FULL-REFERENCE OBJECTIVE QUALITY METRICS FOR VIDEO WATERMARKING, VIDEO SEGMENTATION AND 3D MODEL WATERMARKING

THÈSE N^{O} 3411 (2005)

PRÉSENTÉE À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

Institut de traitement des signaux

SECTION DE GÉNIE ÉLECTRIQUE ET ÉLECTRONIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Elisa DRELIE GELASCA

dottore in ingegneria elettronica, Università degli Studi di Trieste, Italie
et de nationalité italienne

acceptée sur proposition du jury:

Prof. T. Ebrahimi, directeur de thèse
Prof. S. Mitra, rapporteur
Prof. G. Sicuranza, rapporteur
Prof. S. Süsstrunk, rapporteur

Lausanne, EPFL
2005

To my parents,
Gianna and Mario,
for encouraging me
to pursue my
dreams.

He who becomes the slave of habit, who follows the same routes
every day, who never changes pace, who does not risk and
change the color of his clothes, who does not speak and
does not experience, dies slowly.

He or she who shuns passion, who prefers black on white,
dotting ones "i's" rather than a bundle of emotions,
the kind that make your eyes glimmer,
that turn a yawn into a smile, that make the heart pound
in the face of mistakes and feelings, dies slowly.

He or she who does not turn things topsy-turvy,
who is unhappy at work, who does not risk certainty
for uncertainty, to thus follow a dream, those who do not forego
sound advice at least once in their lives, die slowly.

He who does not travel, who does not read,
who does not listen to music, who does not find grace in himself,
she who does not find grace in herself, dies slowly.

He who slowly destroys his own self-esteem,
who does not allow himself to be helped,
who spends days on end complaining about his own bad luck,
about the rain that never stops, dies slowly.

He or she who abandon a project before starting it,
who fail to ask questions on subjects he doesn't know,
he or she who don't reply when they are asked something
they do know, die slowly.

Let's try and avoid death in small doses, reminding oneself
that being alive requires an effort far greater than
the simple fact of breathing.
Only a burning patience will lead to the attainment of
a splendid happiness.

Pablo Neruda

Contents

Contents vii

Acknowledgment xi

Abstract xiii

Version Abrégée xv

Symbols and Acronyms xvii

List of Figures xix

List of Tables xxv

1 Introduction 1
- 1.1 Motivations 1
- 1.2 Investigated Approach 2
- 1.3 Organization of the Thesis 3
- 1.4 Main Contributions 4

I Image and Video Processing Quality Assessment 5

2 Objective and Subjective Quality 7
- 2.1 Introduction 7
- 2.2 Subjective Quality Assessment 8
  - 2.2.1 Subjective Measurement Methods 9
  - 2.2.2 Grading Scales 10
  - 2.2.3 Experiment Requirements and Limitations 11
- 2.3 Subjective Data Analysis 12
  - 2.3.1 Processing the Subjective Data 12
  - 2.3.2 Psychometric and Fitting Functions 13
  - 2.3.3 Minkowski Metric 14
- 2.4 Objective Quality Assessment 15
  - 2.4.1 Objective Models and Metrics 16

- 2.4.2 Objective Model Assessment . . . 17
- 2.5 Conclusions . . . 18

3 Video Watermarking Quality Assessment 19
- 3.1 Introduction . . . 19
- 3.2 Video Watermarking - Motivation and Context . . . 20
- 3.3 Proposed Approach . . . 21
- 3.4 Watermarking Algorithms and Artifacts . . . 22
- 3.5 Experimental Method . . . 24
  - 3.5.1 Test Clips . . . 24
  - 3.5.2 Subjective Experiments . . . 24
- 3.6 Proposed Objective Metrics . . . 25
- 3.7 Experimental Results . . . 27
- 3.8 Conclusions . . . 29

II Video Object Segmentation Quality Assessment 31

4 Subjective Segmentation Evaluation 33
- 4.1 Introduction . . . 33
- 4.2 Motivations . . . 34
- 4.3 Subjective Segmentation Evaluation Methods - Background . . . 36
  - 4.3.1 Still Images . . . 36
  - 4.3.2 Image Sequences . . . 38
- 4.4 Proposed Method for Subjective Evaluation . . . 40
  - 4.4.1 Basic Methodology . . . 41
  - 4.4.2 Procedure . . . 41
  - 4.4.3 Types of Experimental Tasks . . . 43
- 4.5 From Subjective to Objective Evaluation . . . 45
  - 4.5.1 Segmentation Errors . . . 46
  - 4.5.2 Generation of Synthetic Segmentation Errors and Test Sequences . . . 49
- 4.6 Conclusions . . . 55

5 Objective Segmentation Evaluation 57
- 5.1 Introduction . . . 57
- 5.2 Objective Segmentation Evaluation Methods - Background . . . 57
  - 5.2.1 Still Images . . . 58
  - 5.2.2 Video Object Segmentation . . . 59
- 5.3 Proposed Objective Evaluation Metric . . . 67
  - 5.3.1 Spatial Artifacts . . . 67
  - 5.3.2 Temporal Artifacts . . . 69
- 5.4 Perceptual Impact of Segmentation Artifacts . . . 70
  - 5.4.1 Perceptual Spatial Errors . . . 71
  - 5.4.2 Perceptual Temporal Errors . . . 78
  - 5.4.3 Perception of Combined Artifacts . . . 79
- 5.5 Overall Perceptual Objective Metric . . . 86
- 5.6 Conclusions . . . 90

6 Objective Metric Performance and Applications 93

6.1 Introduction 93
6.2 Segmentation Assessment - General Framework 93
6.2.1 Video Object Segmentation Algorithms 94
6.2.2 Subjective Experiment Results 95
6.2.3 Objective Metric Results 102
6.3 Segmentation Assessment - Application Dependent 103
6.3.1 Video Object Segmentation Applications 103
6.3.2 Subjective Experimental Methodology 107
6.3.3 Application Dependent Objective Metric 109
6.4 Conclusions 109

III 3D Object Watermarking Quality Assessment 113

7 Subjective and Objective 3D Watermarking Evaluation 115

7.1 Introduction 115
7.2 Related Work 116
7.2.1 HVS and Perceptual Image Watermarking 117
7.2.2 Mesh Simplification 117
7.2.3 Perceptually-Guided Rendering 119
7.2.4 Proposed Approach 120
7.3 3D Watermarking Algorithms and Artifacts 120
7.4 Experimental Method 121
7.4.1 Rendering Conditions 121
7.4.2 Experimental Procedure 124
7.4.3 Experiment I 127
7.4.4 Generation of Stimuli 127
7.4.5 Experiment II 129
7.4.6 Generation of Stimuli 129
7.5 Data Analysis 130
7.5.1 Normalization 130
7.5.2 Data Evaluation 133
7.6 Proposed Perceptual Metric 135
7.6.1 Roughness Estimation 136
7.7 Experimental Results 139
7.7.1 Hausdorff distances 139
7.7.2 Roughness-based Metrics Results for Experiment I 140
7.7.3 Objective Metrics Performances 140
7.8 Conclusions 142

8 Conclusions 147

8.1 Summary and Contributions 147
8.2 Perspectives 149

A Video Object Segmentation Artifacts 153

B Experiment Scripts **161**
B.1 Sample Instructions for Segmented Object Annoyance Task 161
B.2 Sample Instructions for Segmented Object Strength Task 164
B.3 Sample Instructions for Compression Application Annoyance Task 167
B.4 Sample Instructions for Augmented Reality Application Annoyance Task 169
B.5 Sample Instructions for Video Surveillance Application Annoyance Task 171
B.6 Sample Instructions for Watermarked 3D Model Distortion Task 173

C 3D Model Rendering **177**
C.1 Illumination Models 178
C.2 Light Sources 180
C.3 Basic Shading Techniques 181
C.4 Texturing 182
C.5 Global Illumination 184

Bibliography **187**

Curriculum Vitae **201**

Acknowledgment

A PhD thesis is far from being only a personal work and its accomplishment requires the help, the support and the collaboration of several other people. Among them, my gratitude goes to Prof. Murat Kunt who gave me the opportunity to join the Signal Processing Institute where I spent four years in a unique and enriching research environment. My greatest thank goes to my advisor Prof. Touradj Ebrahimi for providing inspiration and instilling confidence, as he does in all his students, particularly when it was most needed during the last year.

I would also like to express my gratitude to the members of my jury, Prof. Mitra, Prof Sicuranza and Prof. Süsstrunk who took the time to carefully read this thesis and provided me with with a very valuable feedback.

I am grateful to Prof. Sicuranza who gave me the opportunity to carry out my diploma project at EPFL. Thanks to this event, to Roberto and to Elena I started and pursued a PhD work.

Special thanks go to Prof. Mitra for welcoming me for several research visits at his Lab. where I enjoyed doing research.

Ladies first: Dr. Elena, Dr. Merixtell, Roberto, Marco, Dr. Stefan and Zenichi have provided valuable inputs and comments to help me in writing and editing this thesis. I would also like to thank all the people at UCSB for their helpful advices and for generously allowing me to use their subjective testing facilities: Dr. Mylene Farias, Prof. John Foley, Prof. Sanjit Mitra.

My gratitude goes also to three great researchers I had the chance to meet: Mylene Farias who supported me with helpful advices during my four-month staying at UCSB and Massimiliano Corsini for the nice and fruitful collaboration we had during his seven-month staying at EPFL. Thank you both for the great job and friendship. Marco Zuliani for all his encouragement, helpful technical advices and continuous lovely support along the last hardest period without which I would have never managed to carry the burden of work.

I am also indebted to the staff people at LTS, Marianne and Gilles, who have helped me during my staying here. For their contributions I would like to thank the students that I supervised during my PhD especially Danko Tomasic and Gaia Arrigoni. Many thanks to all the nice colleagues with whom I shared the office during these 4 years, they were patient when I was stressed out: Dr. Xavier and Dr. Lorenzo who did appreciate artistic calendars, Dr. Christophe who had patience with my Italian French, Dr. Mattia with whom I shared the last period full of doubts, difficulties but also nice jokes and candies. Special thanks to Dr. Lam and Jean Marc's boys for rushing into my office and having fun while I was trying to write my thesis and preparing my exam.

Doing a thesis involves more than just algorithms, papers and experiments: one must try to maintain sanity as a background process. For this reason I am especially grateful to my family, friends and water-polo team. My parents providing continuing love and nice jokes so that even in the hardest moments I was laughing at the Amanda's and Nicholas' funny stories. My warmest thank goes to my tupper ladies who assisted me in the finest details as if they were my step-family here at

Lausanne: Nawal, Maria Eugenia and Valerie, Meri the mom and Ruth the dad (always busy). All the LTS girls and boys, Rosa and Oscar for their generosity in throwing me the vegetables I needed through the window, Plamen and Ulrich I enjoyed to scare with my driving style, both Lorenzos, Patricia and Lisa for the nice time shared together. Special thanks go to my Italian connections: Matteo the most handsome 'ciociaro' in the Lab., Gianluca and Roberto who strongly encouraged and supported me even in the most stressful periods when I was hard to stand, and Elena, Laura, Giulia, Fabiana and Annetta back in Italy.

Last but by no means least, I thank God for the special two things that Marco knows.

Abstract

Quality assessment is a central issue in the design, implementation, and performance testing of all systems. Digital signal processing systems generally deal with visual information that are meant for human consumption. An image, a video, or a 3D model may go through different stages of processing before being presented to a human observer, and each stage of processing may introduce distortions that could reduce the quality of the final display. To conceive quantitative metrics that can automatically predict the perceived quality, the way humans perceive such distortions has to be taken into account and can be greatly beneficial for quality assessment. In general, an objective quality metric plays an important role in a broad range of applications, such as visual information acquisition, compression, analysis and watermarking. Quality metrics can be used to optimize algorithm parameter settings and to benchmark different processing systems and algorithms.

In this dissertation, new objective quality metrics that take into account how distortions are perceived, are proposed and three different signal processing systems are considered: video watermarking, video object segmentation and 3D models watermarking.

First, two new objective metrics for watermarked video quality assessment are proposed. Based on several different watermarking algorithms and video sequences, the most predominant distortions are identified as spatial noise and temporal flicker. Corresponding metrics are designed and their performance is tested through subjective experiments.

Second, the problem of video object segmentation quality evaluation is discussed, proposing both subjective evaluation methodology and perceptual objective quality metric. Since a perceptual metric requires a good knowledge of the kinds of artifacts present in segmented video objects, the most typical artifacts are synthetically generated. Psychophysical experiments are carried out to study the perception of individual artifacts by themselves or combined. A new metric is proposed by combining the individual artifacts using the Minkowski metric and a linear model. An in-depth evaluation of the performance of the proposed method is carried out. The obtained perceptual metric is also used to benchmark different video object segmentation techniques for general frameworks as well as specific applications, ranging from object-based coding to video surveillance.

Third, two novel metrics for watermarked 3D model quality assessment are proposed on the basis of two subjective experiments. The first psychophysical experiment is carried out to investigate the perception of distortions caused by watermarking 3D models. Two roughness estimation metrics have been devised to perceptually measure the amount of visual distortions introduced on the model's surface. The second psychophysical experiment is conducted in order to validate the two proposed metrics with other watermarking algorithms.

All of the proposed metrics for the three kinds of visual information processing systems are based on the results of the psychophysical experiments. Subjective tests are carried out to study and characterize the impact of distortions on human perception. An evaluation of the performance of these perceptual metrics with respect to the most common state of the art objective metrics is performed. The comparison shows a better performance of the proposed perceptual metrics than that of the state of the art metrics. The performance is investigated in terms of correlation with subjective opinion. The results demonstrate that including the perception of distortions in objective metrics is a reliable approach and improve the performance of such metrics.

Version Abrégée

Le contrôle de qualité est un problème essentiel dans l'implémentation, la conception et les tests de performance de tout système. L'analyse du signal numérique traite généralement d'information visuelle destinée à l'oeil humain. Avant d'être présenté à l'oeil humain, une image, une vidéo ou un modèle 3D passe par différentes étapes de traitement, ce qui conduit à ajouter d'une étape à l'autre des distorsions qui peuvent altérer la qualité du signal final. Afin de développer une métrique quantitative qui peut automatiquement évaluer la qualité perçue, la manière dont sont perçues les distorsions par l'oeil humain doit être prise en compte et peut s'avérer très bénéfique pour le contrôle de qualité. En général une mesure objective de qualité peut jouer un rôle important sur une grande variété d'applications telles que l'acquisition, la compression, l'analyse et le tatouage numérique d'informations visuelles. Les mesures de qualité peuvent être utilisées d'une part pour optimiser le réglage de paramètres de l'algorithme et d'autre part pour tester et évaluer différents algorithmes et systèmes de traitement. Dans ce mémoire, on propose de nouvelles mesures objectives de qualité qui prennent en compte la perception humaine des distorsions.

On considérera dans cette thèse différents systèmes de traitement du signal: tatouage numérique vidéo, segmentation d'objet vidéo et des modèles de tatouage numérique 3D.

Dans un premier temps, deux nouvelles mesures pour le contrôle de qualité du tatouage numérique vidéo sont proposées. En se basant sur différents algorithmes de tatouage numérique et de séquences vidéo, les distorsions prédominantes sont issues du bruit spatial et temporel. Ces mesures sont construites et leur performance est testée à travers des expériences subjectives.

En deuxième lieu, l'évaluation de la qualité de la segmentation d'objet vidéo est discutée à l'aide d'une méthodologie d'évaluation subjective et une mesure de qualité perceptuelle objective. Sachant qu'une mesure perceptuelle demande une bonne connaissance des types d'artéfact présents dans l'objet vidéo segmenté, quatre des artéfacts les plus courants sont générés synthétiquement. Des expériences psychophysiques sont menées pour l'étude de la perception d'artéfacts isolés ou combinés. La mesure perceptuelle est testée sur différentes techniques de segmentation d'objet vidéo que ce soit dans un cadre général ou dans des applications plus spécifiques allant du codage basé objet à la vidéo surveillance.

Pour terminer, deux nouvelles mesures de contrôle de qualité pour les modèles 3D de tatouage numérique, basées sur des expériences subjectives, sont proposées. La première expérience psychophysique est menée pour étudier la perception des distorsions causées par les modèles de tatouage numérique 3D. Deux métriques grossières d'estimation ont été élaborées pour mesurer perceptuellement la quantité de distorsion visuelle introduite sur la surface du modèle. La deuxième expérience psychophysique valide les deux métriques proposées avec d'autres algorithmes de tatouages numériques.

Toutes les métriques proposées pour les trois types de systèmes de traitement d'information visuelle sont basées sur les résultats d'expériences psychophysiques. Les tests subjectifs sont réal-

isés pour étudier et caractériser l'impact des distorsions sur la perception humaine. On réalise l'évaluation de la performance de ces métriques perceptuelles par rapport à l'état de l'art. L'étude comparative montre une meilleure performance des métriques perceptuelles proposées dans cette thèse par rapport à l'état de l'art. On évalue la performance en terme de corrélation avec l'opinion subjective. Les résultats démontrent que l'approche consistant à inclure la perception des distorsions dans les métriques objectives est fiable.

Symbols and Acronyms

Principal Symbols

$\mathbf{B}$	indicator function for expectation effect
δ_j	confidence interval for the test scene j
$\mathcal{A}_b$	added background
$\mathcal{A}_r$	added region
$\mathbf{F}(k)$	flickering artifact at frame k
$\mathcal{H}_b$	boundary hole
$\mathcal{H}_i$	inside hole
MPEGqm	MPEG quality metric
mqm	matching quality metric
M^W	watermarked model
$\mathcal{N}$	false negative pixels
μ_j	subjective MOS for the test scene j
$\mathcal{P}$	false positive pixels
PST	perceptual spatio temporal objective metric
r	goodness of fit, correlation coefficient
r_P	Pearson correlation coefficient
r_S	Spearman correlation coefficient
$\mathcal{R}(M, M^W)$	total roughness of the model M vs. the watermarked model M^W
ST	spatio temporal objective metric
$w_t(k)$	perceptual temporal weight at frame k
wqm	weigthed quality metric

Acronyms

art.live	ARchitecture and authoring Tools prototype for Living Images and new Video Experiments
BER	Bit Error Rate
CIF	Common Interchange Format
CSF	Contrast Sensitivity Function
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale Method
DVD	Digital Versatile Disc or Digital Video Disc
DWT	Discrete Wavelet Transform
HVS	Human Visual System
IST	Information Society Technologies
ITU	International Telecommunications Union
JND	Just Noticeable Difference
JPEG	Joint Photographic Experts Group
MAV	Mean Annoyance Value
MOS	Mean Opinion Score
MPEG	Motion Picture Experts Group
MSE	Mean Squared Error
MSV	Mean Strength Value
NBA	Normal Bin Coding
PSNR	Peak Signal to Noise Ratio
PQoS	Perceptual Quality of Service
PMAV	Perceptual Mean Annoyance Value
QoS	Quality of Service
RMS	Root Mean Square
ROC	Receiver Operating Characteristics
SSCQE	Single Stimulus Continuous Quality Evaluation
VFA	Vertex Flood Algorithm
VQEG	Video Quality Experts Group
YUV	Full-color video signal format: Y (luminance), U and V (chrominance)

List of Figures

2.1 Continuous grading scales: (a) five point continuous scale, and (b) categorical continuous scale. 10

3.1 General scheme of watermarking system. 20
3.2 Sample frames from the test clips. 22
3.3 Intensity errors produced by the watermark obtained by four different algorithms for the frame shown in Fig. 3.2(b). Dark pixels denote negative values, bright pixels denote positive values, medium gray denotes no change. The images were normalized individually to enhance the visibility of the watermarks. 24
3.4 Visibility problem in consecutive frames. 25
3.5 Perceived noise vs. Noise metric and PSNR (subjective data are shown with 95%-confidence intervals). 27
3.6 Perceived flicker vs. Flicker metric and PSNR (subjective data are shown with 95%-confidence intervals). 28

4.1 An example of edge segmentation evaluation sheet from Heat *et al.*'s method [96]. . 37
4.2 Test images for evaluating the best segmentation [137]: (a) shows the first stage of the trial with the original image and (b) shows the second stage in which the segmentation schemes are represented by their outline and masks with the original in the center. . 38
4.3 Each image has been segmented by 3 different people from the Berkeley segmentation dataset [90]. 39
4.4 Typical display configuration adopted in subjective experiment: in the center the segmented video under test. 42
4.5 Display configurations for training stage: (a) the subject is told about how an ideal segmentation looks like, (b) typical segmentation errors are displayed along with the ideal segmentation, (c) the worst segmentations are shown to establish a subjective range of the annoyance scale. 44
4.6 Dialog boxes used for the experimental tasks: (a) annoyance and (b) strength. . . . 45
4.7 Block diagram of the proposed approach: from subjective to objective video object segmentation quality evaluation. 46
4.8 Various segmentations that have equal pixel distance measures and also the same number of misclassified pixels [189]. 47
4.9 Reference segmentation overlapped with the resulting segmentation, at frame k: (a) shows the two kinds of subsets of false positives, $\mathcal{P}(k)$, (b) shows the two kinds of subsets of false negatives, $\mathcal{N}(k)$ 48

4.10 Sample frames of original video sequences: (a) ‘Highway’, (b) ‘Group’, (c) ‘Hall monitor’, (d) ‘Coastguard’. 50
4.11 Temporal insertion of artifacts during 10 frames in different moments of the video sequence. 52

5.1 Samples of ground truth for tracking evaluation through bounding box. ‘Highway’ video sequence: (a) frame #89, (b) frame #101. 60
5.2 Sample of ground truth for tracking evaluation through center of gravity. 60
5.3 Sample of original video sequence ‘Group’ in (a) and the corresponding ideal object segmentation (ground truth) in (b). 61
5.4 Weighting functions for false positives and false negatives for the method of Villegas *et al.* [154]. 65
5.5 Example of border holes with the same amount but different distance from the ideal contour (a) large spatial errors (b) small spatial errors. 68
5.6 Weighted function considering human memory in video quality evaluation proposed in [63]. 70
5.7 Values of annoyance predicted with artifact metric ST_{A_r} for both added region shapes, square s_1 and circle s_2, vs. observed subjective annoyance. In (a) the confidence interval at 95% are plotted along with subjective data. In (b) the fitting Weibull functions are depicted for each shape. 72
5.8 Values of annoyance predicted with artifact metric ST_{A_r} vs. observed subjective annoyance, fitted with an overall Weibull function. 73
5.9 Values of annoyance predicted with artifact metric ST_{A_r} vs subjective annoyance for added regions at different positions from the ground truth: p_1, p_2 and p_3 slightly further. (a) shows the three fitting curves for the three different positions and (b) the unique fitting for all the data. 74
5.10 Values of annoyance predicted with artifact metric ST_{A_b} vs subjective annoyance for added background: (a) shows the fitting curve and its confidence interval for Experiment I and (b) shows the correlation of the derived perceptual metric in Experiment II. 75
5.11 Annoyance curves predicted with artifact metric ST_H vs. observed subjective annoyance for both kind of holes: inside holes H_i and border holes H_s 76
5.12 Correlation of flickering metrics with subjective annoyance: (a) proposed flickering metric for Added Region $\mathbf{F}_{a_r}$, (b) MPEG temporal coherence metric, $MPEGTqm$. . . 78
5.13 Experiment on Expectation effect: (a) temporal defect **B** vs. perceived annoyance for ‘Coastguard’ and ‘Hall’ video sequences; (b) temporal weighting function taking into consideration the Expectation effect. 80
5.14 Mean Strength Value (MSV) obtained in combined artifact experiment for all the video sequences. 81
5.15 Mean Strength Value (MSV) obtained in combined artifact experiment for all the video sequences. 82
5.16 Mean Strength Value (MSV) obtained in combined artifact experiment for all the video sequences. 83
5.17 Subjective values for MSV vs. MAV and their correlation through the Minkowski metric: the fitting parameter and the correlation are reported for each sequence: (a) ‘Coastguard’, (b), ‘Group’, (c) ‘Hall’, (d) ‘Highway’. 85
5.18 Subjective values for MSV vs. MAV and their correlation through the Minkowski metric: the fitting parameter and the correlation are reported. 85

5.19 Correlation of the proposed perceptual metric **PST** with subjective annoyance: first row 'Coastguard', second row 'Group', third row 'Hall monitor'. (a), (c), (e) with the optimal p and (b), (d), (f) with linear fitting $p = 1$. 87

5.20 Correlation of the proposed perceptual metric **PST** with subjective annoyance for 'Highway' and all the video sequences: (g) and (i) with the optimal p and (h) and (l) with linear fitting $p = 1$. 88

5.21 State of the art objective metrics vs. subjective annoyance: (a) MPEG metric, (b) Villegas' metric, (c) Nascimento's metric. 89

6.1 MAV values obtained for each segmentation algorithm and averaged on the three tested video sequences. 97

6.2 Sample frames for reference segmentation of the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105). 97

6.3 Sample frames for the *Kim*'s algorithm segmentation results for the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105). 98

6.4 Sample frames for the *Image Differencing* algorithm segmentation results for the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105). 99

6.5 Sample frames for the *McKenna*'s algorithm segmentation results for the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105). 99

6.6 Sample frames for the *François*'s algorithm segmentation results for the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105). 100

6.7 Sample frames for the *Horprasert*'s algorithm segmentation results for the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105). 100

6.8 Sample frames for the *Jabri*'s algorithm segmentation results for the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105). 101

6.9 Sample frames for the *Shen*'s algorithm segmentation results for the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105). 101

6.10 Objective metrics vs. Subjective Scores (MAV) and correlation coefficients. 104

6.11 Sample frames for video coding segmentation applications 'Hall monitor', 'Group', and 'Highway'. 105

6.12 Sample frames for video surveillance segmentation application 'Group', 'Hall monitor' and 'Highway'. 106

6.13 Sample frames for video augmented reality segmentation application 'Group', 'Hall monitor' and 'Highway'. 107

6.14 Graphical interface example for application dependent segmentation evaluation: tutorial stage. 107

6.15 Graphical interface examples for application dependent segmentation evaluation: experiment trial. 108

6.16 Objective metric **PST** vs. Subjective Scores (MAV) and correlation coefficients for different segmentation applications: with no optimized (a), (c), (e) and optimized (b), (d), (f) Minkowski parameters. 110
6.17 State of the art objective metrics vs. Subjective Scores (MAV) and correlation coefficients for different segmentation applications: (a), (b), (c) compression, (d), (f), (g) surveillance and (h), (i), (l) augmented reality. 111

7.1 Digital Michelangelo Project. (Left) David's acquisition. (Right) The reconstructed three-dimensional model. 116
7.2 Image distortions and RMS metric (from Teo and Heeger [147]). 119
7.3 Geometric defects introduced by 3D watermarking algorithms. 122
7.4 The subject during the interaction with the model. 126
7.5 The subject during the selection of the part where the distortions are more evident. 126
7.6 Rendering Conditions. 128
7.7 Subject data correction results for Experiment I: (a) the mean scores values and (b) the confidence intervals 134
7.8 Subject data correction results for Experiment II: (a) the mean scores values and (b) the confidence intervals 134
7.9 Dihedral angle. 136
7.10 $G(.)$ and $V(.)$ descriptions. 136
7.11 Bumps with different scale. 137
7.12 Smoothing-based Roughness Estimation. 138
7.13 Geometric Hausdorff distance vs Subjective MOS. 140
7.14 Experiment I: Subjective MOS versus objective metrics curves fits. 141
7.15 Experiment II: Subjective MOS vs objective metric curves. The parameters of the fitting curve are the same of the Experiment I. 143
7.16 Experiment II: Comparison between models' impairment ranking. On the left the models are ranking by subjective MOS. On the right the models are ranking by smoothing-based roughness. 144

A.1 Sample frames for different amounts and shapes of the **added region** artifact for the video sequence "Group". 153
A.2 Sample frames for different amount of the **added background** artifact (c is equal to number of pixels on the contour object) for the video sequence "Hall monitor". . . . 154
A.3 Sample frames for different amount of the **border hole** artifact for the video sequence "Group". 155
A.4 Sample frames for different amount of the **inside hole** artifact for the video sequence "Group". 156
A.5 Sample frames for **flickering artifact** with period 3 for the video sequence "Hall". . 157
A.6 Sample frames for **expectation effect** with condition $\mathbf{B}_6$ for the video sequence "Coastguard". 158
A.7 Sample frames for **combined artifacts**. Examples of one, two, three, four different artifacts (corresponding to combinations 4, 17, 26 and 44) for the video sequence "Highway". 159

C.1 Ray Tracing schematization. 178
C.2 Light-material interaction. 179
C.3 Microfacets model of a reflecting surface. 179

C.4 Phong model schematization. 180
C.5 Lighting models (from [83]). 180
C.6 Different kind of light sources. 181
C.7 Basic shading techniques for polygonal meshes. 182
C.8 Different texture projections. 182
C.9 An example of Bump Mapping. 183
C.10 Gloss mapping. The shininess of the shield on the right is modulated in order to appear more realistic. 183
C.11 Global Illumination Effects. (Left) Ray Tracing can be used to reproduce “hard” shadows and reflections. (Right) Photon Mapping can simulate any kind of visual effects produced by illumination in real scenes. 184
C.12 Rendering conditions schematization. 185

List of Tables

2.1 Definitions used in developing quality assessment systems. 9

3.1 Tested watermarking algorithms and their parameters (n is the length, α the strength of the embedded watermark). 23

4.1 Evaluation criteria in McKonen *et al.* [95]'s method. 40
4.2 Viewing conditions during subjective test. 40
4.3 Summary of original video sequences used in subjective experiments. 50
4.4 Tested segmentation artifacts and their values. 53
4.5 Set of amount combinations (pixels) for added regions, added background, inside holes and border holes used in the experiment. 54

5.1 Objective Measures used in evaluating video tracking systems. 61
5.2 Evaluation criteria used in empirical and analytical evaluation of image and video object segmentation systems. 66
5.3 Summary of evaluation criteria used by each reviewed method. 66
5.4 Objective Measures used in evaluating image and video object segmentation systems. 67
5.5 Fitting Parameters S and k for Weibull function. 72
5.6 F values to test if different fitting curves are needed to describe the perceived annoyance for different shapes and positions of added regions. 74
5.7 MSVs and MAVs values for all segmented video sequences and all combination used in combined artifact experiment. the experiment. 91
5.8 Subjective fitting for combined artifact experiment: Minkowsi fitting parameter and correlation values for all the video sequences. 92
5.9 F values to test if different fitting curves are needed to describe the perceived annoyance for different shapes and positions of added regions. 92
5.10 Minkowski parameters and correlation (a) best p fitted, (b) p=1. 92

6.1 MAV values obtained for each segmentation algorithm and the correspondent confidence interval δ for all the test video sequences. 96
6.2 Description of artifacts introduced by the real segmentation algorithms and their perceived strengths gathered in the interview stage. 96
6.3 Correlation coefficients for the averaged objective metrics for all the sequences vs. averaged subjective annoyance. 103
6.4 Viewing conditions during subjective test. 108
6.5 Correlation coefficients for the objective metrics vs. subjective annoyance. 112

7.1 Viewing conditions during subjective tests. 127
7.2 Experiment I: Watermarking parameters values used for each model. 129
7.3 Experiment II: Watermarking parameters values. 130
7.4 Experiment I: ANOVA analysis results. 131
7.5 Experiment II: ANOVA analysis results. 131
7.6 Experiment I: ANOVA analysis results for $\ln(m_{ij} + 1)$. 132
7.7 Experiment II: ANOVA analysis results for $\ln(m_{ij} + 1)$. 132
7.8 ANOVA analysis results after normalization (F-test values for subject Offset and Gain dependency). Probabilities near one means that there is no difference across subjects. 133
7.9 Perceptual metrics performances. 142

Introduction

1

1.1 Motivations

The field of digital data processing deals, in large part, with signals that are meant to convey reproductions and manipulations of visual information for human consumption. A visual data may go through many stages of processing before being presented to a human observer, and each stage of processing may introduce distortions that could reduce the quality of the final display. For example, a compression algorithm can be applied to visual data to reduce the bandwidth requirements for storage or transmission. In watermarking systems, imperceptible data can be inserted in the visual information to authenticate the ownership. In various areas of multimedia, such as content-based information retrieval, remote surveillance and entertainment, analysis systems are developed for the access and manipulation of multimedia content. The amount of distortions that each of these systems could add depends mostly on its intrinsic characteristics and the physical properties of the processed visual data.

One obvious way of determining the quality of visual information is to measure it by means of psychophysical experiments with human subjects. After all, these signals are meant for human consumption. However, such subjective evaluations are not only time-consuming and expensive, but they also cannot be incorporated into systems that adjust themselves automatically based on the feedback of measured output quality.

The goal of this research is to develop quantitative measures for objective quality assessment that can automatically predict the quality perceived by human subjects. Generally speaking, an objective quality metric can play an important role in a broad range of applications, such as image acquisition, compression, communication, analysis and watermarking. First, it can be used to dynamically monitor and adjust visual quality. Second, it can be used to optimize algorithms and parameter settings of visual data processing systems. Third, it can be applied to benchmark processing systems and algorithms.

In short, the objective quality measurement (as opposed to subjective quality assessment by human observers) seeks to determine algorithmically the quality of visual data. The purpose of this thesis is *to design algorithms whose quality prediction is in good agreement with subjective scores* *from human observers for different kinds of visual information.*

A growing interest has emerged over the last few years within the computer vision community in the investigation of quality of new kinds of visual information. In this thesis, particular focus will be given to the investigation of the quality of three types of visual data:

- *watermarked video sequence*;
- *segmented video object*;
- *watermarked 3D object*.

Quality assessment is of great impact within different research areas. It can be considered the focal point of a converging series of multidisciplinary researches such as vision, image and video analysis and experimental psychology.

1.2 Investigated Approach

Traditionally, researchers have focused on measuring signal fidelity as the means of assessing visual quality. Signal fidelity is measured with respect to a reference signal that is assumed to have “perfect” quality. In this thesis, *full reference* quality assessment methods are adopted to assess visual quality of the different kinds of visual data. During the design or evaluation of a system, the reference signal is typically processed to yield a distorted (or test) visual data, which can then be compared to the reference using so-called full reference methods. Typically, this comparison involves measuring the “distance” between the two signals in a *perceptually* meaningful way. This can be achieved by studying, characterizing and deriving the perceptual impact of the distorted signal to human viewers by means of subjective experiments. Our approach consists in taking into account the most common artifacts produced by processing algorithms and carrying out subjective experiments in order to:

1. study and characterize the impact of different distortions;
2. propose an automatic procedure for evaluating the performance of the processing algorithms;
3. validate the proposed automatic method on the basis of correlation with the human perception of visual quality.

ITU-T Recommendation [65] describes standard methods for subjectively testing the quality measurement of processed images and video. Such methods based on psychophysical experiments can be applied in image and video processing systems like encoding and watermarking. However, there are image and video analysis systems such as image and video segmentation where these standard procedures for subjective testing cannot be straightforwardly applied. In this case, methods for subjective quality assessment have to be adapted to the fact that the output, the segmented region/object, constitutes only a part of the input, the original video. For example, how to display (e.g. on what background) the segmented region/object has to be carefully studied in the subjective evaluation procedure.

For different reasons, these standard methods have to be adapted also to 3D object subjective quality evaluation. In fact, the way the visual inspection of a 3D object is performed is different from that of a processed image/video. Displaying 3D objects includes rendering condition setting (such as illumination) and human interaction with the object (such as zoom and rotation), which do not need to be considered in processed image or video quality assessment. Therefore, in this thesis, methods to investigate the quality perceived by human observers are proposed both for segmented video objects and 3D objects.

In order to design objective quality assessment methods whose predictions are in agreement with human perception, two different approaches can be followed. Both approaches are based on the Human Visual System (HVS) since all visual perception tasks have the HVS in common. The first approach tries to model the HVS. If the visual system model is accurate, the model can be used for designing reliable objective metrics. However, the human visual system is extremely complex, and many of its properties are not entirely understood till today. Thus, another approach is used in this thesis to develop perceptual quality assessment metrics. *A priori* knowledge about the signal processing and analysis methods as well as the pertinent types of introduced distortions is used. In order to investigate the perceptual impact of different kinds of distortions a series of psychophysical experiments is performed in this thesis. On the basis of these experiments, the perceptual objective metrics are derived. Although such metrics are not as versatile, they normally perform well in a given application area. Their main advantage lies in the fact that they often allow a computationally more efficient implementation than approaches which directly model the HVS.

To the best of our knowledge, a comparison among different objective methods for quality assessment in the addressed visual information areas has received little attention by the image processing community so far, as well as the study of their performances on real processing algorithms. In this thesis, the performances of the proposed objective metrics on several processing algorithms, for the different kinds of data addressed, are evaluated. A comparison with the state of the art objective methods used in the literature is carried out.

1.3 Organization of the Thesis

The remainder of this dissertation is divided in three parts according to the kind of visual information processing system it deals with. Image and video processing quality assessment is discussed in **Part I**. Background knowledge related to this research is reviewed in *Chapter 2*. First, standard techniques for image and video subjective quality assessment are presented. Then, the approaches to derive objective quality assessment methods in agreement with human perception are discussed. *Chapter 3* focuses on a particular type of video processing, the watermarking. Different watermarking techniques are applied to a set of video sequences and subjective experiments are carried out to investigate the perceptual impact of distortions due to the insertion of the watermark. Then, two new quality assessment objective metrics are proposed and discussed in terms of their correlation with subjective data with respect to a state of the art simple metric.

In **Part II**, quality assessment of segmented video objects is addressed. *Chapter 4* describes the framework for segmentation quality assessment. First, the state of the art methods for image and video subjective evaluation are reviewed. Then, the proposed experimental method for subjective tests, the instructions, the experimental tasks and the synthetically generated segmentation errors are described. Finally, the novelty of the proposed approach for segmentation evaluation that consists in deriving a *perceptual objective metric* from subjective experiments is explained. *Chapter 5* introduces the new perceptual metric proposed for quality assessment of segmented video sequences. First, the objective metrics found in literature for segmentation evaluation are described. Then, a new objective metric is proposed and discussed. Finally, the results of the subjective experiments are presented, the overall perceptual objective metric is derived and its performance, also compared to the state of the art methods, is described. In *Chapter 6*, the proposed metric is tested on video sequences with real artifacts produced by several segmentation algorithms. A general framework is then considered to discuss the correlation between the subjective and objective results. Furthermore, some of the most common applications of video object segmentation are illustrated and subjective experiments are proposed for different applications. Finally, according to the particular application, the parameters of the proposed metric are tuned on the basis of subjective results and then the performance compared to the state of the art objective metrics. *Appendix A* shows the errors synthetically introduced for video object segmentation quality assessment. *Appendix B* reports the scripts for all the subjective experiments carried out in this work.

In **Part III**, methods for quality assessment of watermarked 3D objects are investigated. *Chapter 7* reviews previous works related to perceptual image watermark insertion, to mesh simplification and perceptually-guided rendering. Distortions introduced by common 3D watermarking algorithms are described. On the basis of subjective experiments two new metrics based on the roughness estimation of the model's surface are proposed. The performance of the introduced metrics is compared to two of the most common geometric metrics used to measure the similarity between two 3D objects.

In order to provide an overview on all the possible viewing conditions for 3D model quality assessment, *Appendix C* describes the background on 3D rendering conditions.

1.4 Main Contributions

The significant contributions of the work presented in **Part I** are summarized below:

- **design** and development of two new objective metrics for the quality assessment of watermarked video. Flicker and noise effects are identified by means of subjective experiments and the two objective metrics are proposed and compared to a simple state of the art metric.

The main contributions of **Part II** are:

- **an extensive survey** of literature methods both for subjective and objective quality assessment. Both image and video segmentation quality assessment methodologies are presented and their advantages and disadvantaged are discussed;
- **realization** of several psychophysical experiments to study and characterize the distortions introduced by segmentation algorithms on the basis of a proposed methodology for subjective experiments;
- **design** and development of a new perceptual objective metric. In order to assess the quality of segmented video objects, a metric based on the perception of errors in the segmentations is proposed;
- **application** of the proposed metric to the evaluation of real segmentations in different applications, such as video compression, video manipulation and video surveillance.

Part III dedicated to 3D watermarking quality assessment is a jointly work with Dr. Massimiliano Corsini [25]. The main contributions are:

- **realization** of two psychophysical experiments on the basis of a proposed methodology for collecting subjective data for 3D model quality assessment. The first one is for designing two perceptual objective metrics and the second one is for validating the proposed metrics with different watermarking algorithms;
- **design** and development of two new perceptual objective metrics for watermarked 3D object quality assessment on the basis of the roughness surface estimation. By means of subjective experiments the perceived distortion amount is included to derive these two perceptual metrics.

Part I

Image and Video Processing Quality Assessment

Objective and Subjective Quality

2

2.1 Introduction

Digital data are subject to a wide variety of distortions during acquisition, analysis, processing, compression, storage, transmission and reproduction, any of which may cause a degradation of the *visual quality*. Visual quality plays an important role in various applications [163]. But what do we mean by *quality*?

Every person has a notion of quality that may depend on the context. It is difficult to find a general definition of quality applicable in all contexts. A definition of *quality* can be found in a dictionary [98] that is:

> *Quality, (i.e. the degree of excellence which a thing possesses), refers to a characteristic (physical or nonphysical, individual or typical) that constitutes the basic nature of a thing or is one of its distinguished features.*

As an example, in the context of videoconferencing applications, the term 'quality' typically refers to Quality of Service (QoS) [158] which is described as "good picture quality, good sound, etc...,". The biggest difficulty lies in the fact that there is no quantification or evaluation of what *good* is.

For this reason, when evaluating the visual quality of data a large panel of human observers is needed to produce a Mean Opinion Score (MOS). Standard procedures [65, 66] for psychophysical* experiments have been established as valuable research tools in the image and video processing field for a better understanding of how humans judge quality and the perceived distortions. In order to develop the subjective quality evaluation methods proposed in the following chapters for new kinds of processing systems, the techniques available today for compressed image and video quality assessment [56, 65, 66] are reviewed in this chapter.

As the nature of image/video processing used for compression is different from that for image/video segmentation (or the way the visual inspection of a 3D object is performed is different from that of an image/video) the subjective methods for quality assessment will also be different from those of the more investigated systems, such as video compression systems [48, 102, 175].

---

*Psychophysics is the branch of psychometric (the study of human response to various stimuli) which deals with stimuli that can be expressed in terms of physical parameters.

Once the procedure for subjective experiments is established, the second step is to consider the subjective ratings/responses obtained from psychophysical experiments as benchmarks for the development of objective quality metrics. In fact, the goal of this research is to develop objective measures that can automatically predict the quality perceived by human observers. In order to assess how distortions are perceived by humans and to build a *perceptual objective metric*, subjective experiments have to be carried out. In such a way, a relation between the perceptual quality of image/video sequences and the objective measure of the distortions can be derived. When this relation, the perceptual objective metric, is found, further subjective experiments which are an expensive and time consuming practice to evaluate the visual quality can be avoided.

A great effort has been made in recent years to develop objective quality metrics that correlate with subjective quality measurements in image and video processing systems [40, 45, 63, 105, 152, 156, 161, 164, 168]. The knowledge developed in such a field is taken into account in order to propose objective metrics for the different kinds of visual information processing systems considered in this thesis.

In general, in order to design reliable perceptual objective quality assessment metrics that mimic the subjective responses, it is necessary to investigate:

- the methodology of subjective experiments,
- the significances of subjective data by means of statistical analysis,
- how to model the processed subjective data using fitting functions,
- the models to simulate the Human Visual System (HVS) responses in objective metrics, and
- the performance of the objective methods by their correlation with subjective scores.

These important issues for the quality assessment of any type of visual information are tackled in this chapter. Section 2.2 discusses the subjective quality assessment procedures, the standard methods and the grading scales for carrying out subjective tests, and investigates their requirements and limitations. Section 2.3 presents how to perform the statistical analysis of subjective data according to ITU Recommendations [65, 66], the psychometric functions usually adopted to fit the processed data and the combination rule used in the literature [30] to build quality assessment models. Section 2.4 describes how to develop and evaluate objective quality assessment models based on HVS. Section 2.5 draws the conclusions.

2.2 Subjective Quality Assessment

The benchmark for any kind of visual quality assessment are subjective experiments, where a number of subjects are asked to watch the test images or video and to rate their quality. As already mentioned, quality is rather a nebulous concept. For this reason, when evaluating the quality of visual data, a good compromise is to provide a term of comparison. The comparison may be explicit when the two data are observed side by side and a choice of which one is better (or worse) can be made. The comparison can also be implicit; in this case the judgment will depend on an internal reference. In order to remove this internal reference, the subjects are usually shown with explicit reference. Moreover, there can always be data with higher or lower quality. Hence, the judgment scale should not be limited, giving room for better or worse quality responses. Subjective tests may measure *impairment scores* rather than quality scores; or they can be asked to rate the degree of distortion, the amount of defects or the strength of artifacts (see definitions in Tab. 2.1).

Impairment (or *defect*) is a subjective measure of the degradation of visual information quality. In this case, the reference is assumed to have no impairments and any difference from the reference represents a decreasing quality or an increasing impairment. Subjects are asked to rate the image/video in terms of impairment. Sometimes distortions can be clearly visible but not so objectionable. Thus, subjects may be asked to rate the video in terms of *annoyance*, that is how much the impairment bothers the viewer.

Subjective experiments are carried out to provide the Mean Opinion Score (MOS) computed by averaging all the gathered subjective measurements. The subjective measurement of quality, impairment or annoyance may be carried out with different methodologies. The standard subjective assessment methods [65, 66] for image/video quality assessment are described in Sec. 2.2.1. The different judgment scales adopted in these subjective experiments are presented in Sec. 2.2.2. Further requirements for subjective evaluation and some limits are presented in Sec. 2.2.3.

Table 2.1: Definitions used in developing quality assessment systems.

Term	**Definition**
Distortion	Any measurable change in the form of the original signal during capture, processing, transmission, display.
Impairment or *defect*	An error or measure of the degradation of the signal.
Perceived defect	Perceptual change due to a defect (perceptual dimension).
Artifact	Relatively pure perceptual feature of an impairment.

2.2.1 Subjective Measurement Methods

Standardized subjective quality assessment techniques are well established in the specific frame of coded images and video sequences. ITU-T Recommendation P.910 [65] and ITU-R Recommendation BT.500-6 [66] are the commonly used standards for the subjective assessment of still images and video. As in this thesis other kinds of visual information will be taken into account, such as video objects and 3D models, the proposed methodology for each case will be based on methods and considerations addressed in ITU-R and ITU-T Recommendations listed below:

- Double Stimulus Continuous Quality Scale (DSCQS). This method involves two images or sequences, one of which is a reference. The images or sequences are shown alternatively in random order. The subjects are not told which one is the reference and are asked to rate each picture independently. They rate each of the two separately on a continuous quality scale ranging from “bad” to “excellent”.

- Double Stimulus Impairment Scale Method (DSIS). This method is intended for images or sequences which cover a wide range of impairments. Subjects are shown the original picture followed by the impaired picture, and are asked to rate the quality of the impaired picture with respect to the original. Results are indicated on a discrete five-grade impairment scale: imperceptible, perceptible but not annoying, slightly annoying, annoying, and very annoying.

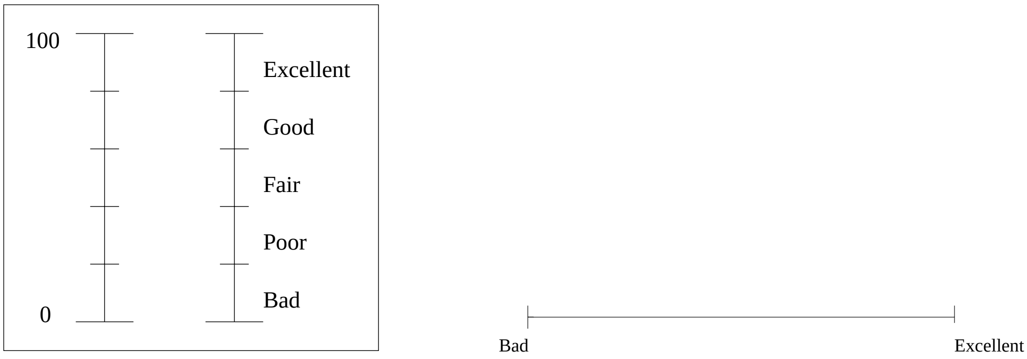
- Stimulus Comparison Method. In this method, subjects are shown the two distorted scenes in a random order and asked to rate the quality of one scene with respect to the other by using a discrete seven level scale (much worse, worse, slightly worse, the same, slightly better, better, and much better). Continuous scales may also be used.

- Single Stimulus Method. This method does not use a reference. Subjects are shown the test image or sequence and are asked to rate its quality. A number of different ways of recording observer results are possible (continuous or discrete scales) and are described in details in Sec. 2.2.2.

These four methods have generally different applications. The choice of a specific method depends on the context, the purpose and where, in the development process of the data the test is to be performed. DSCQS is the preferred method when the quality of the test and the reference sequence are similar, because it is quite sensitive to small differences in quality. The DSIS method is better suited for evaluating clearly visible impairments, such as artifacts caused by transmission errors. Single Stimulus Method is useful when the effect of one or more factors need to be assessed. The factors can either be tested separately or can be combined to test interactions. The Single Stimulus Continuous Quality Evaluation (SSCQE) method relates well to the time varying quality of today's digital video systems. The Stimulus Comparison Method is useful when two impaired images or sequences are required to be compared directly. This is the case for example when different image or video processing systems are compared on the basis of the visual quality of their results.

2.2.2 Grading Scales

Grading scales can be *continuous* or *discrete*, *categorical* or *numerical* [66]. Discrete five-grade impairment scale (5 imperceptible, 4 perceptible but not annoying, 3 slightly annoying, 2 annoying, 1 very annoying) or the discrete five-grade quality scale (5 excellent, 4 good, 3 fair, 2 poor, 1 bad) are usually applied in the DSIS method, according to ITU-T Recommendations.



(a) (b)

Figure 2.1: Continuous grading scales: (a) five point continuous scale, and (b) categorical continuous scale.

Continuous rating system can be adopted to avoid quantization errors. The continuous scales used in the DSCQS method are divided into five equal lengths which correspond to the normal ITU-R five point quality scale. Figure 2.1 (a) shows a typical score sheet. The pairs of assessment (reference and test) for each test condition are converted from measurements of length on the score sheet to normalized score in the range of 0 to 100.

Both *categorical* scales and *numerical* scales are usually adopted in the Single Stimulus Method. Categorical scales assess image quality and image impairment with the five-grade impairment scale or the five-grade quality scale cited above. Numerical scales are also used in Single Stimulus procedures using a discrete 11-grade numerical scale (from 0 to 10). In continuous scaling, a variant of the categorical method is that of assigning each image or image sequences to a point on a line between two semantic labels (e.g. “bad” and “excellent” as in Fig. 2.1 (b)). In numerical scaling there is the option to use a restricted scale (e.g. 0-100) and the variant with an open scale. Sometimes, the value describes the judged level in “absolute” terms (without direct reference to the level of any other image or image sequence). In other cases, the number describes the judged level relative to that of a previously defined “*ground-truth*” (*reference* or gold standard).

2.2.3 Experiment Requirements and Limitations

The Recommendation BT.500 standard [66] specifies several other features to be considered in the testing. They are listed below. Besides them, privacy issues should also be addressed when carrying out subjective experiments.

- Number of subjects: at least 30, and preferably more. They should have normal or corrected-to-normal vision, and should preferably be non-expert.
- Test scenes: these should be critical to the impairment being tested.
- Test session: the experimental part should take no longer than 30 minutes due to the possibility of fatigue when evaluating images. In total, with the instruction part the test should not be longer than 1 hour.
- Viewing conditions: specifications have been established for the room environment, ambient lighting conditions and viewing distance.
- Stimulus presentation: random presentation of image or sequences is recommended.

Subjective testing is currently the accepted method for establishing the quality of a particular processing algorithm. However, there are several difficulties associated with performing subjective quality tests [112, 162]. These intrinsic difficulties have driven the research on objective quality assessment. Some of these limitations are listed below:

- Most psychophysical experiments are conducted on simple patterns. But it is not known if a limited number of simple-stimulus experiments are sufficient to build a model that can predict the visual quality of complex structured natural images.
- Interactive visual processing (e.g. eye movement) influences the perceived quality. For example, subjects will give different quality scores if they are provided with different instructions. Prior information regarding the image content, or attention and fixation, may also affect the evaluation of the image or image sequence quality.
- Subjective tests are extremely time consuming and costly. Many groups do not possess the required equipment and have to conduct tests under non-standard conditions or in other laboratories. It is also difficult to obtain a large number of subjects. The process of subjective testing may take weeks or months, thus becoming a big limitation in the research of subjective quality assessment.
- A large number of subjects is required since there may be a large variation in individual viewing opinions, depending on the subject’s age, sex, motivation and other personal factors.

- Subjective quality may vary depending on the length of the representation. Hamberg and de Ridder [58] found that subjects take around 1 second to react to a particular distortion in a scene and a further 2-3 seconds to stabilize their responses. Horita *et al.* [63] showed that distortions in the first or in the last part of the representation jeopardize the overall quality more than those appearing in the central part.

- The scale used by subjects can also introduce problems. For example, discrete scale with few levels asks for many subjects to reduce the variance. Subjects are usually reluctant to give very high or very low scores. For this reason open-ended scales may be used.

Objective methods which use subjective data as ground truth to assess the accuracy of their technique should take into account that their results are not perfect. Thus, perfect correlation with subjective quality data may not be realistically reachable.

2.3 Subjective Data Analysis

Once the experimenter has performed the subjective experiment with an appropriate tool, chosen among those described in the previous section, and performed the test by taking into account all the requirements, he/she has finally gathered a collection of subjective data to be statistically analyzed. At this point, the experimenter has to determine to which point the data can be trusted and determine the confidence interval. Then, the performance of individual subject data has to be investigated and statistical analysis can be applied to screen the outliers with standard methods [66] as presented in Sec. 2.3.1.

After being analyzed, the data provided by the test subjects, have to be modeled by standard functions (see Sec. 2.3.2) able to describe and to fit the mean opinion scores. Moreover, the goodness of the fit has to be considered. In such a way the human perception of quality (or annoyance) can be predicted in terms of psychometric functions.

Finally, since one or more artifacts can be present at the same time, a combination rule has to be defined to obtain the overall quality (or annoyance). This rule in psychophysical experiments is often based on the Minkowksi metric [30, 41, 89, 129, 162] which is described in Sec. 2.3.3

2.3.1 Processing the Subjective Data

In the first step, the subjective score values are combined into a single overall score for each test scene using the sample mean. The subjective Mean Opinion Score (MOS) for the test scene j is given by:

$$\mu_j = \frac{1}{N_j} \sum_{i=1}^{N_j} m_{ij}, \tag{2.1}$$

where N_j is the number of test subjects in the experiment. This measure represents the averaged subjective scores m_{ij} obtained from subject i after viewing the test scene j. By presenting the results of a subjective test, all mean scores μ_j must be associated to a confidence interval which is derived from the standard deviation and the size of each sample [92, 142]. Confidence intervals for the mean subjective scores are usually calculated using Student's t-distribution. The t-distribution is appropriate when only a small number of sample is available [142]. The sample standard deviation s_j is calculated for each sequence j using μ_j and the confidence intervals are computed as:

$$\delta_j = t_{(\alpha/2, N_j-1)} s_j / \sqrt{N_j}, \tag{2.2}$$

where $t_{(\alpha/2,N_j-1)}$ is the t-value associated with the desired significance level α for a two-sided test with $N_j - 1$ degrees of freedom. As the number of observations N_j increases, the confidence interval decreases. The final results are the j Mean Opinion Score values with an associated $1-\alpha$ confidence interval, $\mu_j \pm \delta_j$.

The second step in the analysis refers to the *normalization* to compensate for the *boundary effects* of the voting scale on the subjective scores. This effect has been identified as the one in which observers tend not to use the extreme values of the judgment scale, in particular for high quality scores. This may arise from a number of factors, including psychological reluctance to make extreme judgments [66, 102].

Finally, the MOS values and sample standard deviations are used in the *subject screening procedure*. In the screening procedure usually adopted (Annex 2 of ITU BT.500 Recommendation [66]) an expected range of values is calculated for each model. A subject is not rejected for always being above the expected range or always being below the expected range but for being erratic on both sides of the expected range. This procedure is appropriate to reduce the variability of the data in small sample sets. If necessary, after the screening procedures, the MOS and their relative confidence intervals are recalculated without the data from rejected test subjects.

2.3.2 Psychometric and Fitting Functions

A *psychometric function* describes the relation between the physical intensity of a stimulus and the subject's ability to detect or respond correctly to it. The measurements are based on a number of discrete trials at a number of different stimulus intensities. The psychometric function usually increases monotonically with stimulus intensity. Sigmoidal functions such as the **logistic** function, normal cumulative functions such as **Gaussian** and **Weibull** functions are commonly fitted to the data by a non-linear least-squares method. All these psychometric curves are cumulative distribution functions with a range [0,1].

These psychometric functions are in common use for describing the relation between some physical measure of a stimulus and the probability of a particular psychophysical response. But they can be adapted and used for fitting MOS on quality or annoyance. Scaling $(M - m)$ and offset parameters, m (where m is the minimum on the grading scale and M is the maximum) can adjust the psychometric functions to an arbitrary scale of quality or annoyance as follows.

The **logistic** function is symmetric and becomes

$$logistic(a, b, x) = m - \frac{M - m}{1 + e^{-(a+bx)}}, \tag{2.3}$$

where the parameters a and b control the curve location and steepness, respectively, and x is the objective measurement of the stimulus intensity.

The **Weibull** function is not symmetric and its shape depends on two parameters, S and k,

$$Weibull(S, k, x) = m + (M - m) \cdot \left(1 - e^{-(S\cdot x)^k}\right). \tag{2.4}$$

The **Gaussian** psychometric function is the normalized cumulative distribution function for a Gaussian distribution, that is

$$Gaussian(a, b, x) = m + (M - m) \cdot \sqrt{2\pi} \cdot \int_{-a+bx}^{\infty} e^{(-\frac{t^2}{2})} dt, \tag{2.5}$$

where a shifts the curve along the x-axis and b controls the steepness of the function.

One of these functions is commonly fitted to the data by a least-squares method and it is assumed that the correct form of the fitting curve is to be known and that the data are uncorrelated and representative [92]. Moreover, the x-values are assumed to be known without uncertainty.

In the case that the mean opinion value's variance is not the same for all data points, a weighted non-linear least-squares method can be used. Weights are calculated for each mean opinion value such that they are inversely proportional to their standard deviation and multiplied by a normalization constant chosen so that the sum of all weights is equal to one [56].

After the best parameters of the fitting curves are found in a least squares sense, the 95% confidence bounds on the curve fits are estimated. After all these parameters have been estimated (the best fitting curve, the parameters that give the best fit and the confidence bounds of the curve), it is necessary to measure how good the fit is. Standard correlation coefficients used for this purpose will be described in Sec. 2.4.2. Another way to represent the fit is to square the correlation coefficient, also called the *goodness of fit*, r [61]:

$$r = \frac{(std(\mathbf{y}))^2 - (std(\mathbf{x} - \mathbf{y}))^2}{(std(\mathbf{y}))^2}, \tag{2.6}$$

where $\mathbf{x}$ and $\mathbf{y}$ are respectively the vectors of the values predicted by the model and the observed ones, and std is the standard deviation. This correlation coefficient is also used in this thesis to estimate the goodness of fit.

2.3.3 Minkowski Metric

In both subjective and objective image/video quality models, the relationship between individual artifacts and the produced overall annoyance is often estimated by the Minkowski metric [89, 105, 162]. If an image/video is affected by one or more types of artifacts, the total impairment can be estimated by knowing the individual artifact amounts and their individual perceptual contributions to the produced overall annoyance.

The perceptual image quality Q (usually it is defined in a interval [0,1] for convenience) is linearly related to the perceptual impairment [129] I by:

$$I = 1 - Q. \tag{2.7}$$

Besides judging the total perceptual impairment, subjects can also distinguish among the various underlying artifacts and are able to judge the perceptual impairments of these underlying artifacts separately [30].

Minkowski metrics have been used as a combination rule for different artifacts in psychophysical experiments [48, 89, 102, 105, 162]. In the Minkowski metric, the perceptual impairments are combined into the total impairment I by means of:

$$I^p = \sum_{i=1}^{M} I_i^p, \tag{2.8}$$

where M is the number of the underlying artifacts, p is the Minkowski parameter.

The use of the Minkowski parameter p has its roots in multidimensional scaling [88] where Minkowski metric is used as a distance measure. A distance interpretation of the Minkowski metric formalism is not the only possibility. It can be generalized by suggesting that observers take some form of average when evaluating image quality. The application of this mathematical formalism has been successful in both the image/video impairment and quality approaches.

The perceptual impairment of the underlying artifacts is found to be linearly related to the perceptual strength S of these artifacts [105]:

$$I_i = a_i \cdot S_i. \tag{2.9}$$

The interpretation is that the strengths of the perceptual attributes are attributes of low level vision, whereas the impairments are cognitive attributes. The constants a_i represent the relative weights of the artifacts in the total perceptual impairment. On the basis of Eq. (2.7), the difference in the perceptual quality between an original image/video and an impaired version equals the total impairments caused by the underlying artifacts. This difference can be rewritten as:

$$I = (\sum_{i=1}^{M} w_i \cdot S_i^p)^{\frac{1}{p}}, \tag{2.10}$$

where w_i are the reformulated perceptual weights. The interpretation of the error measure as a distance measure in perceptual space is straightforward. If the artifacts are visible, the resulting loss in perceptual quality is proportional to the p-norm of a vector in a perceptual multidimensional space (dimension given by M) spanned by these artifacts.

Different Minkowski exponents p have been found to yield good results for different experiments and implementations in the literature [30]. As an example, for subjective experiments with coding artifacts, $p = 2$ (Euclidean metric) was found to give good results. Intuitively, a few high distortions may draw the viewer's attention more than many lower ones. This behavior has been observed and generalized with higher exponents ($p = 4$) in proposed combinations for large impairments in [30].

