*.

V.2.3.2. Quantifying the semantic impact

The impact of the semantic labels is quantified by finding the uneven partition $\lfloor y_0 = 0, y_1, y_2, ..., y_q = M \rfloor$ which leads to statistical identity between Y and Z *pdf*s and by subsequently computing the underlying ρ coefficient.

The experimental conditions are kept the same as in Chapter V.2.3.1. The results are presented in Table V.1 and Table V.2 and illustrated in Figures V-3 – V-6 (corresponding to the 4 types of investigated content). The tops (upper parts) of the Figures V-3 – V-6 represent the case of q = 3 while their bottoms (lower parts) represent the case q = 5.

In order to explain the experiments and the meaning of Figures V-3 – V-6 in the sequel the case of high quality stereoscopic content and of q = 5 is considered.

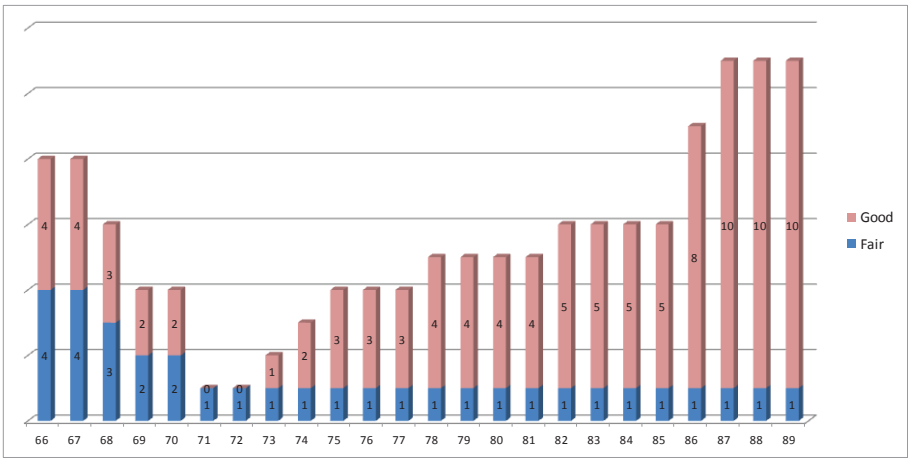
Chapter V.2.3.1 brings to light that, for high quality stereoscopic video content, differences between the *Y* and *Z pdf*s are only related to the *Good* and *Excellent* labels. Hence, the comparison between Y and Z are resumed on uneven partitions of the type $[y_0 = 0; y_1 = 20; y_2 = 40; y_3 = 60; y_4; y_5 = 100]$, with $y_4 \in \{80, 81, ..., 100\}$. Thus, the binomial tests consider only the values $P(Y = 4)$ *vs*. $p_Z(4)$ and $P(Y = 5)$ *vs*. $p_Z(5)$.

The investigated results are presented in Figure V-4. The horizontal axis gives the y_4 values. The vertical axis gives the number of binomial tests which did not pass, represented by bars: the low (blue) bar corresponds to $i = 4$ (*Good)* while the top (pink) bar corresponds to $i = 5$ (*Excellent*). It can be noticed that the differences between Y and Z are minimal when $y_4 \in \{86, 87, 88\}$; actually, this time, the differences are at the limit of the statistical significance. Note that the binomial test is applied at $\alpha = 0.05$ and that it is repeated 48 times, for two type of tests; hence, 2 or even 3 failed tests cannot be considered as a proof of difference between two probabilities.

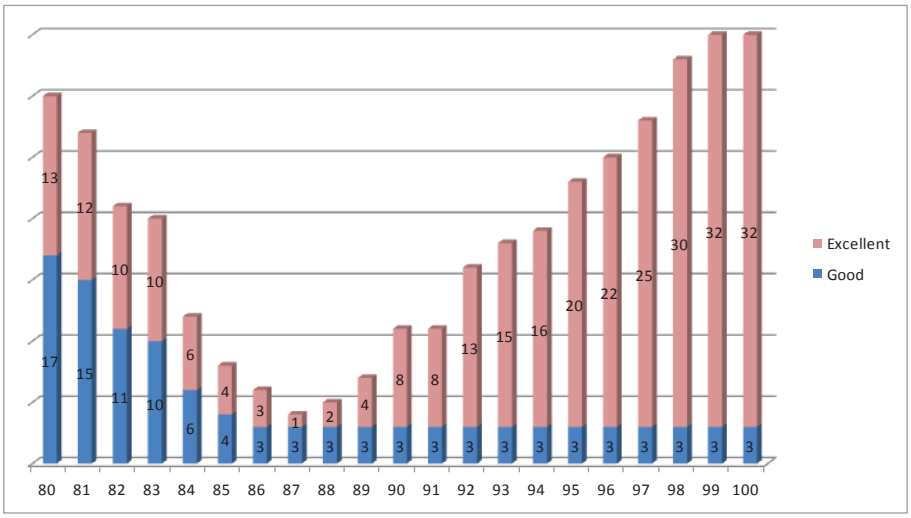
Figure V-4 shows that some psycho-cognitive mechanisms make the observers associate to *Excellent* an interval $(86; 100]$, $(87; 100]$ or $(88; 100]$, instead of an interval $(80; 100]$ that would have been related to a discrete scale with no semantic labels. In other words, the observers are reluctant in assigning the label *Excellent* and prone to assign the label *Good*.

A measure of this phenomenon can be the coefficient ρ , defined in Chapter V.2.2 as the ratio of the actual size of the interval on which the labels are assigned to the size of an unlabeled scale interval.

In the experiments, q=5, M=100 and $y_4 \in \{87,88\}$; hence $\rho_5 \in \{0.65 ; 0.6\}$.

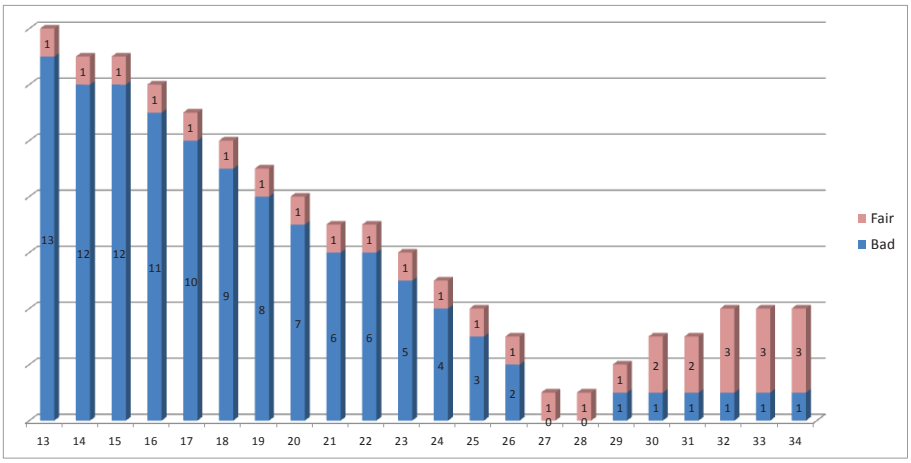


(a) q = 3 quality levels

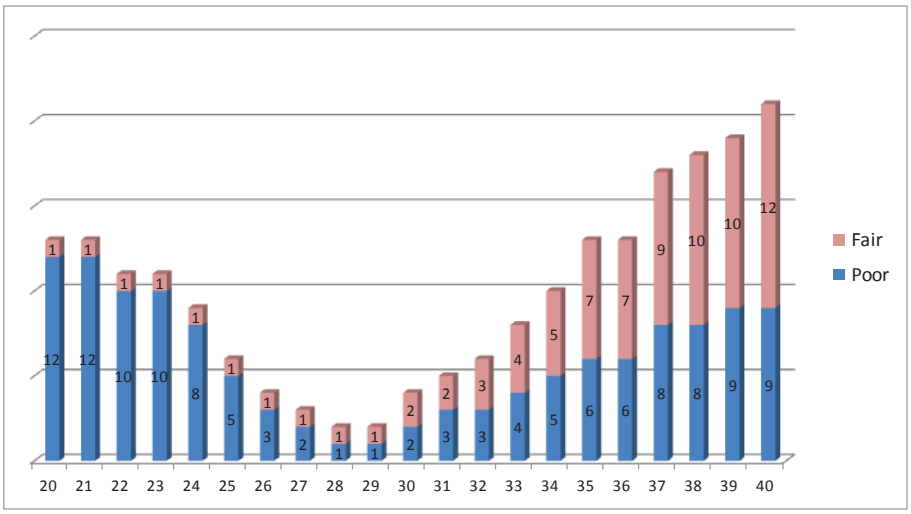


(b) q = 5 quality levels

Figure V-7: The number of failed binomial tests when assessing high quality stereoscopic video content: (a) *P(Y=2) vs.*$p_z(2)$ (blue bottom bar) and *P(Y=3) vs.*$p_z(3)$ (top pink bar) for q = 3; (b) *P(Y=4) vs.*$p_z(4)$ (blue bottom bar) and *P(Y=5) vs.*$p_z(5)$ (top pink bar) for q = 5.

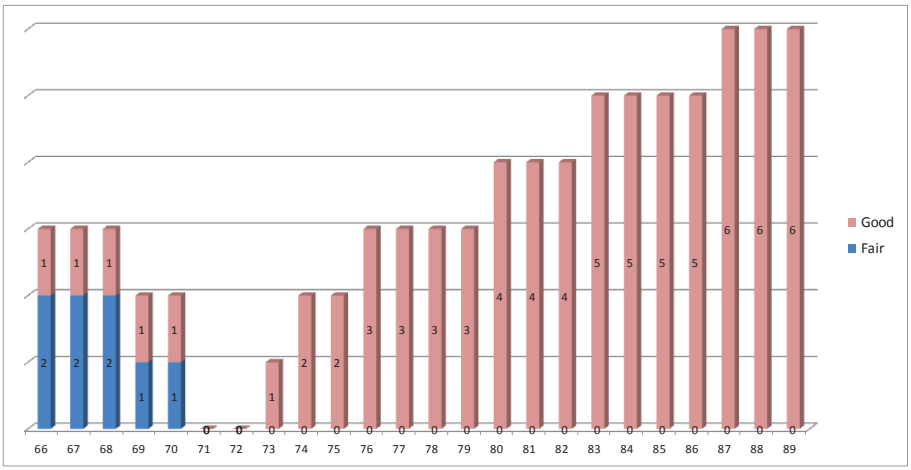


(a) q = 3 quality levels

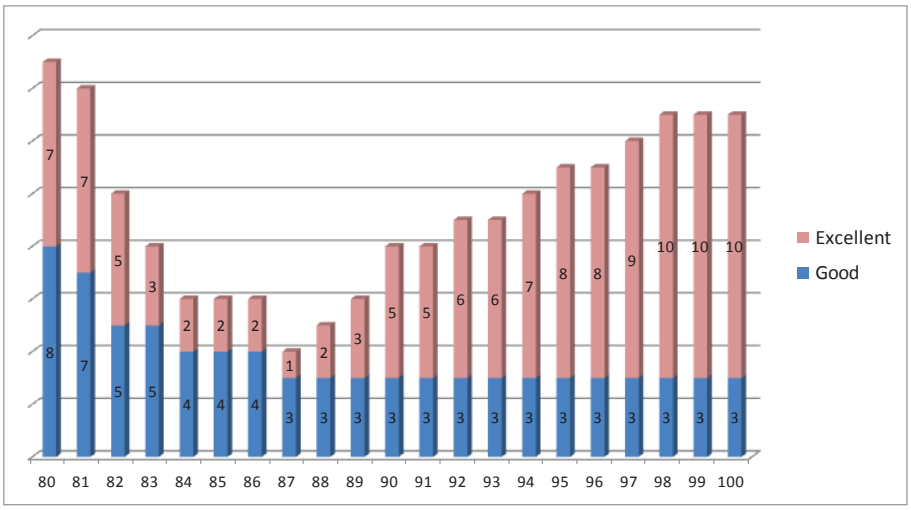


(b) q = 5 quality levels

Figure V-8: The number of failed binomial tests when assessing low quality stereoscopic video content: (a) *P(Y=1) vs.*$p_z(1)$ (blue bottom bar) and *P(Y=2) vs.*$p_z(2)$ (top pink bar) for q = 3; (b) *P(Y=2) vs.*$p_z(2)$ (blue bottom bar) and *P(Y=3) vs.*$p_z(3)$ (top pink bar) for q = 5.

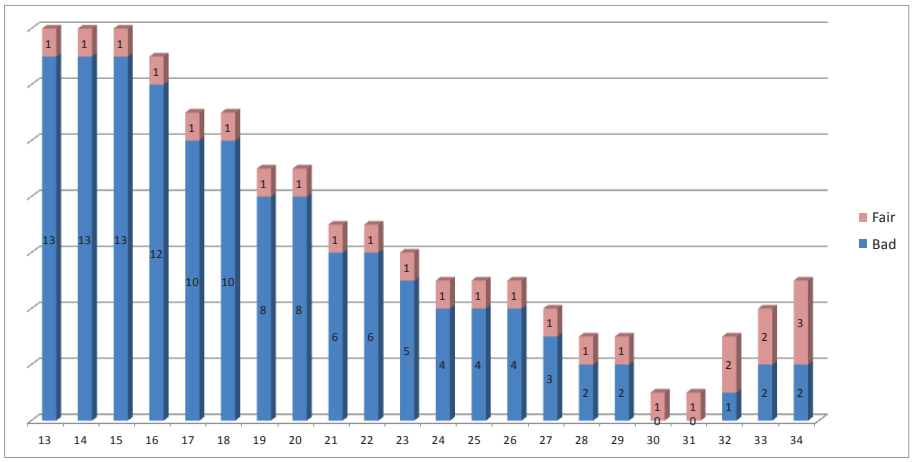


(a) q = 3 quality levels

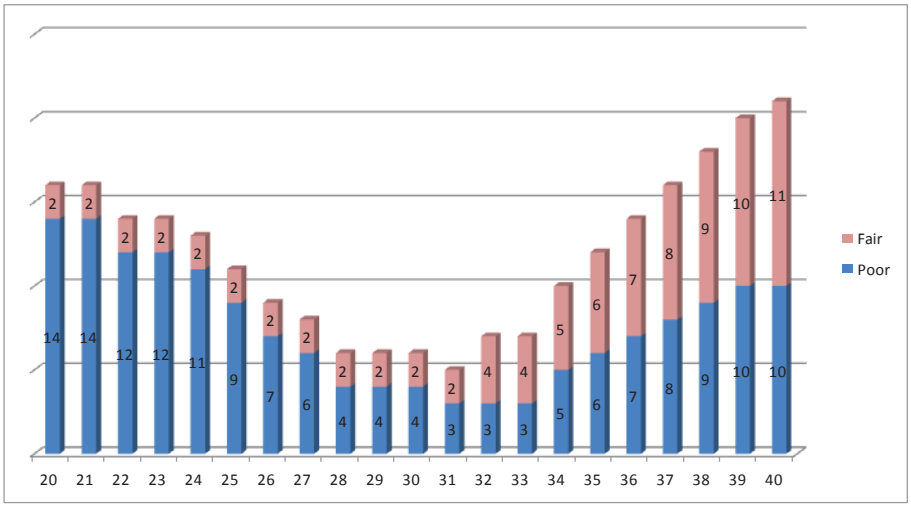


(b) q = 5 quality levels

Figure V-9: The number of failed binomial tests when assessing high quality 2D video content: (a) *P(Y=2) vs.*$p_z(2)$ (blue bottom bar) and *P(Y=3) vs.* $p_z(3)$ (top pink bar) for q = 3; (b) *P(Y=4) vs.*$p_z(4)$ (blue bottom bar) and *P(Y=5) vs.*$p_z(5)$ (top pink bar) for q = 5.



(a) q = 3 quality levels



(b) q = 5 quality levels

Figure V-10: The number of failed binomial tests when assessing low quality 2D video content: (a) *P(Y=1) vs.* $p_z(1)$ (blue bottom bar) and *P(Y=2) vs.* $p_z(2)$ (top pink bar) for q = 3; (b) *P(Y=2) vs.* $p_z(2)$ (blue bottom bar) and *P(Y=3) vs.* $p_z(3)$ (top pink bar) for q = 5.

The conclusion from Figure V-7–V-10 is presented in Table V-1 and V-2. These two tables present the ρ values, for each of the 4 types of the investigated content and for *q*=3 and *q*=5, respectively.

Table V-1: Quantifying the semantic impact for *q=3*

		limits	ρ
High quality stereoscopic video content	*Bad*	[0..34)	$\rho_1 = 1$
	Fair	[34..72)	$\rho_2 = 1.15$
		[34..73)	$\rho_2 = 1.18$
	Good	[72..100]	$\rho_3 = 0.85$
		[73..100]	$\rho_3 = 0.82$
Low quality stereoscopic video content	*Bad*	[0..27)	$\rho_1 = 0.82$
		[0..28)	$\rho_1 = 0.85$
	Fair	[27..67)	$\rho_2 = 1.18$
		[28..67)	$\rho_2 = 1.15$
	Good	[67..100]	$\rho_3 = 1$
High quality 2D video content	*Bad*	[0..34)	$\rho_1 = 1$
	Fair	[34..71)	$\rho_2 = 1.12$
		[34..72)	$\rho_2 = 1.15$
	Good	[71..100]	$\rho_3 = 0.88$
		[72..100]	$\rho_3 = 0.85$
Low quality 2D video content	*Bad*	[0..30)	$\rho_1 = 0.9$
		[0..31)	$\rho_1 = 0.93$
	Fair	[30..67)	$\rho_2 = 1.1$
		[31..67)	$\rho_2 = 1.07$
	Good	[67..100]	$\rho_3 = 1$

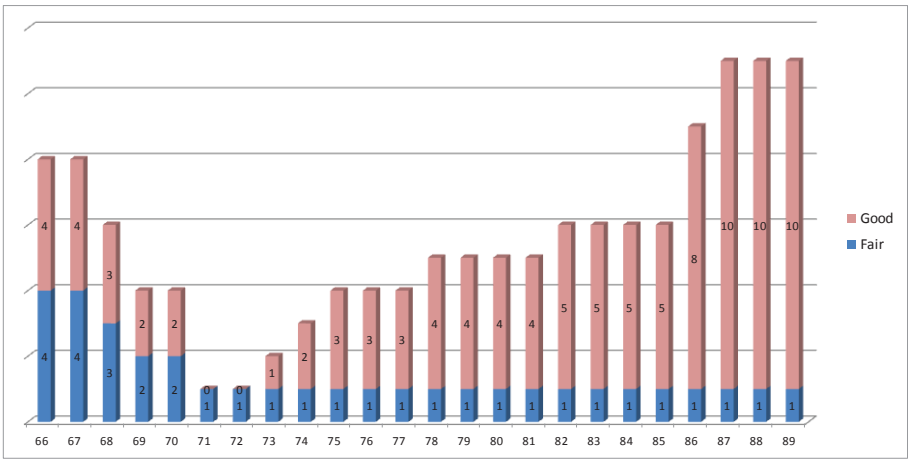
Table V-2: Quantifying the semantic impact for *q=5*

		limits	ρ
High quality stereoscopic video content	*Bad*	[0..20)	$\rho_1 = 1$
	Poor	[20..40)	$\rho_2 = 1$
	Fair	[40..60)	$\rho_3 = 1$
	Good	[60..87)	$\rho_4 = 1.35$
		[60..88)	$\rho_4 = 1.4$
	Excellent	[87..100]	$\rho_5 = 0.65$
		[88..100]	$\rho_5 = 0.6$
Low quality stereoscopic video content	*Bad*	[0..20)	$\rho_1 = 1$
	Poor	[20..29)	$\rho_2 = 0.45$
		[20..30)	$\rho_2 = 0.5$
	Fair	[29..60)	$\rho_3 = 1.55$
		[30..60)	$\rho_3 = 1.5$
	Good	[60..80)	$\rho_4 = 1$
	Excellent	[80..100]	$\rho_5 = 1$
High quality 2D video content	*Bad*	[0..20)	$\rho_1 = 1$
	Poor	[20..38)	$\rho_2 = 0.9$
	Fair	[38..60)	$\rho_3 = 1.1$
	Good	[60..87)	$\rho_4 = 1.35$
		[60..88)	$\rho_4 = 1.4$
	Excellent	[87..100]	$\rho_5 = 0.65$
		[88..100]	$\rho_5 = 0.6$
Low quality 2D video content	*Bad*	[0..20)	$\rho_1 = 1$
	Poor	[20..31)	$\rho_2 = 0.55$
	Fair	[31..60)	$\rho_3 = 1.45$
	Good	[60..83)	$\rho_4 = 1.15$
	Excellent	[83..100]	$\rho_5 = 0.85$

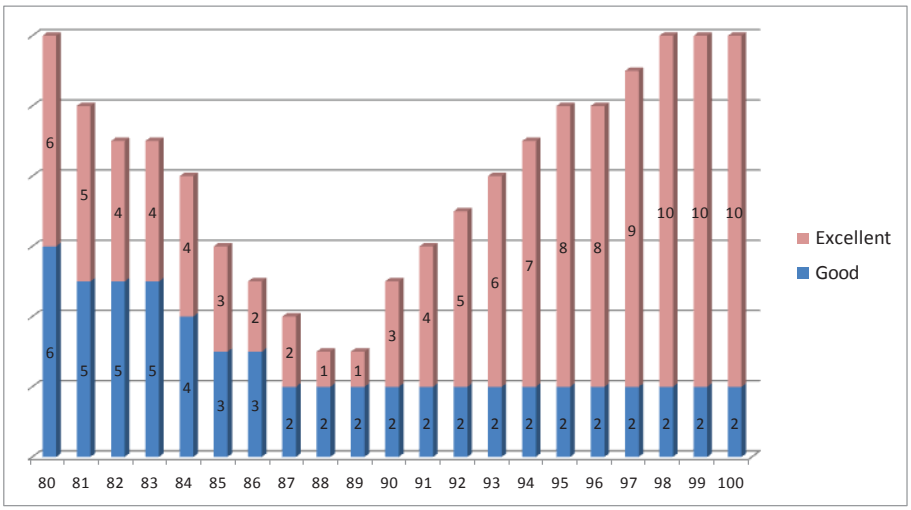
V.2.3.3. The generality of the results

Of course, as in any statistical study, one key issue is the generality of the results. Consequently, the experiments reported in Chapter V.2.3.3 are resumed: the *reference* panels is kept unchanged but the *validation* panels are replaced by new panels of 25 observers each (the so-called *cross-checking panels,* Chapter III.5). All the other experimental conditions are kept unchanged with respect to Chapter V.2.3.2.

The new results are presented in Tables V.3 - V.4 and Figures V-11 – V-14, which are organized the same way as Tables V.1 – V.2 and Figures V-11 – V-14.

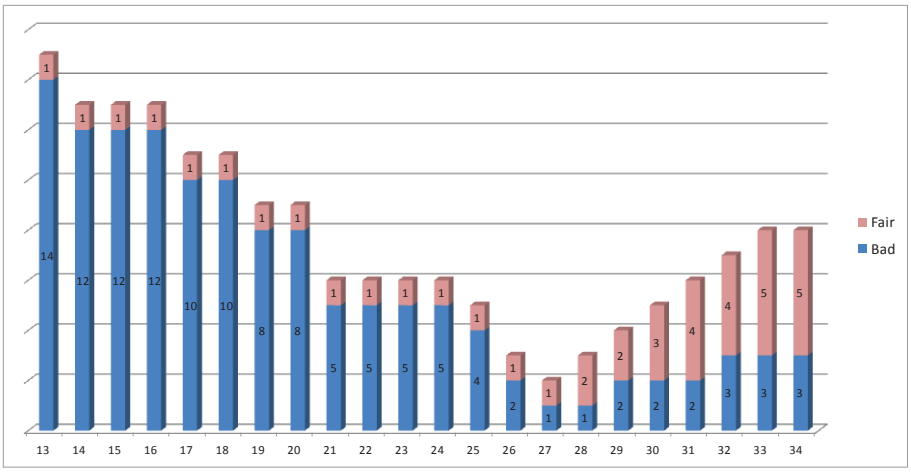


(a) q = 3 quality levels

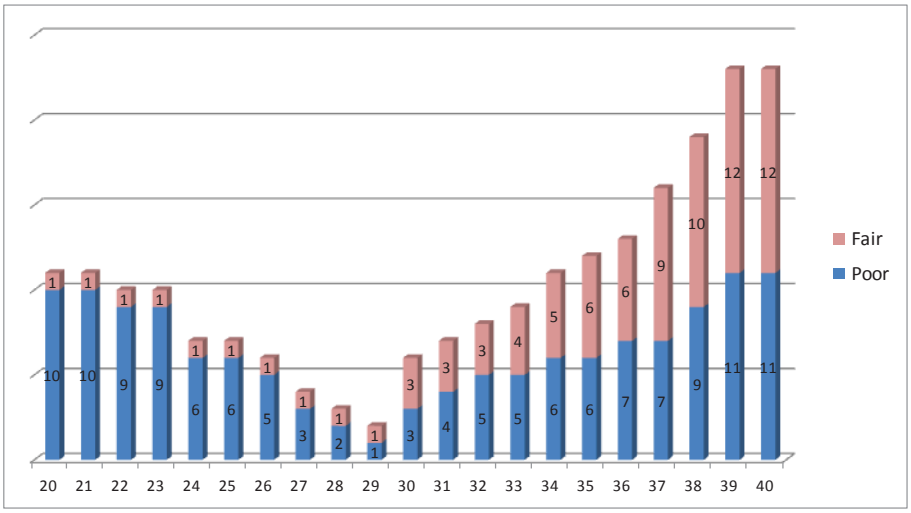


(b) q = 5 quality levels

Figure V-11: The number of failed binomial tests when assessing high quality stereoscopic video content: (a) *P(Y=2) vs.* $p_z(2)$ (blue bottom bar) and *P(Y=3) vs.* $p_z(3)$ (top pink bar) for q = 3; (b) f *P(Y=4) vs.* $p_z(4)$ (blue bottom bar) and *P(Y=5) vs.* $p_z(5)$ (top pink bar) for q = 5.

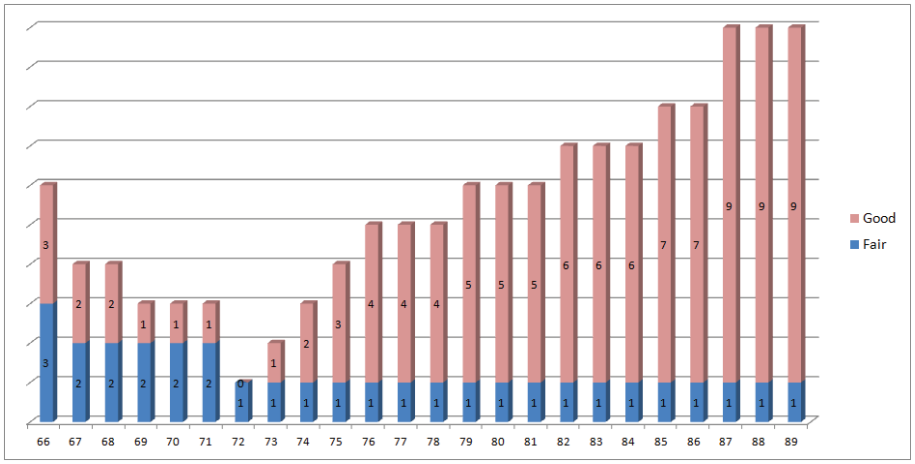


(a) q = 3 quality levels

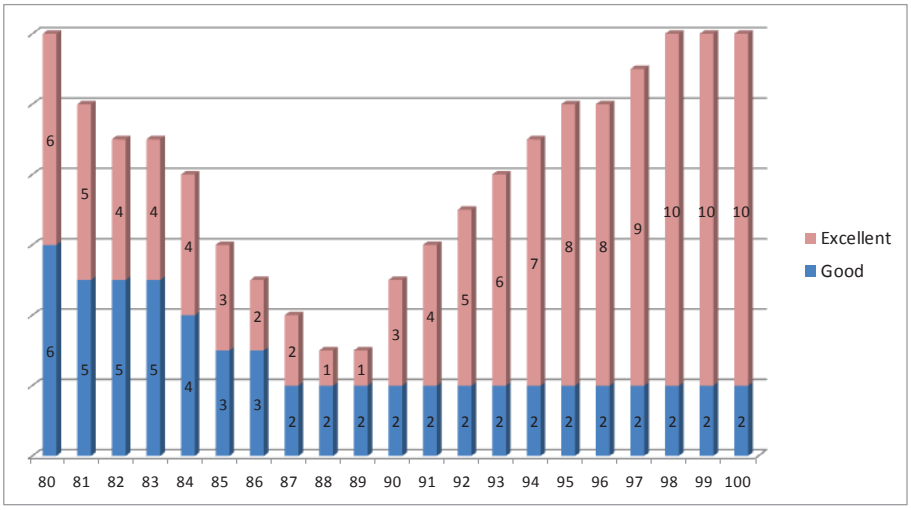


(b) q = 5 quality levels

Figure V-12: The number of failed binomial tests when assessing low quality stereoscopic video content: (a) *P(Y=1) vs.* $p_z(1)$ (blue bottom bar) and *P(Y=2) vs.* $p_z(2)$ (top pink bar) for q = 3; (b) *P(Y=2) vs.* $p_z(2)$ (blue bottom bar) and *P(Y=3) vs.*$p_z(3)$ (top pink bar) for q = 5.

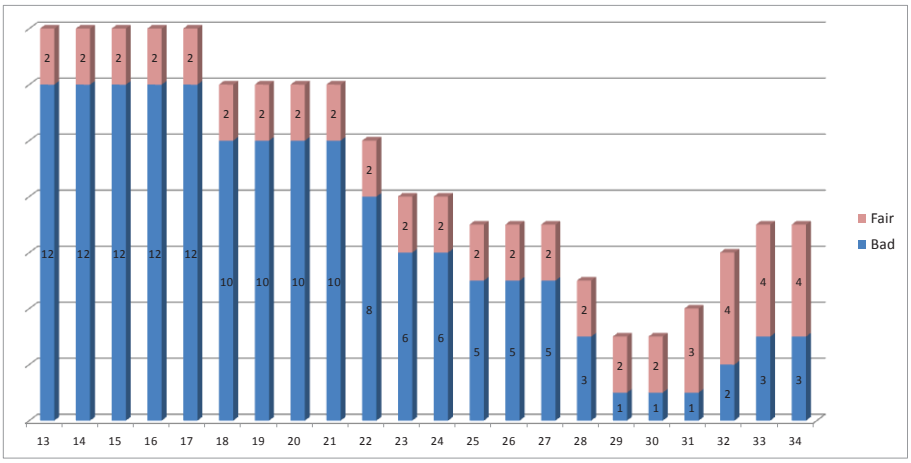


(a) q = 3 quality levels

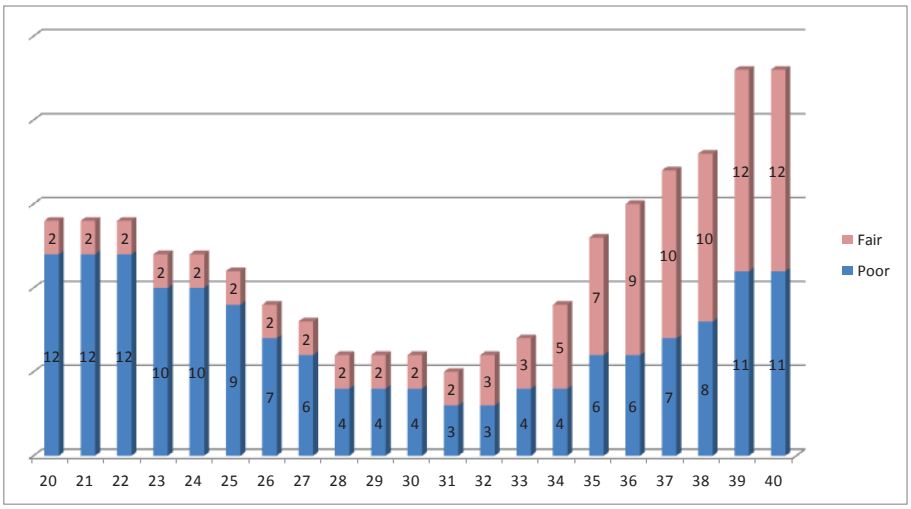


(b) q = 5 quality levels

Figure V-13: The number of failed binomial tests when assessing high quality 2D video content: (a) *P(Y=2) vs.* $p_z(2)$ (blue bottom bar) and *P(Y=3) vs.* $p_z(3)$ (top pink bar) for q = 3; (b) *P(Y=4) vs.* $p_z(4)$ (blue bottom bar) and *P(Y=5) vs.* $p_z(5)$ (top pink bar) for q = 5.



(a) q = 3 quality levels



(b) q = 5 quality levels

Figure V-14: The number of failed binomial tests when assessing low quality 2D video content: (a) *P(Y=1) vs.* $p_z(1)$ (blue bottom bar) and *P(Y=2) vs.* $p_z(2)$ (top pink bar) for q = 3; (b) *P(Y=2) vs.* $p_z(2)$ (blue bottom bar) and *P(Y=3) vs.* $p_z(3)$ (top pink bar) for q = 5.

Table V-3: The generality of the quantization of the semantic impact for *q=3*

		limits	ρ
High quality stereoscopic video content	*Bad*	[0..34)	$\rho_1 = 1$
	Fair	[34..71)	$\rho_2 = 1.12$
		[34..72)	$\rho_2 = 1.15$
	Good	[71..100]	$\rho_3 = 0.88$
		[72..100]	$\rho_3 = 0.85$
Low quality stereoscopic video content	*Bad*	[0..27)	$\rho_1 = 0.82$
	Fair	[27..67)	$\rho_2 = 1.18$
	Good	[67..100]	$\rho_3 = 1$
High quality 2D video content	*Bad*	[0..34)	$\rho_1 = 1$
	Fair	[34..72)	$\rho_2 = 1.15$
	Good	[72..100]	$\rho_3 = 0.85$
Low quality 2D video content	*Bad*	[0..29)	$\rho_1 = 0.88$
		[0..30)	$\rho_1 = 0.91$
	Fair	[29..67)	$\rho_2 = 1.12$
		[30..67)	$\rho_2 = 1.09$
	Good	[67..100]	$\rho_3 = 1$

Table V-4: The generality of the quantization of the semantic impact for *q=5*

		limits	ρ
High quality stereoscopic video content	*Bad*	[0..20)	$\rho_1 = 1$
	Poor	[20..40)	$\rho_2 = 1$
	Fair	[40..60)	$\rho_3 = 1$
	Good	[60..88)	$\rho_4 = 1.4$
		[60..89)	$\rho_4 = 1.45$
	Excellent	[88..100]	$\rho_5 = 0.6$
		[89..100]	$\rho_5 = 0.55$
Low quality stereoscopic video content	*Bad*	[0..20)	$\rho_1 = 1$
	Poor	[20..29)	$\rho_2 = 0.45$
	Fair	[29..60)	$\rho_3 = 1.55$
	Good	[60..80)	$\rho_4 = 1$
	Excellent	[80..100]	$\rho_5 = 1$
High quality 2D video content	*Bad*	[0..20)	$\rho_1 = 1$
	Poor	[20..38)	$\rho_2 = 0.9$
	Fair	[38..60)	$\rho_3 = 1.1$
	Good	[60..88)	$\rho_4 = 1.4$
		[60..89)	$\rho_4 = 1.45$
	Excellent	[88..100]	$\rho_5 = 0.6$
		[89..100]	$\rho_5 = 0.55$
Low quality 2D video content	*Bad*	[0..20)	$\rho_1 = 1$
	Poor	[20..31)	$\rho_2 = 0.55$
	Fair	[31..60)	$\rho_3 = 1.45$
	Good	[60..82)	$\rho_4 = 1.1$
	Excellent	[82..100)	$\rho_5 = 0.9$

V.2.3.4. Reference values for the semantic impact coefficient

When comparing the results presented in Chapter V.2.3.3 to the ones presented in Chapter V.2.3.2, it can be noticed that small differences exist; yet, each and every time common values exist in the corresponding cells in Tables V.1 – V.2 and V.3 – V.4.

This demonstrates that the study carried out in this thesis allows the computation of some *reference* value for the semantic impact coefficient, see Tables V.5 and V.6. The word *reference* means here that such values are validated by all the three types of panels organized in the study.

Table V-5: Reference values for the semantic impact for *q=3*

		limits	ρ
High quality stereoscopic video content	*Bad*	[0..33)	$\rho_1 = 1$
	Fair	[34..72)	$\rho_2 = 1.15$
	Good	[72..100]	$\rho_3 = 0.85$
Low quality stereoscopic video content	*Bad*	[0..27)	$\rho_1 = 0.82$
	Fair	[27..67)	$\rho_2 = 1.18$
	Good	[67..100]	$\rho_3 = 1$
High quality 2D video content	*Bad*	[0..33]	$\rho_1 = 1$
	Fair	[34..72)	$\rho_2 = 1.15$
	Good	[72..100]	$\rho_3 = 0.85$
Low quality 2D video content	*Bad*	[0..30)	$\rho_1 = 0.91$
	Fair	[30..67)	$\rho_2 = 1.09$
	Good	[67..100]	$\rho_3 = 1$

Table V-6: Reference values for the semantic impact for *q=5*

		limits	ρ
High quality stereoscopic video content	*Bad*	[0..19]	$\rho_1 = 1$
	Poor	[20..40)	$\rho_2 = 1$
	Fair	[40..60)	$\rho_3 = 1$
	Good	[60..88)	$\rho_4 = 1.4$
	Excellent	[88..100]	$\rho_5 = 0.6$
Low quality stereoscopic video content	*Bad*	[0..20)	$\rho_1 = 1$
	Poor	[20..29)	$\rho_2 = 0.45$
	Fair	[29..60)	$\rho_3 = 1.55$
	Good	[60..80)	$\rho_4 = 1$
	Excellent	[80..100]	$\rho_5 = 1$
High quality 2D video content	*Bad*	[0..20)	$\rho_1 = 1$
	Poor	[20..37)	$\rho_2 = 0.85$
	Fair	[37..60)	$\rho_3 = 0.15$
	Good	[60..88)	$\rho_4 = 1.4$
	Excellent	[88..100]	$\rho_5 = 0.6$
Low quality 2D video content	*Bad*	[0..20)	$\rho_1 = 1$
	Poor	[20..31)	$\rho_2 = 0.55$
	Fair	[31..60)	$\rho_3 = 1.45$
	Good	[60..83)	$\rho_4 = 1.1$
	Excellent	[83..100)	$\rho_5 = 0.9$

V.3. Conclusion

The present chapter addresses the second main challenge of the thesis, namely the would-be semantic impact of the ITU labels in the evolution scores: the existence of the semantic impact is first investigated, then the possibility of its assessment is addressed.

The existence of the semantic impact is verified (Chapter V.1) by a comparison (based in the Student's paired test) between the average values (representing the *MOS*) corresponding to the continuous (unlabeled) and discrete, semantically labeled scales. The experiments bring to light that such an semantic impact exist for all the investigated cases, *i.e.* for the 4 types of content and for the two numbers of quality levels on the discrete scale (q = 3 and q = 5). Yet, this is only a preliminary result and does not bring any in-depth information about where this impact comes from.

The main contribution of the chapter is the definition a methodological framework for quantifying the semantic impact. The investigation method is composed of three main steps. First, by using the results presented in Chapter IV, the Y *r.v*. modeling the scores assigned by the observers on an discrete unlabeled scale is computed. Secondly, the Z *r.v*. modeling the scores assigned by the observers on a discrete labeled scale is estimated. Finally, the assessment of the semantic impact is achieved by defining a coefficient (denoted by ρ) expressing the differences between the Y and Z.

The experimental results have two-folded outcomes. First, reference values for the semantic impact (i.e. general values, independent with respect to the observers) are computed for each type of content (stereoscopic and 2D video, high and low quality) and for the two investigated numbers of quality levels (3 and 5). Secondly, the fact that practically all the labels (and not only *Excellent*) are involved in the overall semantic impact is demonstrated. Of course as expected, the extent to which each label impact the overall result, depend on the type of content.

VI. Conclusion and Future work

Quality evaluation is an ever-fascinating field, covering at least a century of research works emerging from psychology, psychophysics, sociology, marketing, medicine, ...

While for visual quality evaluation, the ITU Recommendations pave the way towards well-configured, consensual evaluation conditions granting reproducibility and comparability of the experimental results, an in-depth analysis of the state-of-the-art studies shows at least three open challenges, related to the: (1) the continuous *vs*. discrete evaluation scales, (2) the statistical distribution of the scores assigned by the observers and (3) the usage of semantic labels on the grading scales. Thus, the present thesis turns these challenges into three research objectives:

1. bridging at the theoretical level the continuous and the discrete scale evaluation procedures and investigating whether the number of the classes on the discrete scales is a criterion meaningful in the results interpretations or just a parameter; studying the theoretical influence of the statistical model of the evolution results and of the size of the panel (number of observers) in the accuracy of the results are also targeted;
2. quantifying the bias induced in subjective video quality experiments by the semantic labels (e.g. *Excellent*, *Good*, *Fair*, *Poor*, and *Bad*) generally associated to the discrete grading scales;
3. designing and deploying and experimental test-bed able to support their precision and statistical relevance.

With respect to these objectives, the main thesis contributions (detailed in the conclusive sections included in each chapter) are at theoretical, methodological and experimental levels.

First, at the theoretical level, the continuous and the discrete scale evaluation procedures are bridged by a non-linear random variable transformation formula. The instantiations of this formula for two opposite cases of practical relevance, namely the case in which the scores according on a continuous scales are Gaussian distributed and the case in which these scores are not Gaussian (but estimated *via* a Gaussian mixture) brings to light that the *MOS* and the related confidence limits solely depend on the average and variances of the continuous scale models and of the q number of quality levels on the discrete scale.

Secondly, at the methodological level, the main contribution consists in the definition of a methodological framework for quantifying the semantic impact induced by the ITU semantic labels. In this respect, an auxiliary discrete random variable is defined: it is characterized by uneven partition but by equal a posteriori probabilities. The semantic impact is quantified by assessing the differences in the partition class length between this auxiliary random variable and the random variable corresponding to the semantically labeled scale. The experimental results not only identify *reference* values for the semantic impact, according to the type of content (stereoscopic and 2D video, high and low quality) and to the two investigated numbers of quality levels (3 and 5) but also bring to light that practically all the labels (and not only *Excellent*) are involved in the overall semantic impact.

A test-bed designed and deployed for investigating the evaluation procedure *per-se* is also a result of the thesis. It considers a total of 640 human observers, about 80 minutes of video content, and scores assigned on three types of scales (continuous, 5 level labeled scale and 3 level labeled scale).

Hence, as a global conclusion, in order to converge to a unique answer to the controversial issue related to the number of quality levels on a stereoscopic video grading scale, the thesis suggests to perform the evaluation on a continuous grading scale, with no semantic labels associated to the scores and to subsequently map these values on the discrete scales. The ITU labels *Excellent*, *Good*, *Fair*, *Poor*, and *Bad* should be avoided or, at least, during the result interpretation, it should be kept in mind that semantic impacts as large as 0.55 or 0.6 (while a non-semantic behavior would have been denoted by a value of 1) can bias the results.

The future work will be structured at three levels.

First, a richer interpretation of the scores already gathered thanks to the experiments carried out in the thesis is targeted, as for examples studies on: inter-gender variability, on the correlation among the scores corresponding to the two views (stereoscopic content) *vs*. a single view, or on the dependency of the semantic impact coefficient with the *MOS*.

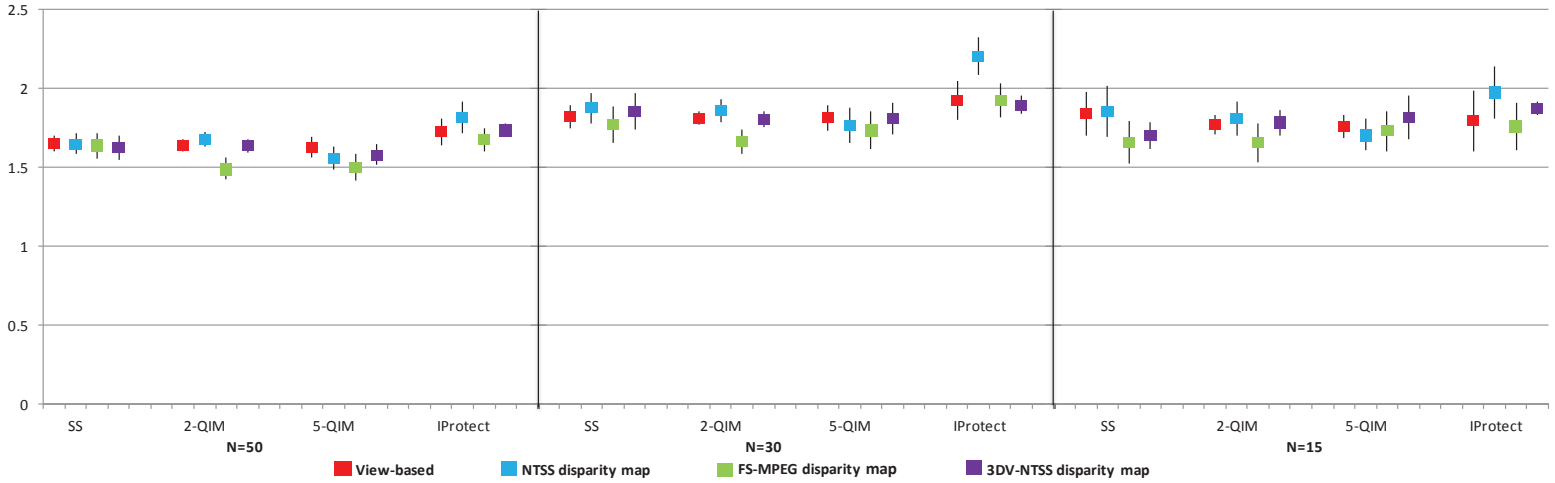
Work will be also devoted to provide the video corpus and the scoring application as open research resources. This way, the research community can capitalize on our results and bring additional results related to various aspects, from language dependability to multimodal investigations.

Finally, the theoretical results presented in Chapter IV and methodological framework in Chapter V.2 will be reverse investigated in order to find out how *correction* formulae for the semantic impact can be defined. It should be emphasized that the It should be emphasized that while the semantic impact does depend with the type of content (stereoscopic or 2D video) and its quality, it is always important. Such a result will emphasis the thesis pragmatic relevance and will turn it into a support for day-by-day subjective quality evaluation.

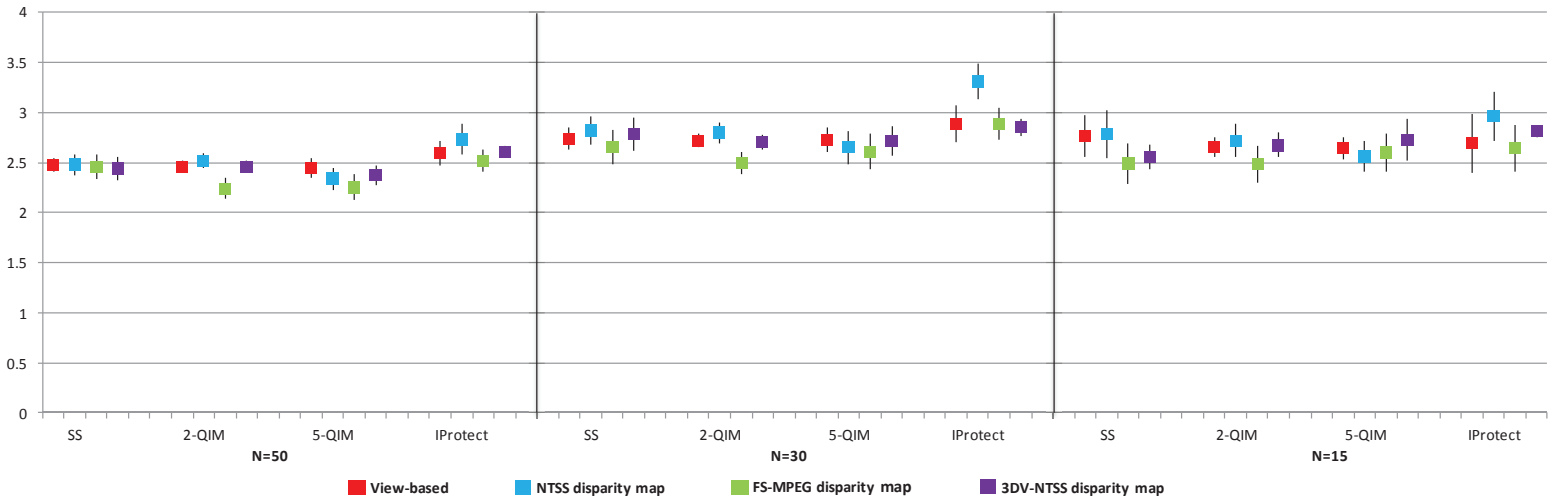
VII. Appendixes

VII.1. Appendix A Discrete unlabeled scale simulation (Gaussian distribution)

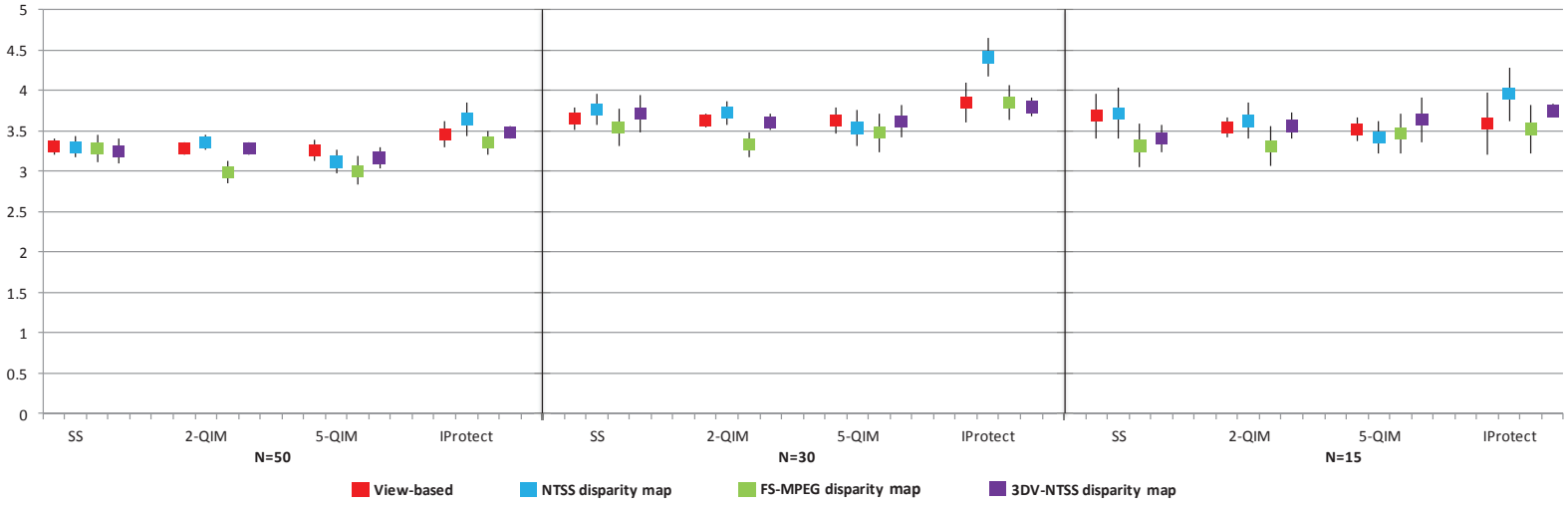
High quality stereoscopic video content (Image Quality)



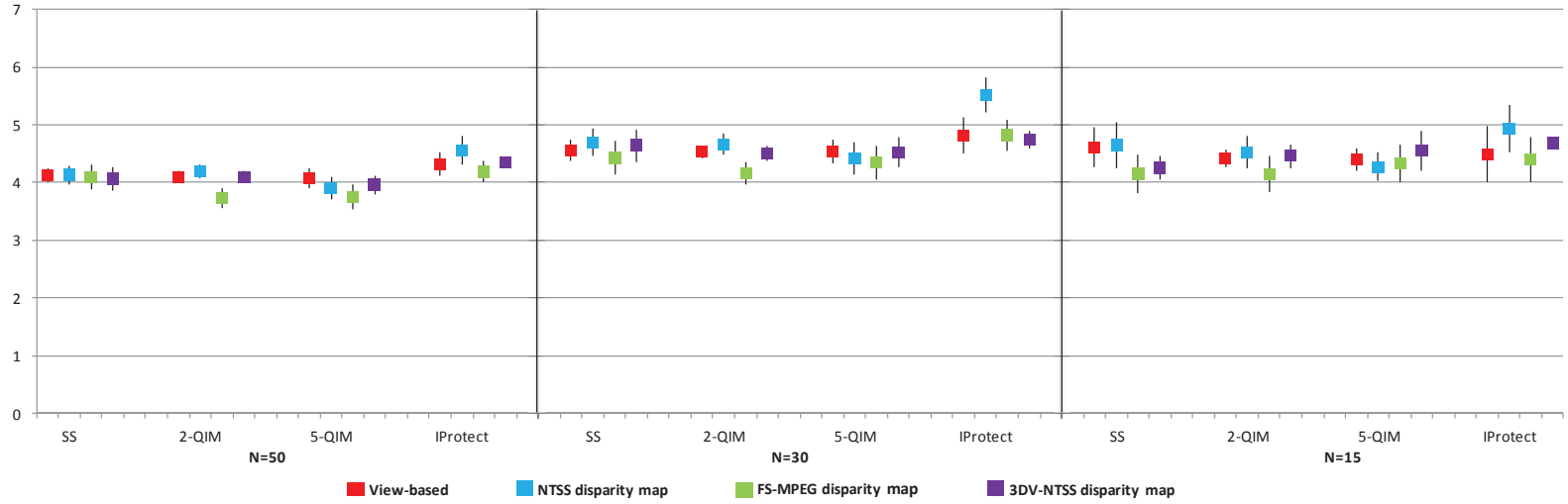
(a)



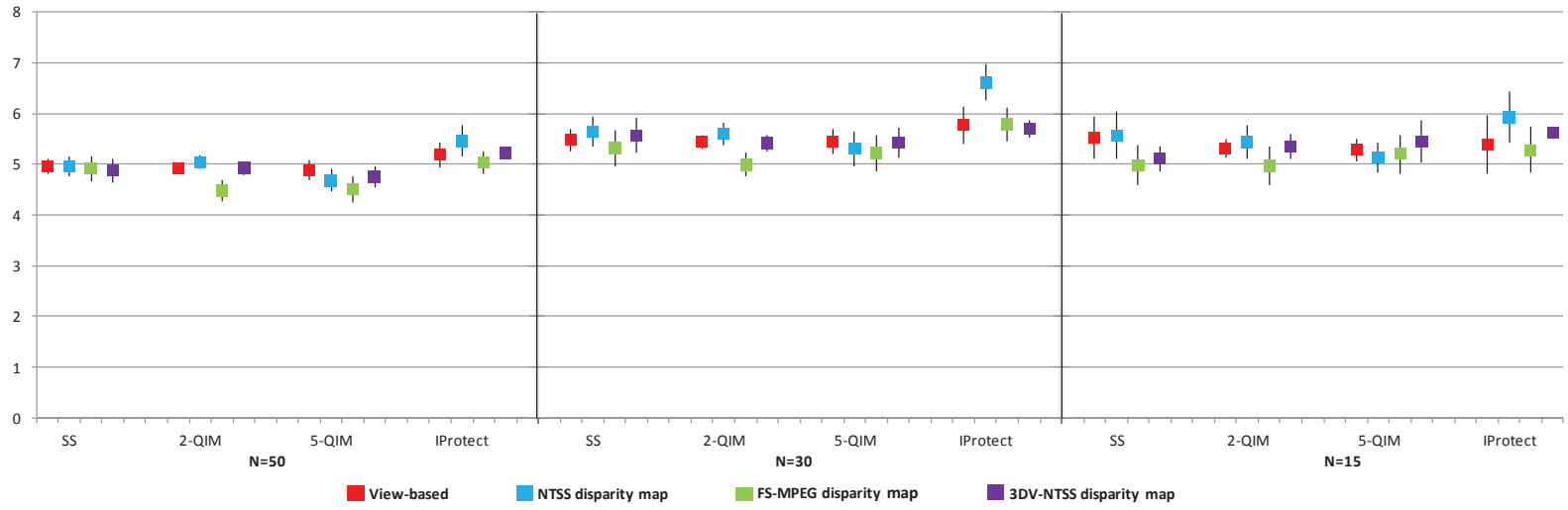
(b)


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

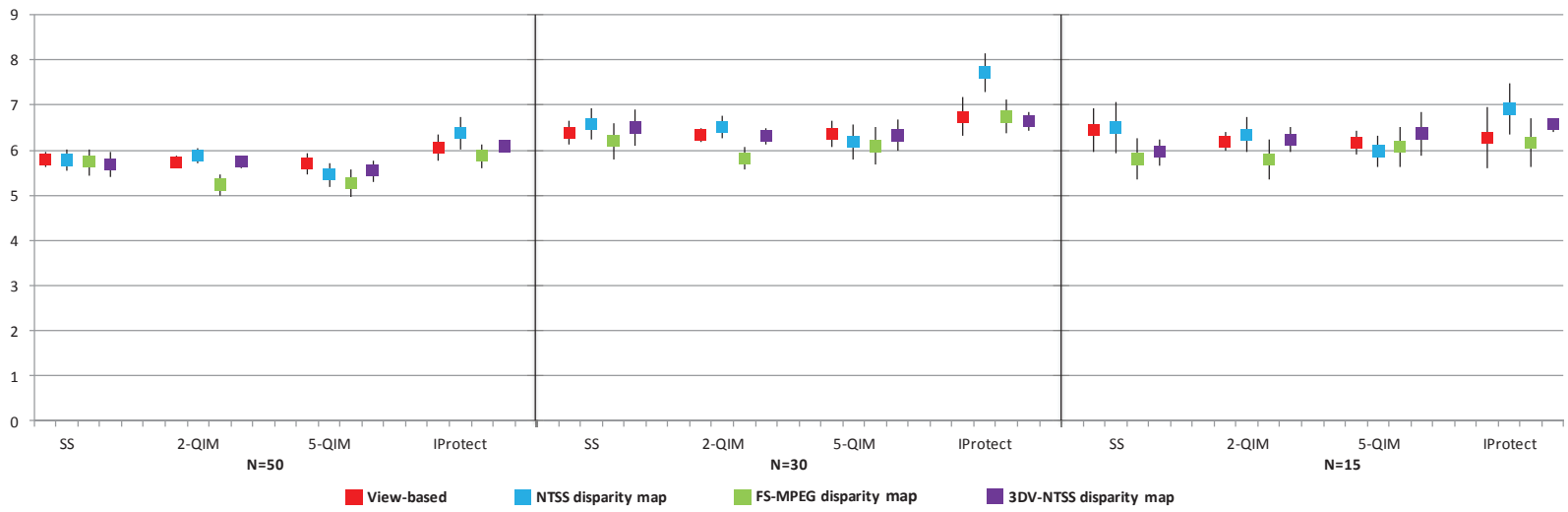
(c)


7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

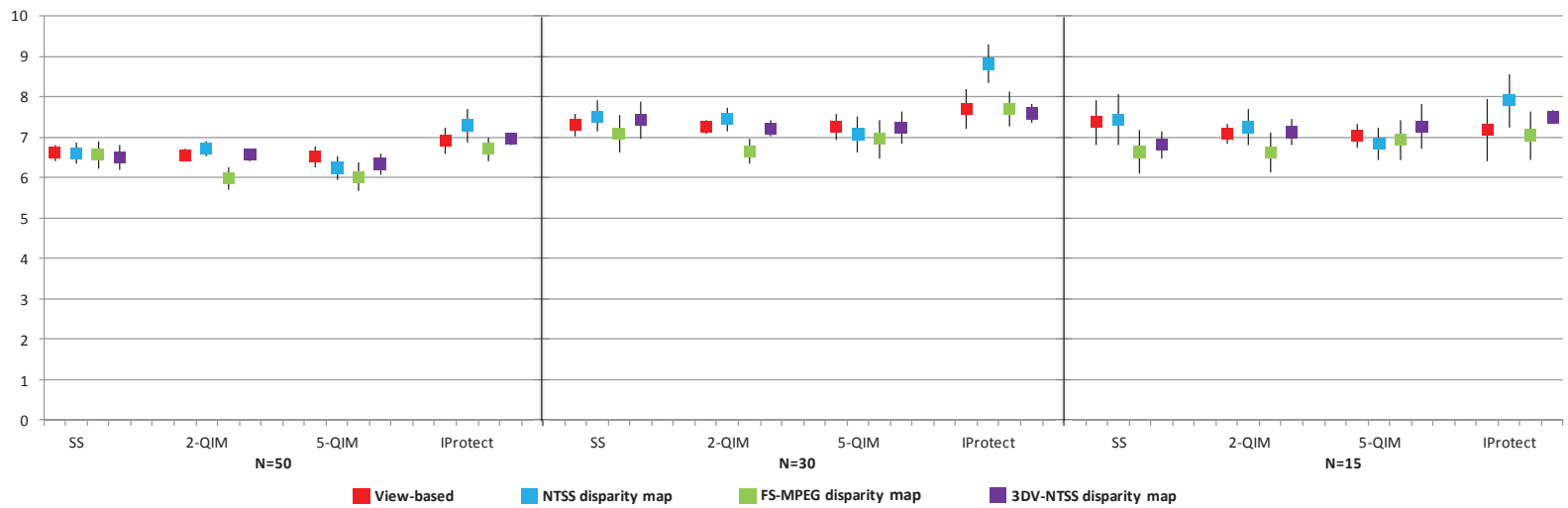
(d)


8
7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

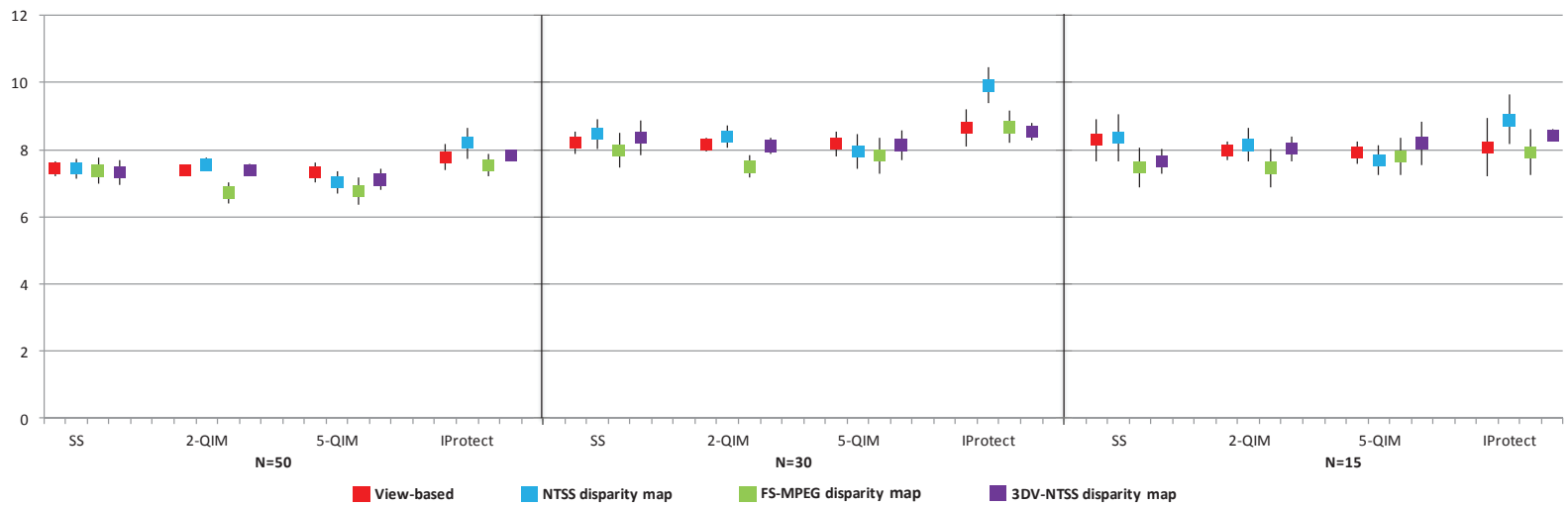
(e)


9
8
7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)

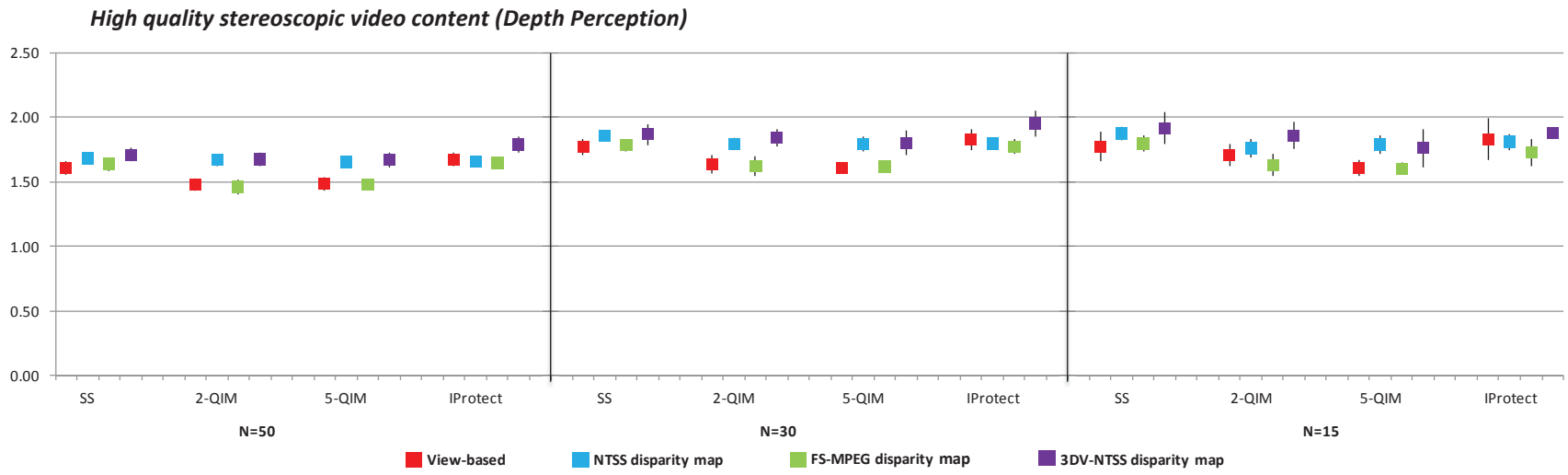


(g)

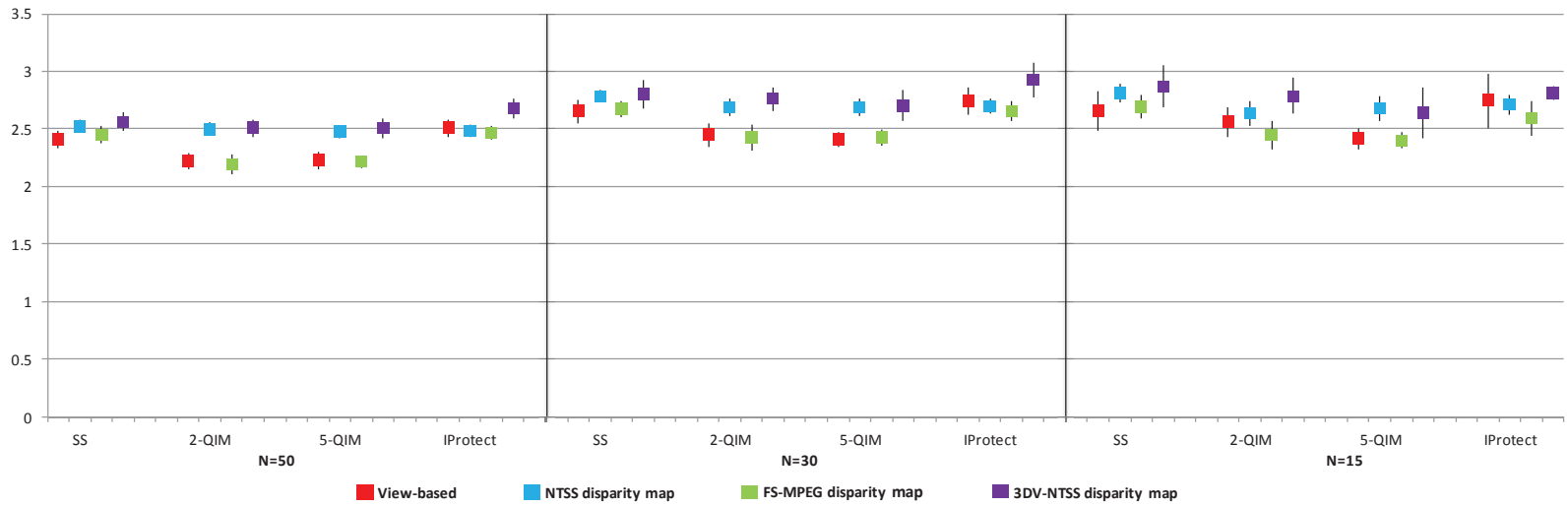


(h)

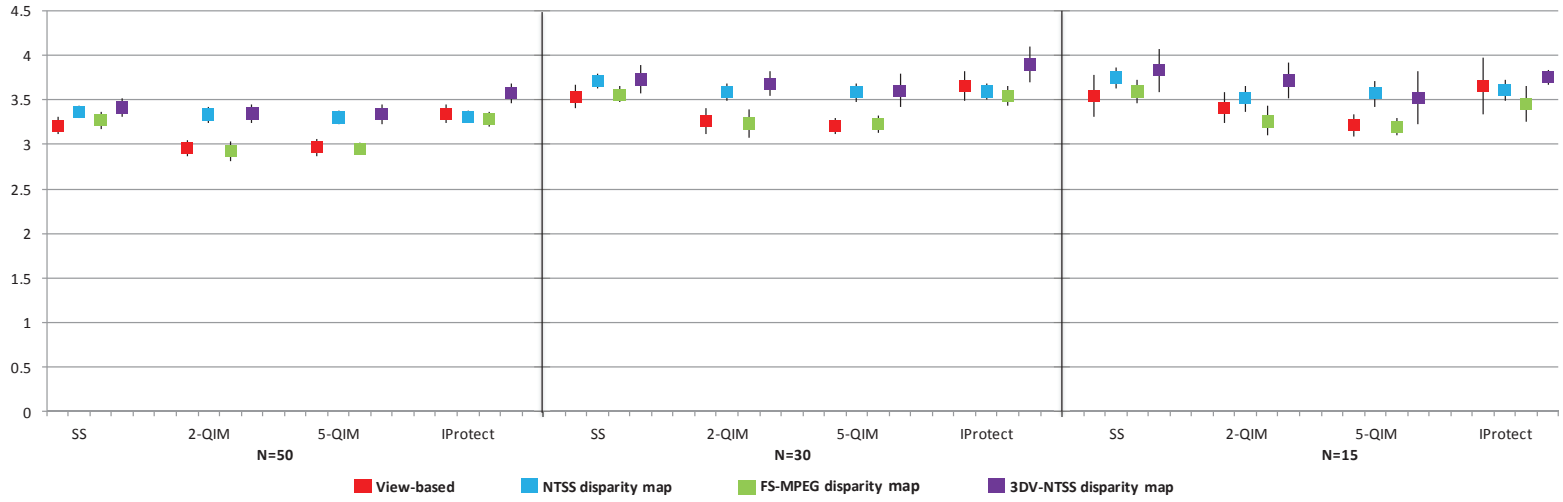
Figure A1-1: Subjective evaluations for high-quality stereoscopic video content (Image Quality), for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*


High quality stereoscopic video content (Depth Perception)
2.50
2.00
1.50
1.00
0.50
0.00
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

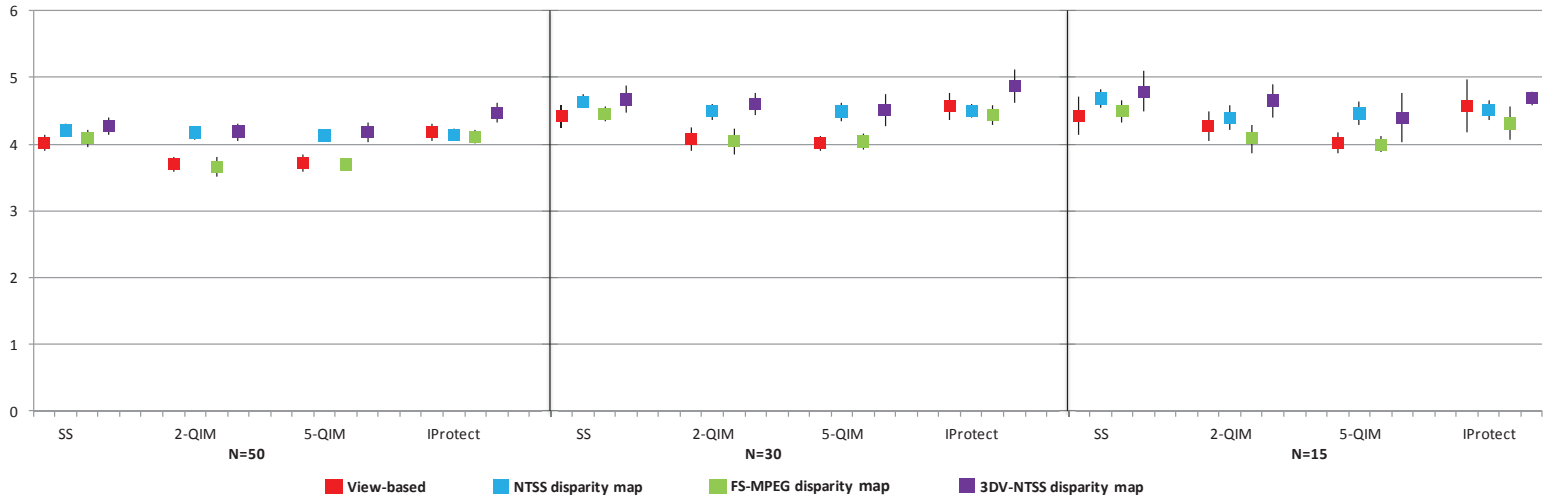
(a)


3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

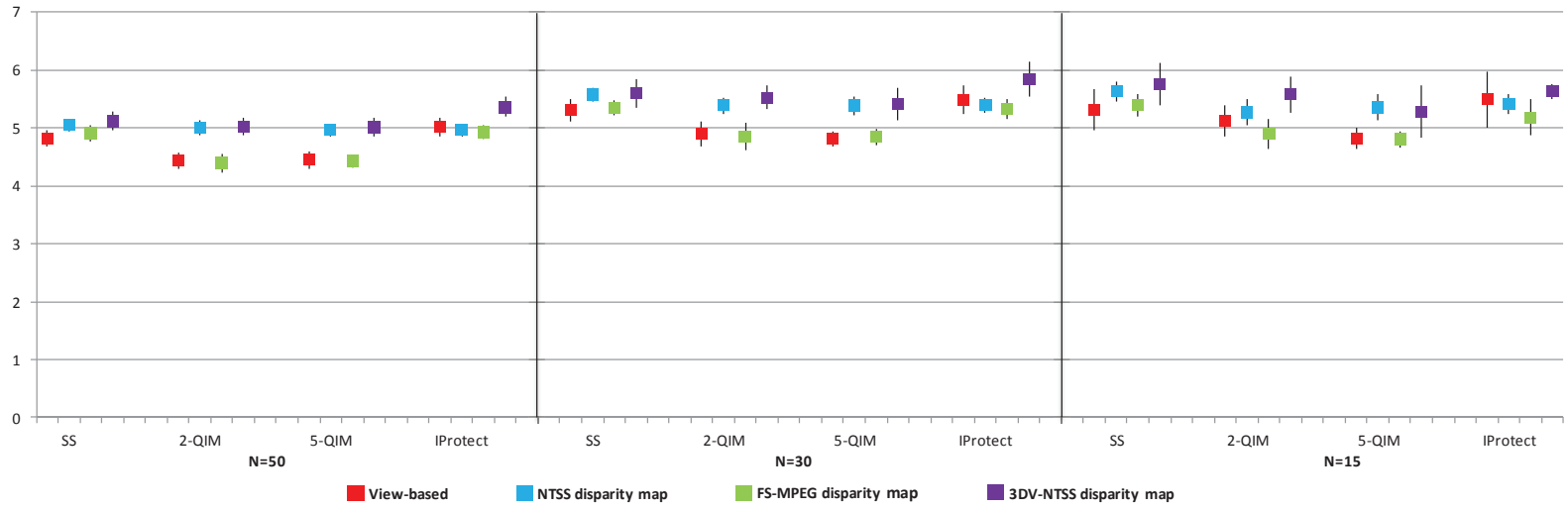
(b)


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

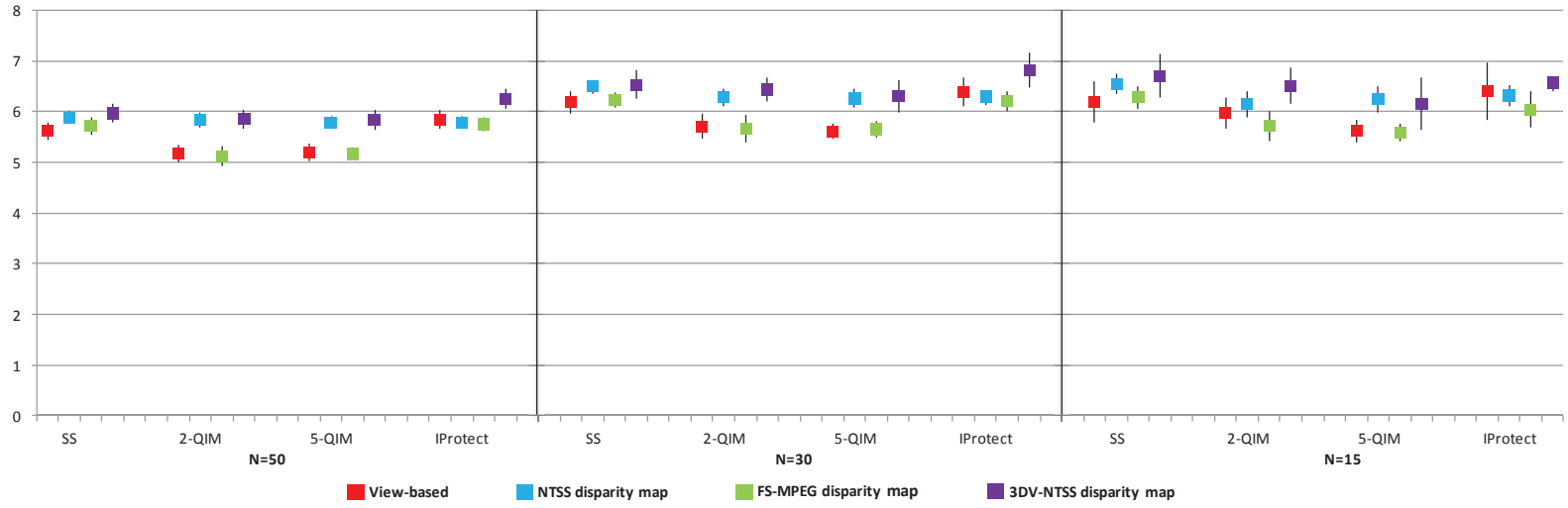
(c)


6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

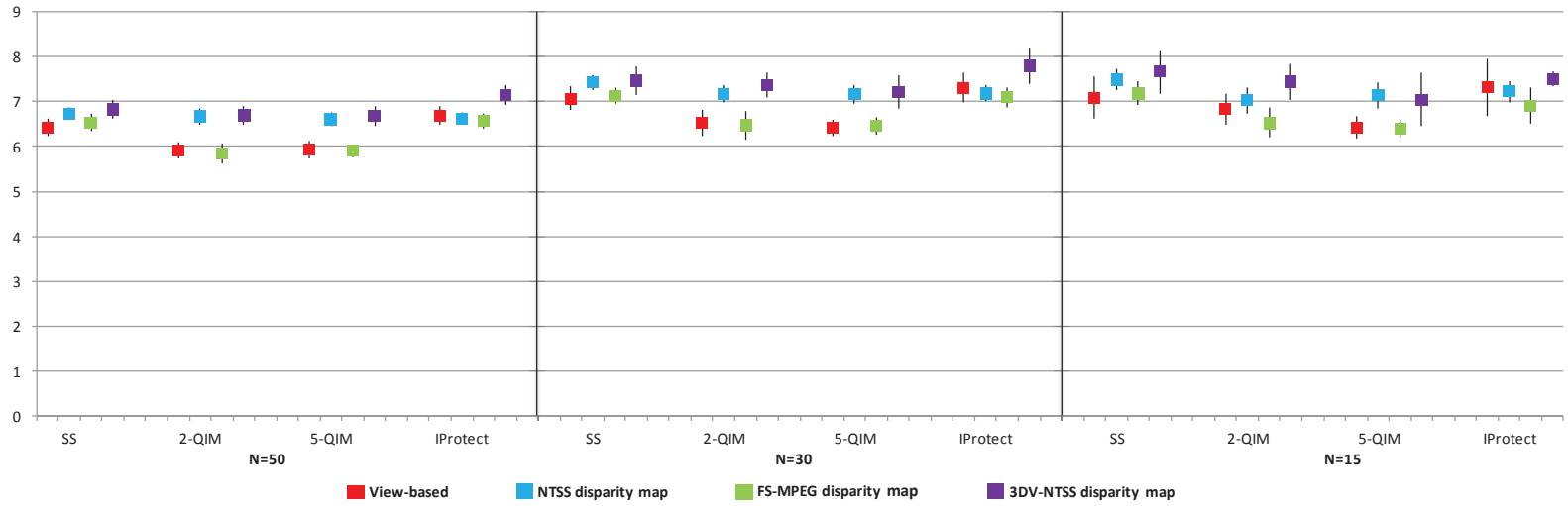
(d)


7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

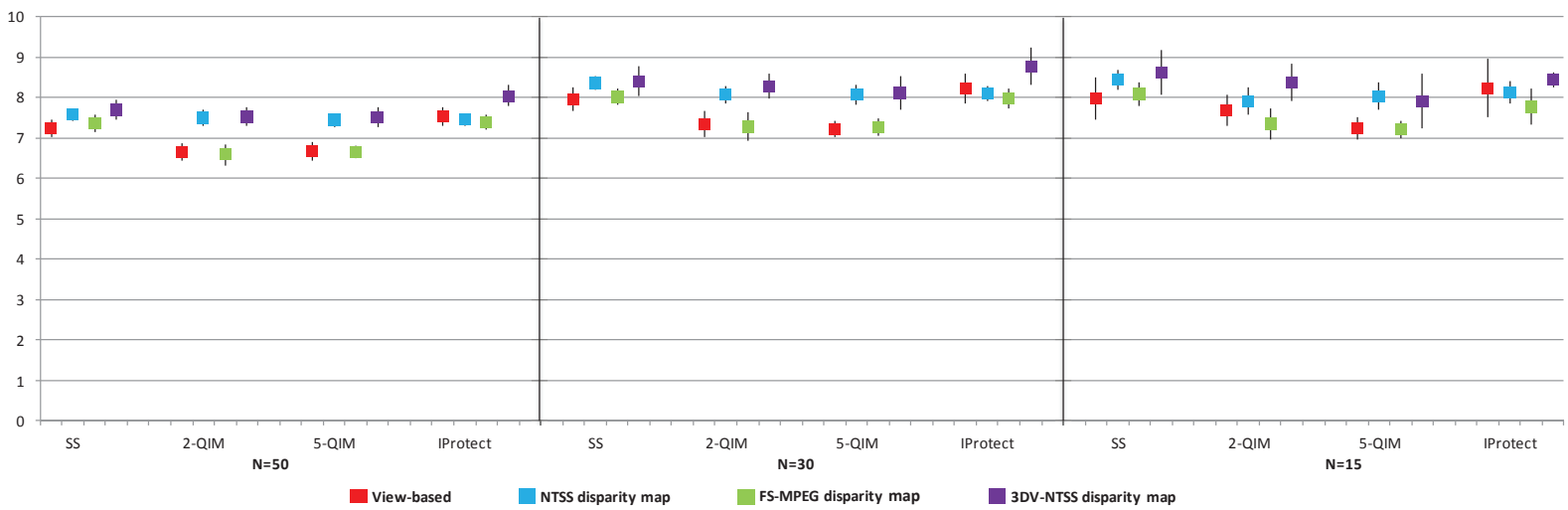
(e)


8
7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)



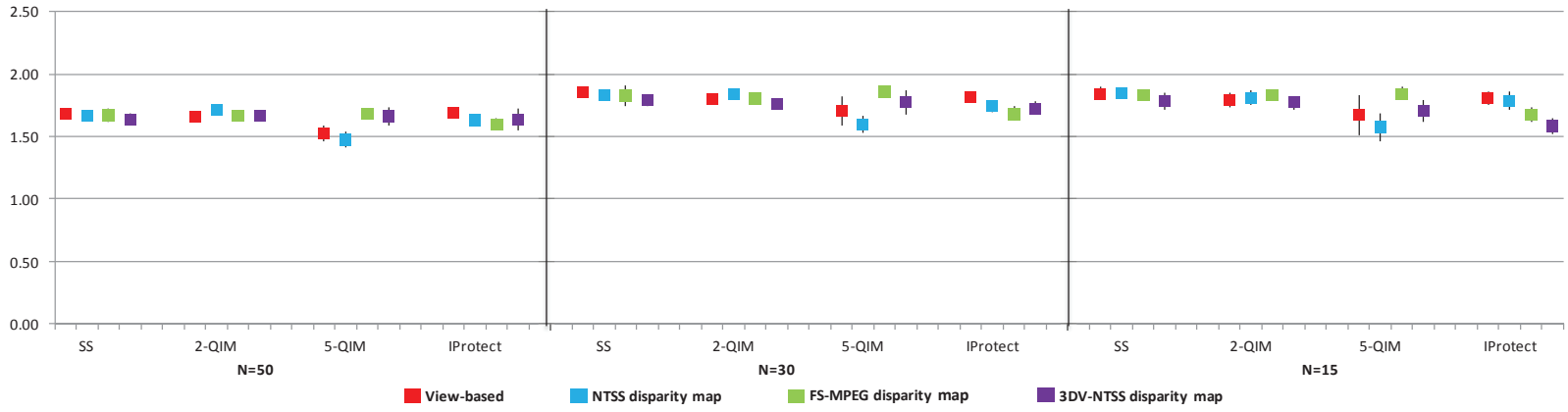
(g)



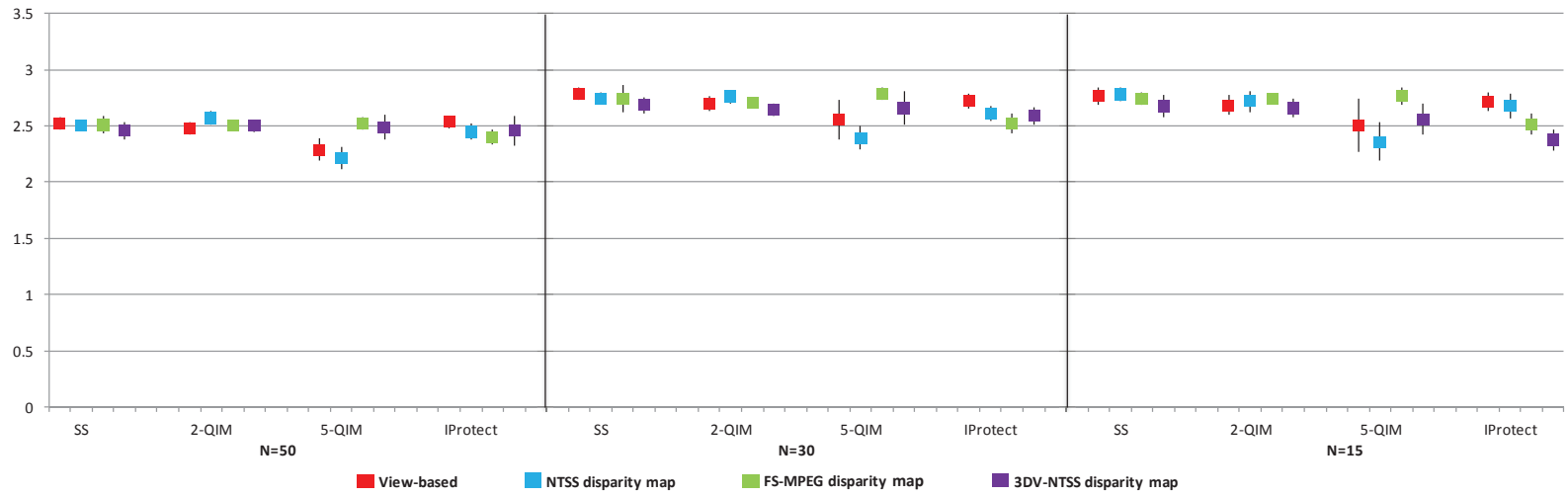
(h)

Figure A1-2: Subjective evaluations for high-quality stereoscopic video content (Depth Perception), for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15..*

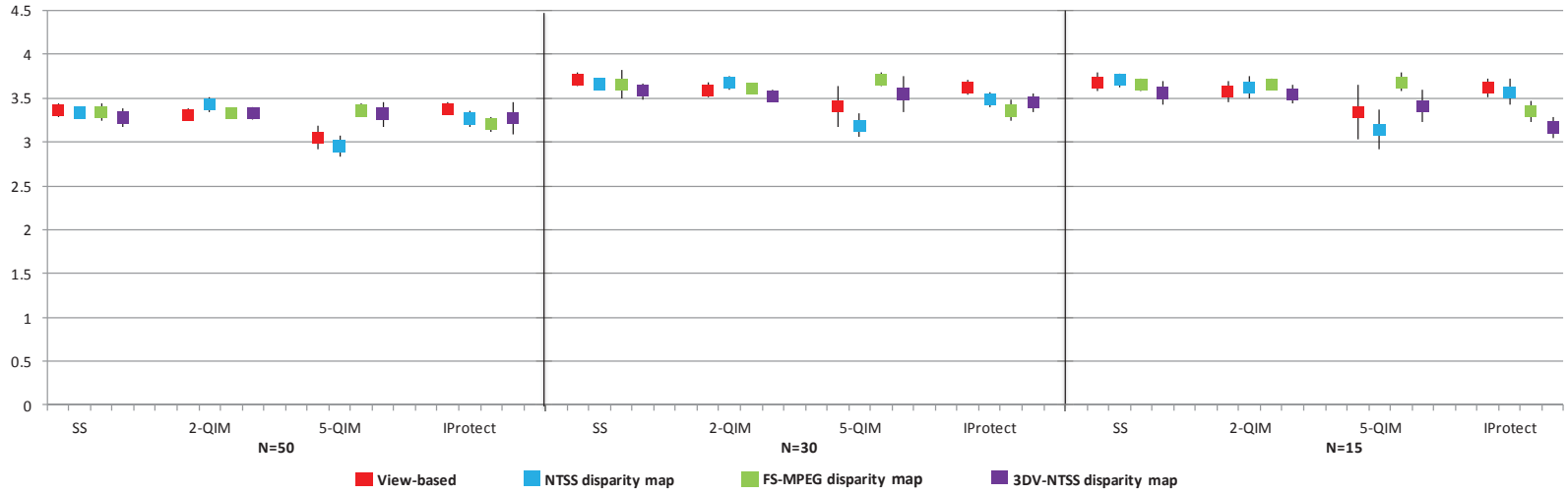
High quality stereoscopic video content (Visual Comfort)



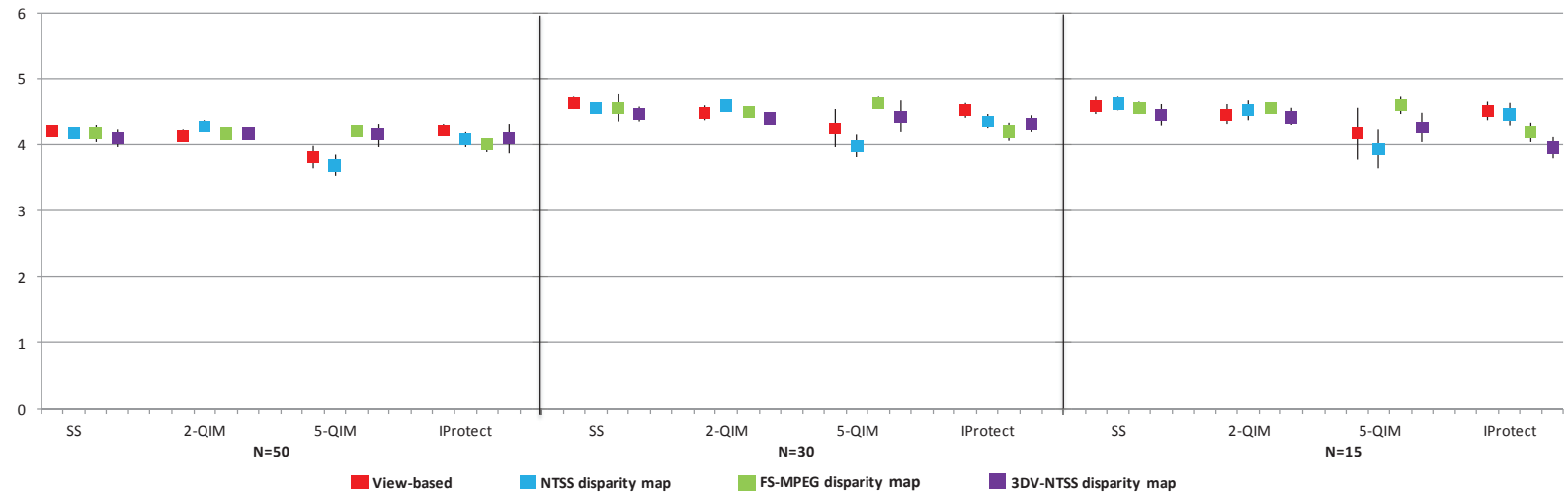
(a)