The Minkowski metric will be used to analyze the data from subjective experiments where subjects were asked individual artifact annoyance and overall annoyance. In such a way, the perceptual weights will be derived in order to build the subjective quality model. The same rule will be used to combine the individual objective measures and the overall annoyance measured in subjective experiments in order to derive the perceptual weights in the objective metric and build the objective quality model.

2.4 Objective Quality Assessment

Images, video sequences and 3D objects are ultimately viewed by humans, thus the only "correct" method of quantifying visual quality is through subjective evaluation. However, subjective evaluation is usually too inconvenient, time-consuming and expensive. The goal of objective quality assessment is to find a way to predict what people will say about the image/video quality without performing any subjective test. In order to reach this goal, the first step is to investigate what method people use to judge the quality and whether they agree on the judgment. For this reason, in the previous sections, we have presented how a subjective experiment has to be run, which are the different experimental conditions that should be followed and how the gathered data should be processed. The second step is to model the way humans perceive image/video distortions in order to develop quantitative measures that can automatically predict perceived image/video quality. This section addresses the issues involved in designing *objective quality assessment* methods.

The development of an objective quality metric is important since it can be used in different stages of image/video processing and analysis systems as follows.

- *Monitoring*. It can be used to monitor and adjust on-line the image/video quality.
- *Optimization*. It can be used off-line to optimize algorithms and parameter settings of image/video processing and analysis systems [35].
- *Benchmarking*. It can be used to compare and rank different image/video processing systems and algorithms.

The objective metrics developed in this thesis will be used to benchmark different processing algorithms and will be compared to the state of the art objective metrics currently in use.

Objective quality metrics can be classified according to the availability of the original (impairment-free) image/video with which the impaired version has to be compared. In the approach called *full reference quality assessment*, the complete reference image/video (ground truth) is assumed to be known. If the reference is not available a *no reference* or “blind” quality assessment is desirable. In a third type of method, the reference is only partially available in the form of a set of extracted features as side information to help to evaluate the quality of the visual data. This is referred to as *reduced reference* quality assessment. This thesis focuses on *full reference* quality assessment.

The objective model should mimic the human visual and perceptual system, so that the measured quality agrees with the subjective quality as perceived by the viewer. Section 2.4.1 describes the models that can be used to build an objective metric based on the HVS. Moreover, the most common objective metrics used in image/video quality assessment are mentioned in this section. Section 2.4.2 presents the attributes that characterize an objective quality metric in terms of its prediction performance with respect to subjective scores.

2.4.1 Objective Models and Metrics

Several studies about the HVS behavior have been carried out, especially with respect to video quality assessment metric development. The objective metric for video quality evaluation can be very application dependent (e.g. regarding the relevant artifacts), and thus a variety of HVS models need to be considered. Two main families of models have been identified by Lubin [84]: *performance modeling* and *mechanistic modeling*. Both models aim at modeling psychophysical quantities but with different approaches:

- *Performance modeling*. These models do not try to directly simulate the human visual functions, but instead provide input-output functions that have a behavior comparable to that of the visual system [145, 169, 176]. They usually involve a lower computational complexity than perceptual models and generate a numerical value for the quality evaluation.

- *Mechanistic modeling*. These models describe the HVS and its functional behavior using the available knowledge about the processing of visual information by a human viewer [85, 174]. Examples of functions used to model this behavior are: contrast sensitivity, masking and luminance adaptation. The output of the perceptual models can take the form of a spatial map of just noticeable differences, or the values in the map can be combined to produce a single value.

Until now, both the mechanistic and performance modeling have produced objective measures of the perceived quality in compressed or watermarked images or video [85, 145, 169, 174, 176]. Some guidelines for developing quality metrics for the different kinds of visual information addressed in this thesis can be extracted from these models. Furthermore, some of the objective metrics proposed in the literature may serve as a basis for developing new ones, more appropriate for 3D models or video objects quality assessment.

There are various quality metrics in the literature to objectively evaluate the quality of an image or video. A simple classification groups the objective metrics into *mathematical metrics* and *perceptual metrics*. Mathematical metrics measure quality in terms of relative simple mathematical functions, usually with pixel-by-pixel weighted differences between the reference R and the distorted image/video D. Peak Signal to Noise Ratio ($PSNR$) and Mean Squared Error (MSE) are the most widely used mathematical metrics which are defined as:

$$MSE = \frac{1}{I}\sum_{i=1}^{I}(R_i - D_i)^2, \tag{2.11}$$

$$PSNR = 10 \cdot log_{10}\frac{R_M^2}{MSE}, \tag{2.12}$$

where I is the total number of pixels in the image and R_M is the maximum possible reference intensity value. These metrics are the most common but are widely criticized as well, for not correlating with the perceived quality measurement. Perceptual metrics are used in mechanistic modeling incorporating HVS characteristics [161] such as luminance contrast sensitivity (Weber's law), frequency contrast sensitivity (Contrast Sensitivity Function) and masking effects [112, 143]. The Just-Noticeable-Difference (JND) is a very important concept in objective metrics using HVS features and ideally it provides each signal being represented with a threshold level of error visibility, below which errors are imperceptible. The JND is adopted in the Sarnoff Visual Discrimination Model proposed by Lubin [85]

In this thesis, the performance modeling approach will be used. The reason why we chose the aforesaid approach for the segmentation quality evaluation for example, is that while in image compression systems the output and the input should be similar as much as possible, in segmentation systems they are quite different. A segmentation system has a restructuring function with the output being organized differently than the input: a pixel matrix is transformed in a label matrix. While in the case of compression the degradation in the texture quality is evaluated, in the second case no degradation is involved. Segmentation quality is more related to the structural matching between the real-world objects and the segmented objects. Therefore, low-level HVS characteristics do not need to be incorporated into the quality assessment model.

2.4.2 Objective Model Assessment

The goal of objective quality assessment is to design algorithms whose quality prediction is in good agreement with subjective scores from human observers. There are different attributes that characterize an objective quality model in terms of its prediction performance with respect to Mean Opinion Score, MOS [56]. Two of these attributes are accuracy and monotonicity which are defined as follows.

- *Accuracy* is the ability of a metric to predict subjective ratings with minimum average error and can be determined by means of the *Pearson* linear correlation coefficient. For a set of N data pairs (x_i, y_i), it is defined as follows:

  $$Pearson = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}}, \tag{2.13}$$

  where $\overline{x}$ and $\overline{y}$ are the means of the respectively objective and subjective data.

  This assumes a linear relation between the data sets, which may not be the case. Therefore, in this thesis psychometric functions will be used to take into account the HVS behavior such as saturation effects. Then, linear correlation will be used to obtain relative comparisons between subjective and objective data.

- *Monotonicity* measures if increases (decreases) in one variable are associated with increases (decreases) in the other variable, independently of the magnitude of the increase (decrease).

Ideally, differences of a metric's rating between two sequences should always have the same sign as the differences between the corresponding subjective ratings. The degree of monotonicity can be quantified by the *Spearman* rank-order correlation coefficient, which is defined as:

$$Spearman = \frac{\sum_{i=1}^{N}(\chi_i - \overline{\chi})(\gamma_i - \overline{\gamma})}{\sqrt{\sum_{i=1}^{N}(\chi_i - \overline{\chi})^2}\sqrt{\sum_{i=1}^{N}(\gamma_i - \overline{\gamma})^2}}, \tag{2.14}$$

where χ_i is the rank of x_i and γ is the rank of y_i in the ordered data series. $\overline{\chi}$ and $\overline{\gamma}$ are the respective mid-ranks. The Spearman rank-order correlation makes no assumption about the relationship between x_i and y_i.

Both correlation coefficients will be used in this thesis to investigate the performance of the proposed objective metrics.

2.5 Conclusions

In this chapter, standard methods for carrying out the subjective experiments, statistical methods for data analysis and objective models and metrics proposed in image/video compression quality assessment have been presented. Even if other processing systems different from compression will be considered in the next chapters, some of the methods and considerations remain valid in 3D or object segmentation quality assessment.

Subjective rating of annoyance of visual impairments is sometimes used instead of quality rating which is rather a nebulous concept. Unrestricted grading scales are used to leave always room to human judgments for something better or something worse. Statistical analysis of subjective results, normalization and screening procedures are performed according to standard procedures.

Standard psychometric functions used for fitting the subjective data have been described. The Minkowski metric usually adopted in psychometric experiments to combine different artifacts in the overall impairment model has been presented. Performance modeling can be used to build objective quality assessment methods that try to indirectly simulate the HVS. The Spearman and Pearson correlation coefficients are applied to test the performance of the proposed objective quality assessment methods versus subjective ratings.

These tools will be used to propose subjective experiment methodologies and objective metrics for the visual information tackled in this thesis.

3

Video Watermarking Quality Assessment

3.1 Introduction

In the previous chapter, we presented standard methods to carry out subjective tests and to derive objective metrics for evaluating the quality of processed images and video. In this chapter, as a first contribution, we present a subjective and objective study of quality on a particular type of digital processing: *video watermarking*.

The rapid spread of digital media (audio, images, video and 3D models) and the ease of their reproduction and distribution has created a need for copyright enforcement schemes in order to protect content creators and owners. In recent years, *digital watermarking* has emerged as an effective way to prevent users from violating copyrights. More precisely digital watermarking regards the embedding of a digital information, called *watermark* in a host multimedia content, such as image, video or even 3D model. A general scheme of a watermarking system is depicted in Fig. 3.1. In brief, a secret *key* is used to embed and extract the *watermark* inserted in the *multimedia content*. *Embedding* and *recovering* are the procedures that enable the insertion and extraction of the watermark. The basic idea of watermarking is to associate the digital information to the digital host in imperceptible way to the human eye yet resistant to (intentional or unintentional) alterations of the watermarking. This association can be used for several applications. For example, in the case of copyright protection, the watermark conveys information about the owner/creator of the digital multimedia content, and it can be used to proof ownership in legal dispute.

The assessment and the comparison of the performance of watermarking procedures are complex since various aspects have to be evaluated as described in the next section. The *visibility* of the watermark is an important aspect in this process.

In this chapter, we propose two new metrics for evaluating the visual impact of video watermarks. Based on several different watermarking algorithms and video sequences, the most prominent impairments are identified as *spatial noise* and *temporal flicker*. We design the corresponding measurement algorithms and corroborate their performance through subjective experiments.

The chapter is organized as follows. In Sec. 3.2 we describe the motivations of digital watermarking and present the state of the art benchmarks for watermarking algorithms. Section 3.3

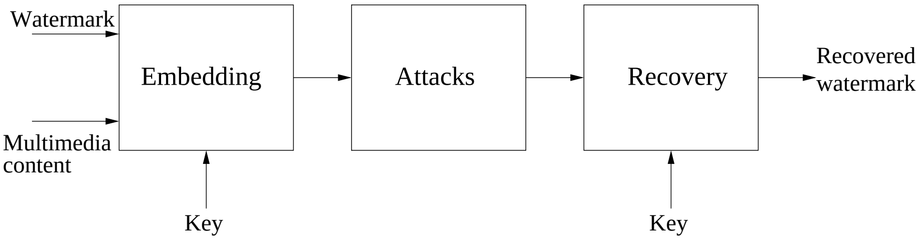


Figure 3.1: General scheme of watermarking system.

outlines the proposed approach to evaluate the visual distortion given by watermarking algorithms. In Sec. 3.4 the watermarking algorithms used in the experiment are described along with the artifacts caused by their insertion. Section 3.5 presents the video sequences used to test the proposed metrics as well as the design of the subjective experiments. Section 3.6 describes the new metrics proposed in this thesis to assess the watermarked video quality. The results are reported and discussed in Sec. 3.7. Section 3.8 presents some conclusions.

3.2 Video Watermarking - Motivation and Context

Video watermarking is mainly utilized in three frameworks: copyright control, broadcast monitoring and more recently quality assessment [16].

As far as copyright protection for DVD is concerned, the watermark can be used to attach information to the content about the restrictions on copying. For example, if a DVD player detects a "never copy" statement on a recorded disk (which can be distinguished from an original), it can be designed to reject it.

Broadcast monitoring systems include video watermarking embedder and detectors that add identifying data to the video signal prior to transmission by terrestrial, cable and satellite broadcasters. Detectors have access to the right key and can detect the digital identifier from the video signal. These identifiers are imperceptible to television audiences and survive all common video processing operations. Broadcasts can then be monitored using these detectors to verify that program and advertisement transmission comply with contractual requirements and do not occur without the permission of the broadcast owner.

Not all watermarking technologies are aimed at robust proof of copyright ownership. Some are intended for adding value (like embedding lyrics in a song). Some are not intended to be robust, but fragile, for data authentication or quality assessment in compression and transmission techniques. The latter has recently been proposed in literature [16] and it consists in making use of a data hiding techniques to embed a fragile mark into perceptually important areas of the video frame. At the receiver, the mark is extracted from the perceptually important areas of the decoded video. In such a way, a quality measure of the video is obtained by computing the degradation of the extracted mark.

In any of the above mentioned applications, the embedded watermark must be imperceptible enough to remain invisible during normal viewing. On the other hand, the more information we want to add to the digital host, the more difficult it becomes to hide the information without introducing a relevant distortion. The best developed watermarking algorithms adapt to the video content in order to achieve the best possible compromise. However, the predominant requirement is that watermarks must work *without bothering* the user.

Besides *visibility* (i.e. how easily the watermark can be discerned by the user) two other factors must be considered in watermarking:

- *capacity*, i.e. the amount of information that the watermark can convey and be recovered without errors;

- *robustness*, i.e. the resistance of the watermark to alterations of the original content such as compression, filtering or cropping.

These three factors are inter-dependent; for example, increasing the capacity will decrease the robustness and/or increase the visibility. Therefore, it is essential to consider all three factors for a fair evaluation or comparison of watermarking algorithms. Various consortia such as Certimark (http://www.certimark.org) or the Content ID Forum (http://www.cidf.org) have been working on the definition of procedures for such evaluations.

In particular, a series of benchmarks has been studied and developed to test the robustness of watermarking algorithms. One of the first benchmarks for image watermarking has been the StirMark [119] package, developed by Petitcolas [120, 121]. StirMark benchmark is one of the most used benchmarks for still image watermarking. Some of the image alterations implemented in StirMark are: cropping, flip, rotation, rotation+scaling, Gaussian filtering, sharpening filtering, linear transformations, random bending, aspect ratio changes, line removal and color reduction. Jane Dittmann *et al.* have also been working on an audio version of StirMark called AudioStirMark [79]. Other benchmarks for still image watermarking are the Optimark [111] software, and the Checkmark [117, 118] developed at the Computer Vision and Multimedia Laboratory of the University of Geneva. Optimark benchmark includes several image attacks plus some watermarking algorithms performance evaluation methods such as statistics to evaluate detector performances (e.g. Bit Error Rate (BER), probability of false detection, probability of missing detection), the estimation of mean embedding and detection time. The CheckMark software is one of the most recent benchmarks for still images and includes some new classes of attacks such as non-linear line removal, collage attack, denoising, wavelet compression (JPEG2000), projective transformations, copy attack, etc. The CheckMark package was primarily developed by Shelby Pereira [118].

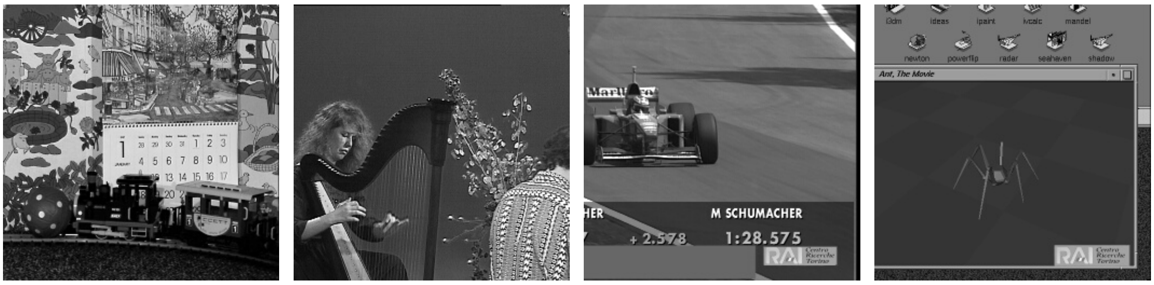
While benchmark tests have already been proposed for the robustness of watermarking algorithms, much less attention has been directed at measuring the *visual effects* of the watermarking process.

3.3 Proposed Approach

An accurate measurement of quality as perceived by a human observer is a great challenge in image or video processing in general. The reason for this is that the amount and visibility of distortions, such as those introduced by watermarking, strongly depend on the actual image/video content. In order to have at disposal a variety of content, we chose to use some of the VQEG [157] test scenes depicted in Fig. 3.2.

The benchmark for any kind of visual quality assessment are subjective experiments, where a number of people are asked to watch the test clips and to rate their quality. Formal procedures for such experiments have been described in the previous chapter. Subjective experiments have been carried out according to the described methods which defined the viewing conditions, criteria for the selection of observers and test material, assessment procedures, and data analysis methods.

Once the subjective results have been obtained, they are compared to the error measures traditionally used in this field to evaluate the quality. In fact, engineers have turned to basic error measures such as mean squared error (MSE) or peak signal-to-noise ratio (PSNR), assuming that



(a) mobile and calendar **(b)** harp and piano **(c)** race car **(d)** computer graphics

Figure 3.2: Sample frames from the test clips.

they would yield quality indications comparable to human perception. However, these simple measures operate solely on the basis of pixel-wise differences and neglect the important influence of video content and viewing conditions on the actual visibility of the artifacts. Therefore, they cannot be expected to be reliable predictors of perceived quality.

The shortcomings of these methods have led to the generation of two perceptual quality metrics presented in the remaining of this chapter [177]. On the basis of the subjective experiments the artifacts introduced in the watermarked video sequences have been identified and a performance model has been applied to build the objective metric. This approach to derive the objective metric is based on *a priori* knowledge about the processing methods as well as the pertinent types of artifacts (see Sec. 2.4.1). It is often a computationally more efficient implementation than models which directly include the HVS characteristics [85, 174].

In the following, we first identify the artifacts caused by watermarking by means of subjective experiments, and then we try to find objective metrics to measure their perceptual annoyance. The proposed metrics are full reference metrics: they compare the video under test with a reference video to measure the quality of the degraded video with respect to the reference.

3.4 Watermarking Algorithms and Artifacts

Most video watermarking techniques today are derived from algorithms for still images. Therefore, we adopt a number of watermarking schemes for still images and apply them to each frame of a video sequence. We chose four algorithms from the literature*, as well as a genuine video watermarking algorithm for video sequences developed by AlpVision†. A brief description of each of these algorithms is given in the following.

The scheme of Cox *et al.* [27] is based on the discrete cosine transform (DCT). The DCT of the entire image is computed, and a sequence of n real numbers is generated from a uniform distribution of zero mean and unit variance, which is then placed into the n highest magnitude coefficients of the transform matrix. Additionally, a scaling parameter α can be specified to determine the amplitude of the watermark.

Dugad *et al.* [39] used a three-level discrete wavelet transform (DWT) with an eight-tap Daubechies filter. The watermark is generated by a sequence of n real numbers and is added to the coefficients

---

* The source code for these algorithms can be downloaded from http://www.cosy.sbg.ac.at/~pmeerw/Watermarking/source/.

† See http://www.alpvision.com for more information.

Table 3.1: Tested watermarking algorithms and their parameters (n is the length, α the strength of the embedded watermark).

Watermarking Algorithm	**Parameters** (default)
Cox	$n = 100$, $\alpha = 0.3$
Wang	$n = 1000$, $\alpha = 0.3$
Xia	$n = 1000$, $\alpha = 0.2$
Dugad	$n = 1000$, $\alpha = 0.2$
AlpVision	$n = 100$

above a given threshold in all sub-bands except the low-pass band. The watermark amplitude can again be controlled by a scaling parameter.

Wang *et al.* [160] adopted a successive subband quantization scheme in the multi-threshold wavelet codec to choose perceptually significant coefficients for watermark embedding. The watermark is inserted in the coefficients above a certain threshold in the current subband while taking into account the scaling factors α and β, which are adjustable by the user.

Xia *et al.* [182] decomposed an image into several bands by a DWT. The watermark is added to the largest coefficients in the high- and middle-frequency bands of the DWT. A parameter α is tuned to control the level of the watermark. The output of the inverse DWT is modified in such a way that the resulting image has the same dynamic range as the original.

The video watermarking scheme developed by AlpVision is based on a technique initially proposed for still images [76, 77]. It uses spread-spectrum modulation to insert a watermark with a variable amplitude and density in the spatial domain. In contrast to the other four algorithms, it considers the temporal content changes in the video.

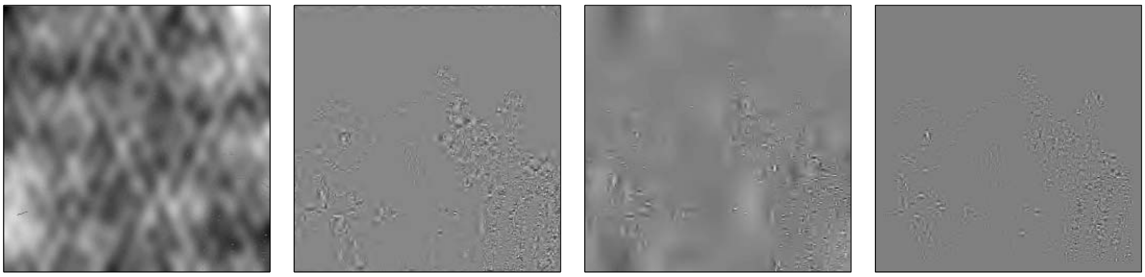
The default settings of each algorithm were used for all parameters (see Tab. 3.1). The resulting watermarks are shown in Fig. 3.3 for a sample frame from one of our test clips. As can be seen, some watermarking algorithms take into account masking phenomena in the human visual system to a certain extent and insert their watermarks mainly in image regions with high spatial activity (bottom right part of the frame).

The transition of watermarking from still images to video sequences would require some changes to the algorithms described above. However, the goal of this chapter is not to provide an algorithm for watermarking video sequences but to assess the perceptual quality of watermarked video sequences.

The most prominent feature in video sequences is the increased sensitivity to changes introduced by the watermarking process. Figure 3.4 illustrates the visibility problem. The block position is no longer restricted to a single image (x and y axes) but extends to the time axis, t. The modification of blocks that are close to each other in x and y as well as in t can result in *flickering* effects.

Homogeneous areas within frames are particularly sensitive to this type of degradation as are regions containing sharp edges. Two criteria for checking sensitive areas before actually computing the error introduced by watermarking have to be considered: *edge* and *smooth* area detections. In the edge areas the noise effect can be noticed. This is due to the fact that the watermarking signal can be considered as a noise-like signal, caused by the noisy nature of the pseudo signal embedded in the original signal. On the other hand, the flickering artifact introduced by the visibility problem in consecutive frames (see Fig. 3.4) affects smooth areas. It is caused by modification of closely spaced blocks on the sequential frames and the homogeneous areas are more sensitive to these effects.

On the basis of these two visual effects, two objective metrics will be proposed in Sec. 3.6.



(a) Cox et al. [27] **(b)** Dugad et al. [39] **(c)** Wang et al. [160] **(d)** Xia et al. [182]

Figure 3.3: Intensity errors produced by the watermark obtained by four different algorithms for the frame shown in Fig. 3.2(b). Dark pixels denote negative values, bright pixels denote positive values, medium gray denotes no change. The images were normalized individually to enhance the visibility of the watermarks.

3.5 Experimental Method

A set of standards and grading techniques to subjectively evaluate quality of processed video and multimedia content have been defined by ITU-R [66] and ITU-T [65] and have already been described in Sec. 2.2. In this section, the video clips used in the subjective experiments and the method used to carry out the tests are presented.

3.5.1 Test Clips

Four different test clips were watermarked for this analysis. These clips were selected from the set of scenes in the first VQEG test [157] to include spatial detail, motion, and synthetic content. Each of them has a length of 8 seconds with a frame rate of 25 fps. They were de-interlaced and sub-sampled from the interlaced ITU-R Rec. BT.601 format [64] to a resolution of 360×288 pixels for progressive display. The implementations of some watermarking algorithms mentioned in the previous section are limited to frame sizes of powers of 2, therefore a 256×256 pixel region was cropped from each frame in the video for watermarking and subsequent quality evaluation. A sample frame from each of the four scenes is shown in Fig. 3.2. In the "mobile and calendar" sequence a toy train is running and in the background a calendar is moving up and down on a textured wallpaper; the "harp and piano" sequence is another indoor sequence: a man is playing a piano and a woman is playing a harp in a smooth background while the camera is zooming on the woman; "race car" is an outdoor sequence in which the driver is racing and goes off the road, finally, "computer graphics" has been obtained synthetically in which a graphic ant is walking on a uniformed background.

3.5.2 Subjective Experiments

For the evaluation of our metrics, subjective experiments were performed. Non-expert observers were asked to rank a total of 20 watermarked test clips from best to worst according to the perceived noise and flicker in two separate trials. The test sheet was like that used in "categorical continuous scale" depicted in Fig. 2.1 (b). The viewing order of the clips was not fixed; observers could freely choose between clips and play them as often as they liked. They could also watch the original clips for comparison. Five observers participated in the noise trial, and six in the flicker trial. For

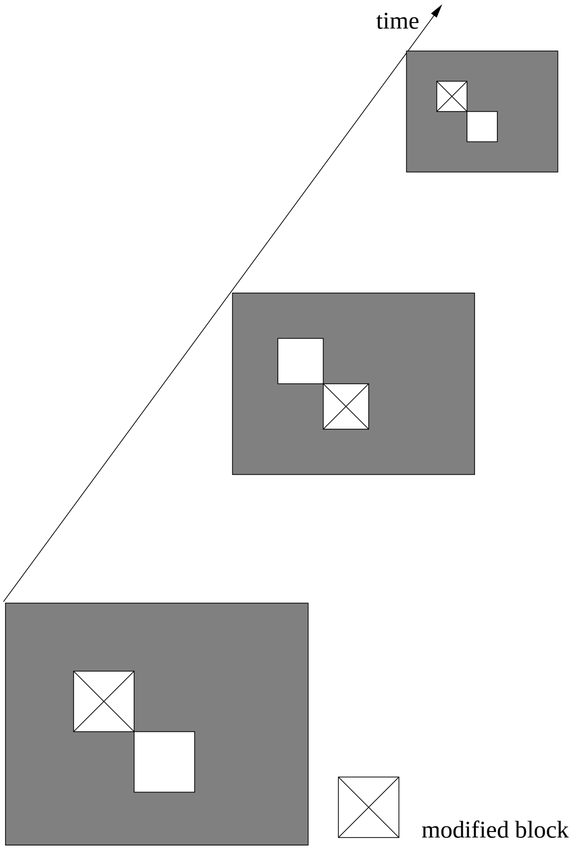


Figure 3.4: Visibility problem in consecutive frames.

comparison with the objective metrics, the data obtained from the subjective ratings were combined to an average rank.

According to the subjective experiments, the most annoying artifacts in video are produced by watermarking algorithms that add noise patterns with relatively low spatial frequencies, which change from frame to frame and thus create clearly visible flicker. Algorithms that add mainly high-frequency noise or temporally unchanging patterns to the video exhibit much less visible distortion.

3.6 Proposed Objective Metrics

From subjective experience with the numerous tested video watermarking algorithms, mainly two kinds of impairments have been seen:

- Spatial noise, which is the fundamental fingerprint of most watermarks;
- Temporal flicker, which results from visible changes of the watermark pattern between consecutive frames.

Based on these observations, we have designed objective metrics that measure the perceptual impact of these two impairments, which we refer to as Noise metric and Flicker metric, respectively in the following. These metrics became part of Genista's *Video PQoS*^{TM} software,[*]. *Video PQoS* is an

---

[*] See http://www.genista.com for more information.

application for the measurement of artifacts affecting the perceptual quality of digital video. It works with full reference quality metrics, i.e. it compares the video under test with a reference video to measure the quality of the degraded video. In addition to the watermarking metrics, *Video PQoS* also provides perceptual metrics for compression artifacts, metrics as defined by ANSI T1.801.03 [144], and fidelity metrics such as PSNR.

Noise Metric

For the computation of the noise metric, the watermark is first extracted as the difference d between a frame in the processed sequence and the corresponding frame in the reference sequence: $d(x,y) = Y_{\mathrm{prc}}(x,y) - Y_{\mathrm{ref}}(x,y)$.

Let $D(u,v)$ be the coefficients of the two-dimensional discrete Fourier transform of $d(x,y)$. Based on the vector feature proposed in section 6.1.2 of ANSI T1.801.03 [144], the radial average r_d of the 2D-DFT coefficients is computed as the absolute sum over the Fourier coefficients inside a ring $\mathcal{R}_k$, defined by $k-1 < \sqrt{u^2+v^2} < k$ for each k:

$$r_d(k) = \frac{1}{N_{\mathcal{R}_k}} \sum_{(u,v)\in\mathcal{R}_k} |D(u,v)|, \tag{3.1}$$

where $N_{\mathcal{R}_k}$ denotes the number of coefficients within ring $\mathcal{R}_k$.

Finally, the sum over the higher frequency range $(f_M \dots f_H)$ of this radial spectrum $r_d(k)$ is computed to yield the Noise metric:

$$\text{Noise} = \frac{1}{f_H - f_M} \sum_{k=f_M}^{f_H} r_d(k). \tag{3.2}$$

We empirically choose the frequency limits to be $f_M = 16\%$ and $f_H = 80\%$ of the maximum spatial frequency.

Flicker Metric

As before, the watermark is extracted as the difference $d(x,y)$ between a frame in the processed sequence and the corresponding frame in the reference sequence. This is done for two consecutive frames, giving d_n and d_{n+1}. The change of this watermark from one frame to the next is computed as $c(x,y) = d_{n+1}(x,y) - d_n(x,y)$.

We again compute the radial frequency spectrum as described above, but this time using the 2D-DFT of $c(x,y)$:

$$r_c(k) = \frac{1}{N_{\mathcal{R}_k}} \sum_{(u,v)\in\mathcal{R}_k} |C(u,v)|. \tag{3.3}$$

The sum over the low frequencies $(f_L \dots f_M)$ of r_c,

$$s_L = \frac{1}{f_M - f_L} \sum_{k=f_L}^{f_M} r_c(k), \tag{3.4}$$

as well as the sum over the high frequencies $(f_M \dots f_H)$ of r_c,

$$s_H = \frac{1}{f_H - f_M} \sum_{k=f_M}^{f_H} r_c(k), \tag{3.5}$$

are calculated. We empirically chose the frequency limits to be $f_L = 1\%$, $f_M = 16\%$, and $f_H = 80\%$ of the maximum spatial frequency.

To take into account spatial and temporal masking by the reference sequence, we estimate the spatial and temporal activity in the reference sequence. This is again based on simple features defined in ANSI T1.801.03 [144], namely the spatial information SI, which is the gradient computed from the horizontally and vertically Sobel-filtered image, as well as the temporal information TI, which is simply the pixel-wise difference between two consecutive frames. We normalize both SI and TI with respect to the maximum possible values. Using the average spatial information (the average gradient magnitude, to be precise) of the reference frame, $\overline{SI} = \sum |SI_r(x,y)|$, and the average temporal information of the reference frame, $\overline{TI} = \sum |TI_r(x,y)|$, a scalar weight m is computed:

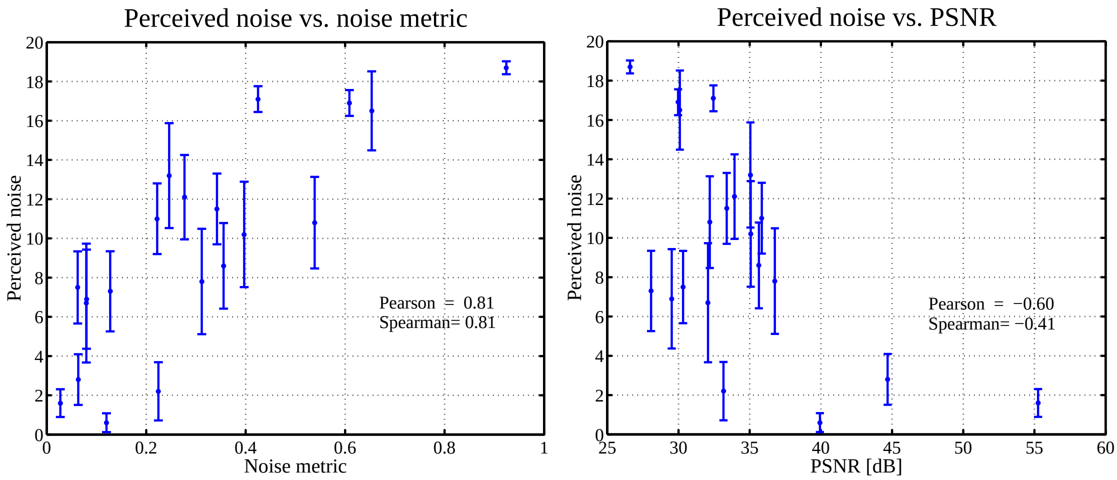
$$m = \max\left(\overline{SI} \cdot \overline{TI}, t\right), \tag{3.6}$$

where t is a threshold to avoid extreme values of masking. We empirically choose $t = 0.007$.

From the above, the Flicker metric is computed as:

$$\text{Flicker} = \frac{s_L + s_H}{m}. \tag{3.7}$$

The proposed noise and flicker metrics will be validated by means of subjective experiments in the next section.



(a) Subjective noise ratings vs. Noise metric.

(b) Subjective noise ratings vs. PSNR.

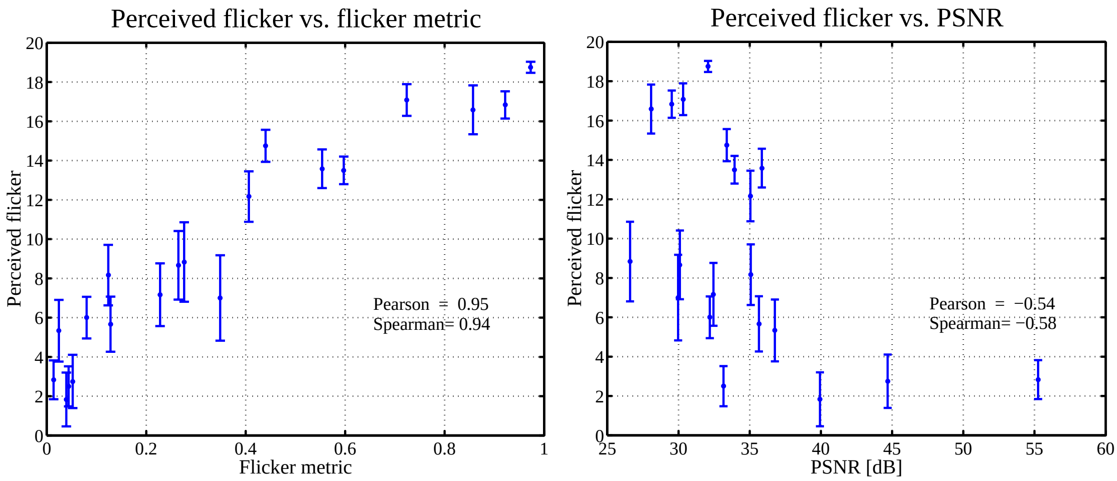
	Noise Metric	PSNR
Pearson	0.81	−0.60
Spearman	0.81	−0.41

(c) Correlations.

Figure 3.5: Perceived noise vs. Noise metric and PSNR (subjective data are shown with 95%-confidence intervals).

3.7 Experimental Results

A statistical analysis of the data was carried out to evaluate the two proposed metrics with respect to the subjective ratings. The subjective scores have to be condensed by statistical techniques



(a) Subjective flicker ratings vs. Flicker metric.

(b) Subjective flicker ratings vs. PSNR.

	Flicker Metric	PSNR
Pearson	0.95	-0.54
Spearman	0.94	-0.58

(c) Correlations.

Figure 3.6: Perceived flicker vs. Flicker metric and PSNR (subjective data are shown with 95%-confidence intervals).

used in standard methods (see Sec. 2.3) to yield results which summarize the performance of the watermarking system under test. The averaged subjective score values, (MOS), is considered as the averaged amount of perceived flicker/noise that anyone can perceive on a particular watermarked video.

Two correlation coefficients are used here to quantify and compare the metrics' performance and have been described in Sec. 2.4.2, namely the (linear) Pearson correlation coefficient and the (non-parametric) Spearman rank-order correlation coefficient.

The scatter plot of perceived versus measured noise for the above-mentioned watermarking algorithms and test clips is shown in Fig. 3.5(a). For comparison, the scatter plot of perceived noise versus PSNR is shown in Fig. 3.5(b). The respective correlation coefficients are shown in the respective Fig. 3.5 (c).

Figure 3.6 shows the same data for perceived flicker, PSNR and the Flicker metric. The proposed metrics clearly outperform PSNR in both cases as shown by the correlation coefficients. The plots show that adding a temporal component such as flicker to the measurements is essential for the evaluation of video watermarks, because PSNR is unable to take this into account. More surprisingly perhaps, PSNR is not well correlated with perceived noise either. This shows the importance of more discriminatory metrics for the perceptual quality evaluation of watermarks.

3.8 Conclusions

In this chapter, the importance of perceptual quality assessment in watermarking has been discussed. While this remains a difficult problem, a possible solution path was presented. We found that watermarked video suffered mostly from added high-frequency noise and/or flicker in the performed subjective tests. The watermarking artifacts, which may be hardly noticeable in still images, become emphasized through the motion effects in video.

Two new metrics that analyze the video by specifically looking for watermarking impairments have been introduced, namely a Noise metric and a Flicker metric, which measure the perceptual impact of these specific distortions. Through subjective experiments we have demonstrated that the proposed metrics are reasonably reliable predictors of perceived noise and perceived flicker and clearly outperform PSNR in terms of prediction accuracy. The developed metrics are now part of Genista's *Video PQoS*^{TM} software.

Part II

Video Object Segmentation Quality Assessment

Subjective Segmentation Evaluation

4

4.1 Introduction

The unsupervised segmentation of digital images is a difficult and challenging task [143] with several key-applications in many fields: image classification, object recognition, remote sensing, medical diagnosis, vision-driven robotics, interactive entertainment, movie production and so on. The performance of algorithms for subsequent image or video processing, compression and indexing to mention a few, often depends on a prior efficient image segmentation. Basically, by segmenting an image, several "homogeneous" partitions are created. The number of homogeneity criteria depends on the particular application and on the *a priori* knowledge of the problem. For example, in a video surveillance application every moving object can be considered as an object of interest and, therefore, this information is used in the segmentation process.

Because of the importance of this task, many segmentation algorithms have been proposed, as well as a number of evaluation criteria. Nevertheless, very few comparative results of segmentation algorithms have been conducted. Many researchers prefer to rely on qualitative human judgment for evaluation. In fact, Pal and Pal [115] say that a "human being is the best judge to evaluate the output of any segmentation algorithm". On the other hand, subjective evaluation asks for a large panel of human observers, thus resulting in a time-consuming and expensive process. Therefore, there is a need for an automatic, objective methodology both to allow the appropriate selection of segmentation algorithms (inter-evaluation) as well as to adjust their parameters for optimal performance (intra-evaluation).

Recent multimedia standards and trends in image and video representation described in the next section, have increased the importance of adequately segmenting semantic "objects" in video, in order to ensure efficient coding, manipulation and identification.

During the last several years, some objective methods for video object segmentation evaluation have been proposed, but no work has been done on studying and characterizing the artifacts typically found in digital video object segmentation. A good understanding of how annoying these artifacts are and how they combine to produce the overall annoyance is an important step in the design of a *perceptual objective quality metric*. To this end, a series of specially designed psychophysical experiments is performed. In these experiments, we use test sequences with synthetic artifacts that look like *real* artifacts, yet simpler, purer, and easier to describe. In this chapter, we present the methodology used for the psychophysical experiments performed in this thesis.

The motivations of this work are described in Sec. 4.2. Methodologies for both image and video subjective evaluation are reviewed in Sec. 4.3. The proposed experimental method for subjective tests, the instructions and experimental tasks are described in Sec. 4.4. Section 4.5 introduces the novelty of the proposed approach that consists in deriving a *perceptual objective metric* from subjective experiments. This section also describes the synthetic artifacts introduced in the test sequences. Section 4.6 includes the conclusions.

4.2 Motivations

The task of extracting objects in video sequences emerges in many applications such as object-based video coding (e.g. MPEG-4) and content-based video indexing and retrieval (e.g. MPEG-7). The MPEG-4 standard provides specifications for the coding of video objects, but does not address the problem of segmenting objects in image sequences. Video segmentation is still a matter of intensive research. Various other applications, such as editing and manipulation of video sequences, video surveillance, or image and video indexing and retrieval applications, are equally dependent on the availability of sophisticated algorithms for content identification, content segmentation, and content description. While powerful solutions exist for some applications, the design of suitable fully automatic algorithms, in particular for image sequence segmentation, still remains an unsolved problem.

The MPEG-4 standard assists the coding of objects in image sequences separately in different object layers. Thus, in MPEG-4, image sequences can be considered to be arbitrarily shaped, in contrast to the standard MPEG-1 and MPEG-2 definitions [74]. Video sequences are decomposed into individual objects in a scene and the objects are encoded entirely separately in individual object layers. In general, this provides to the user an extended content-based functionality (the ability to separately access and manipulate video content) and it is also possible to achieve increased image quality for a number of applications. This requires the segmentation of video sequences into objects of interest prior to coding. However, this can be an extremely difficult task for many applications. If the video was originally shot in a studio environment using the chroma-key technology, image segmentation can be easily performed, e.g. in the case of weather forecasting for news sequences. If no chroma-key segmentation is available, as for most scenes under investigation, the segmentation of the objects of interest needs to be performed using automatic or semi-automatic algorithms. To the best of our knowledge, there is no universal algorithm that could potentially solve the segmentation problem for all tasks. The video segmentation task still remains to a large extent an unsolved problem. It is also an ill-posed problem resulting in a variety of tools and algorithms described in literature, each of them specialized and optimized for a specific segmentation task.

In many applications, a considerable amount of user interaction with the segmentation process is required. The effect of user interaction and subsequent improvements in segmentation quality are not examined in this thesis. Rather, we aim at finding an automatic procedure for evaluating the performance of a fully automated approach which correlates well with the perceived quality of segmentation. The existence of an ideal segmentation –*ground truth*– manually extracted is assumed so that the objective metric allows for the evaluation and ranking of segmentation algorithms on that test data set.

In order to find such a *perceptually driven* objective methodology for segmentation assessment, we have to derive it from subjective experiments. In fact, the perceived quality of a segmentation depends on the annoyance and visibility of the segmentation *impairments* [36, 37, 38]. An *impairment* in a segmentation is any change compared to the ideal segmentation that if sufficiently strong will reduce the perceived quality. Segmentation impairments can be introduced by incapability of segmentation algorithms to deal with noise or shadows [28] present in the original video to as well as by various limitation, artifacts and failures of the image analysis techniques. Most impairments have more than one perceptual feature, but it is possible in segmentation to produce impairments that are relatively pure. We will use the term *artifacts* to refer to the perceptual features of impairments such as spatial and temporal positions, size, number, duration, shape and so on. The novelty of our approach, compared to the state of art evaluation metrics, consists in taking into account the most common artifacts produced by segmentation algorithms and carrying out subjective experiments to 1) study and characterize the impact of different artifacts; 2) derive the perceptual annoyance for each kind of artifact; and 3) find the perceptual interaction among artifacts when they are all combined.

To this purpose, subjective experiments have to be performed and a protocol for subjective evaluation of segmented video objects has to be proposed. The task of defining a formal protocol for subjective tests for video object segmentation quality assessment is very useful, since to the best of our knowledge, only informal tests have been performed [17], [23]. Common practices for evaluating segmentation results are based on human intuition or judgment (*subjective evaluation*) and consist in *ad hoc* subjective assessment by a representative group of observers. To be meaningful, the evaluation must follow precise methodologies, both in terms of test environment set-up as well as of grading techniques. The presented subjective protocol is an effort to make subjective evaluations in this field more reliable, comparable and standardized. Little has been done towards defining a procedure to evaluate the performance of objective metrics for segmentation [51]. Standard subjective evaluation methodologies for image and video segmentation quality evaluation are not yet established but some informal tests have been carried out. These methodologies, both for image and video, are reviewed in Sec. 4.3.

Some works [93, 96, 187] on segmentation evaluation state that, once the specific application is addressed, there is an obvious measure for evaluating the segmentation algorithm. For example, in object recognition, it is typical to use a segmentation algorithm (which could be edge or region based) to partition the image into a number of parts which are then used for object recognition. This means that, in spite of Pal and Pal's statement, it has to be evaluated how well the application does perform using a particular segmentation algorithm, hence "the application is the best judge of any segmentation algorithm".

We do not wish to deny the importance of evaluating segmentation in the context of a task. However, a first point in this thesis is that different segmentations can also be evaluated purely as segmentation results by comparing them to those produced by an ideal segmentation algorithm. In fact, by means of subjective experiments, a considerable consistency among subjective evaluations can be found. Second, it is rarely feasible to build an entire system in order to test different segmentation algorithms because of expense, and because the properties of the segmentation algorithm will often determine what form the subsequent processing should take. Third, application-oriented evaluations might not be able to take into account other functionalities provided by the segmentation, such as individual access to regions of interest (*objects*). Therefore, at first, we are interested in evaluating segmentation without the implementation of subsequent processing and then to focus on specific applications.

4.3 Subjective Segmentation Evaluation Methods - Background

The problem of subjectively assessing the quality of segmentation has been investigated in different contexts in literature: edge based segmentation, region based segmentation, and video object segmentation. Nevertheless, there is no standardized procedure for subjective tests on any of these segmentation methods. In literature, subjective judgments are based on human intuition.

In order to provide a clear overview of the adopted *ad hoc* subjective approaches, a distinction is made between methods which deal with still images (Section 4.3.1) and those used for image sequences (Section 4.3.2). The latter is more relevant to the scope of this thesis.

4.3.1 Still Images

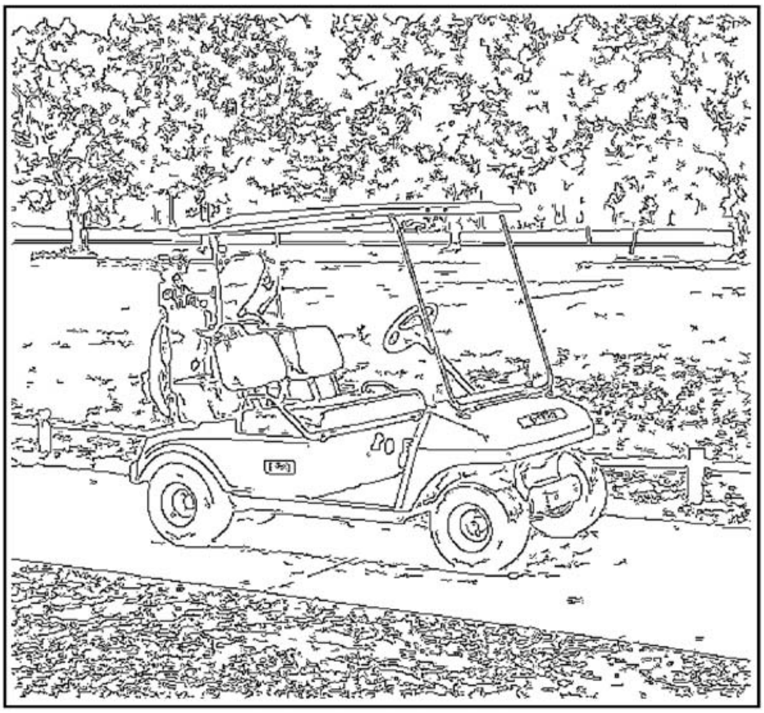
In subjectively evaluating edge detection algorithms [96] and still image segmentation [91, 137, 165, 183], some approaches have been proposed.

There may be some problems in finding agreement among human viewers on how an ideal segmentation should look like.

Heat *et al.* [96] used a subjective evaluation method to compare the outputs of different edge detectors. They asked people to rate the performance of algorithms on a subjective quantitative scale of one to seven. They used the criterion of ease of recognizability of objects (for human observers) in the edge images: a score of seven indicates that the "information allows for easy, quick and accurate recognition of the object", and one indicates that "there is no coherent information from which to recognize the object" (see Fig. 4.1). The authors checked the consistency of the subjects' ratings and found that subjects shared a concept of "edge goodness". On the basis of these scores, they could rate the performance of various algorithms. This approach to subjectively assess the goodness of a segmentation could be used in video object segmentation evaluation in a specific application case: surveillance and with emphasis on the purpose of the surveillance. For example, in intruder detection the most important part to recognize is the face; in highway traffic monitoring the most important part to be recognizable are the plates of the vehicles. On the basis of these specifications this scheme could be used to subjectively evaluate the segmentation algorithms. However, a more general method is needed for any video object segmentation application or to be slightly changed according to the scenario.

In evaluating directly the output of segmentation schemes, Shaffrey *et al.* [137] used psychovisual tests on segmented images. In this subjective evaluation method, the subjects would choose between two kinds of segmentation for each image as shown in Fig. 4.2. The judgment involves both which segmentation the subject prefers and how quickly he/she chooses it. The results confirm that human subjects' judgments agree, thus allowing meaningful subjective evaluation of segmentation algorithms. This approach is not suitable to characterize the perception of different segmentation errors in video sequences: four images may be too many since subjects can concentrate only on one of them.

An alternative approach is to allow human subjects to perform manually the segmentation of the original image for which well-defined semantics exist but ground truth is hard to obtain. This approach consists in asking different subjects to manually segment the image and to see if a reasonable consensus emerge. Warfield *et al.* [165] and Yang *et al.* [183] used multiple expert observers to agree on ground truth in the context of medical imaging. Then, such estimated ground truth can be used as a gold standard for validation. However, such approaches can be adopted only in the cases where experts' segmentations are available, such as specific anatomical or functional important structures in a MR image [5].



Information allows for easy, quick and accurate recognition of the object.

7 6 5 4 3 2 1

No coherent information from which to recognize the object.

Figure 4.1: An example of edge segmentation evaluation sheet from Heat *et al.*'s method [96].

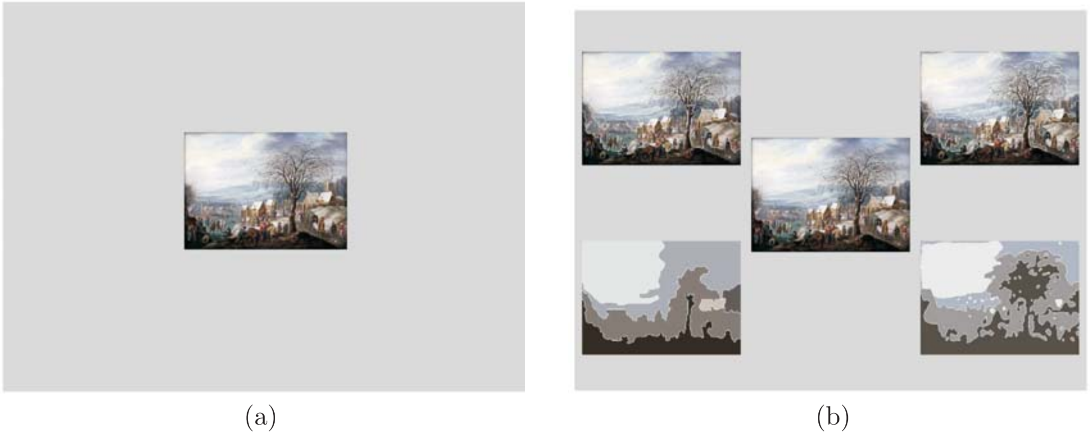

(a) (b)

Figure 4.2: Test images for evaluating the best segmentation [137]: (a) shows the first stage of the trial with the original image and (b) shows the second stage in which the segmentation schemes are represented by their outline and masks with the original in the center.

Martin *et al.* [91] asked subjects to break up the scene in a “natural” manner by dividing the image into pieces (between 2 and 20), where each piece represents a distinguished region in the image and all pieces are approximately of equal importance. They found that different manual segmentations of the same image are highly consistent even when no specific application is explained to the subjects. A large database of natural images segmented by human observers has been made available [90]. Figure 4.3 shows different manual segmentations from the database. This dataset serves as ground truth benchmark to compare different segmentations and boundary finding algorithms. In theory , this approach could be also be adopted in video object segmentation, but practically it is not a feasible method since it requires for too much time of subjects. For example for four video sequences of at least 60 frames each, we found it requires 60 hours time for a subject to perform the manual segmentations. In fact, in a trial we carried out, after the subject acquires some experience in manually segmenting he/she needs in average 15 minutes per frame.

4.3.2 Image Sequences

Very little has been done in the literature to establish an experimental method for subjective tests on image sequences [23, 51, 95]. To make subjective evaluation of video object segmentation more reliable, comparable and standardized, the subjective evaluation must follow precise methodologies, both in terms of test environment set-up as well as scoring techniques. As already mentioned, standard subjective evaluation methodologies for video quality evaluation [65, 66] can provide important guidelines (e.g. for display configuration and experimental conditions) for subjective evaluation of video segmentation quality evaluation.

A set of general guidelines for segmentation quality assessment has been proposed in the COST211 quat European project [51] entitled “Compare your segmentation algorithm to the COST 211 quat analysis model”. These guidelines concern only how the typical display configuration should look like (for further details see [23]), but they do not specify how the test should be carried out. In this framework, the display layout is different whether the ground truth is made available or not. If not available, the layout includes four images: 1) the original image, 2) the segmentation partition (with a different color representing each object), 3) the foreground original segmented object under analysis over a neutral background and 4) vice-versa: the remaining part of the original image with

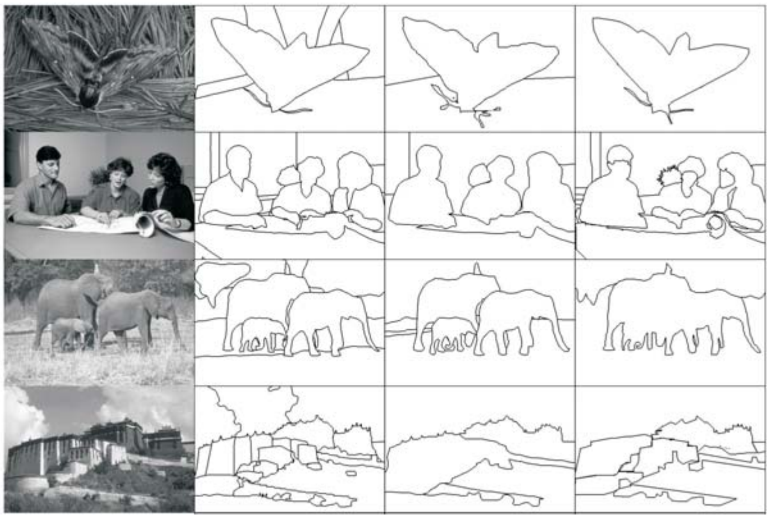


Figure 4.3: Each image has been segmented by 3 different people from the Berkeley segmentation dataset [90].

the object area replaced by a neutral color. For subjective evaluation where the ground truth is used, the original image is replaced by the reference object with original texture over a neutral background, and the segmentation partition by the image with the original image overlaid with the reference object area in a neutral color. Thus, this framework proposes to show people four images at the same time and it does not specify how long the video sequences should be. We performed some informal tests using this display configuration and noticed that for short video sequences (5-10 seconds), four images may be too many since subjects can concentrate only on one of them. Moreover, this layout also shows the original image sequence without any segmentation. This image may not be needed, since the subject, once he/she has learned the task, forms his/her own *implicit* segmentation and does not look any more at the original nor at the reference segmentation. Finally, showing the masks of the object could be of some practical utility for the evaluators if the specific application of the segmentation (e.g. automatic surveillance system) has been specified, otherwise human attention would be attracted to the textured object segmentation.

In [95] some criteria related to the computational complexity of the segmentation system are defined together with a number of questions to investigate subjectively the video object segmentation quality for surveillance applications. For each video sequence, the subject can see the original video sequence as many times as necessary. Then, the segmented video is presented only once and the subject has to answer to 4 evaluation criteria (such as “how well have been important moving objects individually identified?”, or “how well are boundaries provided?”). Table 4.1 reports the segmentation evaluation criteria. The segmented regions provided by the algorithms are represented by the use of colors.

In the informal tests that we performed, we tried to combine the use of different questions to describe the different aspects of segmentation quality. The drawback of this method is that the subjects have to perform a sort of *memory test* given the large number of questions they have to answer after the video is played back. The capacity of a test subject to reliably assess several elements of a video is limited. The memory of a video fades after time. This results in a tiring and too difficult task to be accomplished by a subject.

Table 4.1: Evaluation criteria in McKonen *et al.* [95]'s method.

	Algorithm Evaluation Criteria	**Evaluation Method**
1	Segments moving 'semantic' objects from the background.	Subjective assessment
2	Tracks individual regions throughout the video sequence.	Subjective assessment
3	Provides accurate region, or preferably object boundaries.	Subjective assessment
4	Distinguishes between moving objects and image perturbations.	Subjective assessment
5	Segments objects into associated sub-regions.	Subjective assessment
6	Eliminates or correctly identifies shadows.	Subjective assessment
7	Low computational complexity.	Run-time data
8	Has few configuration parameters.	Run-time data
9	The segmentation is illumination invariant.	Post-assessment analysis
10	The segmentation performs well for outdoor sequences.	Post-assessment analysis

Table 4.2: Viewing conditions during subjective test.

Variable	Values
Peak luminance	≤ 0.04
Maximum observation angle	10 degrees
Monitor resolution	1024×768
Viewing Distance	$35 - 40$ cm
Monitor Size	19"

4.4 Proposed Method for Subjective Evaluation

The goal of this research is *to find a way to predict what people will say about the quality of segmentation without performing any subjective test.* In order to reach this goal, we have to answer to the following questions: 1) What method people use to judge the segmentation quality; 2) Whether people generally agree on the quality of a segmentation (that is not trivial as discussed in Sec. 4.3.1); 3) Whether the expectation of quality affects ratings and other of these unknowns. An experiment should, if designed correctly, at least answer one or more of these questions. To this end, with the help of experts in psychophysical testing, we designed a series of psychophysical experiments. This method aims at making subjective evaluations in segmentation evaluation more standardized. In these experiments, we used test sequences with synthetic artifacts that look like "real" artifacts, but simpler, purer and easier to describe. In the remaining of this chapter, we describe the display layout, the instructions, the process of generating the test sequences and the data analysis used for the experiments. In the next chapter we present the experimental results of the subjective tests on the perceptual impact of different types of artifacts and how these artifacts combine and interact to produce the overall quality. The display layout and viewing distance were in concordance with subjective viewing for CIF format [64] images (see Table 4.2).

4.4.1 Basic Methodology

In general, psychophysical experiments are expensive both in terms of time and resources. The design, execution and data analysis consume a great amount of the experimenter's time. Running an experiment also requires a large investment of time from the subjects. As a result, the number of experiments that can be conducted is limited. An appropriate methodology needs to be developed to maximize the information collected per experiment.

For various reasons, the method of *single stimulus continous annoyance scale* was chosen as the basis for our experiments (see Sec. 2.2.1). First, it was single stimulus, as once subjects have learned the task they form their implicit segmentation, as mentioned in Sec. 4.3.2, and they do not need a term of comparison displayed aside the segmentation under test in the evaluating procedure. Second, continuos rating was used to avoid quantization errors. Finally, an annoyance scale was chosen instead of quality rating rather nebulous concept as discussed in Sec. 2.2.

We aimed at building the experiment with as few *a priori* as possible. The fewer assumptions about the attributes of interest, the better. Several types of questions can be made with this method after each test video is shown. We were generally most interested in knowing how annoying the defects (impairments) and how strong or visible a set of artifacts were in the impairment. A few other questions were asked at the end of each experiment, such as how big the artifacts were, when they occurred and what was the impact of the bad segmentation on the overall annoyance. We could have asked each subject all these questions after every video. However, the ability of a test subject to accurately judge multiple aspects of a video at the same time is limited [66] and not reliable. Therefore, we preferred to ask some questions at the end of the experiment to collect the overall impressions.

The methodology for the experiment is described in the following section. This methodology, with minor variations, was applied to all the experiments carried out during our research. The experiment scripts can be found in Appendix B.

4.4.2 Procedure

Prior to the start of each experiment, several tasks had to be accomplished. The first task was the design and implementation of a graphical user interface in all the experiments. This task was performed only once for all the experiments. Figure 4.4 shows the typical display used in subjective experiments, developed in VISUAL BASIC. The second and generally more challenging task was the generation of segmented test video sequences. This task is described in Sec. 4.5.2. The final task was the establishment of the procedure to carry out the subjective tests that was the result of many informal tests and fruitful discussions with psychophysics experts (Mylene Farias and Prof. John Foley from the University of Santa Barbara, California). In the literature, a set of standards and grading techniques to evaluate quality of video and multimedia content have been defined in the ITU-T [65] and ITU-R [66] Recommendations as presented in Chapter 2. However, there are no prescribed standards for the evaluation of segmented video sequences. The protocol for subjective evaluation of segmented video sequences we propose in this thesis is based on ITU recommendations [65] and [66].

A test session is composed of five stages: 1) oral instructions, 2) training, 3) practice trials, 4) experimental trials, and 5) interview. We will now explain in detail each of these stages.

Oral instructions

In the first stage, the subject was verbally given instructions and was made familiar with the task of segmentation of meaningful moving objects. A script was elaborated to help the experimenter to

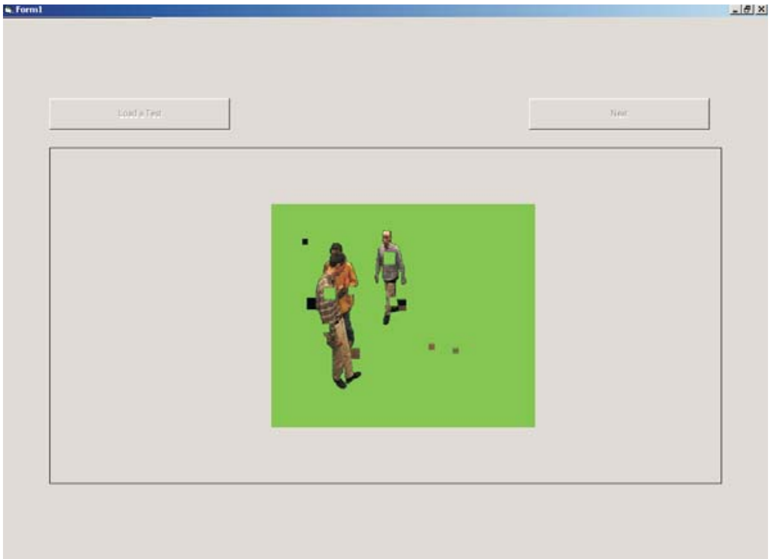


Figure 4.4: Typical display configuration adopted in subjective experiment: in the center the segmented video under test.

perform the experiment. The script contains oral instructions that should be read to the subject to make sure he/she well understands the task to be performed. The scripts varied according to the experiment type (see next section) and are presented in Appendix B.