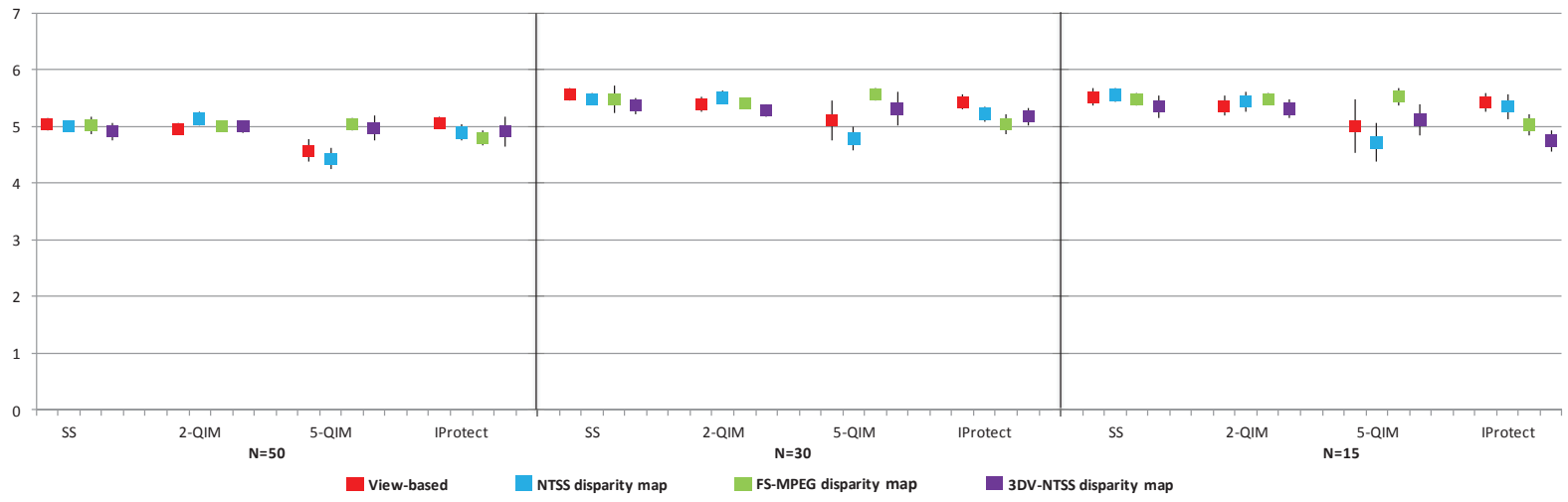
(b)


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

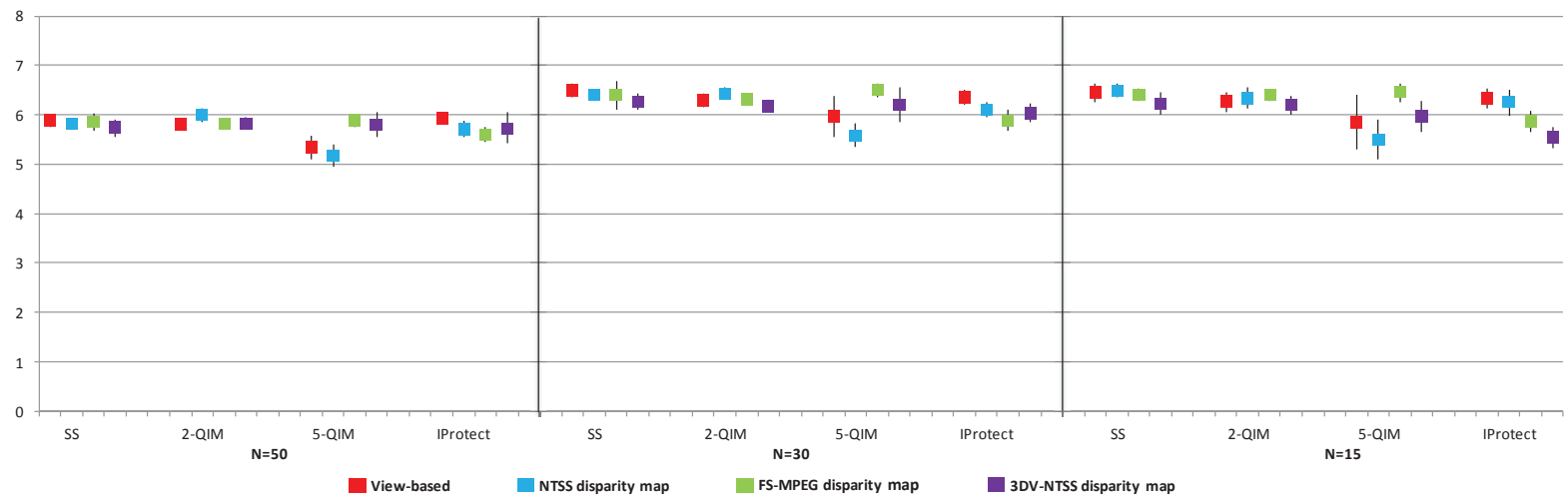
(c)


6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

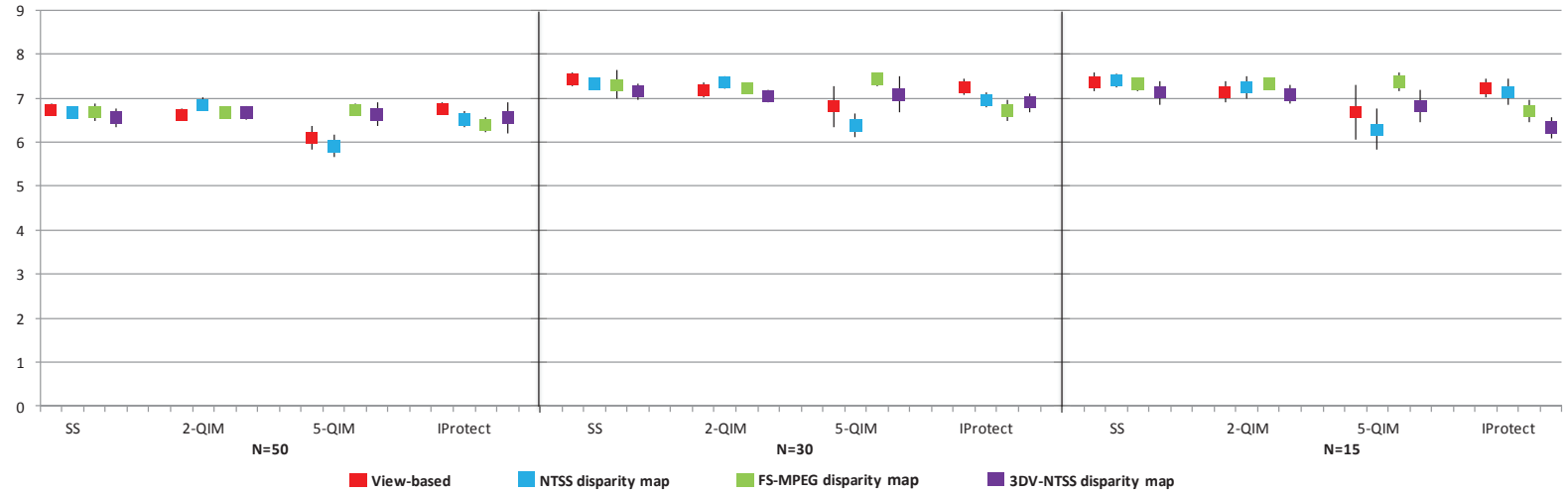
(d)


7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

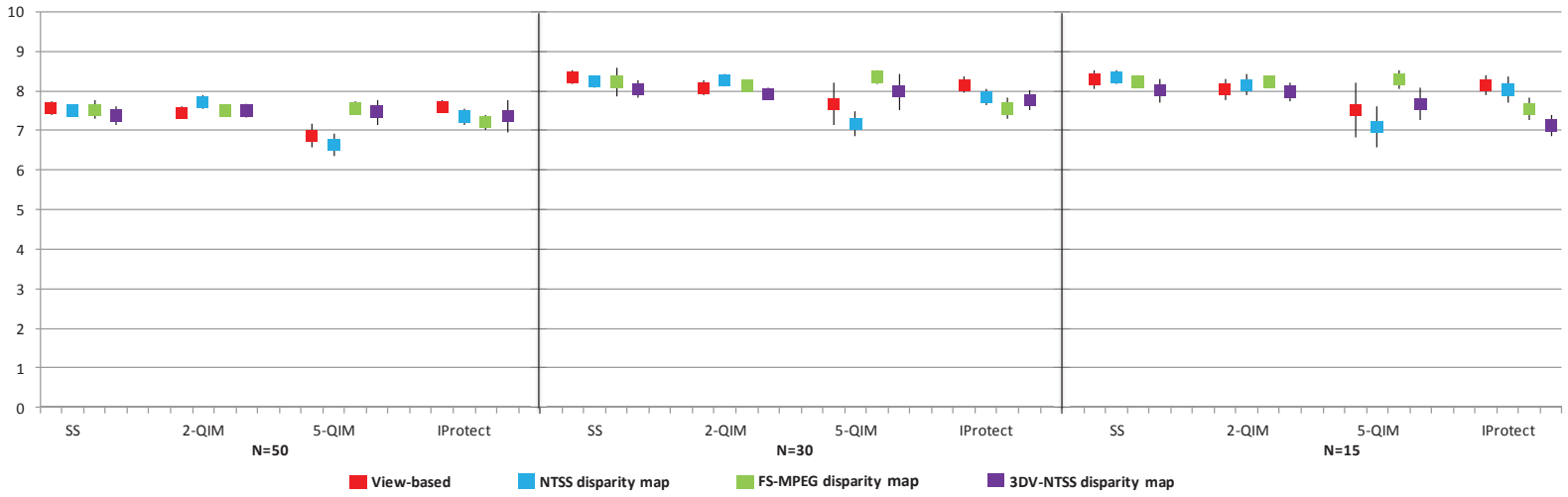
(e)


8
7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)

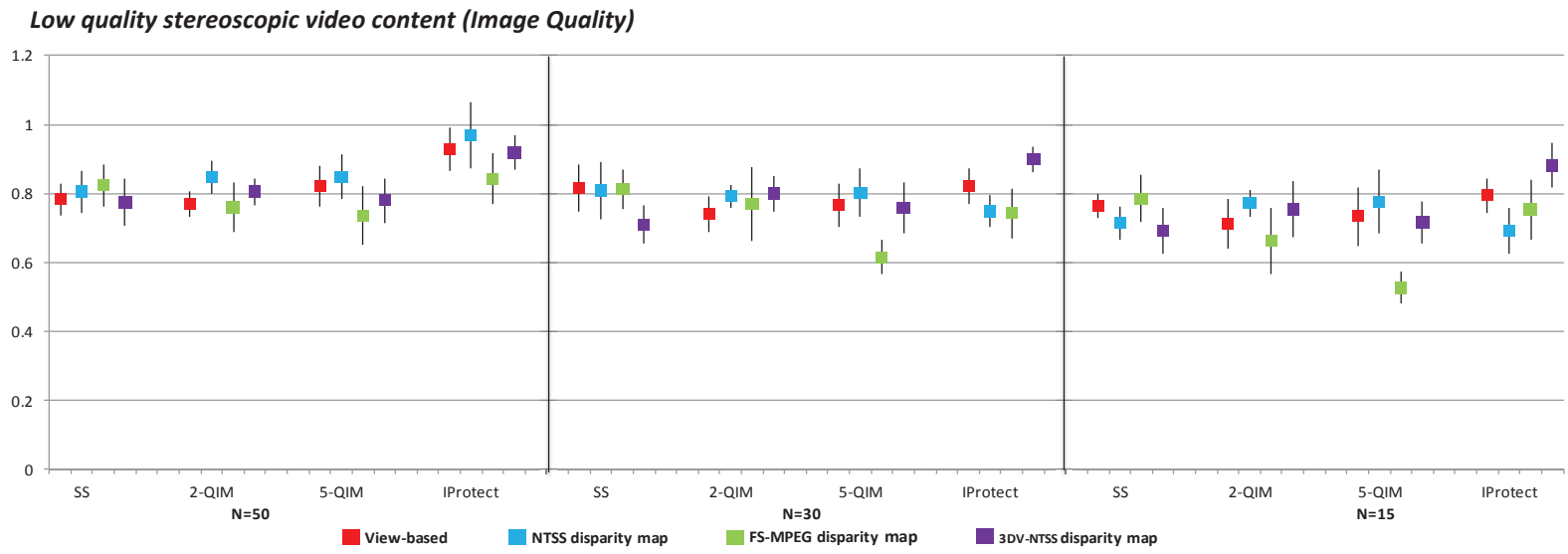


(g)

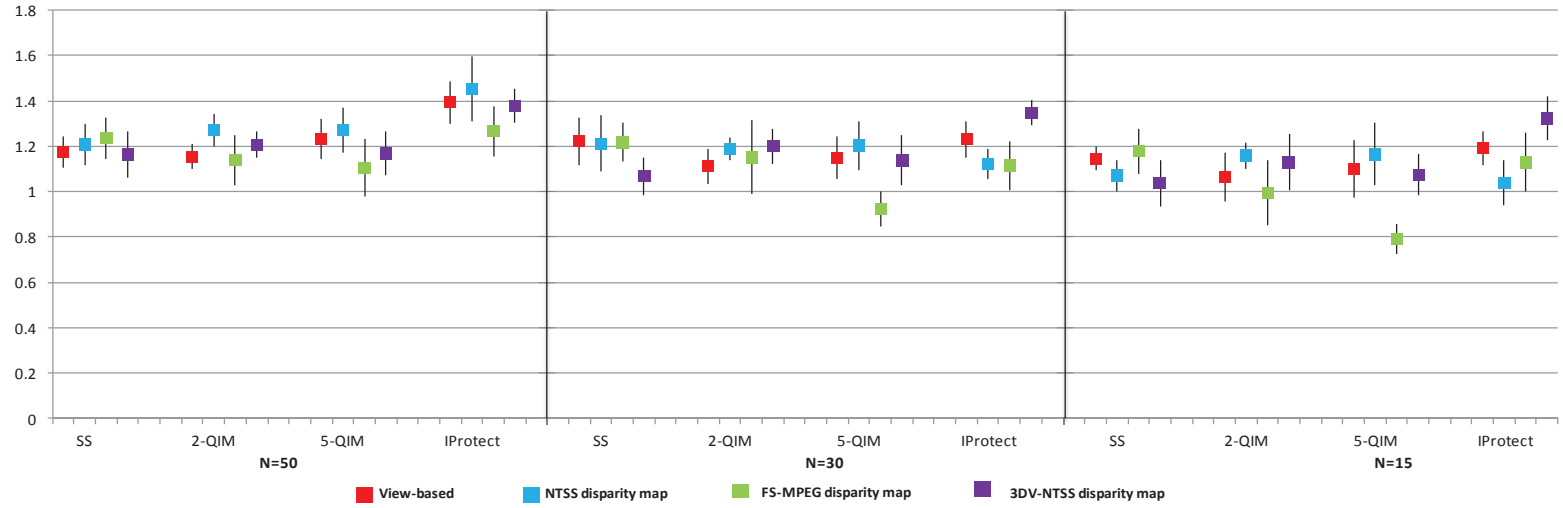


(h)

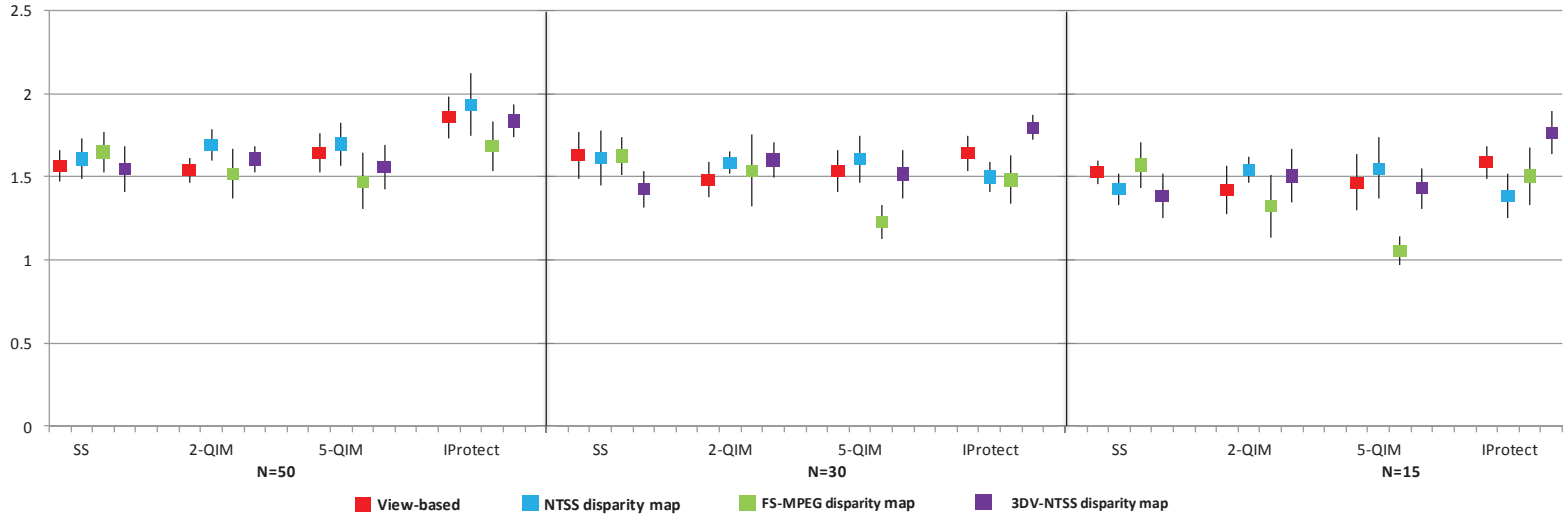
Figure A1-3: Subjective evaluations for high-quality stereoscopic video content (Visual Comfort), for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*.


Low quality stereoscopic video content (Image Quality)
1.2
1
0.8
0.6
0.4
0.2
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

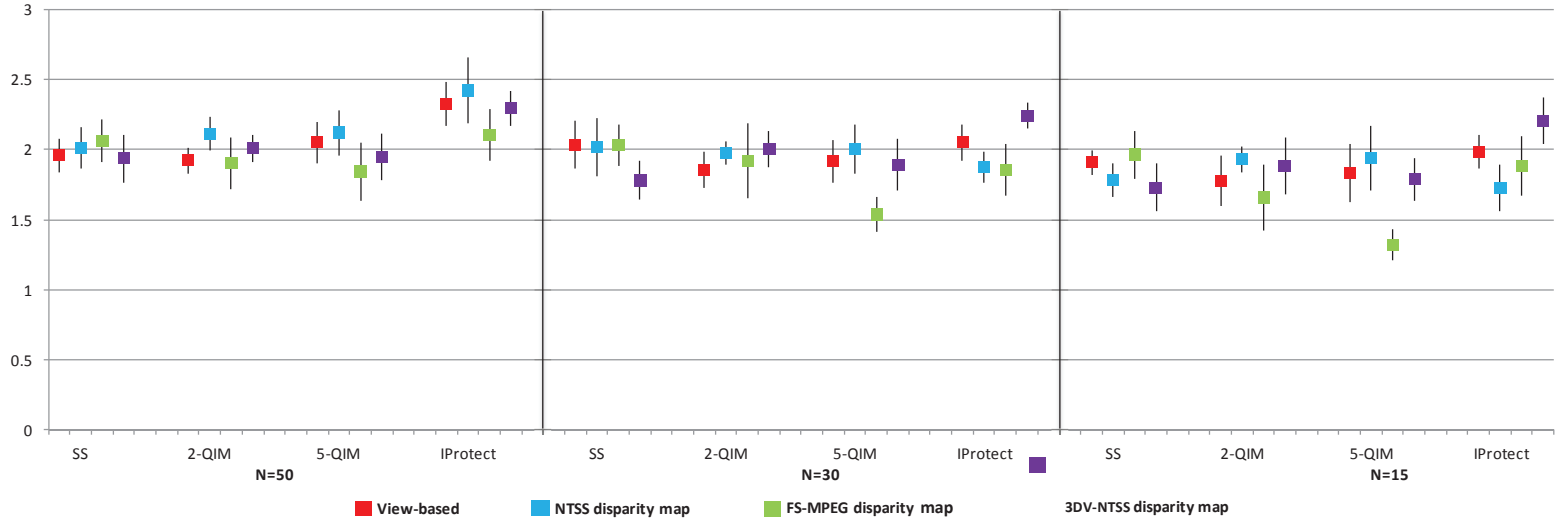
(a)


1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

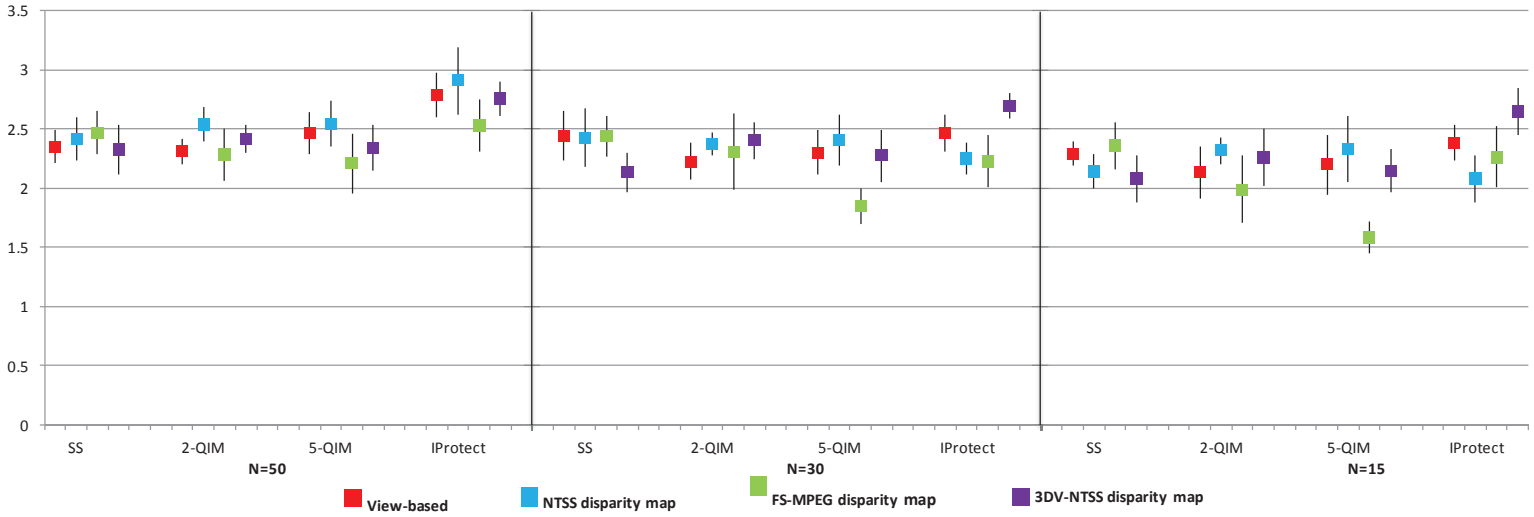
(b)


2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

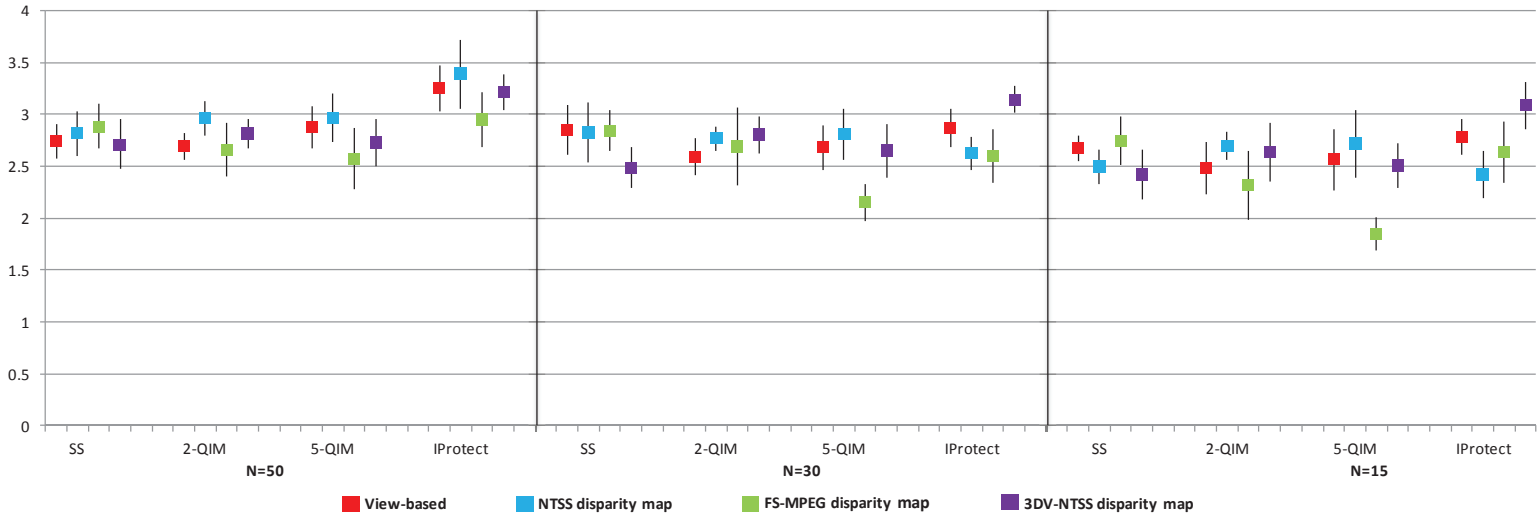
(c)


3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

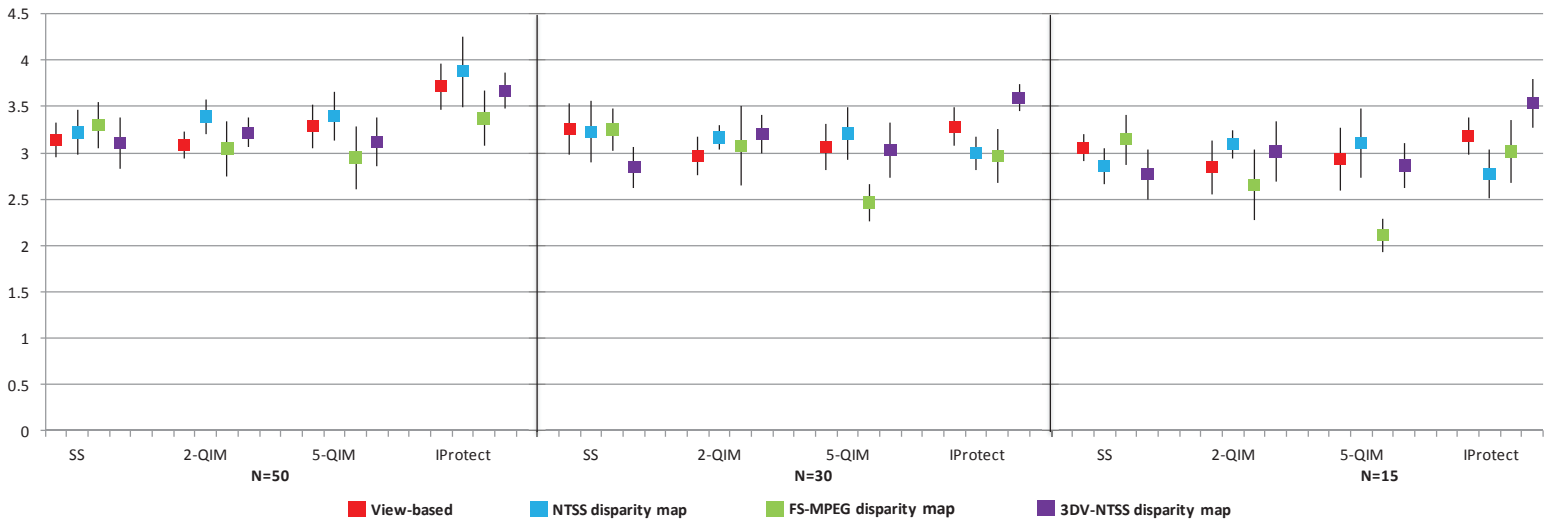
(d)


3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

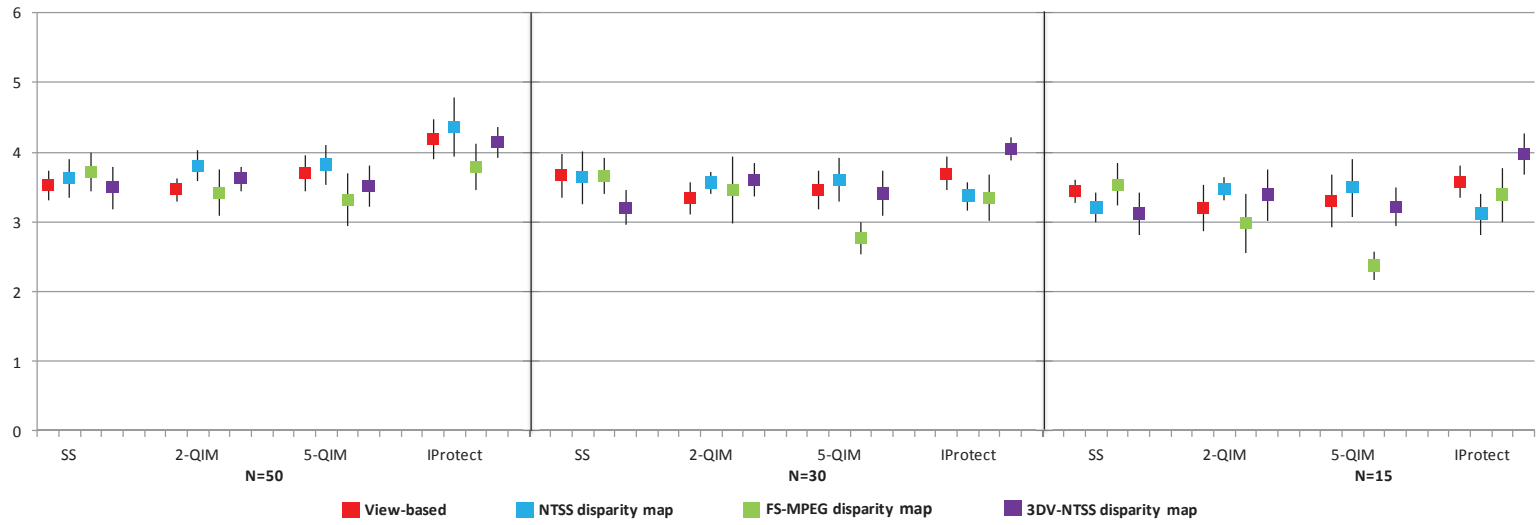
(e)


4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)

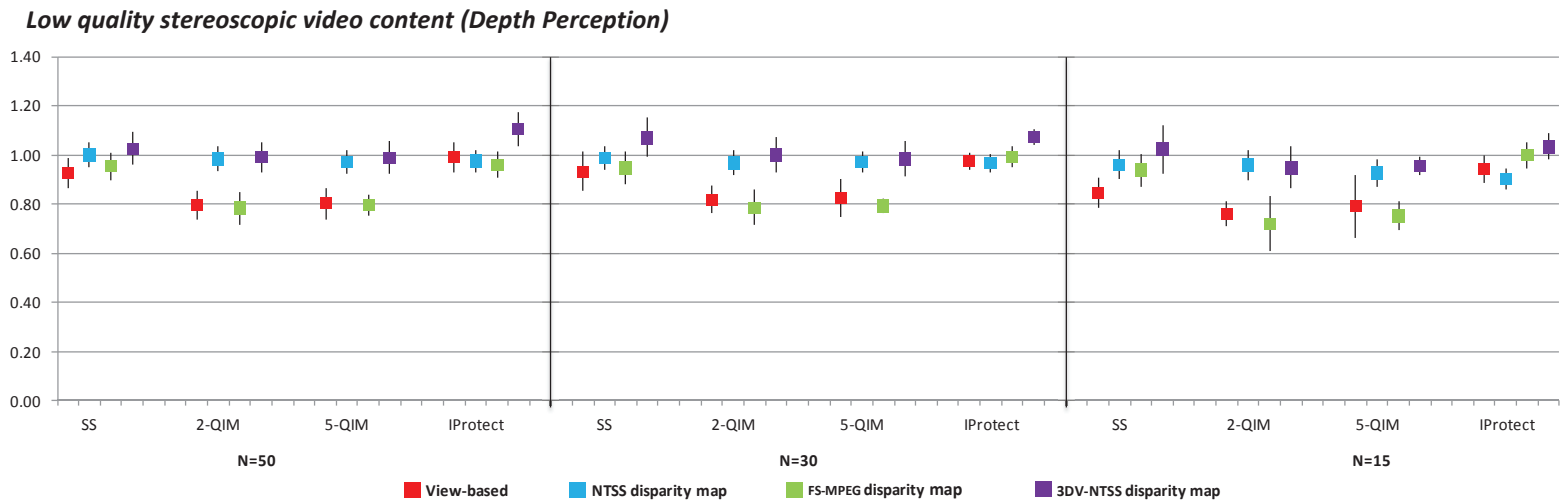


(g)

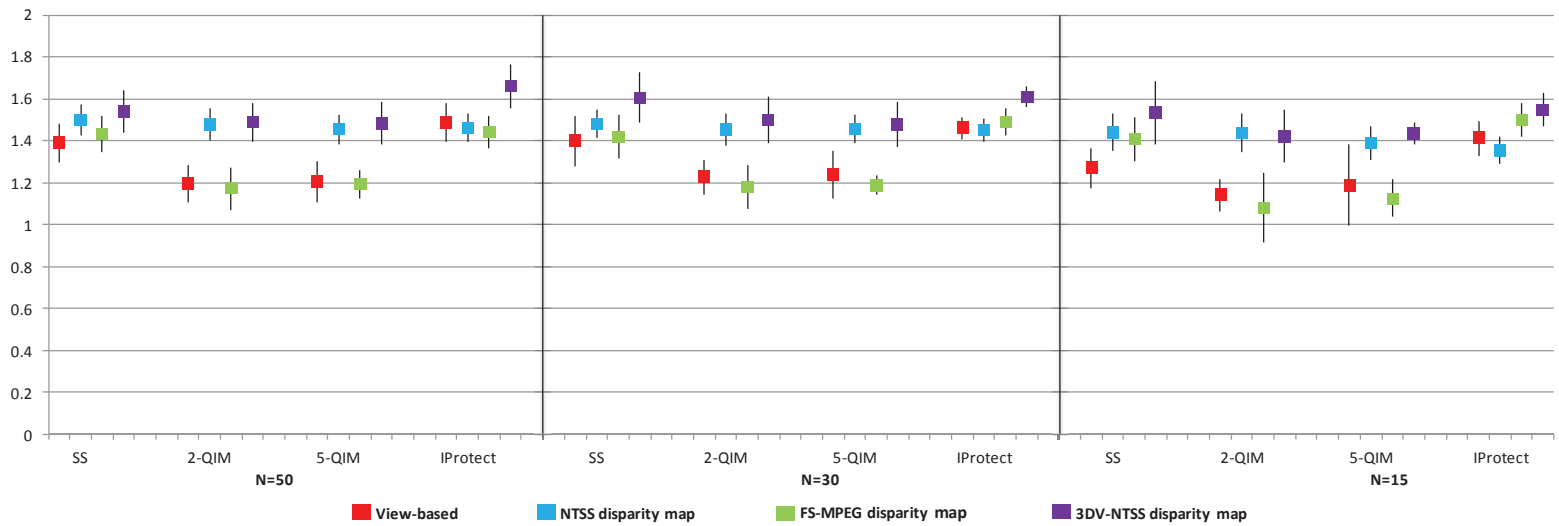


(h)

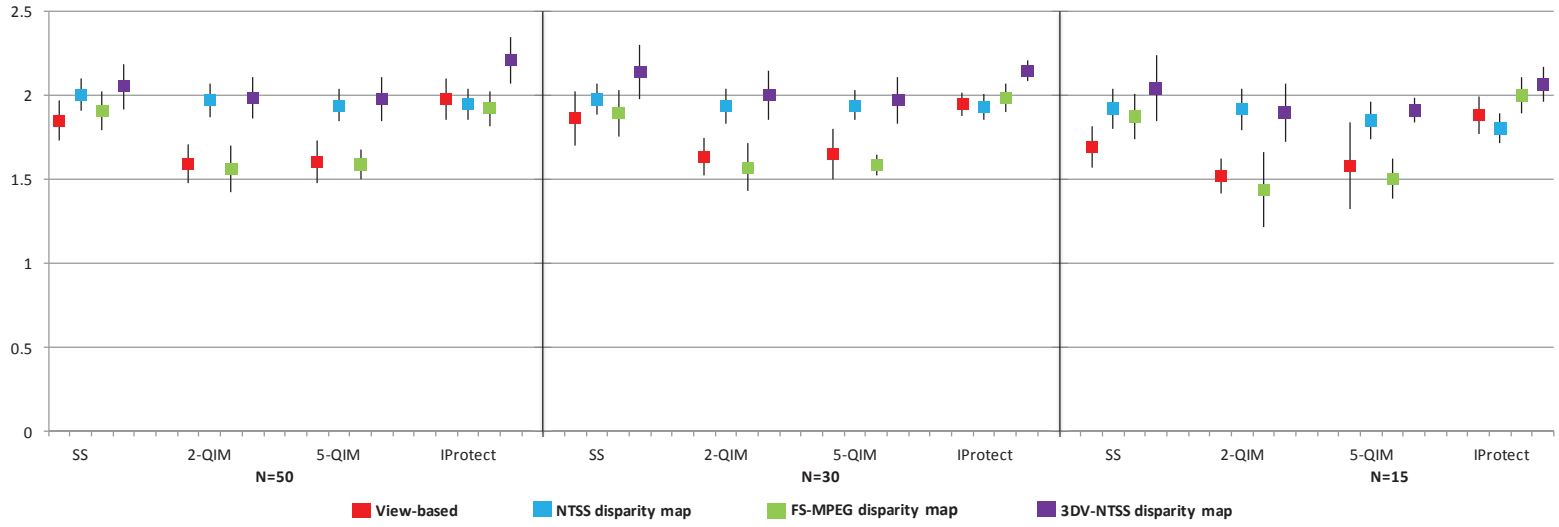
Figure A1-4: Subjective evaluations low-quality stereoscopic video content (Image Quality), for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*


Low quality stereoscopic video content (Depth Perception)
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

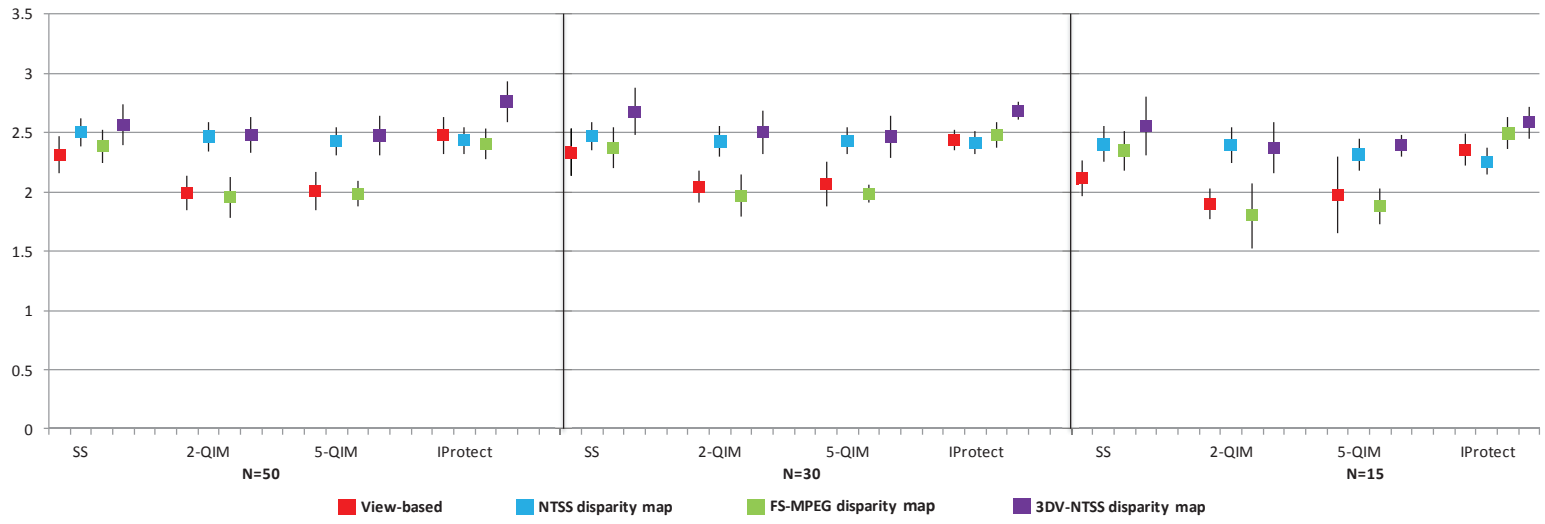
(a)


2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

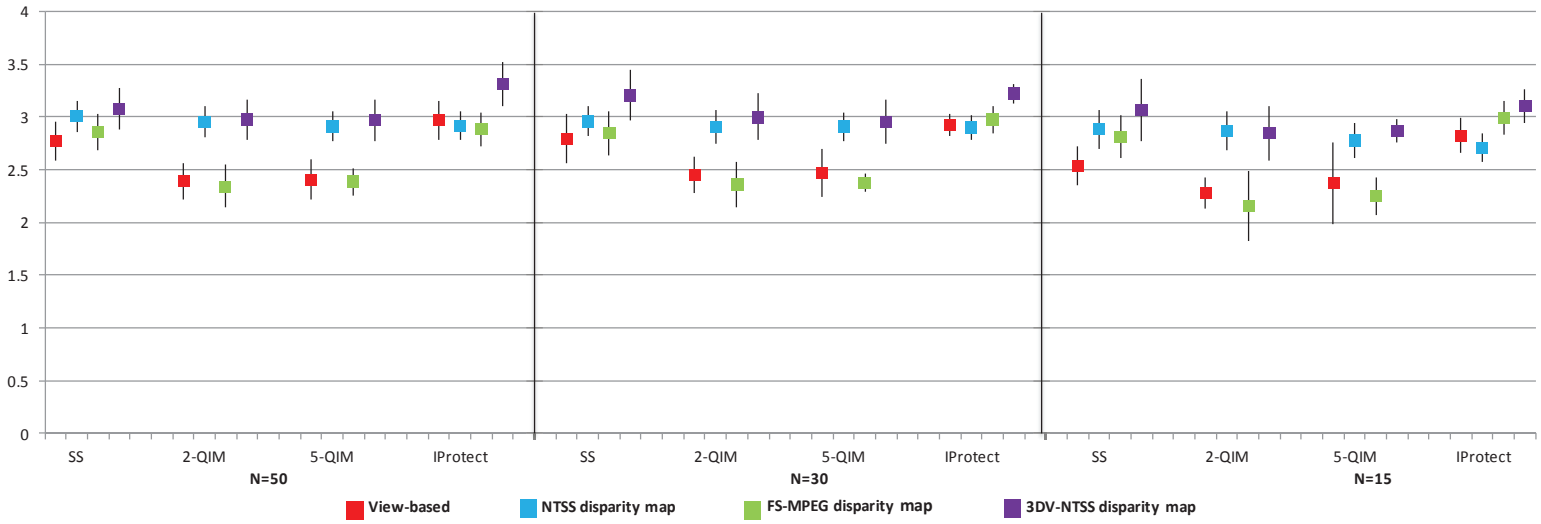
(b)


2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

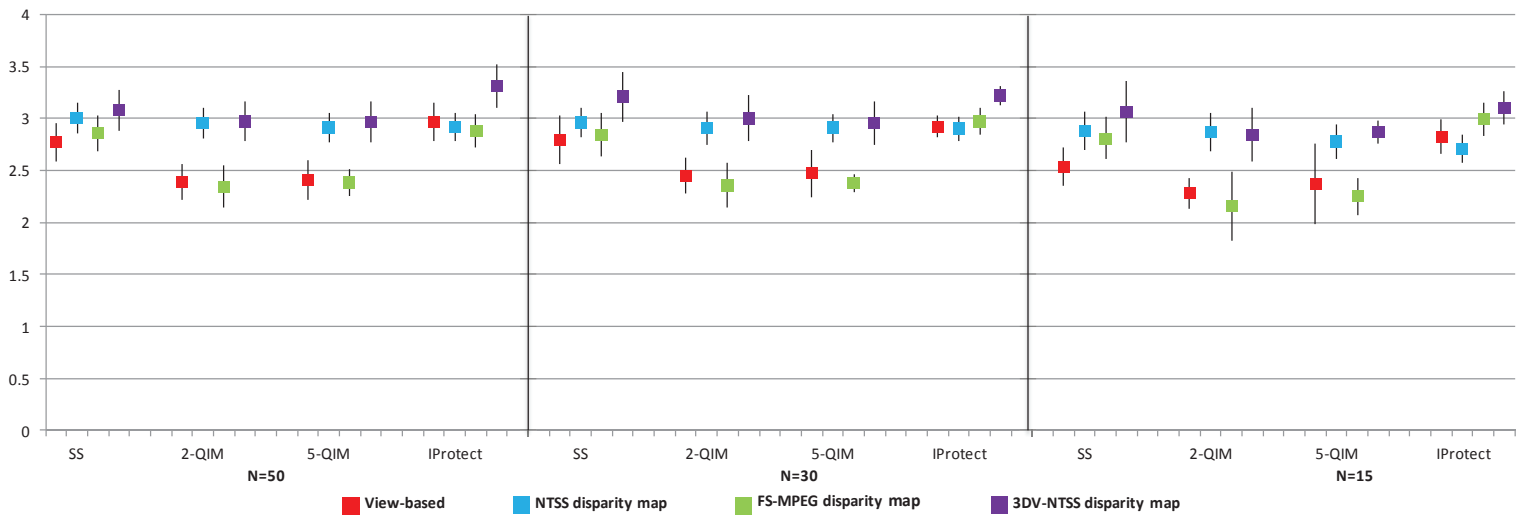
(c)