After the test subject was properly seated at the adequate distance, the tasks to be performed in the experimental trials were explained to the subjects. They were told to disregard the semantic quality of content in the video and to only judge the impairments they see.

Training

The task to be performed in the experiments consists of entering a judgment about an impairment seen in the video. In order to perform this task subjects need to have an idea of how segmentations of a video with no impairments (ideal) compare with that of a video with strong impairments. In the training stage, the original video sequences, the ideal segmentations (reference masks) and sample segmented masks were shown to subjects to establish the subject's range for the annoyance scale (see Fig. 4.5 (a) and (b)). The display configuration showed the texture of the original image in correspondence with the segmented objects/regions over a uniform green background (see Fig. 4.5 (c)). As previously mentioned, the reference segmented masks and the original sequences were only shown in the training stage for two reasons. First, in real applications the reference and the original video are not always available for subjective ratings. Second, in earlier experiments we noticed that subjects do not pay attention to the reference mask or the original video after the training. In fact, they make their own *implicit* segmentation to compare with the segmentation under test. This procedure had the advantage of showing only one video at a time without distracting the subject. This way the subjects' attention was focused on the video to be judged.

In this stage, we chose a subset containing the impairments we believed were the strongest. The subjects were told to *mentally* assign a maximum value of 100 to the worst impairment in the subset (see Fig. 4.5 (c)). As said, explicit reference segmentation masks were not used. Instead, the segmentation representation chosen allowed the original video sequence to be viewed in original texture beneath the uniform background. Viewers were therefore able to see the moving objects and make their judgments as shown in Fig. 4.5.

Practice trials

After the training, in order to familiarize the subject with the experiment and to stabilize the subjects' responses, practice trials were performed with a small subset of the test sequences. Moreover, instead of discarding the first trials as suggested in the ITU Recommendation [66], we included practice trials to eliminate the first erratic answers.

Experimental trials

The experimental trials were performed with the complete set of test sequences presented in a random order. All test subjects saw all the test sequences. The number of test sequences was limited. Previous work in the area of psychovisual testing [48] suggests that a 30-minute time limit should be placed on the length of the test. This is to guard against subject fatigue, which might influence the results in an unpredictable manner. In pilot tests we found that each trial took about 10 seconds. An evaluation set therefore consisted of 150-180 video sequences. For each experiment, several randomly ordered lists of the test sequences were generated. The lists were used sequentially and repeated as necessary. Our test subjects were drawn from a pool of students aged between 22 and 30 years. The number of subjects varied from experiment to experiment but a minimum of 22 subjects were used to guarantee robust results [102].

The subjects were asked one question after each segmented video sequence was presented, *"How annoying was the artifact relative to the worst example in the sample video"*. The subject was instructed to enter a numerical value greater than 0. The value 100 was to be assigned to artifacts as annoying as the most annoying artifacts in the sample video sequences. Although we tried to include the worst test sequences in the sample set, we acknowledge the fact that the subjects might find some of the other test sequences to be worse, and we specifically instructed them to go beyond 100 in those cases. The subjects were then told that artifacts would appear combined or alone and they should rate the overall annoyance in both cases.

Interview

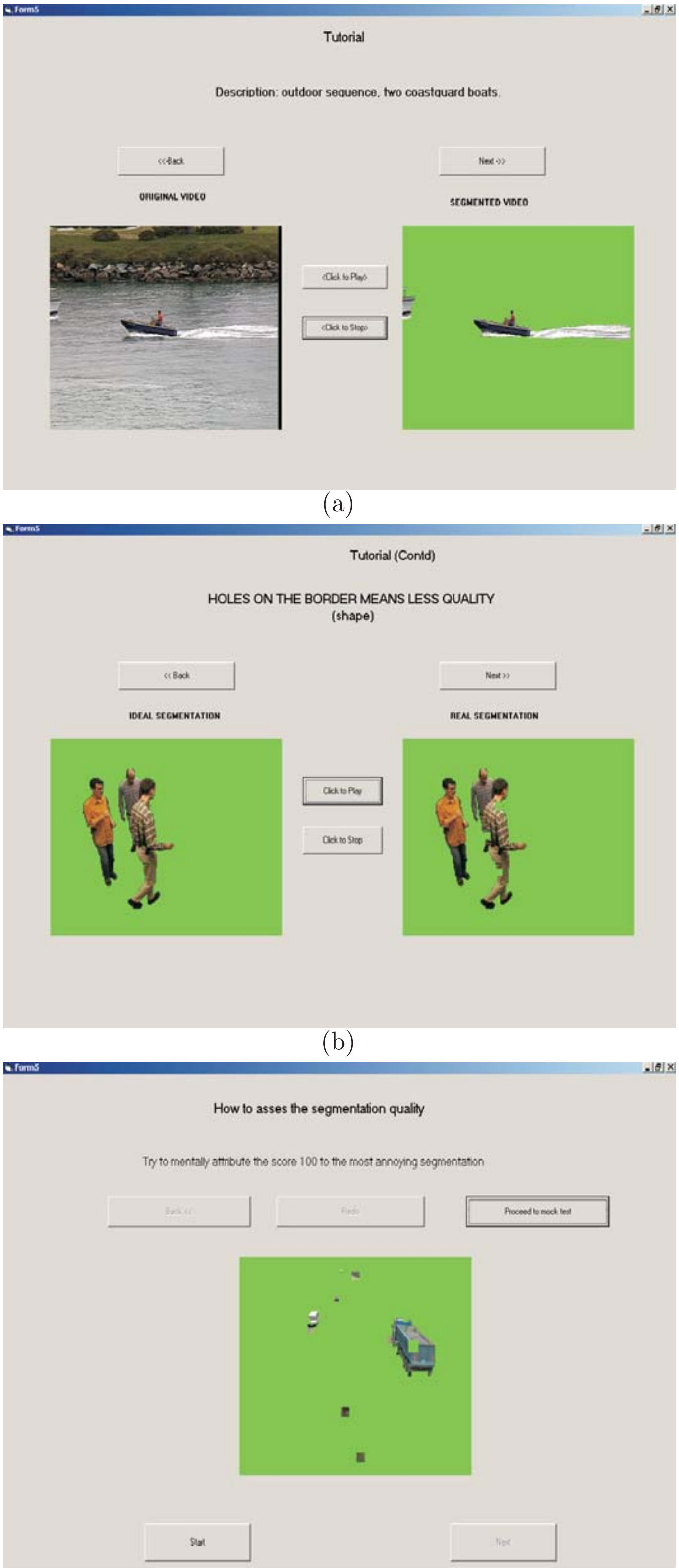
Finally, at the interview stage, we asked the test subjects for qualitative description of the defects that were perceived. The qualitative descriptions are useful for categorizing the defect features seen in each experiment and help in the design of future experiments. In Appendix B a list of interview questions can be found at the end of each script.

4.4.3 Types of Experimental Tasks

According to the goal of the experiment, the subjects were asked to perform one of two different tasks: judging the annoyance of an impairment and judging its strength. In this section, we describe each experimental task. Further details can be found in the scripts in Appendix B.

Annoyance task

The annoyance task consists of giving a numerical judgment of how annoying (bad) the detected impairment is. Examples of original and highly impaired segmentation are shown during the training section. The most annoying segmentations in the training stage should be assigned the value of '100'. The subject is instructed to enter a positive numerical value indicating how annoying the impairment is after each test sequence is played back. Any defect as annoying as the worst impairments in the



Tutorial

Description: outdoor sequence, two coastguard boats.

(a)

Tutorial (Contd)

HOLES ON THE BORDER MEANS LESS QUALITY
(shape)

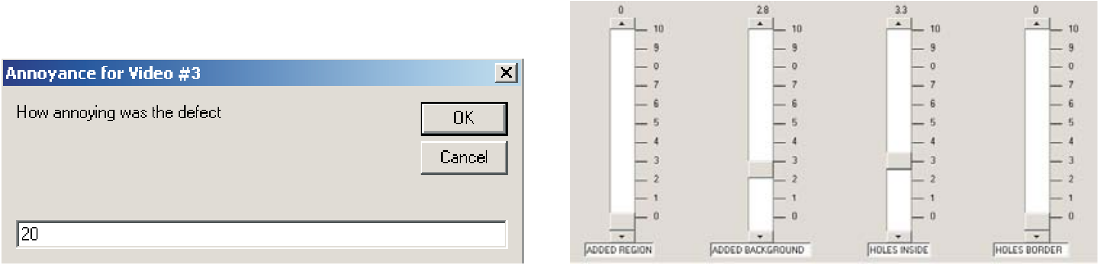
(b)

How to asses the segmentation quality

Try to mentally attribute the score 100 to the most annoying segmentation

(c)

Figure 4.5: Display configurations for training stage: (a) the subject is told about how an ideal segmentation looks like, (b) typical segmentation errors are displayed along with the ideal segmentation, (c) the worst segmentations are shown to establish a subjective range of the annoyance scale.



Annoyance for Video #3

How annoying was the defect

20

0 ADDED REGION

2.8 ADDED BACKGROUND

3.3 HOLES INSIDE

0 HOLES BORDER

(a) (b)

Figure 4.6: Dialog boxes used for the experimental tasks: (a) annoyance and (b) strength.

training stage should be given '100', half as annoying '50', one tenth as annoying '10', and so forth. Although the subjects were asked to enter annoyance values in the range of '0' to '100', they were also told that values greater than 100 can be assigned if he/she thought the impairment was worse than the most annoying impairments in the training stage. Figure 4.6 (a) displays the dialog box used in the experiments. Annoyance values less than zero were not accepted, but the program did not impose any upper limit to the annoyance values. Non-numbers were also rejected. After the value has been accepted, the next video is shown.

Strength task

The strength task consists of asking the subjects for an estimate of how strong or visible a set of artifacts are in the detected impairment. This type of task requires that subjects be taught how each artifact looks like. Therefore, in the training stage subjects were shown a set of sequences illustrating the set of artifacts being measured. In the trials, after the video was played back, the subject was asked to enter a number in a scale with range from '0' to '10' corresponding to the strength of that artifact or feature. If no impairments were seen, subjects were instructed not to enter any number and just click 'Next' to go on to the next trial. Automatically the program set '0' in this case. Figure 4.6 (b) displays the dialog box used for this task.

4.5 From Subjective to Objective Evaluation

In case of subjective evaluations, people watch the segmented images or video sequences and judge the overall quality. To evaluate the segmentation quality objectively, as we aim in this research work, some *objective error measures* related to the artifact such as the number of miss-classified pixels, the distance of miss-classified pixels from the ground truth, etc. are needed. *Perceived errors* are perceptual changes due to defects. In order to assess how objective errors are perceived by humans and to build a *perceptual objective metric*, we use the subjective experiments described in the previous section. We then derive a relation between the perceptual quality of video sequences and the objective features of the artifacts. Since perception is fundamental in judging visual quality, different kinds of artifacts, even with the same amount, are not visually significant at the same degree, as they are perceived differently. Thus to accommodate human perception, different classes of pixels with different relevance must be considered. In Sec. 4.5.1, we describe and provide a mathematical expression for the different classes of segmentation errors we have identified.

If we want to analyze the perception of the identified classes of artifacts, we have to generate different test sequences with various kinds of artifacts. The generated test sequences present pure single synthetic artifacts that look like real artifacts and combinations of them. In Sec. 4.5.2, we

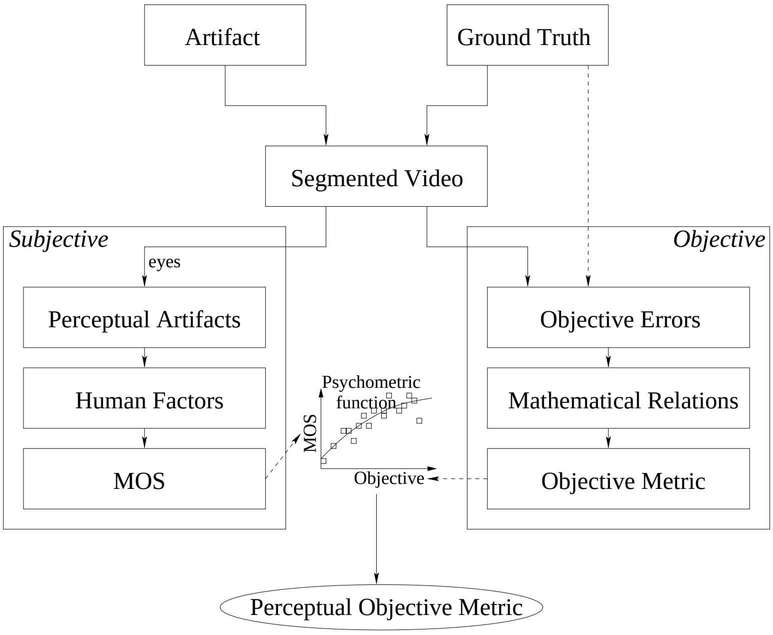


Figure 4.7: Block diagram of the proposed approach: from subjective to objective video object segmentation quality evaluation.

describe the test sequences designed to perform the subjective experiments on artifact perception.

The block diagram of the approach proposed to derive the objective evaluation is depicted in Fig. 4.7. The segmented images or video sequences can be thought as being made of a combination of *ground truth* (reference or ideal segmentation) and *artifacts*. The objective measures are then obtained by subtracting the ground truth from the segmented video under test. In this block diagram, the ground truth link to objective measures is dotted as ground truth may or may be not used to derive these features (see next Chapter). These objective measures are then combined in the overall quality by some mathematical formula forming the *objective metric*.

As mentioned, the goal of our research is to find this mathematical formula which links the objective measures of artifacts to the perceived overall quality of the segmentation. If we find how the segmentation artifacts are perceived by subjects and described by *psychometric functions* (see Section 2.3.2), we can use these functions in the mathematical formula and derive a *perceptual objective metric*. In other words, we derive a relation between amounts of the introduced artifacts (which can be determined objectively) and the perceptual artifacts by taking into account human perception of errors. These errors are then related to the overall quality of the video by a mathematical formula that combines the errors according to their perceptual weights. Finally, we derive a perceptual objective metric which directly correlates the subjective evaluation (Mean Opinion Score, MOS) and the objective features of the artifacts. Hence, the graph plotting these objective and subjective quantities is related by a psychometric fitting curve that produces the *perceptual objective metric* which will be the subject of the next chapter.

4.5.1 Segmentation Errors

It is well known that segmentation errors can affect the quality of a segmented video in two ways: statically (*spatially*) and dynamically (*temporally*) [17, 86]. The *spatial errors* of the segmented video are defined by the amount of mis-segmented pixels that can be easily estimated by a direct

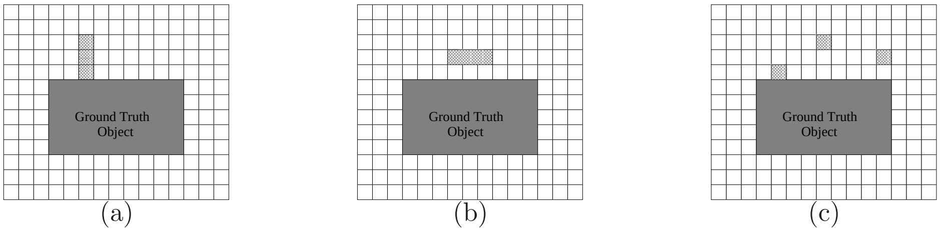


Figure 4.8: Various segmentations that have equal pixel distance measures and also the same number of misclassified pixels [189].

comparison between reference and resulting segmentation mask, for a given frame k. By taking into account the number of mis-classified pixels, an algorithm for object segmentation can in principle be evaluated by estimating only these pixel errors. The sum of the distance between pixels that have been assigned to a wrong class and the nearest pixels that actually belong to the correct class is usually used as evaluation criteria [17, 86]. This simple objective measure suffers from the problem that different configurations can be found for which the same pixel distance error is obtained. Some examples [189] are depicted in Figure 4.8, where the pixel distance errors for the situation (a), (b) and (c) are equal (the number of mis-classified pixels are also equal). Without further processing the three mis-classified pixels in Fig. 4.8 (a), (b) and (c) could be measured as having the same impact (since they are of the same amount and the sum of the distances is the same), whereas the three mis-classified pixels in Fig. 4.8 (a) enlarge the shape of the object by adding some *background* and the three mis-classified pixels in Fig. 4.8(b) are disconnected from the object and perceived as *added region*. The consequences of these two cases are different, for example, both for the influence on the size and the shape of the real objects. Moreover, the pixel distance error cannot distinguish several isolated mis-classified pixels (Fig. 4.8(c)) from a cluster of mis-classified pixels (Fig. 4.8 (a) and (b)), although the two kinds of artifacts are perceived differently. Therefore, we thought to classify different clusters of error pixels according to their perceptual features: size and shape. We group the cluster of error pixels according to the following characteristics: if they do or they do not modify the shape of the object and afterwards their size.

In order to understand how we classified the different clusters of pixel, let us define the different kind of pixel errors. Pixel errors can be divided into two sets [86]: undetected pixels (*false negative*) and incorrectly detected pixels (*false positive*). Let us define a *region* i, $\mathcal{R}_i(k)$, at frame k as a set of pixels with the following properties: 1) $\mathcal{R}_i(k)$ is spatially connected; 2) $\mathcal{R}_i(k) \cup \mathcal{R}_j(k)$ is disconnected $\forall\ i \neq j$.

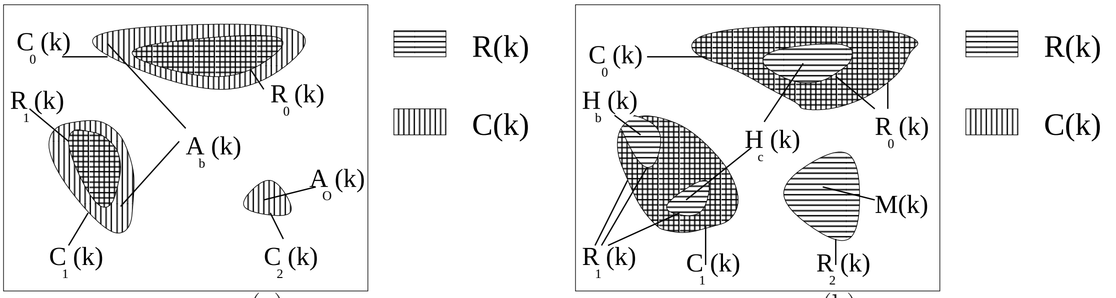
We also indicate $R(k)$ as the set of all the j regions of interest (*objects*) belonging to the reference segmentation, that can be expressed as:

$$R(k) = \bigcup_{0 \le j < J} R_j(k) \quad \text{and} \quad \bigcap_{0 \le j < J} R_j(k) = \emptyset \tag{4.1}$$

where J is the number of reference segmentation objects. J can also take the value zero when no object is present in the reference segmentation. Similarly, the set of pixels segmented at frame k, $C(k)$ is the union of the i regions/objects $C_i(k)$:

$$C(k) = \bigcup_{0 \le i < I} C_i(k) \quad \text{and} \quad \bigcap_{0 \le i < I} C_i(k) = \emptyset \tag{4.2}$$

where I is the number of resulting segmentation regions/objects. In the case I is zero, no region has been segmented in the resulting segmentation.



(a) (b)

Figure 4.9: Reference segmentation overlapped with the resulting segmentation, at frame k: (a) shows the two kinds of subsets of false positives, $\mathcal{P}(k)$, (b) shows the two kinds of subsets of false negatives, $\mathcal{N}(k)$.

The set of *false positive* pixels, $\mathcal{P}(k)$, whose elements are the segmented pixels not belonging to the reference segmentation, can be expressed as:

$$\mathcal{P}(k) = C(k) \cap R'(k) \tag{4.3}$$

where $R'(k)$ denotes the complement of $R(k)$. Similarly, *false negatives* $\mathcal{N}(k)$ appearing in the reference segmentation $R(k)$ and not in the resulting segmentation $C(k)$, can be expressed as:

$$\mathcal{N}(k) = C'(k) \cap R(k) \tag{4.4}$$

A further investigation of the segmentation errors has been carried out. In the following equations, let us define the condition empty intersection $\gamma_{i,j}(k)$ between the $j-th$ object in the reference segmentation and the $i-th$ region in the resulting segmentation as:

$$\gamma_{i,j}(k) = \begin{cases} 1 & if\ \big(C_i(k) \cap R_j(k) = \emptyset\big) \\ 0 & \text{otherwise} \end{cases}$$

The different errors have been mathematically expressed in Eqs. (4.5)-(4.10) and depicted in Figures 4.9 (a) and (b). $\mathcal{P}(k)$ can be divided into two different kinds of subsets: *added background* and *added regions*. The added region set, $\mathcal{A}_r(k)$ is a set of regions in $C(k)$ not present in $R(k)$:

$$\mathcal{A}_r(k) = \bigcup_{i \in Q} C_i(k), \tag{4.5}$$

where $Q = \{i \mid \gamma_{i,j}(k) = 1,\ 0 \leq j < J\ \}$. In the following, let $|\boldsymbol{A}_r(k)|$ denote the cardinality of $\mathcal{A}_r(k)$. $|\boldsymbol{A}_r(k)|$ therefore represents the number of added region pixels at frame k.

Added background $\mathcal{A}_b(k)$ does not constitute a region itself in $C(k)$ but it is a set of false positive pixels erroneously segmented along the boundary of an object which is an object both in $C(k)$ and $R(k)$. $\mathcal{A}_b(k)$ therefore is composed of those pixels that do not satisfy condition in Eq.(4.5.1) and are subsets of $\mathcal{P}(k)$:

$$\mathcal{A}_b(k) = \mathcal{P}(k) \setminus \mathcal{A}_r(k) \tag{4.6}$$

where $\setminus$ denotes a set difference. Let $|\boldsymbol{A}_b(k)|$ denote the cardinality of $\mathcal{A}_b(k)$ that is the total amount of added pixels.

Different classes of sets, depending on the properties of their elements, can also be distinguished inside $\mathcal{N}(k)$. *Missing objects* $\mathcal{M}(k)$ are objects in $R(k)$ not present in $C(k)$:

$$\mathcal{M}(k) = \bigcup_{j \in S} R_j(k) \tag{4.7}$$

where $S = \{j \mid \gamma_{i,j}(k) = 1,\ 0 \leq i < I\ \}$. Holes $\mathcal{H}(k)$ are sets of $\mathcal{N}(k)$ that intersect the reference segmentation and do not satisfy condition in Eq.(4.5.1):

$$\mathcal{H}(k) = \mathcal{N}(k) \setminus \mathcal{M}(k). \tag{4.8}$$

In $\mathcal{H}(k)$ we can differentiate between *holes inside* the object, $\mathcal{H}_i(k)$, and *boundary holes*, $\mathcal{H}_b(k)$, situated on the border of the object. $\mathcal{H}_i(k)$, are sets of those false negative pixels completely inside the objects and satisfy the following condition:

$$\mathcal{H}_c(k) \subset cl\big(R(k)\big) \tag{4.9}$$

where $cl(\cdot)$ is the set closure operator. In the following sections, the total amount of pixels in $\mathcal{H}_c(k)$ is denoted by $|\boldsymbol{H}_c(k)|$.

Boundary holes are sets of false negative pixels that intersect the boundary of the reference object and modify the shape:

$$\mathcal{H}_b(k)\ \cap\ \partial R_j(k) \neq \emptyset \tag{4.10}$$

where ∂ is the boundary set operator. Let $|\boldsymbol{H}_b(k)|$ denote the cardinality of $\mathcal{H}_b(k)$ that is the total amount of pixels of boundary holes.

In the proposed approach, to study and analyze the different impact on the perceptual quality of different artifacts, we consider the above defined specific artifacts that well represent all the segmentation errors: $\mathcal{A}_r(k)$, $\mathcal{A}_b(k)$, $\mathcal{H}_b(k)$ and $\mathcal{H}_i(k)$. The single, pure artifacts and their combination are introduced in test sequences as described in Sec. 4.5.2 and their effect on the overall perceived quality will be investigated by means of subjective experiments (see Sec. 5.4). The objective measures of these artifacts $|\boldsymbol{A}_r(k)|$, $|\boldsymbol{A}_b(k)|$, $|\boldsymbol{H}_b(k)|$ and $|\boldsymbol{H}_i(k)|$ are perceptually weighted to contribute in a perceptual objective measure for segmentation evaluation, as presented in the next chapter.

4.5.2 Generation of Synthetic Segmentation Errors and Test Sequences

Another important step in designing the subjective experiment is to choose a set of original video sequences to be used. A total of four video sequences of assumed high quality are used in this work: 'Hall monitor', 'Group', 'Highway' and 'Coastguard'. In these video sequences we selected 60 frame slots to obtain five seconds long video sequences (12 *fps*). They are in 4:2:0 YUV format, 288 lines × 352 columns. These video sequences are commonly used in the research community to test segmentation algorithms. 'Hall monitor' and 'Coastguard' are MPEG-4 video sequences, 'Group' is an European IST project *Art.live*[*] sequence and 'Highways' is an MPEG-7 test sequence. Representative frames of video sequences used are shown in Figure 4.10.

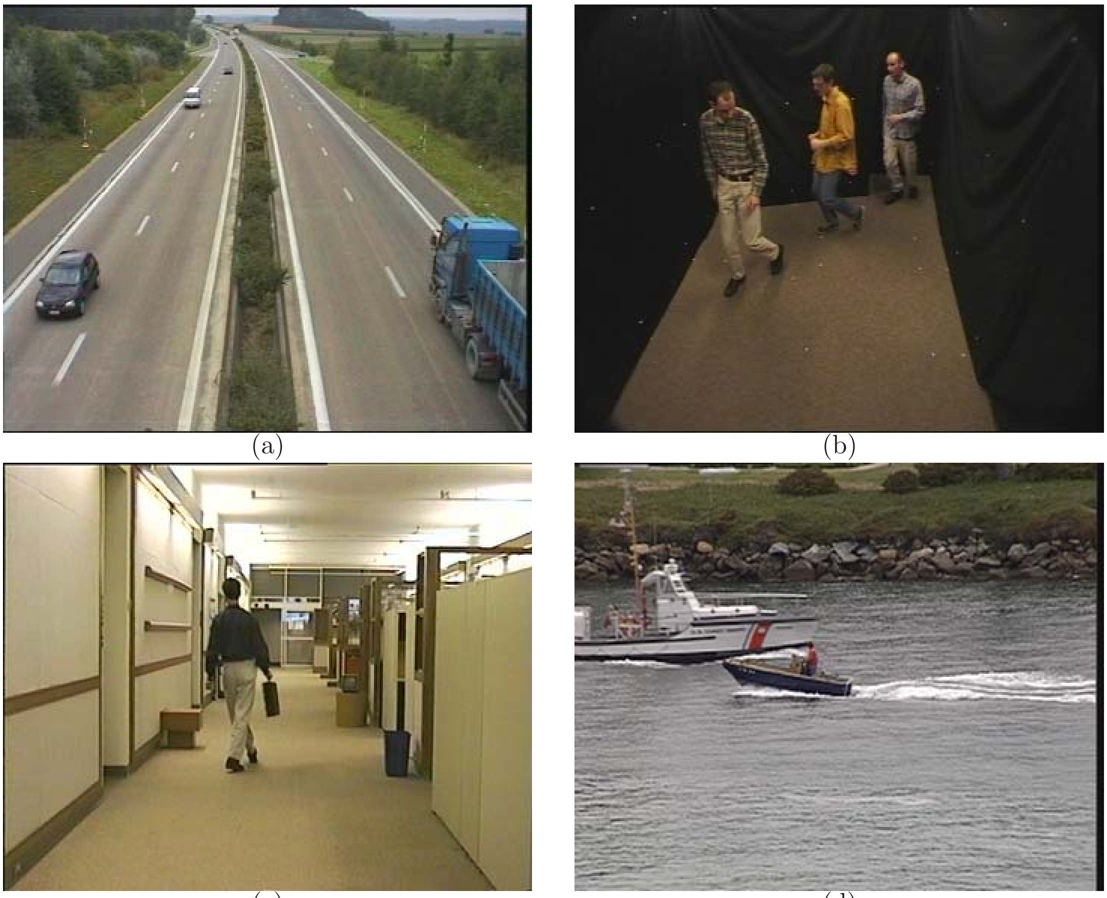
Since only a limited number of sequences can be shown during a 30-minute test session, the total number of originals is kept small. Table 4.3 shows the original used for each experiment.

The second step is to introduce the artifact in the test sequences, by modifying the ideally segmented reference masks, the ground truths. For two of the sequences ('Group' and 'Highway') the reference masks were obtained manually. For the other two sequences, they were obtained from the MPEG website[†].

---

[*] http://www.tele.ucl.ac.be/PROJECTS/art.live/

[†] http://mpeg.telecomitalialab.com



(a) (b) (c) (d)

Figure 4.10: Sample frames of original video sequences: (a) ‘Highway’, (b) ‘Group’, (c) ‘Hall monitor’, (d) ‘Coastguard’.

Table 4.3: Summary of original video sequences used in subjective experiments.

Video	**frames**	**Add reg.**	**Add back.**	**Border Hol.**	**Inside Hol.**	**Flick.**	**Expect.**	**Comb.**
’Coastguard’	1-60	√		√	√		√	√
’Group’	81-140	√		√	√			√
’Hall monitor’	32-91	√	√	√	√	√	√	√
’Highway’	66-125	√		√	√			√

Synthetic spatial errors

The results of segmented images or video sequences can be thought as being made of a combination of the reference segmentations and some errors. In this work, we introduce segmentation errors that are relatively pure and study them both individually as well as combined, for an assessment of their perceptual contribution. Four different kinds of *spatial errors* have been synthesized and combined to the reference segmentations: added background $\mathcal{A}_b(k)$, added regions $\mathcal{A}_r(k)$, holes inside $\mathcal{H}_i(k)$, and boundary holes $\mathcal{H}_b(k)$. Sample frames of the generated spatial artifacts are shown in Appendix A.

The annoyance produced by *added region* artifacts, $\mathcal{A}_r$, was studied by varying their size, position and shape. We artificially mis-segmented three portions of the background completely disconnected from the correctly segmented foreground objects. In a first experiment, the impact of the shape and the position of the added regions with the same size was under investigation. Therefore, we kept the number of regions equal to three and used four different size values, and we varied the position and shape of the artifact for each test sequence. The shape of the *added region* was modeled using a super-ellipsis function. By modifying the super-ellipse parameters, a continuum of several shapes can be formed, ranging from a ellipse to a rectangle. The topology of the reference segmentation was varied in the following way. First, we positioned the group of three added regions in three different random positions (p_1, p_2 and p_3) going from very far from the reference objects to closer. Then, for each of these positions, two different shapes (square and circles) were generated with four different sizes, $|\boldsymbol{A}_r|$ (2×2, 5×5, 10×10, 20×20). The total number of test sequences for this part of the experiment was 75, which included 72 test sequences (3 reference segmentations × 3 positions × 4 sizes × 2 shapes) plus the 3 reference segmentations without any artifact of 'Hall monitor', 'Highway' and 'Group'.

The *added background* test sequence was synthetically generated by adding increasingly more background to the R_j objects. By dilating the reference mask, five levels of dilation were generated. Then the number of pixels added at each frame was c, 3c, 4c, 5c, and 8c, where c is the number of pixels on the ground truth contours. Therefore, five values of *added background* $|\boldsymbol{A}_b|$ were investigated in the experiment and inserted in one reference segmentation, 'Hall monitor'.

In the objective metrics proposed in the literature, holes are only considered in terms of uncorrelated set of pixels and their distances from the reference boundary of the object [17, 23]. According to [86] the more distant a hole is from the boundary of the object, the more annoying the artifact becomes. It has been concluded that as one moves away from the border, holes become more annoying. Boundary holes only make the object thinner. Therefore, they are less annoying for the human observer than inside holes.

In our experiment, we studied if this condition is still valid for large holes. In this case the annoyance caused by a boundary hole could be worse than for a closed hole (completely inside the object). This could be justified by the fact that if the shape of the object is completely modified by a large hole on the boundary, the object can become harder to recognize. On the other hand, in the presence of a large closed hole completely inside an object, the object can be still recognizable and, consequently, this artifact becomes less annoying. For this purpose, we synthetically inserted a group of three holes at three positions: on the contour of the object (boundary hole), and in two inner positions (inside holes). For each position, we generated 4 sizes (3×3, 5×5, 9×9, 13×13) of holes. The total number of test sequences for this part of the experiment was 52 which included 48 test sequences (4 reference segmentations × 3 positions × 4 sizes) plus the 4 original reference segmentations of 'Hall monitor', 'Highway', ' Coastguard' and 'Group'.

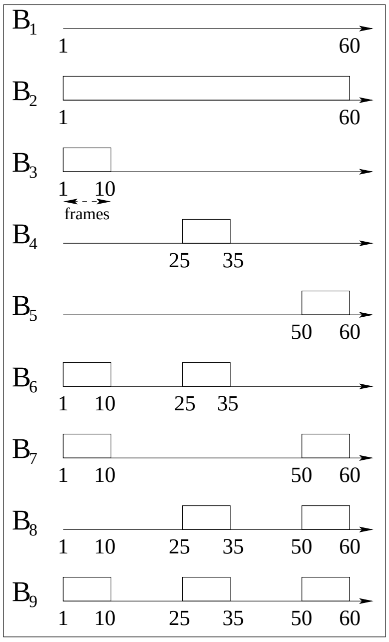


Figure 4.11: Temporal insertion of artifacts during 10 frames in different moments of the video sequence.

Synthetic temporal errors

Since a video is a sequence of images in which spatial errors take place, the temporal effect of segmentation errors must be considered. A given error may be perceived differently, depending on its temporal context. Observers are sensitive to *temporal errors*, i.e., changes in error characteristics along time. In video segmentation, an error may vary its characteristics through time. A non smooth change of any spatial error deteriorates the perception of the error itself. The temporal artifact caused by a variation of the spatial error is called *flickering*. By carrying out subjective tests on real segmentation, flickering has been observed to be one of the most annoying artifacts introduced by segmentation algorithms. In fact, if an imprecise segmentation mask is stable along time, it is perceived less annoying than a more precise segmentation presenting abrupt changes. We performed two kinds of experiments with temporal errors. We tested the *flickering* and the effect produced by a bad (or good) segmentation at the end (or beginning) of segmented video sequences.

In the first experiment, different variations of any spatial error could be implemented to test the flickering perception. We chose to change the position of added regions along the test sequence. The test video sequences with the temporal errors presented the same number of added regions with the same shape and size. But their positions changed every 1, 3, 5, 12 and 30 frames (let $\boldsymbol{f_T}$ denote the flickering period) by starting from a very fast and annoying flickering, and by ending with a temporally smooth change of added region position.

A second experiment on the temporal perception of artifacts was performed. In this experiment, we wanted to find whether there is an *expectation* effect and how this affects the overall perceived quality. By *expectation* we mean the effect that a good segmentation at the beginning could create a good overall impression on assessing the quality of the sequences under test and vice versa. Three regions of the same size (10×10) were added always at the same position along the entire video sequence. The added regions appeared and disappeared along the time causing a temporal artifact.

Table 4.4: Tested segmentation artifacts and their values.

Tested Artifacts	**Amount**
Added region	2×2, 5×5, 10×10, 20×20.
Added background	c, 3c, 4c, 5c, 8c
Border hole	3×3, 5×5, 9×9, 13×13
Inside hole	3×3, 5×5, 9×9, 13×13
Flickering period	1, 3, 5, 12, 30
Expectation	$\mathbf{B}_1$, $\mathbf{B}_2$, $\mathbf{B}_3$, $\mathbf{B}_4$, $\mathbf{B}_5$, $\mathbf{B}_6$, $\mathbf{B}_7$, $\mathbf{B}_8$, $\mathbf{B}_9$

This temporal artifact can be expressed by an indicator function $\mathbf{B}(t_1, t_2)$ whose value is 1 for t in the interval $[t_1, t_2]$ and zero otherwise.

Figure 4.11 shows an illustration of how added regions were inserted in the sequences in order to create the temporal artifacts. Condition $\mathbf{B_1}$ corresponds to the reference sequence, while condition $\mathbf{B}_2$ corresponds to a sequence with the added regions present in all 60 frames. Conditions $\mathbf{B}_3 - \mathbf{B}_5$ are cases where the added regions were inserted in 10 out of 60 frames. They were inserted in different parts of the video sequence: at the beginning ($B(1, 10)$), at the end ($B(50, 60)$), and in the middle ($B(25, 35)$). Conditions $\mathbf{B}_6 - \mathbf{B}_9$ correspond to combinations of these three previous occurrences. A total of 9 test conditions and two test video sequences 'Hall monitor' and 'Coastguard', were used, which means 20 test sequences (2 reference segmentations × 9 conditions) plus 2 reference segmentations.

The spatial and temporal errors and their values are summarized in Table 4.4. Sample frames of the generated synthetic artifacts are given in Appendix A.

Synthetic combined errors

The last experiment was performed to understand how artifacts combine and interact to produce the overall annoyance. This experiment takes into account the interaction of different artifacts and the artifact perception is more complicated to model. The first step generated segmentations with one type of artifact at a relatively high level of annoyance. Three synthetic added regions, added background, boundary holes and inside holes were created. They were added at different amounts as summarized in Table 4.5. For each reference segmentation, 12 test sequences (3 amounts × 4 artifacts) were created: with only added region artifacts $\mathcal{A}_r$ (at low, medium and high level of annoyance), with only added background $\mathcal{A}_b$ (at low, medium and high level of annoyance), with only boundary holes $\mathcal{H}_b$ (at low, medium and high level of annoyance) and with only inside holes $\mathcal{H}_i$ (at low, medium and high level of annoyance). Then, an impaired video was created by varying the combinations of these 12 test sequences, as given by the following equation:

$$\mathcal{S} = (R \backslash \mathcal{H}) \cup \mathcal{A} \tag{4.11}$$

where $\mathcal{S}$ is the segmentation under test, R the reference segmentation, $\mathcal{H}$ the hole artifact, $\mathcal{A}$ the added artifact. The different amounts of artifacts used, $|\boldsymbol{A}_r(k)|$, $|\boldsymbol{A}_b(k)|$, $|\boldsymbol{H}_b(k)|$, $|\boldsymbol{H}_i(k)|$ are indicated in Table 4.5. Depending on the amount of the artifact the appearance of the overall impairment on the segmentation changed, making it more over-segmented or under-segmented. The 45 combinations of $|\boldsymbol{A}_r(k)|$, $|\boldsymbol{A}_b(k)|$, $|\boldsymbol{H}_b(k)|$, $|\boldsymbol{H}_i(k)|$ values used to generate the test sequences are shown in columns 2-5 of Table 4.5. We did not use all possible combinations of the four artifacts since it would have made the experiments too long.

Table 4.5: Set of amount combinations (pixels) for added regions, added background, inside holes and border holes used in the experiment.

combination	Added region, $\|\mathcal{A}_r\|$	Added background, $\|\mathcal{A}_b\|$	Inside hole,$\|\mathcal{H}_i\|$	Border hole, $\|\mathcal{H}_b\|$
1	0	0	0	0
2	180	0	0	0
3	0	180	0	0
4	0	0	180	0
5	0	0	0	180
6	60	0	0	0
7	0	60	0	0
8	0	0	60	0
9	0	0	0	60
10	24	0	0	0
11	0	24	0	0
12	0	0	24	0
13	0	0	0	24
14	60	60	60	60
15	24	24	24	24
16	60	60	0	0
17	0	60	60	0
18	0	0	60	60
19	60	0	0	60
20	60	0	60	0
21	0	60	0	60
22	180	0	24	0
23	0	180	0	24
24	24	0	180	0
25	0	24	0	180
26	180	24	24	0
27	0	180	24	24
28	24	0	180	24
29	24	24	0	180
30	180	0	60	60
31	60	180	0	60
32	60	60	180	0
33	0	60	60	180
34	60	60	60	180
35	180	60	60	60
36	60	180	60	60
37	60	60	180	60
38	24	180	24	24
39	180	24	24	24
40	24	180	24	24
41	24	24	24	180
42	180	24	180	60
43	60	180	24	180
44	180	60	180	24
45	24	180	60	180

As stated, four original video sequences were used in this experiment: ‘Hall monitor’, ‘Group’, ‘Highway’ and ‘Coastguard’. A total of 180 test video sequences were derived from these original video sequences (4 segmentation × 45 combinations) to investigate the relationship between the overall annoyance and the artifact strengths.

4.6 Conclusions

In this chapter, we have underlined why a good image/video segmentation is important. In particular, we have focused on video object segmentation. The task of extracting an object from video sequences has been illustrated for applications such as object-based coding and video object indexing and retrieval.

The state of the art methodologies for psychophysical experiments of segmentation quality assessment have been described both for still images and video sequences. In addition, the advantages and disadvantages of each approach have been presented. On the basis of this discussion a new methodology for subjective evaluation of video object segmentation has been proposed. The method of single stimulus continous annoyance scale has been chosen as the basis for our experiments as a result of many informal tests and discussions with psychophysics experts.

The experiment setup and the instructions given to the subjects have been described. The five stages of the proposed methodology have been illustrated in details: oral instructions, training, practice trials, experimental trials and interview.

The proposed approach to obtain a *perceptual* objective metric from the subjective experiments has been presented. The process of generation of test sequences has been introduced. The synthetic test sequences developed for our experiments have been described. Spatial and temporal artifacts commonly found in video object segmentation have been inserted in the test sequences. The investigated artifacts are: added regions, added background, border holes, inside holes, flickering and expectation effect. Moreover, combinations of all artifacts have been created to study how they interact in the overall annoyance.

In the next chapter, the psychophysical experiments carried out with the described test sequences are presented and their results analyzed. On the basis of these results, the *perceptual* objective metric, for segmentation evaluation, proposed in this thesis will be introduced.

Objective Segmentation Evaluation

5

5.1 Introduction

In the previous chapter, we presented the proposed experimental method for performing subjective experiments. In these experiments we aim at investigating the annoyance and the perceived strength of typical segmentation artifacts. We explained how we generated the synthetic test sequences with typical artifacts under study: added regions, added background, inside and border holes.

In this chapter, our goal is to find, by means of subjective experiments, a *perceptual objective metric*. With various subjective experiments, we show how to derive the psychometric curves for each artifact metric. Since different strengths of artifacts contribute differently to the overall annoyance, subjective experiments are performed to find out how perceptual artifacts combine to produce the overall annoyance. On the basis of this last experiment, a perceptual objective metric that combines all the perceptual artifact metrics is introduced in this chapter. The results are shown and a comparison among the state of the art objective methods is presented.

The objective metrics found in the literature for segmentation evaluation are described in Sec. 5.2. A new objective metric is proposed and described in Sec. 5.3. Section 5.4 describes the results of the subjective experiments and presents the perceptual artifact metrics. Section 5.5 discusses the overall perceptual objective metric and its performance, also compared to the state of the art methods. Section 5.6 draws the conclusions.

5.2 Objective Segmentation Evaluation Methods - Background

To avoid systematic subjective evaluation of segmentation, an automatic procedure is preferred. This procedure is referred to as *objective segmentation evaluation method*. Quality metrics for objective evaluation of segmentation may judge either the segmentation algorithms or their segmentation results. These are referred to as analytical or empirical methods, respectively [186]. *Analytical methods* evaluate segmentation algorithms by considering their principles, their requirements and their complexities. The advantage of these methods is that an evaluation is obtained without implementing the algorithms. However, the properties of algorithms, such as the computational cost, could change quickly if alternative architectures and implementations are carried out. Moreover, because of the lack of a general theory for image segmentation, and because segmentation algorithms may be complex systems composed of several components, not all properties (and therefore strengths) of segmentation algorithms may be easily evaluated. *Empirical methods*, on the other hand, do not evaluate the segmentation algorithms directly, but indirectly through their results. Empirical methods are further divided into *empirical discrepancy*, metrics when the segmentation result is compared to an ideally segmented reference map (ground truth), and *empirical goodness* metrics, when the quality of the segmentation result is based on intuitive measures of goodness such as gray-level or color uniformity, shape regularity or contrast between regions. The advantage of this second class of methods is that it requires only that the user defines a goodness metric. Therefore, they do not require manually segmented images to be supplied as ground truth data. In addition, they can be used in an on-line manner, so that the effectiveness of an algorithm can be monitored during actual application. A major disadvantage of such an approach is that the goodness metrics are at best heuristic, and may exhibit strong bias toward a particular algorithm. For example the intra-region gray-level uniformity goodness metric will cause any segmentation algorithm which forms regions of uniform texture to be evaluated poorly.

Even if goodness methods are less complex than discrepancy methods (they do not require a manual segmentation) and they can be used for on-line evaluation, for algorithms performance comparisons, discrepancy methods can be applied on a predefined data set (they require a manual segmentation). Analogous to the case of the *empirical goodness* methods, a discrepancy measure must be explicitly defined, but this is likely to be easier to do and exhibits less bias when compared to former methods because of the availability of a ground truth. In image compression, the disparity between the original image and the decoded image has often been used to objectively assess the performance of the compression algorithms. A commonly used discrepancy measure is the mean-square signal-to-noise ratio [143]. However, in contrast to image encoding, image segmentation is a process that changes the image units. In other words, image encoding is an image processing process, while image segmentation is an image analysis process, in which the input and the output are different matters. So other specific discrepancy measures have been proposed in the field.

To properly evaluate the performance of segmentation techniques, objective methods have been proposed both for image and video segmentation. They are described respectively in Secs. 5.2.1 and 5.2.2, respectively.

5.2.1 Still Images

Some works dealing with image segmentation assessment have been reported in the literature for evaluating still image segmentation [15, 46, 55, 91, 99, 131, 183]. More attention has been dedicated in the past several years to evaluate edge detection algorithms [32, 49, 184]. However, these techniques are specific for edge detection and cannot be directly applied to video object segmentation evaluation.

In [188], an extensive survey of existing methods for evaluating image segmentation has been published. Goodness methods for image segmentation evaluation have been proposed, among others, by Borsotti *et al.* [15] and Rosenberger *et al.* [131], where metrics for intra-object homogeneity and inter-object disparity are proposed. Borsotti adds to its evaluation measure a penalization related to the number of objects in the segmented image (the quality is lowered in case of over segmentation).

Although goodness evaluation methods can be very useful for on-line evaluation, their results do not necessarily coincide with human perception of the goodness of segmentation. In fact, as mentioned above, in the case of an image presenting uniformly textured regions, an intra-region gray-level uniformity metric would not provide the same results when compared to subjective evaluations. For this reason, when a reference mask is available or can be generated, discrepancy evaluation methods are preferred.

A discrepancy method based on shape features (circularity and elongation), under-merging and over-merging error is proposed in [183]. Recent discrepancy evaluation methods include compactness evaluation of both under and over detected pixels [55] and a discrepancy approach based on spatial accuracy [99] which was originally proposed for the evaluation of video object segmentation by Villegas *et al.* [86].

A potential problem for a measure of consistency between segmentations of still images is that there is no unique segmentation of an image, as mentioned in Sec. 4.3.1. Martin *et al.* [91] proposed two error measures that take small values when one segmentation is simply a refinement of the other. In addition, the two measures are independent from the coarseness of pixelation, robust to noise along region boundaries and tolerant to different number of segments. Then, the authors validate the proposed error measures on a database of manual segmentations. It is shown that the error measures between different segmentations of the same image are low since they are the simple refinement of the each other, while the errors between segmentations of different images are high.

An interesting approach is described in [46], where a measure incorporating multiple measures –both discrepancy and analytical– (e.g. the segmentation accuracy versus the execution time) performs the evaluation in a multi-dimensional fitness/cost space.

According to [49], in order to compare the performance of segmentation algorithms, it is not sufficient to check whether the order proposed by the objective measure coincides with the order proposed by human observers. To be reliable, an error measure has to prove to have no drawbacks in which the measure shows a very important difficulty or anomaly. The authors adopt a simple analysis of quality curves to detect, if in any practical situations, five classical evaluation discrepancy measures show a significant difficulty or anomaly. In particular, they check the quality curves whether they have similar values corresponding to very distant thresholds for ten edge segmentation algorithms under test. If the measure presents such anomaly for one edge detector it is discarded from the comparison of algorithms.

In medical image segmentation other metrics are commonly used in quantitative evaluation. Alonso [2] *et al.* proposed a method incorporating specific metrics that compare the medical image segmentation to the ground truth, such as the preservation of the mean, the standard deviation, the perimeter and the area.

Finally, another way to define an evaluation measure is by considering it as a particular case of shape similarity metric, a problem with a long tradition in the pattern recognition literature [185].

5.2.2 Video Object Segmentation

Although several quality measures have been developed for still image segmentation they are not directly applicable to video object segmentation. In this section we will present the state of the art evaluation metrics for video object segmentation [17, 23, 43, 44, 97, 123, 154, 180] and video object tracking evaluation [11, 104, 106, 110, 135]. In particular we will provide the details of three methods [104, 154, 180] that will be used in the comparison between the proposed metric and the state of the art methods.

We distinguish between video object segmentation and tracking evaluation since they are two different matters. Video tracking is the process of locating a moving object (or several ones) in time using a camera. An algorithm analyzes the video frames and outputs the location, optionally in real time. It is mainly used in video surveillance systems. The issues involved in video object

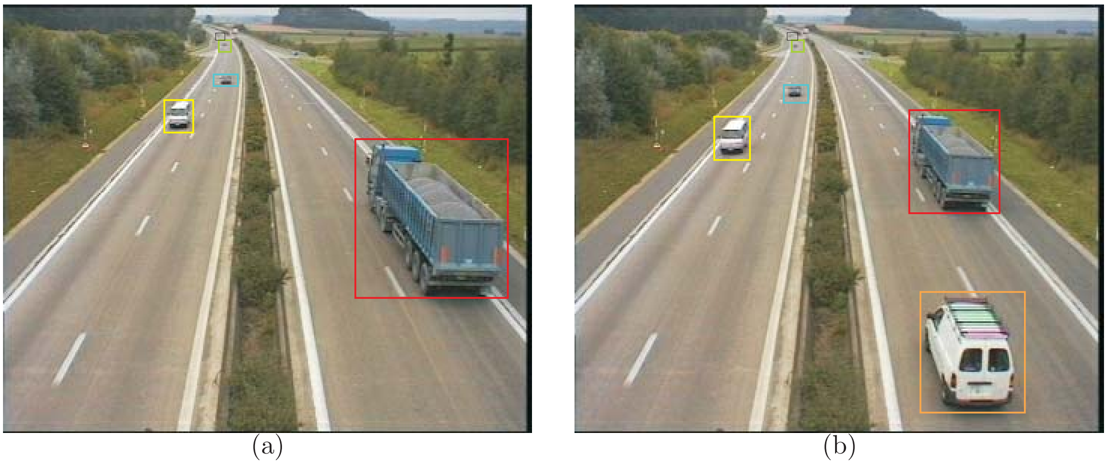


Figure 5.1: Samples of ground truth for tracking evaluation through bounding box. ‘Highway’ video sequence: (a) frame #89, (b) frame #101.

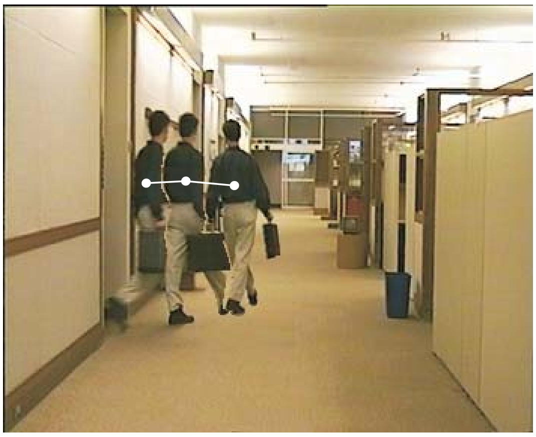


Figure 5.2: Sample of ground truth for tracking evaluation through center of gravity.

tracking are different from those of video object segmentation evaluation since the ground truth on which these algorithms compare their performance is different. In fact, video surveillance systems concern algorithms for detecting, indexing and tracking moving objects and the system has to be characterized and evaluated in other ways. The ideal output (ground truth) of a tracking system can be of two types: bounding box and/or center of gravity. In the former case, regions that contain the detected moving objects of interest are segmented with a set of rectangular areas called bounding boxes as shown in Fig. 5.1. Detection and false alarms rates in this case are derived by counting how many times interesting and irrelevant regions are detected. In the latter, the manual ground truth consists in a set of points that define the trajectory of each object in the video sequence (center of gravity) as depicted in Fig. 5.2. In this case, the motion detection and tracking algorithm is then run on the video sequence and tracking results and ground truth centers of gravity are compared to assess tracking performance. Figure 5.3 (a) depicts the original frame and (b) shows the result of a ideal video object segmentation - ground truth. As depicted, it does not represent a binary detection problem. Several types of errors (such as shape errors along the boundaries of the object, content similarity, etc,.) should be considered (not just mis-detection and false alarms). Thus, proposed tests based on the selection of rectangular regions with and without objects are unrealistic since

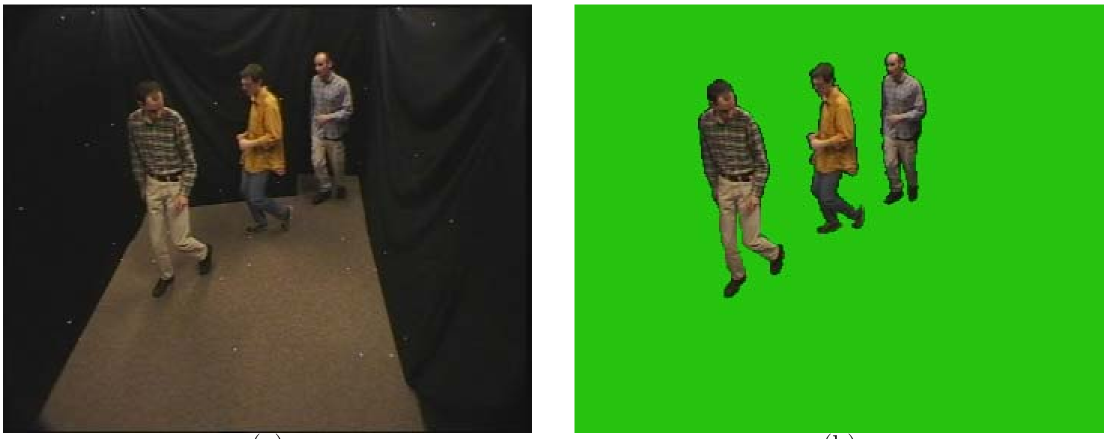

(a) (b)

Figure 5.3: Sample of original video sequence 'Group' in (a) and the corresponding ideal object segmentation (ground truth) in (b).

practical segmentation algorithms have to segment the image into *foreground* –objects of interest– and *background*, and do not have to classify rectangular regions selected by the user.

First, we will present some techniques for tracking evaluation and then we will provide an overview on object segmentation evaluation which is more relevant to the topic of this part of the thesis.

Tracking evaluation criteria

Recently, a number of measures have been proposed for video object tracking evaluation. Since we are interested in how the object is segmented and the evaluation of tracking raises different problems briefly discussed in this section, the reader is introduced to fora such as PETS [155] and CAVIAR [125] for a complete overview on that issue.

Table 5.1: Objective Measures used in evaluating video tracking systems.

Method	Measure	Source
Discrepancy	False Alarm	Ellis [11], Nascimento [104], Oliveira [110], Oberti [106]
Discrepancy	Misdetection	Ellis [11], Nascimento [104], Oliveira [110], Oberti [106]
Discrepancy	Split and/or Merge	Ellis [11], Nascimento [104], Oberti [106]
Discrepancy	Area Matching	Ellis [11], Nascimento [104]
Discrepancy	Occlusion management	Ellis [11]
Discrepancy	Center of gravity	Ellis [11], Senior [135]

In the following, we will refer to some representative works [11, 106, 110, 135] that can be found in the literature and specifically to Nascimento and Marques's metric [104] that can be applied also to a more general object segmentation evaluation case. Table 5.1 shows all the state of the art methods grouped by discrepancy measure.

Standard measures used in communication theory such as mis-detection rate, false alarm rate and Receiver Operating Characteristics (ROC) are used in [106, 110]. An ROC curve is generated by computing pairs (P_d, P_f), where P_d is the probability of correct signal detection and P_f is the false alarm probability. For example, Oberti *et al.* [106] compute the false-alarm (P_f) and the mis-detection probabilities ($1 - P_d$) on the basis of discrepancies between the resulting objects and matching area (false alarm) or between the reference area and the matching one (mis-detection). The global performance curve summarizing the curves obtained under different working conditions is obtained by imposing an operating condition ($P_f = 1 - P_d$) and by plotting the corresponding values against different values of the variable of interest (scene complexity, distance of objects from sensors).

In [110], a specific parameter of the tracking algorithm is varied and the false alarm/detection and split/merge rates are plotted against it. Senior *et al.* [135] employed the trajectories of the centroids of tracked objects and their velocities to evaluate their discrepancy measures.

An interesting framework for tracking performance evaluation uses pseudo-synthetic video [11]. Isolated ground truth tracks are automatically selected from the PETS2001 dataset, according to three criteria: path, color and shape coherence (in order to remove tracks of poor quality). Pseudo-synthetic video are generated by adding more ground truth tracks and the complex object interactions are controlled by the tuning of perceptual parameters. The metrics used are similar to those in the previously described works: tracker detection rate, false alarm rate, track detection rate, occlusion success rate, etc.

However these approaches have several limitations. As already mentioned, object detection can not be considered as a simple binary detection problem. Several types of error should be considered and just mis-detection and false alarms are not enough. For example, the proposed test in [135] is based on employing the centroid and areas of rectangular regions but practical algorithms have to segment the image into background and foreground and do not have to classify rectangular regions selected by the user.

To overcome these limitation Nascimento and Marques [104] used several simple discrepancy metrics to classify the errors into region splitting, merging or split-merge, detection failures and false alarms. In this scenario, the most important thing is that all the objects have to be detected and tracked along time. Object matching is performed by computing a binary correspondence matrix between the segmented and the ground truth images. The advantage of this method is that ambiguous segmentations are considered (e.g., it is not always possible to know if two close objects correspond to a single group or a pair of disjoint regions: both interpretations are adopted in such cases). In fact, by analyzing this correspondence matrix, the following measures are computed: Correct Detection (CD): the detected region matches one and only one region; False Alarm (FA): the detected region has no correspondence; Detection Failure (DF): the test region has no correspondence; Merge Region (M): the detected region is associated to several test regions; Split Region (S): the test region is associated to several detected regions; Split-Merge Region (SM): when the conditions M and S simultaneously occur.

The normalized measures are obtained by normalizing the amount of FA by the number of objects in the segmentation, N_C, and all the others by the number of objects in the reference, N_R, and finally by multiplying the obtained numbers by 100. The **object matching quality metric** at frame k, $mqm(k)$, is finally given by:

$$mqm(k) = w_1 \cdot \frac{CD(k)}{N_R} + w_2 \cdot \frac{FA(k)}{N_C} + w_3 \cdot \frac{DF(k)}{N_R} + w_4 \cdot \frac{M(k)}{N_R} + w_5 \cdot \frac{S(k)}{N_R} + w_6 \cdot \frac{SM(k)}{N_R} \quad (5.1)$$

where w_i are the weights for the different discrepancy metrics. It is evident that this metric is able to describe quantitatively the correct number of detected objects and their correspondence with the ground truth only while the metrics described in the next section are able to monitor intrinsic properties of the segmented objects such as shape irregularities and temporal instability of the mask along time.

Video object segmentation evaluation criteria

Empirical goodness methods have been defined not only for still image but also for video in [23, 44]. In [23], goodness metrics are developed and grouped into two classes: *intra-object homogeneity* (shape regularity, spatial uniformity, temporal stability and motion uniformity) and *inter-object disparity* (local color and motion contrast with neighbors). The goodness metrics are all combined in a composite metric with weights differentiated according to the type of content (stable or moving content). Erdem *et al.* [44] utilized a *spatial color contrast measure*, *color histograms differences* along the temporal axis and *motion vector differences* along the boundaries of the segmented objects, all combined in a single performance measure. Piroddi *et al.* [123] improved Erdem's goodness method in terms of sensitivity as well as noise immunity.

To evaluate a video scene with segmented moving objects by means of discrepancy methods, Erdem and Sankur [43] combined three empirical discrepancy measures into an overall quality segmentation evaluation: *mis-classification penalty*, *shape penalty*, and *motion penalty*. In [23], Correia and Pereira first measured the individual segmentation quality through four spatial accuracy criteria: *shape fidelity*, *geometrical fidelity*, *edge and statistical content similarity* and two temporal criteria: *temporal perceptual information* and *criticality*. Second, they computed the similarity factor between the reference and the resulting segmentation. Furthermore, the multiple-object case was addressed by using the criteria of application-dependent "*object relevance*" [22] to provide the weights for the quality metric of each object. Finally, they combined all these three measures in an overall segmentation quality evaluation.

Another way to approach the problem is to consider it as a particular case of shape similarity as proposed in [97] for video object segmentation. In this method, the evaluation of the spatial accuracy and the temporal coherence is based on the mean and standard deviation of the 2-D shape estimation errors.

We proposed to evaluate the quality of the segmented object through spatial and temporal accuracy joined to yield a combined metric [17]. This work was based on the two other discrepancy methods [86, 180] described below.