3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

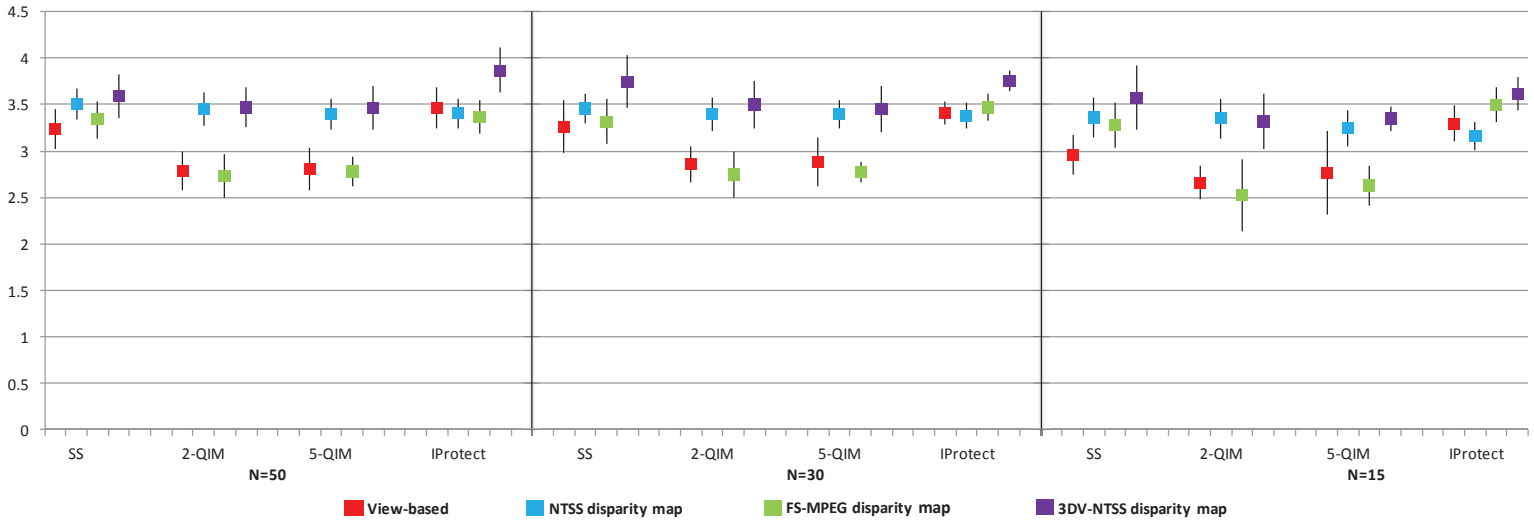
(d)


4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

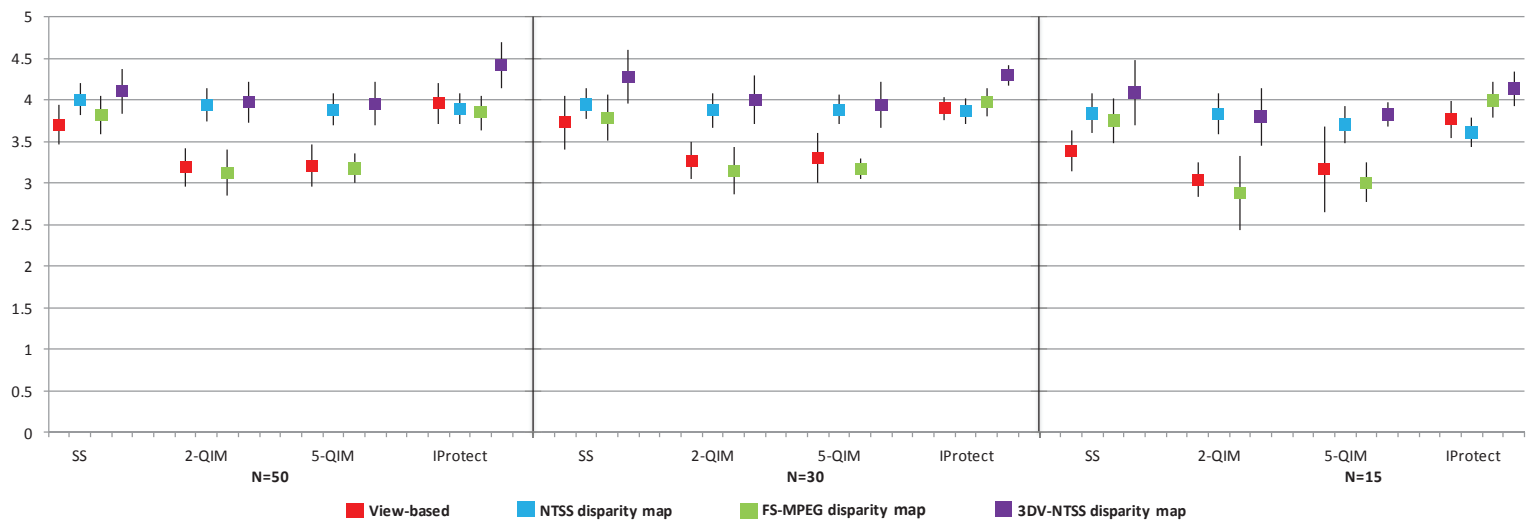
(e)


4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)



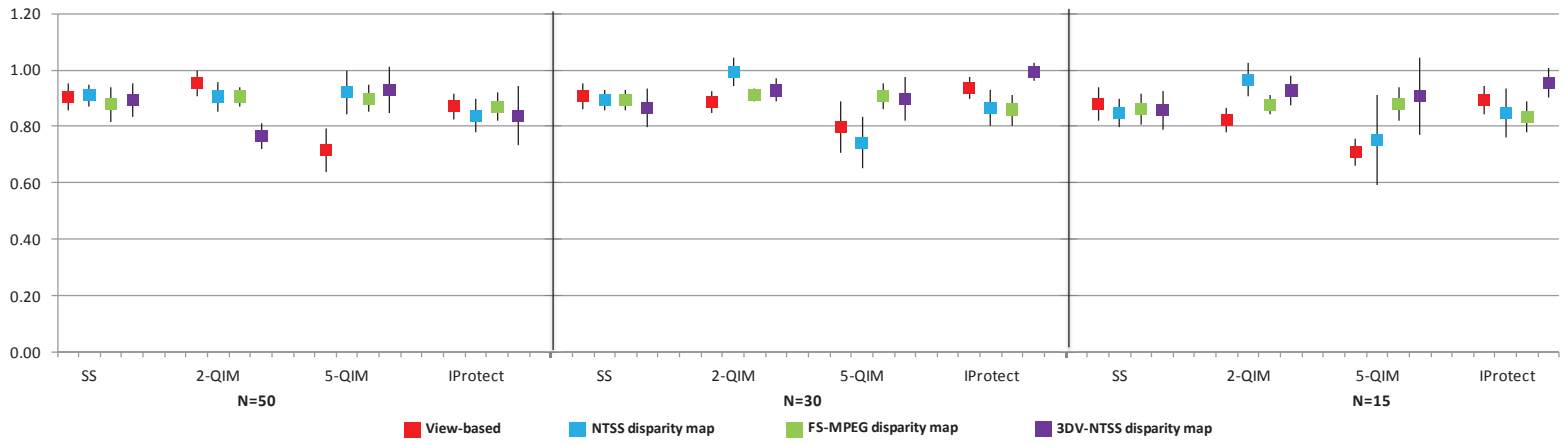
(g)



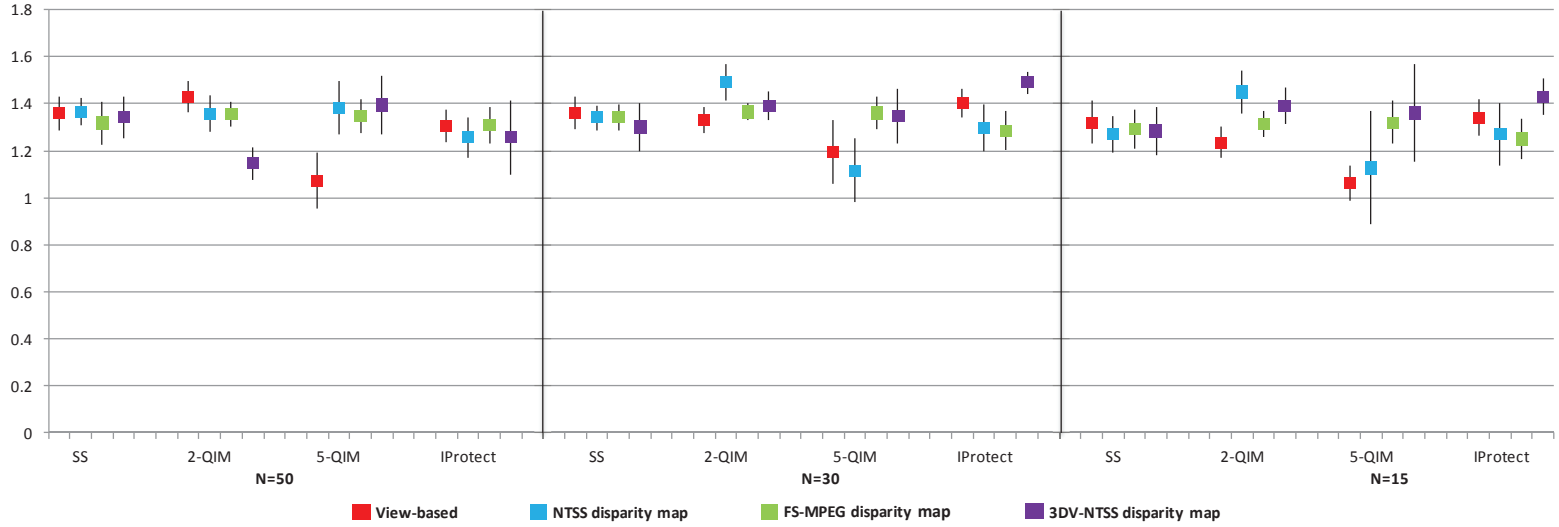
(h)

Figure A1-5: Subjective evaluations low-quality stereoscopic video content (Depth Perception), for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*

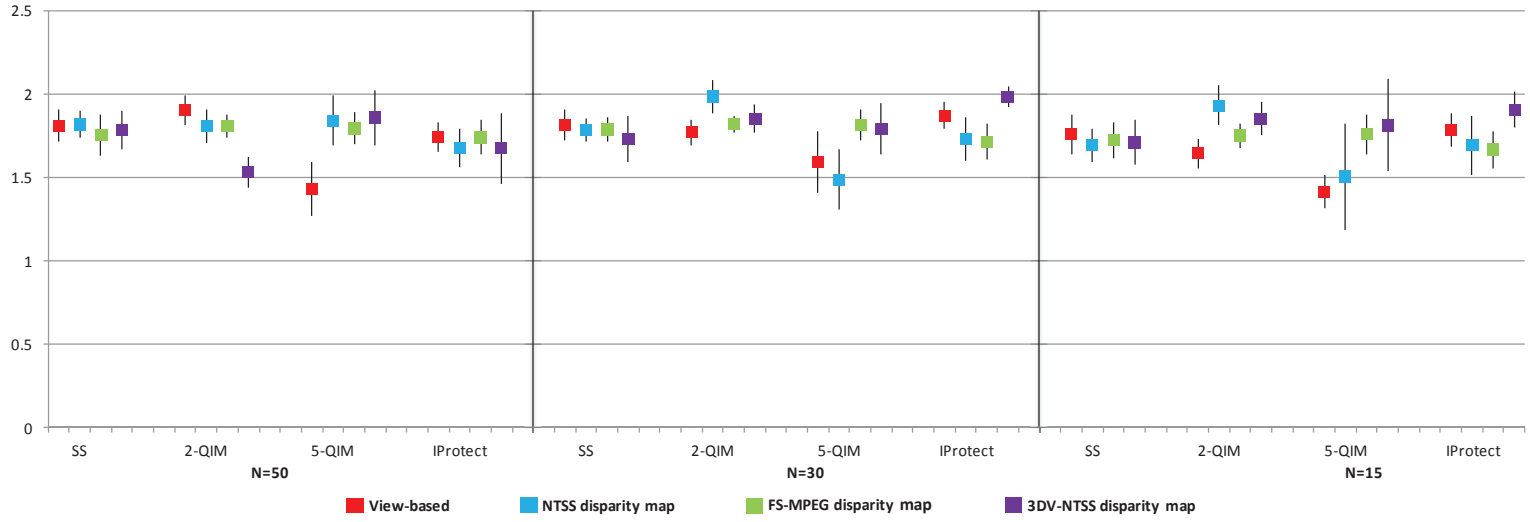
Low quality stereoscopic video content (Visual comfort)



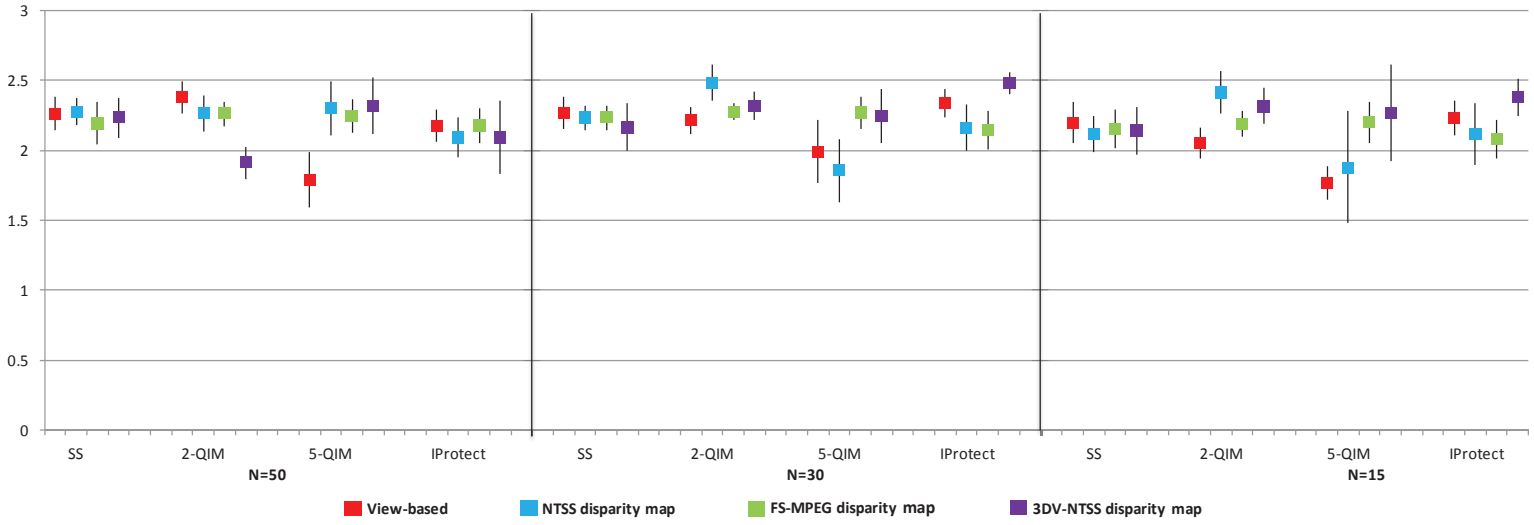
(a)



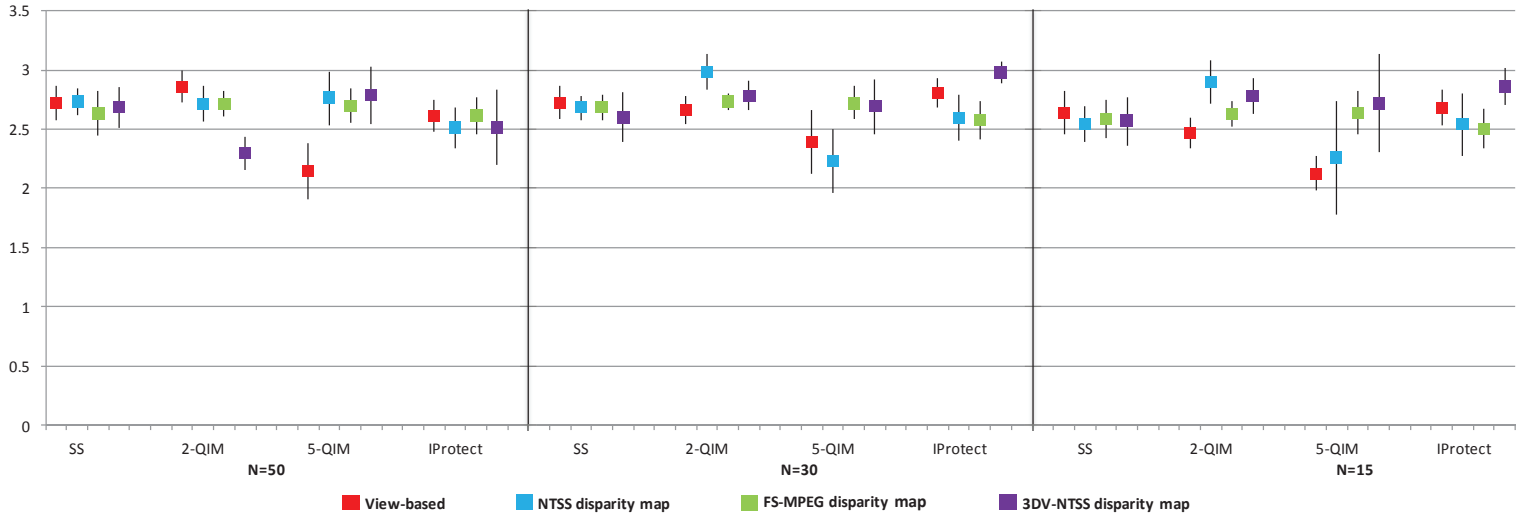
(b)


2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

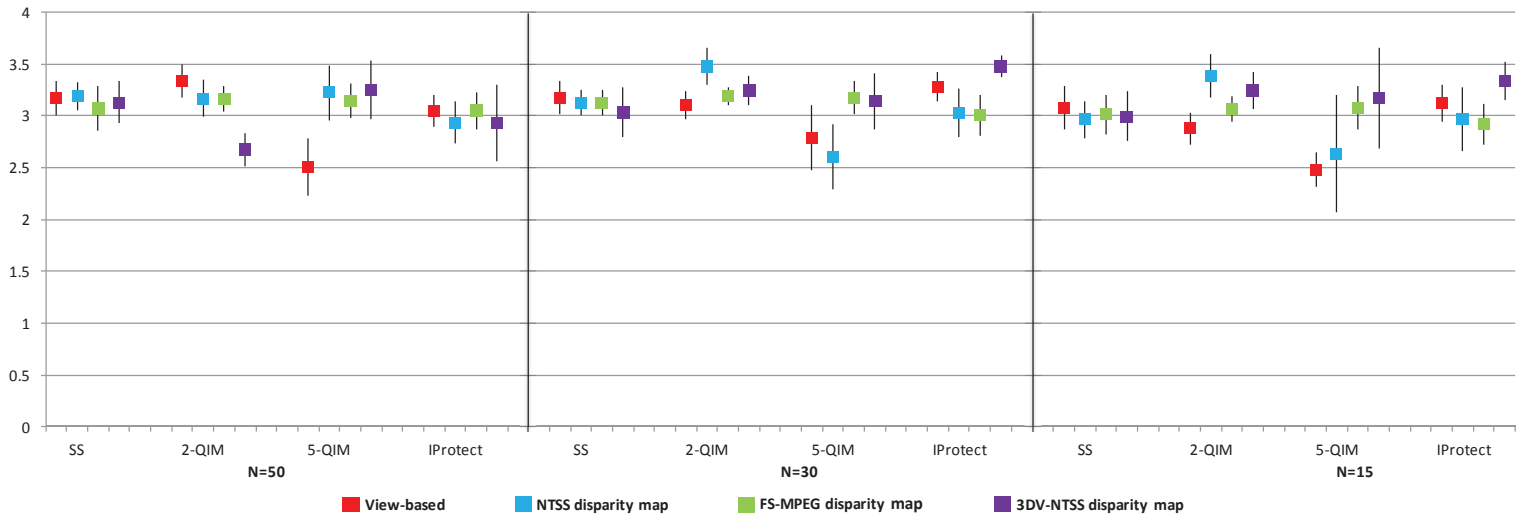
(c)


3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

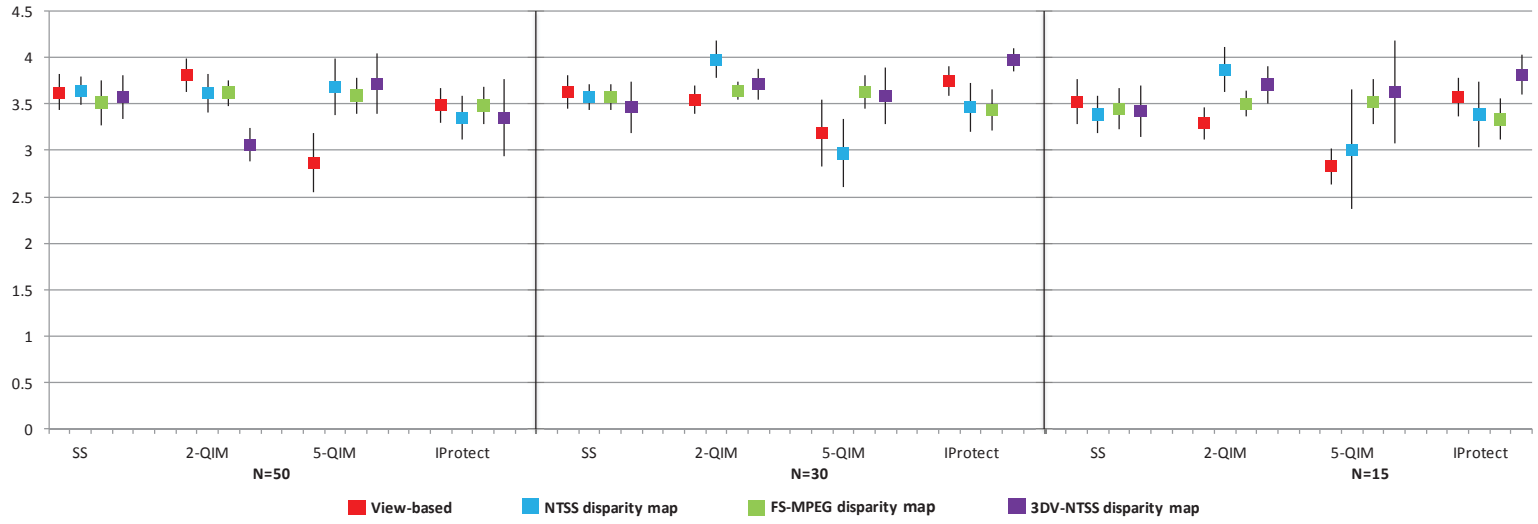
(d)


3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

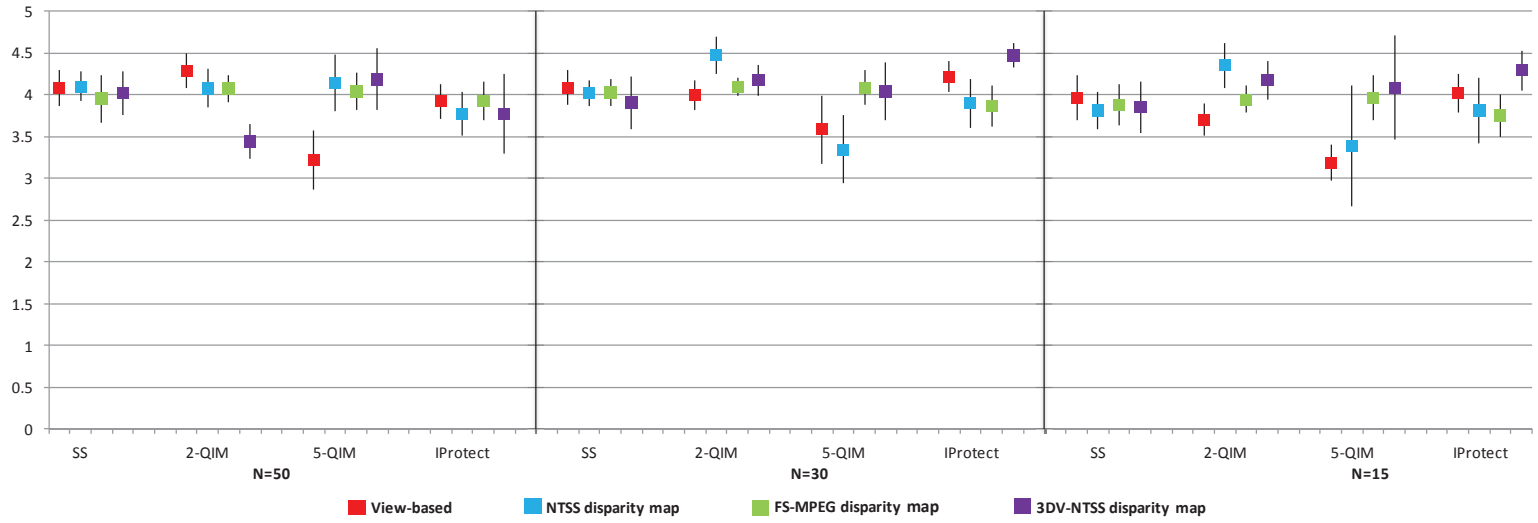
(e)


4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)

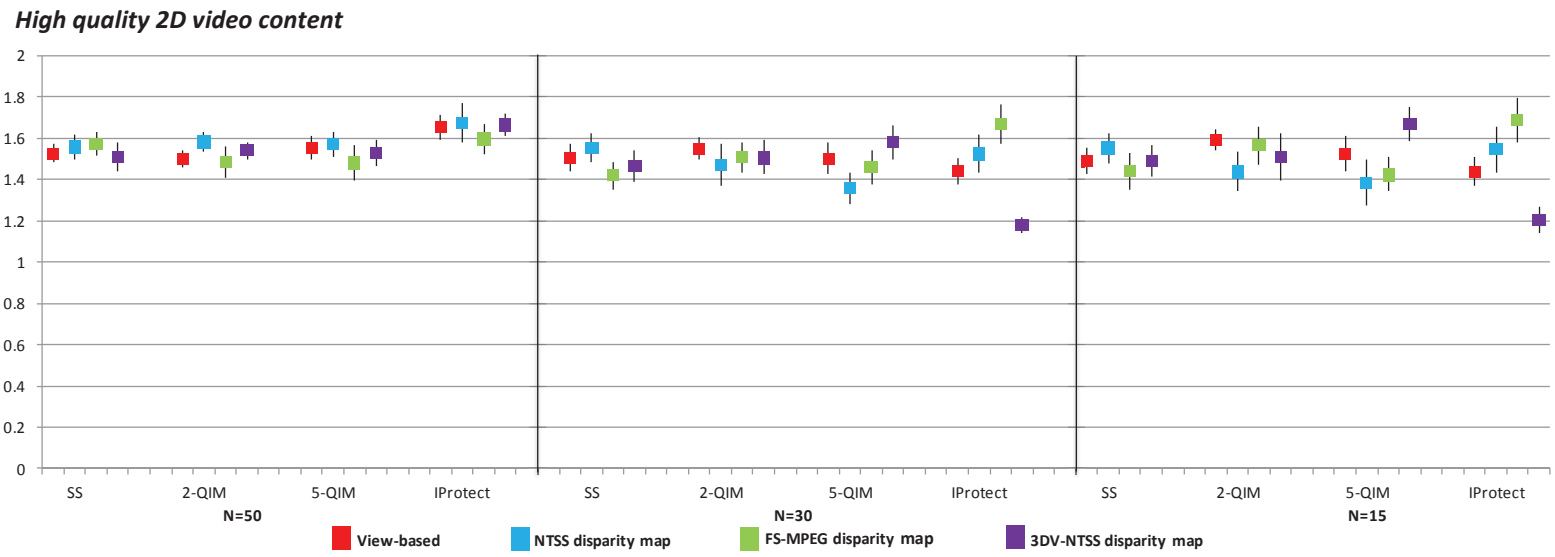


(g)

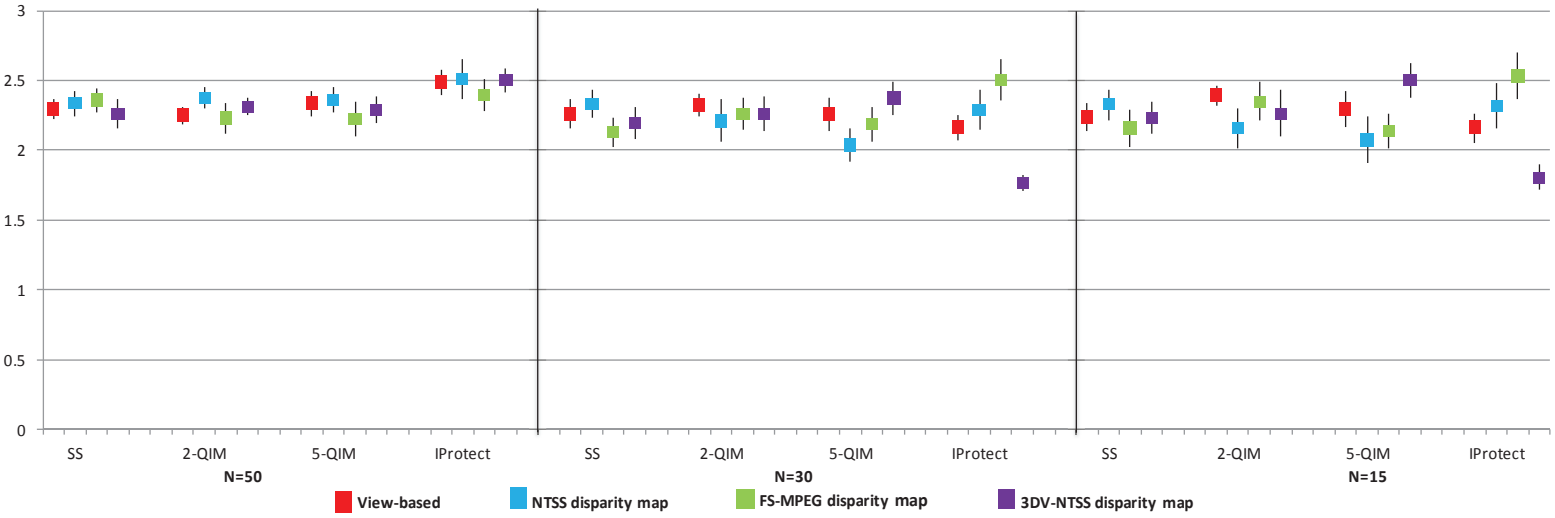


(h)

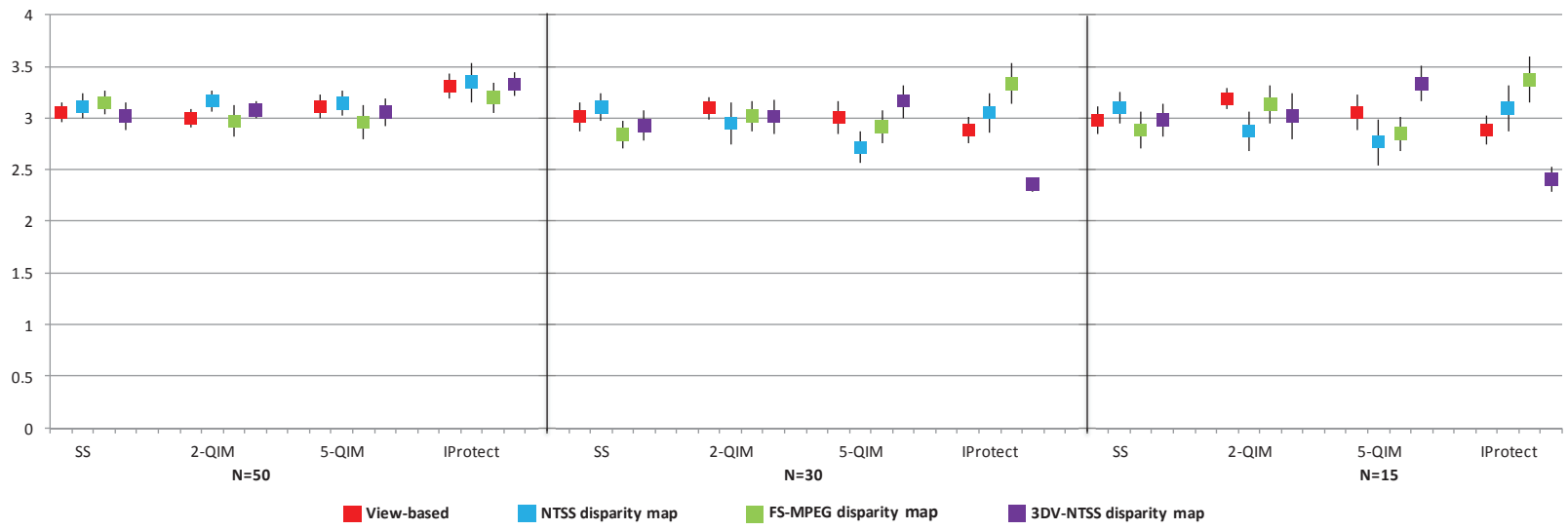
Figure A1-6: Subjective evaluations low-quality stereoscopic video content (Visual Comfort), for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*


High quality 2D video content
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

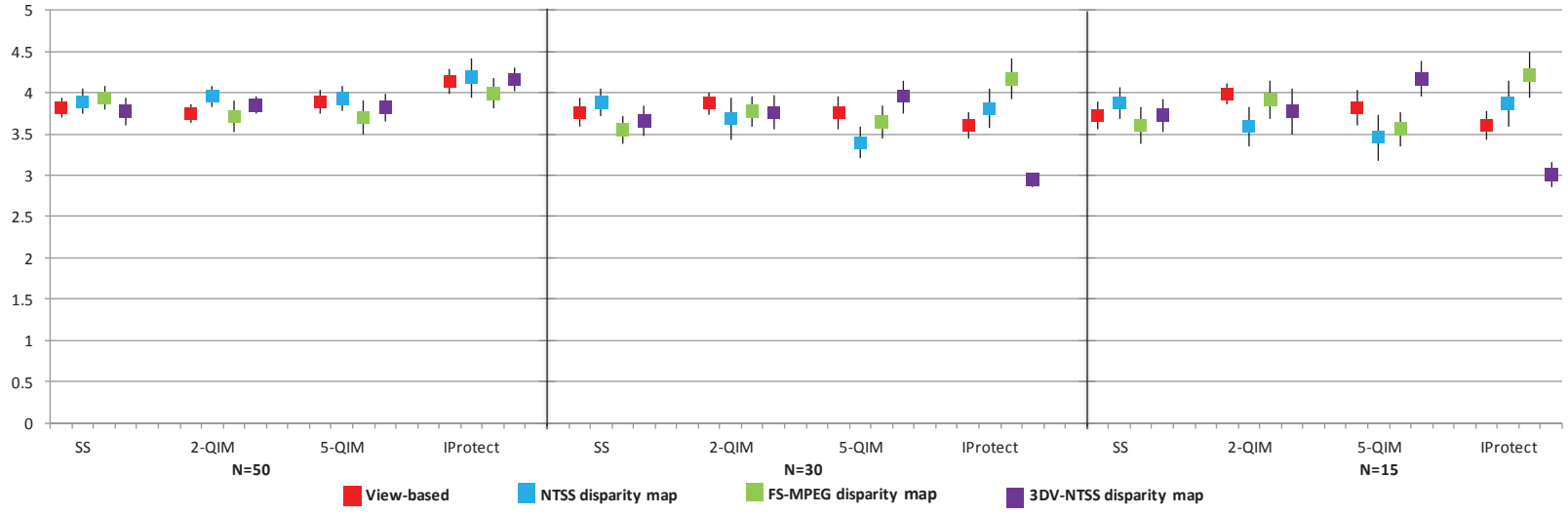
(a)


3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

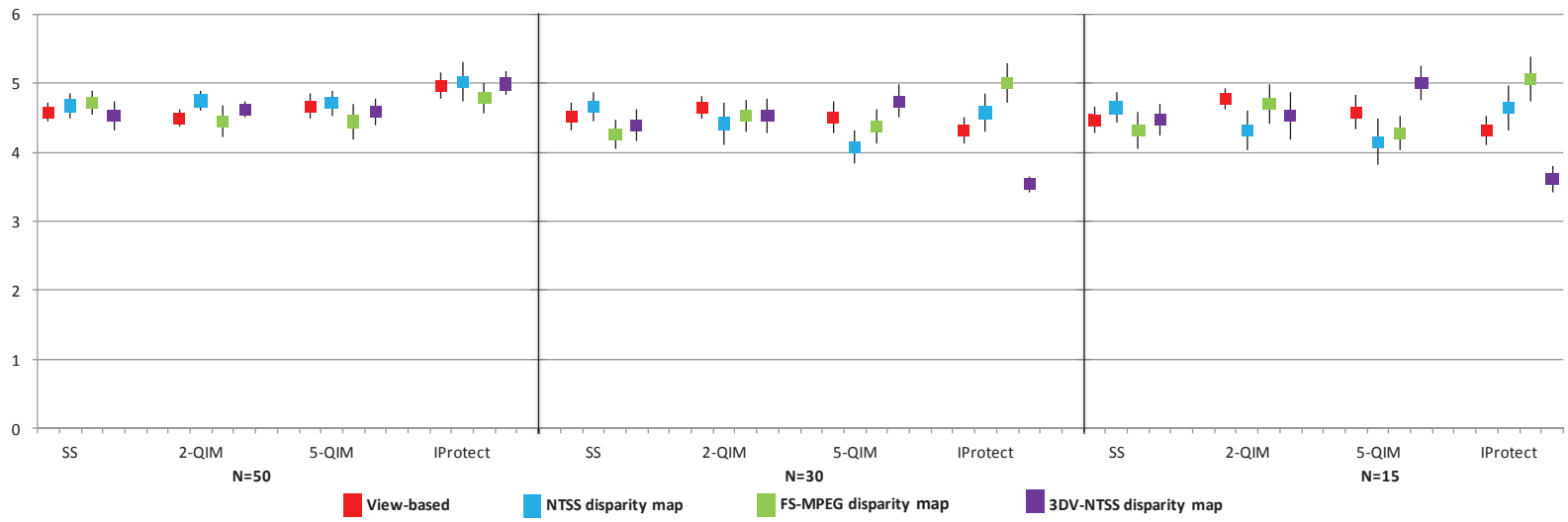
(b)


4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

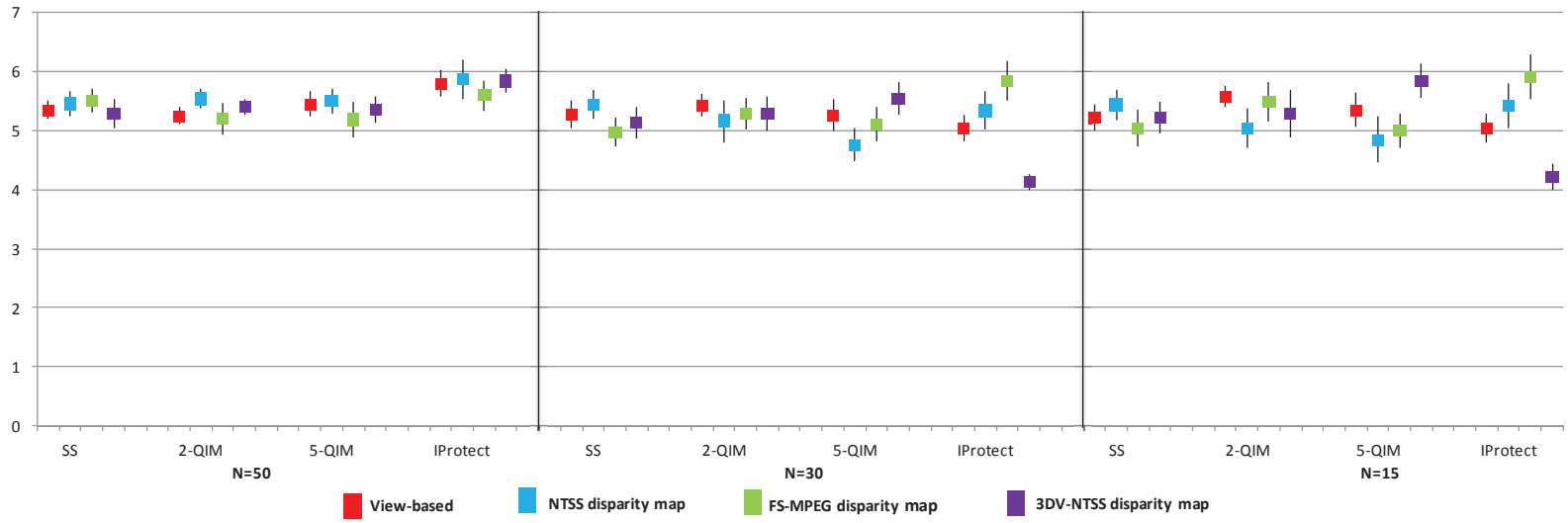
(c)


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

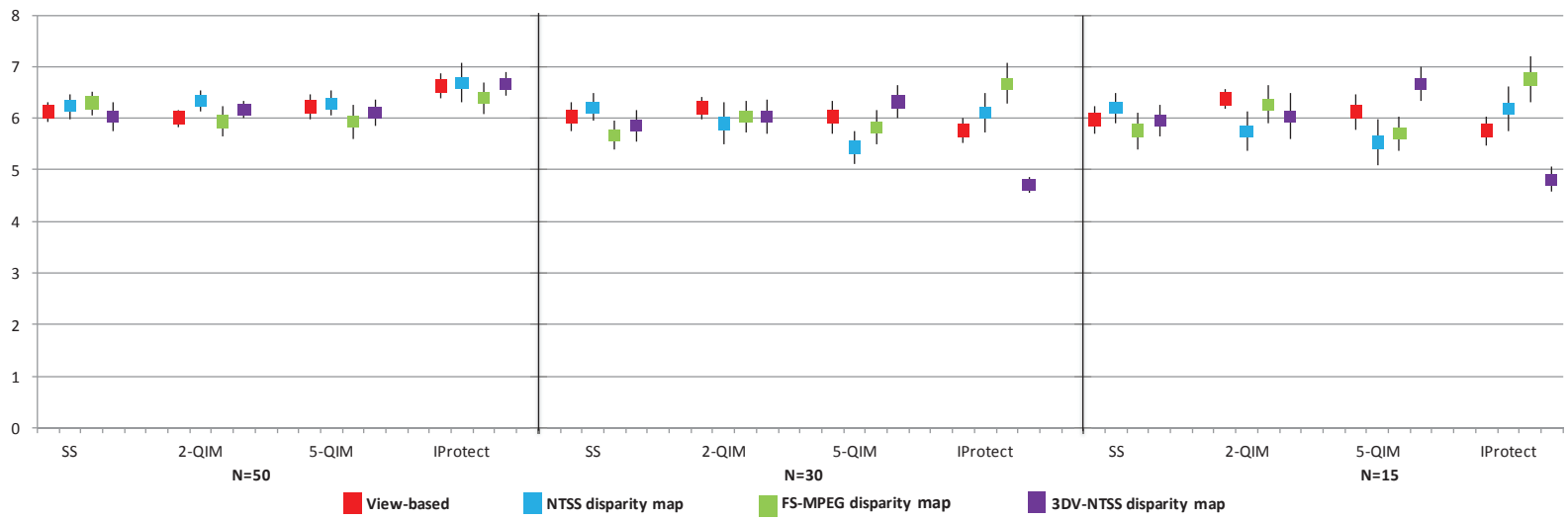
(d)


6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

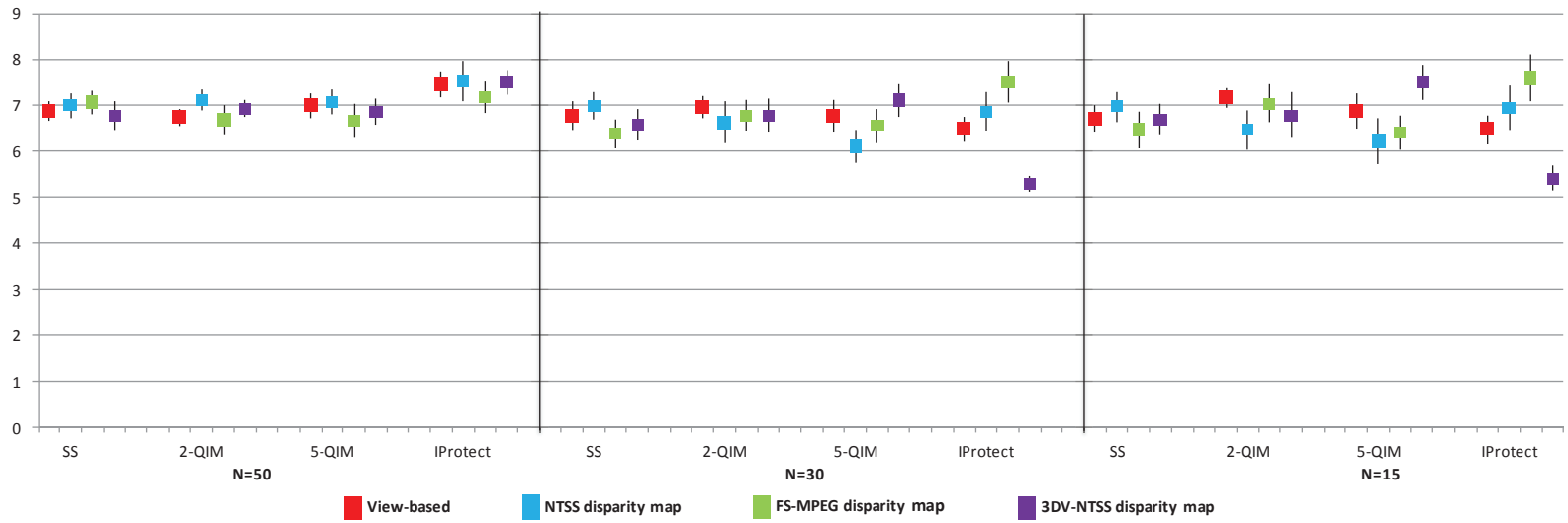
(e)


7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)



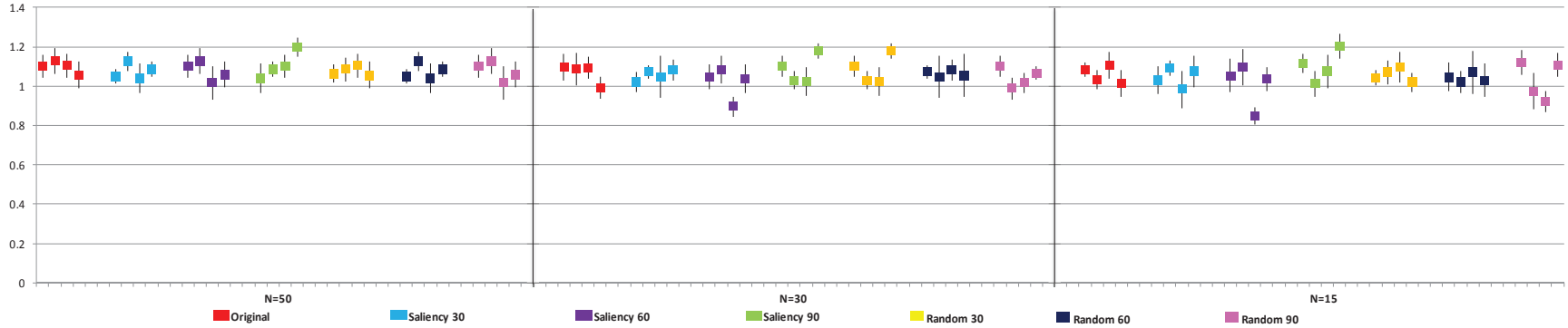
(g)