During the standardization work of ISO/MPEG-4, within the core experiment on automatic segmentation of moving objects, it became necessary to compare the results of different proposed object segmentation algorithms, not only by subjective evaluation, but also by objective evaluation. The proposal for objective evaluation [180] agreed by the working group uses a ground truth in order to evaluate the segmentation results. This metric is usually adopted by the research community due also to its simplicity. A refinement of this metric has been proposed by Villegas *et al.* [86, 154]. These two metrics have been chosen as term of comparison for the new metric proposed in this thesis. Below, the descriptions of these two metrics are provided.

MPEG Evaluation Criteria

A moving object can be represented by a binary mask, called *object mask*, where a pixel has object-label if it is inside the object and background-label if it is outside the object. The objective evaluation approach used in the ISO/MPEG-4 core-experiment has two objective criteria: the *spatial accuracy* and the *temporal coherence*. Spatial accuracy is estimated through the amount of error pixels in the object mask (both false positive and false negative pixels) in the resulting mask deviating from the ideal mask (see Eq. (4.1)-(4.4)):

$$Sqm(k) = \frac{|\mathcal{P}(k)| + |\mathcal{N}(k)|}{|R(k)|}. \tag{5.2}$$

Temporal coherence is estimated by the difference of the spatial accuracy between the mask, M, at the current and previous frame,

$$Tqm_M(k) = Sqm(k) - Sqm(k-1). \tag{5.3}$$

The two evaluation criteria can be combined in a single **MPEG quality measure**, *MPEGqm(k)*, through the sum:

$$MPEGqm(k) = Sqm(k) + Tqm_M(k). \tag{5.4}$$

In this metric, the perceptual difference of different classes of errors, false positive and false negative, is not considered and they are all treated the same in Eq. (5.2). In fact, different kinds of errors should be combined in the metric in correct proportions to match evaluation results produced by human observers.

Weighted Evaluation Criteria

Within the project COST 211 [51] the above approach has been further developed by Villegas and Marichal [86, 154]. For the evaluation of the spatial accuracy, as opposed to the previous method, two classes of pixels are distinguished: those which have object-label in the resulting object mask, but not in the reference mask (false positive) and vice versa (false negative), and they are weighted differently. Furthermore, their metric takes into account the impact of the two classes (see Eqs. (4.3)- (4.4)) on the spatial accuracy, that is, the evaluation worsens with pixel distance d to the reference object contour. The spatial accuracy, qms, is normalized by the sum of the areas of reference objects as follows:

$$qms(k) = \frac{qms^+(k) + qms^-(k)}{\sum_{i=1}^{N_R} R_i(k)} = \frac{\sum_{d=1}^{D^+_{max}} w_+(d) \cdot |\mathcal{P}_d(k)| + \sum_{d=1}^{D^-_{max}} w_-(d) \cdot |\mathcal{N}_d(k)|}{\sum_{i=1}^{N_R} R_i(k)}, \tag{5.5}$$

where D^+_{max} and D^-_{max} are the biggest distance d for, respectively, false positive and false negative; N_R is the total number of objects in the reference R; $\sum_{i=1}^{N^R} R_i(k)$ is the sum of the area of all the objects i in the reference; $w_+(d)$ and $w_-(d)$ are the weighting functions for positive and negative respectively. They are expressed as:

$$w_+(d) = b_1 + \frac{b_2}{d + b_3}, \qquad w_-(d) = f_S \cdot d, \tag{5.6}$$

where the parameters b_i and f_S are fixed empirically [154]: $b_1 = 20$, $b_2 = -178.125$, $b_3 = 9.375$ and $f_S = 2$. These functions represent the fact that the weights for false negative pixels increase linearly and they are larger than those for false positives at the same distance from the border of the object as depicted in Fig. 5.4. In fact, as we move away from the border, missing parts of objects are more important than added background.

Then, in [86, 154], two criteria are used for estimating temporal coherence, the temporal stability $qmt(k)$ and the temporal drift $qmd(k)$ of the mask. First, the variation of spatial accuracy criterion between successive frames is investigated as follows, the temporal stability is equal to the normalized sum of the differences of the spatial accuracy into two consecutive frames for false positive and false negative pixels:

$$qmt(k) = \frac{|qms^+(k) - qms^+(k-1)| + |qms^-(k) - qms^-(k-1)|}{\sum_{i=1}^{N_R} R_i(k)}. \tag{5.7}$$

Second, the displacement of the gravity center, $\overrightarrow{GC}^{x,y}$, of the resulting object and the reference object mask is computed for successive frames to estimate the possible *drifts* of the object mask, $\overrightarrow{qmd}(k)$:

$$\overrightarrow{qmd}(k) = [\overrightarrow{GC}^{x,y}_E(k) - \overrightarrow{GC}^{x,y}_R(k)] - [\overrightarrow{GC}^{x,y}_E(k-1) - \overrightarrow{GC}^{x,y}_R(k-1)] \tag{5.8}$$

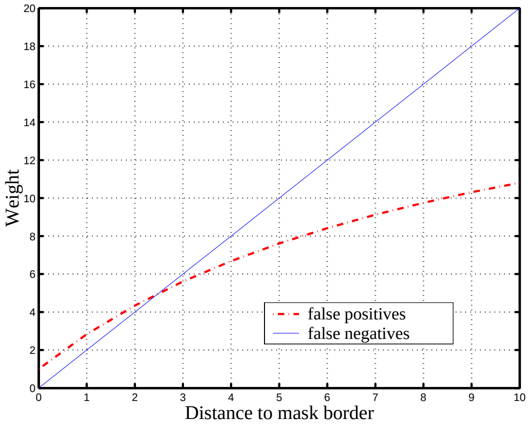


Figure 5.4: Weighting functions for false positives and false negatives for the method of Villegas *et al.* [154].

that is displacement from time $(k-1)$ to time (k) of the centers of gravity of the masks, GC. The value of drift is the norm of the displacement vector normalized by the sum of the reference object bounding boxes,

$$qmd(k) = \frac{||\overrightarrow{qmd}(k)||}{\frac{1}{N_R}\sum_{i=1}^{N_R} BB_i^{x,y}(k)}, \tag{5.9}$$

where $BB_i^{x,y}(k)$ is the bounding box of the object i in the reference mask R at time k. The authors proposed to define a single quality value by linearly combining all the three presented measures as the **weighted quality metric**, $wqm(k)$:

$$wqm(k) = w_1 \cdot qms(k) + w_2 \cdot qmt(k) + w_3 \cdot qmd(k). \tag{5.10}$$

The values of the weights w_i are extremely application dependent. If no application is specified all the three weights can be thought equal to $\frac{1}{3}$.

In this method, the perceptual difference between two kinds of errors is taken into account. The drawback is that the weighting functions defined in Eq. (5.6), that should be ‘perceptual’ weights of the evaluation criteria, are defined by means of empirical tests. These empirical tests are not generally sufficient to guarantee the definition of ‘perceptual’ weights. As well as in all other proposed evaluation criteria in the literature, the relevance and the corresponding weight of different kinds of errors should be supported by formal subjective experiments performed under clear and well defined specifications.

In Tab. 5.2 all the evaluation criteria for still and video sequences are summarized. Table 5.3 reports for each state of the art objective method all the evaluation criteria used in that specific method, whereas, Tab. 5.4 shows the objective methods grouped according to the evaluation criteria.

The averaging of the three quality metrics, **MPEGqm**, **wqm** and **mqm**, over a whole sequence of K frames $(k = 1, .., K)$ processed by the system under test, makes the evaluation criteria more robust:

$$\mathbf{MPEGqm} = \frac{1}{K}\sum_{k=1}^{K} MPEGqm(k) \qquad \mathbf{wqm} = \frac{1}{K}\sum_{k=1}^{K} wqm(k) \qquad \mathbf{mqm} = \frac{1}{K}\sum_{k=1}^{K} mqm(k). \tag{5.11}$$

We will compare our proposed objective metric with the results of these three state of the art metrics averaged over a set of $K = 60$ frames.

Table 5.2: Evaluation criteria used in empirical and analytical evaluation of image and video object segmentation systems.

#	Method group	Measure
G-1	Goodness	Intra-region uniformity
G-2	Goodness	Inter-region contrast
G-3	Goodness	Intra-frame color differences
G-4	Goodness	Inter-frame color histogram differences
G-5	Goodness	Motion differences along the boundaries
D-1	Discrepancy	Positions of mis-segmented pixels
D-2	Discrepancy	Classes of mis-segmented pixels
D-3	Discrepancy	Number of mis-segmented objects
D-4	Discrepancy	Shape changes
D-5	Discrepancy	Temporal stability
D-6	Discrepancy	Temporal drift
A-1	Analytical	Execution time

Table 5.3: Summary of evaluation criteria used by each reviewed method.

Image/Video	Source	Measures
Image	Borsotti [15]	G-1,G-2
Image	Rosenberger [131]	G-1,G-2
Image	Everingham [46]	D-4, A-1
Image	Alonso [2]	D-4
Image	Yang [183]	D-4
Image	Goumeidane [55]	D-2
Image	Mezaris [99]	D-2
Video	Correia [23]	G-1,G-2,G-5,D-3, D-4
Video	Cavallaro [17]	D-1,D-2, D-5
Video	Erdem [44]	G-3, G-4, G-5
Video	Erdem [43]	D-4, D-5
Video	Piroddi [123]	G-3, G-4, G-5
Video	Villegas [154]	D-1,D-2, D-5, D-6
Video	MPEG [180]	D-2, D-5
Video	Mech [97]	D-4

Table 5.4: Objective Measures used in evaluating image and video object segmentation systems.

Criteria	Measure	Objective Metric	Source
G-1	Intra-region uniformity	Borsotti [15], Rosenberger [131]	Image
G-2	Inter-region contrast	Borsotti [15], Rosenberger [131]	Image
G-3	Intra-frame color differences	Erdem [44]	Video
G-4	Inter-frame color histogram differences	Erdem [44]	Video
G-5	Inter-frame motion differences	Correia [23], Erdem [44]	Video
D-1	Positions of mis-segmented pixels	Cav. [17],Erdem [43], Villegas [154]	Video
D-2	Classes of mis-segmented pixels	Cav. [17],Villegas [154], MPEG [180]	Video
D-2	Classes of mis-segmented pixels	Goumeidane [55], Mezaris [99]	Image
D-3	Number of objects	Correia [23]	Video
D-4	Shape changes	Erdem [43], Correia [23], Mech [97]	Video
D-4	Shape changes	Alonso [2],Yang [183], Everingham [46]	Image
D-5	Temporal stability	Villegas [154],MPEG [180],Erdem [43],Cav. [17]	Video
D-6	Temporal drift	Villegas [154]	Video
A-1	Execution time	Everingham [46]	Image

5.3 Proposed Objective Evaluation Metric

The proposed objective metric is defined based on two kinds of errors, namely objective errors and perceptual errors. Objective metrics quantify the deviation (objective error) of the segmentation under test from the ground truth and are described in this section. Perceptual metrics weight these deviations (perceptual errors) according to human perception by means of subjective experiments and are presented in Sec. 5.4.

As represented in Fig. 4.7, the proposed objective metric, described in this section, will produce the objective results for each segmentation that will be then, as discussed in the next section, compared to the MOS (Mean Opinion Score) to provide the final perceptual objective assessment.

5.3.1 Spatial Artifacts

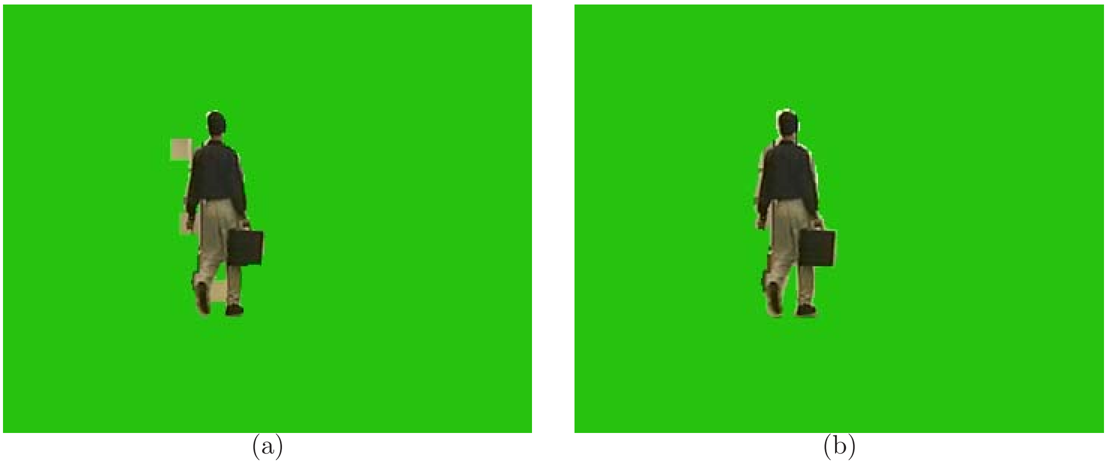
As defined in Sec. 4.5.1, a direct comparison of the results of the segmentation under test with the reference segmentation allows us to identify two types of errors: false positive pixels, $\mathcal{P}(k)$, and false negative pixels $\mathcal{N}(k)$ at frame k. An estimation of absolute spatial error at frame k can be defined as:

$$\mathcal{F}(k) = \mathcal{P}(k) + \mathcal{N}(k). \tag{5.12}$$

A simple normalized spatial error estimate can be computed by normalizing the total amount of false detection, $\mathcal{F}(k)$, by the sum of reference, $R(k)$, and the result segmentation areas, $C(k)$. The relative spatial error, $\mathbf{S}_{error}$ so obtained is given by:

$$\mathbf{S}_{error}(k) = \begin{cases} 0 & \text{if } |\boldsymbol{R}(k)| = 0 \text{ and } |\boldsymbol{C}(k)| = 0, \\ \frac{|\mathcal{F}(k)|}{|\boldsymbol{R}(k)|+|\boldsymbol{C}(k)|} & \text{otherwise.} \end{cases} \tag{5.13}$$

where $|\cdot|$ denotes the cardinality operator; the normalization factor $|\mathbf{R}(k)| + |\mathbf{C}(k)|$ represents the area of the union of both foreground objects. Hence, the relative spatial error can be obtained in the



(a) (b)

Figure 5.5: Example of border holes with the same amount but different distance from the ideal contour (a) large spatial errors (b) small spatial errors.

same manner, for each of the segmentation artifacts we defined in Eqs. (4.5)-(4.6) and Eqs. (4.9)-4.10: $\mathcal{A}_r(k)$, $\mathcal{A}_b(k)$, $\mathcal{H}_i(k)$, $\mathcal{H}_b(k)$.

The relative spatial error $\mathbf{S}_{A_r}(k)$, for all the j added regions , $\mathcal{A}_r^j(k)$, is obtained by simply applying Eq. (5.13) as follows:

$$\mathbf{S}_{A_r}(k) = \frac{\sum_{j=1}^{N_{Ar}} |\boldsymbol{A}_r^j(k)|}{|\boldsymbol{R}(k)| + |\boldsymbol{C}(k)|}, \tag{5.14}$$

where N_{Ar} is the total number of added regions.

Similarly, for all the j holes inside the segmentation, $\mathcal{H}_i^j(k)$, the relative spatial error, $\mathbf{S}_{H_i}(k)$, is given by:

$$\mathbf{S}_{H_i}(k) = \frac{\sum_{j=1}^{N_{Hi}} |\boldsymbol{H}_i^j(k)|}{|\boldsymbol{R}(k)| + |\boldsymbol{C}(k)|}, \tag{5.15}$$

where N_{Hi} is the total number of holes inside the objects.

The spatial error for added background and holes on the border of the object is formulated in a different way. In fact, both kinds of errors are located around the object contours and have to be distinguished from the numerous deviations around the object boundary and a few but larger deviation [97] (as depicted in Fig. 5.5). These two cases (Fig. 5.5(a) and (b)) are perceptually very different. In the first one (a), a part of the the left hand and shoulder and a part between the legs of the person are added. Thus, there are large estimation errors mainly at three regions of the object contour. The second one (b) is a dilated version of the original object, which therefore has a lot of small spatial errors around the object contour. Although the two results look very different, they give similar values of spatial error if evaluated by approaches from the literature such as [17, 43, 104, 154, 180] or by simply applying Eq. (5.13). As in [97], we compute for each error pixel the distance, d, to the ground truth object contour. Moreover, in our approach, we distinguish between the two kinds of error pixel: added background $\mathcal{A}_b(k)$ and holes on the border $\mathcal{H}_b(k)$. In such a way, we will obtain different perceptual weights for different classes of error, since added parts and missing parts are perceived very differently. From the distance values[*] d, we calculate the mean $\overline{d}$ and the standard deviation σ_d, which are then normalized by the maximal diameter, d_{max},

---

[*] For distance computation, 8-connectivity has been used.

of the ground truth object to which the cluster of errors belongs:

$$1+\frac{\overline{d}+\sigma_d}{d_{max}}. \tag{5.16}$$

By combining Eqs. (5.16) and (5.13), we obtain for the border artifacts the corrected relative spatial error. For j added backgrounds, $\mathcal{A}_b^j(k)$, the relative spatial error $\mathbf{S}_{A_b}(k)$, is given by:

$$\mathbf{S}_{A_b}(k)=\big(1+\frac{\sum_{j=1}^{N_{Ab}}(\overline{d}^j_{Ab}+\sigma^j_{dAb})\cdot|\boldsymbol{A}_b^j(k)|}{d^{max}}\big)\cdot\frac{1}{|\boldsymbol{R}(k)|+|\boldsymbol{C}(k)|}, \tag{5.17}$$

and, similarly for j holes on the border, $\mathcal{H}_b^j(k)$, the relative spatial error $\mathbf{S}_{H_b}(k)$ is given by:

$$\mathbf{S}_{H_b}(k)=\big(1+\frac{\sum_{j=1}^{N_{Hb}}(\overline{d}^j_{Hb}+\sigma^j_{dHb})\cdot|\boldsymbol{H}_b^j(k)|}{dmax}\big)\cdot\frac{1}{|\boldsymbol{R}(k)|+|\boldsymbol{C}(k)|}. \tag{5.18}$$

In these measures, while the mean distance is a measure for the average distance between the ground truth and the segmented object contour, the standard deviation gives an idea of how different the measured distances are. The standard deviation is small if the deviation between the original and the estimated contour is quite similar for all the contour pixels. The standard deviation grows with the difference of the measured distance values. This factor is able to take into account the different perceptual impact of the two artifacts depicted in Fig. 5.5 (a) and (b), even if they have the same area.

5.3.2 Temporal Artifacts

The most subjectively disturbing effect is the temporal incoherence of an estimated sequences of object masks. In video segmentation, an artifact often varies its characteristics through time. A non smooth change of any spatial error deteriorates the perceived quality. The temporal artifact caused by an abrupt variation of the spatial errors between consecutive frames is called *flickering*. By carrying out subjective tests on real segmentation, flickering has been observed to be one of the most annoying artifacts introduced by segmentation algorithms. In fact, if an imprecise segmentation is stable along the time, it is perceived less annoying than a more precise segmentation presenting abrupt changes along time. To take this phenomenon into account in the objective metric, we introduce a measure of flickering, $\mathbf{F}(k)$ that can be computed for each kind of *artifact(k)*=[$\mathcal{A}_r(k)$, $\mathcal{A}_b(k)$, $\mathcal{H}_i(k)$, $\mathcal{H}_b(k)$]:

$$\mathbf{F}_{artifact}(k)=\frac{|\mathit{artifact}(k)|-|\mathit{artifact}(k-1)|}{|\mathit{artifact}(k)|+|\mathit{artifact}(k-1)|}, \tag{5.19}$$

where the difference of the amount of an artifact between two consecutive frames is normalized by the sum of the amount of this artifact in current frame k and the previous frame $k-1$. With this formula if the error disappears/appears suddenly it is evenly penalized by the normalization since it causes in the human observer an annoyance due to the unexpected change in the segmentation quality. By doing so, also the *surprise* effect [134] can be taken into account into the metric. This effect is meant to amplify the changes in the spatial accuracy. Moreover, Eq. (5.19) is supported by subjective experiments as we will see in Sec. 5.4.2.

To model this effect, we combine Eqs. (5.13) and (5.19) to construct an objective spatio-temporal error measure $\mathbf{ST}(k)$ at frame k for each artifact:

$$\mathbf{ST}_{artifact}(k)=\mathbf{S}_{artifact}(k)\cdot\frac{1+\mathbf{F}_{artifact}(k)}{2}, \tag{5.20}$$

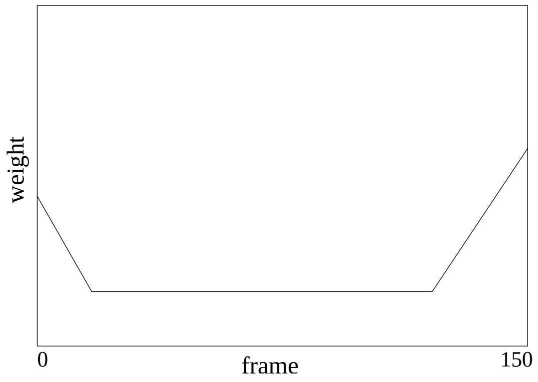


Figure 5.6: Weighted function considering human memory in video quality evaluation proposed in [63].

where we weigh each value of the relative spatial error $\mathbf{S}_{artifact}(k)$ with its correspondent value of flickering, $\mathbf{F}_{artifact}(k)$. This takes into account not only the quality but also the stability of the results.

In modeling the relation between instantaneous and overall quality [58], we can identify two other phenomena related to the temporal context, namely the *fatigue* effect and the *expectation* effect. The fatigue effect is related to the fact that after a while the humman gets used to a certain visual quality thus judging it more acceptable if it persists long enough. In subjective experiments on coded video sequences [63] the characteristics of *short-term* human memory have been studied. Figure 5.6 shows the characteristics of the weighting functions for the short-term characteristics of human memory. The first gradient is called the beginning effect of human memory (it lasts around 50 frames) and presents higher values at the first frames. With our subjective experiments, we aim at finding the weighting function for 60 frame long video sequences.

In fact, our test video were only 5 seconds time long (60 frames) and thus not long enough to cause fatigue effect in the human observers. On the other hand, since they were short video we experienced a different phenomenon: *expectation* effect. By expectation we mean that a good segmentation at the beginning could create a good overall impression on assessing the overall quality of the sequence under test and vice-versa. To model this effect, the overall objective spatio temporal metric, **ST** is formulated as follows:

$$\mathbf{ST}_{artifact} = \frac{1}{K}\sum_{k=1}^{K} w_t(k)\mathbf{ST}_{artifact}(k), \tag{5.21}$$

where the temporal weights $w_t(k)$ that model the expectation effect will be defined by means of subjective results in Sec. 5.4.2. Since our subjective data take values between 0 and 100, $ST_{artifact}(k)$ is scaled by multiplying it by 100.

5.4 Perceptual Impact of Segmentation Artifacts

An automatic method of segmentation typically introduces a combination of errors as described in Sec. 4.5.2. To evaluate objectively the segmentation quality in comparison with a reference segmentation, some features related to the artifact are derived (such as the number of added regions, distance of boundary holes from the ideal contour and so on). As proposed in Sec. 4.5.2, we indentified and analyzed four kinds of typical errors in real object segmentation, namely, added regions, added background, border holes and inside holes.

Many objective segmentation quality metrics have been proposed, as mentioned above, but no effort has been devoted to the study and characterization of typical errors from a perceptual significance point of view. In our approach, in order to study the perceptual impact of segmentation artifacts, we performed different subjective experiments: one for each kind of synthetic artifact and finally one for studying the relationship and their perceptual combination in the overall annoyance.

For each single pure artifact, we tested the proposed artifact metrics (see Eqs. (5.14)- (5.20)). The values for each artifact metric were extracted from the video sequences generated as described in Sec. 4.5.2. The same video sequences were used in the subjective experiments and once the subjective data were processed, the MOS (Mean Opinion Score) was obtained for each test video sequence. The values of the artifact metrics were plotted versus the values of MOS (Mean Opinion Score) as depicted in the block diagram of Fig. 4.7. As described in Sec. 5.4.1, the best fitting psychometric curves were found (among those described in Sec. 2.3.2) that relate the subjective data to the artifact metric output. We use the obtained psychometric curves, one for each artifact, to obtain four *perceptual artifact metrics*.

The results of the perceptual impact of temporal changes in the quality of segmentation and its influence in relation to the length of the video sequence are analyzed in Sec. 5.4.2. The perceptual combination of the four artifacts is analyzed in Sec. 5.4.3 where, the relationship between artifacts and the overall annoyance is found. Finally, in Sec. 5.5, a *perceptual objective spatio temporal metric*, **PST** is proposed on the basis of subjective data.

Standard methods [66] are used to analyze and to screen the judgments provided by the test subjects. Subjective scores are the judgments given by the subjects to each test sequence. The data is first processed by calculating the MOS. Second, outliers are rejected by a screening standard procedure [66]. Depending on the task, the MOS is called MAV (Mean Annoyance Value) since in this case the subjective scores correspond to ‘annoyance’ scores. For strength tasks, the MOS is called MSV (Mean Strength Value), since in this case they correspond to ‘strength’ scores.

5.4.1 Perceptual Spatial Errors

In this section, we present a series of results obtained from psychophysical experiments using *spatial synthetic artifacts*. We define them *spatial* as their features (such as shape, position and area) do not vary along the time. In these experiment, we investigate the annoyance of added regions, added backgrounds, border holes and inside holes. We derive for each artifact a *perceptual artifact metric*, $\mathbf{PST}_{A_b}$, $\mathbf{PST}_{A_r}$, $\mathbf{PST}_{H_i}$, $\mathbf{PST}_{H_b}$, by fitting the artifact metric and subjective results with suitable psychometric curves.

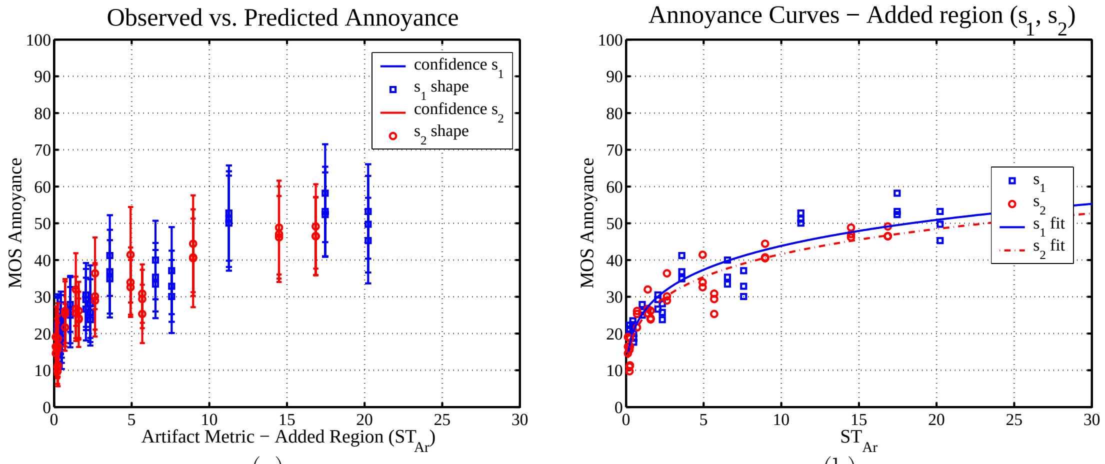
Added region’s experiment

The goal of this experiment was to estimate the annoyance of the added region artifacts by varying its amount. Moreover, we also wanted to test whether different positions and shapes of added region artifacts are perceived the same way. In this experiment, 28 naive subjects were asked to perform the annoyance task. The dialog box used in this experiment is shown in Fig. 4.6 (a).

As described in Sec. 4.5.2, we tested different amounts of added regions with two shapes (square, s_1 and circle, s_2) and three positions slightly further from the ground truth objects (p_1, p_2, and the further p_3). Figure 5.7 (a) shows the $MAVs$ for the two different shapes as a function of the added

Table 5.5: Fitting Parameters S and k for Weibull function.

Artifact	**S**	**k**	r	Pearson	Spearman
s_1	0.0163	0.3030	0.92	0.92	0.92
s_2	0.0127	0.3009	0.91	0.94	0.91
p_1	0.0201	0.3309	0.93	0.88	0.93
p_2	0.0121	0.2869	0.93	0.92	0.93
p_3	0.0127	0.2939	0.93	0.93	0.93
all	0.0148	0.3042	0.94	0.92	0.94



(a) (b)

Figure 5.7: Values of annoyance predicted with artifact metric ST_{A_r} for both added region shapes, square s_1 and circle s_2, vs. observed subjective annoyance. In (a) the confidence interval at 95% are plotted along with subjective data. In (b) the fitting Weibull functions are depicted for each shape.

region artifact metric, $\mathbf{ST}_{Ar}$:

$$\mathbf{ST}_{Ar} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{ST}_{Ar}(k) = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{j=1}^{N_{Ar}}|\boldsymbol{A}_r^j(k)|}{|\boldsymbol{R}(k)|+|\boldsymbol{C}(k)|}\cdot\frac{(1+\mathbf{F}_{A_r}(k))}{2} \tag{5.22}$$

The objective results versus the subjective annoyance and the correspondent 95% confidence intervals are shown in Fig. 5.7 (a). It can be noticed that the perceived annoyance that increases with the size of artifact is not very different for the two different shapes s_1 and s_2. These data were fitted with two Weibull psychometric functions (see Sec. 2.3.2), $w(x,S,k)$:

$$w(x,S,k) = 1 - e^{-(Sx)^k} \qquad where \qquad x = \mathbf{ST}_{Ar} \tag{5.23}$$

and the fitting parameters, S and k, for the two curves s_1 and s_2 are reported in rows 1-2 of Tab. 5.5. The typical correlation coefficients (see Sec. 2.4.2) are reported along with the fitting parameters in columns 4-6 of Tab. 5.5. The two fitting curves depicted in Fig. 5.7 (b) for the different shapes look very similar: the positive trend of the perceived annoyance and the increase of the size of the artifacts are alike.

In Figure 5.8, we plotted an overall fitting curve with its confidence interval for all the data for

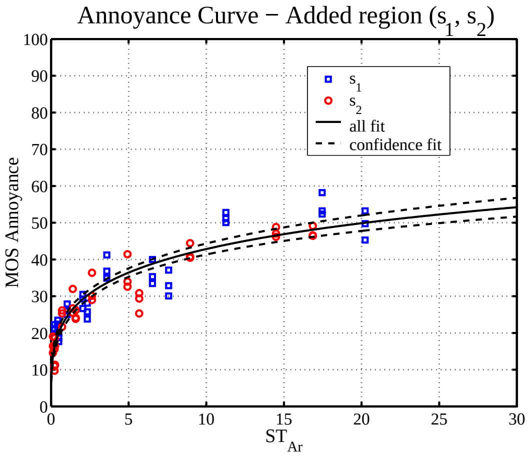


Figure 5.8: Values of annoyance predicted with artifact metric ST_{A_r} vs. observed subjective annoyance, fitted with an overall Weibull function.

both shapes. The question of interest is whether the rate of increase of annoyance with increasing amount of added regions is the same for squares and circles.

To test this hypothesis we used the statistical F test [34]. In Eq. (5.23) we can assume that S and k are constant for the two shapes, yielding a model with just two parameters. In such a case, the model fitting all the data for both shapes has the parameters reported in row 6 of Tab. 5.5 and it is shown in Fig. 5.8. On the other hand, S and k could vary for the two different shapes as depicted in Fig. 5.7 (b) and a total of four parameters (S_{s1}, S_{s2}, k_{s1}, k_{s2}) would better describe the data (see rows 1-2 of Tab. 5.5).

In order to compare the two models describing the data, one with only two parameters and the other with four, the sums of square errors from the fits of the two models were compared. If the errors in the models are independent and normally distributed, the quantity

$$F = \frac{\widehat{\mathcal{S}}_0 - \widehat{\mathcal{S}}_1/(df_0 - df_1)}{\widehat{\mathcal{S}}_1/df_1} \tag{5.24}$$

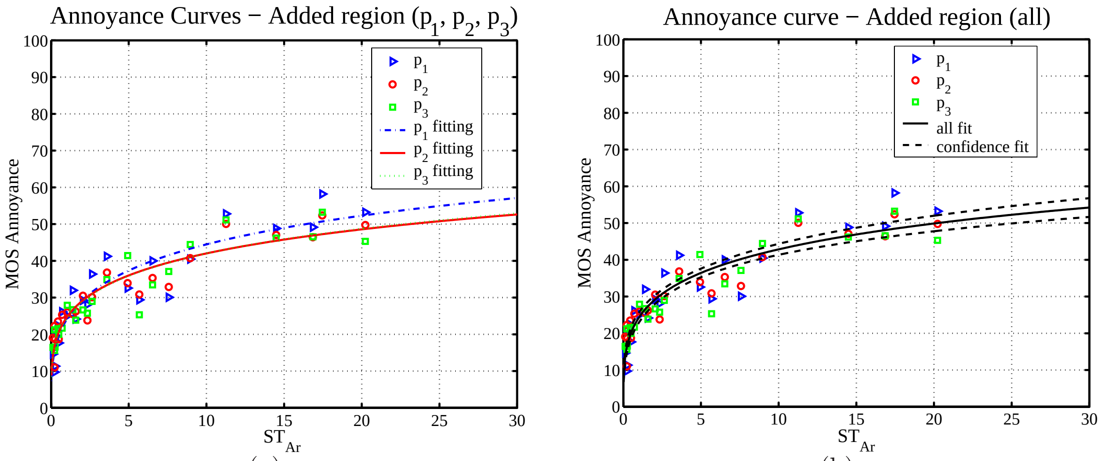
has the central distribution $F_c(df_0 - df_1, df_1)$, where df_0 and df_1 are the degrees of freedom for $\widehat{\mathcal{S}}_0$ and $\widehat{\mathcal{S}}_1$, respectively, which is the sum of squared residual errors under the two hypotheses [34, 92, 101]. The values of F calculated are shown in Tab. 5.6. The value was smaller than the critical F-value indicating that the two-parameter model describes the data as well as the model with four paramters. This means that the rate of increase in annoyance with increasing amount of added region is the same for squares and circles and one curve describes well enough all the data both for squares and circles. The same test was applied to test whether different positions influenced the perception of annoyance of added regions. Figure 5.9 (a) shows the three Weibull curves (see Sec. 2.3.2) for p_1, p_2 and p_3 and Fig. 5.9 (b) shows (the same as in Fig. 5.8) the overall fitting curve plotted for different positions.

In this second case, the hypothesis was that the simple model with two parameters of row 6 of Tab. 5.5 described the data as well as the model with six parameters (S_{p1}, S_{p2}, S_{p3}, k_{p1}, k_{p2}, k_{p3}) given in rows 3-5 in Tab. 5.5. Also in this case the value was smaller than the critical F-value and the model with two parameters ($S = 0.0148$, $k = 0.3042$) was chosen to describe the added region perceptual metric, $\mathbf{PST}_{A_r}$:

$$\mathbf{PST}_{Ar} = w(\mathbf{ST}_{Ar}, 0.0148, 0.3042) = 1 - e^{-(\mathbf{ST}_{A_r} \cdot 0.0148)^{0.3042}} \tag{5.25}$$

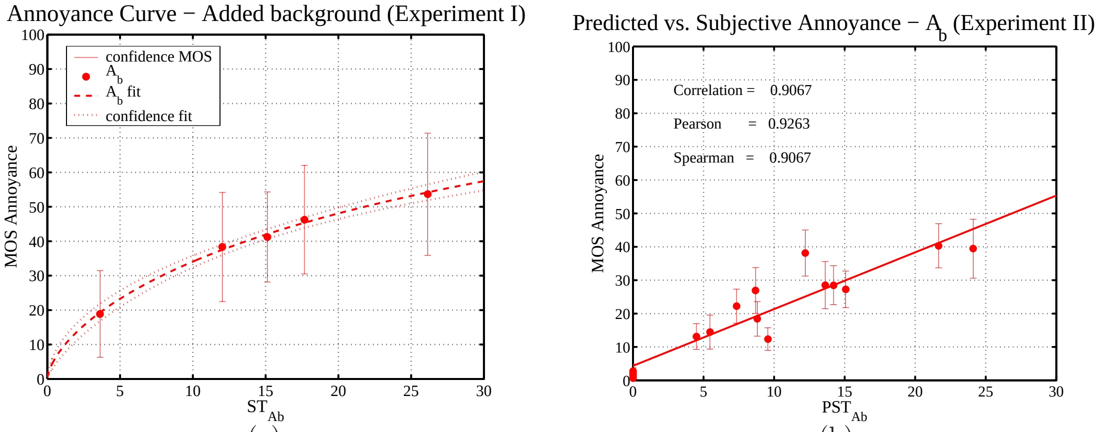
Table 5.6: F values to test if different fitting curves are needed to describe the perceived annoyance for different shapes and positions of added regions.

Artifact model	$\mathbf{F_c}$ (critical)	**F** (value)	$\mathbf{p}(\mathbf{F} < \mathbf{F_c})$
added region shape	F(2,68)=3.13	1.43	0.24
added region position	F(4,66)=2.51	0.64	0.63
inside hole position	F(2,28)=3.34	0.13	0.87
hole distinction	F(2,44)=3.21	5.01	0.01



(a) (b)

Figure 5.9: Values of annoyance predicted with artifact metric ST_{A_r} vs subjective annoyance for added regions at different positions from the ground truth: p_1, p_2 and p_3 slightly further. (a) shows the three fitting curves for the three different positions and (b) the unique fitting for all the data.



(a) (b)

Figure 5.10: Values of annoyance predicted with artifact metric ST_{A_b} vs subjective annoyance for added background: (a) shows the fitting curve and its confidence interval for Experiment I and (b) shows the correlation of the derived perceptual metric in Experiment II.

We have thus proved by means of subjective experiments that shape and position do not influence the perception of added region artifact at a given amount. In several state of the art methods [37, 86, 154], the distance of this kind of error from the border of the ground truth is always taken into account. This experiment showed that added region artifact is perceived independently not only from the distance from the ground truth but also from the shape. In other state of the art metrics [23, 97], this added region error (disconnected from the ground truth) is not even considered.

Added background experiment

The first step was to generate video sequences with only one type of artifact at relatively high level of annoyance: added background, $\mathcal{A}_b(k)$. The synthetic added background was created as described in Sec. 4.5.2. For the video *Hall monitor*, five new segmented video sequences were created by varying the number of dilations of correctly segmented video sequences from one dilation to eight dilations.

Subjects in this experiment (Experiment I) were 8 male students from EPFL, aged between 23-28 and were asked to rate the quality of the segmented video under test [38]. The data gathered from subjects for added background evaluation provided one single value for each test sequence. From these data the values corresponding to the $MAVs$ for added background were obtained.

The performance of the proposed objective metric for added background was tested by means of these subjective results. The specific artifact metric for added background is:

$$\mathbf{ST}_{A_b} = \frac{1}{K}\sum_{k=1}^{K} ST_{A_b}(k) = \frac{1}{K}\sum_{k=1}^{K} (1+\frac{\sum_{j=1}^{N_{Ab}}(\overline{d}^{j}_{Ab}+\sigma^{j}_{dAb}))}{dmax}\cdot\frac{|\boldsymbol{A}_b^j(k)|}{|\boldsymbol{R}(k)|+|\boldsymbol{C}(k)|}\cdot\frac{(1+\mathbf{F}_{A_b}(k))}{2} \tag{5.26}$$

Figure 5.10 (a) displays the results of applying this metric on the test sequences for this experiment containing only added background artifacts. The x-axis corresponds to the values of the added background objective metric $\mathbf{ST}_{A_b}$ and the y-axis corresponds to the subjective MAV values.

As it can be noticed from this figure, $\mathbf{ST}_{A_b}$ increases linearly with an increase in the strength of the artifact increases, as well as the MOS Annoyance. In Fig. 5.10 (a) the MOS Annoyance values have been fitted with a Weibull function:

$$\mathbf{PST}_{A_b} = w(\mathbf{ST}_{A_b}, 0.0262, 0.6533) = 1 - e^{-(\mathbf{ST}_{A_b}\cdot 0.0262)^{0.6533}} \tag{5.27}$$

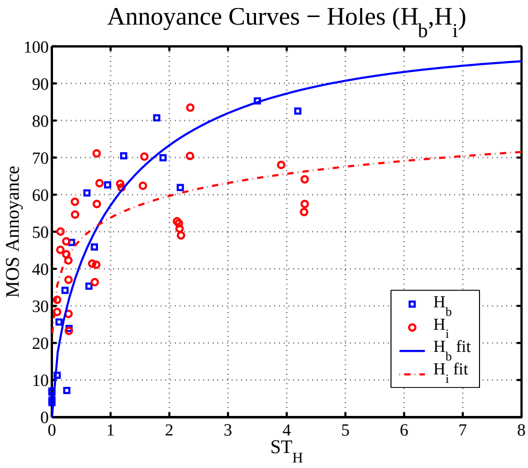


Figure 5.11: Annoyance curves predicted with artifact metric ST_H vs. observed subjective annoyance for both kind of holes: inside holes H_i and border holes H_s

.

where the parameters are equal to $S = 0.0262$, $k = 0.6533$. This psychometric fitting curve correlates very well with the subjective data: the correlation coefficients are $r = 1.00$, Pearson=0.99, and Spearman=1.00.

The high value of correlation shows that $\mathbf{PST}_{A_b}$ is a good perceptual artifact metric to predict subjective annoyance values. However, since this experiment contained only 5 test sequences and 8 subjects, we validated this perceptual artifact metric on other test sequences. We applied the same metric on the test sequences of the combined artifacts experiment that contained only added background artifacts. In this second experiment (Experiment II), 31 subjects aged between 21-30 years performed the annoyance estimation task. The test video sequences are described in combinations 3-7-11 of Tab. 4.5. As described in Sec. 6, there were three amounts of added background inserted in four video sequences for a total of 12 test video sequences. Figure 5.10 (b) displays the correlation of MOS Annoyance values obtained for these test sequences containing only added background with the perceptual metric of Eq. (5.27) derived from Experiment I. As can be observed from this curve the metric has a good fit with the MOS Annoyance values and the psychometric curve of Eq. (5.23) produces a correlation of 90%. This confirms the results of the first experiment on added background and the reliability of the proposed metric $\mathbf{PST}_{A_b}$ in Eq. (5.23) for added background. Such perceptual artifact metric has proved to well describe the subjective perception of added background in both cases (presented in Figs. 5.5 (a) and (b)). That is, when the distance of the added background from the ideal contour has large value of σ_d as in Experiment II in Fig. 5.5 case (a) and for lower value of σ_d as in Experiment I, Fig. 5.5 case (b).

Holes experiment

There were two goals in this experiment. The first goal was to test the two objective metrics, one proposed for border holes:

$$\mathbf{ST}_{H_b} = \frac{1}{K}\sum_{k=1}^{K} ST_{H_b}(k) = \frac{1}{K}\sum_{k=1}^{K} (1+\frac{\sum_{j=1}^{N_{Hb}}(\bar{d}^j_{Hb}+\sigma^j_{dHb}))}{d_{max}}\cdot\frac{|\boldsymbol{H}_b^j(k)|}{|\boldsymbol{R}(k)|+|\boldsymbol{C}(k)|}\cdot\frac{(1+\mathbf{F}_{H_b}(k))}{2} \qquad (5.28)$$

and the second for inside holes:

$$\mathbf{ST}_{H_i} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{ST}_{H_i}(k) = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{j=1}^{N_{Hi}}|\boldsymbol{H}_i^j(k)|}{|\boldsymbol{R}(k)|+|\boldsymbol{C}(k)|}. \tag{5.29}$$

The second goal was to determine the psychometric annoyance functions for the two kinds of synthetic artifacts. Finally, we studied whether the annoyance caused by a boundary hole could be worse than for an inside hole (for large holes). For this purpose we generated 48 test sequences as described in Sec. 4.5.2. In this experiment 28 naive subjects were asked to perform the annoyance task.

Figures 5.11 shows the plots of the *MOS* Annoyance as a function of the proposed objective metric **ST** of **hole** artifacts for all video sequences. The graph shows two curves, one corresponding to the boundary holes *MOS* Annoyance and the other corresponding to the *MOS* Annoyance inside holes. The boundary holes curve increases faster than the inside holes curve. For small values of the size, $\mathbf{H}_i$ is more annoying than $\mathbf{H}_b$, as already reported in the literature [86]. But on the other hand, by increasing the size of the holes, a *point of inversion* can be noticed concerning the annoyance of the two kinds of artifacts [37]. After that point of inversion, $\mathbf{H}_b$ is more annoying than $\mathbf{H}_i$, since the shape of the object becomes less recognizable. For all the sequences tested ('Highway', 'Coastguard', 'Hall monitor' and 'Group'), independently from the content, the point of such inversion starts around the same amounts of errors [37]. This subjective experiment indicates that both the kind and the size of the hole should be jointly taken into account and not only the distance when an objective metric is proposed. Besides, in the objective metrics proposed in the literature, holes are only considered in terms of uncorrelated set of pixels and their distances from the reference boundary of the object [17, 154]. With this experiment we proved that a cluster of error pixels should be distinguished and their characteristics should be thoroughly studied instead of considering each error pixel individually. In other words, in the literature, methods reported in [17, 86, 154] claim that as we move away from the border, holes become more annoying but this depends on also the kind and the size of the hole, as shown in this experiment.

The psychometric curve that best fits the subjective data is still the Weibull function, for both kinds of holes. The perceptual artifact metric for inside hole is given by:

$$\mathbf{PST}_{H_i} = w(\mathbf{ST}_{Hi}, 0.3310, 0.2339) = 1 - e^{-(\mathbf{ST}_{H_i}\cdot 0.3310)^{0.2339}} \tag{5.30}$$

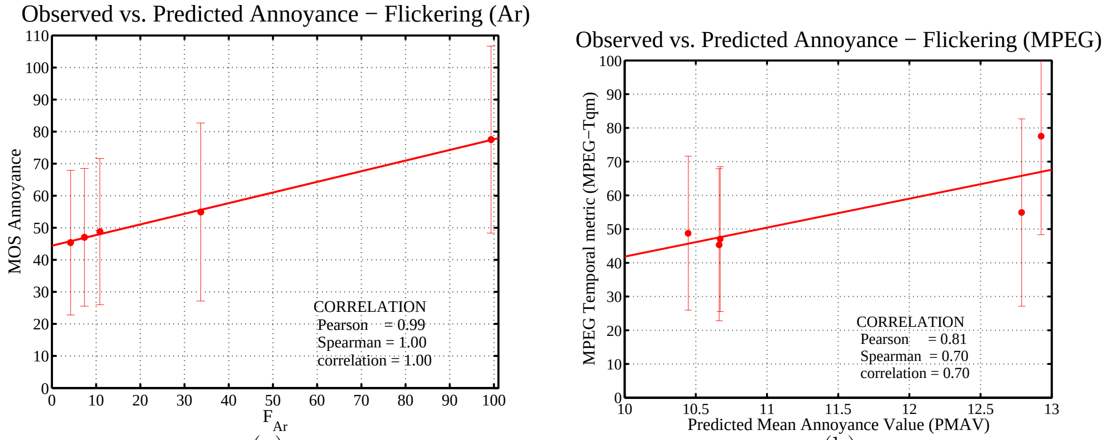
and for border hole, the perceptual artifact metric is:

$$\mathbf{PST}_{H_b} = w(\mathbf{ST}_{Hb}, 0.7716, 0.6416) = 1 - e^{-(\mathbf{ST}_{H_b}\cdot 0.7716)^{0.6416}} \tag{5.31}$$

As can be observed from these curves the $\mathbf{PST}_{H_b}$ metric has a good fit with the *MOS* Annoyance values and the psychometric curve (Eq. (5.31)) produces a correlation of 94% for r, Pearson = 92% and Spearman = 94% . The correlation for the perceptual metric $\mathbf{PST}_{H_i}$ is lower: $r = 0.68$, Pearson = 0.65 and Spearman = 0.68.

Two positions of inside holes have been tested: one further than the other to the object borders. Hence, the F-test has been used to investigate whether the perceived annoyance of these two positions could be described with two different fitting curves. As reported in row 3 of Tab. 5.6 the F value shows that the same curve can be used to fit both positions of inside holes as plotted in Fig. 5.11. This validates the simple characterization that made about inside holes without considering the distance of the inside hole from the border of the ground truth (see Eq. (5.29)).

To further confirm the hypothesis that a distinction between inside holes and border holes has to be made applied the F-test on these two sets of data to see if a unique fitting curve can interpolate both kinds of artifacts (see row 4 of Tab. 5.6). The F-value in this case is equal to 5.01 that is above



(a) (b)

Figure 5.12: Correlation of flickering metrics with subjective annoyance: (a) proposed flickering metric for Added Region $\mathbf{F}_{a_r}$, (b) MPEG temporal coherence metric, $MPEGTqm$.

the threshold of $F(2, 44)$ equal to 3.21. This means that an overall fitting curve is not sufficient to describe both phenomena.

5.4.2 Perceptual Temporal Errors

We tested two different temporal phenomena on the perception of segmentation quality. Both are described in Sec. 4.5.2. The first one aims at validating the proposed *flickering* metric in Eq. (5.19). The second is pointed to find the temporal weights of Eq. (5.21) related to the *expectation* effect.

Flickering experiment

By carrying out subjective tests on the segmentation quality, flickering has been observed to be one of the most annoying artifacts introduced by segmentation algorithms. Different variations of any spatial error could be implemented to test the flickering perception. We chose to change the position of added regions along the test sequence. This segmentation error is typically given by noise introduced by the video camera and changes in illumination. The temporal errors present the same number of added regions by the same shape but their position change every flickering period f_T (reported in Tab. 4.4). The annoyance task was performed with 8 naive subjects aged between 23-28. Figure 5.12 (a) shows the MOS Annoyance values gathered versus the objective metric for flickering $\mathbf{F}_{Ar}$:

$$\mathbf{F}_{A_r}(k) = \frac{|\mathbf{A}_r(k)| - |\mathbf{A}_r(k-1)|}{|\mathbf{A}_r(k)| + |\mathbf{A}_r(k-1)|} \tag{5.32}$$

The correlation with the MAV, as it can be noticed, is extremely high ($r = 100\%$). This shows that the metric proposed for flickering is very suitable to describe the perception of this artifact. In order to compare this metric for capturing the temporal variations present in segmentation we tested the temporal metric MPEG [180], $MPEGTqm$ in Eq. (5.3). Figure 5.12 (b) shows the result of this state of the art metric versus the $MAVs$ gathered for this experiment. The correlation is 70%. This shows that the proposed objective flickering metric outperforms the MPEG temporal coherence measure.

Expectation experiment

In this experiment, 28 subjects were asked to perform the annoyance task. The same amounts of added regions were inserted and varied temporally as depicted in Fig. 4.11. We aimed at finding if there is an *expectation* effect and how this affects the overall perceived quality. Expectation effect tested consists in providing a good overall impression on assessing the quality of the sequence under test if a good segmentation at the beginning is present. We used in this experiment two video sequences: 'Coastguard' and 'Hall monitor'. Even though the contents of the two test sequences are very different, the two curves obtained in Fig. 5.13 (a) for the *MAVs* look quite similar at a first glance. This is especially true for complex temporal artifacts and when the temporal defects become similar to a *flickering*. For both video sequences, the most annoying artifacts are those with more temporal variation of added regions (condition $\mathbf{B}_9$ in Fig. 4.11). A surprising result is that the initial temporal variation is worse than the final temporal variation for both video sequences [37] (see conditions $\mathbf{B}_6$ and $\mathbf{B}_8$ in Fig. 4.11 and MAV values in ig. 5.13 (a)). We explained this effect like a sort of expectation effect. A good segmentation at the beginning creates a good impression. A bad segmentation at the beginning puts the overall impression of the segmentation quality in jeopardy. We wanted to model this temporal effect on the overall impression of segmentation quality. Therefore, we used these three $MAVs$ value $\mathbf{B}_6$, $\mathbf{B}_8$ and $\mathbf{B}_9$ to mimic this temporal perception of error and find the temporal weights $w_t(k)$ of Eq. (5.21) in solving the following system:

$$\left\{ \begin{array}{lcl} \frac{1}{\sum w_t(k)} \sum_{k=1}^{K} w_t(k) \cdot \mathbf{ST}_{B_6}(k) & = & MAV_{\mathbf{B}_6} \\ \frac{1}{\sum w_t(k)} \sum_{k=1}^{K} w_t(k) \cdot \mathbf{ST}_{B_8}(k) & = & MAV_{\mathbf{B}_8} \\ \frac{1}{\sum w_t(k)} \sum_{k=1}^{K} w_t(k) \cdot \mathbf{ST}_{B_9}(k) & = & MAV_{\mathbf{B}_9}, \end{array} \right.$$

where $\sum_k w_t(k) = K$ is the fourth condition, since the overall judgment for the sequence has to be normalized by the total number of frames (60). After some trials, we chose to parametrize $w_t(k)$ with the following function:

$$w_t(k) = a \cdot e^{-\frac{k}{b}} + c. \tag{5.33}$$

We use a nonlinear least-square data fitting by the Gauss-Newton method to estimate the parameters a, b and c. Since we wanted to make it as general as possible and not specific for a kind of artifact, we fixed $\mathbf{ST}_B(k)$ equal to 1 if the stimulus (the added region) at instant k was present and zero otherwise (see Fig. 4.11). Figure 5.13 (b) shows the weighting function and the parameters a, b and c found with the data fitting that will be used in the spatio-temporal metric:

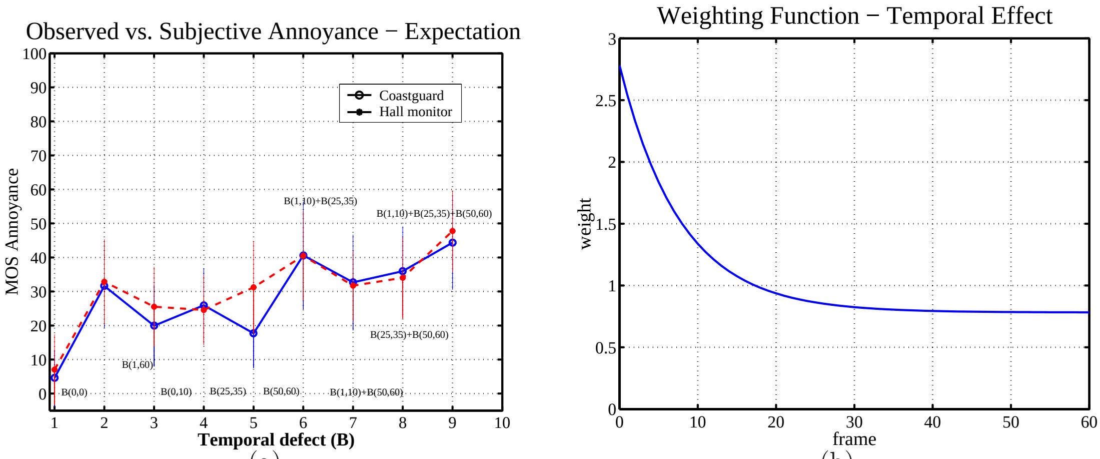
$$\mathbf{ST} = \frac{1}{K} \cdot \sum_{k=1}^{K} (a \cdot e^{-\frac{k}{b}} + c) \cdot \mathbf{ST}_{artifact}(k), \tag{5.34}$$

with $a = 2$, $b = 7.8$, and $c = 0.78$.

It is likely that for longer video sequences the *short term* memory effect [63] affects the expectation as depicted in Fig. 5.6. In our case, by applying this weighting function to the objective metric we observed that the correlation with the MAV values increased with respect to the simple weighting $w_t(k) = 1$ for all k.

5.4.3 Perception of Combined Artifacts

To investigate the relationship between individual artifact strengths and overall annoyance another experiment was devised [48, 102]. The subjects in this experiment were divided into two independent groups. The first group was composed of 31 subjects aged between 21-30 (with 5 females) who performed the annoyance task. The second experiment group was composed of 27 subjects (with 9



(a) (b)

Figure 5.13: Experiment on Expectation effect: (a) temporal defect **B** vs. perceived annoyance for 'Coastguard' and 'Hall' video sequences; (b) temporal weighting function taking into consideration the Expectation effect.

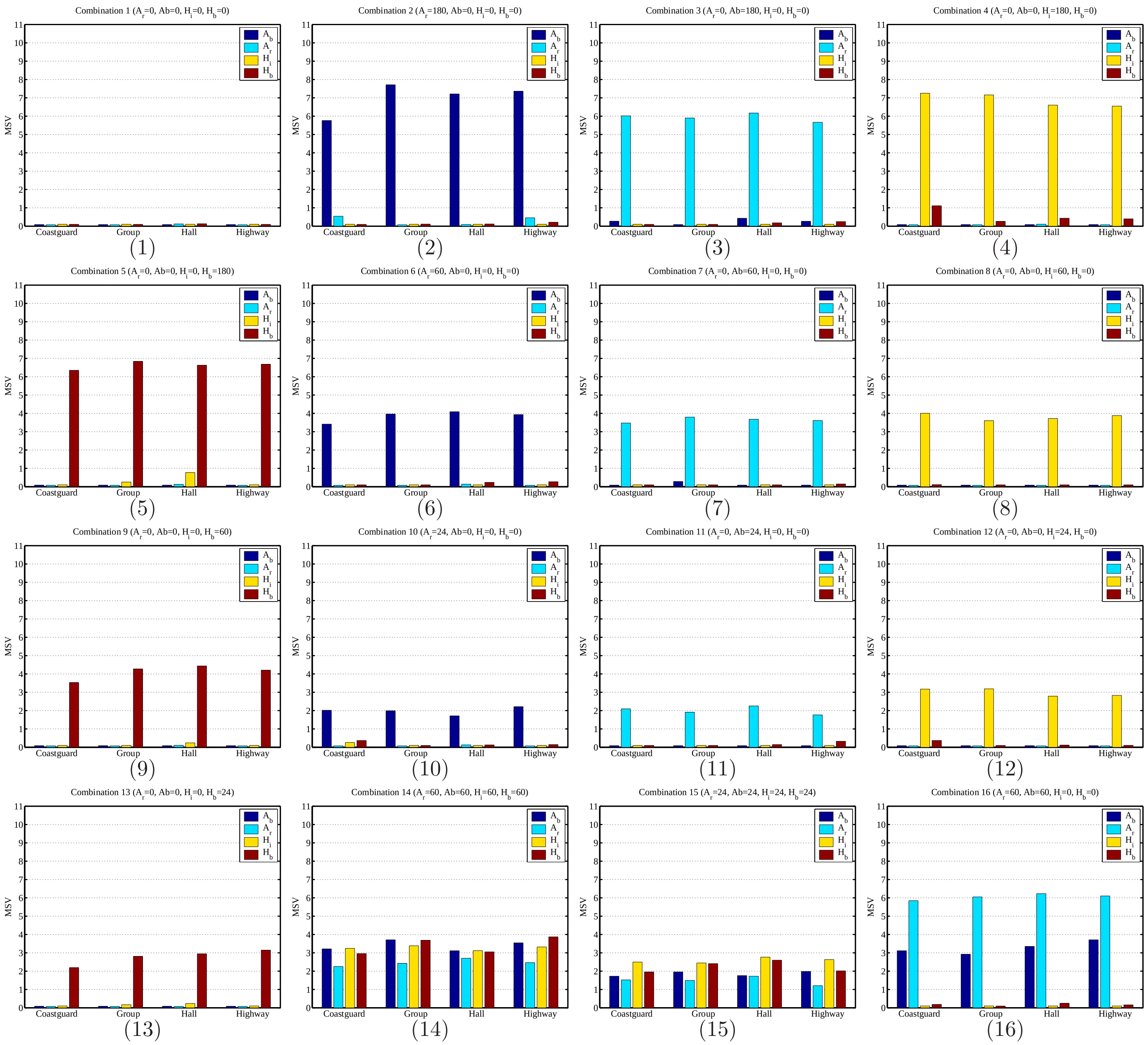
females) aged between 23-32 who performed the strength task (see Sec. 4.4.2). Both groups watched and judged the same test sequences which consisted of 45 combinations of added regions, added background, inside hole and border hole at different amounts as reported in Tab. 5.7.

Four original video sequences were used in this experiment: 'Coastguard', 'Group', 'Hall monitor' and 'Highway'. During the instruction stage, subjects of both groups were told that the test video might contain up to four different types of artifacts - added regions, added background, inside hole and border hole. The training stages were different for each group of subjects. Subjects in the first group were shown two sets of video sequences: reference segmentation and segmentations with very annoying impairments. They were asked to assign a value of '100' to the most annoying impairments in the second set. Subjects in the second group were also shown two sets of segmentations: reference segmentations and segmentation with example of strong pure artifacts. Before the presentation of each type of artifact, subjects were told the name of the artifact type and given a brief description of its appearance. They were asked to assign a value of '10' to the strongest impairments. For both groups of subjects the same test sequences were used for the practice and experimental stages.

Subjective Data Analysis

The data gathered from subjects in the first group provided one single score value for each test sequence. This value corresponded to the Mean Annoyance Value (MAV). The data gathered from subjects in the second group provided four score values for each test sequence. These values corresponded to the mean strength values ($MSVs$) for added regions, added background, inside hole, border hole respectively. The values for the average MAV and MSV for all video sequences are shown in columns 2-6 of Tab. 5.7.

Figures 5.14-5.16 (1)-(45) depict the bar plots of the MSV values obtained for added regions, added background, inside hole and border hole. Each graph shows the $MSVs$ for each of the combinations. The combination 1 corresponds to the reference segmentation since $|\boldsymbol{A}_r| = 0$, $|\boldsymbol{A}_b| = 0$ $|\boldsymbol{H}_i| = 0$, $|\boldsymbol{H}_b|$=0. It is interesting to notice that for some reference segmentations the values for the $MSVs$ and $MAVs$ corresponding to the originals are not zero, indicating that subjects report that these segmentations contained some type of impairment and annoyance/strength levels different from zero.



(1) (2) (3) (4)
(5) (6) (7) (8)
(9) (10) (11) (12)
(13) (14) (15) (16)

Figure 5.14: Mean Strength Value (MSV) obtained in combined artifact experiment for all the video sequences.