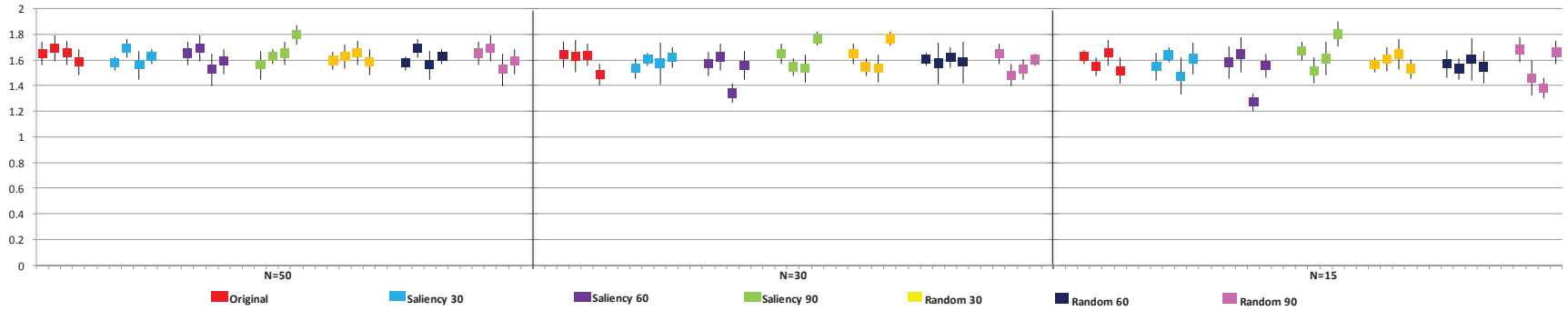
(h)

Figure A1-7: Subjective evaluations high-quality 2D video content ,for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15..*

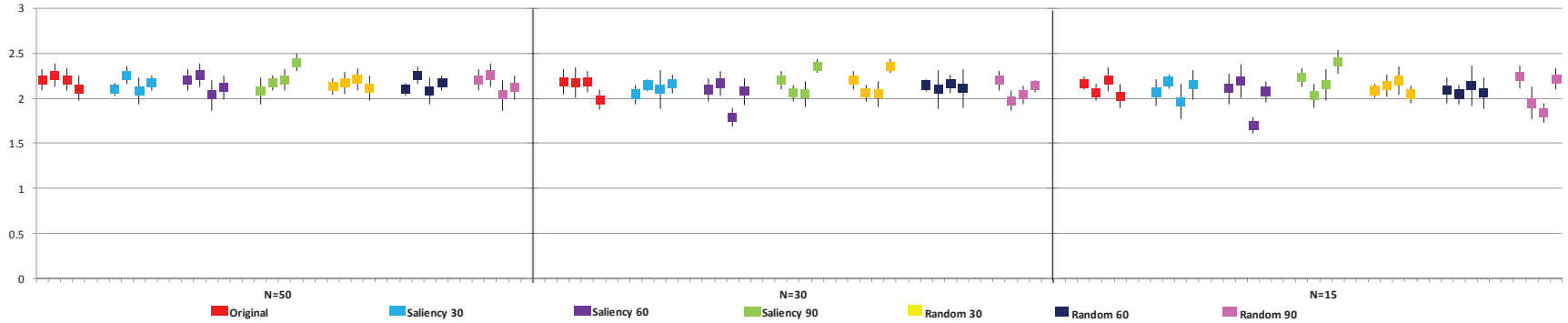
Low quality 2D video content



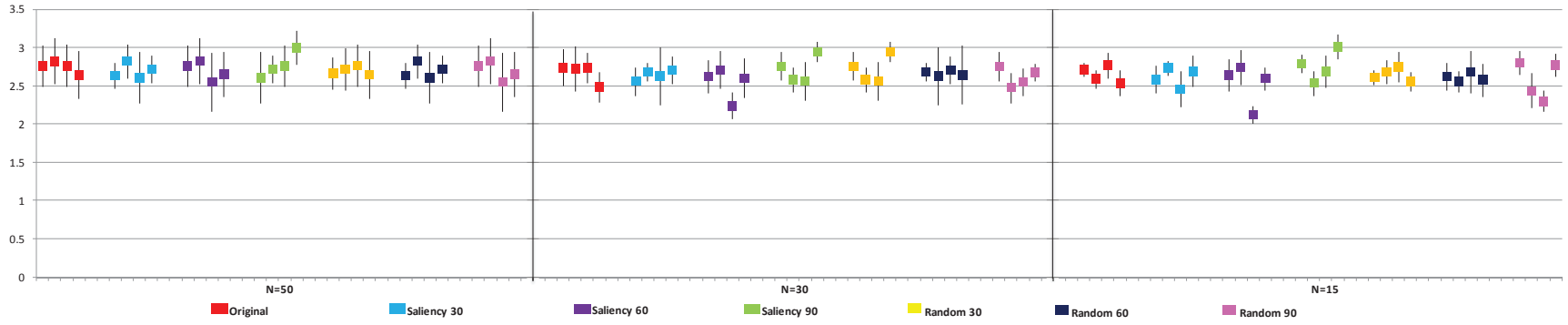
(a)



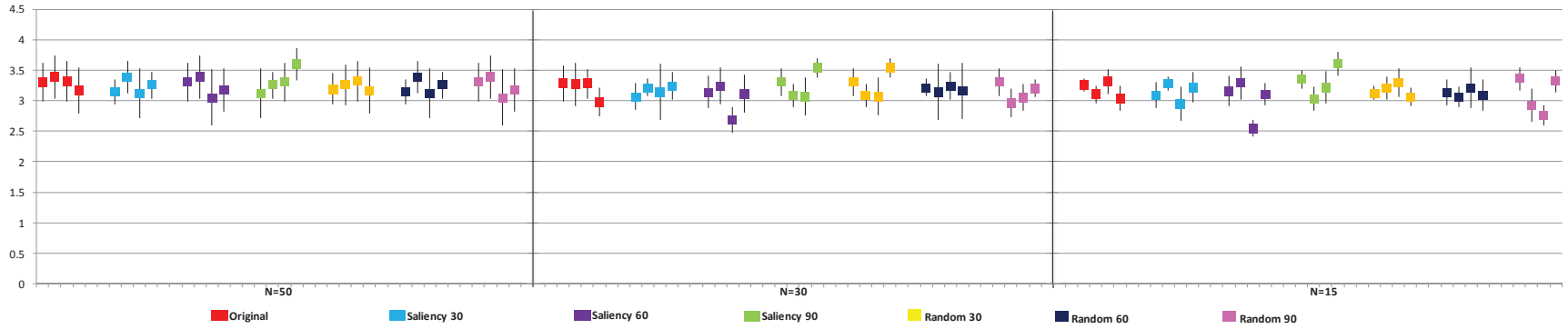
(b)



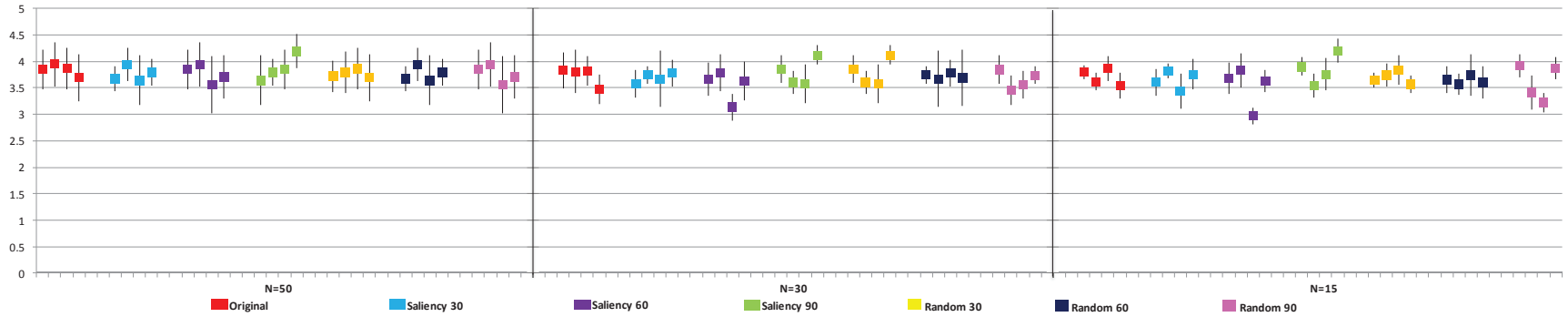
(c)


3.5
3
2.5
2
1.5
1
0.5
0
N=50
N=30
N=15
Original
Saliency 30
Saliency 60
Saliency 90
Random 30
Random 60
Random 90

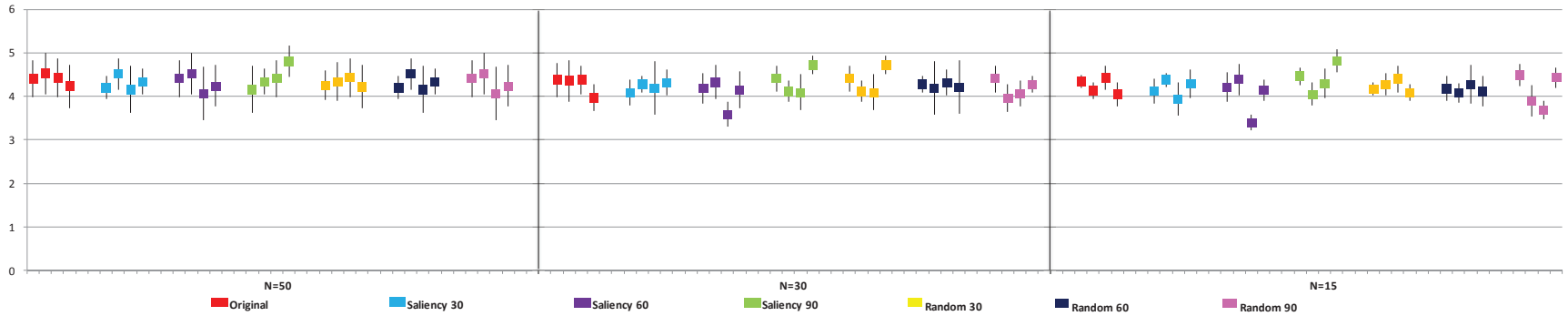
(d)


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
N=50
N=30
N=15
Original
Saliency 30
Saliency 60
Saliency 90
Random 30
Random 60
Random 90

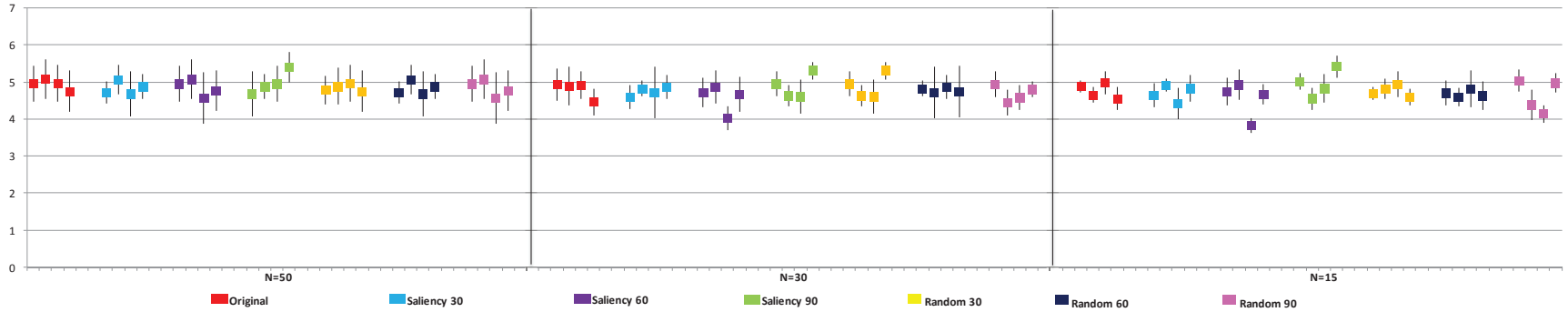
(e)


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
N=50
N=30
N=15
Original
Saliency 30
Saliency 60
Saliency 90
Random 30
Random 60
Random 90

(f)



(g)

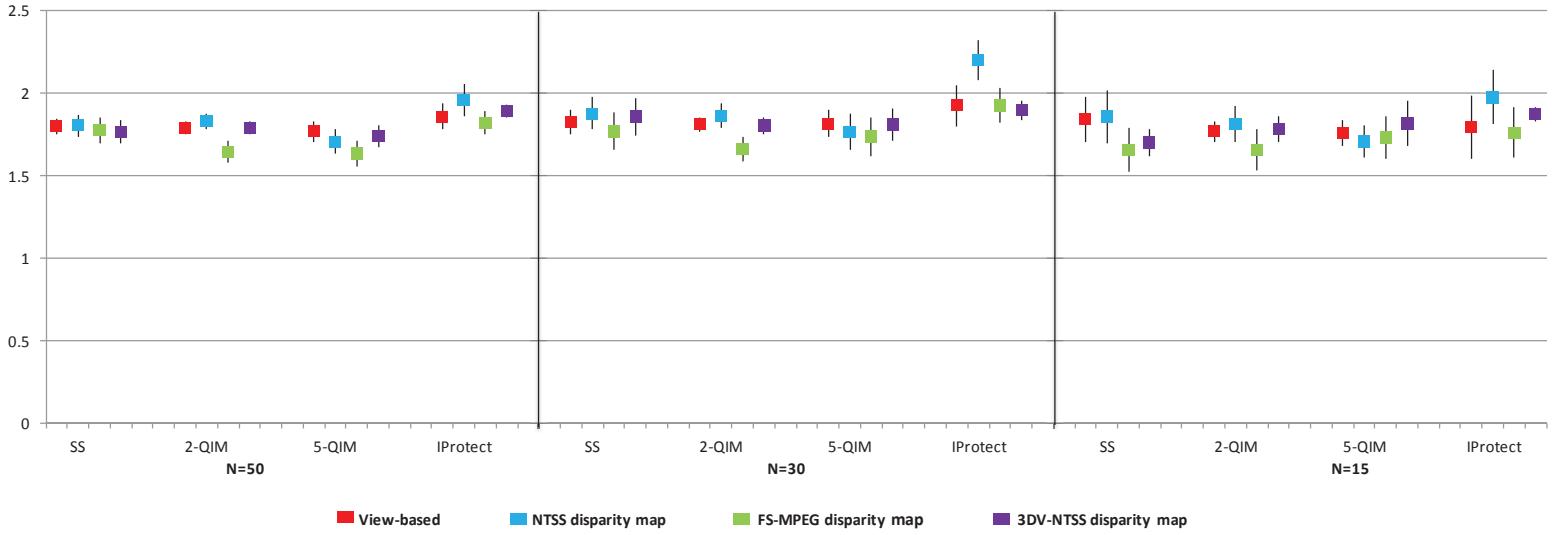


(h)

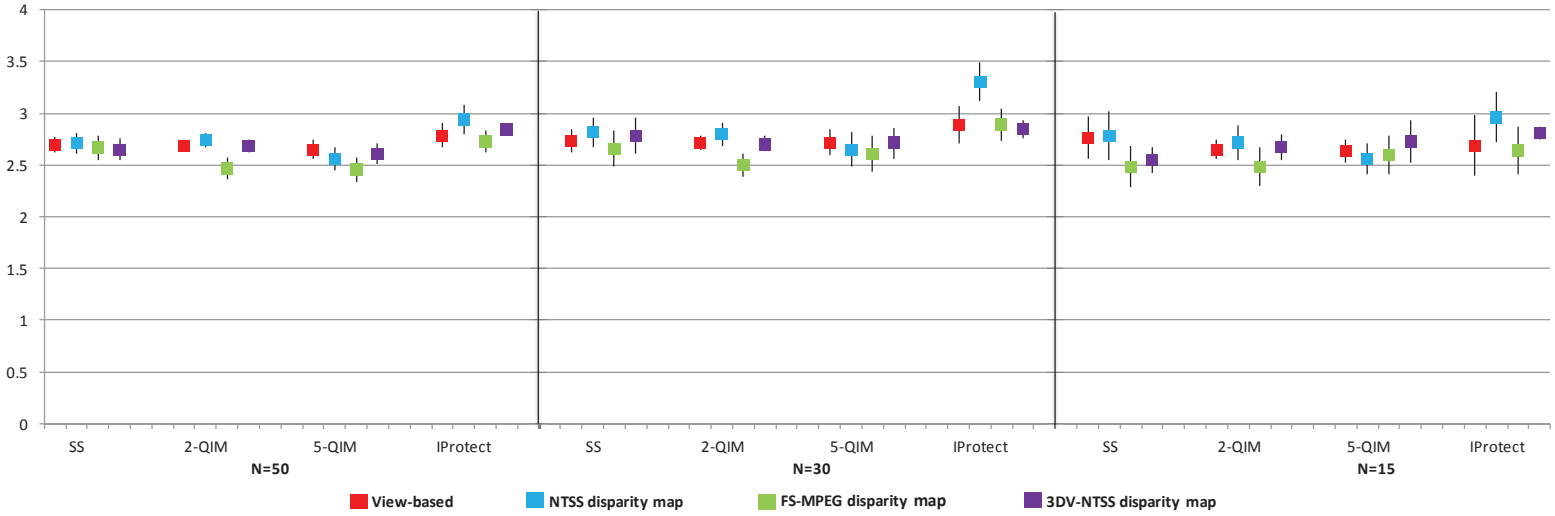
Figure A1-8: Subjective evaluations low-quality 2D video content ,for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*

VII.2. Appendix B Discrete unlabeled scale simulation (Gaussian mixture)

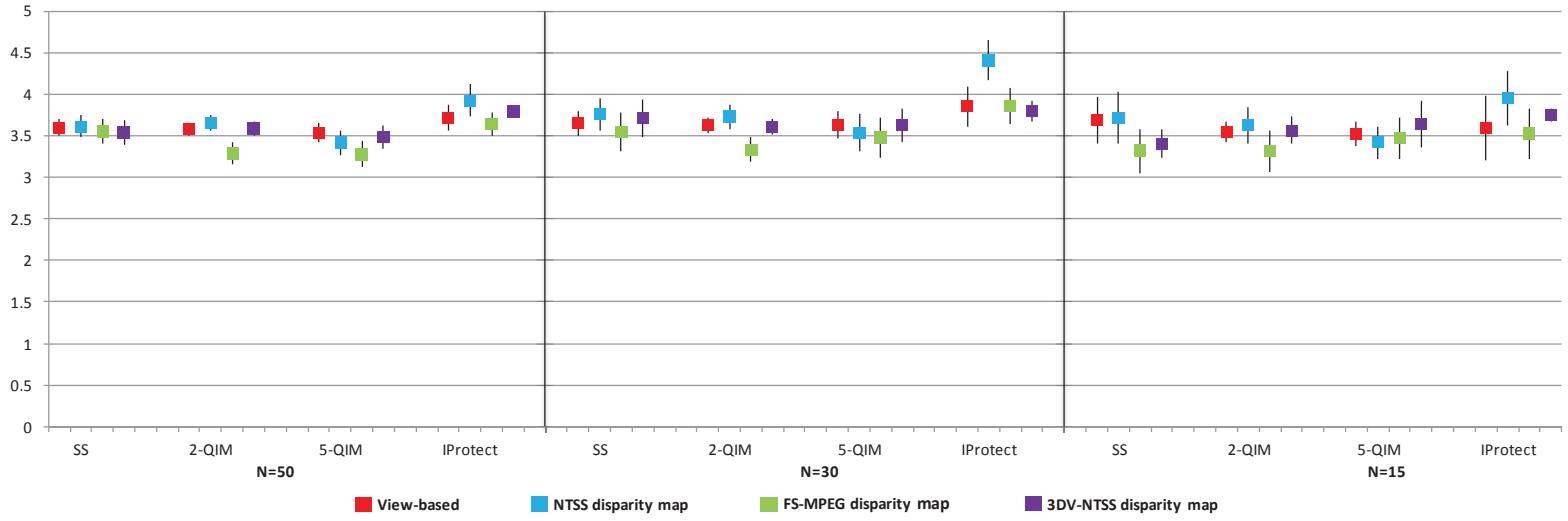
High quality stereoscopic video content (Image Quality)



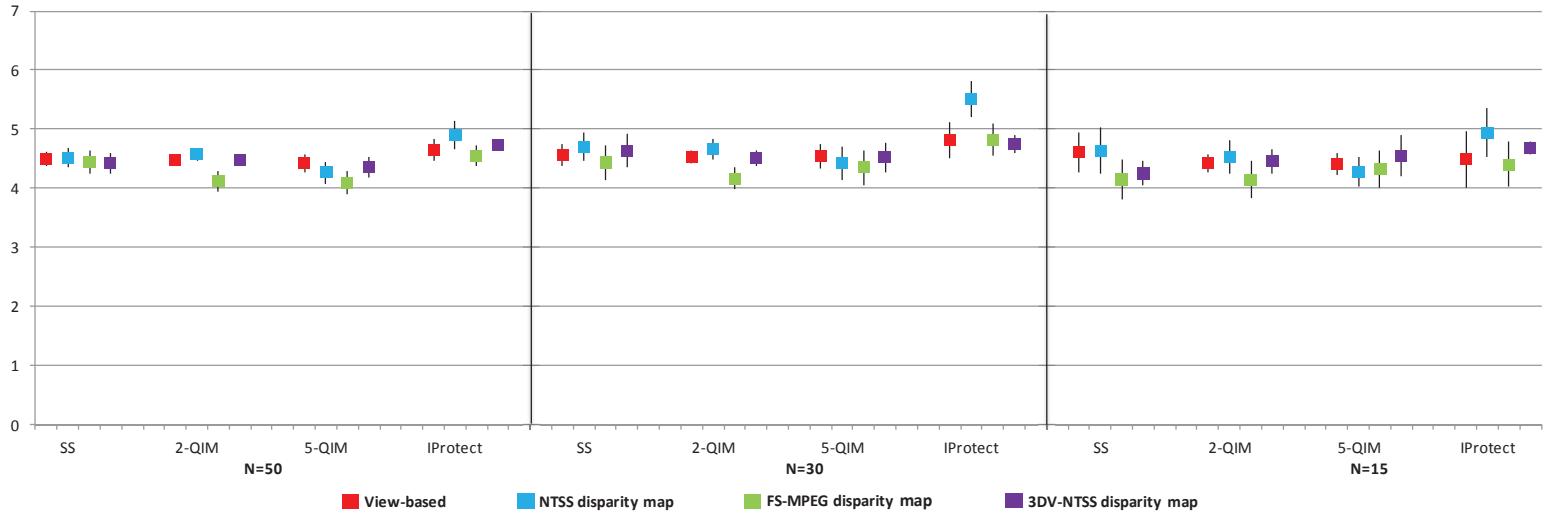
(a)



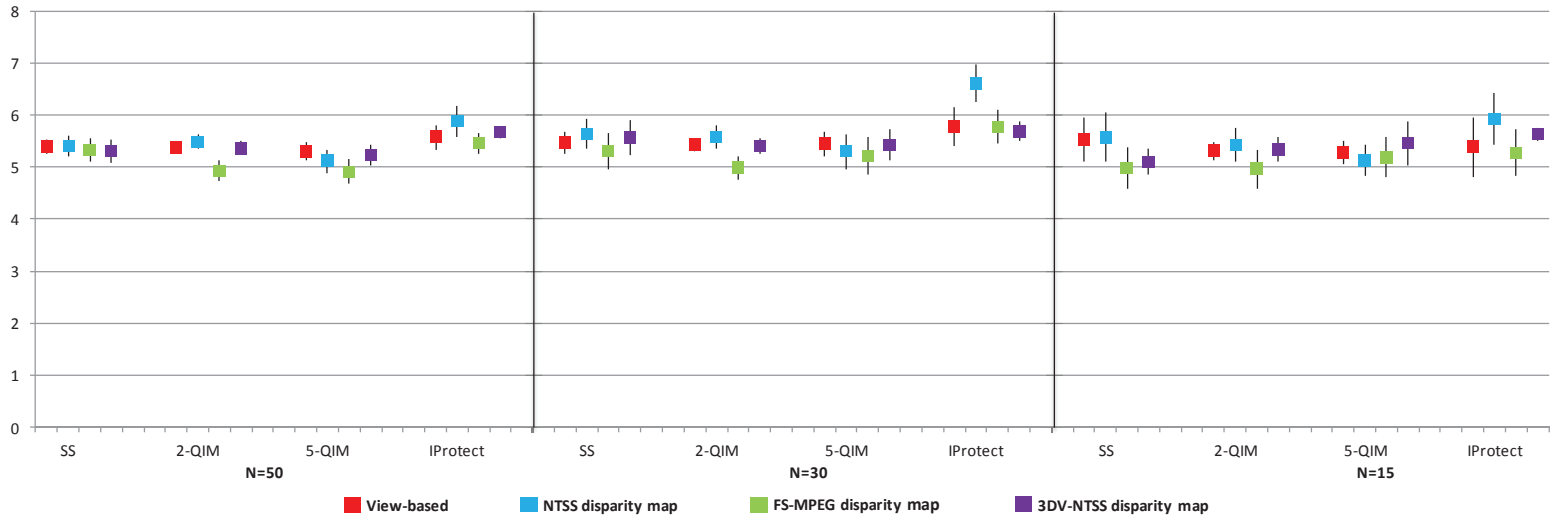
(b)


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

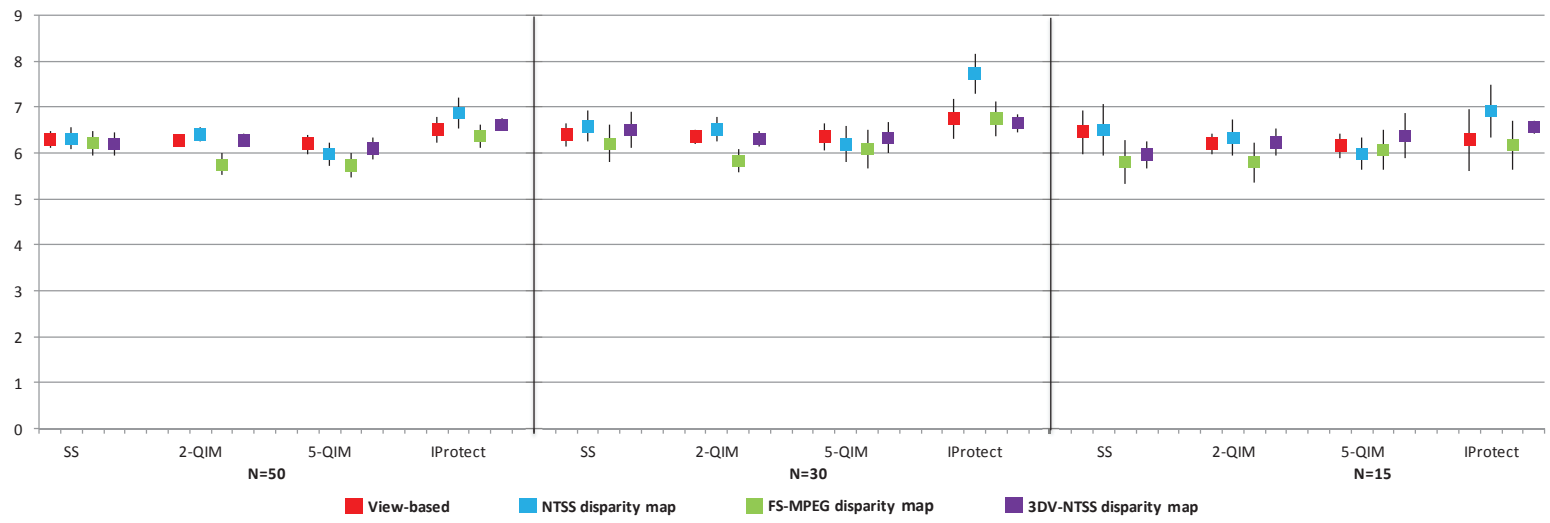
(c)


7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

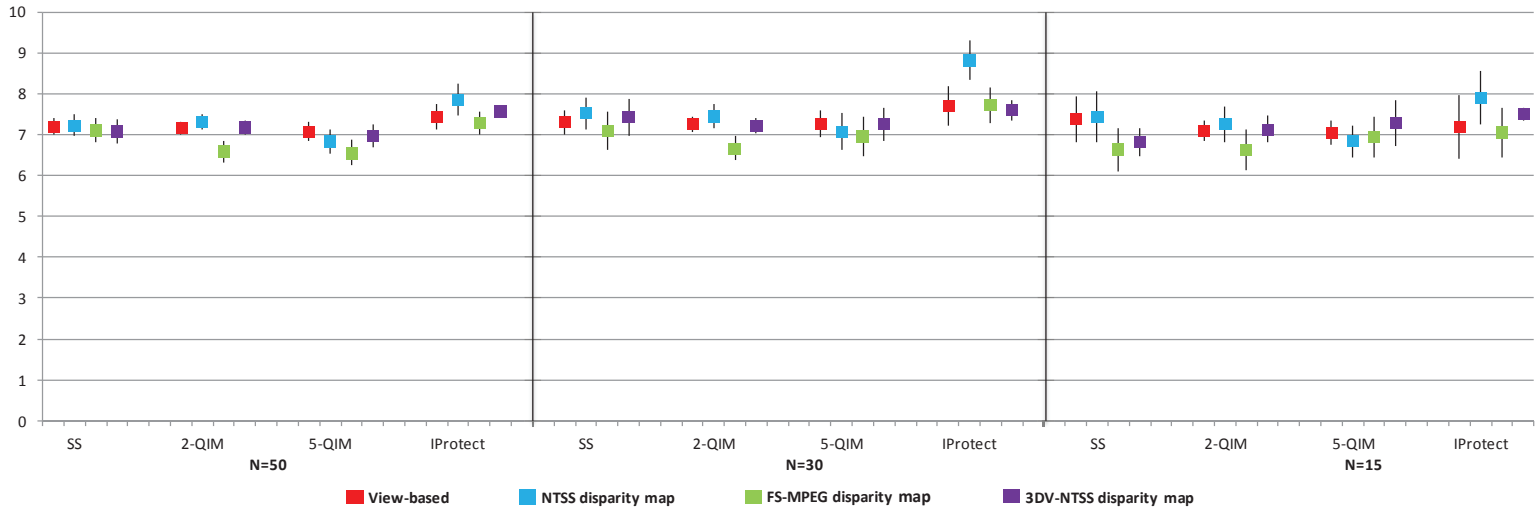
(d)


8
7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
SS
2-QIM
5-QIM
IProtect
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

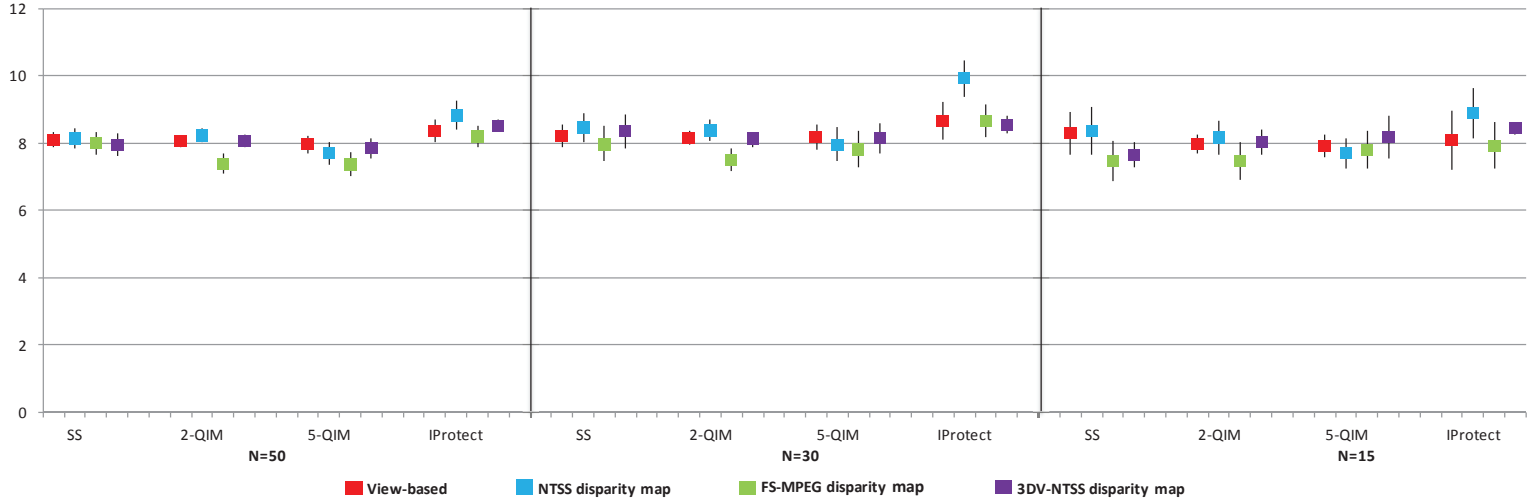
(e)


9
8
7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
SS
2-QIM
5-QIM
IProtect
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)



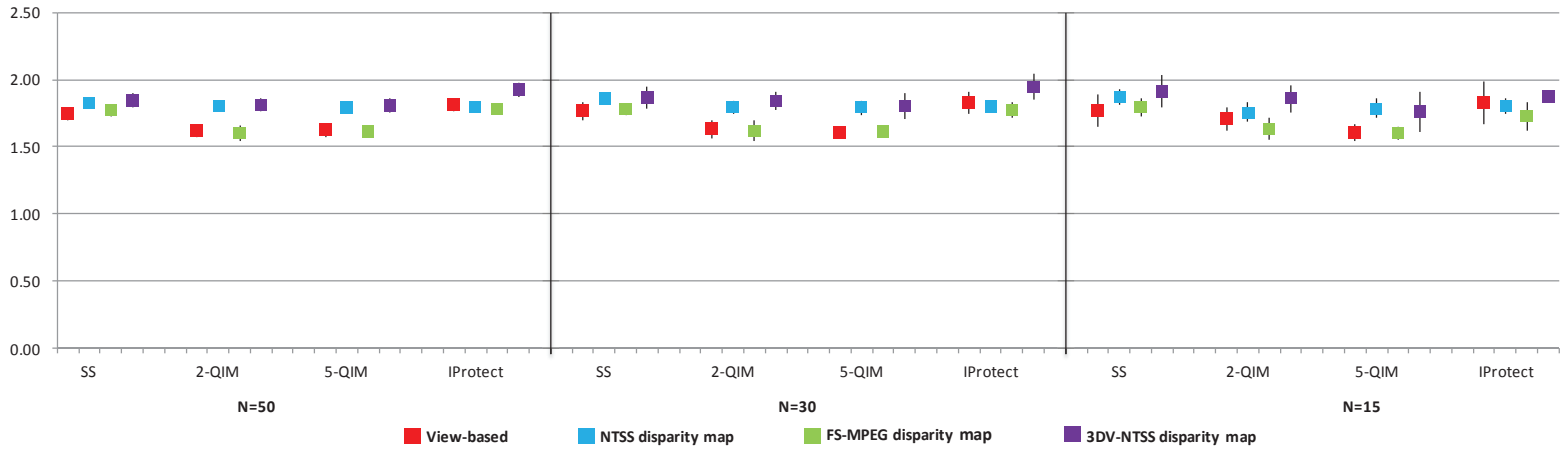
(g)



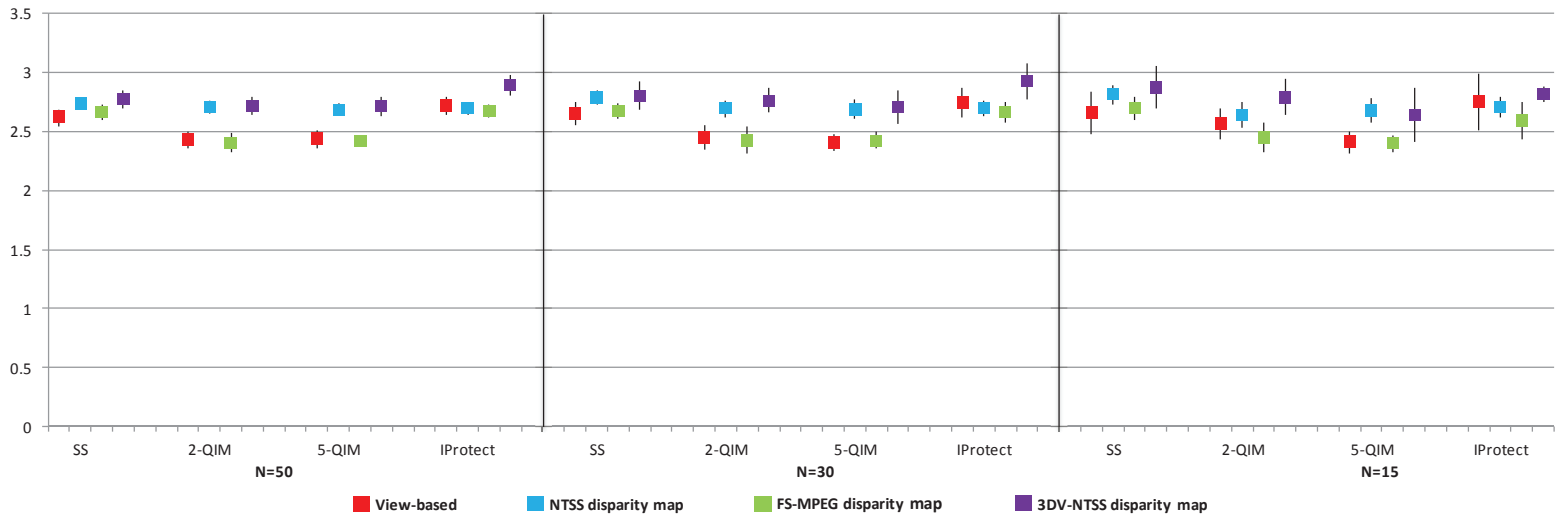
(h)

Figure A2-1: Subjective evaluations for high-quality stereoscopic video content (Image Quality), for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*

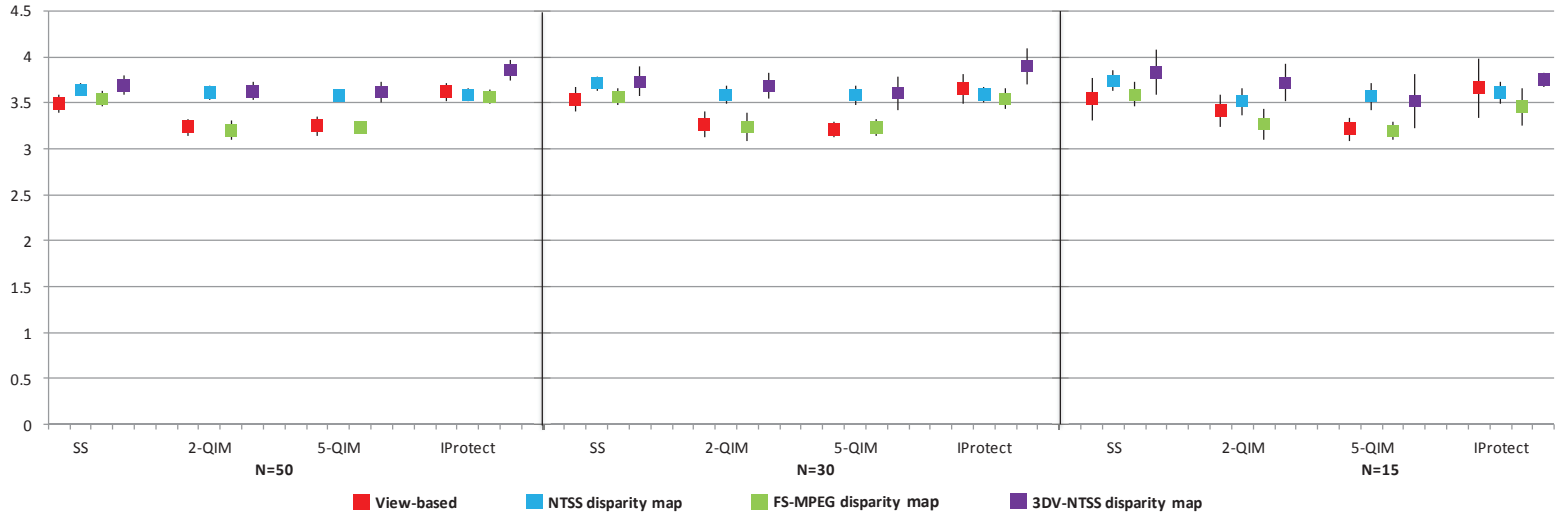
High quality stereoscopic video content (Depth Perception)



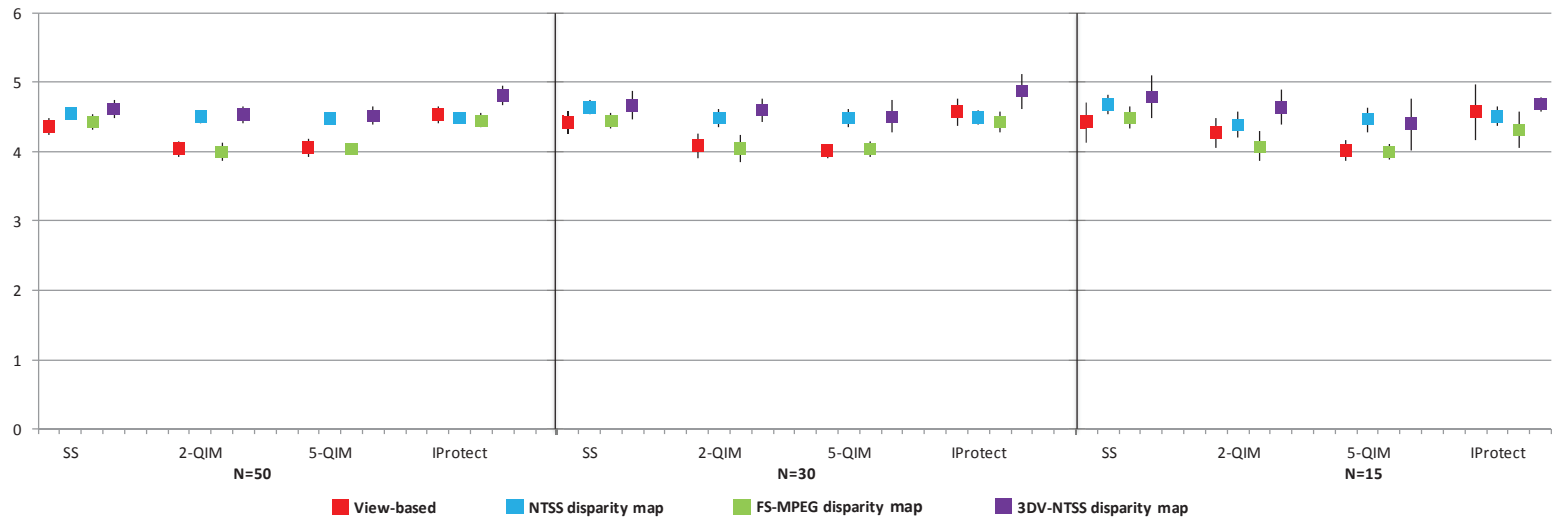
(a)



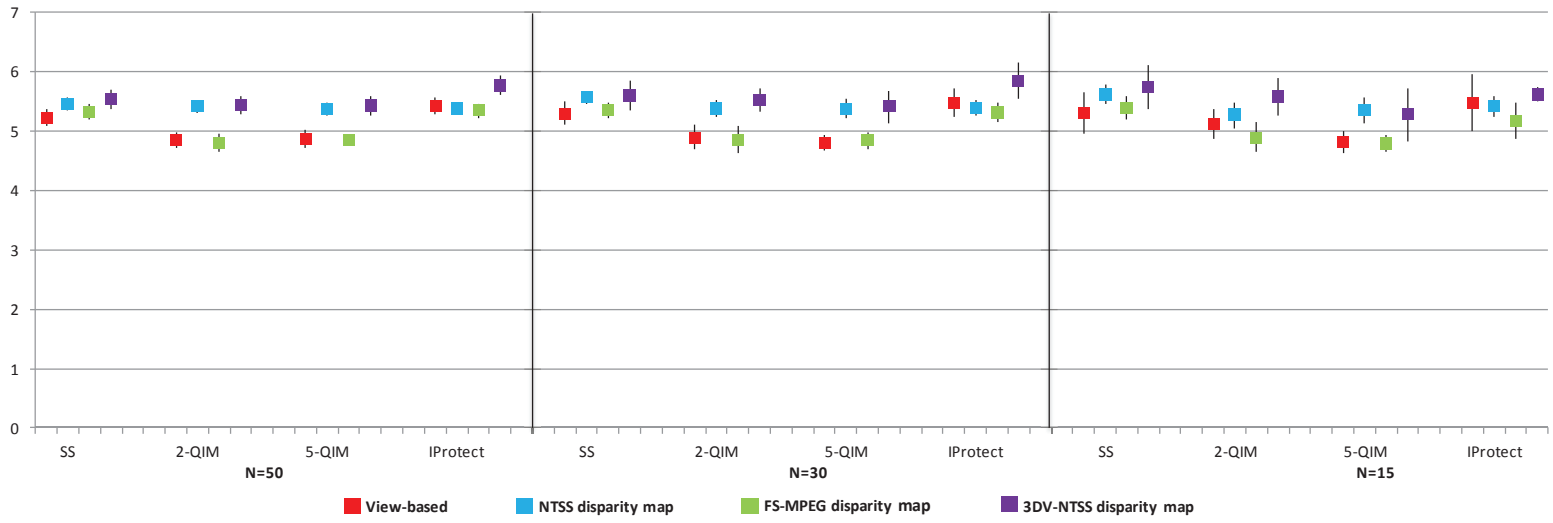
(b)


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

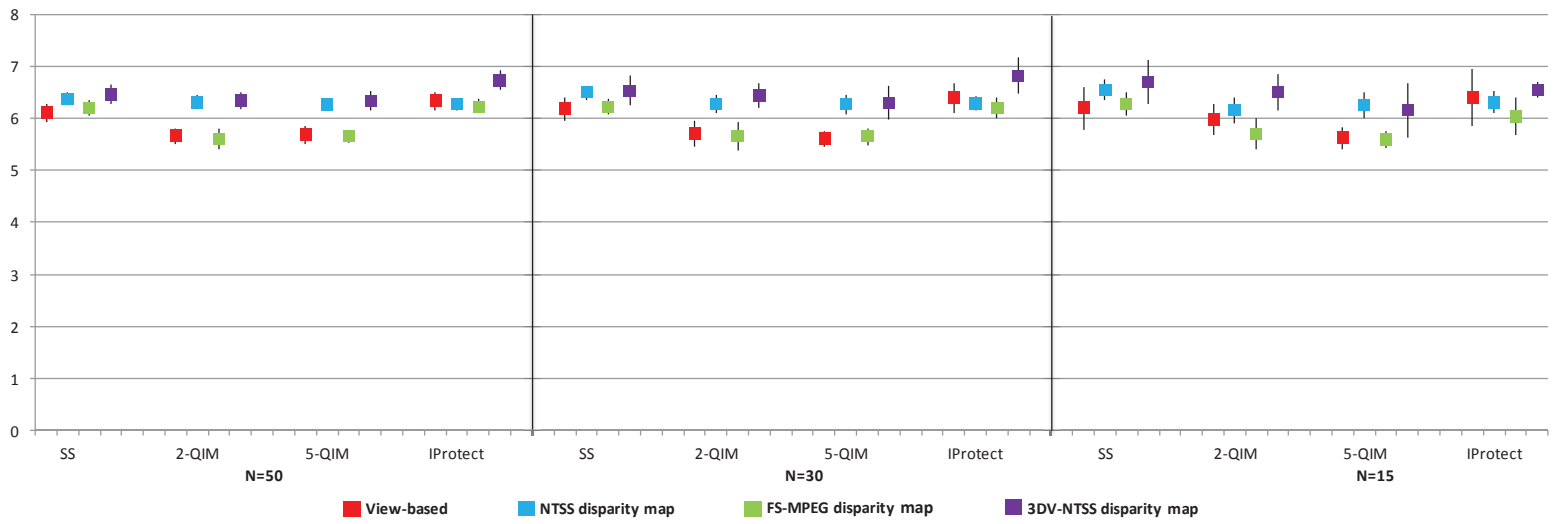
(c)


6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

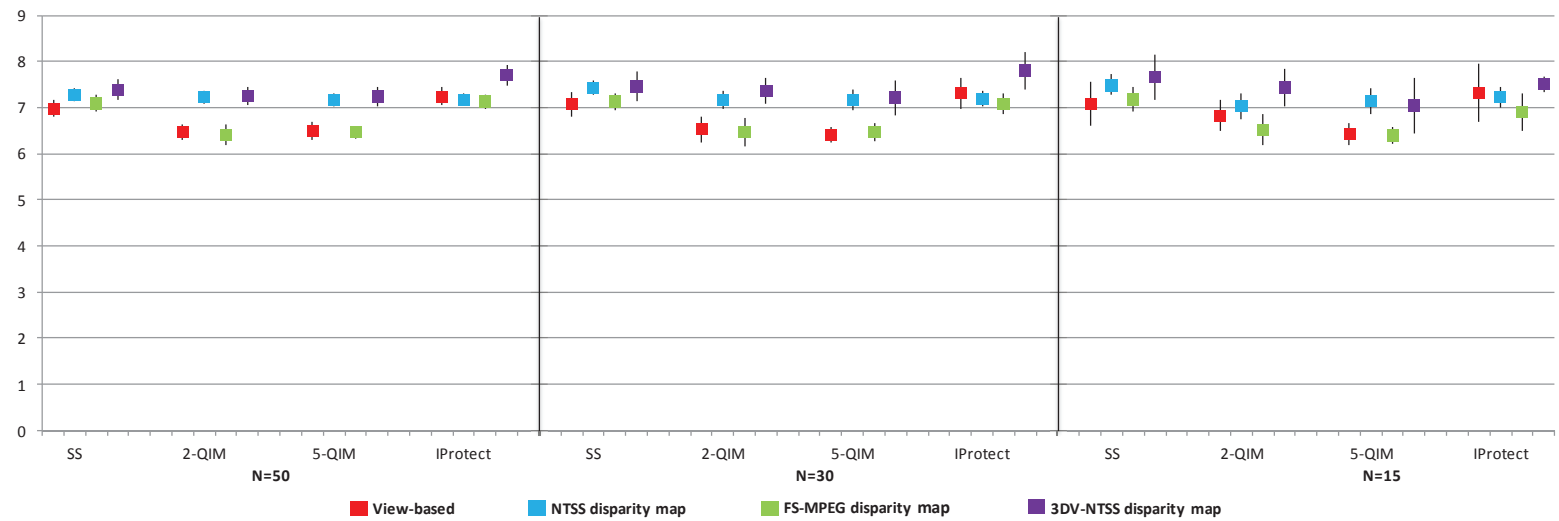
(d)


7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
SS
2-QIM
5-QIM
IProtect
N=30
SS
2-QIM
5-QIM
IProtect
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

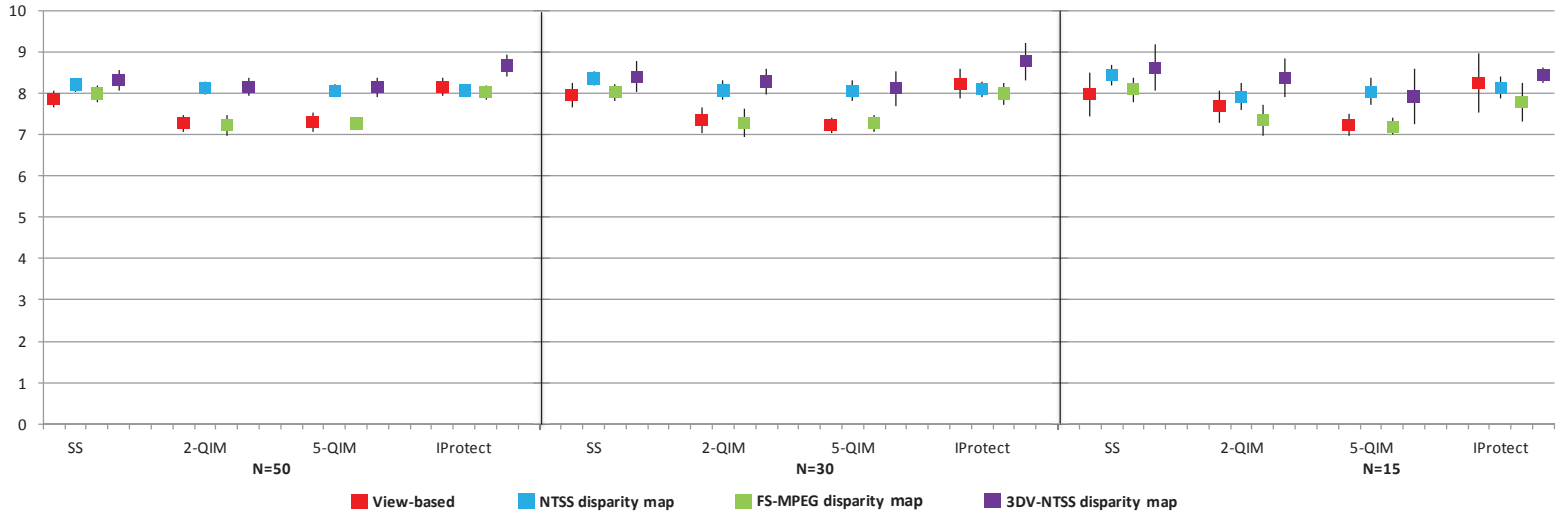
(e)


8
7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
SS
2-QIM
5-QIM
IProtect
N=30
SS
2-QIM
5-QIM
IProtect
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)



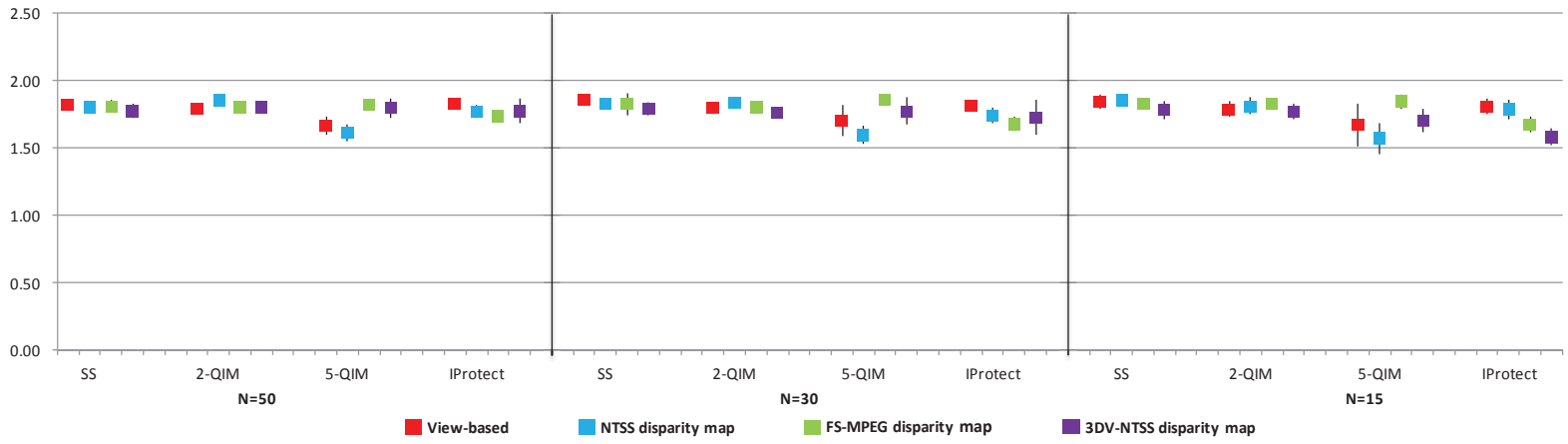
(g)



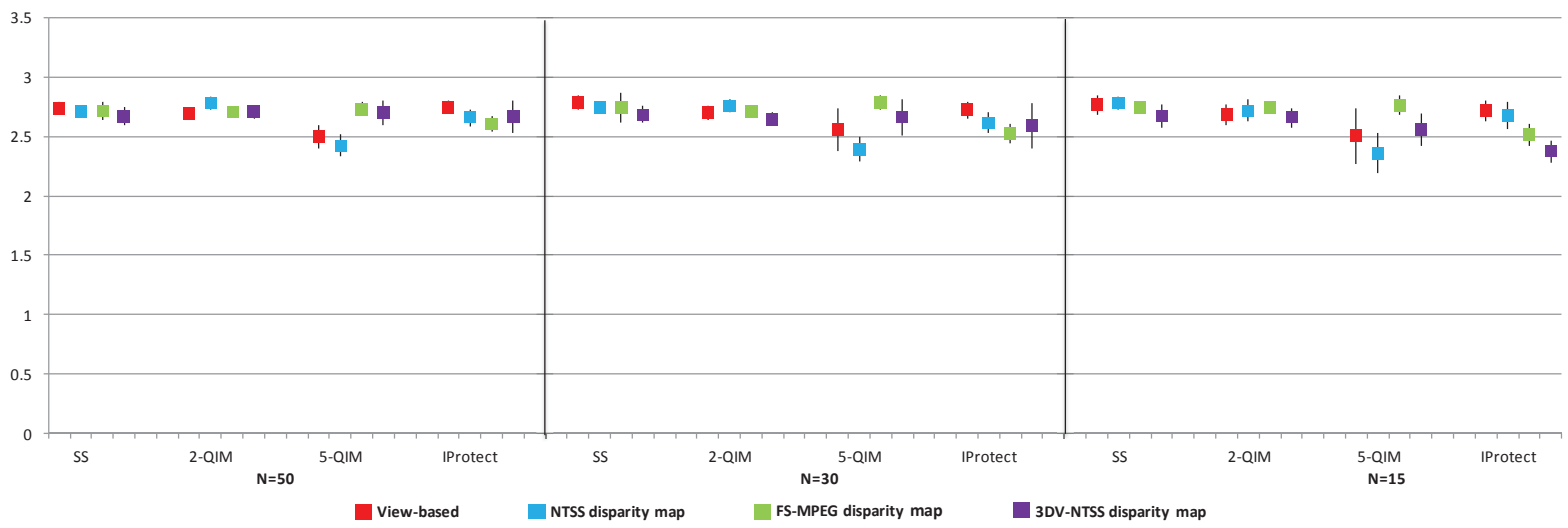
(h)

Figure A2-2: Subjective evaluations for high-quality stereoscopic video content (Depth Perception), for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*

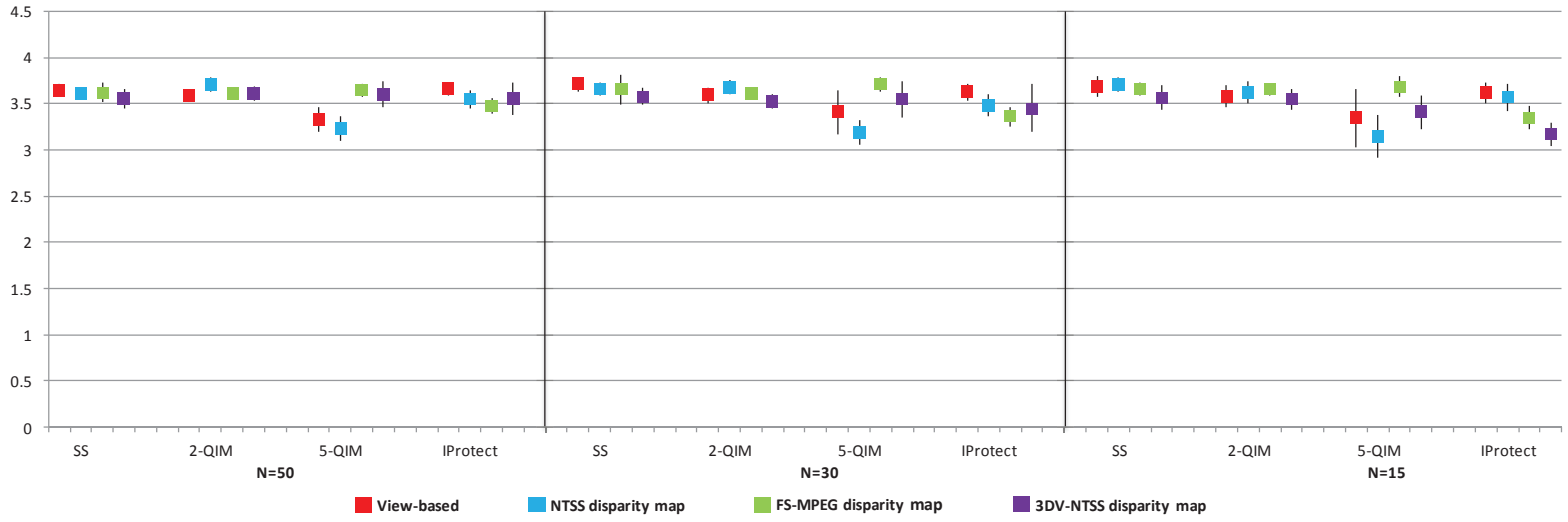
High quality stereoscopic video content (Visual Comfort)



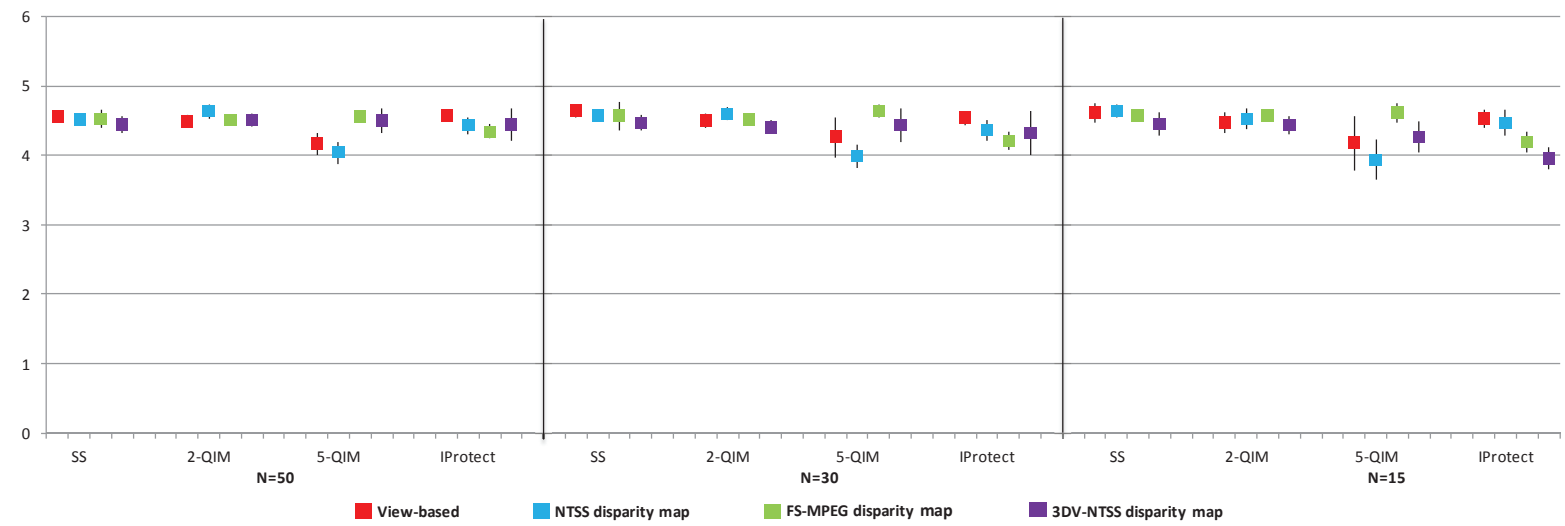
(a)



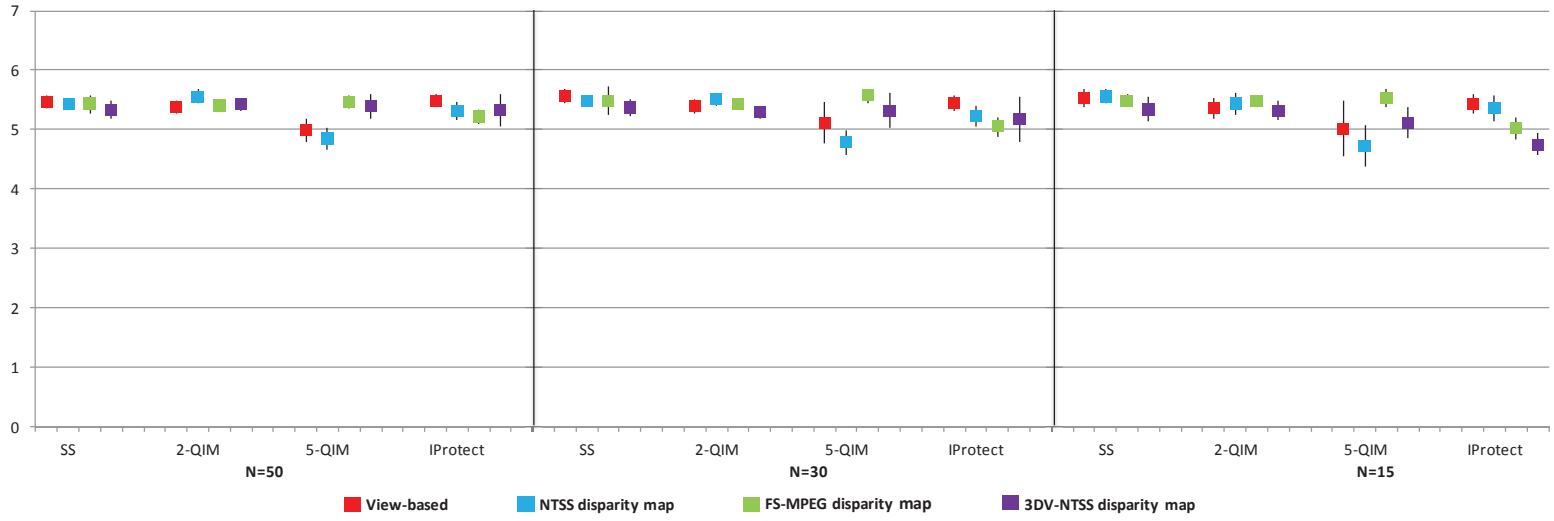
(b)


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

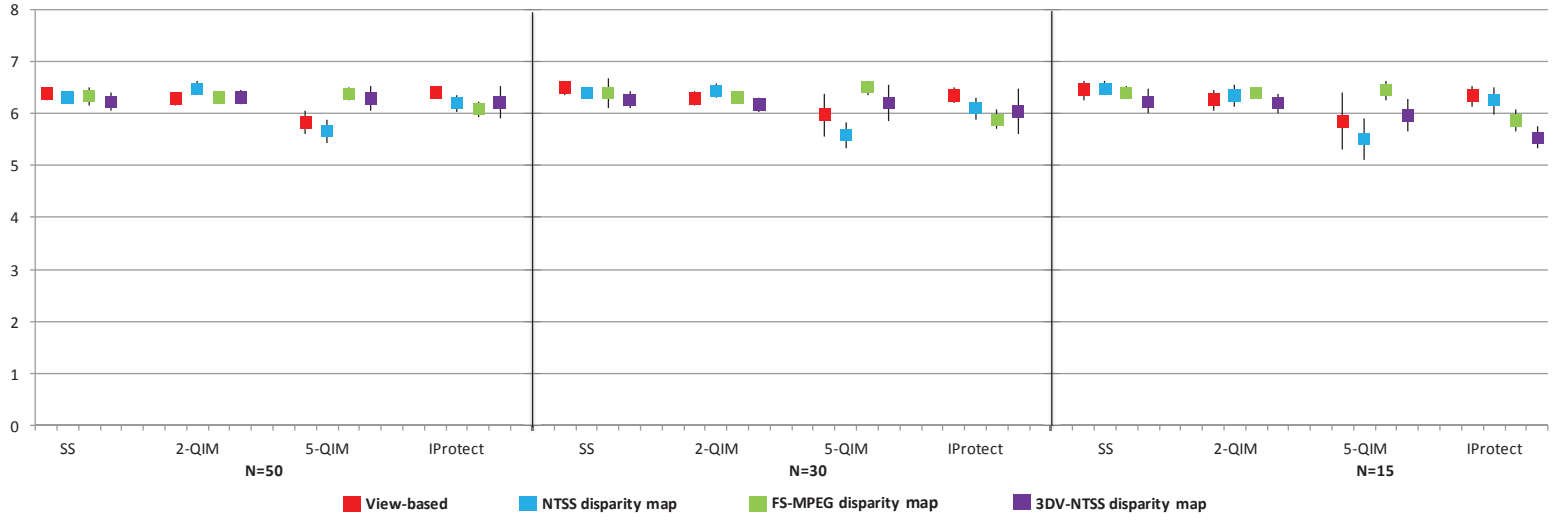
(c)


6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

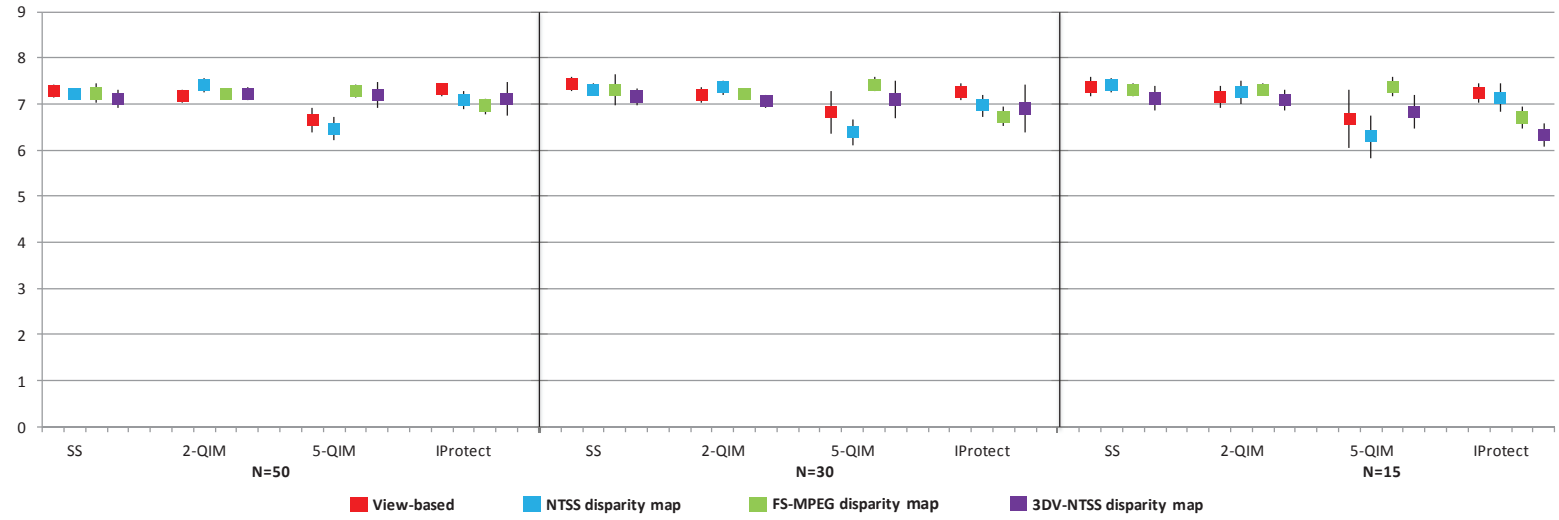
(d)


7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
SS
2-QIM
5-QIM
IProtect
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

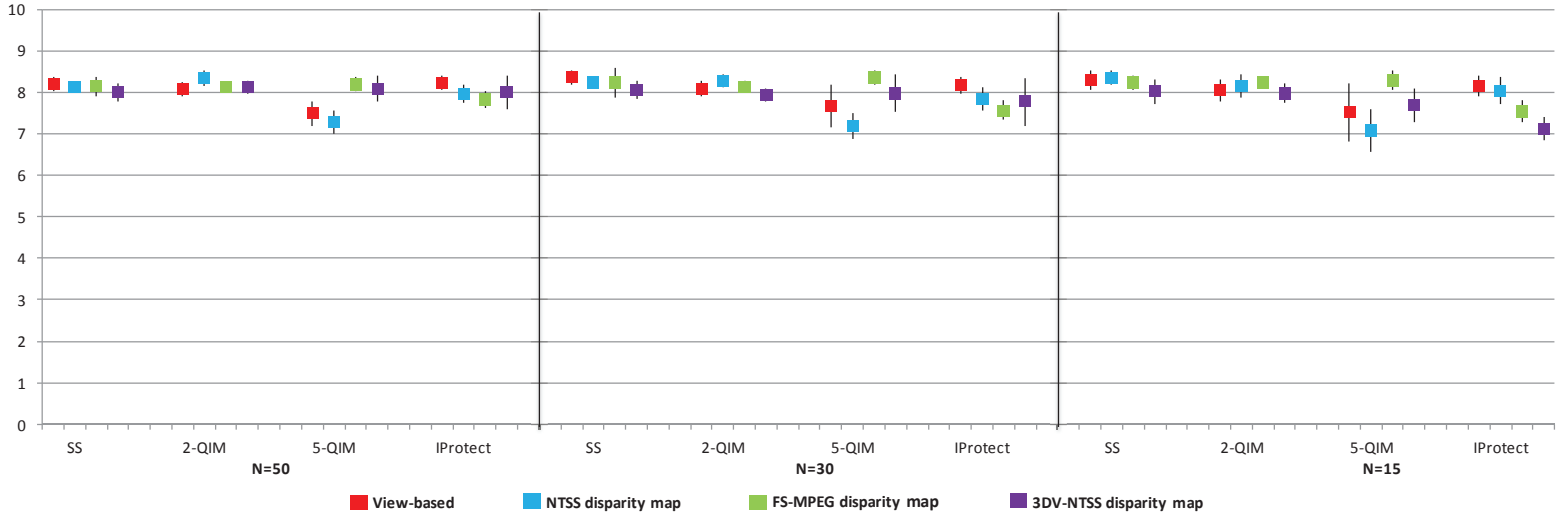
(e)


8
7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
SS
2-QIM
5-QIM
IProtect
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)

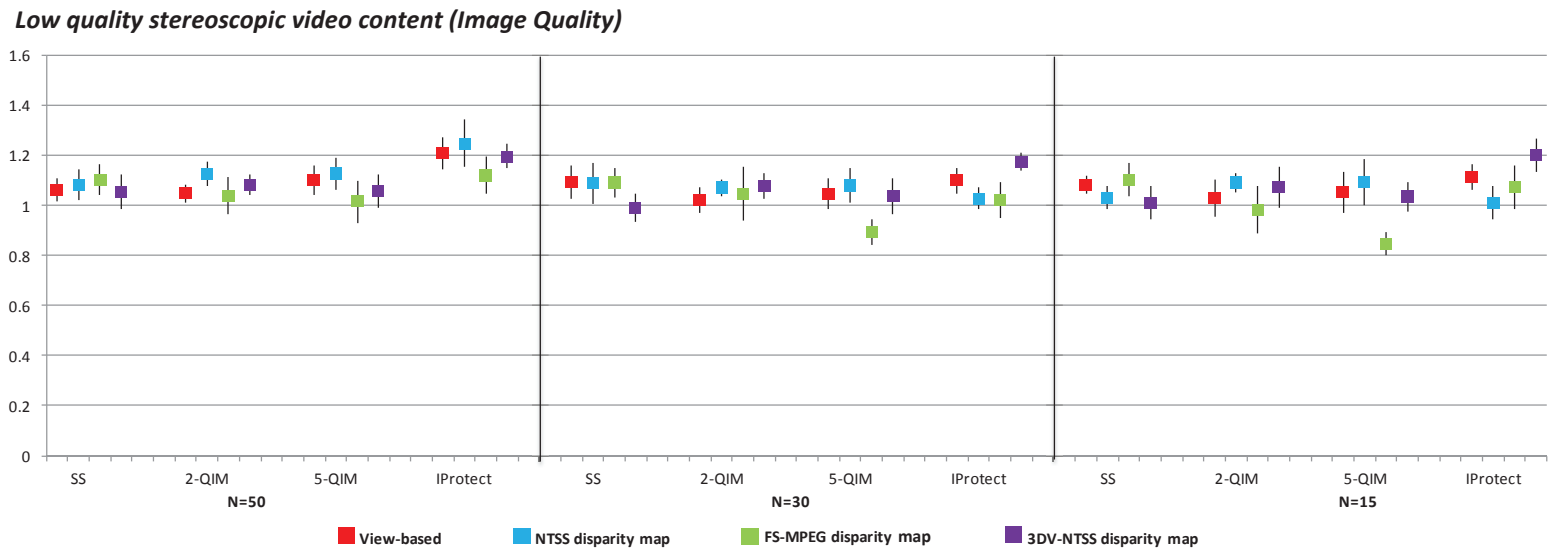


(g)

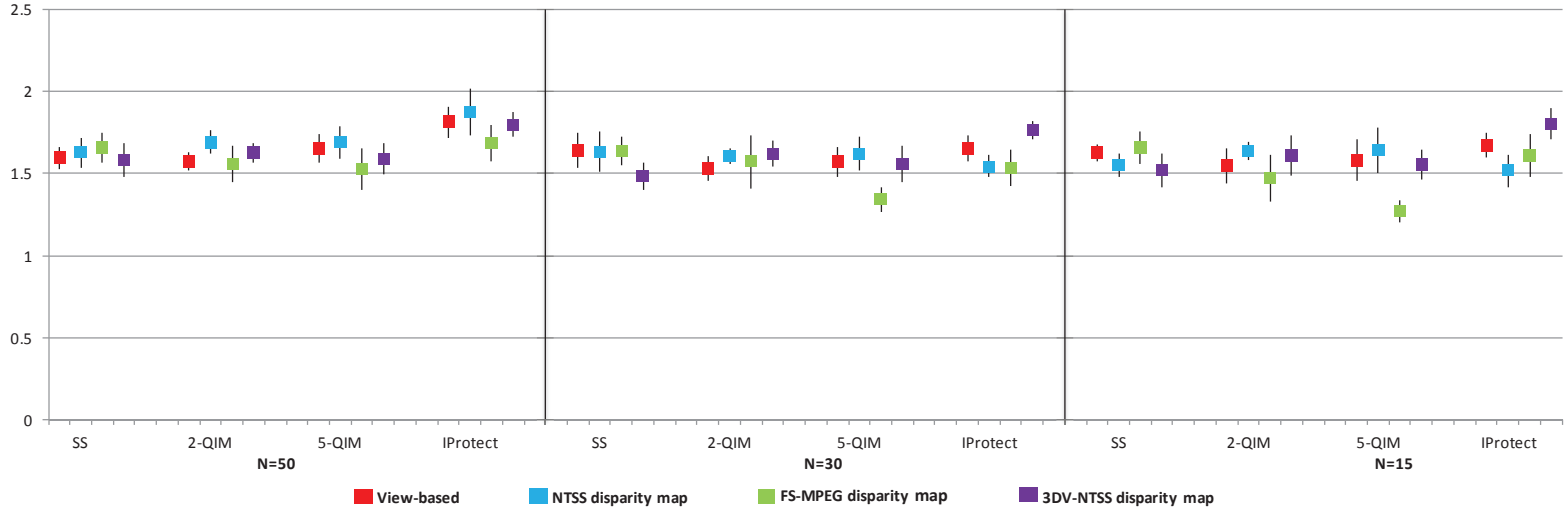


(h)

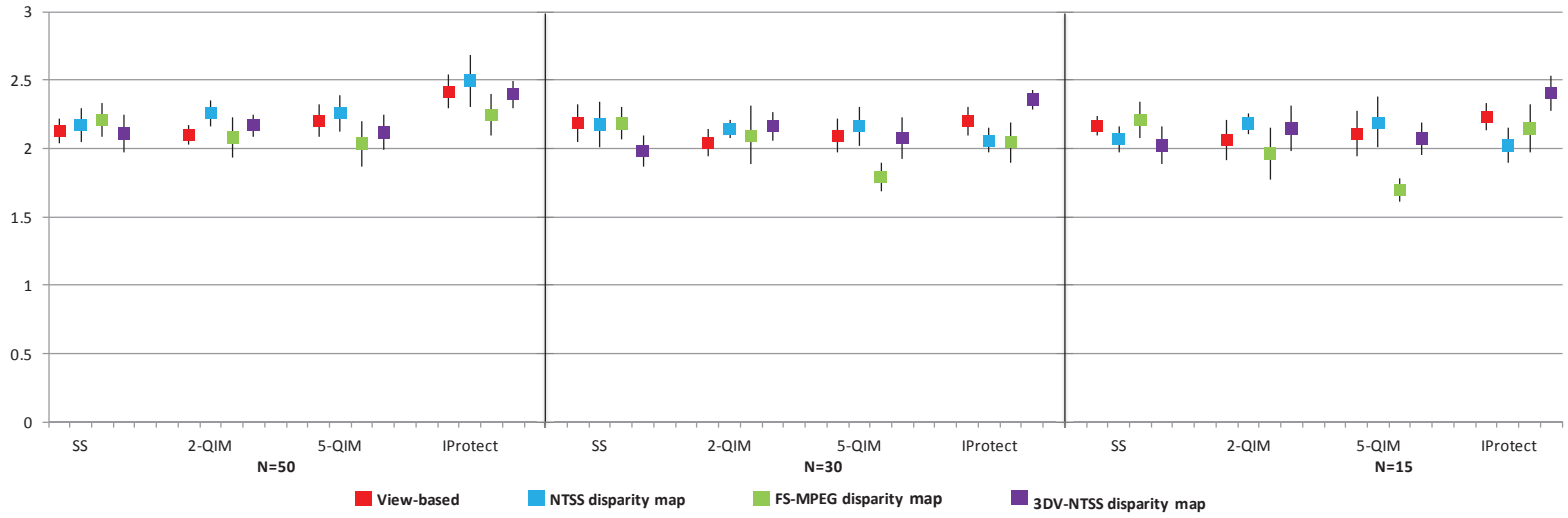
Figure A2-3: Subjective evaluations for high-quality stereoscopic video content (Visual Comfort), for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*


Low quality stereoscopic video content (Image Quality)
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

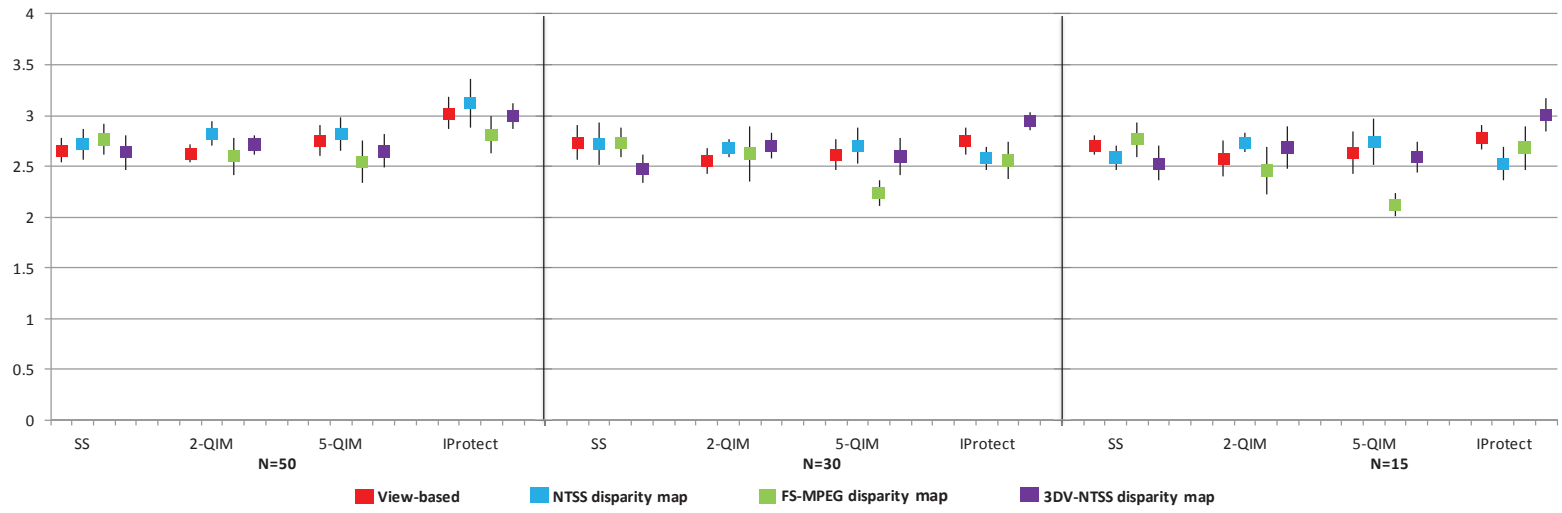
(a)


2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

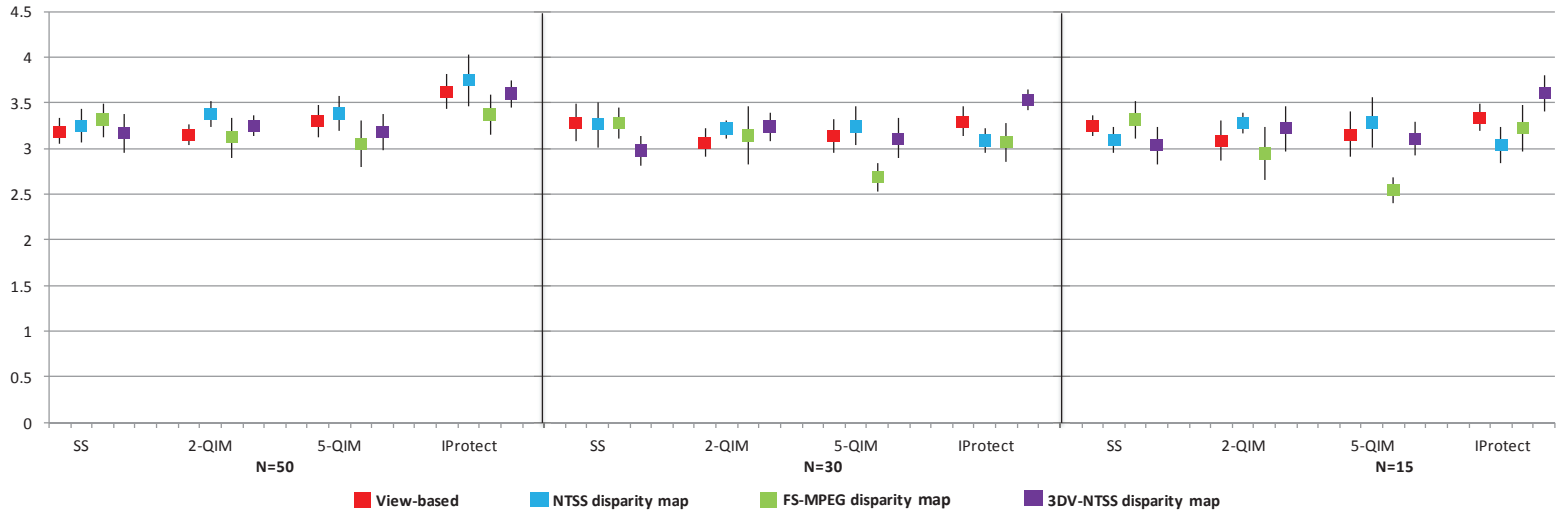
(b)


3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

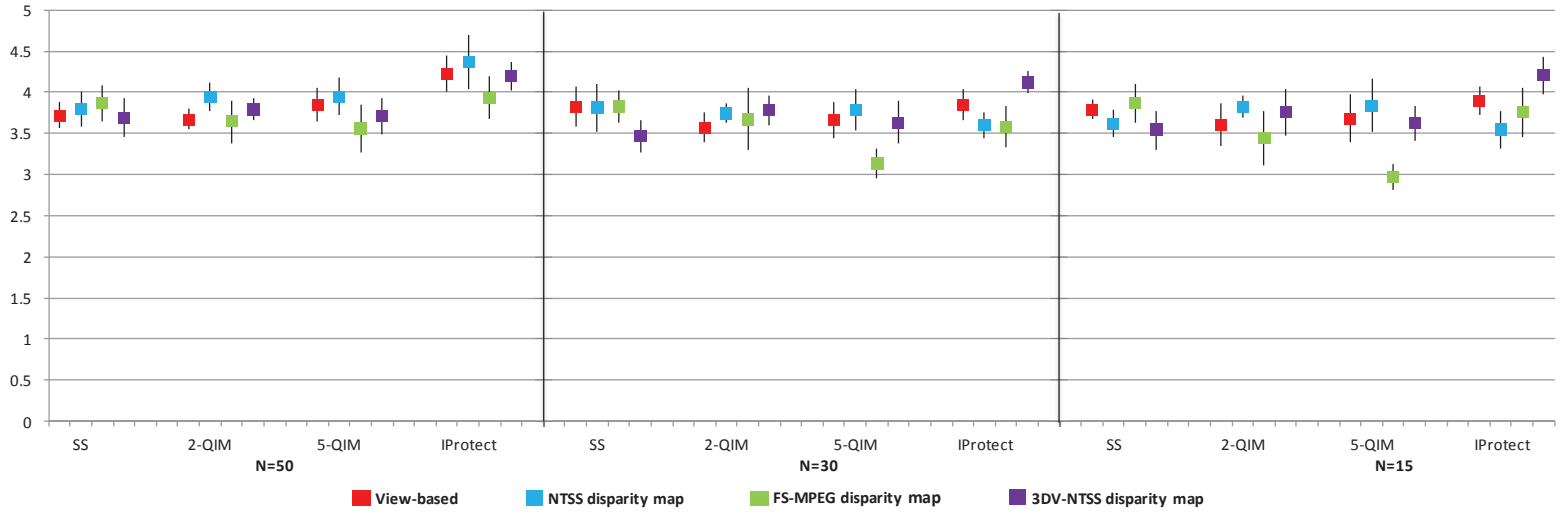
(c)


4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

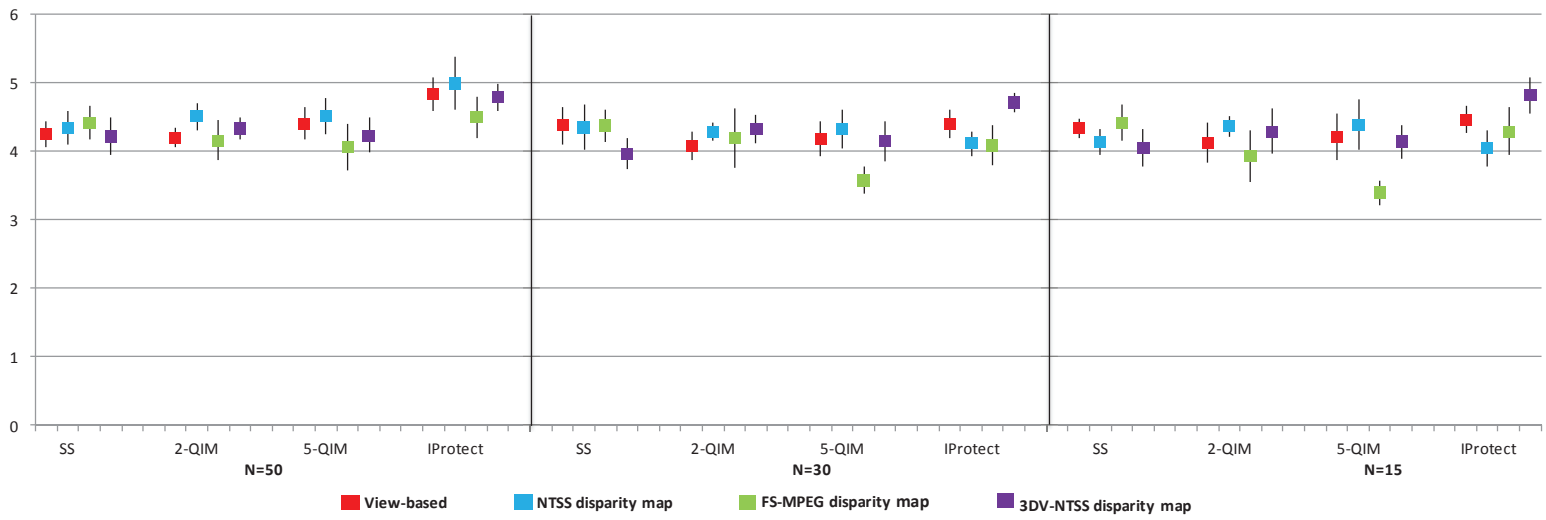
(d)