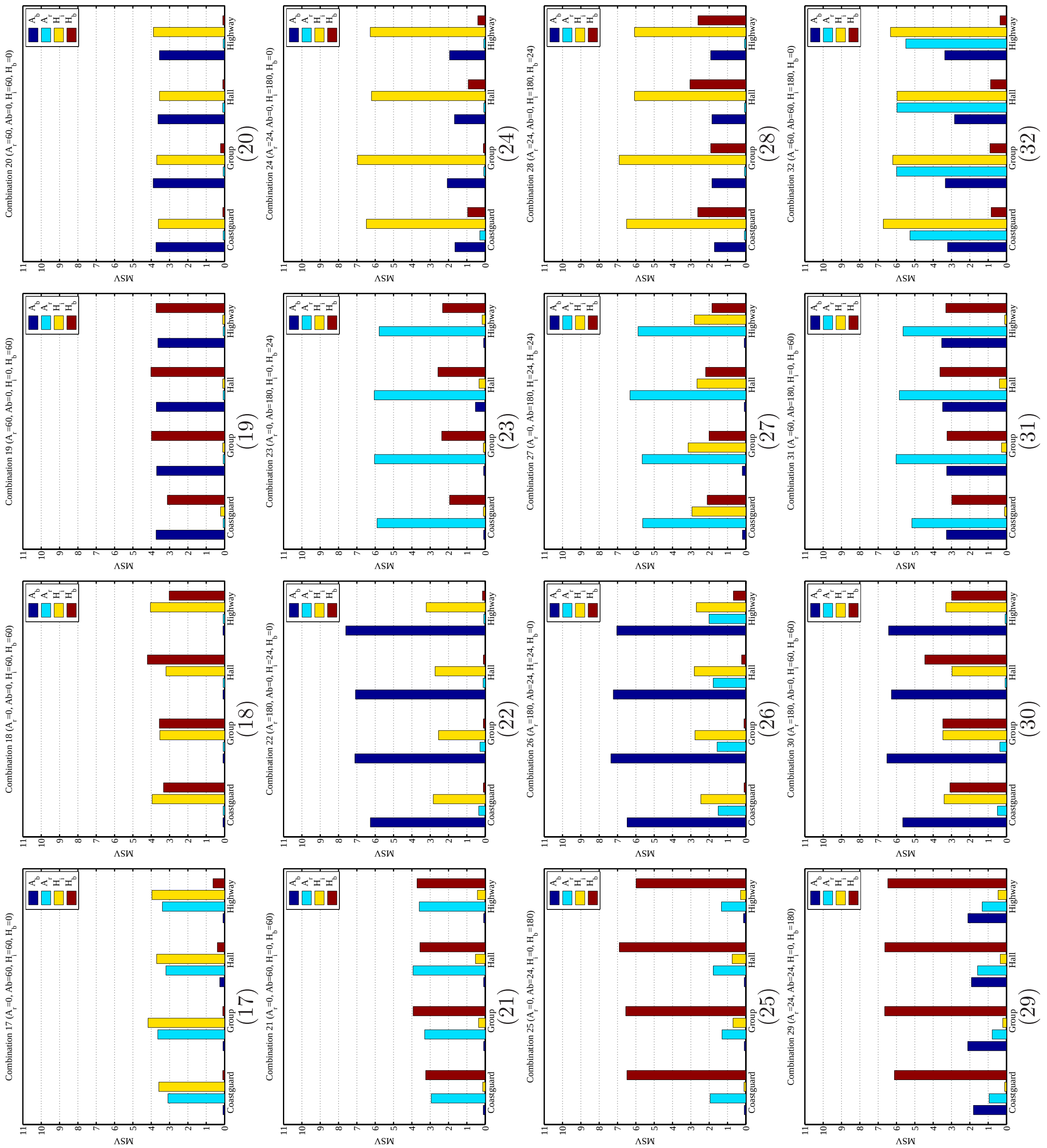


Figure 5.15: Mean Strength Value (MSV) obtained in combined artifact experiment for all the video sequences.

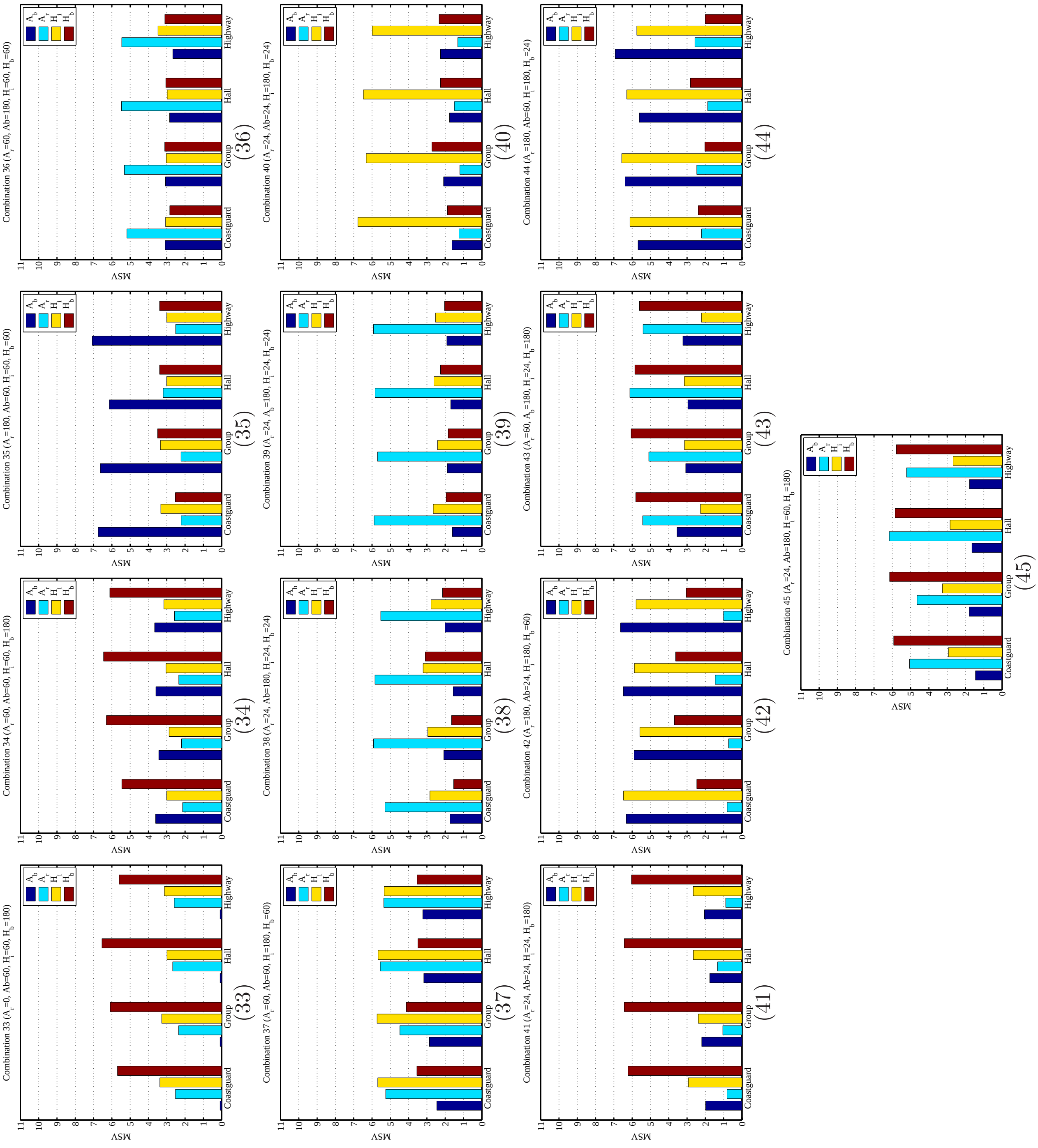


Figure 5.16: Mean Strength Value (MSV) obtained in combined artifact experiment for all the video sequences.

The test combinations 2-6-10, 3-7-11, 4-8-12 and 5-9-13 (defined in Tab. 4.5) in Fig. 5.14 correspond to segmentations with only added regions, added background, inside hole and border hole respectively. For these combinations, the highest $MSVs$ were obtained for the corresponding pure artifacts, while the other three types of artifacts received smaller values. MAV values were highest for segmentation that contained large border holes artifacts. The test combinations 16-25 in Figs 5.14 and 5.15 correspond to segmentations with two types of artifacts. For these combinations, the inserted artifacts with biggest amount received the highest MSV. Equal amounts received more or less the same MSVs. The test combinations 26-33 in Fig 5.15- 5.16 correspond to segmentation with three types of artifacts and the same happens. The test combinations 14-15 in Fig 5.14 and 34-45 in Fig 5.16 correspond to segmentations with the four types of artifacts. Also, for these combinations, the artifacts with the greatest amount received higher $MSVs$. The same effects were observed for these combinations. The presence of all the artifacts seemed to decrease the perceived MSV of added regions (combination 36) while it seemed to increase the perceived strength of inside hole (combination 37). Therefore, there are interactions between the four artifacts in determining the perceived strengths of the artifacts. Our principal interest in measuring the artifacts' strength is to investigate the relationship between the perceptual strengths of each type of artifacts and the overall annoyance. In other words, we want to predict the MAV from the 4 MSV values (MSV_{A_r}, MSV_{A_b} , MSV_{H_i}, MSV_{H_b}). As discussed in Chapter 2, if a video is affected by one or more types of artifacts, the total annoyance can be estimated from the individual artifact perceptual strengths ($MSVs$) using the Minkowski metric [48]:

$$PMAV = (a \cdot MSV_{A_r}^p + b \cdot MSV_{A_b}^p + c \cdot MSV_{H_i}^p + d \cdot MSV_{H_b}^p)^{\frac{1}{p}} \tag{5.35}$$

where $PMAV$ is the predicted value for MAV, p is the Minkowski power, a, b, c, and d are the weighing coefficients for added regions, added background, inside hole and border holes, respectively. Columns 2-7 of Tab. 5.8 show the results obtained for the Minkowski metric fitting. The fits were made both to the data for individual video sequences and to the overall data set. Column 7 of Tab. 5.8 shows the correlation r of the fit. The P-values corresponding to the correlation values of the fit were all roughly equal to zero for all cases. The fit to all the data sets is reasonably good and there is little systematic error in the predictions.

Figures 5.17 (a)-(d) correlate the MAV versus $PMAV$ for the segmented video 'Coastguard', 'Group', 'Hall monitor' and 'Highway' generated using the Minkowski metric. The correlation coefficients for these fits are respectively: 96%, 96%, 95% and 94%. In Fig. 5.18 we plotted the MAV versus $PMAV$ corresponding to the set of all segmentations. The correlation coefficient for the fit is 94%.

In summary, annoyance increases both with the number of artifacts and their strengths. The added region weight is almost the half of inside hole and added background weight and one third of border hole artifact weight (a = 5.50, b = 10.35, c = 10.81, d = 15.30).

Objective data analysis

In the previous section, we presented an analysis of the Minkowski metric using only subjective data from the experiment of combined artifacts. Now, we want to investigate if the same type of model can be used to estimate the overall annoyance by using, instead, individual perceptual artifact metrics. In previous sections of this chapter, we presented several artifact metrics that measured the annoyance of four of the most common spatial and temporal artifacts found in video object segmentation. These metrics were tested using synthetic artifacts developed in Sec. 4.5.2. To evaluate the performance of each artifact metric and to find the psychometric fitting curve that transforms the metric into a perceptual one, we tested its ability for test segmentations containing

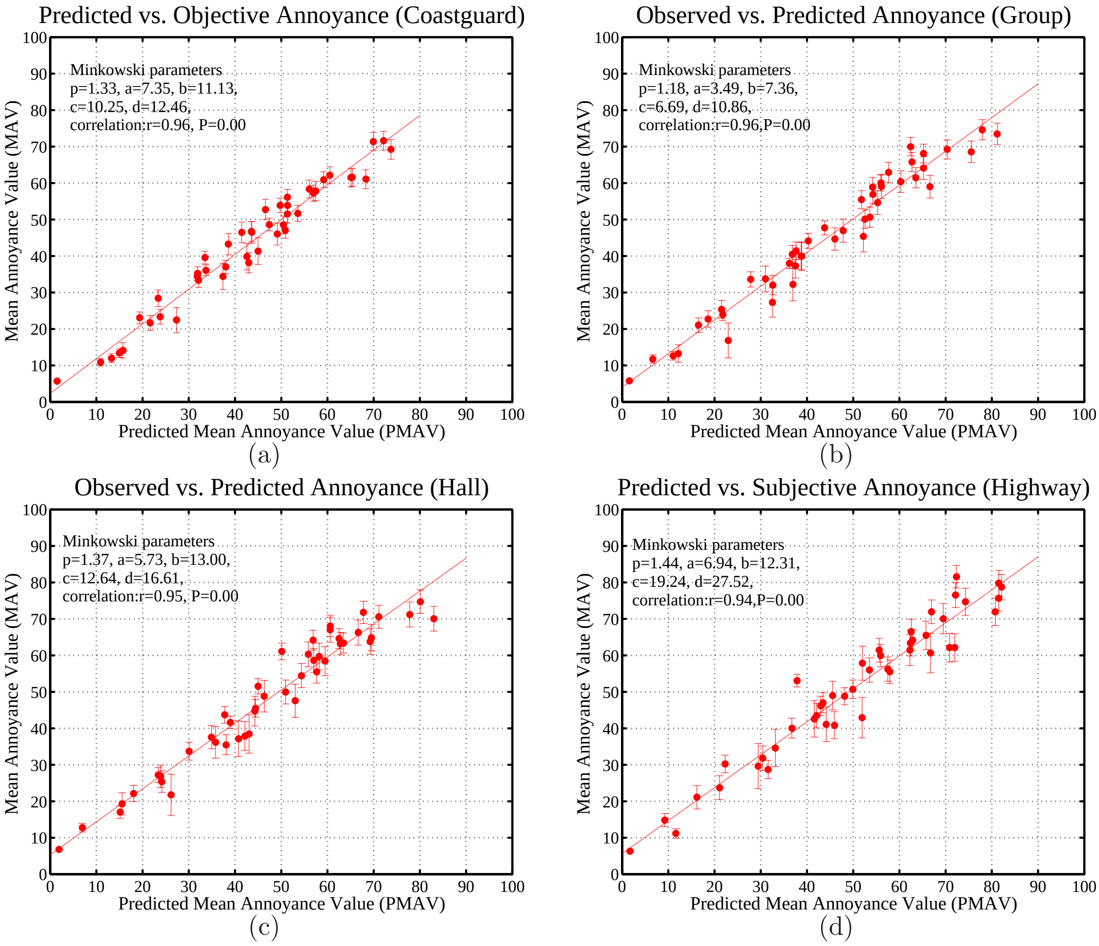


Figure 5.17: Subjective values for MSV vs. MAV and their correlation through the Minkowski metric: the fitting parameter and the correlation are reported for each sequence: (a) 'Coastguard', (b), 'Group', (c) 'Hall', (d) 'Highway'.

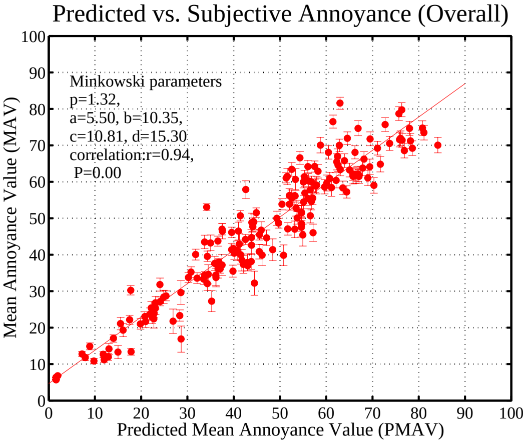


Figure 5.18: Subjective values for MSV vs. MAV and their correlation through the Minkowski metric: the fitting parameter and the correlation are reported.

only the artifact being measured. All these analysis were carried out in Secs. 5.4.1 and 5.4.2, respectively using subjective data from experiments on added regions, added background, inside hole, border hole, flickering and expectation effect.

In order to be able to combine the different proposed perceptual artifact metrics to estimate the overall annoyance, we first need to normalize the metric values to a common range. This is because each metric has its own range of numerical values, according to the specific method employed. The output values of the metrics are normalized to be within 1-10, using the coefficients obtained from a linear fitting of the metric values to the perceptual strength data MSV acquired from the subjective experiment. Table 5.9 shows the normalization expression of each metric, along with the Spearman's rank correlation coefficient for all the fits. The correlation values give a measure of how close the perceptual artifact metrics are to the subjective data for this experiment.

In summary, all the perceptual metrics show a good correlation with the MSV of the subjective experiment. Then, on the basis of these results an overall annoyance metric will be proposed in the next section and the results will be presented and discussed.

5.5 Overall Perceptual Objective Metric

Given the results of the previous sections, we propose a ground truth based perceptual objective metric that uses the metrics for spatial artifacts: $\mathbf{PST}_{A_r}$ of Eq. (5.25), $\mathbf{PST}_{A_b}$ of Eq. (5.27), $\mathbf{PST}_{H_i}$ of Eq. (5.30), and $\mathbf{PST}_{H_b}$ of Eq. (5.31); for temporal artifacts: the **flickering** metric in Eq. (5.19) and the **expectation** effect of Eq. (5.34). Therefore the predicted annoyance by the proposed metric is given by the following expression (Minkowski metric):

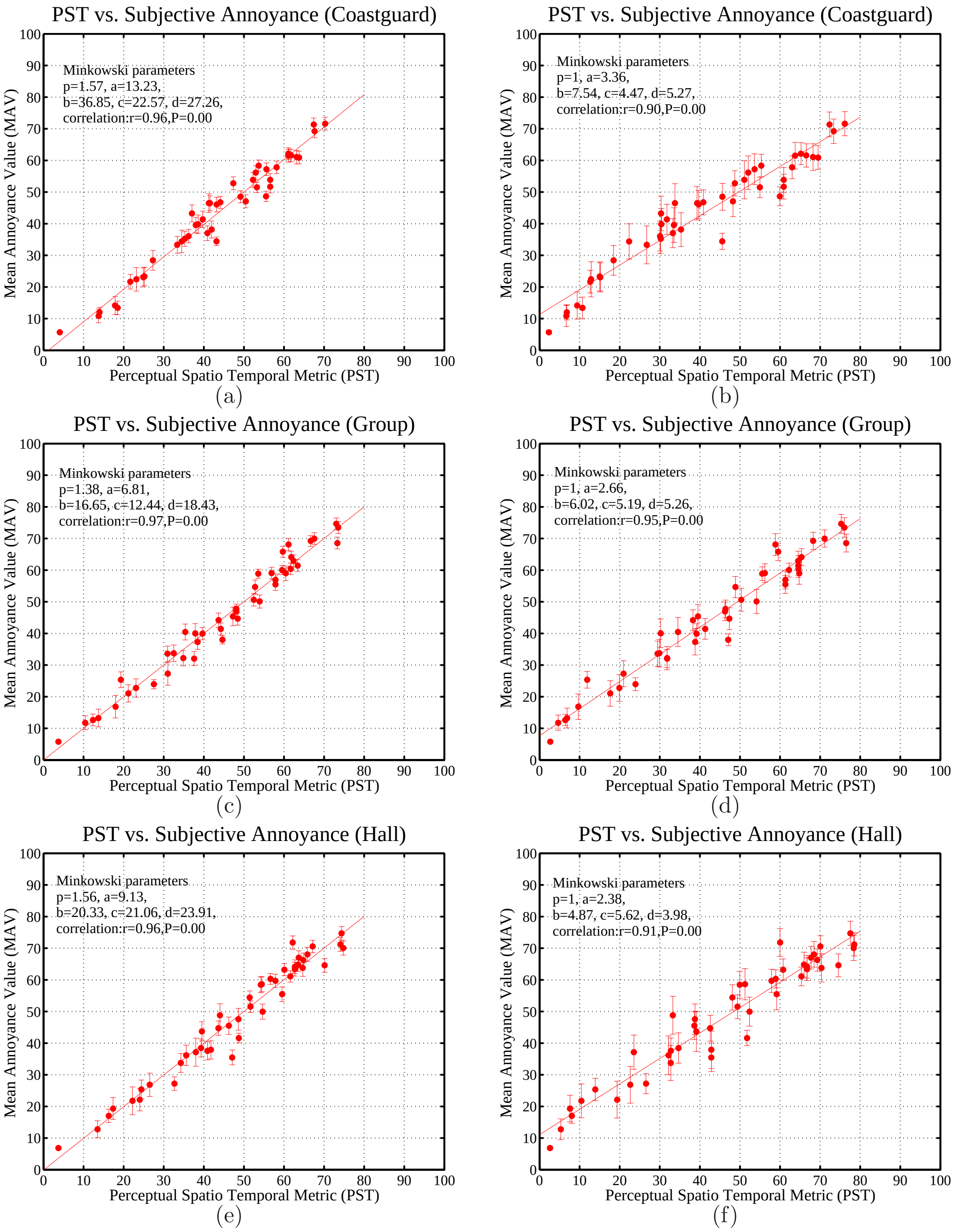
$$\mathbf{PST} = (a \cdot (\mathbf{PST}_{A_r})^p + b \cdot (\mathbf{PST}_{A_b})^p + c \cdot (\mathbf{PST}_{H_i})^p + d \cdot (\mathbf{PST}_{H_b})^p)^{\frac{1}{p}} \tag{5.36}$$

To find the Minkowski coefficients and exponent, we perform a nonlinear least-squares data fitting using the mean annoyance values (MAV) obtained from the subjective experiment of combined artifacts. The correlation coefficient and the Minkowski exponents and coefficients are reported in Tab. 5.10(a) for all the test sequences.

Figures 5.19 (a), (c), (e), (g) show the graphs of the MAV values versus the perceptual spatio temporal objective metric (PST) for the tested video sequences. We have also fitted the results for the linear model with **p=1** in Eq. (5.36) and the results are reported in Figs. 5.20 (b), (d), (f), (h). The graph of the correlation of MAVs versus **PST** for all the video sequences is reported in Fig. 5.19 (i). The correlation coefficients for these fits are reported in Tab. 5.10 (b). The linear model is simpler and more restrictive and, as it can be seen in the graphs in Fig. 5.19 (l), the correlation slightly decreases between the linear (l) ($r = 86$) and the more generic Minkowski model (i) ($r = 0.90$, Pearson = 0.95 , Spearman = 0.94). Moreover, there is not a significant difference between the objective model ($r = 0.90$) and the subjective model depicted in Fig. 5.18 ($r = 0.94$). Also in this case, the added regions weight is almost half of inside hole and added background and one third of border hole artifact (a = 11.36, b = 19.54, c = 26.58, d = 32.52).

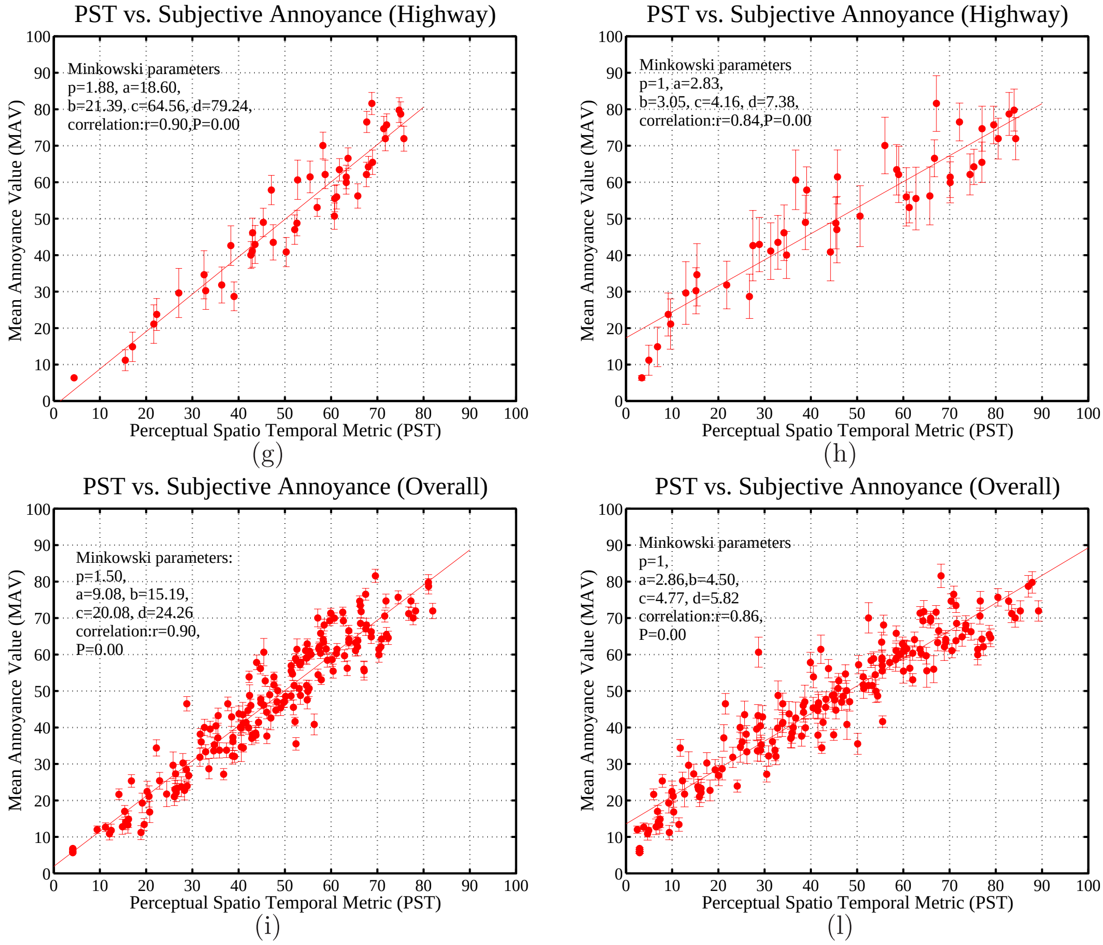
Although in theory different weighting functions could be used for each image size in order to attain the resolution independence, the anisotropic behaviors of pixel distances in rectangular grids makes it computationally difficult to realize. Therefore, a standard image resolution has been defined for comparison of result, namely CIF size (288 lines by 352 columns) and masks of other dimensions have to be scaled to CIF before the metric computation is performed. The viewing conditions are defined in the ITU Recommendations [65, 66].

In order to compare the results of the proposed method to the state of the art metrics, we ran the three metrics described in Sec. 5.2.2 on the 180 test sequences of the combined artifact experiment.



(a) (b) (c) (d) (e) (f)

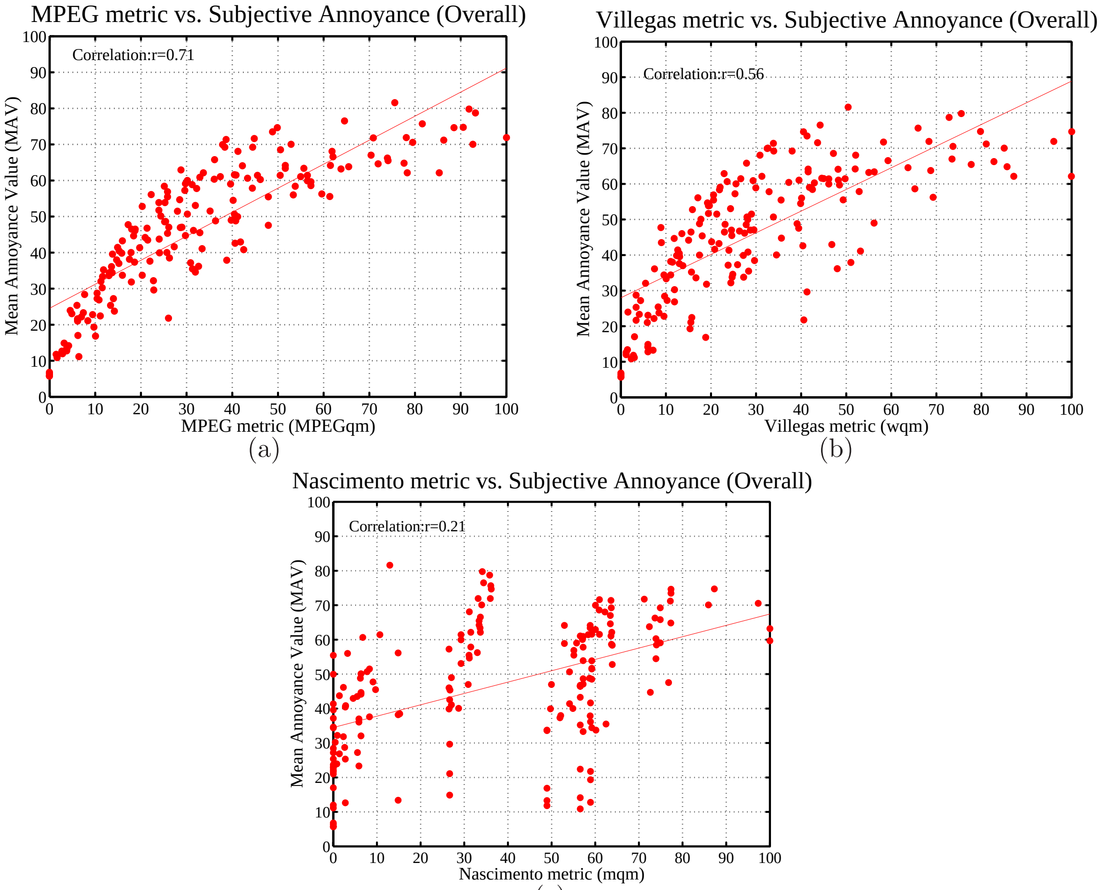
Figure 5.19: Correlation of the proposed perceptual metric **PST** with subjective annoyance: first row ‘Coastguard’, second row ‘Group’, third row ‘Hall monitor’. (a), (c), (e) with the optimal p and (b), (d), (f) with linear fitting $p = 1$.



(g) (h)

(i) (l)

Figure 5.20: Correlation of the proposed perceptual metric **PST** with subjective annoyance for 'Highway' and all the video sequences: (g) and (i) with the optimal p and (h) and (l) with linear fitting $p = 1$.



(a) (b) (c)

Figure 5.21: State of the art objective metrics vs. subjective annoyance: (a) MPEG metric, (b) Villegas' metric, (c) Nascimento's metric.

The state of the art metrics reported in Eq. (5.11) and tested are: the MPEG metric, **MPEGqm**, Villegas' metric, **wqm**, and Nascimento's metric, **mqm** .

Figures 5.21 (a), (b) and (c) show the graphs for respectively, **MPEGqm**, **wqm** and **mqm** metrics. The correlations coefficients are respectively: $r = 0.71$, $r = 0.56$ and $r = 0.21$. Our proposed method with a correlation of $r = 0.90$ outperforms the other state of the art metrics (see Fig. 5.19 (i)).

MPEG metric is the second best metric in fitting the subjective data. This result is surprising since no distinction between different kinds of error is applied in the MPEG metric in contrast with Villegas' and our metrics. However, it has to be mentioned that all the weights for the Villegas' metric in Eq. (5.10) are set the same and maybe by applying a tuning of them a better fit could have been obtained. However, if no specific application is specified, as in these subjective experiments, using equal weights is a good compromise.

As predicted in Sec. 5.2.2, the Nascimento's metric does not provide a good fit with the subjective data. In fact this metric is more suitable to predict object tracking quality than object segmentation quality. In our subjective experiments, subjects were told to judge in general the quality of segmentation without necessarily taking into special account the quality of object tracking. However it is interesting to notice in the graph of Fig. 5.19 (d) the different ranges of errors are grouped together by Nascimento's metric according to the prevalence of the kind of spatial error inserted in the test video sequences.

5.6 Conclusions

The first objective of this chapter is to present the existing methods to objectively evaluate the segmentation quality both for still images and video sequences. Their advantages and disadvantages are discussed and it is pointed out that none of them include the characterization of artifact perception in their models. Video object segmentation evaluation is tackled with particular focus. To this end, three state of the art metrics whose performance were analyzed are described in details. The first state of the art metric, **MPEGqm** is a simple sum of spatial and temporal errors commonly adopted by the research community. The second metric, **wqm** is a refinement of the first one where false positive and false negative errors are distinguished and weighted differently in the final formula. The third state of the art metric, **mqm** combines several simple metrics to classify the errors into split and merge errors, detection failures and false alarms. None of the state of the art objective methods includes the characterization of artifact perception in their models.

The second objective of the chapter is to propose a new objective metric which includes the study and characterization of segmentation artifact perception obtained by means of subjective experiments. Four spatial artifacts are deeply analyzed, namely added regions, added background, inside holes and border holes. Two temporal effects are studied, namely the temporal flickering and the expectation effect. Objective measures are proposed to estimate these artifacts. Through subjective experiments the objective measures are modeled by psychometric curves found to assess the annoyance of the artifact perception.

The subjective experiment results show that added region annoyance perception is not influenced by the shape or the position of the artifact but only by its size; the added background measure matches the human annoyance perception both when the artifact is uniformly distributed along the object boundaries and when it is concentrated in some parts of the object boundaries; inside hole for small sizes are more annoying than holes on the border, but on the other hand by increasing their size border holes become more annoying than inside holes as the shape of the object becomes less recognizable; the proposed flickering measure is more correlated to the subjective annoyance perception of such artifacts than the state of the art **MPEG** temporal metric; expectation effect is obtained in 5-second long test sequences and it consists in providing a good overall impression on assessing the quality of the sequence under test if a good segmentation is presented at the beginning and vice-versa. In the last subjective experiment, the relationship between the individual spatial artifact weights and the overall annoyance is found. The added region weight is half of the inside hole and added background weight, and one third of the border hole artifact weight (the most annoying artifact).

The final objective of this chapter is to propose an overall perceptual objective metric on the basis of the results above described. The performance of the new metric is analyzed in terms of correlation with subjective scores and compared to those of the three considered state of the art metrics. The perceptual objective metric results are comparable to the subjective annoyance model based on perceptual artifact strength and are definitely better than those of the state of the art metrics **MPEGqm**, **wqm** and **mqm**.

Table 5.7: *MSV*s and *MAV*s values for all segmented video sequences and all combination used in combined artifact experiment. the experiment.

Test	$\|\mathbf{A}_r\|$	$\|\mathbf{A}_b\|$	$\|\mathbf{H}_i\|$	$\|\mathbf{H}_b\|$	MSV_{A_r}	MSV_{A_b}	MSV_{H_i}	MSV_{H_b}	MAV
1	0	0	0	0	0.0877	0.0876	0.1021	0.1069	6.1536
2	180	0	0	0	7.0133	0.2924	0.1021	0.1356	22.6652
3	0	180	0	0	0.2645	5.9390	0.1021	0.1563	33.3475
4	0	0	180	0	0.0877	0.0843	6.8909	0.5516	38.7372
5	0	0	0	180	0.0877	0.0892	0.3071	6.6246	483527
6	60	0	0	0	3.8507	0.0922	0.1021	0.1749	16.9553
7	0	60	0	0	0.1372	3.6402	0.1021	0.1105	24.0348
8	0	0	60	0	0.0877	0.0770	3.8014	0.1016	27.4520
9	0	0	0	60	0.0877	0.0841	0.1366	4.1107	34.1164
10	24	0	0	0	1.9842	0.0899	0.1407	0.1857	12.5655
11	0	24	0	0	0.0877	2.0074	0.1021	0.1674	13.2082
12	0	0	24	0	0.0877	0.0761	2.9943	0.1733	24.1132
13	0	0	0	24	0.0877	0.0761	0.1520	2.7757	23.3064
14	60	60	60	60	3.3964	2.4635	3.2683	3.3934	58.4866
15	24	24	24	24	1.8564	1.4877	2.5851	2.2455	41.7692
16	60	60	0	0	3.2743	6.0577	0.1021	0.1737	44.4791
17	0	60	60	0	0.1309	3.3299	3.8570	0.3055	42.4672
18	0	0	60	60	0.0877	0.0761	3.6815	3.5309	44.3497
19	60	0	0	60	3.7006	0.0761	0.1313	3.7160	39.3042
20	60	0	60	0	3.7049	0.0831	3.6841	0.1292	35.6497
21	0	60	0	60	0.0927	3.4445	0.3646	3.6124	46.0307
22	180	0	24	0	7.0127	0.2083	2.8334	0.1115	38.5573
23	0	180	0	24	0.2002	5.9421	0.1780	2.3034	43.4045
24	24	0	180	0	1.8346	0.1310	6.4809	0.5970	47.3065
25	0	24	0	180	0.0985	1.5889	0.4589	6.4816	55.1360
26	180	24	24	0	7.0244	1.7153	2.6863	0.2761	42.6581
27	0	180	24	24	0.1390	5.8701	2.8902	2.0371	54.4130
28	24	0	180	24	1.8309	0.0777	6.3907	2.5492	53.3266
29	24	24	0	180	1.9844	1.1578	0.2771	6.4670	60.3913
30	180	0	60	60	6.2229	0.2530	3.2891	3.5029	53.1470
31	60	180	0	60	3.3900	5.6682	0.2185	3.2926	60.0855
32	60	60	180	0	3.1897	5.6819	6.3051	0.7411	61.4938
33	0	60	60	180	0.0877	2.5356	3.1930	5.9815	69.6631
34	60	60	60	180	3.5758	2.3133	3.0205	6.0809	69.4016
35	180	60	60	60	6.6470	2.5376	3.1695	3.2051	62.6772
36	60	180	60	60	2.9195	5.3585	3.1376	3.0265	66.8558
37	60	60	180	60	2.9301	5.1584	5.6035	3.6762	70.8286
38	24	180	24	24	1.8480	5.6422	2.9430	2.1074	58.5226
39	24	180	24	24	1.7746	5.8348	2.5564	2.0194	58.1906
40	24	180	24	24	1.9364	1.3128	6.3864	2.2994	59.2983
41	24	24	24	180	1.9999	1.0205	2.6602	6.2791	65.9693
42	180	24	180	60	6.3306	1.0030	5.9296	3.2082	65.6139
43	60	180	24	180	3.2002	5.5115	2.6920	5.8289	73.1322
44	180	60	180	24	6.1484	2.2794	6.1829	2.3050	63.5157
45	24	180	60	180	1.6707	5.2750	2.9312	5.9242	74.8563

Table 5.8: Subjective fitting for combined artifact experiment: Minkowsi fitting parameter and correlation values for all the video sequences.

Video	**p**	**a**	**b**	**c**	**d**	**r**
Coastguard	1.33	7.35	11.13	10.25	12.46	0.96
Group	1.18	3.49	7.36	6.69	10.86	0.96
Hall monitor	1.37	5.73	13	12.64	16.61	0.95
Highway	1.44	6.94	12.31	19.24	27.52	0.94
All	1.32	5.50	10.35	10.81	15.30	0.94

Table 5.9: F values to test if different fitting curves are needed to describe the perceived annoyance for different shapes and positions of added regions.

Artifact	**Normalization**	**r**
$\overline{PST}_{A_r}$	$0.16 \cdot PST_{A_r} - 0.15$	0.95
$\overline{PST}_{A_b}$	$0.25 \cdot PST_{A_b} + 0.15$	0.92
$\overline{PST}_{H_i}$	$0.08 \cdot PST_{H_i} - 0.33$	0.94
$\overline{PST}_{H_b}$	$0.06 \cdot PST_{H_b} - 0.28$	0.90

Table 5.10: Minkowski parameters and correlation (a) best p fitted, (b) p=1.

Video	**p**	**a**	**b**	**c**	**d**	**r**
Coastguard	1.57	13.23	36.58	22.57	27.26	0.96
Group	1.38	6.81	16.65	12.44	18.43	0.97
Hall monitor	1.56	9.13	20.33	21.06	23.91	0.96
Highway	1.88	18.60	21.39	64.56	79.24	0.90
All	1.60	11.36	19.54	26.58	32.52	0.90

(a)

Video	**a**	**b**	**c**	**d**	**r**
Coastguard	3.36	7.54	4.47	5.27	0.90
Group	2.66	6.02	5.19	5.26	0.95
Hall monitor	2.38	4.87	5.62	3.98	0.91
Highway	2.83	3.05	4.16	7.38	0.84
All	2.86	4.50	4.77	5.82	0.86

(b)

Objective Metric Performance and Applications

6

6.1 Introduction

In the previous chapter, we described a set of experiments on segmented video sequences with embedded synthetic artifacts. In this chapter we want to test our metric on real artifacts generated by typical segmentation algorithms. We present both objective and subjective studies of the annoyance generated by the real artifacts introduced by video object segmentation algorithms. The metric proposed in this thesis and some state of the art metrics are used to assess the objective quality of these segmentation algorithms. Subjective experiments are performed to validate their correlation with Mean Opinion Scores (MOS).

To the best of our knowledge, a comparison among different objective metrics for video segmentation quality assessment has received little attention by the image processing community so far, as well as the study of their performances on real segmentation algorithms. This study considers both a *general* and an *application* dependent segmentation scenarios, while in the literature no application has been taken into account.

The selected segmentation algorithms are presented in Sec. 6.2 and a general framework is considered to discuss the correlation between the subjective and objective results. In the second part of this chapter, some of the most common applications of video object segmentation are illustrated. In Sec. 6.3, subjective experiments are proposed for different applications. According to the particular application, subjective results are correlated to objective metric results to test their performances. Section 6.4 draws the conclusions.

6.2 Segmentation Assessment - General Framework

As underlined in the previous sections, segmentation of objects in image sequences is a crucial task for a wide variety of multimedia applications. The ideal goal of segmentation is to identify the semantically meaningful components of an image and to group the pixels belonging to such components. While it is very hard to segment static objects in images, it is easier to segment moving objects in video sequences. Once the moving objects are correctly detected and extracted, they can serve for a variety of purposes. In this section, we do not focus on any of the specific purposes of segmentation but provide a general framework. Human viewers are asked to assess the quality of segmented objects for "general purposes" as we did in the two previous chapters. In order to assess if a segmentation is good for general purposes, we asked viewers to mentally compare the results of the segmentation at hand to the ideal (reference) segmentation and formulate their judgments.

Since studying how subjective scores change in relation to the specific segmentation task at hand provides a lot of interesting insights in developing evaluation metrics, in the next section, application dependent segmentation results will be analyzed.

In Sec. 6.2.1, we describe the segmentation algorithms used in testing the performance of the objective metric. The subjective experiment and the data set are presented in Sec. 6.2.2. Section 6.2.3 reports the experimental results. The objective results versus the subjective ones are analyzed in terms of correlation coefficients and our metric is compared to the results obtained using state of the art metrics.

6.2.1 Video Object Segmentation Algorithms

In our experiments we have used seven static background segmentation methods. In the following, the principles on which each technique is based are reported. For further details the reader is invited to refer to the corresponding paper [52, 62, 67, 71, 94, 132, 138]. Fine tuning of parameters has been done on a small data set of each algorithm according to subjective evaluation criteria. Then parameters are left untouched for the remaining image sequences in the test data.

The approaches of the tested algorithms differ in using various features such as color, luminance, edge, motion and combinations of them.

For example, a segmentation method that uses only the edge information is the technique proposed by:

Kim et al. [71] extracted the difference edge map between consecutive frames. This approach is based on gray scale images and it applies the Canny edge operator to the current, background, and successive frames. The motion information obtained by the difference edge map is used for selecting the relevant edges from the current frame. The object mask is achieved by filling the boundaries received by the previous edge results with connecting the first and second occurred edge pixels for each vertical and horizontal line, respectively.

There are three methods based on the color information analysis:

Horprasert et al. [62] used color and illumination information. This method evaluates for each pixel the brightness and the chromaticity distortions between the background image and the current frame. The background image is therefore modeled by four values: the mean and the standard deviation over several background frames and the variation of the brightness and chromaticity distortions. The current frame is subtracted from the modeled background image and each pixel is classified as *original background*, *shadow*, *highlighted background* and *foreground*.

François and Medioni [52] operated in the Hue-Saturation-Value (HSV) color space and models the background pixels by using the mean and standard deviation and updating these values at each frame. The current frame pixels are compared to those of the updated background. The V value is always used and the color information H and S are used in the regions where they are evaluated to be reliable.

Shen et al. [138] used two color spaces: Red-Green-Blue (RGB) and Hue-Saturation-Intensity (HSI). In the first color space the moving objects are detected by using a fuzzy segmentation for each color channel while in the second, shadows are eliminated by using the difference between consecutive frames. This algorithm performs well also under varying illumination conditions since it considers changes of successive frames and updates the background model.

Two of the considered methods use both color and edge information:

Jabri et al. [67] modeled the background in two parts: the color model and the edge model. The background model is trained in both mentioned parts by calculating the mean and standard deviation for each pixel of any color channel. The edge model is built by applying the Sobel edge operator for both horizontal and vertical case. With subtraction of the incoming current image on each channel, confidence maps are generated for both information color and edge. After that a combination of the two maps are utilized by taking its maximum values, a single median filtering step is applied to the resulting confidence map to fill holes and remove isolated pixels. At least this output goes through a hysteresis thresholding for binarization.

McKenna et al. [94] proposed a method similar to *Jabri*, since they use the same information (color and edge) to build the background but introduce a new color model to eliminate cast shadows. In fact, with this method each pixel's chromaticity is modeled using the means and variances of the normalized RGB color components [47]. Therefore, any significant intensity change without significant chromaticity change is detected as shadow.

Finally there is one method that does not correspond to any of the previously defined classes:

Image Differencing by Rosin [132] was the first method applied for video segmentation in case of static or motion compensated camera conditions and is based on basic background subtraction. Gray scale images are used and the results depends only on the applied thresholding method. The segmentation results differ very much since the threshold value is sensitive to environmental conditions, e.g. due to similar colors, illumination changes.

6.2.2 Subjective Experiment Results

The experimental methodology corresponds to the five-step procedure described earlier in Sec. 4.4: oral instructions, training, practice trials, experimental trials and interview.

The test group was composed of 35 subjects aged between 23 and 41 (with 8 females) who performed the annoyance task (see Sec. 4.4.2). The textured video objects have been overlapped on a uniform gray background ($Y = 127$, $U = 127$, $V = 127$). Three original video sequences used in this experiment were 'Highway', 'Group' and 'Hall monitor' (sample frames are shown in Figs. 4.10 (a), (b), (c)). The seven segmentation algorithms described in the previous section have been applied to each original video sequences. A total number of 24 sequences was generated: 21 test segmented sequences (3 original video sequences × 7 segmentation algorithms) plus the 3 reference segmentations of 'Hall monitor', 'Highway' and 'Group' shown in Fig. 6.2.

In Tab. 6.1 the gathered Mean Opinion Scores for the Annoyance Values (MAV) are reported for all the video sequences and algorithms along with the 95% confidence interval δ.

The results of the subjective experiments averaged for all the three video sequences are depicted in bar-graph of Fig. 6.1. In this graph the averaged MOS Annoyance values (MAV) have been plotted for each real segmentation algorithm and the reference segmentation. The subjective results show that the algorithms which on average introduce the most annoying artifacts are the *Kim* (see Fig. 6.3) and *Image Differencing* (see Fig. 6.4). The least annoying artifacts are generated by the *Jabri* (see Fig. 6.8) and *Shen* (see Fig. 6.9). .

Table 6.2 reports the subjective ranking of the tested algorithms from the most annoying to the least annoying and a brief description of the artifacts that are typically introduced. As described in Tab 6.2, the most annoying artifact is flickering. It is usually due to noise, camera jitter and varying illumination, and consists in erroneously segmented regions that are different at each frame. A high value of flickering of added regions is generated by *Kim*'s algorithm and as it has been already

Table 6.1: MAV values obtained for each segmentation algorithm and the correspondent confidence interval δ for all the test video sequences.

	'Group'		'Hall monitor'		'Highway'	
Segmentation	MAV	δ	MAV	δ	MAV	δ
Reference	8.77	2.94	26.74	6.89	15.31	5.22
François	68.57	8.56	61.43	7.52	30.20	6.76
Horprasert	69.94	6.79	57.57	7.18	32.06	6.83
Image Differencing	99.74	4.94	60.00	6.94	67.54	7.29
Jabri	57.46	7.81	40.37	7.11	37.94	7.01
Kim	72.00	7.05	86.89	6.52	71.14	7.43
McKenna	83.36	5.80	56.86	7.47	54.26	8.15
Shen	57.83	7.59	55.26	7.48	54.26	7.16

Table 6.2: Description of artifacts introduced by the real segmentation algorithms and their perceived strengths gathered in the interview stage.

Algorithm	Artifacts	Strength
Kim	added regions	high
	added background	high
	flickering	high
Image Differencing	inside holes	high
	border holes	high
	flickering	medium
McKenna	inside holes	medium
	border holes	medium
	flickering	medium
François	added background	high
Horprasert	border holes	medium
Jabri	added regions	medium
	added background	low
Shen	added background	low
	border holes	low

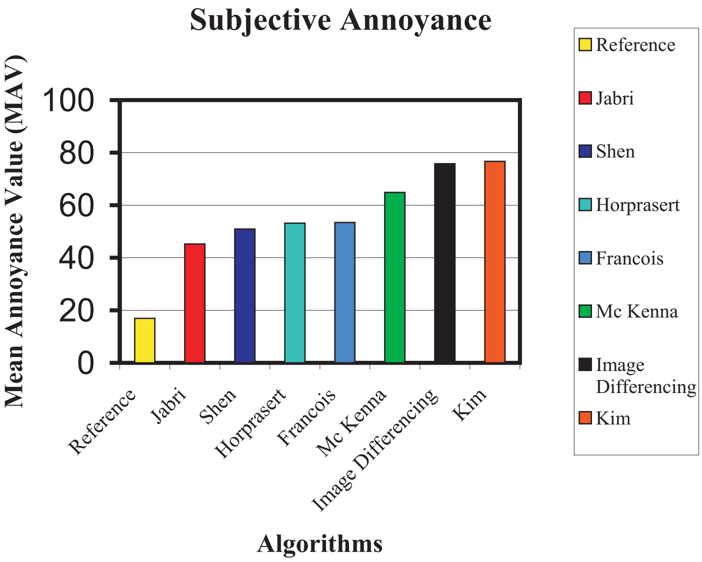


Figure 6.1: MAV values obtained for each segmentation algorithm and averaged on the three tested video sequences.

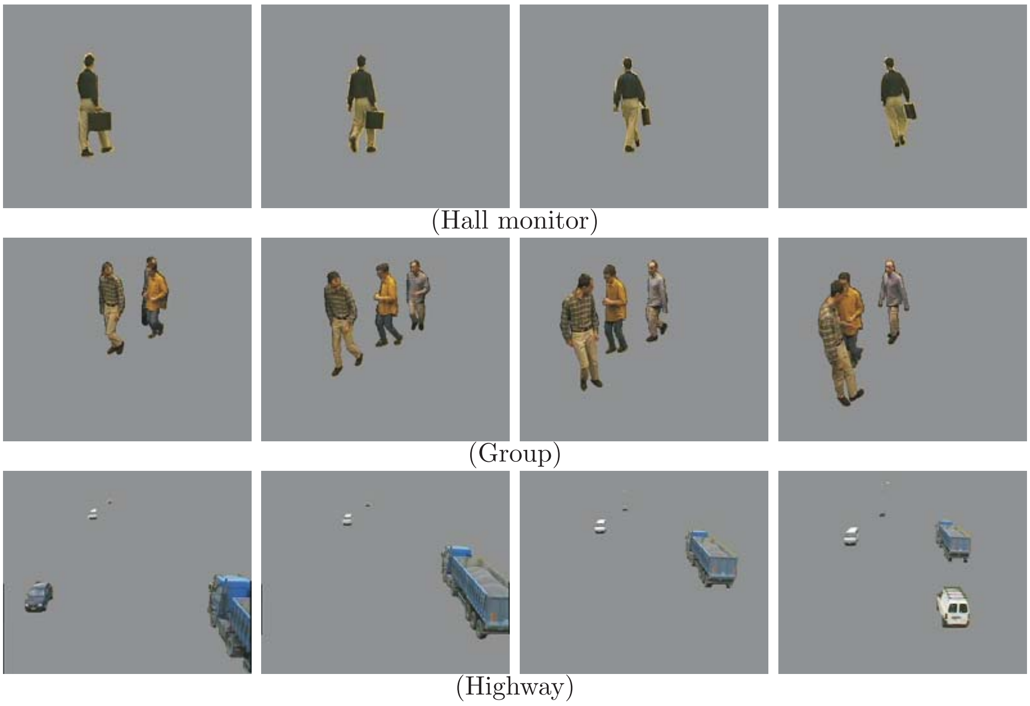
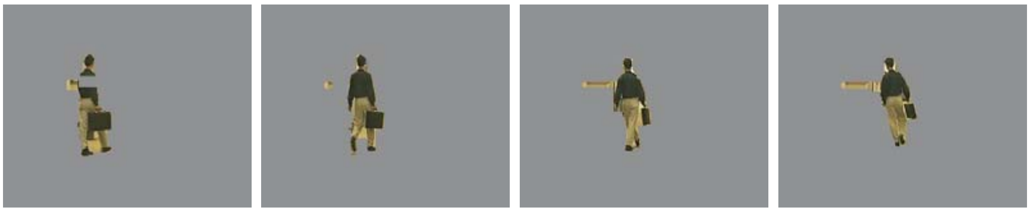
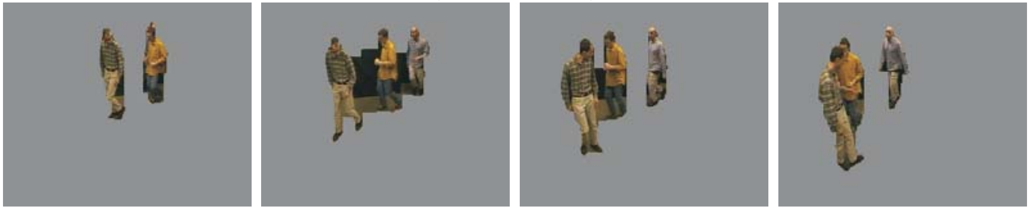


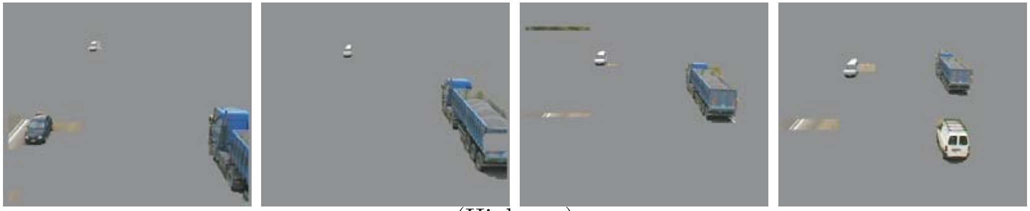
Figure 6.2: Sample frames for reference segmentation of the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105).



(Hall monitor)



(Group)



(Highway)

Figure 6.3: Sample frames for the *Kim*'s algorithm segmentation results for the tested video sequences: ‘Hall monitor’ (frames #40, #50, #60,#70), ‘Group’ (frames #90, #100, #110, #120), ‘Highway’ (frames #75, #85, #95, #105).

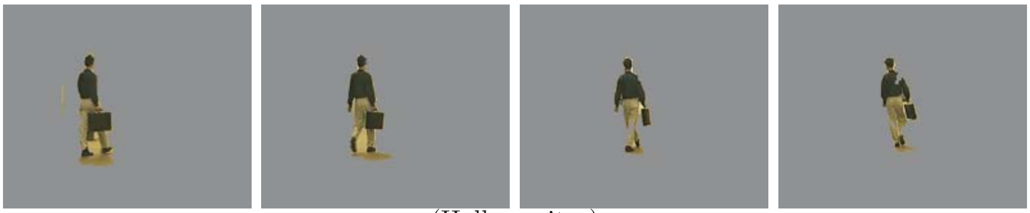
pointed out (see subjective experiments reported in Sec. 5.4.2), it is the most annoying artifact (see Fig. 6.3). In fact, no matter what the size of the artifact is, if the segmentation presents temporal instabilities it will annoy the subject a lot more than any other spatial artifacts.

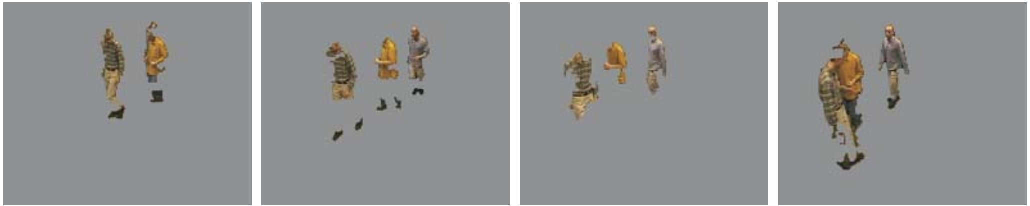
In Tab. 6.2, the second annoying artifact is that one introduced by *Image Differencing* (see Fig. 6.4) due to the large amount of holes and especially border holes. As we commented in the combined artifact experiment discussed in Sec. 5.4.3, this has the biggest weight in terms of annoyance. It is usually due to the algorithm's failures in differentiating the foreground regions from the background since they look very similar in color or texture or other uniformity features that the algorithm exploits to perform the segmentation.

Then the artifacts introduced by *McKenna*'s algorithm (see Fig. 6.5) are rated as the third most annoying. In this case, especially the holes are annoying to human observers, even if they are smaller than those introduced by the *Image Differencing*'s method, but still of considerable amount.

Added background is the fourth annoying artifact and is generated by *François*'s algorithm (see Fig. 6.6). It is mostly caused by erroneously detecting moving shadows as part of the moving foreground objects. Since shadows move along with objects from which they are cast, we observed that this artifact does not annoy too much the human observer and it is subjectively rated better than flickering or missing parts of objects in this general scenario.

The least annoying artifacts are those introduced by the *Horprasert*'s (see Fig. 6.7), the *Jabri*'s (see Fig. 6.8) and the *Shen*'s algorithms (see Fig. 6.9). In fact, these algorithms introduce smaller amounts of artifacts compared to others.


(Hall monitor)


(Group)

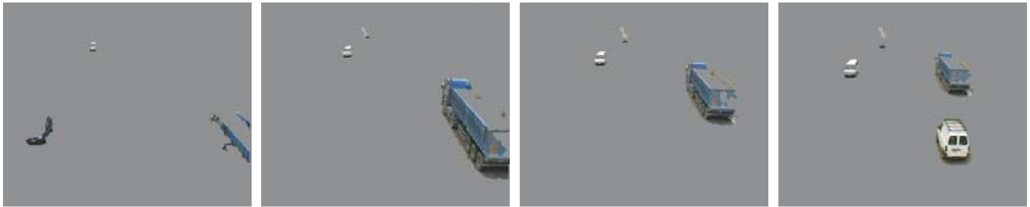
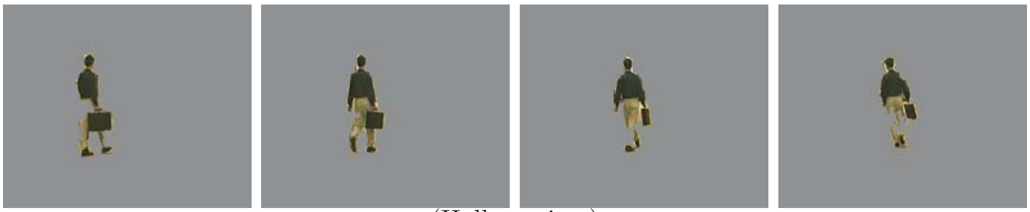
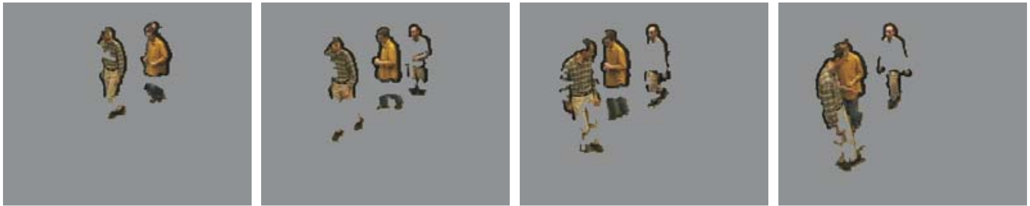

(Highway)

Figure 6.4: Sample frames for the *Image Differencing* algorithm segmentation results for the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105).


(Hall monitor)


(Group)

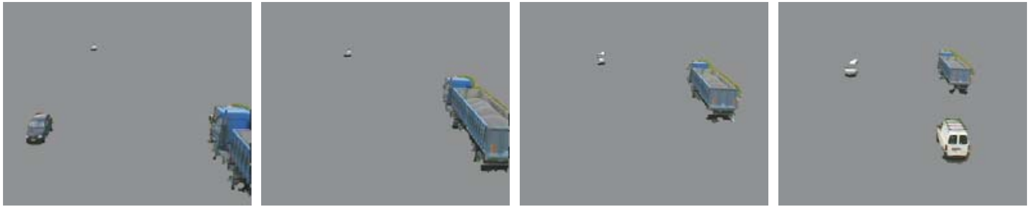
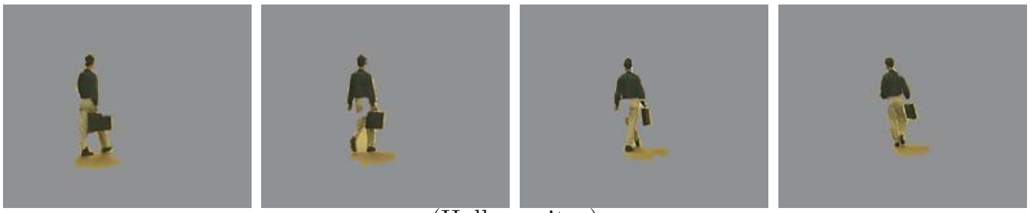
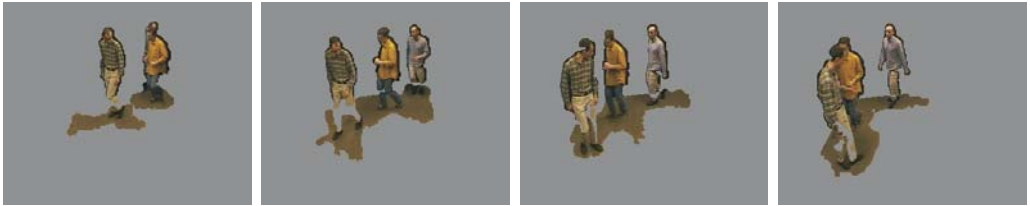

(Highway)

Figure 6.5: Sample frames for the *McKenna*'s algorithm segmentation results for the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105).


(Hall monitor)


(Group)

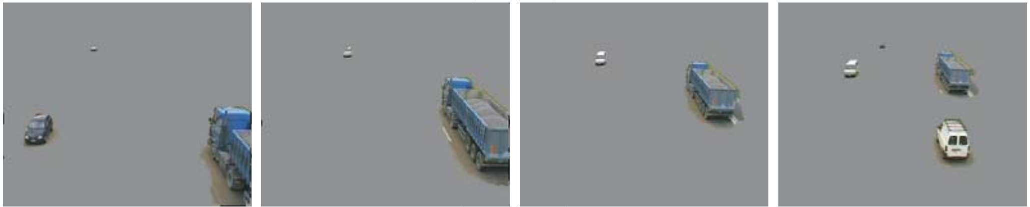
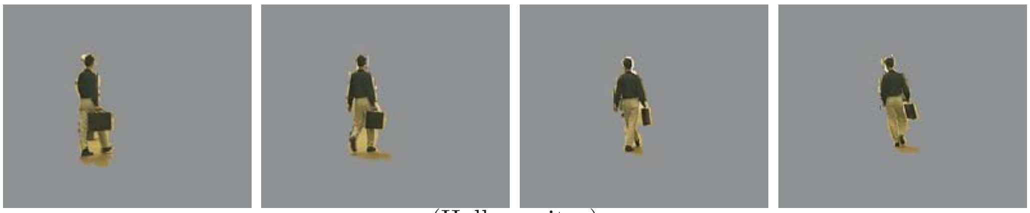
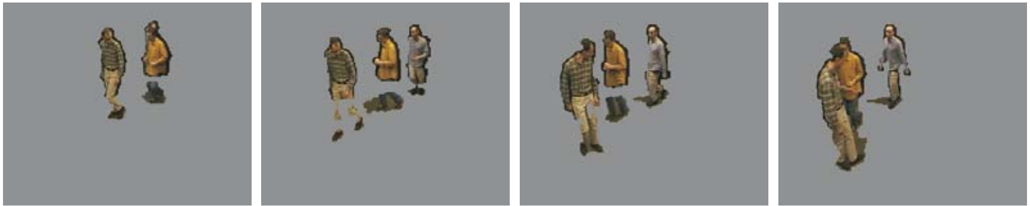

(Highway)

Figure 6.6: Sample frames for the *François*'s algorithm segmentation results for the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105).


(Hall monitor)


(Group)

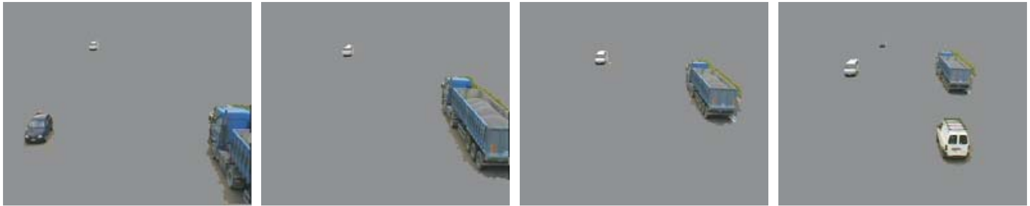
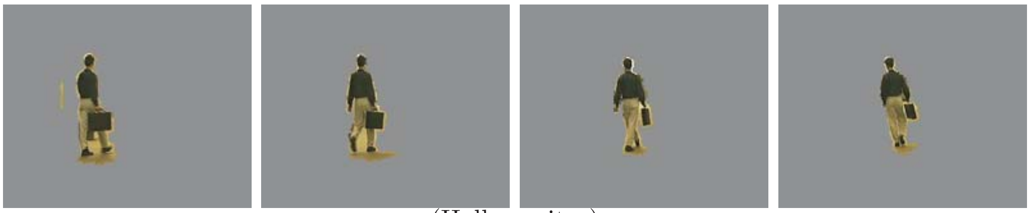
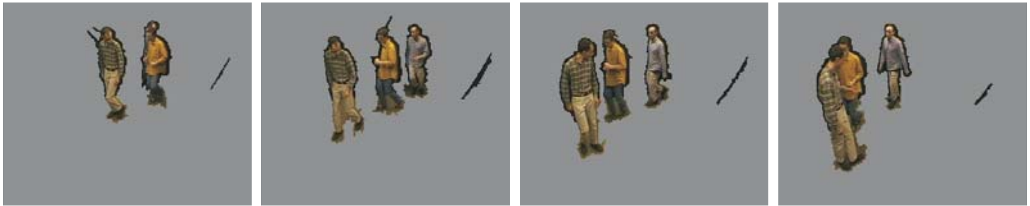

(Highway)

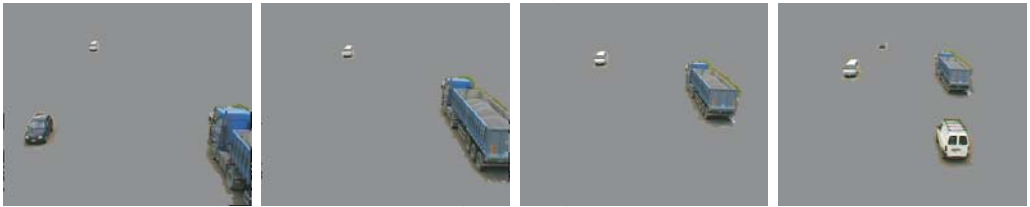
Figure 6.7: Sample frames for the *Horprasert*'s algorithm segmentation results for the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105).



(Hall monitor)

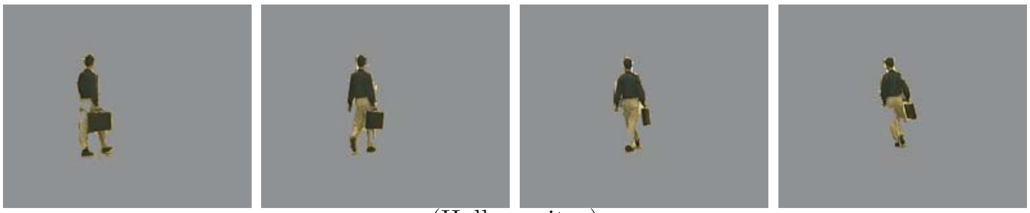


(Group)

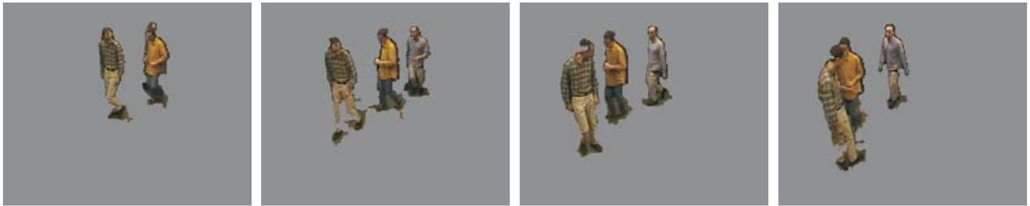


(Highway)

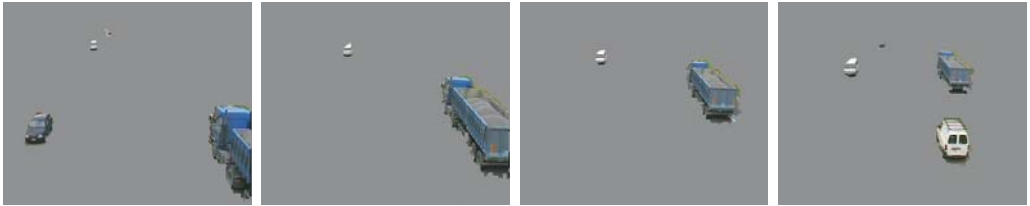
Figure 6.8: Sample frames for the *Jabri*'s algorithm segmentation results for the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105).



(Hall monitor)



(Group)



(Highway)

Figure 6.9: Sample frames for the *Shen*'s algorithm segmentation results for the tested video sequences: 'Hall monitor' (frames #40, #50, #60,#70), 'Group' (frames #90, #100, #110, #120), 'Highway' (frames #75, #85, #95, #105).

6.2.3 Objective Metric Results

We have proposed a perceptual metric **PST** based on subjective experiments with synthetic segmentation artifacts. In this section, we aim at testing the performance of **PST** on the real segmentation artifacts produced by the above mentioned algotihms and compare its performance to those of the other objective metrics presented in the previous chapter: **MPEGqm**, **wqm** and **mqm** (see Sec. 5.2.2).

First, we apply the objective metrics to the seven algorithms segmentation results to obtain the objective evaluations, then we compare these results to those obtained by subjective evaluations and we look at the correlation between the two kinds of data: subjective and objective.

Our perceptual objective metric, explained in Chapter 5, is given by a linear combination of the perceptual metrics for four kinds of artifacts: added region A_r, added region A_b, border holes H_b and inside holes H_i and for sake of convenience is written below:

$$\mathbf{PST} = 2.86\cdot(\mathbf{PST}_{A_r}(\mathbf{ST}_{A_r}))+4.50\cdot(\mathbf{PST}_{A_b}(\mathbf{ST}_{A_b}))+4.77\cdot(\mathbf{PST}_{H_i}(\mathbf{ST}_{H_i}))+5.82\cdot(\mathbf{PST}_{H_b}(\mathbf{ST}_{H_b})), \tag{6.1}$$

where the perceptual metric $\mathbf{PST}_{A_r}$ is given by Eq. (5.25), $\mathbf{PST}_{A_b}$ by Eq. (5.27), $\mathbf{PST}_{H_b}$ by Eq. (5.31) and $\mathbf{PST}_{H_i}$ by Eq. (5.30). The objective metrics **ST** for each artifact are computed as described in Sec. 5.3. The only difference with the metric proposed in the previous chapter is with regards with the temporal weights in Eq. (5.21). Dueto the different environment in which the subjective tests were carried out for this specific experiment, the experimental conditions were found to be slightly different. In fact, the last frame of the segmented video sequence under test remained on the display while the human observer was making his/her judgment. This fact slightly conditioned the human scores and the last frames had more impact on the overall annoyance with respect to the initial ones. Thus, the temporal weights were modified as follows to model this effect:

$$\mathbf{ST} = \frac{1}{K}\cdot\sum_{k=1}^{K}(a\cdot e^{\frac{k-30}{b}}+c)\cdot\mathbf{ST}_{artifact}(k), \tag{6.2}$$

with $a = 0.02$, $b = 7.8$, $c = 0.0078$, and $K = 60$ chosen empirically.

For future subjective experiments, it is advisable to make the last frame of the segmented video sequence disappear after its display to prevent the subject to excessively focus on the last few frames.

The results obtained with the proposed metric, **PST**, for all the video sequences and algorithms are depicted in Fig. 6.10 (a). The results for the other three state of the art metrics are shown in Figs. 6.10 (b), (c) and (d). The correlation coefficients for our metric are larger (Pearson = 0.86, Spearman = 0.79) compared to those of the others objective metrics: **MPEGqm** (Pearson = 0.73, Spearman = 0.67), **wqm** (Pearson = 0.69, Spearman = 0.71) and **mqm** (Pearson = 0.53, Spearman = 0.44). Our metric is consistenly defined on the annoyance scale between 0 and 100.

Since the final goal for an objective metric is to help choosing the best algorithm the best performing algorithm on a given set of data, we have considered, for each algorithm, the $MAVs$ and the objective metric values and averaged them on all the three video sequences to obtain a global subjective and objective judgment on the algorithm under test. From now on, we focus our discussion on the *averaged* results. The *averaged* MAV values are reported in Tab. 6.3. Our purpose was to identify which are the objective metrics that predict more reliably which algorithms perform better on subjective scales. Thus, we computed the correlation coefficients of the *averaged* MAV versus **PST**, **MPEGqm**, **wqm** and **mqm**. Table 6.3 provides these correlation coefficients. From this table, it can be noticed that our metric is able to predict better the *averaged* performance of the segmentation algorithms in terms of subjective assessment. In fact, the correlation coefficients

Table 6.3: Correlation coefficients for the averaged objective metrics for all the sequences vs. averaged subjective annoyance.

	Correlation	
Objective Metric	Pearson	Spearman
PST	0.91	0.81
MPEGqm	0.82	0.66
wqm	0.84	0.71
mqm	0.83	0.76

(Pearson = 0.91, Spearman = 0.81) are larger than those of the other state of the art objective metrics reported in Tab. 6.3. In this table, we also see that Villegas' metric **wqm** seems to perform better than **MPEGqm** metric. This shows the utility of introducing different categories of pixel errors which is not done for **MPEGqm**.

Our evaluation metric has been proposed for general purpose segmentation. Therefore, during the subjective experiments, the subject had in mind an ideal segmentation which was displayed during the training stage.

Segmentation is an ill-posed problem if no application is defined. Therefore, it is important when evaluating the performance of an algorithm to have *a priori* knowledge on the specific application it is addressing. In the next section we explore how segmentation is perceived differently according to particular applications and how the objective metrics perform in such cases.

6.3 Segmentation Assessment - Application Dependent

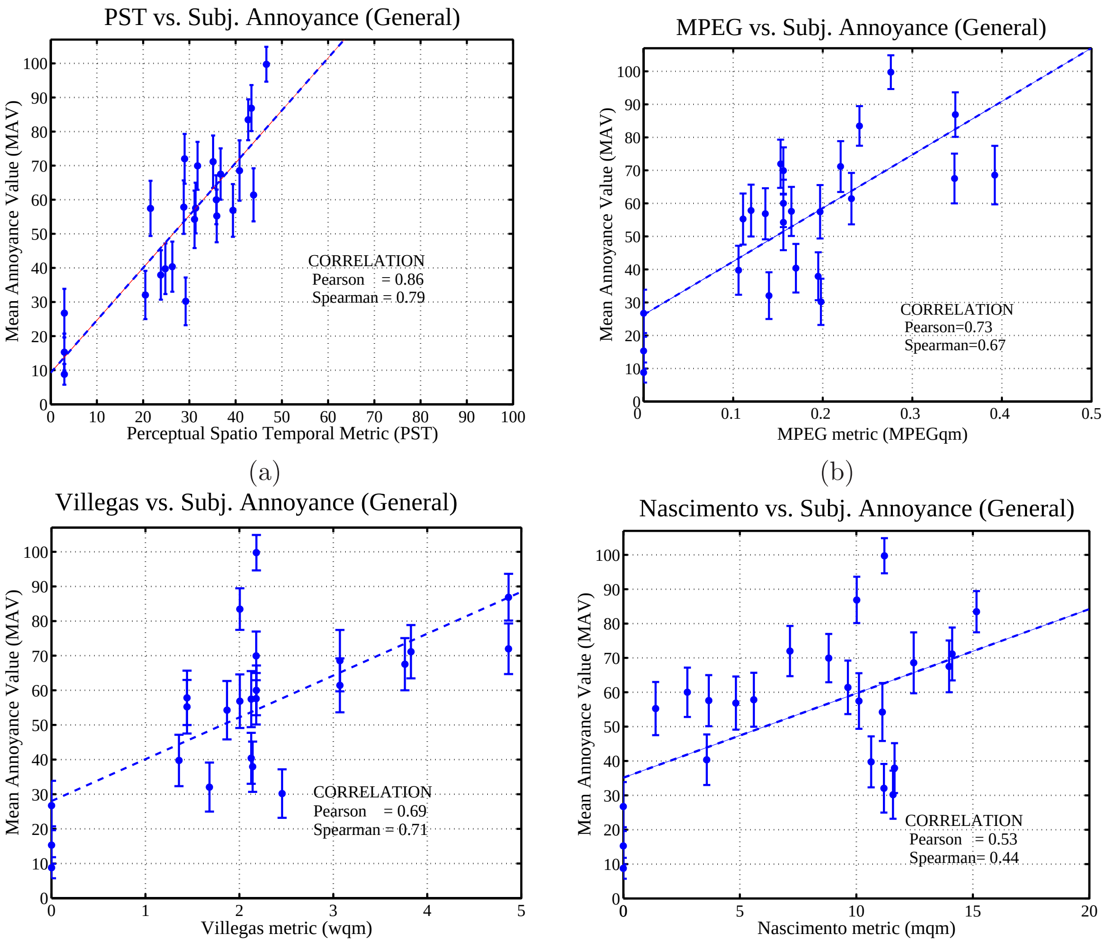
The expected segmentation quality for a given application can often be translated into requirements related to the shape precision and the temporal coherence of the objects to be produced by the segmentation algorithm. Video sequences segmented with high quality should be composed of objects with precisely defined contours, having a perfectly consistent partition along time.

A large number of video segmentation applications can be considered and typically they have different requirements. A full classification of segmentation applications into a set of scenarios, according to different application constraints and goals can be found in [24]. The setting up of a subjective experiment differs for each kind of application. Therefore, we have focused our experiments on three kinds of applications that are described in Sec. 6.3.1.

Section 6.3.2 presents how the subjective experiments have been carried out differently for each specific application. The correlation between the subjective scores and the objective results are analyzed in Sec. 6.3.3. In that section, an analysis is carried out to determine how to tune the metric parameters according to the specific application.

6.3.1 Video Object Segmentation Applications

In coding, the development of segmentation techniques for moving objects has mostly been driven by the so-called second generation coding [140, 141] (see Sec 4.2). The second generation coding techniques use image representations based on the Human Vision System (HVS) rather than the conventional canonical form which is based on the concept of pixel or block of pixels as the basic entities that are coded. As a result of including the human visual system, natural images are treated as being not a composition of objects defined by a set of pixels regularly spaced in all



(a) (b)

(c) (d)

Figure 6.10: Objective metrics vs. Subjective Scores (MAV) and correlation coefficients.

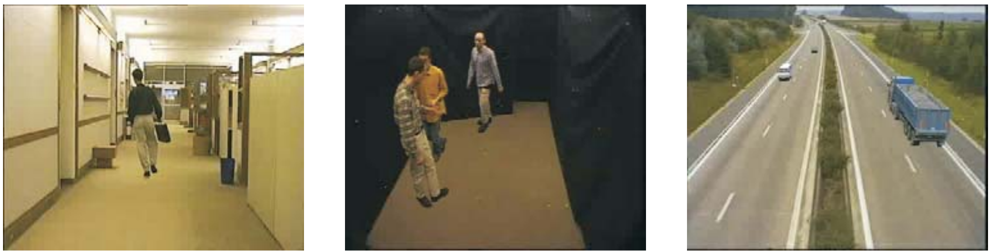

(Hall monitor) (Group) (Highway)

Figure 6.11: Sample frames for video coding segmentation applications ‘Hall monitor’, ‘Group’, and ‘Highway’.

dimensions but by their shape and color. MPEG-4 standard supports the coding of video sequences that are pre-segmented based on video content to allow a separate and flexible reconstruction and the manipulation of content at the decoder. Moreover, there are some special scenarios in which automatic detection of moving video objects is strongly required. For istance, it is fundamental in object-based video surveillance systems, which need to be implemented with some event detection schemes. Thus, a prior decomposition of sequences into video objects becomes an important issue in video analysis, video coding and video manipulation applications. We have chosen one application for each of these three broad fields.

Video coding

The different segmentations (see Sec. 6.2.1) of the scene into meaningful objects have been tested in the *compression* scenario. Segmentation can improve the coding performance over a low-bandwidth channel. We have used the MPEG-4 codec scheme in the object-based video compression mode to compress objects separately from the rest of the scene. This application is useful for applications where the bandwidth is limited. In fact, the compression rate of the foreground objects and background can be a function of their different importance in the scene and a lowered bit rate at the same perceived quality can be obtained. The decomposition can be obtained with an object-based coder (*object based mode*) as well as with a traditional coder (*frame based mode*). In order to evaluate different segmentation results in a video coding application, the MPEG-4 encoder (Miscrosoft’s MPEG VM software encoder & decoder[*]) was used in the experiments. According to previous work on object based coding [153], we compressed a single background image for each test sequence using MPEG-4 frame based-coding and the sequences of segmented foreground objects for each algorithm using MPEG-4 object-based coding. All the quantization parameters Q for the background coding were chosen to be equal to 10 [153]. Since we only want to study the segmentation artifact perception, the compression artifacts were not included the test sequences. The segmented video objects were not compressed. In such a way, the compressed background can be transmitted only once and the video objects corresponding to the foreground (moving objects) can be transmitted and added on top of it so as to update the scene. Samples of object-based coding test sequences is shown in Fig.6.11. Eight types of segmentation maps are applied, including the ideal (reference) segmentation obtained by hand and the results of the seven segmentation algorithms described in the previous section.

---

[*] Version: FDAMI 2-3-001213, integrator: Simon Winder, Microsoft Corp.

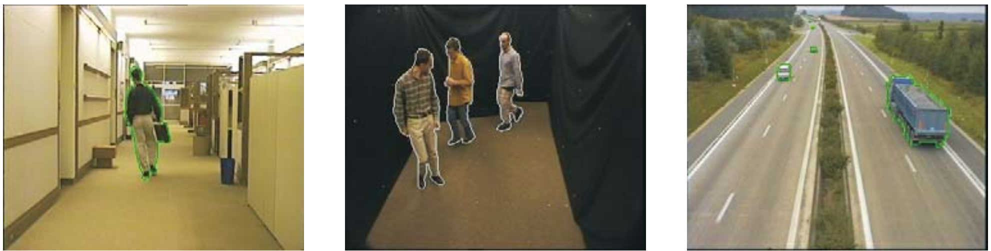

(Hall monitor) (Group) (Highway)

Figure 6.12: Sample frames for video surveillance segmentation application ‘Group’, ‘Hall monitor’ and ‘Highway’.

Video analysis

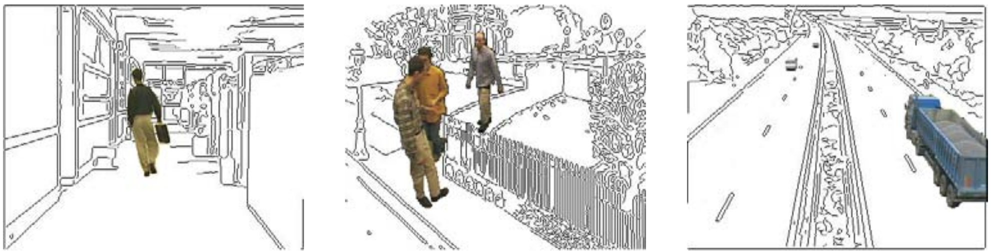
Video surveillance is a particular application of video analysis [18]. Until recently, surveillance has been performed entirely by human operators, who interpreted the visual information presented to them on one or more monitors. Sometimes, the fatigue due to several work hours compromised the ability to give the alarm appropriately. Therefore, in the last decade, a considerable effort has been devoted to developing automatic or semi-automatic video-based surveillance systems able to alert human operators when something unusual occurs in the environment under surveillance. Such systems are used to perform different tasks, such as object detection, vehicle tracking on highways for traffic control purposes, analysis of human behavior, or people counting in public environments. All the systems must detect unusual situations. The definition of a performance evaluation procedure can be helpful, in that it would allow one to select the best segmentation parameter values and to provide useful guidelines for the installation of a particular system. However, performance evaluation of complex systems remains an open problem.

Video surveillance is a typical case where knowledge of the specific application can be used to tune the parameters of the evaluation metric: undetected objects or over segmentation will have a bigger impact on the overall annoyance than changes in the shape of the correctly detected objects.

In order to evaluate different segmentation algorithms in the context of a video surveillance applications, the segmentation results (see Sec. 6.2.1) and the reference segmentation have been used to produce test video sequences where the object boundaries detected by the segmentation algorithm have been underlined on the original video sequence by a colored contour as depicted in Fig. 6.12.

Video manipulation

The goal of video manipulation is to put together video objects from different sources in order to create new video content. In particular, in the *augmented reality* application [100] considered here, video segmentation serves to extract real objects that are then inserted in a virtual background. One of the possible application is to create narrative spaces and interactive games and stories [3, 87]. In order to evaluate different segmentation results in augmented reality scenario, we created a virtual background for each original sequence: we extracted the contour of the background image to recall a virtual background in black and white as in comics scenarios. For the test sequence ‘Group’ we applied a virtual background created in the context of the European Project art.live [3] processed the same way to extract only the contours. Figure 6.13 shows a sample frame for each video sequence.



(Hall monitor) (Group) (Highway)

Figure 6.13: Sample frames for video augmented reality segmentation application 'Group', 'Hall monitor' and 'Highway'.

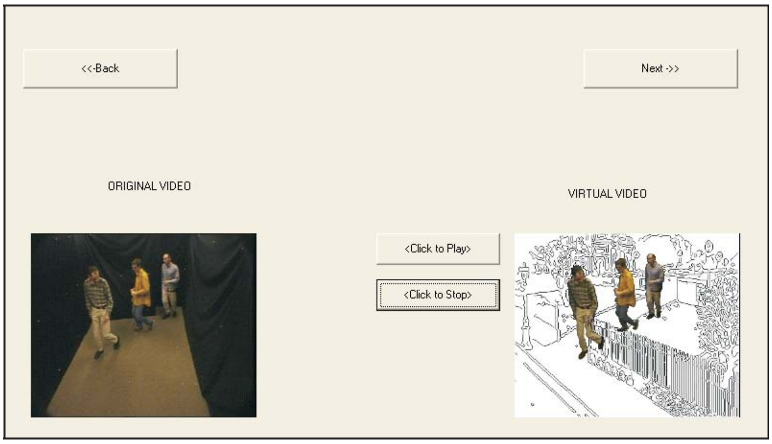


Figure 6.14: Graphical interface example for application dependent segmentation evaluation: tutorial stage.

6.3.2 Subjective Experimental Methodology

The experimental methodology is composed of a five-step procedure as described in Sec. 4.4: oral instructions, training, practice trials, experimental trials and interview. After a general introduction on segmentation, the typical artifacts are shown and the original video with the correspondent segmented video are shown as in Fig. 4.5 (a). After this introduction the three different applications are explained and the corresponding segmentations are shown in the training stage as depicted in Fig. 6.14.

Each application has specific protocols that are reported in Appendix B. The test group was composed of 35 subjects aged between 23 and 41 (with 8 females) who performed the annoyance task. During the experimental trials, subjects were asked to evaluate one application at a time for the tested segmentation algorithms (see Fig. 6.15 for surveillance application). The total number of test sequences for this part of the experiment was 82 which included 3 original video sequences ('Hall monitor', 'Highway','Group') × 8 segmentation algorithms (the reference and the above described segmentation algorithms) × 3 applications (compression, augmented reality, video surveillance).

The display layout and viewing distance were in concordance with the subjective viewing for CIF format [64] images (see Table 6.4).

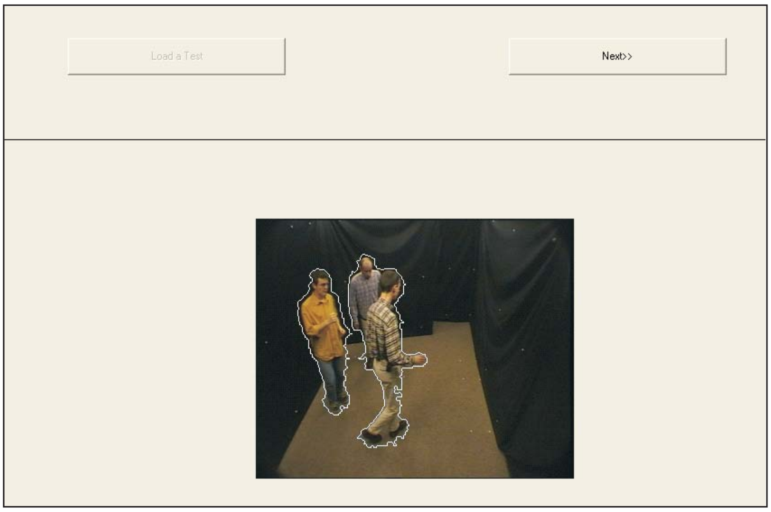


Figure 6.15: Graphical interface examples for application dependent segmentation evaluation: experiment trial.

Table 6.4: Viewing conditions during subjective test.

Variable	Values
Monitor type	WXGA Color Shine LCD
Monitor resolution	1280 × 800
Viewing distance	60 cm
Monitor size	15.4”
Room illumination	dark room

6.3.3 Application Dependent Objective Metric

In this section we investigated two different issues. First, we find the best parameters for our metric depending on the segmentation application. To find the a, b, c, d coefficients for the proposed metric, we performed a nonlinear least-squares data fitting using the mean annoyance values (MAV) obtained from the subjective experiment. Second, we analyze our metric performance compared to those of other state of the art metrics according to the tested application.

Figures 6.16 (a), (c), (e) show the results of the proposed **PST** metric versus the subjective annoyance values MAV in relation with the application.

The best performance, when comparing the three applications is obtained for the augmented reality application. This can be explained by the fact that this application is the most similar to the general purpose segmentation evaluation developed to design the metric. In fact, cutting and pasting objects into a virtual background is more or less what we have done with a plain uniform background for the general framework. Therefore, the perception of objects in these two different scenarios is similar and since our metric has been built on the basis of experiments in the general case, it seems to work better for the augmented reality applications.

An interesting aspect of this research is that the parameters of our metric can be easily adjusted according to the different kinds of applications. Thus, on the basis of the subjective data we defined *a posteriori* the weights for each artifacts to achieve a better fit with the subjective values. We obtained good results that are depicted in Fig. 6.16 (b), (d), (f).

In the compression scenario, the weights obtained for added regions and background were really small compared to those for inside and border holes. In fact, in this application we have preserved the quality of the objects and compressed the background. Thus, the parts of the object that have been erroneously segmented as part of the background will be compressed and will annoy the subjects more than having segmentation artifacts like added region or background that have not be compressed.

In the surveillance application the biggest weights are given to added regions and inside holes. This can be explained by the fact that human viewers in the surveillance scenario pay attention to mis-detected or over-detected objects that could lead to dangerous situations of false alarms (in case of erroneus detection of background parts as moving objects) and missed alarms (in case of mis-detection of moving objects).

Finally, in the augmented reality application the most important weights were for added background, and inside and border holes. In fact, every artifact that changes the shape or allows to see the virtual background beneath the real objects causes a lot of annoyance in the subjects that are focusing their attention on the virtual story or the interactive game.

Next, we need to find the correlation with the subjective data for the other state of the art metrics. We plotted the objective results versus the subjective annoyance values in Figs. 6.17 (a), (b), (c) for **MPEGqm**; (d), (f), (g) for **wqm**; and (h), (i), (l) for **mqm** in all the three applications. The correlation coefficients reported in Fig. 6.17 are lower than those of our **PST** metric of Fig. 6.16. However, MPEG metric, **MPEGqm**, outperforms both Villegas's (**wqm**) and Nascimento's (**mqm**) metrics in surveillance and augmented scenarios. No state of the art objective metric performs well in the case of compression, and the surveillance application is the only case where Nascimento's metric (**mqm**) performs reasonably well. The correlation coefficients are given in Tab.6.5.

6.4 Conclusions

In this chapter a study on real artifacts produced by typical video object segmentation algorithms has been carried out to test the proposed objective metric for segmentation quality assessment.

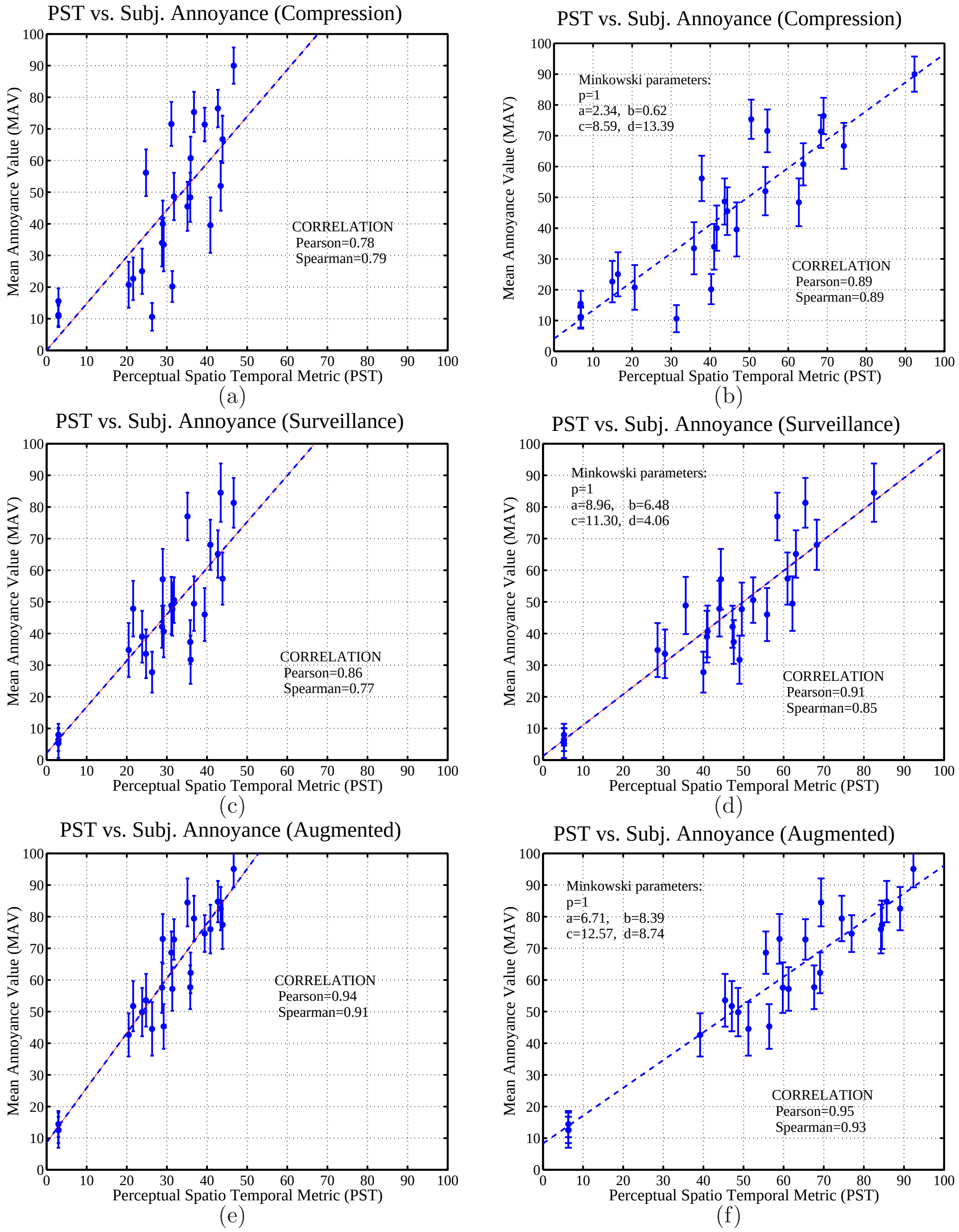


Figure 6.16: Objective metric **PST** vs. Subjective Scores (MAV) and correlation coefficients for different segmentation applications: with no optimized (a), (c), (e) and optimized (b), (d), (f) Minkowski parameters.

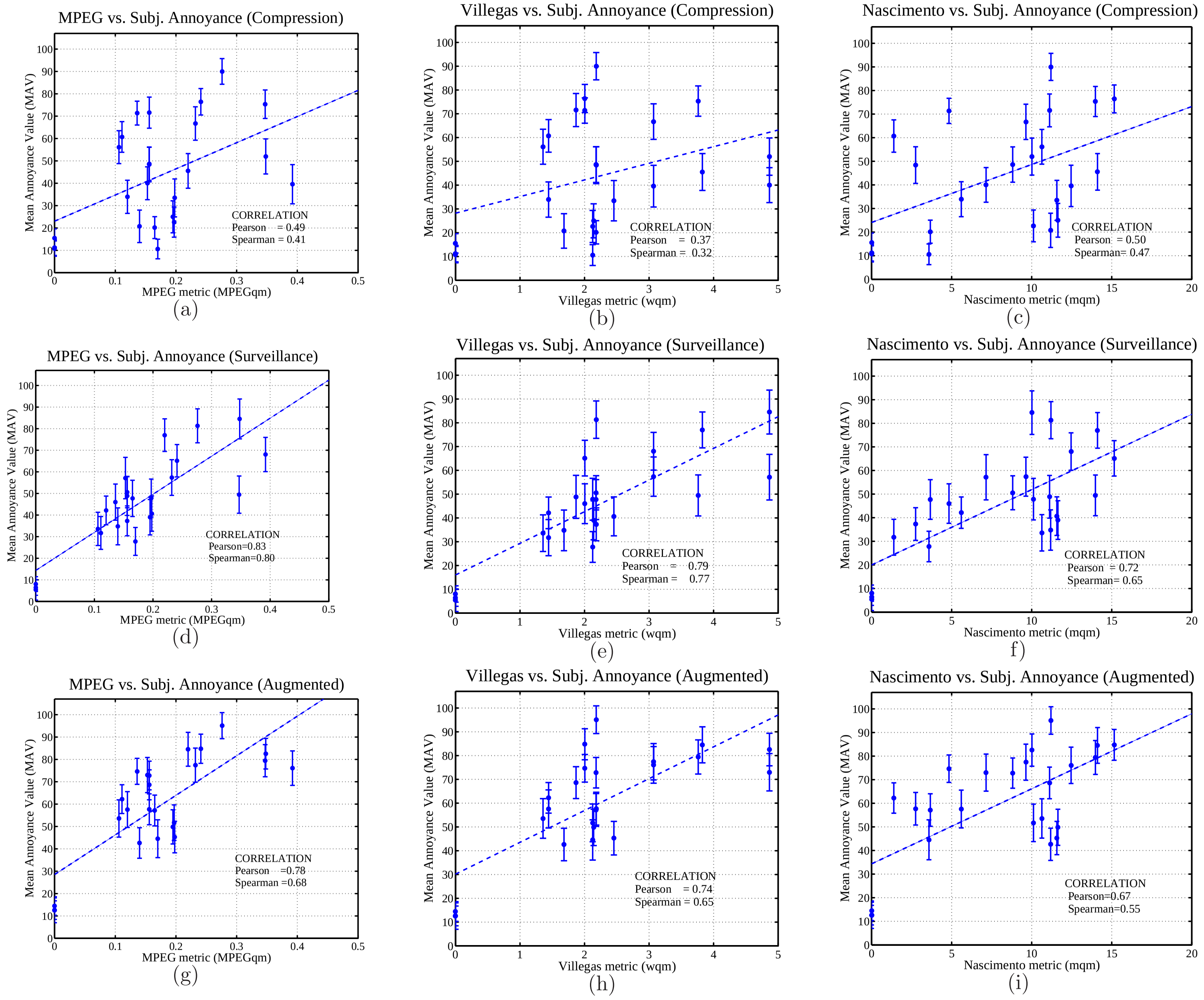


Figure 6.17: State of the art objective metrics vs. Subjective Scores (MAV) and correlation coefficients for different segmentation applications: (a), (b), (c) compression, (d), (f), (g) surveillance and (h), (i), (l) augmented reality.

Objective Metric	Augmented		Surveillance		Compression	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
PST (general)	0.94	0.91	0.86	0.77	0.78	0.79
PST (app. dep.)	0.95	0.93	0.91	0.85	0.89	0.89
MPEGqm	0.78	0.68	0.83	0.80	0.49	0.41
wqm	0.74	0.65	0.79	0.77	0.37	0.32
mqm	0.67	0.55	0.72	0.65	0.50	0.47

Table 6.5: Correlation coefficients for the objective metrics vs. subjective annoyance.

Seven segmentation algorithms were chosen as typical and tested both objectively and subjectively. First, a classification of the real artifacts introduced by typical segmentation algorithms has been provided according to subjective perception. Second, an analysis of the performance of the objective metrics, including ours, has been performed.

To the best of our knowledge, a comparison among different objective metrics for video segmentation quality assessment has received little attention by the image processing community so far, as well as the study of their performances on real segmentation algorithms. Moreover, our study considered both *general* and *application* dependent segmentation scenarios, while in the literature, applications are neglected in the evaluation.

Real artifacts have been subjectively tested in both general and application dependent scenarios. The segmentation applications have been chosen in the field of video coding, video analysis and video manipulation. A subjective test has been proposed and designed to test each of them.

In both scenarios, the proposed metric outperformed the state of the art metrics.

In addition, it has been discussed how appropriate parameters can be chosen for our proposed metric for each of the considered applications. In fact, we found that when developing a segmentation evaluation for a specific application, the characteristics of the application provide valuable information for the selection of appropriate segmentation artifact weights. For the considered applications, especially compression, augmented reality and video surveillance, different *perceptual* weights have been found on the basis of subjective experiments. In the compression scenario, the perceptual weights obtained for inside and border holes are larger than those for added regions and background. This is due to the fact that parts of objects are erroneously considered as belonging to the background (holes) and thus compressed. In the surveillance application, added region weights are larger than those of background since thay can be confused as objects and thus causing false-alarms. Moreover, the weights of inside hole are larger than those of border holes since they could cause dangerous missed alarm situations. In the augmented reality scenario, shape artifacts (border holes and added background) have the largest weights since they compromise the overall impression of interactive story in which the characters are cut and pasted in the virtual background.

Part III

3D Object Watermarking Quality Assessment

Subjective and Objective 3D Watermarking Evaluation

7

7.1 Introduction

Nowadays, a new kind of multimedia data has reached the same level of diffusion as audio, still images and video: *geometric data*. Architecture, Design, Mechanical Engineering, Entertainment and Cultural Heritage are some of the main important areas in which three-dimension (3D) models are widely applied. Often, the creation of a 3D model, in particular in certain fields, like in Cultural Heritage, requires a lot of resources in terms of time and costs. The Digital Michelangelo Project [126] of the University of Stanford, for example, has required 30 nights of scanning and 22 people only to acquire a high-quality 3D model of the David of Michelangelo (Figure 7.1) and about 1 year to create a complete archive of the principal statues and architectures of the great artist. These two factors strongly motivate the demand of robust watermarking methods for 3D objects. Relatively recently, researchers in digital watermarking have moved their attention to this problem, and some algorithms to embed information within geometric data have been developed. However, no benchmark to test digital watermarking systems of geometric data has been reported.

In this part of the thesis, we have concentrated our efforts in the study of the visual distortions introduced by typical watermarking systems for 3D models and the development of objective metrics apt to perceptually assess the quality of watermarked 3D models. In fact, one of the fundamental requirements of a watermarking system is the *imperceptibility* of the inserted watermark (see Sec. 3.2).

In this chapter, we propose two perceptual metrics for the quality assessment of watermarked 3D objects. The reasons for proposing perceptual metrics are the evaluation and comparison of perceptual artifacts introduced by 3D watermarking algorithms. The final aim of evaluation is to minimize extraneous details introduced by watermarking by modulating the watermark insertion in order to obtain little or no perceptual artifacts. The second is to use such metrics for comparing the performance of different 3D watermarking algorithms on the basis of the artifacts perceived on the 3D model.

A possible approach could be to simply apply well-known image-based perceptual metrics to the final rendered images of the 3D model. The main problem of this approach is that the perceived

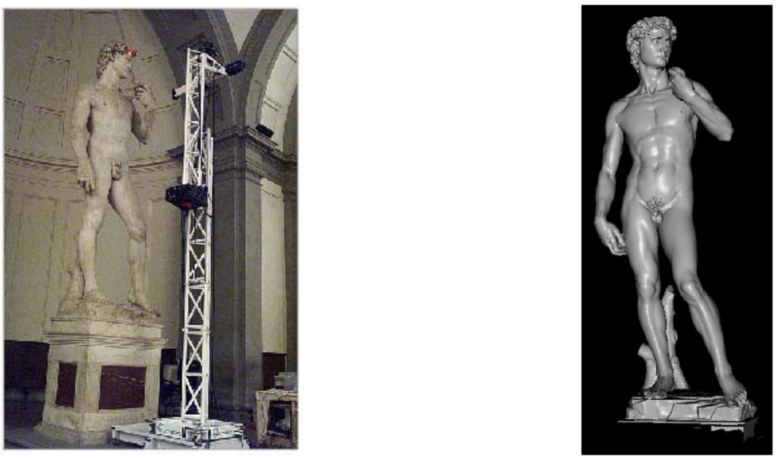


Figure 7.1: Digital Michelangelo Project. (Left) David's acquisition. (Right) The reconstructed three-dimensional model.

degradation of still images may not be adequate to evaluate the perceived degradation of the equivalent 3D model. Hence, the approach we chose is to evaluate the human perception of geometric defects of watermarked models and then to build an ad-hoc perceptual metric that works directly on the model's surface. In such a case, subjective experiments dealing directly with the 3D models are needed. In particular, we propose two subjective experiments with different purposes. The first experiment (Experiment I), is carried out to investigate the perception of artifacts caused by a watermarking algorithm on 3D models and to find suitable metrics to measure artifacts' perceptual severity. On the basis of the subjective data collected with this experiment two metrics based on roughness estimation of the model's surface have been devised to perceptually measure the amount of visual distortions introduced by the watermarking algorithm over the surface of the model. Then, a second experiment (Experiment II) is conducted in order to validate the proposed metrics with other watermarking algorithms.

This chapter is organized as follows. Previous works on perceptual image watermark insertion, mesh simplification and perceptually-guided rendering are reviewed in Section 7.2. In Section 7.3 we describe the artifacts introduced by common 3D watermarking algorithms. Our experimental methodology to carry out subjective experiments on 3D model quality evaluation is described in Section 7.4. Subjective data analysis is performed in Section 7.5. Section 7.6 describes the proposed metric. Finally, results are presented and discussed in Section 7.7.

7.2 Related Work

The knowledge of the human visual system (HVS) has been widely applied in *perceptual image watermarking* to obtain high quality watermarked images, i.e. watermarked images indistinguishable from the original ones. Our investigation concerns the extension of this idea to 3D watermarking. The goal is to develop a perceptual metric to estimate the perception of visual artifacts introduced by watermarking. The evaluation of the visual impairment introduced by a watermarking algorithm can be used to adjust the watermarking parameters in order to obtain a watermarked version that looks like the original one. Perceptual metrics are not limited to perceptual watermarking, but they have also been used in two other fields of Computer Graphics: *mesh simplification* and *perceptually-guided rendering*. The three issues related to our investigations, concerning perceptual image watermarking, mesh simplification and perceptually-guided rendering, will be discussed in the following.

7.2.1 HVS and Perceptual Image Watermarking

It is widely known among researchers working in Digital Watermarking that HVS characteristics have to be carefully considered in order to minimize the visual degradation introduced by the watermarking process while maximizing the robustness [26, 148, 178]. Considering a noisy image, some aspects of human visual perception are as follows: *1)* distortions in the uniform regions of an image are more visible than those in textured regions, *2)* noise is more easily perceived around edges and *3)* the human eye is less sensitive to distortions in very dark and very bright regions. These basic mechanisms of the human visual perception can be mathematically modeled considering two main concepts: the *Contrast Sensitivity Function* (CSF) and the *contrast masking* model. CSF is a measure of the responsiveness to contrast for different spatial frequencies. Typically, CSF models the capability of the human eye to perceive sinusoidal patterns on a uniform background. The contrast perception varies with the frequency of the sinusoidal pattern, the orientation of the pattern, the observer's viewing angle and the luminance of the background where the stimulus is presented. Many analytical expressions of CFS can be found in literature, one of the most used is the Barten's model [6].

The *masking effect* concerns the visibility reduction of one image component due to the presence of other components. In other words, while CSF considers the visual perception of a sinusoidal pattern on a uniform background the *visual masking model* considers the perception of a sinusoidal pattern over spatially changing background. The non-uniform background may be modeled with another sinusoidal pattern with different properties. Some models of visual masking have been developed by Watson [166, 167] and by Legge and Foley [80].

Many methods have been proposed so far to exploit the models of the HVS to improve the effectiveness of existing watermarking systems [124, 178]. We can divide the approaches proposed so far into theoretical [75, 124, 179] and heuristic [7, 133]. Even if theoretically grounded approach to the problem would be clearly preferable, heuristic algorithms sometimes provide better results due to some problems with the HVS models currently in use [7, 31].

7.2.2 Mesh Simplification

Mesh simplification is concerned with the reduction of the number of vertices and triangles of a polygonal mesh while preserving its visual appearance. In general, the simplification process is driven by a similarity metric that measures the impact of the changes of the model after each simplification step. So, one of the most important consideration of a simplification method is the *error metric* it uses. Two kinds of metrics are considered for simplification: geometric metrics and (perceptual) image-based metrics.

Geometry-based metrics

Metrics for simplification are commonly used for two distinct purposes; evaluating the quality of the final model and determining where and how to simplify the model. The most used global geometry-based metrics for off-line quality evaluation of 3D models are based on the Hausdorff distance.

The *Hausdorff distance* is one of the most well-known metrics for making geometric comparisons between two point sets. Assuming that the shortest distance between a point x and a set of points Y (e.g. the vertices of the 3D model) is the minimum Euclidean distance:

$$d(x, Y) = \min_{y \in Y} d(x, y), \tag{7.1}$$

the asymmetric Hausdorff distance between two point sets is defined as:

$$\vec{d}_\infty(X,Y) = \max_{x \in X} \min_{y \in Y} d(x,y). \tag{7.2}$$

Since $\vec{d}_\infty(.)$ is not symmetric , i.e. $\vec{d}_\infty(X,Y) \neq \vec{d}_\infty(Y,X)$, this distance is not a metric in mathematical sense. To obtain symmetry it can be redefined as:

$$d_\infty(X,Y) = \max\left(\vec{d}_\infty(X,Y), \vec{d}_\infty(Y,X)\right). \tag{7.3}$$

The quantity $d_\infty(X,Y)$ in Eq.(7.3) is usually referred to as the *maximum geometric error*. This metric is not able to catch well geometric similarity since a single point of the set X, or Y, can determine the Hausdorff error. One possible alternative based on the average deviation that best measures geometric similarity is given by:

$$\vec{d}_1(X,Y) = \frac{1}{\mathcal{A}_X} \int_{x \in X} d(x,Y)dX \tag{7.4}$$

where $\mathcal{A}_X$ is the area of the surface X. This metric is also asymmetric. The symmetric version of this metric assumes the following form:

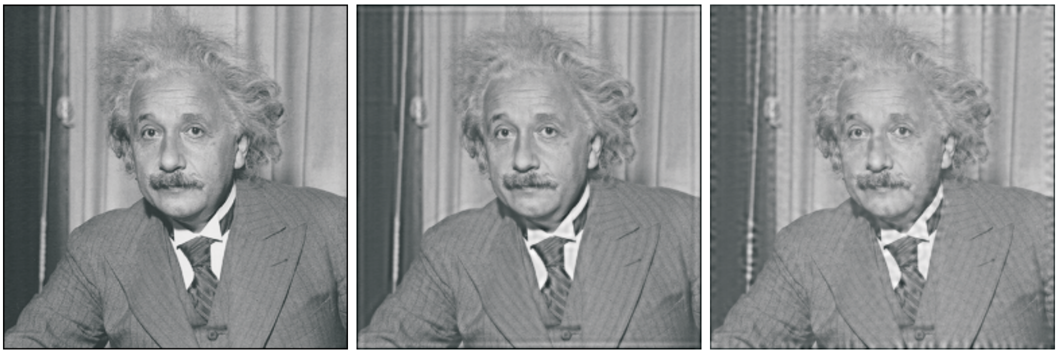
$$d_1(X,Y) = \frac{\mathcal{A}_X}{\mathcal{A}_X + \mathcal{A}_Y}\vec{d}_1(X,Y) + \frac{\mathcal{A}_Y}{\mathcal{A}_X + \mathcal{A}_Y}\vec{d}_1(Y,X) \tag{7.5}$$

and it is usually referred to as the *mean geometric error*. Two tools for geometric meshes comparison based on the maximum Eq.(7.3) and on the mean geometric error Eq.(7.5) are the Metro [113] and the Mesh [4] tool. Several researchers have proposed other geometry-based metrics to evaluate 3D model quality. Most of them are variations of the $d_\infty(.)$ and $d_1(.)$ metrics. In Section 7.7, we will analyze the performance of these two geometric metrics in the case of 3D watermarking quality evaluation.

Image-based metrics

Image metrics are adopted in several graphic applications. In fact, since most computer graphics algorithms produce images, it makes sense to evaluate their results using image differences instead of metrics based on geometry. Many simple image metrics such as the Root Mean Square (RMS) and the Peak Signal Noise Ratio (PSNR) have been widely used in the past, but such metrics are not able to measure the differences between two images as perceived by a human observer [147]. For example, Fig. 7.2 shows that the values of RMS do not correlate with the perception of image distortions. For this reason, nowadays, most applications move to perceptual-based image metrics. Two of the most perceptually accurate metrics for comparing images are the *Visual Difference Predictor* by Daly [29] and the *Sarnoff Model* developed by Lubin [85]. Both of these metrics include models of different stages of the human visual system, such as opponent colors, orientation decomposition, contrast sensitivity and visual masking.

Concerning *perceptually-based mesh simplification*, Lindstrom and Turk [82] proposed an image-driven approach for guiding the simplification process: the model to be simplified is rendered by considering several viewpoints and an image quality metric, based on a simplified version of the Sarnoff Model [85] is used to evaluate the perceptual impact of the simplification operation. More recently, Luebke *et al.* [173] developed a view-dependent simplification algorithm based on a simple model of CSF that takes into account texture and lighting effects. This method provides also an accurate modeling of the scale of visual changes by using parametric texture deviation to bound the size (represented as spatial frequencies) of features altered by the simplification. Other studies



Original image Minimum Perceptual Distortion (RMS = 9.0) Maximum Perceptual Distortion (RMS = 8.5)

Figure 7.2: Image distortions and RMS metric (from Teo and Heeger [147]).

related to perceptual issues in mesh simplification have been conducted by Rogowitz and Rushmeier [130] and by Yixin Pan *et al.* [116]. In particular, Rogowitz and Rushmeier analyzed the quality of simplified models perceived by human observers in different lighting conditions by showing to the observers still images and animations of the simplified objects. From the experiments they draw several interesting conclusions. The most important one is that the perceived degradation of the still images is not adequate to evaluate the perceived degradation of the equivalent animated objects. This result suggests that an experimental methodology to evaluate the perceived alterations of 3D objects should rely on the interaction with the model.

7.2.3 Perceptually-Guided Rendering

The aim of perceptually-guided rendering is to accelerate photo-realistic rendering algorithms in order to avoid computations for which the final result will be imperceptible.

One of the first work of this type has been done by Reddy [128], who analyzed the frequency content of 3D objects in several pre-rendered images and used these results to select the "best" version of the objects from a pool of models representing the same shape with different levels of details in order to speed-up the visualization of a virtual environment. If the high-resolution version of the model differs only at frequencies beyond the modeled visual acuity or the greatest perceptible spatial frequency, the system selects a low-resolution version of the model.

Other remarkable works in this field include the work of Bolin and Meyer [14] who used a simplified Sarnoff Visual Discrimination Model [85] to speed-up the rendering techniques based on sampling (e.g. Monte Carlo Ray Tracing), Myszkowski *et al.* [103] who incorporated the spatio-temporal sensitivity in a variant of Daly Visual Difference Predictor [29] to create a perceptually based animation quality metric (AQM) to accelerate the generation of animation sequences and Ramasubramanian *et al.* [127] who applied perceptual models to improve global illumination techniques used for realistic image synthesis.

Another excellent work related to study of human visual perception in rendering is the one by Ferwerda and Pattanaik [50]. In this work a sophisticated perceptual metric for the evaluation of how much a visual pattern, i.e. a texture, hides geometry artifacts is proposed. The visual masking effect caused by texturing is taken into account by analyzing the final rendered images.

7.2.4 Proposed Approach

Our goal is to develop a perceptual metric that measures the human perception of geometric artifacts introduced over a 3D surface by watermarking. Two approaches to develop a perceptual metric for 3D watermarking are possible. The first one follows the (perceptual) image-based approach for simplification seen before [82, 173]. Instead of driving the simplification process, the perceptual image metric can be used to evaluate the visual effects of the watermark insertion by computing the perceptual differences between several images rendered from the original and the watermarked model. This approach presents two advantages. First, since it is rendering-dependent, complex lighting and texturing effects can be taken into account in a natural way. The second advantage is that all possible kinds of visual artifacts can be evaluated with the same approach. The main disadvantage is that the rendering conditions must be known in advance. The other possible approach is to evaluate then how the human visual system perceives geometric distortions on the model surface and build an ad-hoc perceptual metric for geometric artifacts. Moreover, this approach is more interesting from a research viewpoint, since no similar studies have been conducted so far. The potential field of applications is not limited to 3D watermarking, but other Computer Graphics applications can also benefit from them. For these reasons we decided to follow the second approach, i.e. to work directly on the geometry of the 3D model.

7.3 3D Watermarking Algorithms and Artifacts

Digital watermarking algorithms can be classified according to the domain they work: hybrid domain and transformed domain. Here, for each class of algorithms, we describe the geometric artifacts that they introduce in the watermarked model.

First, we consider the algorithms working in the asset domain. The algorithms based on topological embedding [19, 107] produce small geometric distortions that can be described by the addition of a small amount of *noise* to the position of the mesh vertices. When only the connectivity of the mesh is used to embed the watermark, such as in the TSPS and in the MDP algorithms [107], the amount of introduced distortions is imperceptible, since topology changes usually do not produce noticeable visual effects. Concerning geometric features embedding we have to distinguish between those algorithms that embed the watermark by using vertices position and those algorithms that are based on shape-related features, such as vertex normals. Changes in the vertices position produce the same effect of topology-driven embedding, i.e. a “noisy” watermarked surface, but, in this case the amount of distortion may be considerably high, due to the vertices displacements needed to embed the watermark. For example the Vertex Flood Algorithm (VFA) [8] may introduce moderate-to-strong distortions depending on the embedding parameters. In the same manner, the method of Harte and Bors [59] may produce perceptible distortions depending on how many vertices are watermarked and on the dimension of the bounding volume used. Shape-related algorithms, like the Normal Bin Encoding (NBE) [9] and the method proposed by Wagner [159], instead, introduce artifacts that look very different from the noisy effect of the other techniques. This kind of surface alterations produces soft changes in the shape of the model thus resulting in artifacts difficult to perceive.

The algorithms that work in the hybrid domain are able to spread the distortions smoothly over the whole surface of the model by introducing the watermark in the low resolution of the model. Typically, this permits the reduction of the previously described ”noise” effect. The amount of distortion produced by the Uccheddu *et al.* technique [150], that works in the hybrid domain, heavily depends on the level of resolution used to embed the watermark. In particular, for a fixed watermark strength, the higher the level of resolution used, the stronger the amount of visual impairment of the watermarked model. In the same way the algorithm by Kanai *et al.* [69], which is based on wavelet decomposition, may introduce geometric artifacts as several levels of resolution are used to embed the watermark. The authors propose a geometric tolerance threshold to limit the introduction of these visual artifacts.

Concerning the transformed domain, mesh spectral methods [20, 108, 109] cause a vertices perturbation due to the modifications of the mesh spectral coefficients, thus resulting in a moderate "noisy" watermarked surface. Ohbuchi [109] suggests to reduce this effect by watermarking those mesh spectral coefficients that are related to the low frequencies content of the model.

Summarizing, we observe that, in general, 3D watermarking algorithms produce "noisy" surfaces. The characteristics of the noise depend on the specific algorithms; noise can have different granularity and size, and may be uniform or not over the model surface. The watermarking techniques that do not introduce perceptible artifacts are typically those techniques that have relaxed robustness requirements. In our subjective experiments that will be described in the next Section, we have implemented four different watermarking algorithms: the Vertex Flood Algorithm (VFA) [8], the Normal Bin Encoding (NBE) [9], the method by Kanai *et al.* [69], and Uccheddu *et al.* algorithm [150]. The algorithm by Kanai et al. and the Uccheddu *et al.* will be indicated in the following using the initials of the authors, i.e. KDK and UCB respectively. Figure 7.3 shows the artifacts introduced by these watermarking algorithms.

7.4 Experimental Method

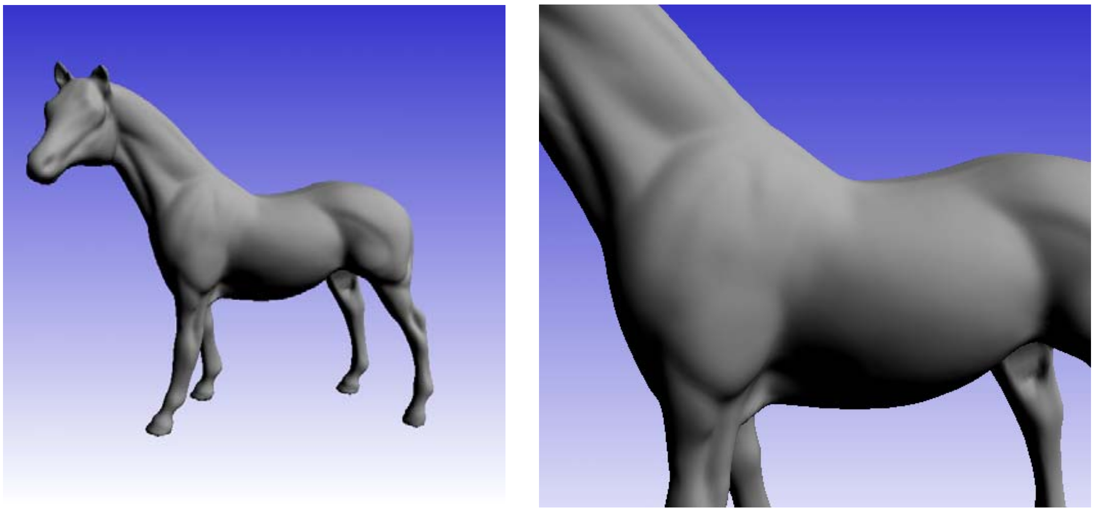
A set of standards and grading techniques to evaluate the quality of video and multimedia content have been defined in ITU-R [66] and ITU-T [65]. However, there are no prescribed standards for the evaluation of 3D objects with impairments. In this chapter, we propose a *method for subjective evaluation of 3D watermarked objects*. This experimental methodology attempts to make subjective evaluations in this field more reliable, comparable and standardized.

The starting point for the design of a subjective experiment for the quality evaluation of 3D objects is to define how to render the object under examination. By specifying appropriate rendering conditions, we aim at putting the human observer in favorable conditions to make a fair judgment on the three-dimensional object. The rendering conditions should not bias the human perception of the 3D model by choosing, for example, one view of the 3D object rather than another one.