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

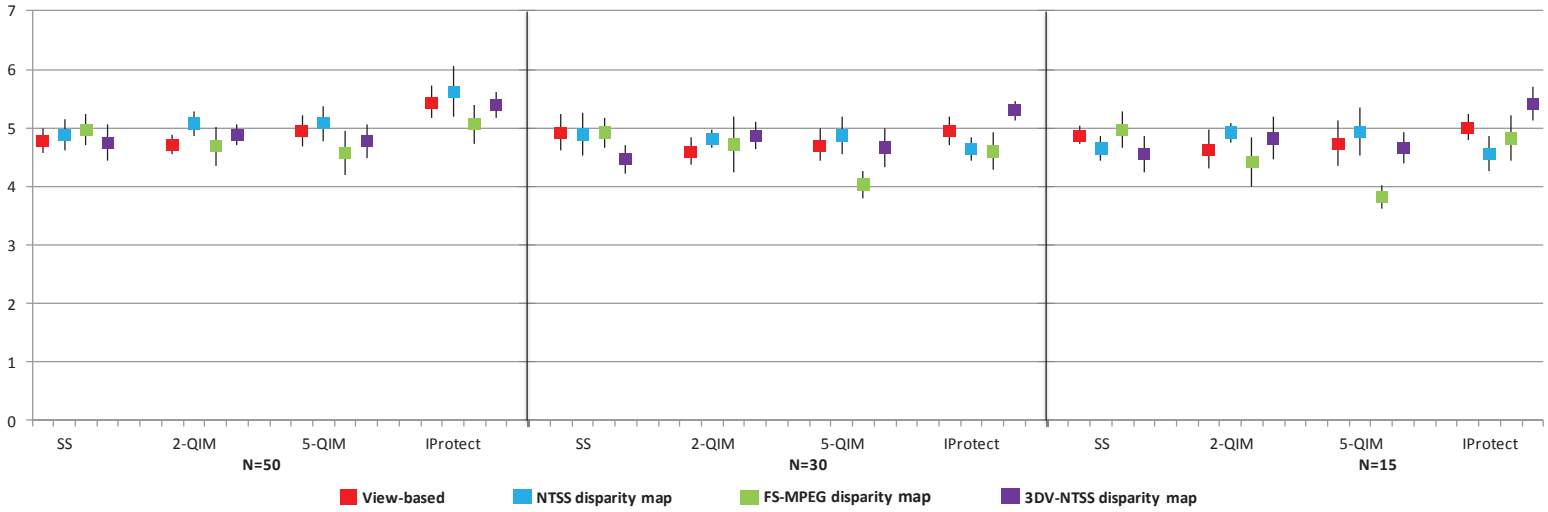
(e)


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)



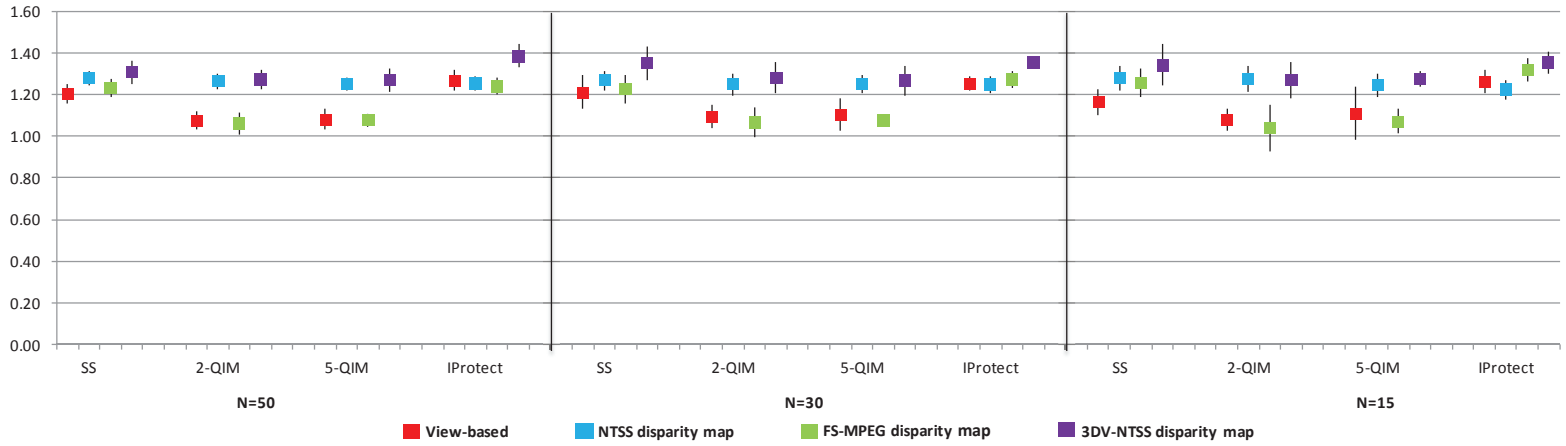
(g)



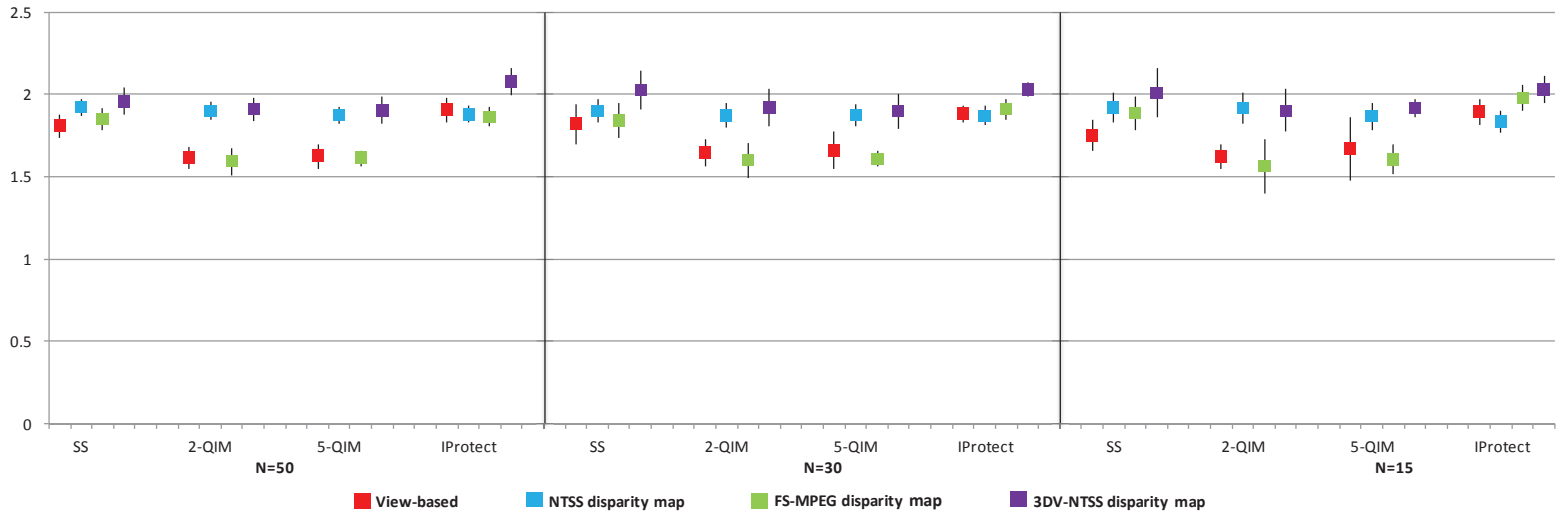
(h)

Figure A2-4: Subjective evaluations low-quality stereoscopic video content (Image Quality), for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*

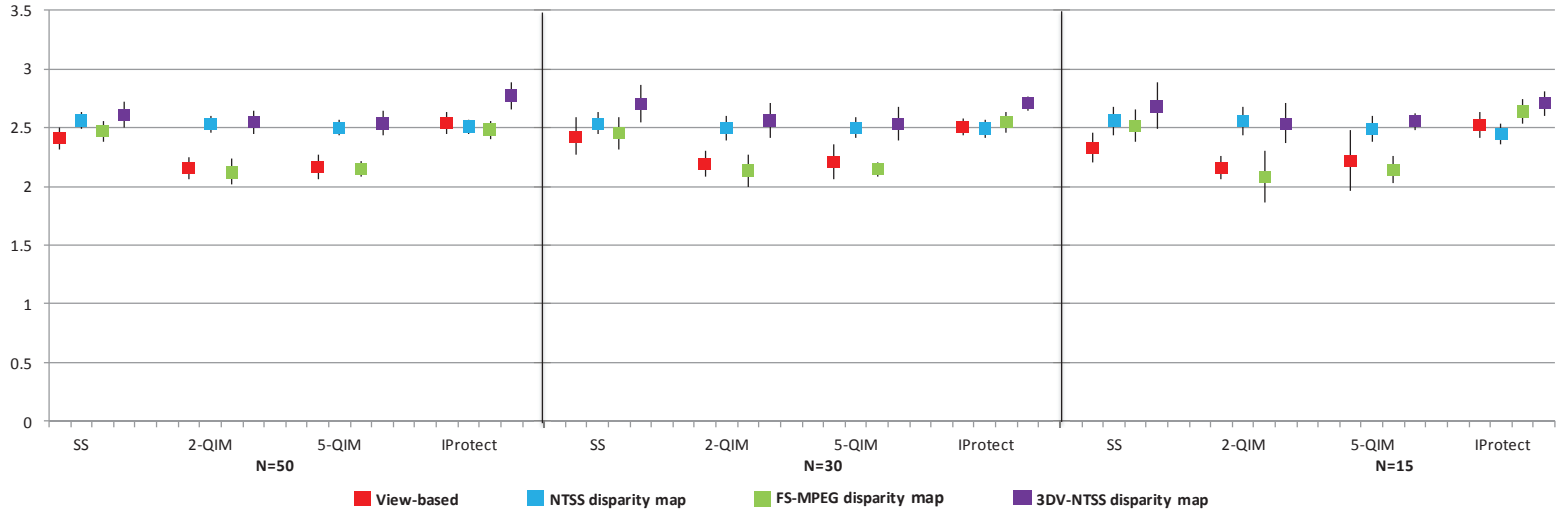
Low quality stereoscopic video content (Depth Perception)



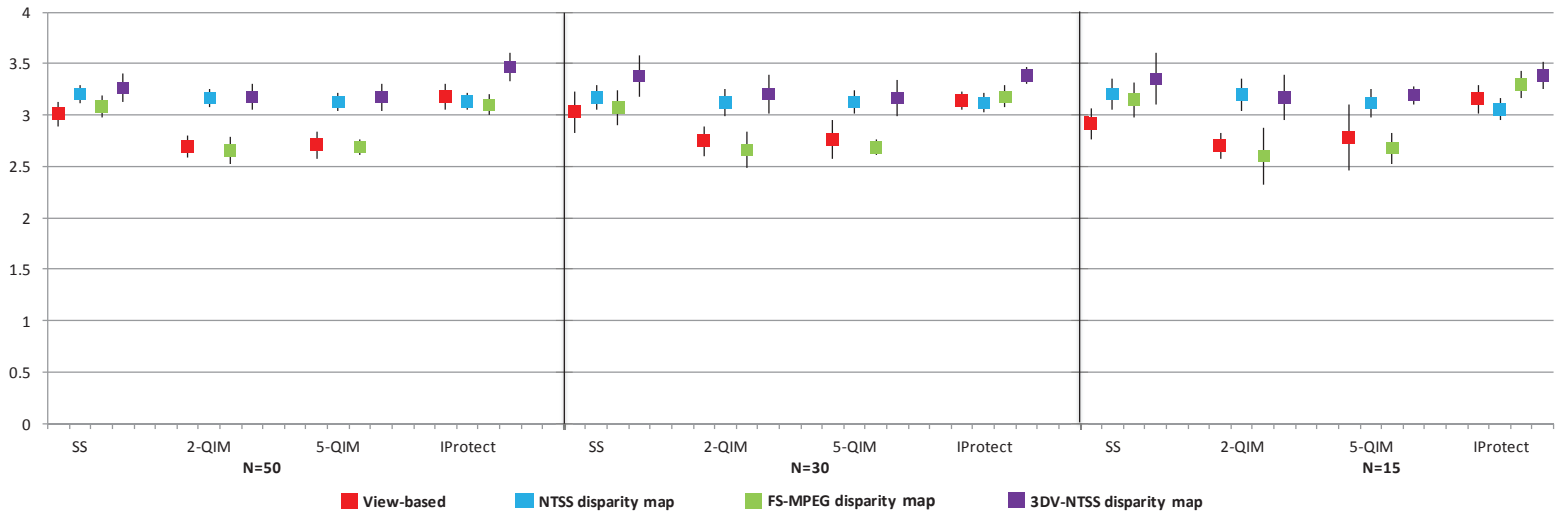
(a)



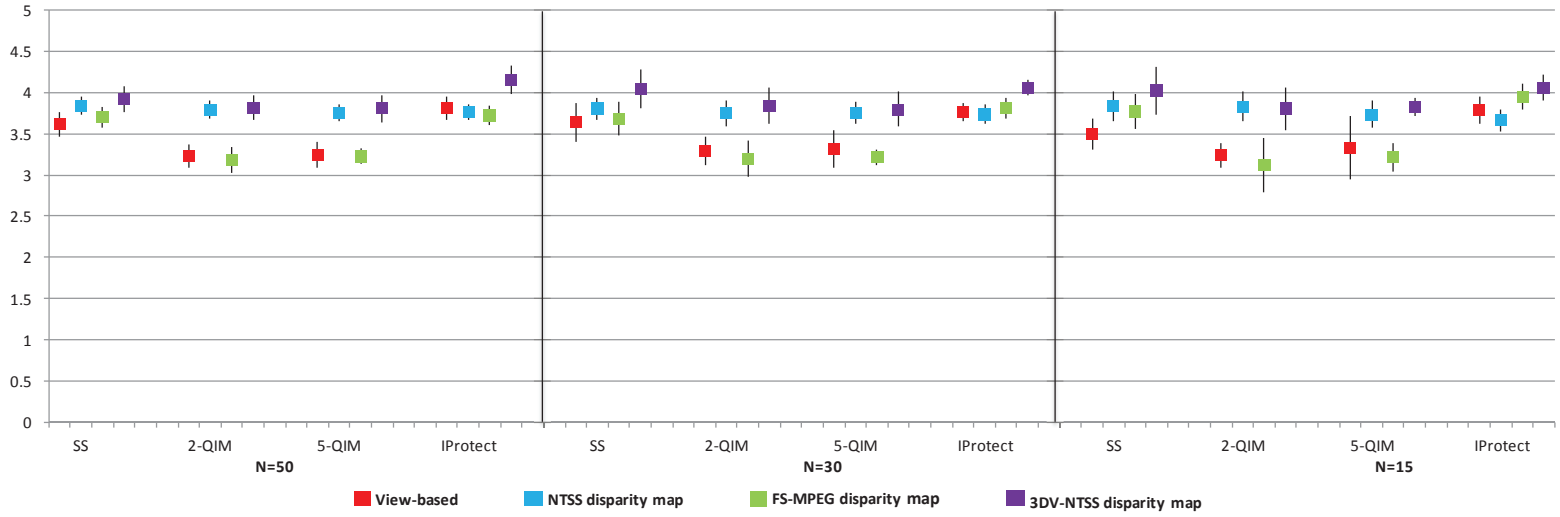
(b)


3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

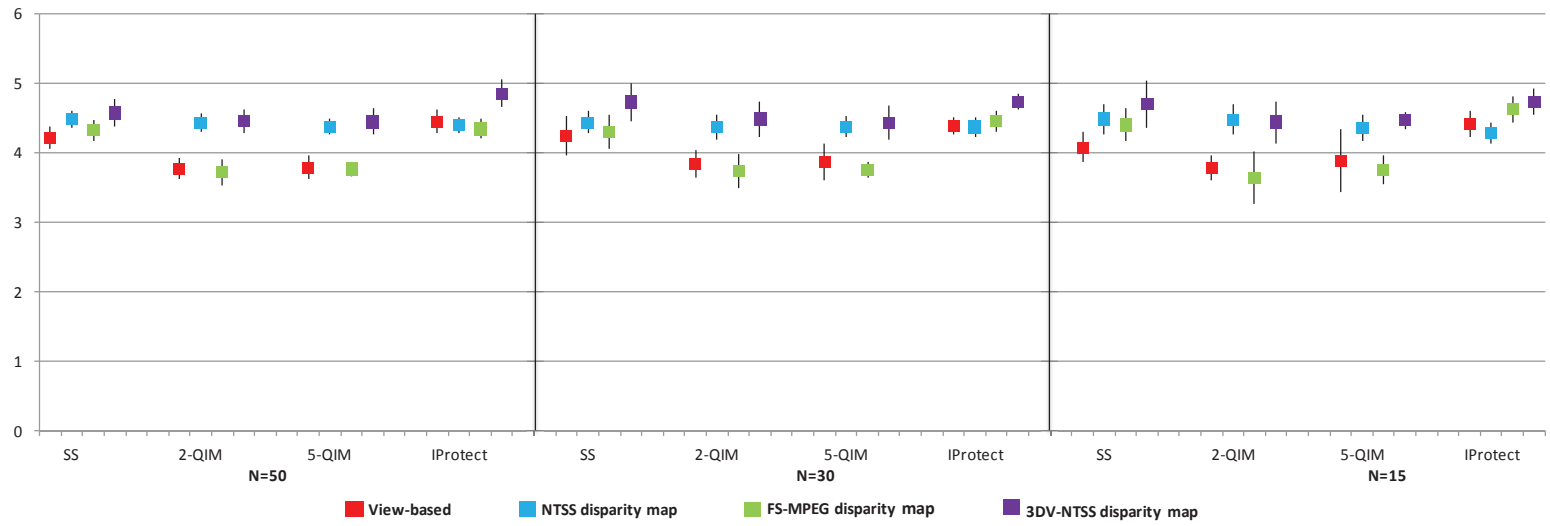
(c)


4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

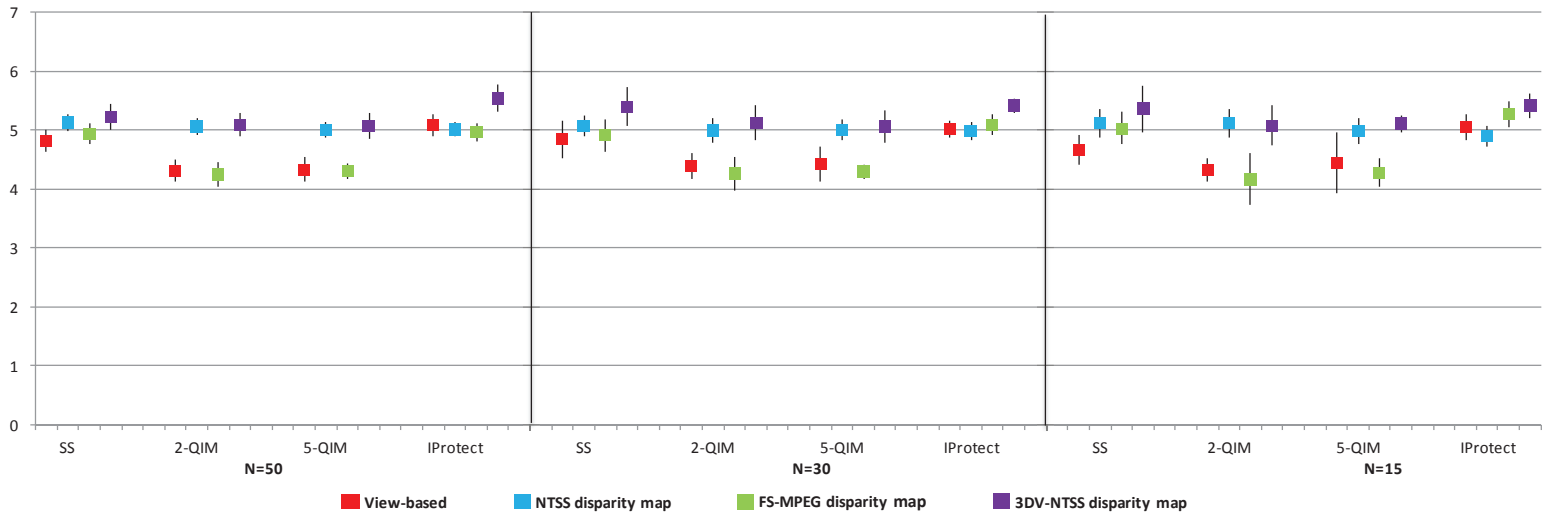
(d)


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

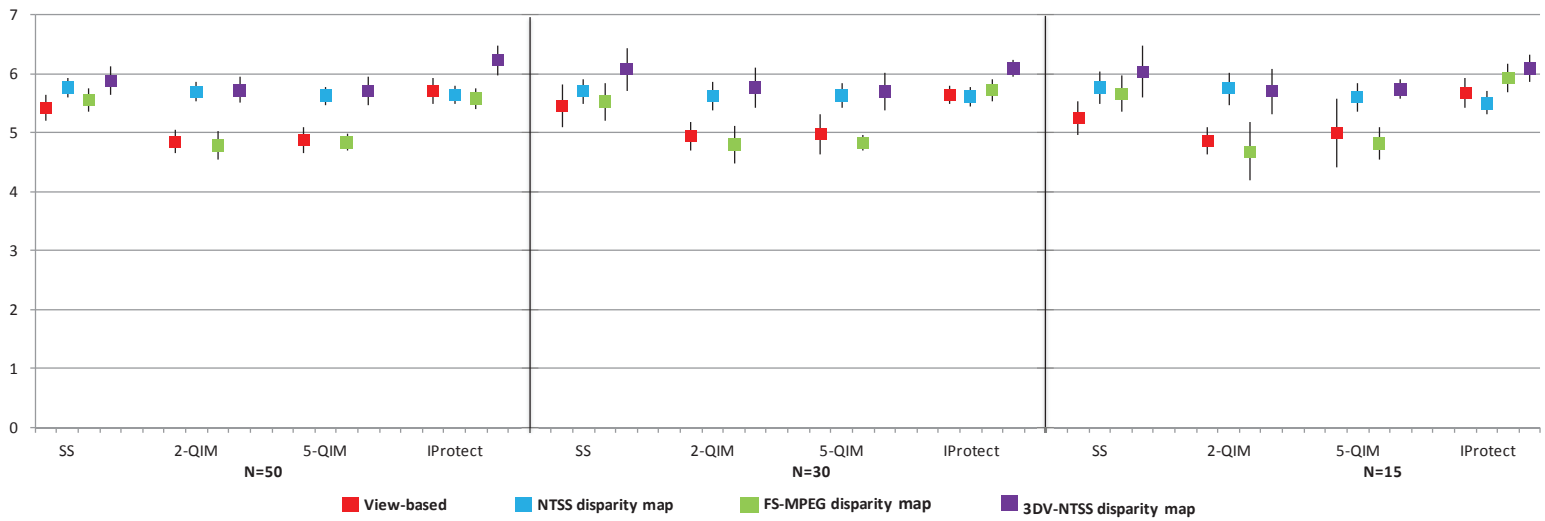
(e)


6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)



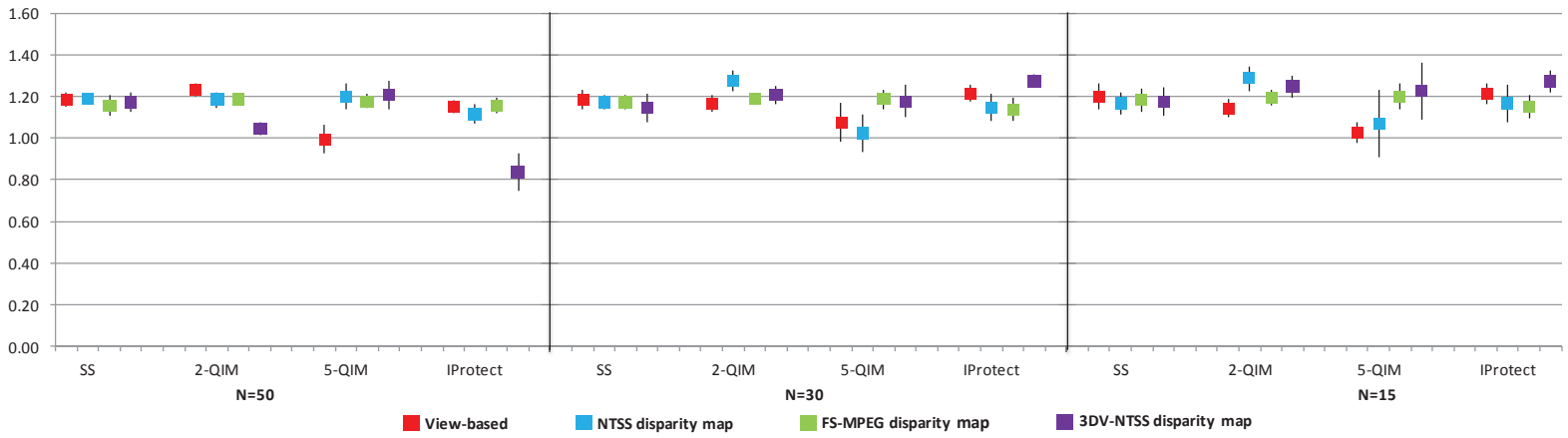
(g)



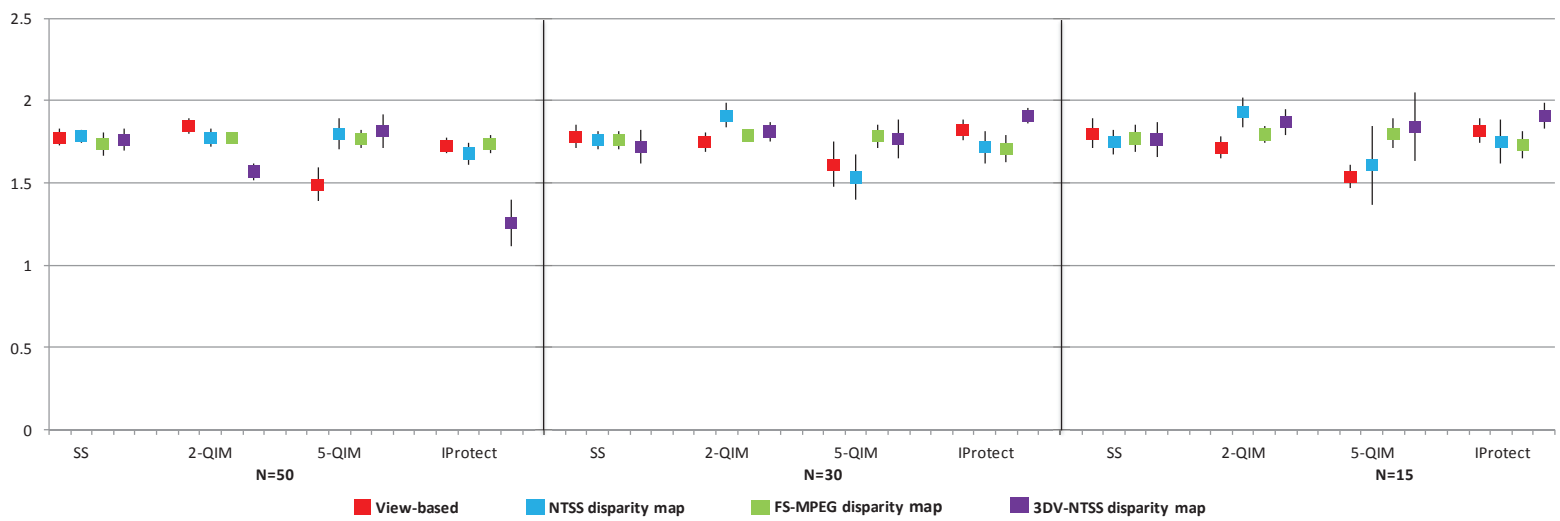
(h)

Figure A2-5: Subjective evaluations low-quality stereoscopic video content (Depth Perception), for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*

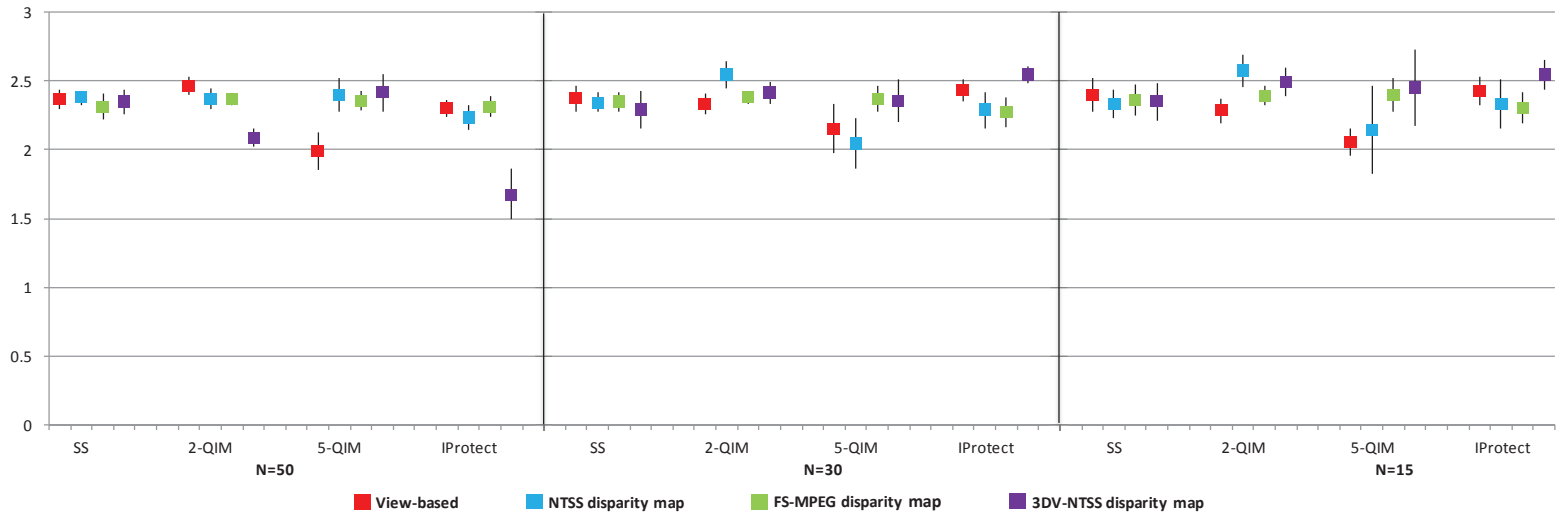
Low quality stereoscopic video content (Visual comfort)



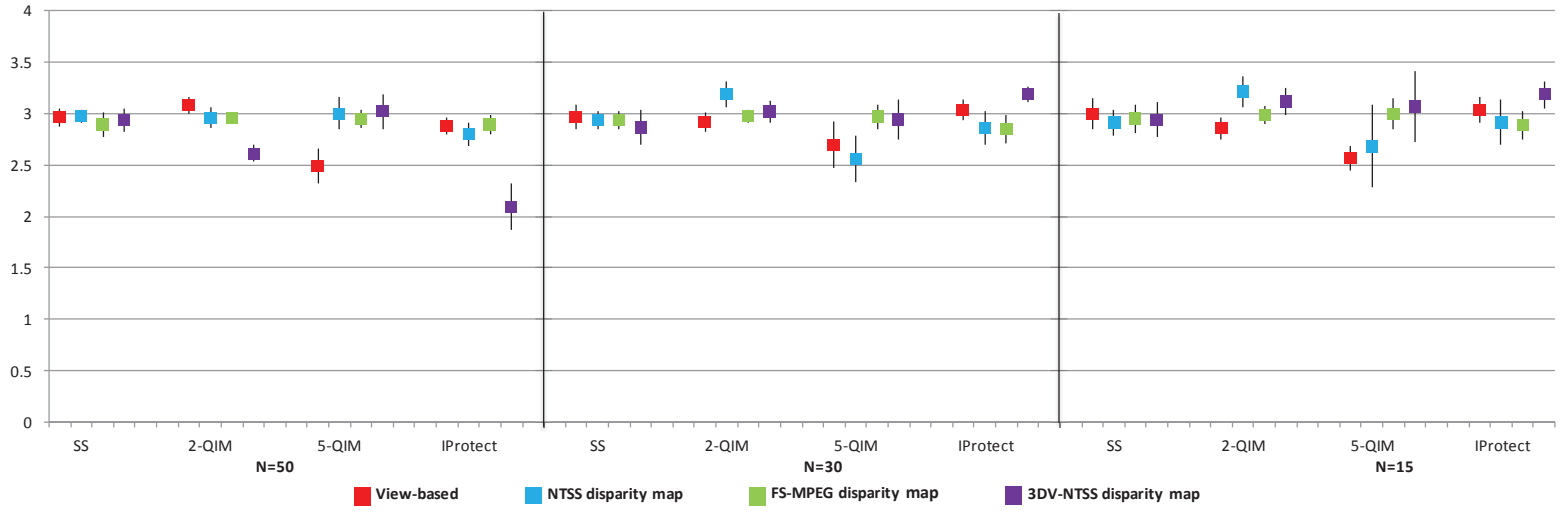
(a)



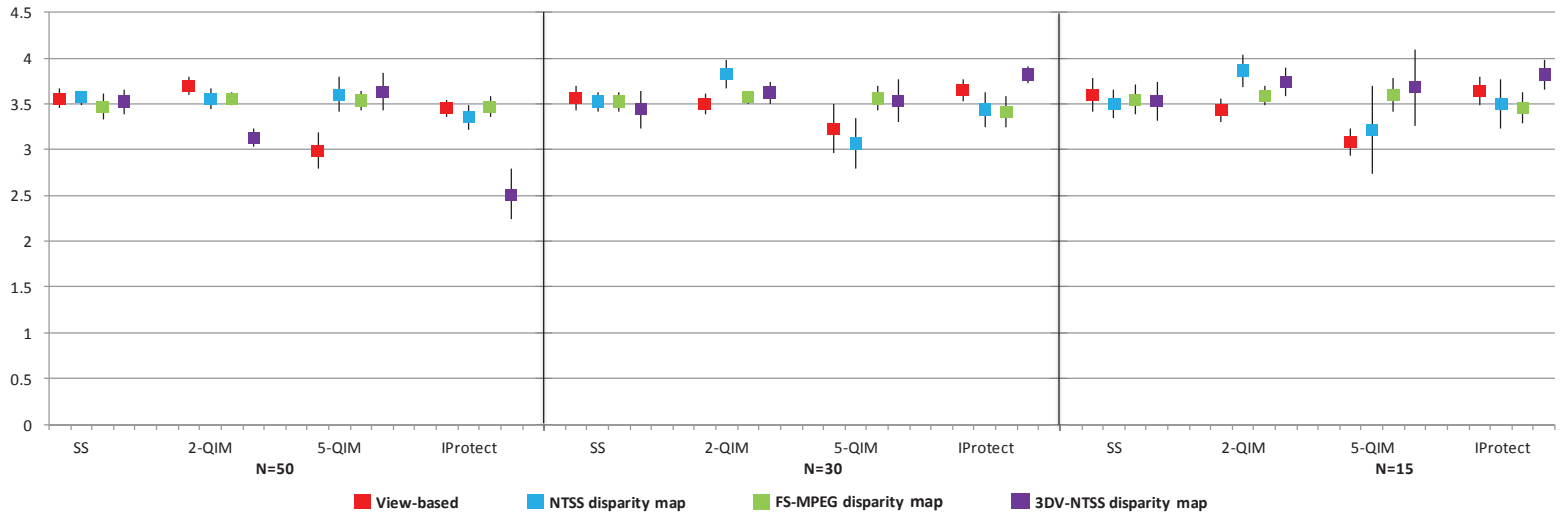
(b)


3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

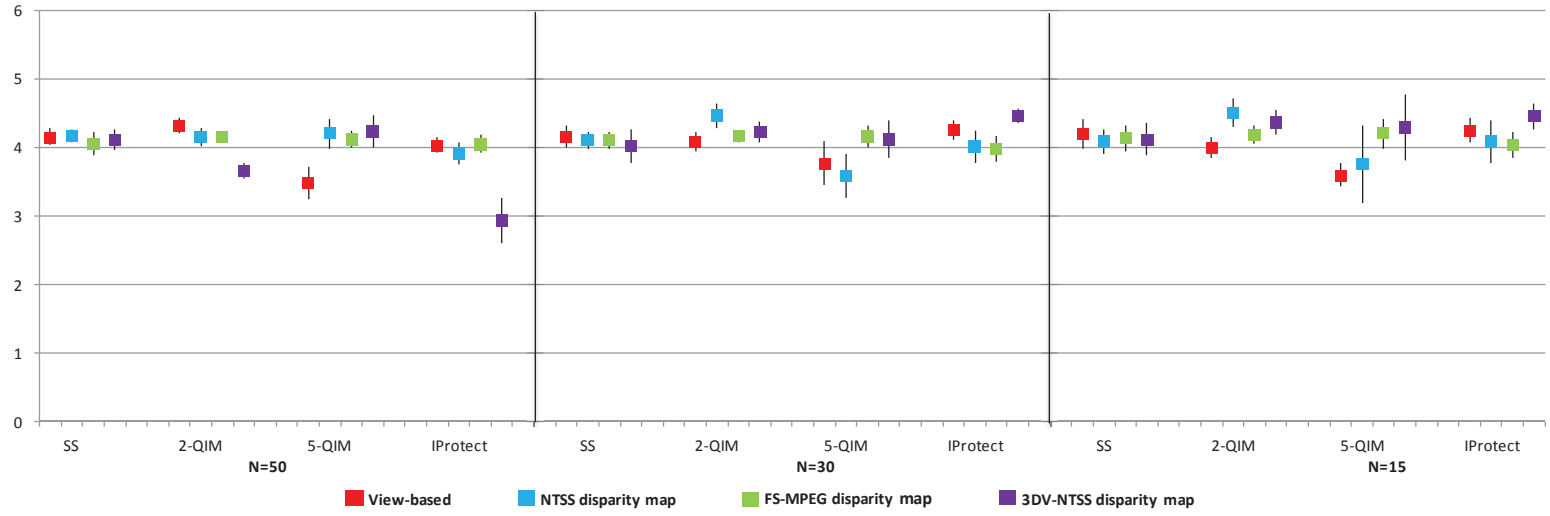
(c)


4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

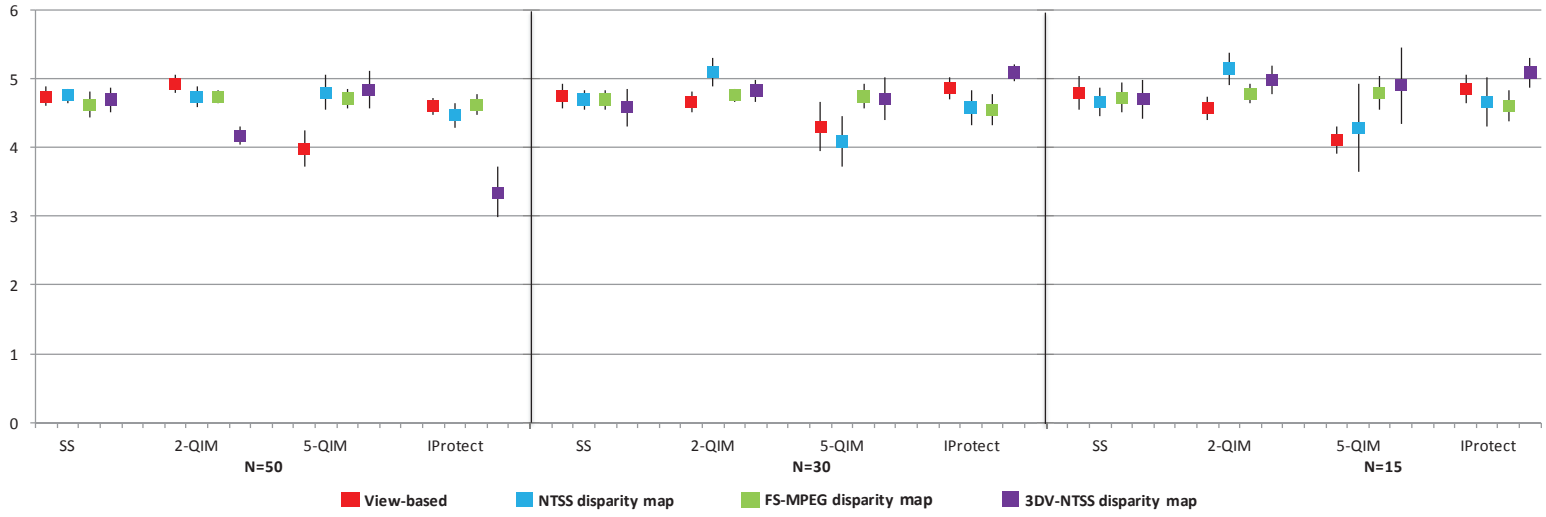
(d)


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

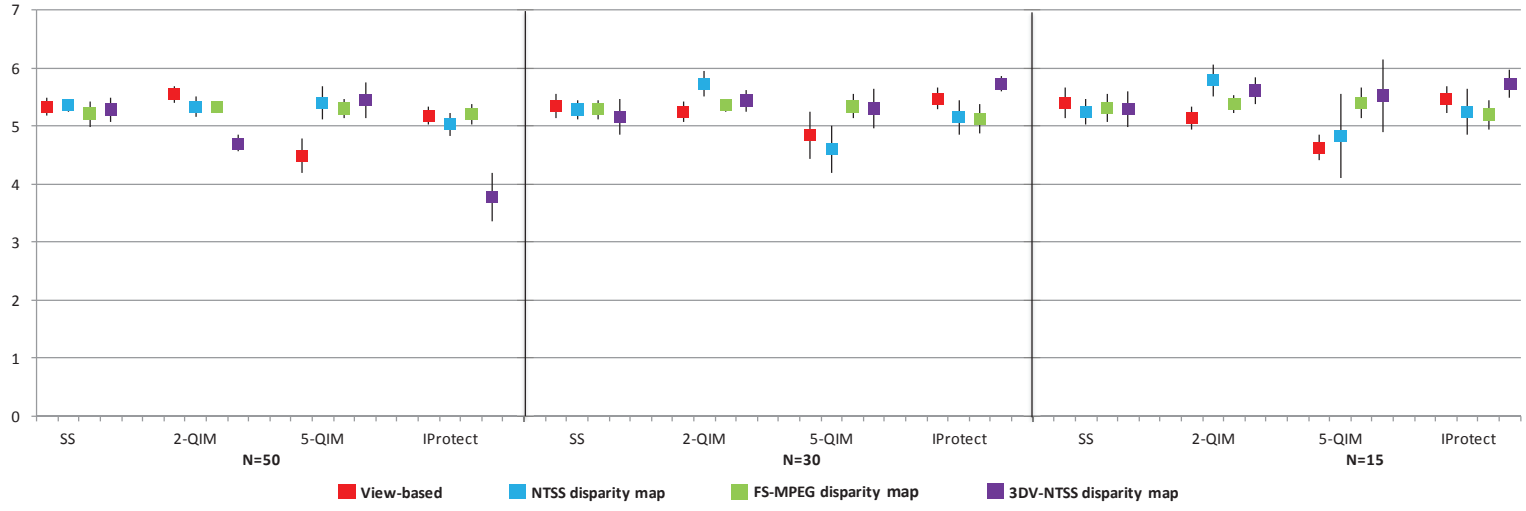
(e)


6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)



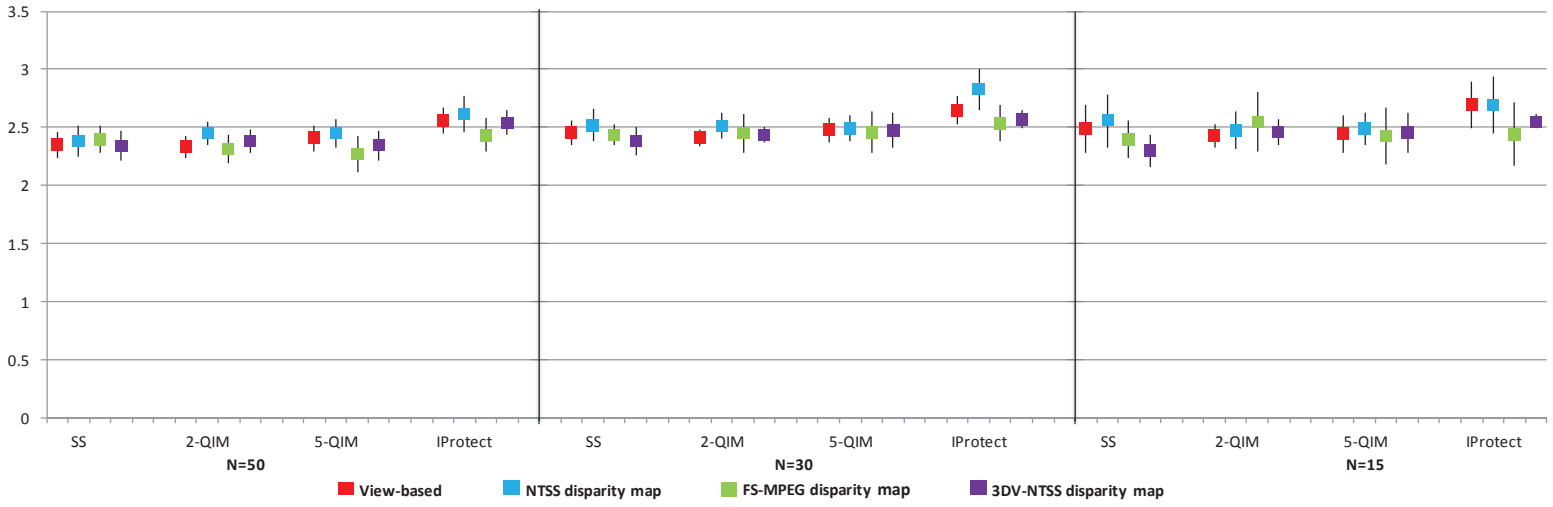
(g)



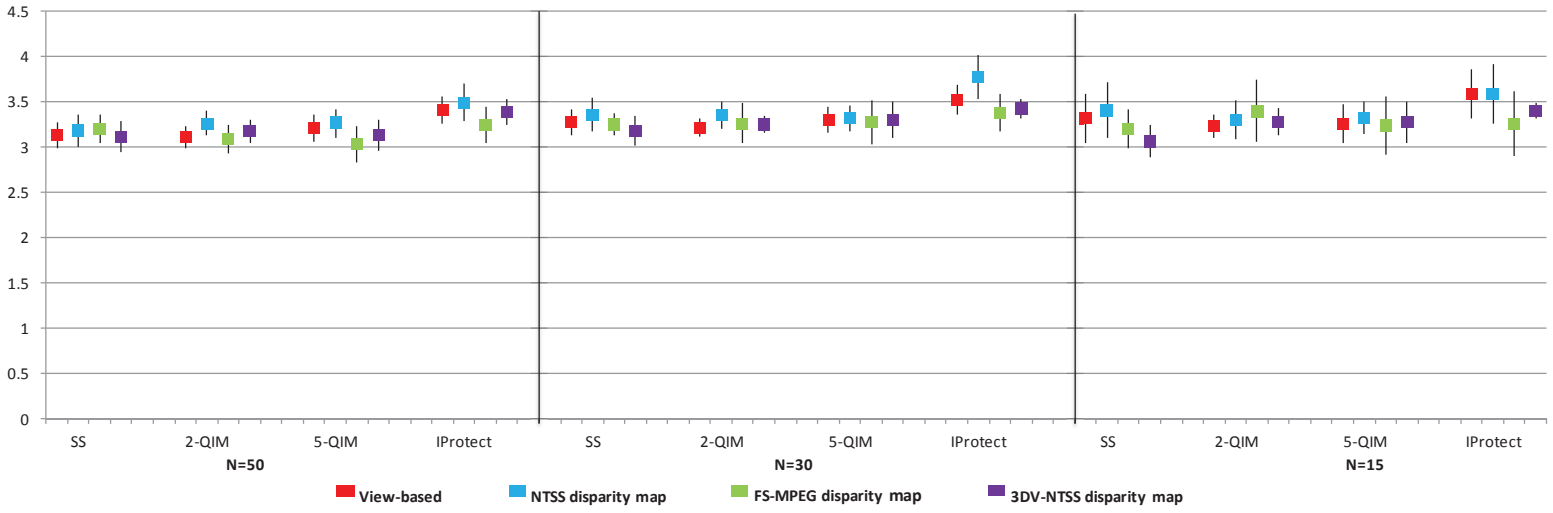
(h)

Figure A2-6: Subjective evaluations low-quality stereoscopic video content (Visual Comfort), for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*

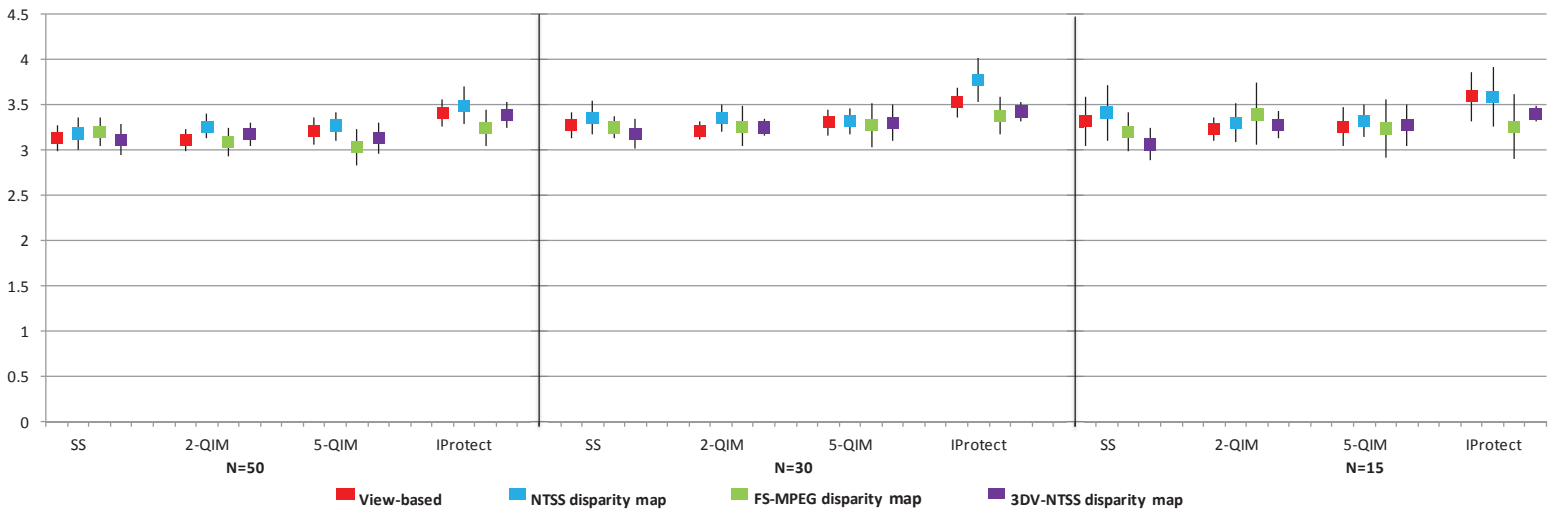
High quality 2D video content



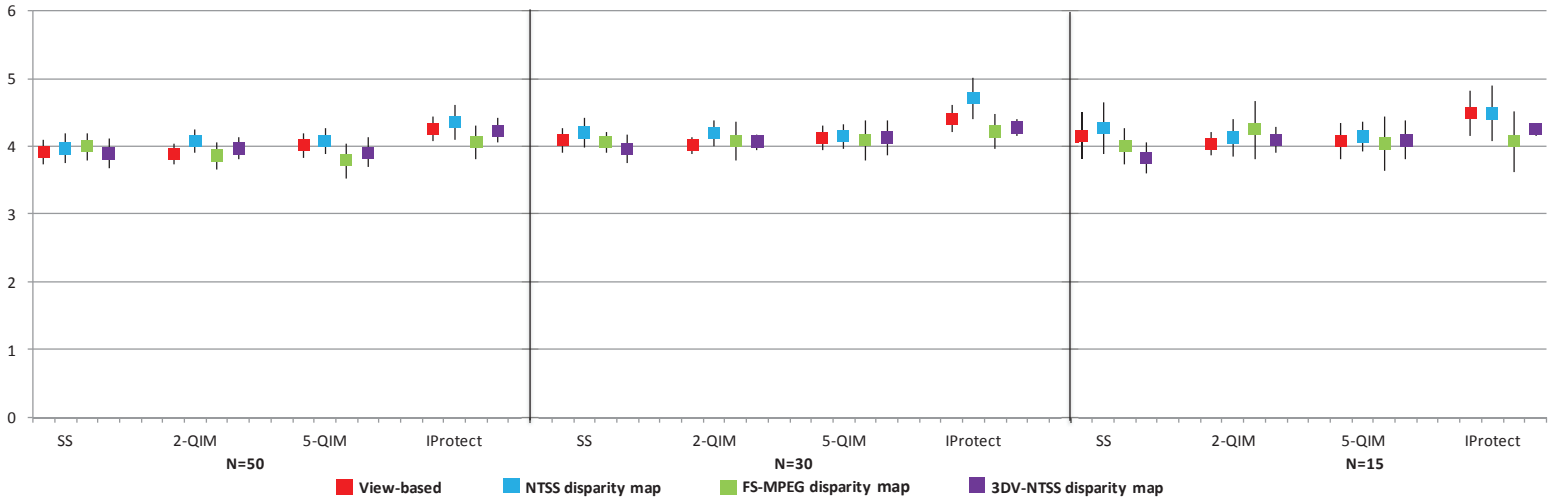
(a)



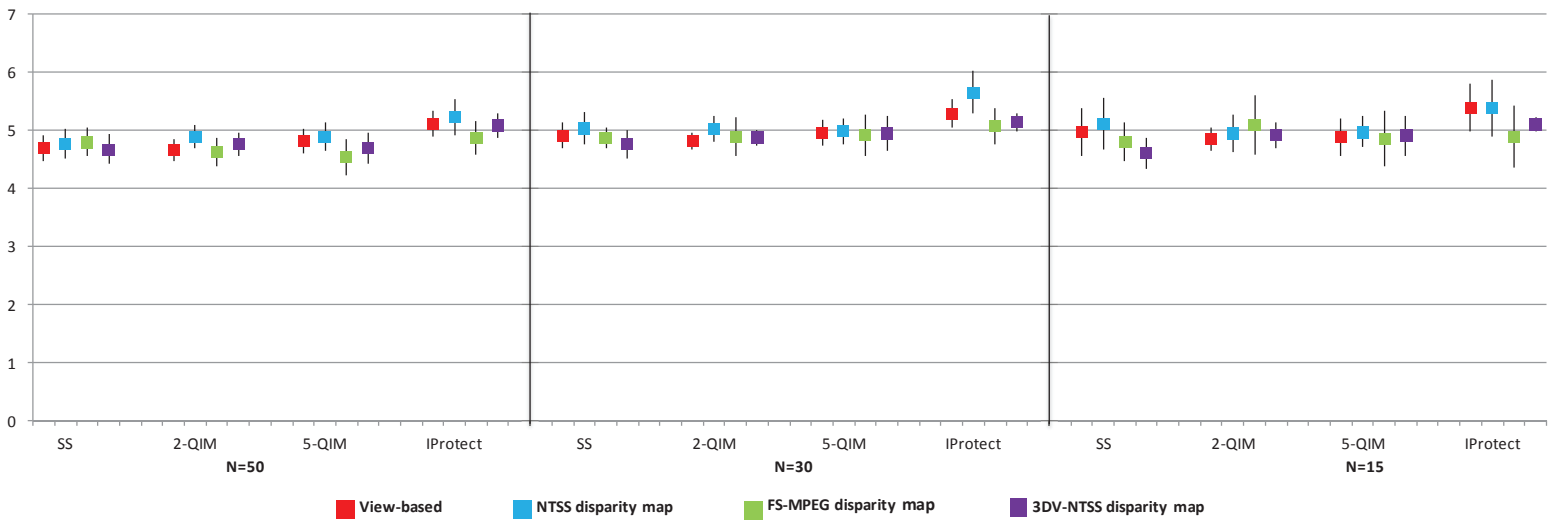
(b)


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

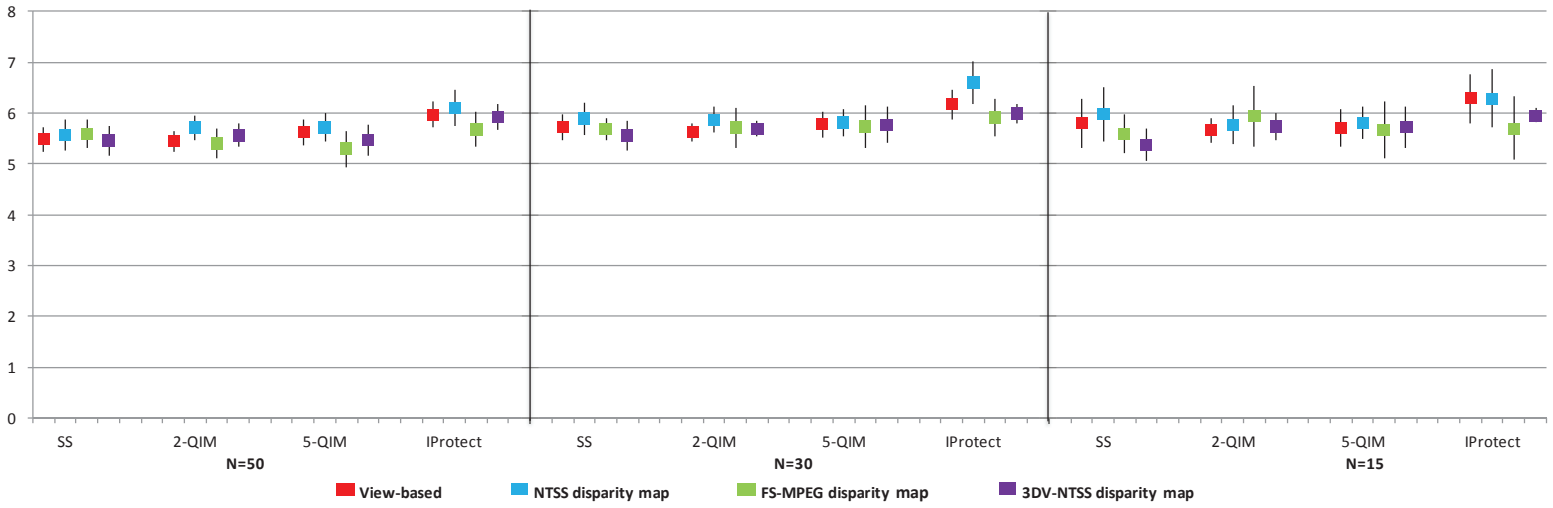
(c)


6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

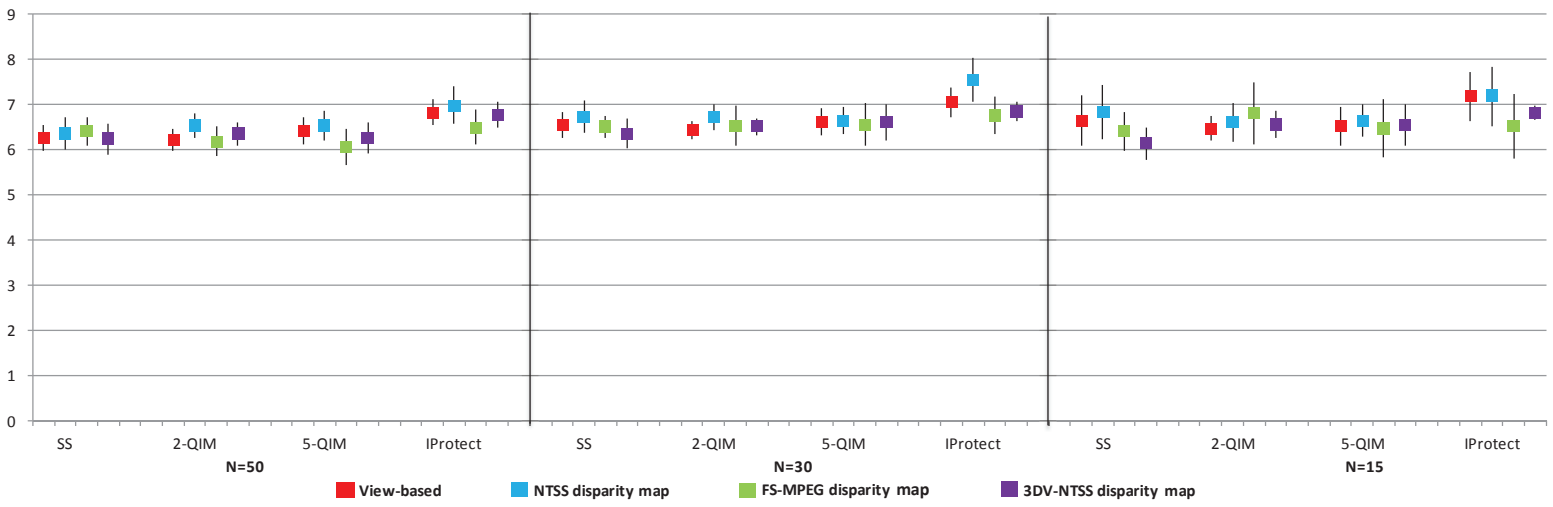
(d)


7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
SS
2-QIM
5-QIM
IProtect
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

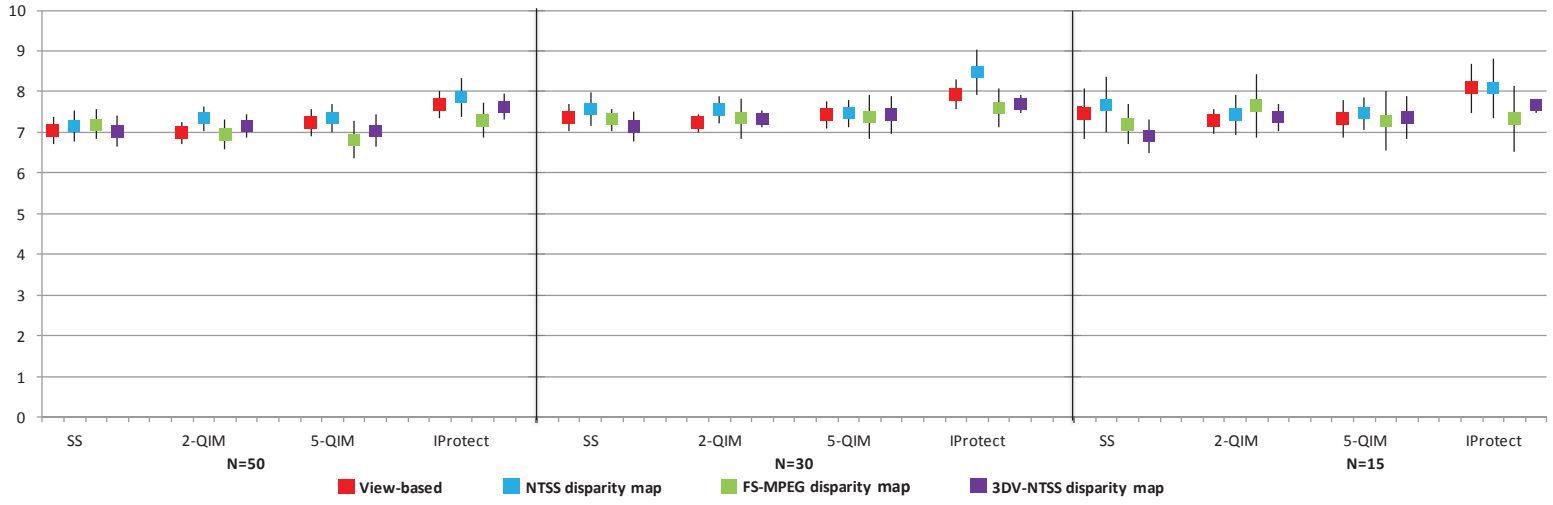
(e)


8
7
6
5
4
3
2
1
0
SS
2-QIM
5-QIM
IProtect
SS
2-QIM
5-QIM
IProtect
SS
2-QIM
5-QIM
IProtect
N=50
N=30
N=15
View-based
NTSS disparity map
FS-MPEG disparity map
3DV-NTSS disparity map

(f)



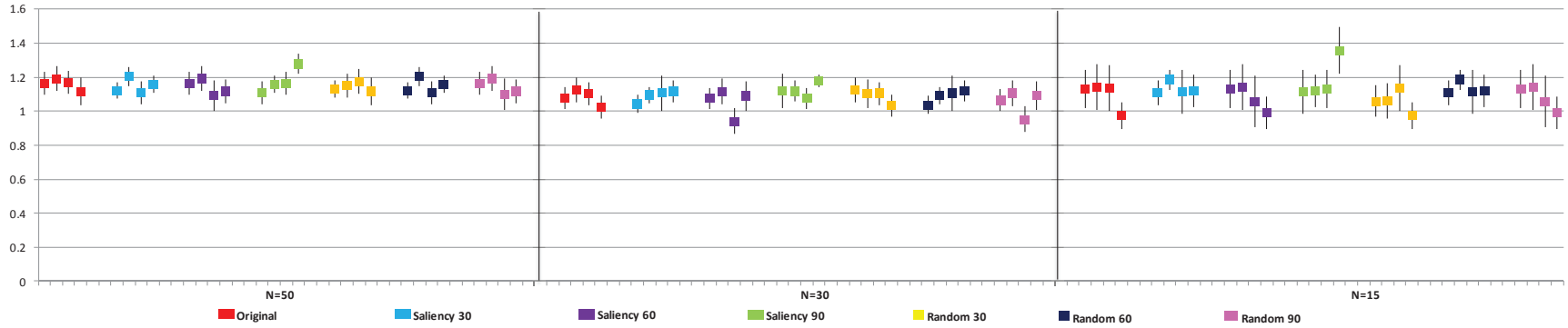
(g)



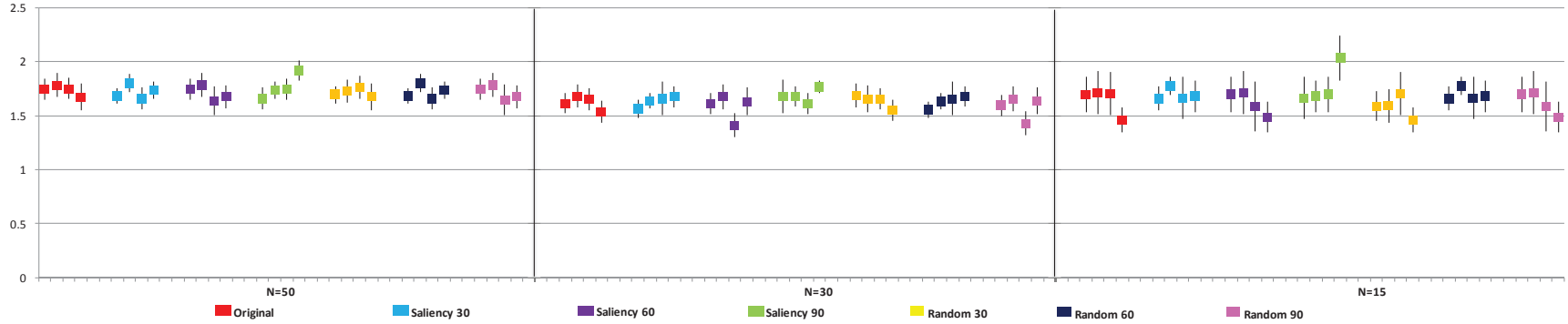
(h)

Figure A2-7: Subjective evaluations high-quality 2D video content ,for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*

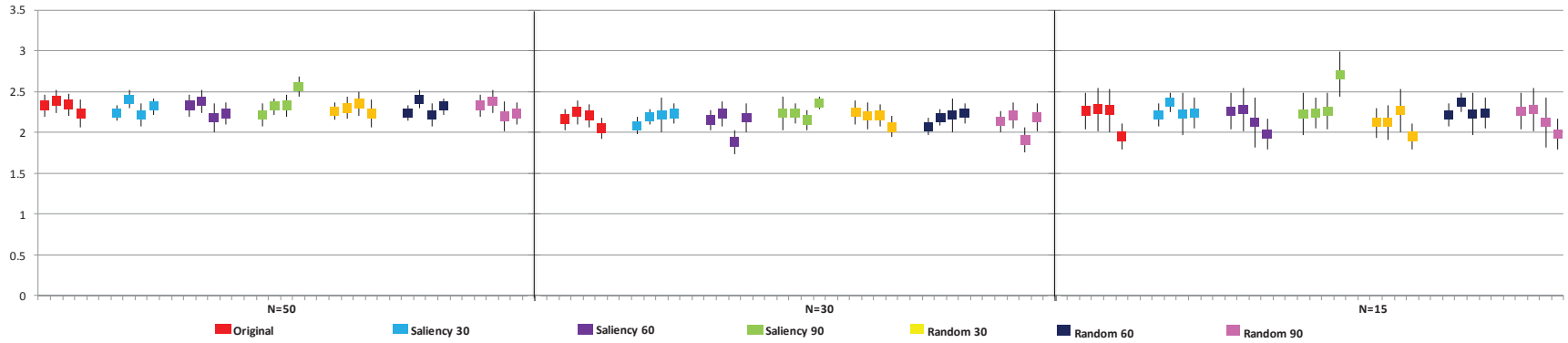
Low quality 2D video content



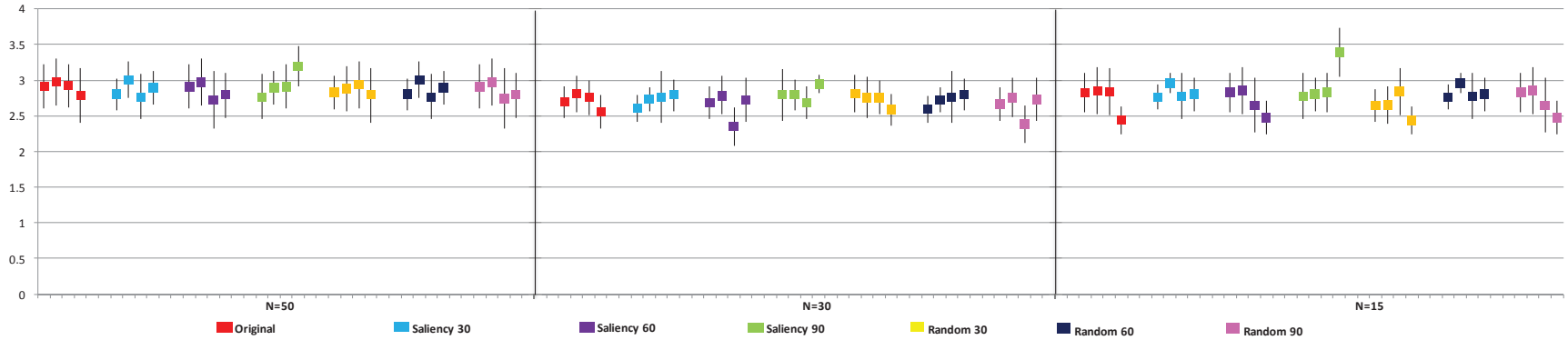
(a)



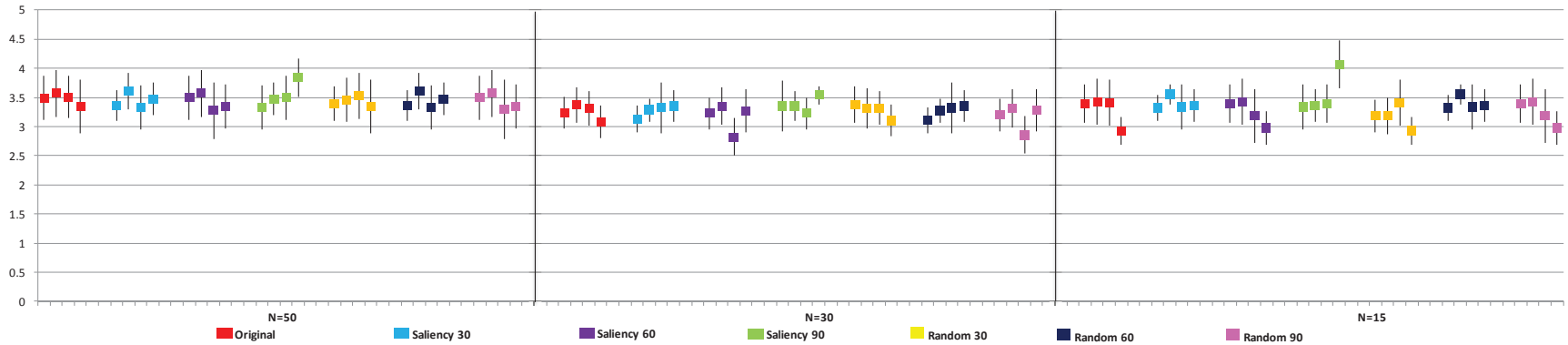
(b)


3.5
3
2.5
2
1.5
1
0.5
0
N=50
N=30
N=15
Original
Saliency 30
Saliency 60
Saliency 90
Random 30
Random 60
Random 90

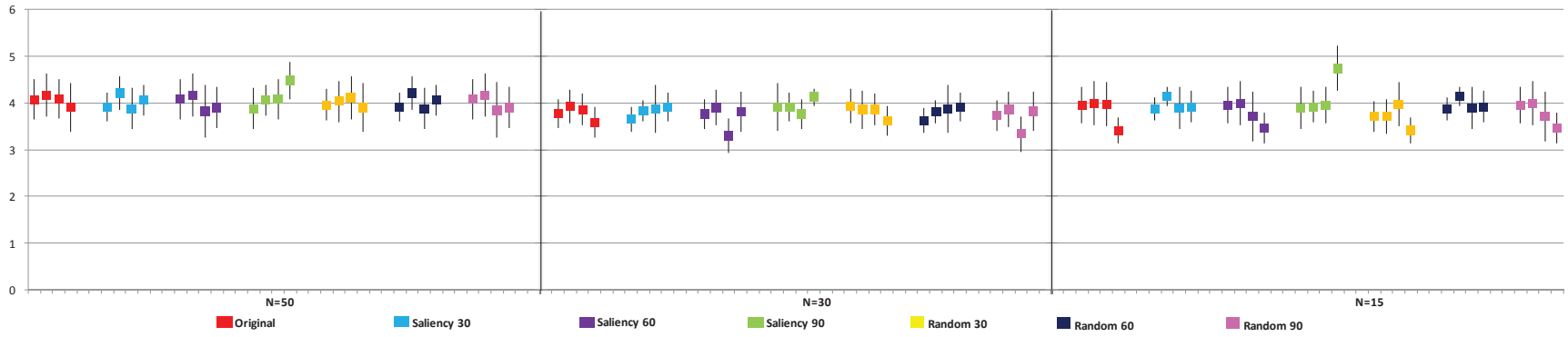
(c)


4
3.5
3
2.5
2
1.5
1
0.5
0
N=50
N=30
N=15
Original
Saliency 30
Saliency 60
Saliency 90
Random 30
Random 60
Random 90

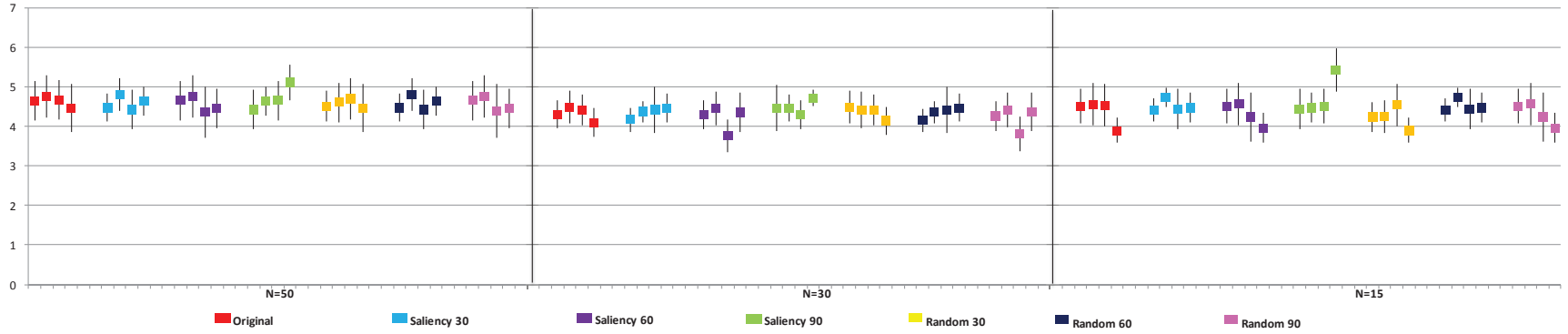
(d)


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
N=50
N=30
N=15
Original
Saliency 30
Saliency 60
Saliency 90
Random 30
Random 60
Random 90

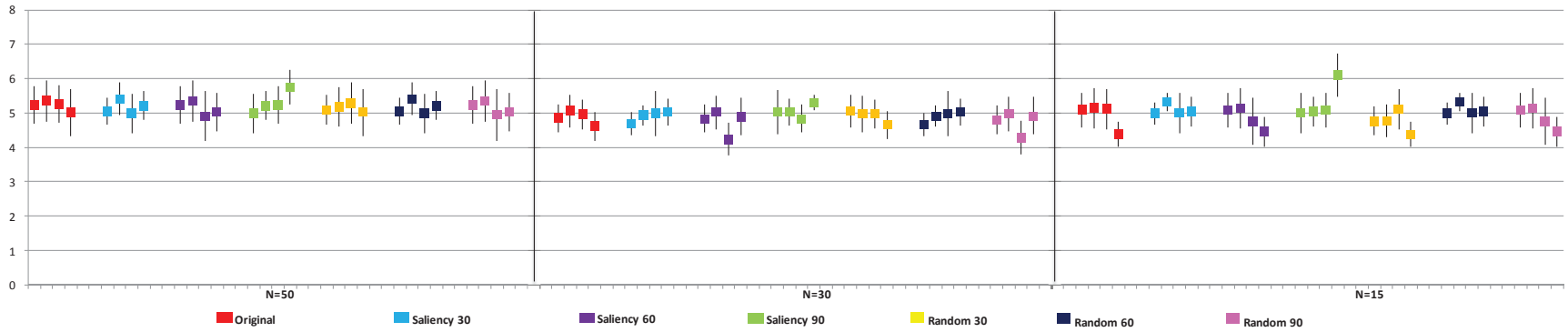
(e)


6
5
4
3
2
1
0
N=50
N=30
N=15
Original
Saliency 30
Saliency 60
Saliency 90
Random 30
Random 60
Random 90

(f)


7
6
5
4
3
2
1
0
N=50
N=30
N=15
Original
Saliency 30
Saliency 60
Saliency 90
Random 30
Random 60
Random 90

(g)



(h)

Figure A2-8: Subjective evaluations low-quality 2D video content ,for grading scales of: (a) $q = 2$, (b) $q = 3$, (c) $q = 4$, (d) $q = 5$, (e) $q = 6$, (f) $q = 7$, (g) $q = 8$, (h) $q = 9$ quality levels and for a number of observers *N=50, N=30 and N=15.*

References

[AIT69] Aitken R.C., “Measurement of feelings using visual analogue scales”. Proc R Soc Med, 62989- 993, (1969).

[ALB81] Albaum G., Best R., and Hawkins D., “Continuous vs. discrete semantic differential rating scales”. Psych.Reports, vol.49, pp.83–86, (1981).

[AMM17-01] Ammar M., Mitrea M., Hasnaoui M., Le Callet P., “MPEG-4 AVC stream-based saliency detection.Application to robust watermarking” Signal Processing: Image Communication Volume 60, Pages 116-130, (2018).

[AMM17-02] Ammar M.,” Visual saliency extraction from compressed streams”. Phd Thesis, (2017).

[BEG03] Beghdadi A. and Pesquet-Popescu B., “A new image distortion measure based on wavelet decomposition”. 7th International Symposium and Its Applications, Paris, France ,(2003).

[BEN08] Benoit A., Le Callet P., Campisi P., and Cousseau R., “Quality assessment of stereoscopic images”. EURASIP Journal on Image and Video Processing, pp. 1-13, (2008).

[BEN14] Bensaied R., Mitrea M., Chammem A., and Ebrahimi T. “Subjective quality assessment for stereoscopic video: case study on robust watermarking”. IS&T/SPIE Electronic Imaging, (2014).

[BON74] Bond A. and Lader M., “ The use of analogue scales in rating subjective feelings”. British Journal of Medical Psychology 47, 211–217, (1974).

[BOS11] Bosc E., Pépion R., Le Callet P., Koppel M., Ndjiki-Nya P., and al., “Towards a New Quality Metric for 3-D Synthesized View Assessment”. IEEE Journal on Selected, Topics in Signal Processing, (2011).

[BOS13] Bosc E., Hanhart P., Le Callet P., and Ebrahimi T., “A quality assessment protocol for

Free-viewpoint video sequences synthesized from decompressed depth data". Fifth International Workshop on Quality of Multimedia Experience (QoMEX), 100-105, (2013).

[CAM07] Campisi P., Le Callet P., and Marini E., "Stereoscopic images quality assessment". Proc. European Signal Processing Conference (EUSIPCO), Poznan, Poland, (2007).

[CAM07] Campisi P., Le Callet P., and Marini E., "Stereoscopic images quality assessment". Proc. 15th European Signal Processing Conference (EUSIPCO), Poznan, Poland, (2007).

[CHA13] Chammem A., "Robust watermarking techniques for stereoscopic video protection". Phd Thesis, UPMC, (2013).

[CHE12] Chen M.J., Kwon D.C., and Bovik A.C., "Study of subject agreement on stereoscopic video quality". IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), p 173-176, (2012).

[CHE12-01] Chen M.J., Kwon D.K., Bovik A.C., "Study of subject agreement on stereoscopic video quality". Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation, (2012).

[CIS15] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white paper c11-520862.pdf

[CIS17] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white paper c11-520862.pdf

[COX02] Cox I., Miller M., Bloom J., and Miller M., "Digital Watermarking". (2002).

[COX80] Cox E.P., "The optimal number of response alternatives for a scale: A review". J. Marketing Res., vol. 17, no. 4, pp. 407–422, (1980).

[CRO82] Crosby P., "The puzzlement of quality". Technical report, Philip Crosby Associates II, (1982).

[DEM77] Dempster A.P., Laird N.M., and Rubin D.B., "Maximum Likelihood from Incomplete Data via the EM algorithm", J. Roy. Stat. Soc. (B), 38-39, (1977).

[DIN06] Ding L., Chien S., and Chen L., "Joint prediction algorithm and architecture for stereo

video hybrid coding systems". IEEE Trans. on Circuits Syst. Video Technol, pp. 1324-1337, (2006).

[EMI11] Bosc E., Pepion R., Le Callet P., Koppel M., and Ndjiki-Nya P.,"Towards a New Quality Metric for 3-D Synthesized View Assessment". IEEE Journal on Selected Topics in Signal Processing, pp. J-STSP-ETVC-00048-(2011).

[ENG99] Engeldrum P.G., "Image quality modeling: Where are we?". In Image Processing, Image Quality, Image Capture, Systems Conference (PICS), pages 251–255, Savannah, GA, (1999).

[FAN15] Fang Y., Ma K., Wang Z, Lin W., Fang Z., and Zhai G., "No-reference quality assessment of contrast-distorted images based on natural scene statistics," IEEE Signal Process. Lett., vol. 22, no. 7, pp. 838–842, (2015).

[FRE23] Freyd M., "The graphic rating scale". Journal of Educational Psychology 14, 83-102, (1923).

[FRO89] Froberg D.G, and L. KANE R., "Methodology for measuring health-state preferences-ii: scaling methods", Division of Human Development and Nutrition, School of Public Health, University of Minnesota, Minneapolis, MN 55455, U.S.A., (1988)

[GAO13] Gao F., Tao D., Gao X., and al., "Learning to rank for blind image quality assessment" .Xiv preprint arXiv:1309.0213 , (2013).

[GOL10-01] Goldmann L., De Simone F., Ebrahimi T., "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video". In IS&T/SPIE Electronic Imaging, (2010).

[GOL10-02] Goldmann L., De Simone F., Ebrahimi T., "Impact of acquisition distortions on the quality of stereoscopic images". In Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics-VPQM, (2010).

[HAN13] Hanhart H., and Ebrahimi T., "On the evaluation of 3D codecs on multiview autostereoscopic display". International Workshop on Hot Topics in 3D, (2013).

[HAN15] Hanhart P., Bernardo M.V., Pereira M., Pinheiro A.M.G, and Ebrahimi T., "Benchmarking of objective quality metrics for HDR image quality assessment," EURASIP

Journal on Image and Video Processing, vol. 39, pp. 1–18, (2015).

[HAS14] Hasnaoui M., Mitrea M., “Multi-symbol QIM video watermarking Signal”. Process. Image Communication, vol. 29, no. 1, pp.
107–127. (2014).

[HEW08] Hewage C.T.E.R, Worrall S.T, Dogan S., and Kondoz A.M, “Prediction of stereoscopic video quality using objective quality models of 2-D video”. Electronics Letters, pp. 963-965, (2008).

[HUY11] Huynh-Thu, Q., Garcia, M., Speranza, F., Corriveau, P., Raake, A.: Study of rating scales for subjective quality assessment of High-Definition video. Broadcasting, IEEE Transactions on 57(1), 1–14 (2011).

[HUY07] Q. Huynh-Thu, M. Brotherton, D. Hands, K. Brunnström, and M. Ghanbari, “Examination of the SAMVIQ methodology for the subjective assessment of multimedia quality,” in Proc. 3rd Int. Workshop Video Process. Consum. Electron., Scottsdale, AZ, USA, Jan. (2007).

[ISO11] ISO/IEC JTC1/SC29/WG11, “Report of Subjective Test Results from the Call for Proposals on 3D Video Coding Technology,” Doc. N12347, Geneva, CH, (2011).

[JAH97] Jahne B.,” Digital Image Processing: Concepts, Algorithms, and Scientific Applications”. Springer-Verlag New York, Inc. Secaucus, NJ, USA, (1997).

[JAN99] Janssen T. J.W. M., “Computational Image Quality”. PhD thesis, Technische Universiteit Eindhoven, (1999).

[JON86] Jones B.L., and McManus P.R., “Graphic scaling of qualitative terms”. SMPTE Journal, 1166–1171, (1986).

[KUN16] Kundu D., "Subjective and objective quality evaluation of synthetic and high dynamic range images". (2016).

[LI94] Li R., Zeng B., and Liou M.L., “A new three-step search algorithm for block motion estimation”. IEEE Trans. on Circuits Syst. Video Technol, 438-442, (1994).

[MAR07] Marini E., Autrusseau F., Le Callet P., Campisi P., “Evaluation of standard watermarking techniques”. International Society for Optics and Photonics in Electronic Imaging, 65050O-65050O, (2007).

[MAR11] Marini F.,”Content Based No-Reference Image Quality Metrics”.(2011).

[MAT71] Matell M.S., and Jacoby J., “Is there an optimal number of alternatives for Likert scale items? Study 1: reliability and validity”, Educational and Psychological Measurement, vol. 31, pp. 657–674, (1971).

[MCG84] McGuire D.B., “The measurement of clinical pain”. Nurs Res, pp. 152-156, (1984).

[MCK78] McKelvie S.J., “Graphic rating scales—How many categories?”. British J. Psych., vol. 69, no. 2, pp. 185–202, (1978).

[MIT13] Mitrea M., Chammem A., and Prêteux F., “ Tatouage stéréoscopique “, [Vidéo et TVHD 3D … de la capture à la diffusion Principe, tendances et perspectives], Laurent Lucas, Céline Locos et Yannick Remion Editors, Lavoisier, Paris, Chapter 13, 251-217, (2013).

[MSU13] MSU Video Group “MSU Video Quality Measurement Tool” http://www.compression.ru/ , (2013).

[MUE02] Mued L., Lines B., Furnell S. and Reynolds P., “Investigating Interaction of Audio and Video Quality as Perceived in Low-Cost Multimedia”. pp 181-189, (2002).

[NAR93] Narita N., “Graphic scaling and validity of Japanese descriptive terms used in subjective evaluation tests”. SMPTE J., vol. 102, no. 7, pp. 616–622, (1993).

[NIN06] Ninassi A., Le Meur O., Le Callet P., Barba D., and Tirel A., “Task impact on the visual attention in subjective image quality assessment”. European Signal Processing Conference, France, Invited paper, (2006).

[NIN09] Ninassi A., Le Meur O., Le Callet P., and Barba D., “Considering temporal variations of spatial visual distortions in video quality assessment”. IEEE Journal Of Selected Topics In Signal Processing : Special Issue On Visual Media Quality Assessment, pp.253-265, (2009).

[PEC08] Péchard S., Pépion R., Le Callet P., “Suitable methodology in subjective video quality assessment: a resolution dependent paradigm”. Proceedings of the Third International Workshop on Image Media Quality and its Applications, IMQA2008, (2008).

[PIE14] Piedade D., Bernardo M., Fiadeiro P., Pinheiro A., Pereira M., ”Chromatic variations on 3d video and qoe”.European Signal Processing Conference, Lisbon, Portugal, pp.221-225, (2014).

[PRE00] Preston C.C. and Colman A.M, “Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences”. Acta Psychologica, vol. 104, no. 1, pp. 1–15, (2000).

[PRE15] Preiss J., “Color-Image Quality Assessment: From Metric to Application”. (2015).

[PEC08] Stephane Pechard, Romuald Pepion, and Patrick Le Callet, “Suitable methodology in subjective video quality assessment: a resolution dependent paradigm,” in Proc. of International Workshop on Image Media Quality and its Applications (IMQA2008), Kyoto, Japan, (2008).

[SES10] Seshadrinathan K., Soundararajan R., Bovik A.C., Cormack L.K., "Study of subjective and objective quality assessment of video". IEEE Trans. Image Process., vol. 19, no. 6, pp. 1427-1441,(2010).

[SIM09] Simone D.F., Goldmann L., Baroncini V., and Ebrahimi T., “Subjective evaluation of JPEG XR image compression”. In Applications of Digital Image Processing, vol. 7443 of Proceedings of SPIE, (2009).

[SIM11] Simone D.F., Goldmann L., Lee J.C., and Ebrahimi T., "Towards high efficiency video coding: Subjective evaluation of potential coding technologies", J. Vis. Commun. Image R., vol. 22, no. 8, pp. 734-748, (2011).

[SVE00] Svensson E., “Comparison of the quality of assessments using continuous and discrete ordinal rating scales”. Biometrical J., vol. 42, no. 4, pp. 417–434, (2000).

[TEU96] Teunissen K., “The validity of CCIR quality indicators along a graphical scale”, SMPTE J., vol. 105, no. 3, pp. 144–149, (1996).

[UDA13] Chaitanya dornadula U., “Implementation of No-Reference Image Quality Assessment in

Contourlet Domain". Thesis manuscript, (2013).

[URV12] Urvoy M., Barkowsky M., Cousseau R., Koudota Y., and Ricordel V., "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences". QoMEX - Fourth Inter-national Workshop on Quality of Multimedia Experience, Jul 2012, Yarra Valley, Australia, pp.1-6, (2012).

[VIR95] Virtanen M.T., Gleiss N., and Goldstein M., "On the use of evaluative category scales in telecommunications," Proc. Human Factors Telecommun., 253–260, (1995).

[WAL93] Walpole R.E., Myers R.W, Myers S.L, and Ye K., "Probability & statistics for engineers and scientists". NewYork: Macmillan, (1993).

[WAS98] Watson A., and Sasse A., "Measuring perceived quality of speech and video in multimedia conferencing applications". Proc. ACM Multimedia Conf., pp. 55–60, (1998).

[WEB01] https://www.oxforddictionaries.com/definition/quality

[WEB02] https://www.iso.org/iso-9001-quality-management.html

[WEB03] https://www.wikipedia.org

[WEB04] https://www.statistica.com

[WEB05] "ACR (Absolute Category Rating) Test Interface "http://www.acceptv.com, (2012)

[CHE12-02] Chen w., Fournier J., Barkowsky M., Le Callet P., "Quality of experience model for 3DTV". SPIE Stereoscopic Displays and Applications XXIII, Jan 2012, San franscisco, United States. 8288 (59), pp.1-6, (2012).

[WIN03] Winkler S. and Campos R., "Video quality evaluation for Internet streaming applications". Proc. SPIE Human Vision and Electronic Imaging, Santa Clara, CA, vol. 5007, pp. 104–115, (2003).

[WIN09] Winkler S., "On the properties of subjective ratings in video quality experiments". In Proc. Int. Workshop Quality Multimedia Exper.(QoMEX), San Diego, CA, (2009).

[ZIE07] Zieliñski S., Brooks P., and Rumsey F., “On the use of graphic scales in modern listening tests”. Proc. 123rd AES Convention, New York, (2007).

List of publications

Published papers

BENSAIED R., MITREA M., CHAMMEM A., EBRAHIMI T., “Subjective quality assessment for stereoscopic video.Case study on robust watermarking”, Proc. SPIE 9011, March 2014.

BENSAIED R., MITREA M., CHAMMEM A., EBRAHIMI T.,” Continuous vs. discrete scale stereoscopic video subjective evaluation: case study on robust watermarking”, in Proc. of Quality of Multimedia Experience (QoMEX),Sixth International Workshop on, pp. 238 - 244, DOI: 10.1109/QoMEX.2014**.**

BENSAIED R., MITREA M., “Assessing the impact of the semantic labels in subjective video quality evaluation”. IMA 2016. 11th IMA International Conference on Mathematics in Signal Processing, Dec 2016, Birmingham, United Kingdom. Institute of Mathematics and its Applications (IMA), pp.16-20, 2016, 11th IMA International Conference on Mathematics in Signal Processing

BENSAIED R., MITREA M., « L'impact des étiquettes sémantiques dans l'évaluation subjective de la qualité des séquences vidéo 2D et stéréoscopiques »TAIMA 2018.

List of acronyms

DMOS	Differential Mean Opinion Scores
MOS	Mean Opinion Scores
CCIR	Consultative Committee on International Radio
ITU	International Telecommunication Union
3DV-NTSS	3 Dimension Video New Three Step Search
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
FS-MPEG	Full Search MPEG
HD 3D TV	High Definition 3 Dimension Television
ITU	International Telecommunication Union
JPEG	Joint Photographic Experts Group
LCD	Liquid crystal Display
QIM	Quantization Index Modulation-
SS	Spread Spectrum
SSCQE	Single Stimulus for Continuous Quality Evaluation
SSIM	Structural SIMilarity
NCC	Normalized Cross Correlation