7.4.1 Rendering Conditions

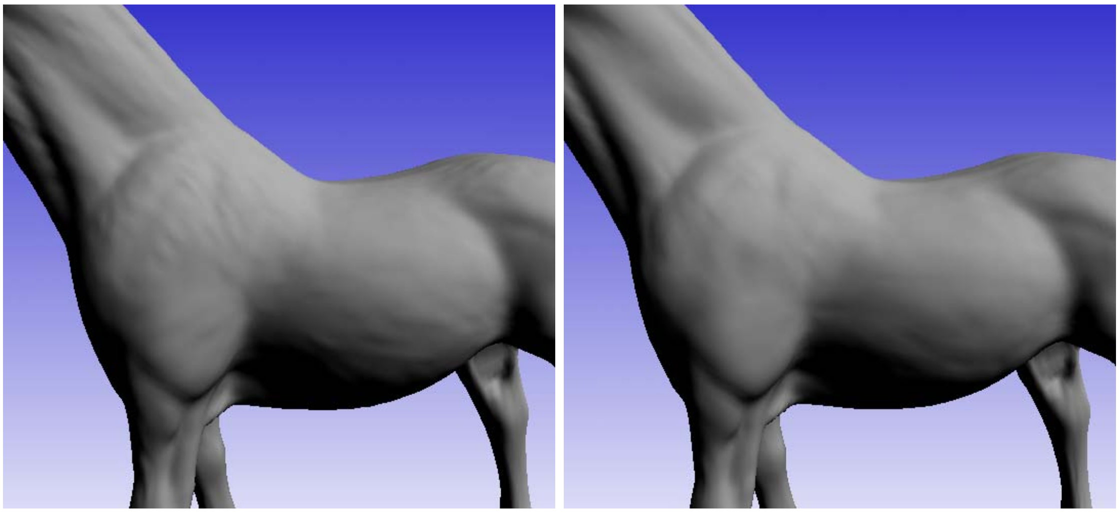
The rendering of a three-dimensional model is accomplished via a combination of techniques such as associating a material to each surface of the model, applying various kinds of light sources, choosing a lighting model, adding textures and so on. In our investigations we assumed that the rendering conditions have to be as simple as possible, because very few works have dealt with psychophysical tests of 3D object perceived quality as reported in Section 7.2 and no experimental data are available. Moreover, too many or complicated rendering effects would involve many and mutually linked aspects of spatial vision that have to be avoided to obtain more reliable results. Such results can be further extended by taking into account more aspects of visualization techniques, such as the role of photorealism in the perception of impairments. In fact, by keeping plain but effective rendering conditions, we do not influence or bias the human perception and, as such, the subjects' evaluation. The rendering conditions that we have chosen are described below.

- *Light sources.* Humans can see an object because photons are emitted from the surface of the object and reach the eyes of the viewer. These photons may come from light sources or from



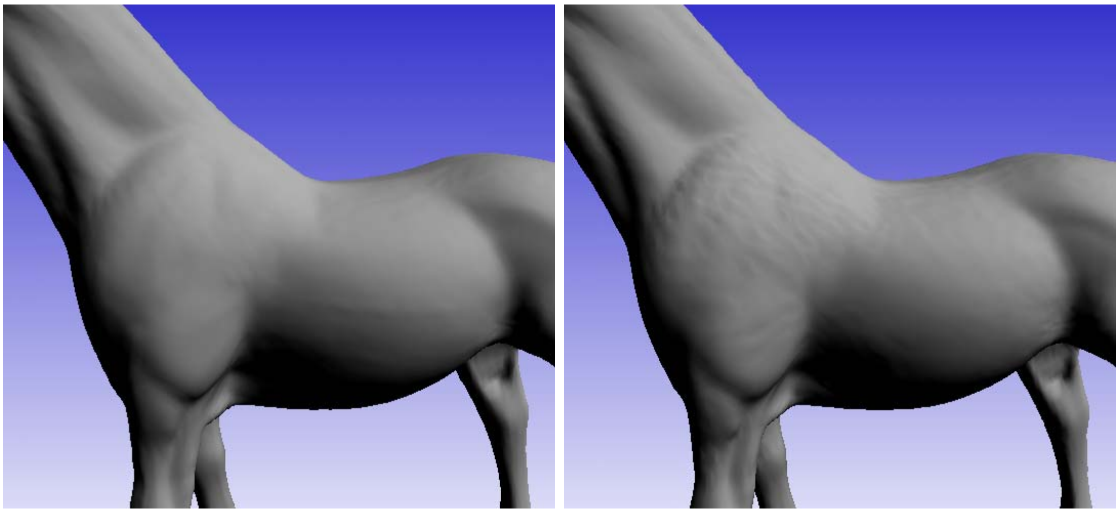
Original

Original - detail



UCB algorithm

KDK algorithm



Normal Bin Encoding (NBE)

Vertex Flood Algorithm (VFA)

Figure 7.3: Geometric defects introduced by 3D watermarking algorithms.

other objects. The three common types of light sources are the directional, the point, and the spot light sources. Point and spot light sources are also called *positional lights* because they are characterized by a location in space. Spotlights are not suitable for our purposes since this kind of light source could make some parts of the model better illuminated with respect to others. Multiple lights can cause effects that may confuse the human observer and provide contradictory results more complex to evaluate [136]. Additionally, the HVS tends to assume that the scene is illuminated by a single light source and that light illuminating the scene is coming from above. For all of these reasons, in our experiments, each model is illuminated with one white point light source located in the top corner of the Object Bounding Box (OBB) of the 3D object. Achromatic light is used in order to preserve the colors of the material.

- *Lighting and shading.* A good choice of the lighting model and of the shading method succeeds in effectively communicating to a human observer the 3D shape and the fine geometric details of a 3D object. The influence of the lighting model is very important since it may affect the perceived quality of the 3D model considerably. Ideal lighting and shading conditions are very hard to find and some methods have been proposed to optimize rendering conditions in order to improve the perceptual quality of the rendered model [136]. To narrow the scope, we use a simple local illumination lighting model where only the diffusive component of the reflected light is considered. In fact, the diffusive component is well-connected to physical reality since it is based on Lambert's Law (Eq. (C.2)) which states that for surfaces that are ideal diffusive (totally matte, without shininess), the reflected light is determined by the cosine between the surface normal and the light vector. For this reason, the diffuse component does not depend on the viewer's position making this model suitable to unbias the human perception of the 3D object under examination. The specular component of the reflected light is not considered even if it would improve the photorealism of the objects. In fact, while the diffuse component catches the behavior of matte surfaces, the specular component models the shininess of the surfaces. The highlights created by the specular component help the viewer to better perceive the surface's curvature. Moreover, other more complex photorealistic issues such as self-shadowing are not introduced in our model, as they would unnecessarily complicate the experimental method and introduce too many variables to evaluate during result analysis. Thus, the implemented lighting model is:

  $$I_r = I_{\mathtt{amb}} K_a + I_i K_d \min(0\,,\,\vec{N}\cdot\vec{L}) \tag{7.6}$$

  where $\vec{N}$ is the surface normal at the considered point, $\vec{L}$ is the incident light direction vector and the constant K_a and K_d depend on the material properties. About shading methods that deal with triangular meshes, we can choose among flat, Gouraud and Phong shadings (see Sec. C.1). Flat shading is not suitable for our purposes since it produces the well-known unnatural faceting effect. Since both Gouraud and Phong shadings produce almost the same visual effects if the model resolution, i.e. the number of triangles of the model, is high, we decided to use the Gouraud shading that is more common and less computationally expensive than the Phong method. Finally, we have decided to show the model on a non-uniform background since a uniform background highlights too much the countour edges of 3D objects.

- *Texturing.* We want to evaluate the perception of artifacts on the surface of the 3D objects, hence textures or other effects are avoided as they usually produce a masking effect on the perceived geometry [50]. In fact, image texture mapping, bump mapping, and other kind of texturing may hide the watermark artifacts. This is partially due to the *visual masking* perceptual effect, in which frequency content in certain channels suppresses the perceptibly of other frequencies in that channel [50]. We do not account for visual masking, leaving that as an important and interesting area for future researches.

- *Material properties.* The color of a surface is determined by the parameters of the light source that illuminate the surface, by the lighting model used and by the properties of the surface's material. We consider only gray, *stone-like* objects. This choice is made for different reasons: first, if all models are seen as "statues" the subjects perceive all the models in the same manner and naturally enough; and second, in this way we avoid the *memory color* phenomenon experimented in psychology studies [54]. This phenomenon regards the fact that an object's characteristic color influences the human perception of that object's color, e.g. shape such as heart and apple are characteristically red. A particular choice of a specific color for all the models could mis-lead the perceived quality of the object and introducing too many colors for different objects would have made the experimental method less general by introducing too many degrees of freedom.

- *Screen and Model Resolution.* The monitor resolution used in the experiments was 1280 × 600 and each model displayed a window of 600 × 600 pixels. The model occupies around 80% of the window and the resolution of the models used in the experiments ranged between 50.000 and 100.000 triangles. This screen resolution and the level of details of the models allow a good visualization of the model details, and hence of the model distortions. In particular the blurring effect of the Gouraud shading interpolation is negligible. Such blurring effect increases when the subject observes the model closely. Moreover, the complexity of the models used allows us to render them quickly (using a powerful graphics accelerator board) making fluid user-interaction possible. A minimum frame rate of 50 fps for each of the visualized models is guaranteed.

- *Interaction.* One essential feature of an interactive application is that objects are observed in motion. In our experimental method, we decided to allow the subjects to interact with the model by rotation and zoom operations. The user interacts with the model by using a mouse. Three-dimensional interaction is achieved by *ARCBALL rotation control* [139]. The motion of the 3D object is then interactively driven by the subject and not pre-registered like in other subjective experiments in the literature [116, 130]. This avoids the detection of less details in frames that pass quickly. It has to be mentioned that in previous works [130, 170], 2D images of 3D objects have been used for subjective experiments. The problem is that different static views of 3D objects can have significantly different perceived quality depending on the direction of illumination. Subjective experiments that were conducted to address this question suggest that judgments on still images do not provide a good predictor of 3D model quality [130]. Rogowitz's work confirmed that the use of still images produce results that are view-dependent and not well correlated with the real perception of the 3D object.

7.4.2 Experimental Procedure

Our test subjects were drawn from a pool of students from the Ecole Polytechnique Fédérale de Lausanne (EPFL). They were tested one at a time and not specifically tested for visual acuity or color blindness. The 3D models were displayed on a 17-inch LCD monitor, with participants sitting approximately 0.4 meter from the display. The experiment followed a five-stage procedure [102]. The stages were: (1) oral instructions, (2) training, (3) practice trials, (4) experimental trials, (5) interview.

Oral Instructions

Test subjects are told how to perform the experiment. Prior to each experiment, the instructions, also known as *experiment scripts*, are elaborated to help the experimenter defining the task. The script contains details of what the experimenter should do at each step of the experiment. More importantly, it contains oral instructions that are given to the subject to make sure he/she understands the task to be performed. An introductory explanation about what 3D models are and what watermarking is, is given. Different sections of the instructions actually apply to all the various stages of the experiment. However, the most important part of the instructions comes before the training stage. After the subject is properly seated and made comfortable, the main task is explained. The instructions for both experiments can be found in Appendix A.

Training

In each experiment the subject is asked to perform a task which consists of entering a judgment about an impairment detected in the 3D object. In order to complete this task, subjects need to have an idea of how the original 3D objects with no impairments look like. Therefore, a training session is included in the procedure which consists of displaying the four original models used to embed the watermark. In this phase, only the experimenter interacts with the graphical user interface, the subject is asked to look carefully to the models displayed on the screen. In the next phase, a set of 3D models with the typical distortions introduced by watermarking is shown.

Another set of 3D objects is required to set a value on the scale of judgements. The end of the scale is set by 3D objects with the strongest defect (in this case the perceptually strongest watermarking). A total of 12 and 16 models were shown as worst examples for the Experiments I and II, respectively. Therefore, the test subjects are instructed to pick the strongest stimulus in the training set and assign to that stimulus a number from the upper end of the scale. In our experiments, the subject is asked to assign 10 to the worst example (on a discrete scale ranging from 0, implying no distortions are perceived, to 10). However, due to visual masking and to the variety of the originals, it is not possible to anticipate which defects the subjects will consider worst or strongest. As a result, the subjects are asked to record values greater than the nominal upper value if 3D objects are observed to exceed the expectations established in the training phase. For example, if a test subject perceives a defect twice as bad as the worst in the training set, he/she is asked to give it a value of 20. Finally, in the last phase of the training stage, subjects are told how to use the graphical user interface to interact with the 3D models.

Practice Trials

In the practice trial stage, subjects are asked to make judgments for the first time. Because of the initial erratic responses of the subjects, ITU Recommendation [65] suggests to throw away the first five to ten trials of an experiment. In our case, instead of discarding the responses of the first trials, we included the practice trial phase. This also gives other benefits. It exposes the test subject to 3D models throughout the impairment stage. It gives the test subject a chance to try the data entry and above all the chance to get familiar with the graphical interface for the virtual interaction with the 3D object (rotation and zooming). The number of practice trials is six. The subject has to perform three tasks at most. The first one is to detect the distortion and he/she has to answer to the question *did you notice any distortion?*. In case of positive answer, the subject has to give a score to indicate *how much the distortions are evident.* The subject has 30 seconds at disposal to interact with the model and to make his/her judgement. Then, the subject has to input the score in a dialog box. Figure 7.4 shows the interface for the subjective tests. On the left a time bar advices

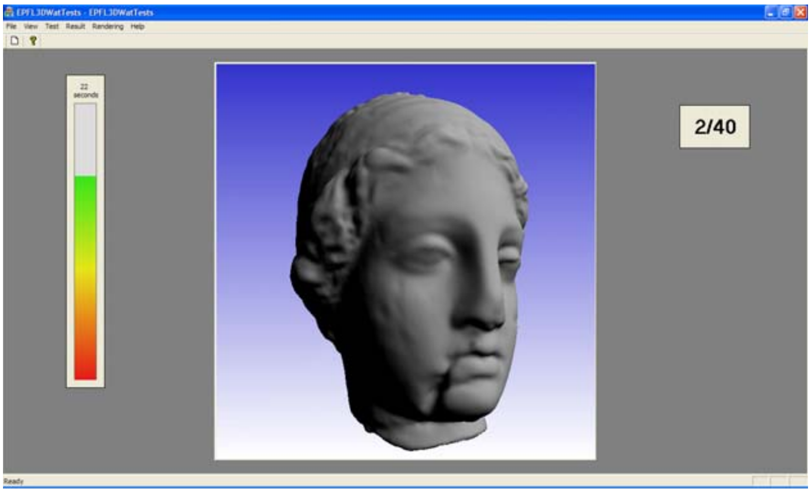

EPFL3DWatTests - EPFL3DWatTests
File View Test Result Rendering Help
22 seconds
2/40
Ready

Figure 7.4: The subject during the interaction with the model.

the user of the remaining interaction time. The box indicates the progression of the test by showing the number of the model under examination. The model is displayed in the center of the screen. Finally, the third question is *where* he/she noticed the distortion on the 3D model. To answer this question he/she has to indicate, by selection, the part of the model where the distortions are the most evident (Fig. 7.5).

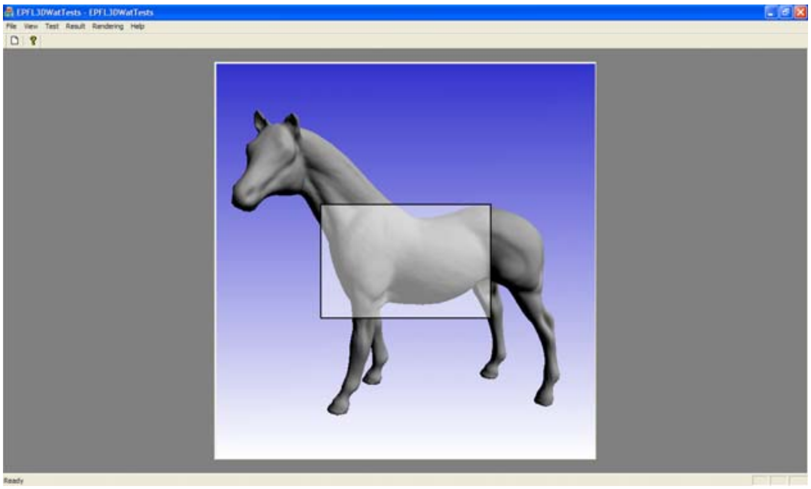

EPFL3DWatTests - EPFL3DWatTests
File View Test Result Rendering Help
Ready

Figure 7.5: The subject during the selection of the part where the distortions are more evident.

Experimental Trials

The subjective data is gathered during the experimental trials. In this stage, a complete set of 3D objects is presented in random order. To be more specific, for each experiment, several random-ordered lists of watermarked 3D test objects are shown. In this way the results of the test are made independent of the order in which the models are presented. All the test subjects see all the 3D objects. The number of test 3D objects is limited so that the whole experiment lasts no more than one hour. This limit translates to 40 models in Experiment I and 48 in Experiment II.

Table 7.1: Viewing conditions during subjective tests.

Variable	Values
Peak luminance	≤ 0.04
Maximum observation angle	10 degrees
Monitor resolution	1280×600
Interaction window resolution	600×600
Viewing Distance	$35 - 45$ cm
Monitor Size	17”

Interview

After the trials are completed, the test subjects are asked a few questions before they leave. The kinds of question depend on the experiment, mainly test subjects are asked for qualitative descriptions of the impairment. The questions asked in our case are:

1. Did you experience any problem correlated to a specific model in identifying the distortion?
2. How would you describe the distortions that you saw?
3. Have you general comments or remarks about the test?

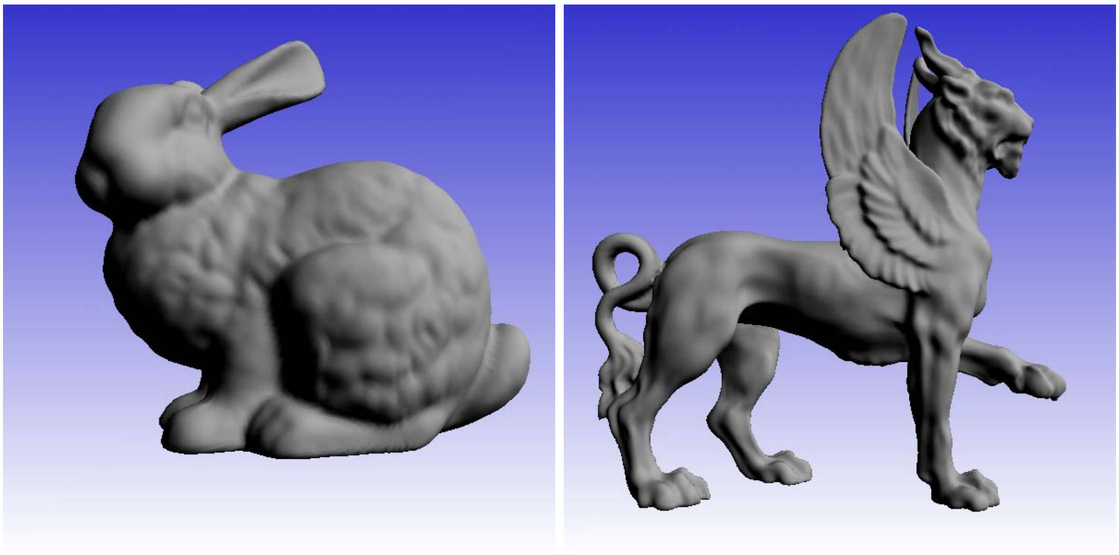
These questions gather interesting answers. They are for example useful for categorizing the distortion features seen in each experiment and for helping in the design of next experiments.

7.4.3 Experiment I

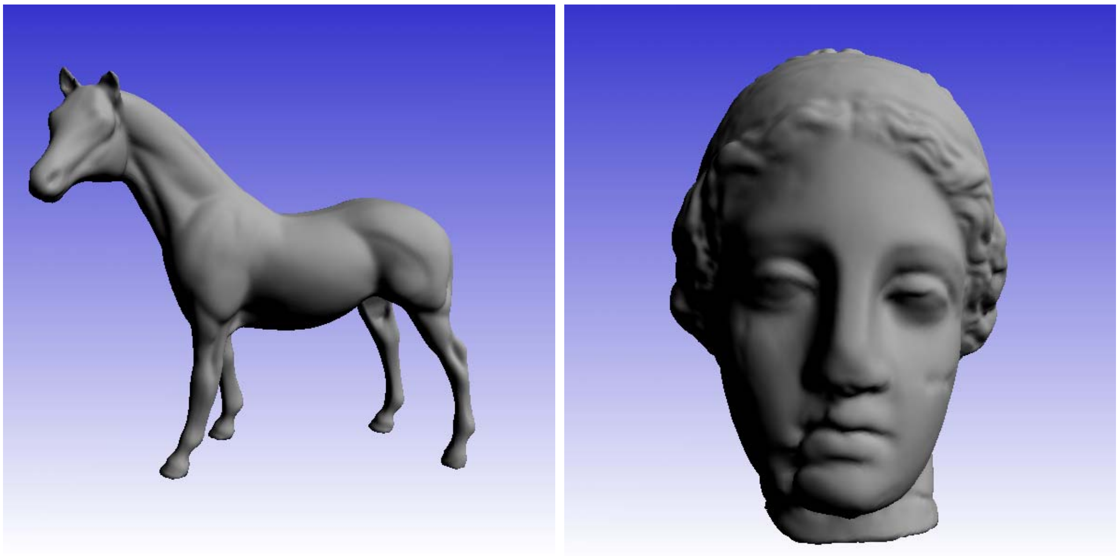
The goal of the first experiment was to make an initial study about the perception of artifacts caused by watermarking on 3D models and to find suitable metrics to measure the perceptual severity of such artifacts. The output of the experiment was a collection of subjective evaluations of a set of watermarked 3D objects. The watermarking artifacts, varying in strength and resolution, were generated using Uccheddu *et al.*’s watermarking algorithm [150]. This subjective data set allowed us to confirm some basic findings like that commonly used geometric-based metrics (already mentioned in Section 7.2) are not good measures for subjective quality evaluation of watermarked 3D objects. Additionally, we were interested in how the test subjects would describe the watermarking defects produced for this experiment. In particular, the appearance-related questions included in the interview provided us some directions to design a perceptually driven objective metric. The experiment have been performed by 11 subjects. The methodology for the experiment has been described in Section 7.4.2.

7.4.4 Generation of Stimuli

The test models for this experiment were generated by applying the previously described UCB algorithm to the *“Bunny”*, *“Feline”*, *“Horse”* and *“Venus”* models. Figure 7.6 shows these models rendered with the rendering conditions used in the experiments. These models are suitable for perceptual studies due to the wide range of characteristics presented by their surfaces. For example, the Bunny model surface is full of bumps, most parts of the Horse model are smooth, Feline model presents a wide range of characteristics such as parts with high curvature, or low curvature, moderate bumps,



Bunny Feline



Horse Venus

Figure 7.6: Rendering Conditions.

smoothed parts and several protrusions, and the Venus model has the same range of characteristics as the Feline but without consistent protrusions.

The watermark was uniformly distributed over all the surface of the 3D objects. The amount of distortions introduced by the watermarking varied according to two parameters: *i)* the resolution level l that hosts the watermark and *ii)* the coefficient γ determining the strength of the watermark. Table 7.2 shows the values of the watermarking parameters used for the experiment. Three amounts of watermarking strength (low, medium and high) and three levels of resolution (low, medium and high) were applied to each model. In addition to the watermarked models, the 4 original models were included in the complete model set. In fact, the original may present impairments unrelated to the watermarking ones deliberately inserted into the test models. To separate the effects of the deliberate and the pre-existing defects, the originals had to be inserted. A total of 40 (4 originals × 3 watermarking strength × 3 resolution level + 4 originals) test models were used in the experiment.

Table 7.2: Experiment I: Watermarking parameters values used for each model.

Level of Resolution (l)	**Corresponding Value**	**Watermarking Power (γ)**	**Corresponding Value**
Low	4	Low	0.0003
Medium	3	Medium	0.0015
High	2	High	0.003

7.4.5 Experiment II

In Experiment I, test subjects evaluated differently watermarked models ranging from severe down to weak visual impairments. Those different distortions' strengths were generated using a specific watermarking algorithm, i.e. the UCB algorithm. With this experiment we wanted to test by means of subjective validation the perceptually-based objective metrics for the quality assessment of 3D watermarking we obtained from the subjective data of the Experiment I. Therefore, we chose three different watermarking algorithms: NBE, VFA and KDK. Technically, the defects inserted are slightly different from the ones studied in the Experiment I (see Fig. 7.3). In fact, while the UCB algorithm produces a uniform kind of noise that can be described as an increase of the roughness of the watermarked surface, VFA produces a kind of noise that looks like marble streak, depending on the viewpoint. The artifacts of the KDK algorithm are the same of the UCB algorithm but due to the geometric tolerance introduced by Kanai to limit the visual impact of the watermark, the final visual effects of such distortions are not uniformly distributed over the model's surface. Concerning NBE, the visual aspect of its artifacts is very different from those of UCB, VFA and KDK and more difficult to perceive. The methodology for this experiment is practically the same as Experiment I. The only difference is that no location information was gathered since the metric developed on the basis of the data collected in Experiment I does not take into account the location information.

7.4.6 Generation of Stimuli

The test models for this experiment were generated using the same four original models of Experiment I. As just stated, the three watermarking algorithms used to generate the watermarked models are the VFA, the NBE and the KDK algorithms. Each watermarking algorithm is characterized by its own embedding parameters that are qualitatively and quantitatively different. For an exact description of each parameter we refer to the literature. The watermarking parameters of the VFA are the number of clusters used to embed the bits (one bit for each cluster) and the maximum allowable distance (D_{MAX}) from the starting application points. NBE is characterized by the feature type used, the number of bins (N_B), the search range (Δ_R) and the number of iterations (n_I) of the optimization process. KDK parameters are the selection threshold (δ_1), the geometric tolerance threshold (δ_2) and the least significant decimal digits used to embed the watermark (d_{w}); if $d_{\mathrm{w}} = 2$ then the second least significant decimal digit of the wavelet coefficients is modified to embed the watermark, $d_{\mathrm{w}} = 3$ indicates the third least significant decimal digit, and so on. The watermarking parameters for the three algorithms are reported in Tab. 7.3. In our test set, we tried to range from severely to weakly watermarked model as in Experiment I. The 11 level of impairment are also reported in Tab. 7.3. A total of 48 test models (4 models × 11 watermarking settings + 4 originals) were used in this experiment.

Table 7.3: Experiment II: Watermarking parameters values.

Algorithm	**Watermarking Parameters**	**Impairment**
KDK1	$\delta_1 = 0.001$, $\delta_2 = 0.005$, $d_{\mathrm{w}} = 2$	medium-strong
KDK2	$\delta_1 = 0.001$, $\delta_2 = 0.008$, $d_{\mathrm{w}} = 3$	medium
KDK3	$\delta_1 = 0.001$, $\delta_2 = 0.02$, $d_{\mathrm{w}} = 3$	medium-strong
NBE1a	Feature Type I, $N_B = 80$, $\Delta_R = 0.0015$, $n_I = 3$	medium
NBE1b	Feature Type I, $N_B = 80$, $\Delta_R = 0.0008$, $n_I = 3$	weak-medium
NBE2a	Feature Type II, $N_B = 80$, $\Delta_R = 0.0001$, $n_I = 1$	weak
NBE2b	Feature Type II, $N_B = 20$, $\Delta_R = 0.0004$, $n_I = 1$	medium
VFA1	600 clusters, $D_{\mathrm{MAX}} = 1.8$	strong
VFA2	960 clusters, $D_{\mathrm{MAX}} = 1.8$	medium
VFA3	1320 clusters, $D_{\mathrm{MAX}} = 1.8$	weak
VFA4	200 clusters, $D_{\mathrm{MAX}} = 0.6$	medium

7.5 Data Analysis

During the experiments, if a subject notices surface defects, he/she is supposed to enter a value proportional to the amount of distortions perceived on the model surface. In the following we refer to these values as *subjective scores*. The subjective scores have to be condensed by statistical techniques used in standard methods [66, 102] to yield results which summarize the performance of the system under test. The averaged score values, *Mean Opinion Score* (*MOS*), are considered as the amount of distortions that anyone can perceive on a particular watermarked 3D object. However, impairment is measured according to a certain scale, and such scale may vary from person to person. In this section, we report the methods used for matching the scales of our test subjects. Then, we describe the methods used to combine the subjective data and evaluate the precision of the estimates. The subjects are screened and outliers are discarded. Finally, the results are checked for error due to the methodology of the experiment.

7.5.1 Normalization

As a measurement device, a test subject may be susceptible to both systematic and random errors. The purpose of the normalizing procedure is to compensate for any systematic error. This procedure [102] is applied to the measurement gathered from each test subject prior to the combination of measurements across all subjects. The unscaled annoyance value, m_{ij}, obtained from subject i after viewing the test object j, can be represented by the following model:

$$m_{ij} = g_i a_j + b_i + n_{ij} \tag{7.7}$$

where let a_j be the true annoyance value for the test object j in the absence of any error, g_j is a gain factor, b_i is an offset, and n_{ij} is generally assumed to be a sample from a zero-mean, white Gaussian noise.

In this model, the gain and offset could vary from subject to subject. If the variations are large across the subjects or the number of subjects is small, normalization procedure can be used to reduce the gain and the offset variations among test subjects. In order to check if the offset and gain factors vary significantly from subject to subject, a two-way analysis of variance (ANOVA) approach was used [142]. A two-way ANOVA divides the total variations of a table of data into two parts: the variation that is attributed to the columns of the data table, and the variation that is attributed to the rows of the data table. Specifically, the F-test can be used to determine the likelihood that the means of the columns, or the means of the rows, are different. The experimental data, m_{ij} is arranged so that each row represents data for one test object and each column represents all the data for one test subject. The analysis assumes that the data can be modeled as [142]:

$$m_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \tag{7.8}$$

where μ is the overall mean, α_i stands for the subject effect, β_j stands for the model effect, and ϵ_{ij} are the experiment errors. In Tabs.7.4 and 7.5 the ANOVA results are indicated for objects and subjects from Experiment I and Experiment II respectively. The *F-values* for both subjects ($F = 12.22$ and $F = 18.3$ for Experiment I and Experiment II respectively) and objects ($F = 23.01$ and $F = 21.16$) were large. The right-most column of the table contains the probabilities that the subject effect and the object effect are constant, i.e. that there are no differences among the subjects or among the test objects. It is important to underline that the object effect is not expected to be constant since it depends on the particular 3D object and the 3D objects were deliberately varied.

Table 7.4: Experiment I: ANOVA analysis results.

Source	*SoS*	*df*	*MS*	*F*	*p*
Subjects	372.22	10	37.24	12.22	0
Models	2735.54	39	70.14	23.01	0
Error	1188.83	390	3.04		
Total	4296.82	439			

Table 7.5: Experiment II: ANOVA analysis results.

Source	*SoS*	*df*	*MS*	*F*	*p*
Subjects	382.84	10	38.28	18.3	0
Models	2080.63	47	44.26	21.16	0
Error	983.34	470	2.09		
Total	3446.82	527			

Table 7.6: Experiment I: ANOVA analysis results for $\ln(m_{ij}+1)$.

Source	*SoS*	*df*	*MS*	*F*	*p*
Subjects	21.79	10	2.18	8.92	0
Sequences	151.454	39	3.88	15.89	0
Error	95.33	390	0.24		
Total	268.58	439			

Table 7.7: Experiment II: ANOVA analysis results for $\ln(m_{ij}+1)$.

Source	*SoS*	*df*	*MS*	*F*	*p*
Subjects	21.18	10	2.11	14.54	0
Sequences	132.57	47	2.82	19.35	0
Error	68.50	470	0.14		
Total	222.26	527			

For the F-values in Tabs. 7.4 and 7.5, the probabilities were zero. This means that there is a significant variation in the subjective value means from subject to subject. To check if the variation is caused by variations in the gain factor g_i, an ANOVA is also informed for the natural logarithm of m_{ij}. In fact, by taking the logarithm of the Eq. (7.7) we obtain:

$$\ln(m_{ij}) = \ln(g_i a_j + b_i + n_{ij}) \approx \ln(g_i) + \ln(a_j) + b_i/g_i a_j + n_{ij}/g_i a_j \tag{7.9}$$

In this equation $\alpha_i = \ln(g_i)$, $\mu + \beta_j = \ln(a_j)$ and $\epsilon_{ij} \approx b_i/g_i a_j + n_{ij}/g_i a_j$ (see Eq. (7.8)). ϵ_{ij} is no longer independent from the other factors. However, if b_i and n_{ij} are small, this will not matter much in the analysis. Tables 7.6-7.7 contain the ANOVA results for $\ln(m_{ij}+1)$. Here, we decided to use the $\ln(m_{ij}+1)$ instead of the $\ln(m_{ij})$ to avoid numerical problems to the presence of zero scores. The F-values are large and the probabilities near zero. This means that there were significant subject-to-subject variations in the gain factors and then some form of subject-to-subject correction was required. Several methods for estimating the offsets and the gains are possible. The probability of the null hypothesis (no variation in the subject means) for each correction method and the ANOVA results are summarized in Tab. 7.8. The measurements are adjusted prior to combination in the following way:

$$\hat{m}_{ij} = \frac{1}{\hat{g}_i}(m_{ij} - \hat{b}_i) \tag{7.10}$$

where $\hat{g}_i$ is the corrected gain, $\hat{b}_i$ is the corrected offset and $\hat{m}_{ij}$ is the normalized score.

In the first correction method the offsets are estimated using the mean of all measurements made by each subject:

$$\hat{b}_i = \frac{1}{J}\sum_{j=1}^{J} m_{ij} - \mu \tag{7.11}$$

Since the mean is not a robust estimator of the center of a distribution the median is also tried as an estimate of the offset.

$$\hat{b}_i = \mathsf{median}\left\{m_{ij}\ ,\ \forall j \in J\right\} - \mu \tag{7.12}$$

where J is the set of test models. The results are shown in Tab. 7.8. The mean estimate removes the subject to subject variations for both Experiments.

To adjust the gain as well, two gain estimation methods have been considered. The first gain estimation evaluates the measurements in terms of the experiment instructions and corrects the gain if the instructions were not followed exactly. In fact, the test subjects are told to assign a value of 10 to the worst of the test models seen during the training session. The corrected gain is set to make this true:

$$\hat{g}_i = \frac{1}{K} \max_{j \in J}(m_{ij}) \tag{7.13}$$

where K is equal to the upper end of the scale (10) in our testing procedure. The second method for correcting the gain variation relies on a statistical estimate of each test subject's range. The standard deviation of the values is used to estimate the range. In this case, the gain factor becomes:

$$\hat{g}_i = \frac{4\delta_i}{K} \tag{7.14}$$

where δ_i is the standard deviation of all the values recorded by the i-th subject. The results are summarized in Tab. 7.8. The gain correction (Eq. (7.14)) combined to the mean offset correction provides the best results in term of F-test [34]. Hence, after the normalization the collected data depend on the model but do not depend on the subject. This fact indicates that *the experiment is well-designed*, i.e. the experimental methodology is not affected by any systematic error.

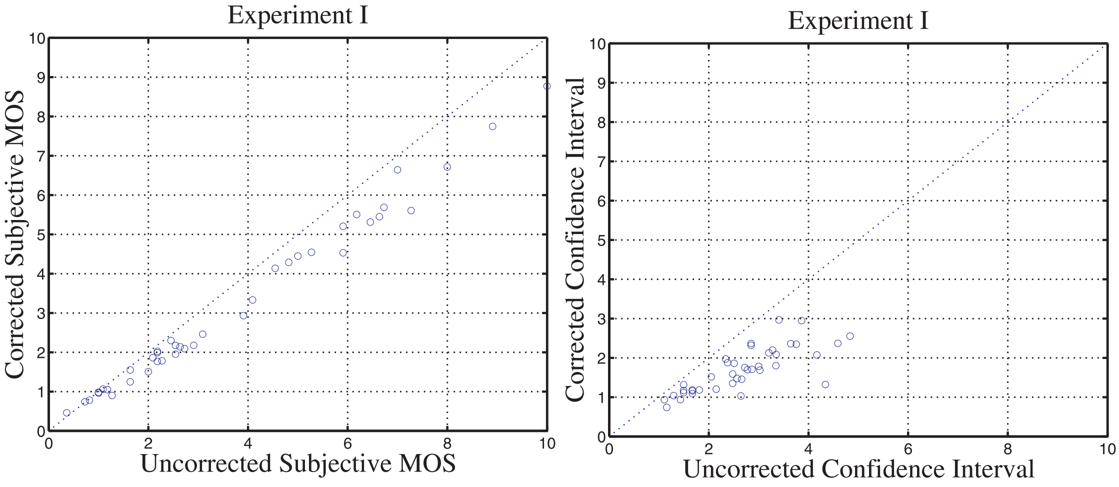
Table 7.8: ANOVA analysis results after normalization (F-test values for subject Offset and Gain dependency). Probabilities near one means that there is no difference across subjects.

	Experiment I		**Experiment II**	
Correction	**Offset**	**Gain**	**Offset**	**Gain**
None	0	0	0	0
Mean	1	0.5557	1	0.6903
Median	0.0006	0.0001	0.0006	0
Mean + max	0.998	0.7962	1	0.8286
Mean + std	1	0.9996	1	0.999

7.5.2 Data Evaluation

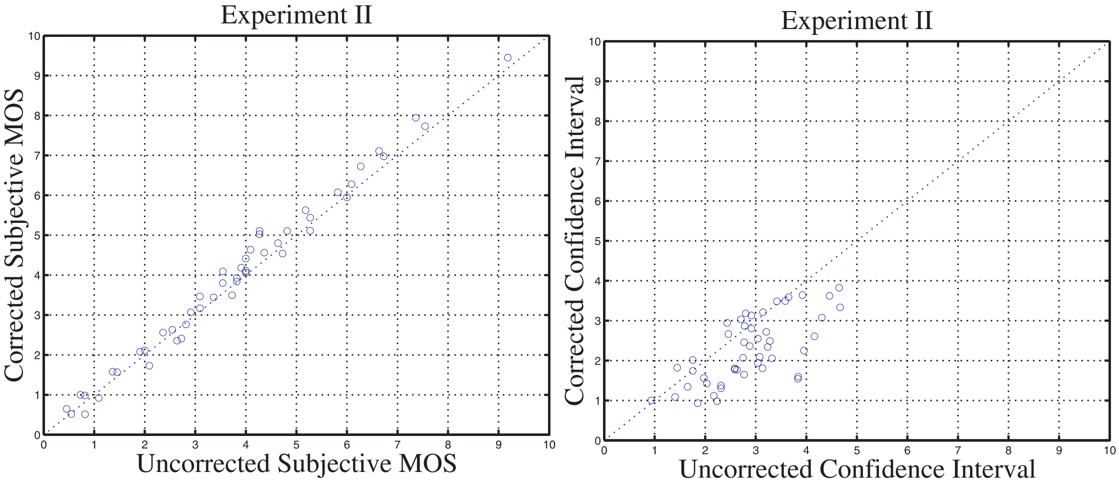
After the normalization process, the subjective score values are screened and combined into a single overall score for each test model using the sample mean described in Sec. 2.3.1. After the normalization and screening, for Experiment I the data of one subject was discarded, and for Experiment II the data of three subjects were rejected. Then, the subjective data were used to test the proposed objective metric of the watermarked 3D object defects. In this section we check the validity of the obtained data and evaluate whether the methodology can be improved. Figures 7.7 and 7.8 show the overall data spread for Experiment I and II. Note that the data spread was good for all experiments. For Experiment I the MOS values ranged from 0.36 to 10.0 before the normalization and screening and ranged from 0.45 to 8.76 after (Fig. 7.7 (a)). For Experiment II the ranges before and after are (0.45 , 9.17) and (0.64 , 9.44) respectively (Fig. 7.8 (a)). The large number of data points (test 3D models) compensated for the lack of precision of individual points.

In summary, the experiments provided good data for most of the test models. The confidence intervals were reduced after the normalization and screening procedure (Figs. 7.7 (b) and 7.8 (b)). The confidence intervals were large for the test models that were hard to examine. This situation could be improved in future experiments by ensuring that the weakly watermarked models are closer



(a) (b)

Figure 7.7: Subject data correction results for Experiment I: (a) the mean scores values and (b) the confidence intervals



(a) (b)

Figure 7.8: Subject data correction results for Experiment II: (a) the mean scores values and (b) the confidence intervals

to perceptual threshold or using many more test subjects. In conclusion, the data were good for the overall fits.

7.6 Proposed Perceptual Metric

Perceptual metrics that compute predictions of human visual task performance from input images are usually based on a vision model. For any application in which a vision model produces reliable performance predictions, its use is almost always preferable to psychophysical data collection. One reason for this preference is that running a model generally costs much less than running a psychophysical experiment to generate the same system performance information, especially when the system evaluation in question is still in the design phase as in our case. There are two approaches [84] to model psychophysical quantities: *performance modeling* and *mechanistic modeling*. Although the distinction is more a continuum than a strict dichotomy, the performance models tend to treat the entire visual system as a *"black box"* for which input/output functions need to be specified. For the mechanistic model, physiological and psychophysical data are used to open the black box. As a result, input/output functions are needed not only for the system as a whole but for a number of component mechanisms within. These components of the model have the same functional response as physiological mechanisms of different stages of the HVS. It is important to underline that, at this time, it does not appear feasible to build a sophisticated model of visual perception of geometric artifacts since such models could become too complex to be handled in practice. In fact, one essential feature of any interactive applications is that 3D objects are observed in motion, so the classical visual models used for still images should be integrated with other perceptual models that take into account the behavior of the human perceptions in a dynamic scene. Additionally, this model should take into account a lot of parameters that depend on the rendering, such as the lighting model, the texturing, and so on. For all of these reasons we opt for the "black box" approach. In particular, we use an objective metric based on surface roughness estimation combined with a standard psychometric function to model the black box. The purpose of a psychometric curve is to associate the values given by the objective metric to the subjective score values provided by the subjects. In this way a match between the human perception of geometric defects and the values provided by the objective metric is established obtaining a perceptual metric. Three kinds of psychometric functions are commonly used [42]: the cumulative Gaussian distribution, the logistic psychometric function, and the Weibull psychometric function. In particular, we use the *Gaussian psychometric function* defined in Sec. 2.3.2:

$$g(a,b,x) = \frac{1}{2\pi}\int_{a+bx}^{\infty} e^{-\frac{t^2}{2}}\,dt \tag{7.15}$$

where a and b are the parameters to be estimated by fitting the objective metrics values as a function of the subjective data and x is the *objective metric* used to measure the visual distortion. To estimate such parameters we use a nonlinear least-squares data fitting by the Gauss-Newton method. We chose this psychometric function since it provided the best fit between our objective metrics and the subjective data.

The intuition and the interviews in Experiment I and II confirm that the watermarking artifacts that produce different kinds of noise on the surfaces can be described essentially with *roughness*. Hence, the objective metric which we chose to measure the strength of the defects is based on an estimation of the surface roughness.

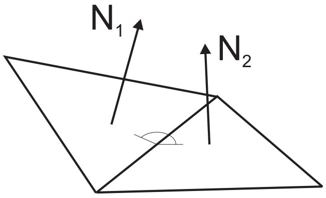


Figure 7.9: Dihedral angle.

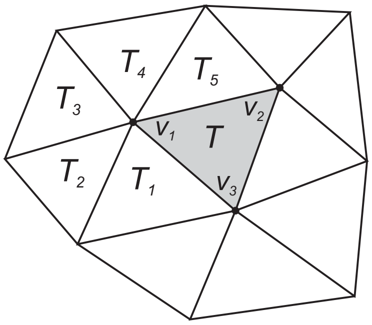


Figure 7.10: $G(.)$ and $V(.)$ descriptions.

7.6.1 Roughness Estimation

With the help of previous studies on 3D watermarking [151], we have realized that a good measure of the visual artifacts produced by watermarking should be based on the amount of roughness introduced on the surface. Moreover, as just said, the interview phase of the two experiments confirmed that "roughness" is a good term to describe, in a general way, the defects introduced over the surface of the model. Hence, two objective metrics based on roughness estimation of the surface have been developed. In the following we give a detailed description of these metrics.

Multi-scale Roughness Estimation

The first roughness measure we propose is a variant of the method of Wu et al. [181]. This metric measures the per-face roughness by making statistical considerations about the dihedral angles associated to each face. Wu *et al.* developed this measure in order to preserve significative shape features in mesh simplification algorithm.

The *dihedral angle* is the angle between two planes. For a polygonal mesh, the dihedral angle is the angle between the normals of two adjacent faces (Fig. 7.9). The basic idea of this method is that the dihedral angle is related to the surface roughness. In fact, the face normals of a smoothed surface vary slowly over the surface, consequently the dihedral angles between adjacent faces are close to zero. To be more specific, Wu *et al.* associated to each dihedral angle an amount of roughness given by the quantity $1 - (\vec{N}_1 \cdot \vec{N}_2)$, where $\vec{N}_i)$ is the normal to the surface. Given a triangle T with vertices v_1, v_2 and v_3, its roughness is computed as:

$$\mathcal{R}_1(T) = \frac{G(v_1)V(v_1) + G(v_2)V(v_2) + G(v_3)V(v_3)}{V(v_1) + V(v_2) + V(v_3)} \tag{7.16}$$

Referring to Fig. 7.10, $G(v_1)$ is the average of the roughness associated to the dihedral angles $T - T_1$, $T_1 - T_2$, $T_2 - T_3$, $T_3 - T_4$, $T_4 - T_5$ and $T_5 - T$. In the same way $G(v_2)$ and $G(v_3)$ are the mean roughness associated to the dihedral angles of the faces adjacent to the vertices v_2 and v_3. Instead, $V(v_1)$, $V(v_2)$ and $V(v_3)$ are the variance of the roughness associated to the dihedral angles of the faces adjacent to the vertex v_1, v_2 and v_3.

A rough surface can be considered as a surface with a high concentrations of bumps of different sizes over it. This metric is able to measure 'bumpiness' of the surfaces at face level, but, if the granularity of the surface roughness, i.e. the size of the bumps, is higher than the medium dimension of one face, this metric fails to measure them correctly. In other words this measure does not take into account the *scale* of the roughness. Our idea is to modify Eq. (7.16) in order to account for different bump scales. The first step to achieve this goal is to transform this per-face roughness estimation in a per-vertex roughness estimation in the following way:

$$\mathcal{R}_1^N(v) = \frac{1}{|S_T^N|} \sum_{i \in S_T^N} \mathcal{R}_1(T_i)\mathcal{A}_{T_i} \tag{7.17}$$

where S_T^N is the set of the faces of the *N-ring** of the vertex v, $|.|$ is the usual cardinality operator and $\mathcal{A}_{T_i}$ is the area of the face T_i. The reason to consider the *N-ring* in the roughness evaluation accounts for different scales of bumpiness. Referring to Fig. 7.11; the bump of size equivalent to the 1-ring (A) is well measured by $\mathcal{R}_1^1(v)$, a correct value of roughness for the vertex v in the case (B) is provided by $\mathcal{R}_1^2(v)$. Approximatively, we can state that the roughness of a vertex v centered on a bump of area close to the area of the faces that form the N-ring is well measured by $\mathcal{R}_1^N(v)$. This approximation could not be valid in certain cases, for example for high values of N, or when a surface presents high curvature. Hence, a real multi-scale measure of bumpiness would require further developments but we assume that this approximation is sufficient. In order to obtain a single value of roughness for each vertex that accounts for the roughness evaluated at several scales we take the maximum value produced by N-ring of different sizes. In particular, in our objective metric we have chosen 3 scales of roughness:

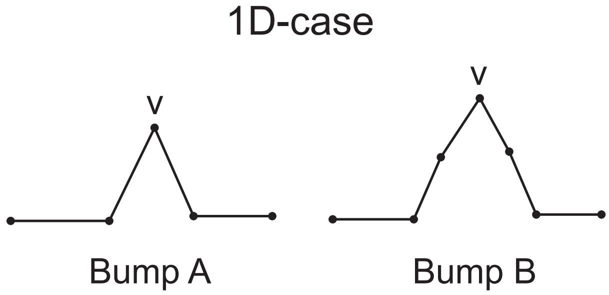


Figure 7.11: Bumps with different scale.

$$\mathcal{R}_1(v) = \max\{\mathcal{R}_1^1(v), \mathcal{R}_1^2(v), \mathcal{R}_1^4(v)\} \tag{7.18}$$

The total roughness of the 3D object is the sum of the roughnesses of all vertices:

$$\mathcal{R}_1(M) = \sum_{i=1}^{N_v} \mathcal{R}_1(v_i) \tag{7.19}$$

where N_v is the total number of mesh vertices. In the following we will see how to transform this multi-scale roughness estimation in an objective metric that correlates well to the human perception of geometric defects.

Smoothing-based Roughness Estimation

The second method we developed to measure surface roughness is based on considerations arising during the subjective tests. Since most of the subjects have said during the interview that the

---

*The *N-ring* neighborhood vertices of a vertex v is an extension of the 1-ring neighborhood. A 2-ring neighborhood is created from the 1-ring by adding all of the vertices of any face containing at least one vertex of the 1-ring. Additional rings can be added in the same way to form the 3-ring, the 4-ring and so on.

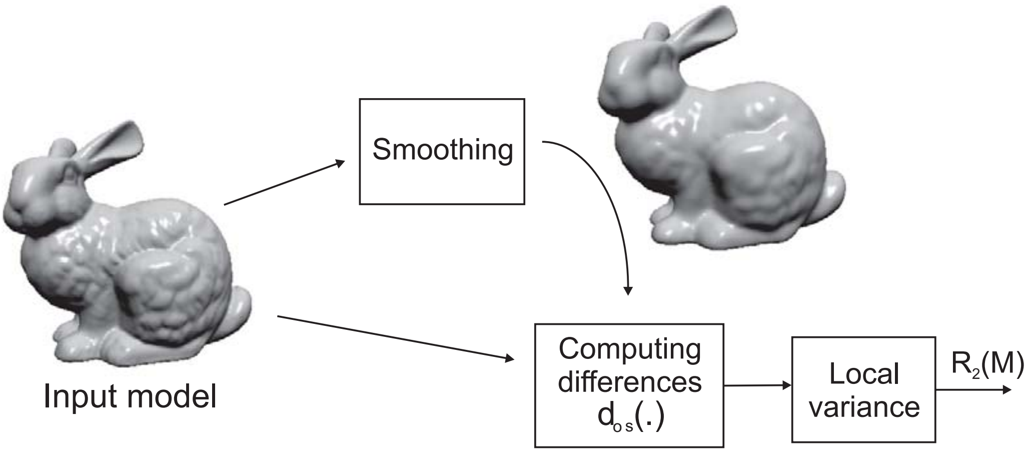


Figure 7.12: Smoothing-based Roughness Estimation.

defects are perceived better on smooth surfaces, we decided to develop a *smoothing-based roughness estimation*. The basic idea of this approach is to apply to the model a smoothing algorithm and then to measure the roughness of the surface as the variance of the differences between the smoothed version of the model and the original one. A sketch of the smoothing-based roughness is depicted in Fig. 7.12.

The first step of this approach is to build a smoothed version of the model (M^S) by applying a *smoothing* algorithm to the input model (M). Several possibilities for smoothing exist [33, 68, 73, 146]. Here, we decided to use the *Taubin filter* [146] for its simplicity of implementation. The parameters of the Taubin filter used are the usual $\lambda = 0.6307$, $\mu = -0.6352$. This filter is iterated 5 times. When the smoothed model is obtained, the distance between each vertex v of M and v^S of M^S is computed in the following way:

$$d_{OS}(v, v^S) = \mathtt{proj}_{\vec{n}_v^S}(v - v^S) \tag{7.20}$$

where $\mathtt{proj}(.)$ indicates the projection of the vector $(v - v^S)$ on the vertex normals of the smoothed surface ($\vec{n}_v^S$). At this point the per-vertex roughness is computed by evaluating the local variance of the distances $d_{OS}(.)$ around each vertex. To be more specific, for each vertex v, the set of distances associated to its *2-ring* ($S_d^2(v)$) is built and the variance of this set evaluated. Then, the per-vertex smoothing-based roughness is computed by:

$$\mathcal{R}_2(v) = \frac{V(S_d^2(v))}{\mathcal{A}_{S^2}} \tag{7.21}$$

where A_{S^2} is the area of the faces that form the 2-ring of v. This area is used as the denominator since surfaces with the same local variance of the distances but smaller area are assumed to be rougher. The roughness of the input model is the sum of the roughnesses of all vertices of the model:

$$\mathcal{R}_2(M) = \sum_{i=1}^{N_v} \mathcal{R}_2(v_i) \tag{7.22}$$

where N_v is the number of vertices of the model.

Objective Metrics

Now, we describe how to use the roughness estimation to predict the visual distortions produced by a certain 3D watermarking algorithm. On the basis of several evaluations we decided to define our objective metric as the increment of roughness between the original and the watermarked model. This increment, $\mathcal{R}(M, M^w)$, is normalized with respect to the roughness of the original model. In formula:

$$\mathcal{R}(M, M^w) = \log\left(\frac{\mathcal{R}(M) - \mathcal{R}(M^w)}{\mathcal{R}(M)} + k\right) - \log(k) \tag{7.23}$$

where $\mathcal{R}(M)$ is the total roughness of the original model and $\mathcal{R}(M^w)$ is the total roughness of the watermarked model. Both $\mathcal{R}_1(.)$ and $\mathcal{R}_2(.)$ can be used to obtain two different objective metrics for 3D watermarking quality evaluation. The logarithm is employed to better discriminate low values of relative roughness increments. The constant k is used to avoid the numerical instability of (Eq. (7.23)) since the logarithm tends to $-\infty$ for M^w very close to M. In particular the value of k has been set to normalize the values provided by the metric between 0 and 10, that is the same range of values used by the subjects during the experiments. In the following, we indicate with $\mathcal{R}_1(M, M^w)$ the objective metric based on the multi-scale roughness and with $\mathcal{R}_2(M, M^w)$ the objective metric based on the smoothing-based roughness estimation.

7.7 Experimental Results

In this section we analyze the performances of the two proposed objective metrics and we compared them with geometric metrics usually adopted in literature for model quality evaluation. First, the correlation between the subjective Mean Opinion Score (MOS) collected in Experiment I and the distances given by two geometric metrics based on the Hausdorff distance for model similarity is evaluated. In this way we obtain a term of comparison for the evaluation of our metrics. Then, the objective metrics are fitted with the Gaussian psychometric curve of Eq. (7.15) to match the subjective data collected in the first experiment. The performances of the *perceptual metrics* so obtained are evaluated using the subjective MOS provided by the Experiment II. In other words the subjective data of the Experiment II are used to *validate* the developed perceptual metrics. The results obtained will be discussed at the end of this section.

7.7.1 Hausdorff distances

As previously stated (Section 7.2) two of the most common geometric metrics used to measure the similarity between two 3D objects are the Maximum (Eq. (7.3)) and the Mean Geometric Error (Eq. (7.5)). These two metrics are based on the Hausdorff distance between models' surfaces. Here, we want to evaluate if the distance between the original and the watermarked model could be a reliable metric for perceptual watermarking impairments prediction. To do this, the Hausdorff distances of each watermarked models from the original are plotted versus the MOS provided by Experiment I. At this point, the linear correlation coefficient of Pearson (r_P) [172] and the non-linear (rank) correlation coefficient of Spearman (r_S) [81] are calculated in order to evaluate the global performances of the metric obtained by fitting these geometric data with a cumulative gaussian (Eq (7.15)). The Spearman rank correlation coefficient is a measure of the strength of monotone association between two variables. Even if a psychometric curve is used to fit the geometric measures, the results do not correlate well with subjective MOS. This underlines the fact that $d_\infty(.)$ and $d_1(.)$ are not designed on the basis of how humans perceive geometric defects. The results are summarized in Fig.7.13. Such results will be used as a reference to compare the performances of the perceptual metrics based on roughness estimation.

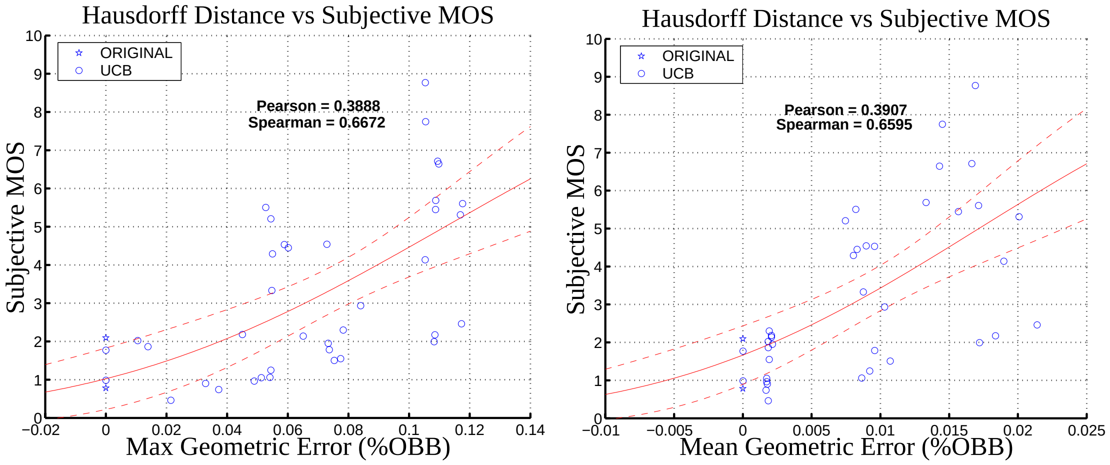


Figure 7.13: Geometric Hausdorff distance vs Subjective MOS.

7.7.2 Roughness-based Metrics Results for Experiment I

As stated in Sec. 7.5, the goal of the first experiment was to make an initial study on the perception of the geometric defects caused by watermarking algorithms. The experimental data confirm that the subjective perception of the impairments is well-described by a measure of roughness. The subjective data of this experiment are used to obtain two perceptual metrics, named $\mathcal{R}_1^*(M, M^w)$ and $\mathcal{R}_2^*(M, M^w)$, from the corresponding two proposed objective metrics $\mathcal{R}_1(M, M^w)$ and $\mathcal{R}_2(M, M^w)$. Those perceptual metrics are obtained by fitting these subjective data with a gaussian psychometric curve (Eq. (7.15)). In this way two kinds of perceptual metrics are obtained, one based on multi-scale roughness measure and another one based on smoothing-based roughness estimation. The parameters of the Gaussian psychometric curve after the fitting are $a = 1.9428, b = -0.2571$ for $\mathcal{R}_1(M, M^w)$ and $a = 2.0636, b = -0.2981$ for $\mathcal{R}_2(M, M^w)$. The smoothing-based one provides a better fit ($r_P = 0.8286$, $r_S = 0.8919$) than the multi-scale one ($r_P = 0.6730$, $r_S = 0.8680$) as depicted in Fig. 7.14. The 95% confidence intervals of the subjective MOS versus the roughness metric are depicted in Fig. 7.14 (Top). Few confidence intervals are large, approximatively 20% of the maximum range scale. Note that the width of the intervals would have been reduced if the experiment had been carried out with more subjects. On the right top of the graphs it is possible to notice some points outside the fitting curve. Most of these outliers correspond to the Venus model. This is due to the fact that the Venus model represents a human face. Human face images are well-known in subjective experiments as a high level factor attracting human attention, i.e. people are more able to deal with human faces, so the distortions on the Venus head are perceived as more visible and annoying with respect to the other models.

7.7.3 Objective Metrics Performances

As discussed in the previous section, the two proposed objective metrics have been transformed into two corresponding perceptual metrics using the data from Experiment I. In order to evaluate such metrics, Experiment II was carried out with three other watermarking algorithms: KDK, NBE and VFA (described in Sec. 7.3). The validation is very simple: the perceptual metrics obtained in Experiment I are used to predict the MOS obtained in the second experiment and their correlation coefficients are computed. The correlation coefficients r_P and r_S are reported in Tab. 7.9. The rows indicate the watermarking algorithm groups. The first two columns of this table report the

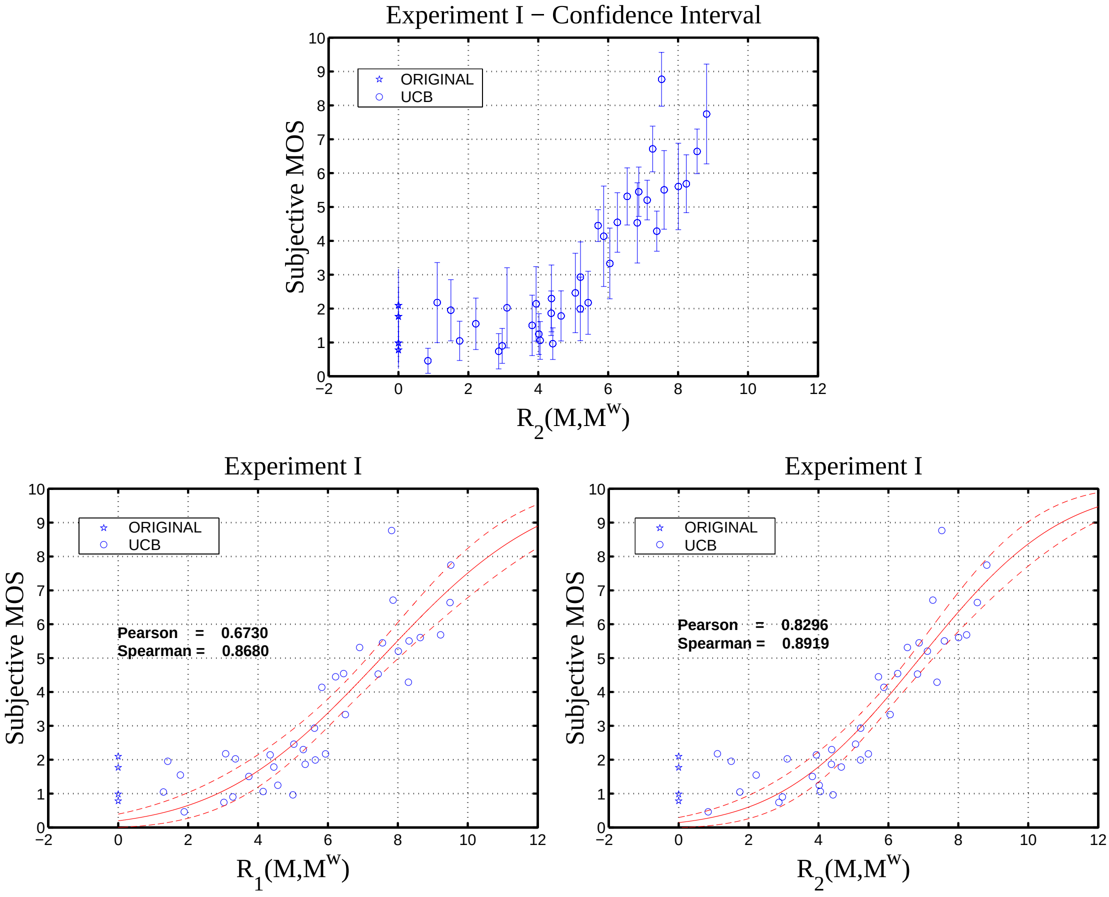


Figure 7.14: Experiment I: Subjective MOS versus objective metrics curves fits.

Spearman correlation coefficient of the Maximum and Mean Geometric Error for comparison. The third and the fourth column shown the values of r_P and r_S for $\mathcal{R}_1^*(M, M^w)$, while the last two columns are the r_P and r_S values for $\mathcal{R}_2^*(M, M^w)$. Referring to this table we can make the following important considerations:

- Overall, both geometric metrics based on the Hausdorff distance do not correlate well with the subjective data. On the other hand the developed metrics exhibit strong correlation with the subjective data, in particular concerning the Spearman's coefficient.

- The Spearman's coefficients for the NBE and VFA algorithms (third and fourth rows respectively) demonstrate that both metrics are able to predict impairment introduced by these two algorithms.

- The worst performances of the proposed metrics are obtained for the KDK algorithm. This can be explained by considering that the distortion produced by the KDK algorithm on the surface are *non-uniform*.

- The results of Experiment II are reported in the 5th row of the table. The values of correlation coefficients ($r_S = 0.7062$ for the first metric, $r_S = 0.6929$ for the second metric) outperform the results provided by the state of the art metric ($r_S = 0.3759$ for the Maximum Hausdorff metric, $r_S = 0.4853$ for the Mean Hausdorff metric).

- The overall performances of the perceptual metrics for the watermarking algorithms that introduce *uniform* distortions on the surface shown are reported in the 6th row of the table. The values of the correlation coefficients ($r_P = 0.6455$ and $r_S = 0.8416$ for the first metric, $r_P = 0.7383$ and $r_S = 0.8954$ for the second metric) are very high. Hence, the developed metrics provide a very good prediction of the impairment caused by 3D watermarking.
- The overall performances of the perceptual metrics considering all the *uniform* and *non-uniform* watermarking algorithms tested are reported in the last row of the table. Despite the presence of the KDK algorithm, for which the performance are not high, the global prediction of the metrics still remains good. In particular, such performances are excellent comparing with the ones of the two geometry-based metrics.

Table 7.9: Perceptual metrics performances.

	Hausdorff Distance		$\mathcal{R}_1^*(M, M^{\mathbf{w}})$		$\mathcal{R}_2^*(M, M^{\mathbf{w}})$	
Algorithms	**Max (r_S)**	**Mean (r_S)**	r_P	r_S	r_P	r_S
UCB	0.6672	0.6595	0.6730	0.8680	0.8296	0.8919
KDK	0.6904	0.3230	0.6154	0.7171	0.5514	0.7111
NBE	0.7087	0.7026	0.5597	0.7917	0.6240	0.8146
VFA	0.4951	0.8815	0.7472	0.9389	0.7763	0.9147
KDK, NBE, VFA	0.3759	0.4853	0.4877	0.7062	0.4982	0.6929
UCB, NBE, VFA	**0.5219**	**0.6183**	**0.6455**	**0.8416**	**0.7383**	**0.8954**
ALL	**0.4993**	**0.5352**	**0.6098**	**0.8122**	**0.6487**	**0.8380**

In order to visualize the results of Tab. 7.9, the graphs of Fig. 7.15 show the values of the objective metrics plotted versus the subjective MOS for several watermarking algorithm groups. The curve drawn on this figure does not represent the result of a fit; the same gaussian curve obtained with the data of Experiment I are drawn for all the pictures. In other words, these graphs visualize the behavior of KDK, NBE and VFA algorithms with respect to the perceptual metrics developed after Experiment I, that is represented by the red curve (the dashed line is the confidence interval for that curve).

Since the non-linear correlation coefficient of Spearman is based on the rank of the data instead of the data values themselves like the Pearson's coefficients, it is interesting to compare the watermarked models ranking by the impairments perceived by the subjects and by the impairments predicted by the metrics. An example of this comparison is reported in Fig. 7.16 where the Bunny and the Feline models are considered. It is possible to see that the smoothing-based perceptual metric, that has values of r_S slightly higher with respect to the multi-scale metric, is able to rank the watermarked models in a way very close to the subjective rank.

7.8 Conclusions

In this work, our investigations about the extension of the ideas of perceptual image watermarking to 3D watermarking have been presented. In particular, a new experimental methodology for subjective quality assessment of watermarked 3D objects has been proposed. The analysis of the data collected by two subjective experiments that use this methodology demonstrates that such methodology is well-designed and provides reliable subjective data about quality evaluation of watermarked 3D objects. Moreover, two perceptual metrics for 3D watermarking impairment prediction have been

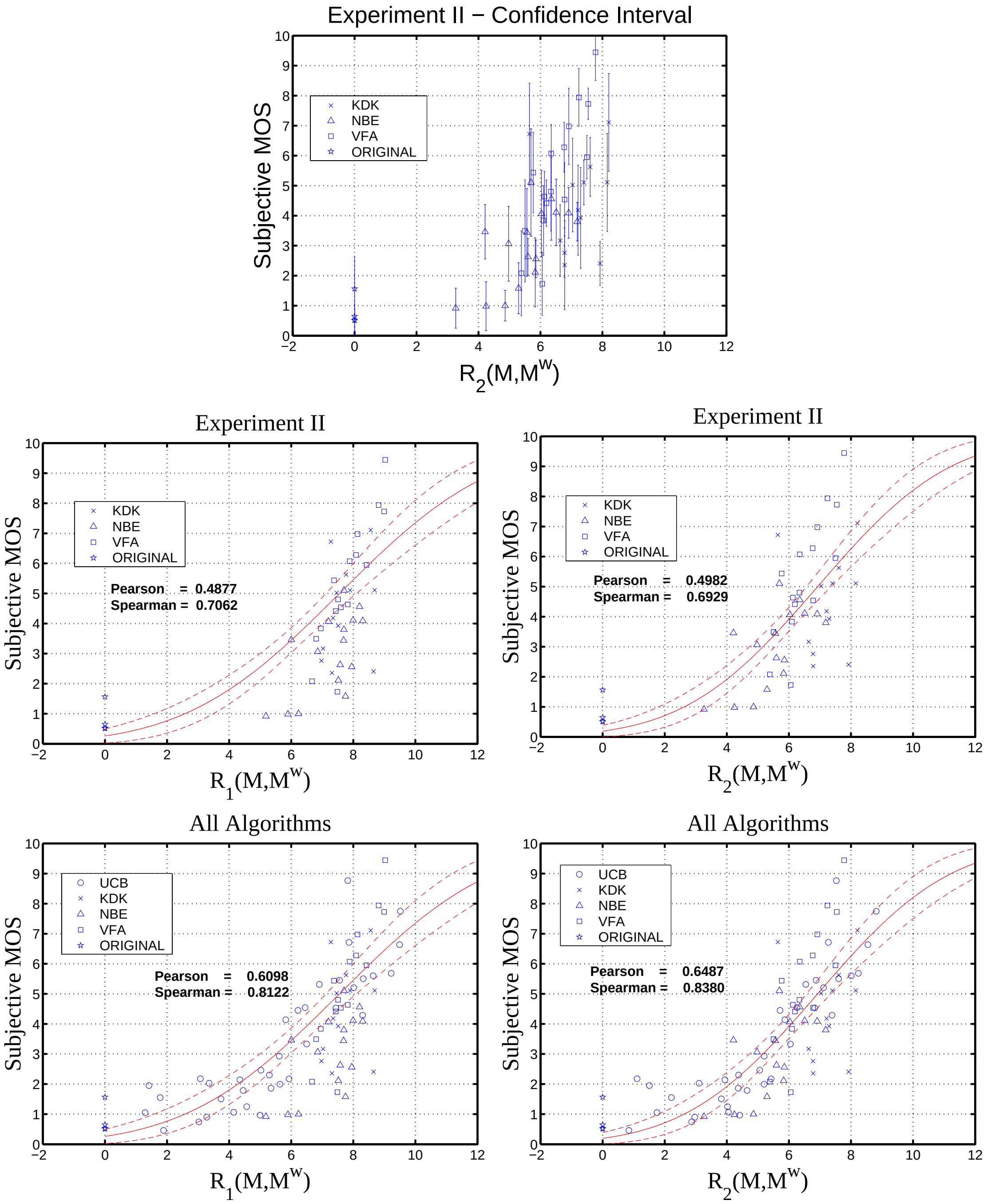


Figure 7.15: Experiment II: Subjective MOS vs objective metric curves. The parameters of the fitting curve are the same of the Experiment I.

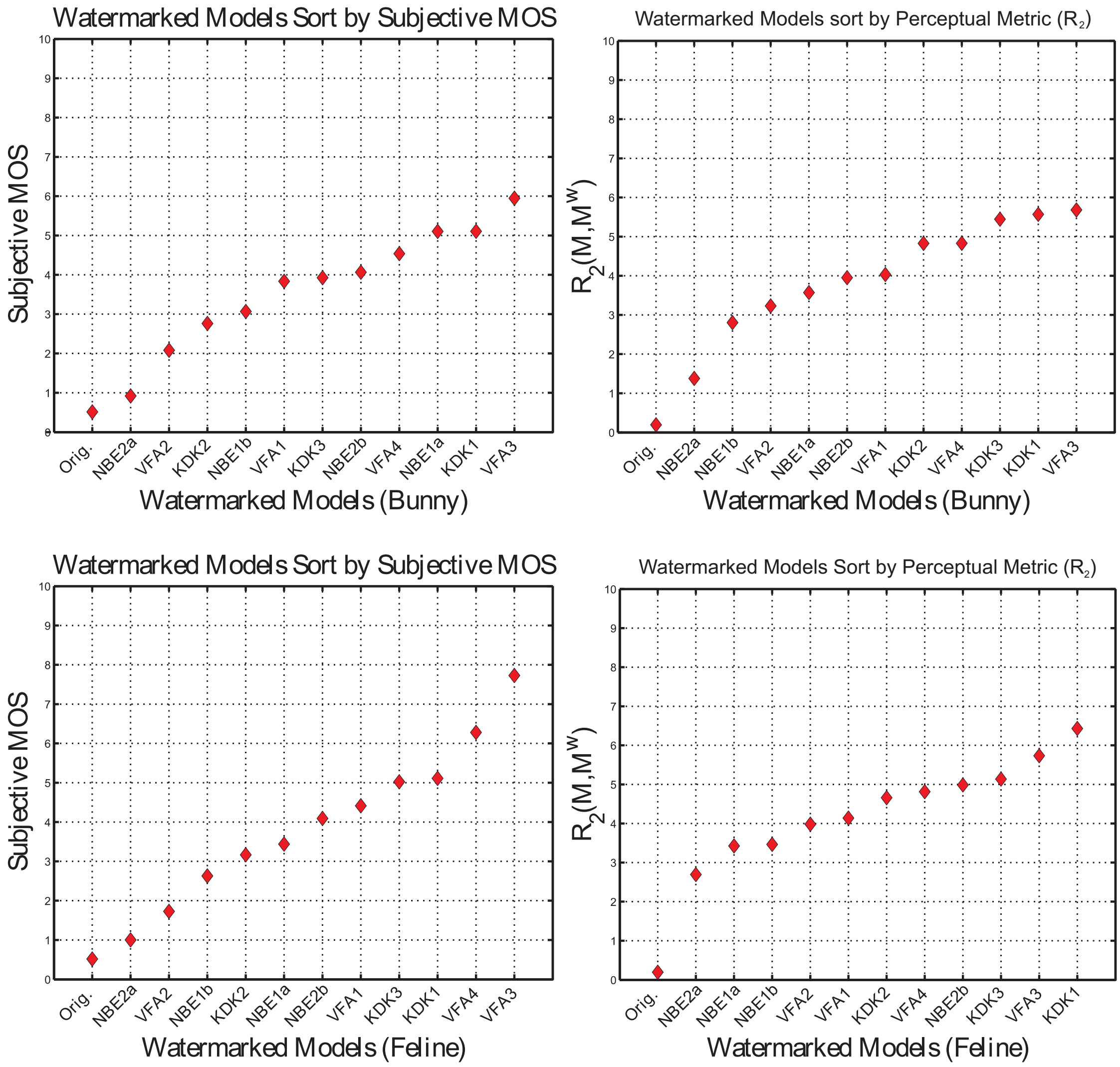


Figure 7.16: Experiment II: Comparison between models' impairment ranking. On the left the models are ranking by subjective MOS. On the right the models are ranking by smoothing-based roughness.

developed by combining roughness estimation with subjective data. The performances of these metrics have been deeply analyzed. The results of this analysis demonstrate the effectiveness of the proposed perceptual metrics with respect to the state-of-the-art geometric metrics commonly used for models comparison. More important, the experimental results show that the proposed metrics provide a good prediction of the human perception of the distortions introduced by 3D watermarking over the model's surface. Hence, these metrics could be used in a feedback mechanism to tune the watermarking parameters of 3D watermarking algorithms optimizing the watermark insertion. For example, referring to UCB watermarking algorithm, for each level of resolution the maximum amount of watermark strength before reaching watermark perceptibility can be easily computed using these metrics, thus improving the robustness of the algorithm while ensuring imperceptibility.

Concluding we can state that, despite the fact that the perceptual evaluations of geometric defects is a very difficult task due to the enormous number of influencing factors, these first results are very encouraging. Further researches can regard the evaluation of the performances of the proposed metrics under different rendering conditions and the extension of the proposed metrics by taking into account the influence of the local properties of the surface (e.g. curvature, protrusions) on the perception of the geometric artifacts.

8 Conclusions

8.1 Summary and Contributions

Quality assessment is a central issue in the design, implementation, and performance testing of all systems. This thesis has discussed how to automatically estimate the visual quality in several visual information processing systems similarly to the way humans perceive it. For the evaluation of the different processing systems, *full reference* quality assessment methods in which the processed (distorted) signal is compared to the reference signal have been adopted. *Performance modeling* has been used to build the quality assessment methods that try indirectly to simulate the Human Visual System. This model treats the entire visual system as a “black box” for which input/output need to be specified. The inputs are the objective measurements of distortions introduced with respect to the reference signal. The outputs are the human responses to such distortions in terms of perceived quality. By finding the output functions that model how human perceive different kinds of distortions through psychophysical experiments, the objective measurements have been transformed into perceptual metrics.

In this work, new objective quality metrics have been proposed for three different visual information processing systems: video watermarking, video object segmentation and 3D model watermarking. The combination of objective measures with perceptual factors represents an element of originality and improvement with respect to the state of the art metrics proposed in these three domains.

In **Part I** standard methods for carrying out the subjective experiments, statistical methods for data analysis and objective models, and metrics proposed in image and video compression quality assessment have been presented. These tools have been used to propose subjective experiment methodologies and objective metrics for the different visual information processing systems considered in this thesis.

Two new objective metrics have been proposed for video watermarking quality assessment. Different watermarking algorithms and video sequences have been judged by means of subjective experiments. Watermarked video sequences were found to suffer mostly from added high-frequency noise and/or flicker in performed subjective tests. These two metrics have been proposed to analyze the video by specifically looking for watermarking impairments, namely the noise metric and the flicker metric, which measure the perceptual impact of these specific distortions. The Spearman and Pearson correlation coefficients are applied to test the performance of the proposed objective quality assessment methods versus the subjective opinion score. Through subjective experiments it has been demonstrated that the proposed metrics are reliable predictors of perceived noise and perceived flicker and clearly outperform PSNR measures in terms of prediction accuracy.

In **Part II**, the importance of efficient image/video segmentation has been discussed, and in particular, video object segmentation. This thesis has provided an original and complete picture of the existing solutions concerning subjective and objective segmentation quality assessment. In addition, the advantages and disadvantages of each approach have been analyzed in depth. On the basis of the discussed approaches, a methodology for subjective evaluation of video object segmentation has been proposed. In the field of segmentation quality evaluation, no formal subjective test procedures have been defined until now, and thus no mean opinion scores are readily available for setting precise targets for the objective metrics to meet. The proposed methodology can be used to subjectively rank different segmentation results and thus serve as reference target for the objective evaluation procedures to be developed. According to the proposed methodology, synthetic test sequences were generated with four types of spatial artifact commonly found in video object segmentation, (added regions, added background, border holes, inside holes) and a temporal artifact (flickering). Moreover, combinations of all artifacts were simulated to study how they interact in the temporal dimension and in the overall annoyance. To obtain a *perceptual* objective metric, several subjective experiments were performed to characterize the different individual segmentation artifacts and their combination. A set of new perceptual objective metrics for estimating the annoyance of the typical segmentation artifacts was derived. The final metric has been proposed by combining the individual artifact metrics using the Minkowski metric and a linear model. Both models presented a very good correlation with the subjective data with no statistical difference in performance. The insertion of the perceptual aspect in such a metric represents an element of originality in video object segmentation evaluation criteria.

An in-depth evaluation of the performance of the proposed method was carried out. The perceptual metric was tested on different video object segmentation techniques for general frameworks as well as specific applications, ranging from object-based coding to video surveillance. To the best of our knowledge, a comparison among different objective metrics for video segmentation quality assessment has received little attention by the image processing community so far, as well as the study of their performances on different segmentation techniques. Moreover, in the literature, segmentation applications are often neglected in the performance analysis of objective quality assessment methods. In all these contexts, the proposed perceptual metric has proved its reliability and good performance.

In addition, the choice of appropriate metric parameters according to the specific application is an element of originality in the proposed approach, too. In fact, when developing segmentation evaluation criteria for specific applications, the characteristics of the application itself have provided valuable information for the selection of appropriate segmentation artifact weights. For the considered applications, especially compression, augmented reality and video surveillance, different *perceptual* weights have been found on the basis of subjective experiments. In the compression scenario, the perceptual weights obtained for missing part of the objects are larger than those for added ones. This is due to the fact that parts of objects are erroneously considered as belonging to the background and thus compressed. In the surveillance application, added object weights are large since they can be confused as objects and thus causing false-alarms. Furthermore, missing objects weights are also large since they could cause dangerous missed alarm situations. In the augmented reality scenario, shape artifacts have the largest weights since they compromise the overall impression of the interactive story in which the characters are cut and pasted on the virtual background.

In **Part III**, the subjective and objective quality assessment of watermarked 3D models has been investigated. In particular, a new experimental methodology for subjective quality assessment of watermarked 3D objects has been proposed. The analysis of the data collected by two subjective experiments that use this methodology demonstrates that such methodology is well-designed and provides reliable subjective data about quality evaluation of watermarked 3D objects. Moreover, two perceptual metrics for 3D watermarking impairment prediction have been developed by combining roughness estimation with subjective perception of distortions. The results of performance analysis demonstrate the effectiveness of the proposed perceptual metrics with respect to the state-of-the-art geometric metrics commonly used.

Concluding, we can state that, despite the fact that perceptual evaluations of visual quality is a very difficult task due to the enormous number of influencing factors, the achieved results in these relatively unexplored fields are very encouraging. The performances of the proposed metrics in all the addressed fields have been deeply analyzed. The integration of perceptual factors has allowed the methods introduced in this thesis to achieve an improvement with respect to the state of the art methods, as demonstrated by the comparisons with different existing metrics. The new metrics have been tested on several processing techniques allowing the characterization of a broad range of artifacts. The experimental results showed that the proposed metrics provide a good prediction of the human perception of the introduced distortions. Hence, these metrics have been used to benchmark different processing techniques respectively for video watermarking, video object segmentation and 3D watermarking algorithms. In conclusion, instead of time-consuming and expensive subjective evaluation, these perceptually driven objective evaluations can be used to provide guidelines for optimizing visual information processing systems, their parameter settings, and to benchmark systems and algorithms.

8.2 Perspectives

Although the subjective testing has confirmed the performance of the objective metrics presented in this thesis, there are many areas which still require further research. This section discusses some of the improvements and extensions which may be made to each of the objective metrics presented in this thesis.

Objective metrics

- The introduced metrics presented good correlation with the perceptual amount of distortion. However, no human vision model was directly included in the proposed metrics. Including simple models like contrast sensitivity, temporal and spatial masking could improve the metric performance in watermarking system quality evaluation.

- The presented quality evaluation methods were all based on *full reference* methods. The development of *no reference* methods considering the distortion perception still remains an interesting and unexplored topic in the evaluation of the considered visual information processing systems.

- Conclusions drawn using a testbed on specific image, video sequences or 3D models do not *a priori* generalize to other types of content. Additional tests could give better indications of how well they generalize to different contents. Nevertheless, the analysis of the results from any testbed may suggest avenues of exploration in the research of both image/video/3D processing techniques and of their evaluation methodologies.

- The proposed metrics could be used in a feedback loop to tune the visual information processing system parameters in order to optimize the output visual quality.

Video watermarking quality evaluation

Future extension of the work developed in video watermarking quality assessment are listed below.

- A number of watermarking schemes for still images was adopted and they were applied to each frame of the video sequences. More genuine video watermarking schemes should be used to evaluate the metrics in realistic conditions;
- The adopted watermarking schemes used only the luminance channel, so the evaluation was carried out on monochrome video sequences. An extension of the metrics to color is necessary for a reliable evaluation of watermarking algorithms that use all three color channels.

Video object segmentation quality evaluation

The development of objective metrics for video object segmentation evaluation is a relatively new area of research. Several areas still remain for further work, and some of them are discussed below.

- Video object segmentation ground truth was manually obtained for each test sequence. Generation of various segmentations for each test sequence could be carried out using multiple expert observers. The final ground truth could be statistically obtained on the agreement of the different segmentations in the context of the addressed application.
- The investigation on the perceptions of artifacts in the temporal dimension requires further experiments. The perception of artifacts changing in time was characterized for only one kind of artifact and then extended to the other artifacts. More subjective experiments including the temporal variation of the other artifacts should be carried out. Furthermore, how to model the relation between instantaneous and overall quality needs further investigations. In subjective experiments regarding video sequence quality assessment, the characteristics of human memory in relation to the length of the representation have to be studied more deeply.
- Weights for the quality metric of each object part could be provided by using application dependent criteria of region importance in the scene. Such as a good segmentation of heads and faces could be the most important criteria to judge a good segmentation in video surveillance applications.

3D watermarking quality evaluation

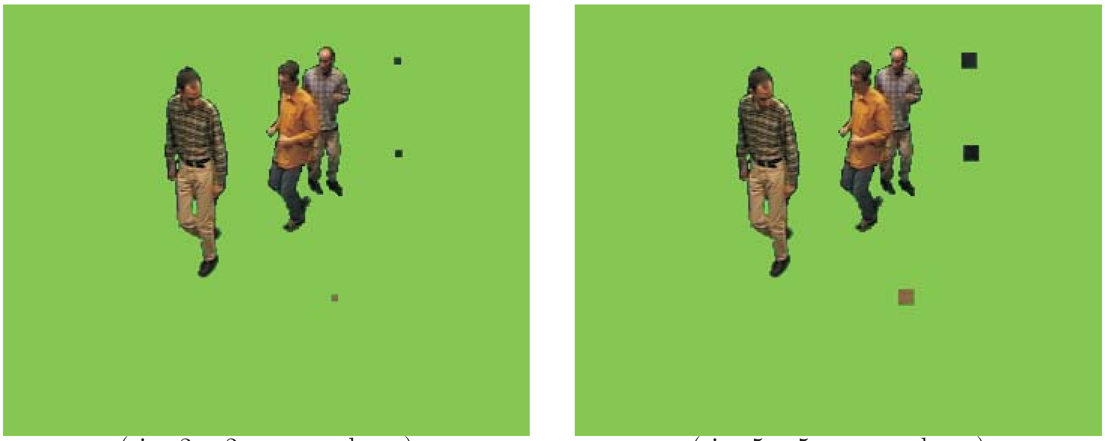
Further researches on 3D watermarking quality assessment can regard:

- in order to avoid complicated rendering effects, plain but effective rendering conditions were kept in the subjective experiments. Such results can be further extended by taking into account more aspects of visualization techniques, such as the role of photorealism in the perception of impairments (see Appendix C);
- extension of the proposed metrics by taking into account the influence of the local properties of the surface (e.g. curvature, protrusions) on the perception of the geometric artifacts.

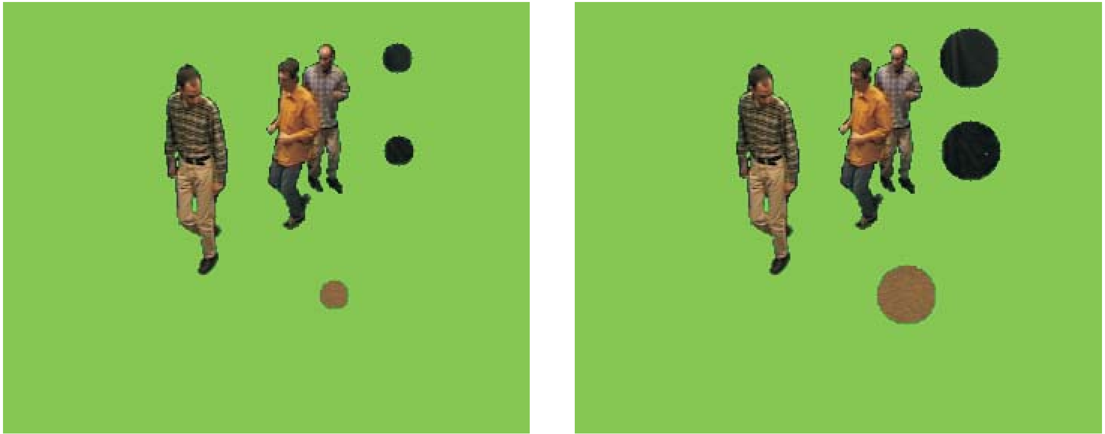
APPENDIX

A

Video Object Segmentation Artifacts

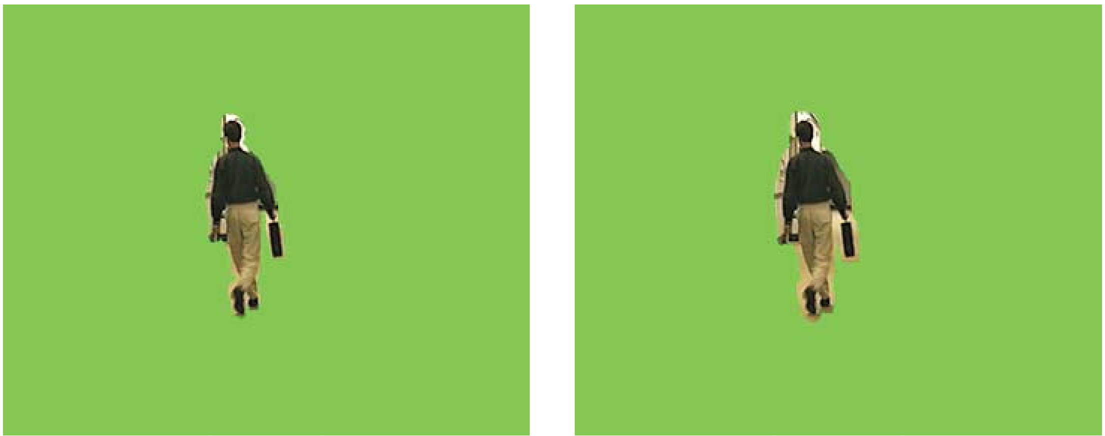


(size 2 × 2, square shape) (size 5 × 5, square shape)

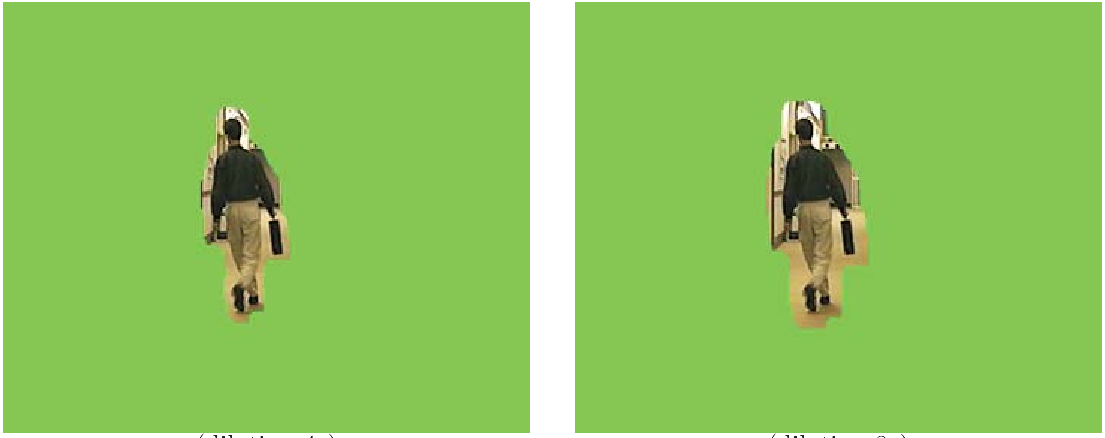


(size 10 × 10, circle shape) (size 20 × 20, circle shape)

Figure A.1: Sample frames for different amounts and shapes of the **added region** artifact for the video sequence “Group”.

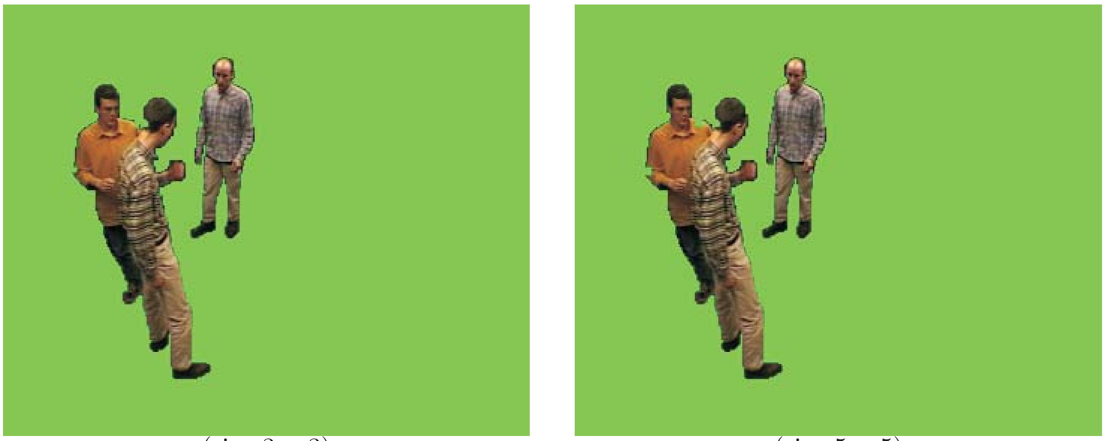


(dilation c) (dilation $3c$)

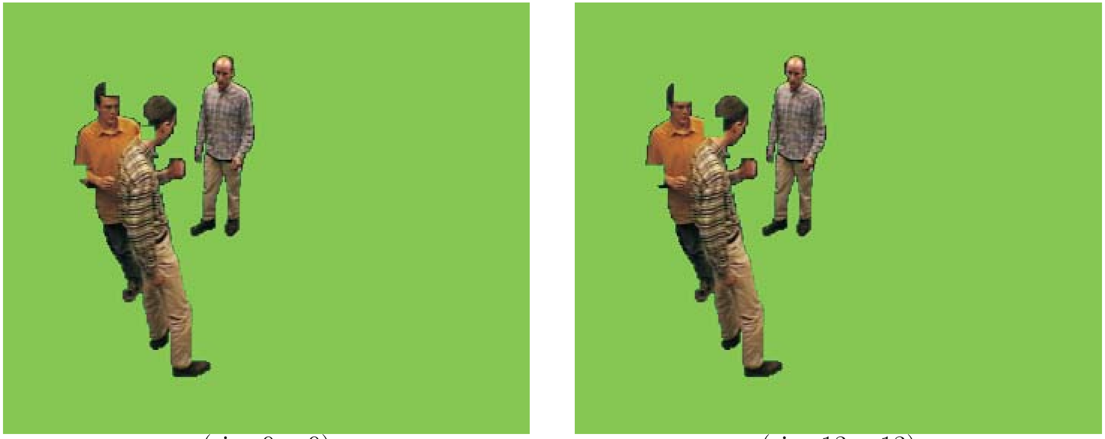


(dilation $4c$) (dilation $8c$)

Figure A.2: Sample frames for different amount of the **added background** artifact (c is equal to number of pixels on the contour object) for the video sequence "Hall monitor".

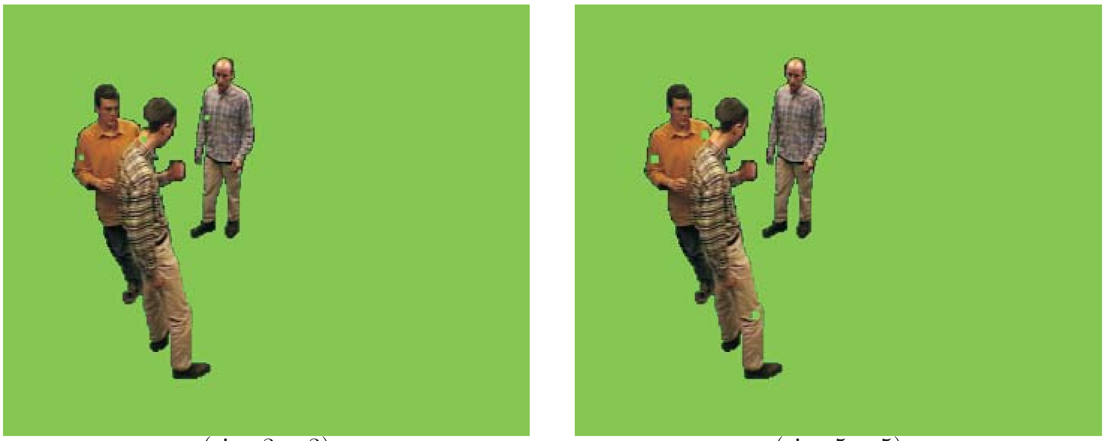


(size 3 × 3) (size 5 × 5)

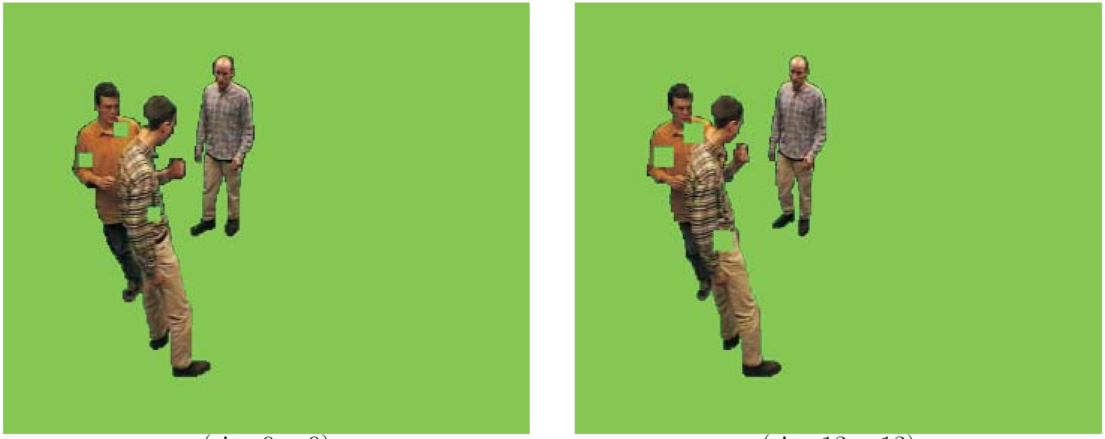


(size 9 × 9) (size 13 × 13)

Figure A.3: Sample frames for different amount of the **border hole** artifact for the video sequence "Group".

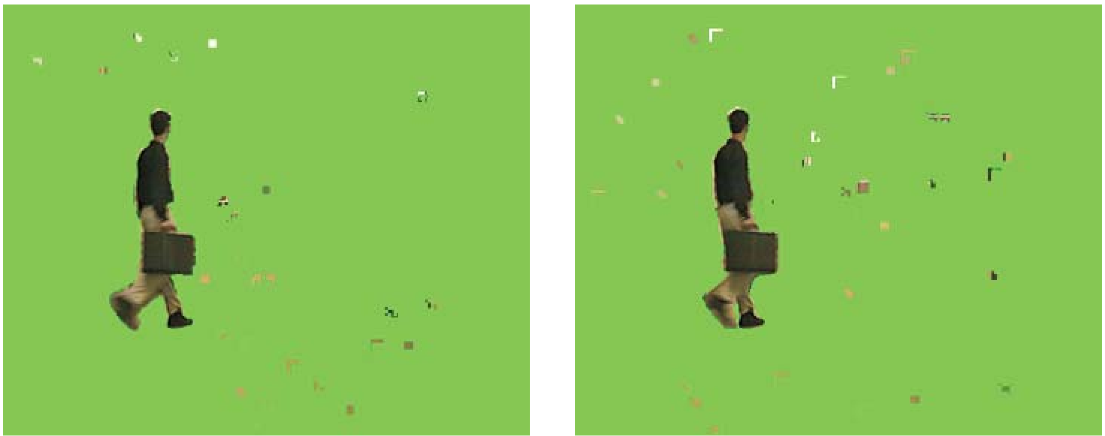


(size 3 × 3) (size 5 × 5)

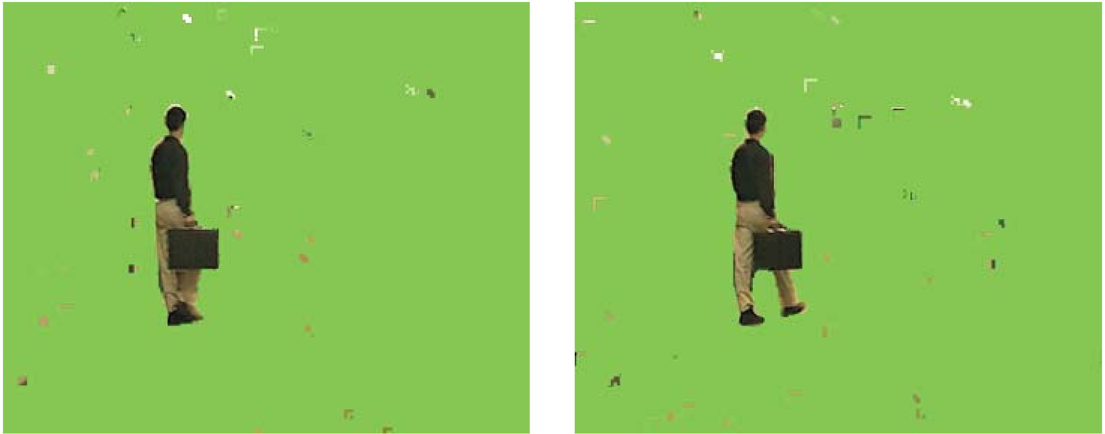


(size 9 × 9) (size 13 × 13)

Figure A.4: Sample frames for different amount of the **inside hole** artifact for the video sequence "Group".

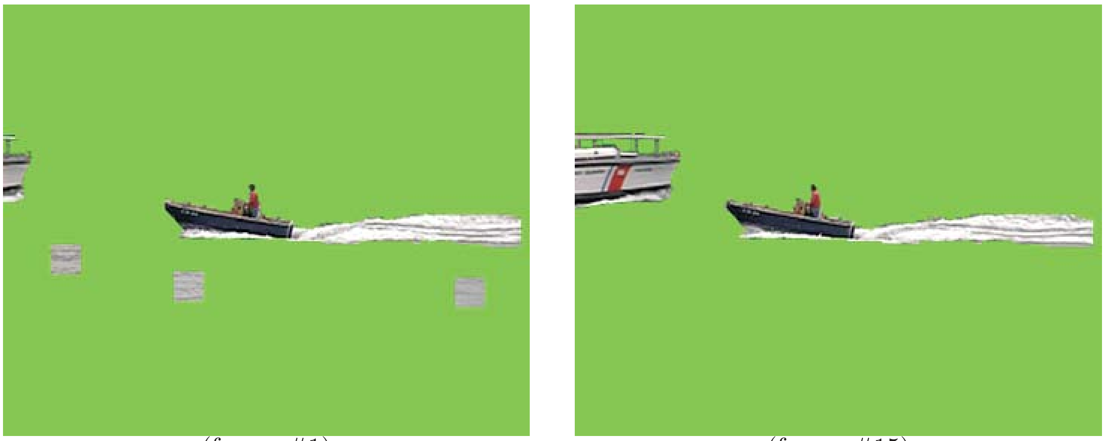


(frame #31) (frame #34)

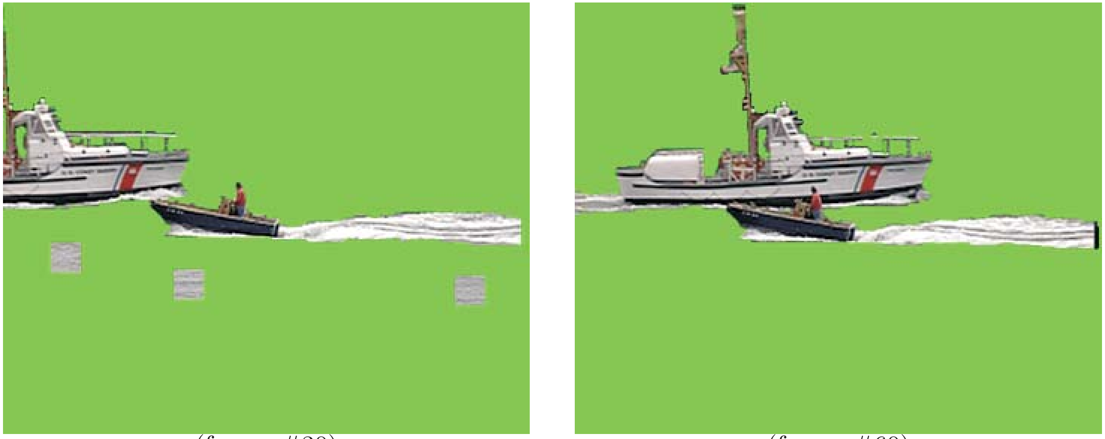


(frame #36) (frame #37)

Figure A.5: Sample frames for **flickering artifact** with period 3 for the video sequence “Hall”.

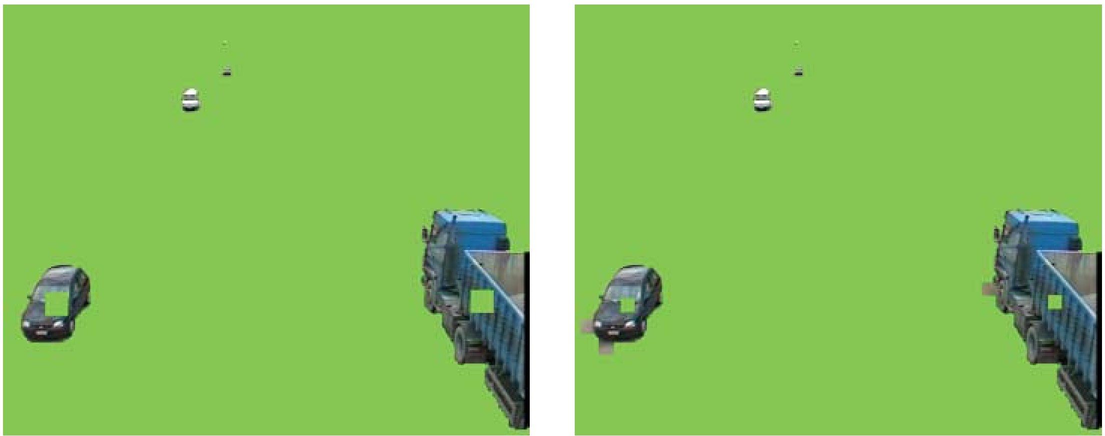


(frame #1) (frame #15)

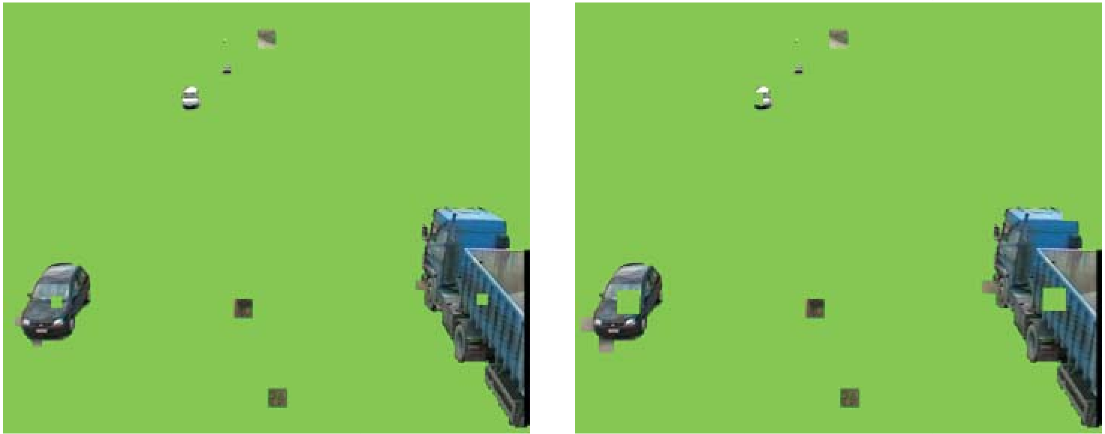


(frame #30) (frame #60)

Figure A.6: Sample frames for **expectation effect** with condition $\mathbf{B}_6$ for the video sequence "Coastguard".



(Combination 4) (Combination 17)



(Combination 26) (Combination 44)

Figure A.7: Sample frames for **combined artifacts**. Examples of one, two, three, four different artifacts (corresponding to combinations 4, 17, 26 and 44) for the video sequence "Highway".

Experiment Scripts B

B.1 Sample Instructions for Segmented Object Annoyance Task

“Before the subject arrives:

1. Set the lighting in the room to the standard illumination.
2. Log in to the server.
3. Double-click on the Segmentation Experiment icon.
4. Click on start.

After the subject arrives, read the following instructions:

- Sit the subject in the chair, centered in front of the video monitor. The subject should be adjusted backward or forward to get a distance of 45 cm from the video monitor screen. The most comfortable position for the subject tends to be leaning forward slightly with forearms or elbows on the table.
- ”Enter your name and hit <**NEXT**>. This study is concerned with defects or impairments in segmented video images and their effect on human viewers. We are not concerned with the content of the videos. We are interested in knowing how you judge any defects or impairments in the videos and, how annoying they are. These are videos of natural scenes in which something moves. The moving object is cut out or segmented from the background and most of the background is uniform green. Let me show an example of the type of video that will be shown. “
- [The experimenter click on <**Play**>].
- ”In these segmented videos only moving objects are extracted and the background is left out. The green color for the background has been chosen only for a better visualization of the

segmented objects. These segmented videos that are displayed, are segmented by hand, but typically automatic methods of segmentation do not work so well. Automatic segmentation often leaves defects outside, inside or on the border of segmented objects. Now, let me show you the typical errors introduced by an automatic method of segmentation."

- [The experimenter click on <**NEXT**>].
- "Do you have any questions? Do you want to see it again? .Now I will explain you how to perform the task. . Prior to each trial, you will look at the center of the screen of the computer monitor. You may move your eyes during the presentation of the clip. You will be presented with one video clip on each trial: the segmented video under test. Each trial will last 5 seconds. The distance from the monitor to your eyes is important during the presentation. Try not to lean forward or backward. The indicator to your right shows the distance at which we would like you to have your eyes. At the end of the presentation, a question will appear on the computer monitor The same question will be asked after every trial. Do not spend a lot of time thinking about your responses. We want to know your initial impressions. You will be asked to indicate how annoying the defect was by entering a number that is proportional to its annoyance value. You are to indicate only the degree of annoyance produced by the defect; you will not be judging the entire clip. Here is how you will determine the annoyance value. I am about to show you a set of sample clips. This will give you an idea of the range of image quality that you will be seeing. Some of the video clips will have one defect at a time, some of them will have two defects at the same time. Moreover, one defect could appear and disappear along the video sequence. You are to assign an annoyance value of 100 to the most annoying defect seen among the sample clips. If the annoyance value of a defect in the experiment is half of the worst sample clip, give it a 50; if it is 1/10th as bad, give it a 10; if it is 1.5 times as bad, give it a 150. If the defect did not annoy you at all, call it 0 (zero). I will show you the sample clips now."
- [Click on <**START**> to proceed with the rest of the example clips.]
- "Did you see any defects? Remember that the most annoying defect that you have seen is to have a value of 100. Use the numeric keypad to enter the annoyance value. After you finish entering the annoyance value, click on <**OK**> or hit <**ENTER**> on the numeric keypad. The computer will start showing the next video clip as soon as you click on <**NEXT**>. Before we start the experiment, you will have nine practice trials to be sure that you understand the task. You will respond in these trials just like you will in the main experiment. Remember that you have one question to answer. How annoying was the defect relative to the worst example? Remember to press <**OK**> or <**ENTER**> after entering the annoyance value and to press <**NEXT**> when you are ready for the next video. We will not use the data from the practice trials, so don't be concerned if you make a mistake here."
- [Start the practice trials by hitting the Mock Test button.]
- "Do you have any questions?
- You can take a break at any time by entering your answers for the most recent video, but waiting to hit <**NEXT**> until you are ready to go on. You should stop if you are confused about what to do, if you realize you have entered data incorrectly, or if you need a break. You cannot stop the video from playing or go back and fix the data from a previous clip after you hit <**NEXT**>. So if something goes wrong, watch the video and then tell the experimenter. We will go back and fix it later. The question will come up on the computer screen after the

video is displayed. The question will not change. There are 196 clips in the experiment and it takes approximately 35 minutes to complete, if you do not take any breaks.

- Do you have any questions?
- At the end of the experiment I will ask a few questions. Start the experiment by hitting the <Proceed to Real Test> button. Finally, when you are ready to start the experiment, let me know so that I can upload the test."
- [Start the experiment.]
- At the end of the experiment, ask the following questions and write down the answers: "How would you describe the defects that you saw? What made a defect the most annoying for you?"

B.2 Sample Instructions for Segmented Object Strength Task

Sit the subject in the chair, centered in front of the video monitor. The subject should be adjusted backward or forward to get a distance of 45 cm from the video monitor screen. The most comfortable position for the subject tends to be leaning forward slightly with forearms or elbows on the table.

- Enter your name and hit <**NEXT**>.
- “This study is concerned with defects or impairments in segmented video images and their effect on human viewers. We are not concerned with the content of the videos. We are interested in wheter or not you see a defect in the segmentation that we will show, and if so, how strong each type of artifact is. These are videos of natural scenes in which something moves. The moving object is cut out or segmented from the background and most of the background is uniform green. Let me show an example of the type of video that will be shown.”
- [The experimenter click on <**Play**>]
- “In these segmented videos only moving objects are extracted and the background is left out. The green color for the background has been chosen only for a better visualization of the segmented objects. These segmented videos that are displayed, are segmented by hand, but typically automatic methods of segmentation do not work so well. Automatic segmentation often leaves defects outside, inside or on the border of segmented objects. Now, let me show you the typical errors introduced by an automatic method of segmentation:
  - The next video segmentation contains a strong added background defect.[Wait until clip finishes]
  - The next two segmented video sequences also contain added background defects. However, in each clip the defect gets weaker and weaker.[Watch the next two clips]
  - The next video segmentation contains a strong added region defect. [Wait until clip finishes]
  - The next two segmented video sequences also contain added region defects. However, in each clip the defect gets weaker and weaker.[Watch the next two clips]
  - The next video segmentation contains a strong holes inside the objects defect. [Wait until clip finishes]
  - The next two segmented video sequences also contain holes inside the objects defects. However, in each clip the defect gets weaker and weaker.[Watch the next two clips]
  - The next video segmentation contains a strong holes on the border of the objects defect. [Wait until clip finishes]
  - The next two segmented video sequences also contain holes on the border of the objects defects. However, in each clip the defect gets weaker and weaker.[Watch the next two clips]
- [The experimenter click on <Next>].
- Do you have any questions? Do you want to see it again?
- Now I will explain you how to perform the task. . Prior to each trial, you will look at the center of the screen of the computer monitor. You may move your eyes during the presentation of the clip. You will be presented with one video clip on each trial: the segmented video under

test. Each trial will last 5 seconds. The distance from the monitor to your eyes is important during the presentation. Try not to lean forward or backward. The indicator to your right shows the distance at which we would like you to have your eyes.

- At the end of the presentation, a question will appear on the computer monitor The same question will be asked after every trial. Do not spend a lot of time thinking about your responses. We want to know your initial impressions.
- You will be asked to estimate the strength of each type (added region, added background, holes inside and holes on the border) of each kind of artifact . The defect can be found in any region of the image and in any time during the clip. If you do not see any defect, call it 0 (zero) for all the defects and do not enter any number. Just press NEXT and then look back at the video screen.
- You will be asked to rate the strength of each kind of artifact by clicking in a number of the correspondent scale. You are to indicate only the strength of the specified type of defect; you will not be judging other types of defects that may be present in the video. If you see a defect which does not fit the categories given, please ignore it. You will be asked to rate the strength of each kind of impairments using four scale bars. Each bar will be labeled with an eleven points scale (0-10). However, each bar contains far more than eleven points and intermediate values are allowed. You will enter the scores by using the mouse to click on each scale. The scale bars will be updated to show on the top the entered strength. Until you click on Next video, the entered values can be adjusted by re-clicking on the scale bars.
- Here is how you will determine the annoyance value. I am about to show you a set of sample clips. This will give you an idea of the range of image quality that you will be seeing. The sample clips include five sets of videos. The first set has five high quality segmentations. The second set has three videos with strong added regions defect. You are to assign a strenght value of 10 to the most annoying defect seen among the segmented video clips. If the strength of a defect in the main experiment is half of the worst sample clip, give it a 5; if it is 1/10th as bad, give it a 1. If the defect did not appear, call it 0 (zero). I will show you the sample clips now.The third set has three segmented videos with strong added background defect. You are to assign a strenght value of 10 to the most annoying defect seen among the segmented video clips. If the strength of a defect in the main experiment is half of the worst sample clip, give it a 5; if it is 1/10th as bad, give it a 1. If the defect did not appear, call it 0 (zero). I will show you the sample clips now. The fourth set has three segmented videos with strong hole inside defect. You are to assign a strenght value of 10 to the most annoying defect seen among the segmented video clips. If the strength of a defect in the main experiment is half of the worst sample clip, give it a 5; if it is 1/10th as bad, give it a 1. If the defect did not appear, call it 0 (zero). I will show you the sample clips now. The fifth set has three segmented videos with strong hole border defect. You are to assign a strenght value of 10 to the most annoying defect seen among the segmented video clips. If the strength of a defect in the main experiment is half of the worst sample clip, give it a 5; if it is 1/10th as bad, give it a 1. If the defect did not appear, call it 0 (zero). I will show you the sample clips now."
- [Click on <**START**> to proceed with the rest of the example clips.]
- Did you see any defects? Remember that the most annoying defect that you have seen is to have a value of 10.

- Use the numeric keypad to enter the annoyance value. After you finish watching the segmentation click on <OK > or hit <**ENTER**> to enter the strength values with the mouse. on the numeric keypad. The computer will start showing the next video clip as soon as you click on <**NEXT**>. If you took more than 6-8 seconds to answer the questions, the next clips will start showing immediately.
- Before we start the experiment, you will have twelve practice trials to be sure that you understand the task. You will respond in these trials just like you will in the main experiment. Remember that you have one question to answer. How strong was the defect relative to the worst example? Remember to press <OK> or <**ENTER**> after entering the annoyance value and to press <**NEXT**> when you are ready for the next video. We will not use the data from the practice trials, so don't be concerned if you make a mistake here.
- [Start the practice trials by hitting the Mock Test button.]
- Do you have any questions?
- You can take a break at any time by entering your answers for the most recent video, but waiting to hit <**NEXT**> until you are ready to go on. You should stop if you are confused about what to do, if you realize you have entered data incorrectly, or if you need a break. You cannot stop the video from playing or go back and fix the data from a previous clip after you hit <**NEXT**>. So if something goes wrong, watch the video and then tell the experimenter. We will go back and fix it later.
- The question will come up on the computer screen after the video is displayed. The question will not change.
- There are 180 clips in the experiment and it takes approximately 35 minutes to complete, if you do not take any breaks.
- Do you have any questions?
- At the end of the experiment I will ask a few questions. Start the experiment by hitting the <Proceed to Real Test> button. Finally, when you are ready to start the experiment, let me know so that I can upload the test.
- [Start the experiment.]
- [At the end of the experiment, ask the following questions and write down the answers:]
- "How would you describe the defects that you saw?
- What made a defect the most annoying for you?"

B.3 Sample Instructions for Compression Application Annoyance Task

- This part of the trial is concerned with the quality of the segmentation in connection with the application of video compression. You will see the sequences obtained using different segmentation algorithms and you should judge them giving a valuation that considers all the defects and the impairments noticed. Remember that we are not concerned with the content of the videos: we want to know how you judge any imperfections in the segmentations and how annoying they are. The videos that you will examine contain one or more objects extracted from the original by the segmentation algorithms and put over a compressed version of the background. Let me show an example of the type of video that will be shown.
- [Click on <**Play**>].
- This kind of application is related to the preceding one of video-surveillance, whose systems often include blocks for the shots recording or transmission (to a security office) that are activated in the presence of anomalous situations. Introducing compression is useful to ensure a real time transmission or to occupy less storage capacity, but it is also important that the people, the things, the shapes and the parts of the objects are recognizable: a solution is encoding the segmented objects with the best possible quality and the background with a lower quality (introducing compression). To judge the segmentations for this application you should look at the defects and the imperfections also considering these requirements about the identification and the recognizability of the objects. Keep in mind that what you have to evaluate is the quality of the segmentation, not how the video are compressed.
- Now I will explain you how to perform the task. Prior to each trial, you will look at the center of the screen of the computer monitor. You may move your eyes during the presentation of the clip. You will be presented with one video clip on each trial: the segmented video under test. Each trial will last 5 seconds. It is important that the distance from the monitor to your eyes is preserved during the presentation, so try not to lean forward or backward. The indicator to your right shows the distance at which we would like you to have your eyes.
- At the end of the presentation, a question will appear on the computer monitor. The same question will be asked after every trial. Do not spend a lot of time thinking about your responses: we want to know your initial impressions. You will be asked to give a global judgment about the observed video by entering a number that is proportional to the annoyance degree perceived as a result of the imperfections in the segmentation.
- Here is how you will determine the annoyance value. I am about to show you a set of sample clips. This will give you an idea of the range of image quality that you will be seeing. Among these videos you should identify the worst segmentation and mentally assign an annoyance value of 100 to it. Then, in the experiment, you will use that video as a term of comparison to judge the other ones: for instance if the annoyance value of a defect in the segmentation is half of the worst sample clip, give it a 50; if it is 1/10th as bad, give it a 10; if it is 1.5 times as bad, give it a 150. If the defect did not annoy you at all, call it 0 (zero). I will show you the sample clips now.
- [Click on <**START**>]
- Did you see any defects? Do you want to see the segmentations again? Remember that the most annoying defect that you have seen is to have a value of 100.

- Before we start the experiment, you will have nine practice trials to be sure that you understand the task. You will respond in these trials just like you will in the main experiment. Remember that you have one question to answer: how annoying was the defects in the segmentations relative to the worst example? Use the numeric keypad to enter the annoyance value, after you finish entering the annoyance value, click on <OK > or hit <**ENTER**> on the numeric keypad and press <**NEXT**> when you are ready for the next video. We will not use the data from the practice trials, so don't be concerned if you make a mistake here.
- [Start the practice trials]
- Do you have any questions?
- You can take a break at any time by entering your answers for the most recent video, but waiting to hit <**NEXT**> until you are ready to go on. You should stop if you are confused about what to do, if you realize you have entered data incorrectly, or if you need a break. You cannot stop the video from playing or go back and fix the data from a previous clip after you hit <**NEXT**>. So if something goes wrong, for example if you enter 600 instead of 60 and hit <**NEXT**>, watch the video and then tell me what happened specifying the number of the video (it appears up in the window where you put your answer), I will go back and fix it later.
- This part of the experiment includes 30 videos and it takes approximately 5 minutes to complete: the same question will come up on the computer screen after the video is displayed. Remember that we want to know your initial impressions so you should answer soon enter a judgment about the entire clip, and not only about the last frames.
- Do you have any questions?
- [Start the experiment.]

B.4 Sample Instructions for Augmented Reality Application Annoyance Task

- This part of the trial is concerned with a completely different field of application of the segmentation: you will judge the quality of the segmentation in connection with the application of augmented reality. Augmented reality is a technology by which one can insert objects from the real world in a virtual environment. You will see the sequences obtained using different segmentation algorithms and you should judge them giving a valuation that considers all the defects and the impairments noticed. Remember that we are not concerned with the content of the videos: we want to know how you judge any imperfections in the segmentations and how annoying they are. The videos that you will examine contain one or more objects extracted from the original by the segmentation algorithms and put over a virtual background in black and white. Let me show an example of the type of video that will be shown.
- [Click on <**Play**>].
- This application allows the integration between the real world and the virtual world. In particular the segmentation can be used to extract the objects from the real world and then put them over virtual background. The final purpose is entertainment: it could be to create an interactive comic where the people get themselves together with the other things (placed in front of the camera and subsequently segmented) immersed in a virtual environment. Now, you should judge the segmentation considering the observed defects and imperfections and evaluating the suitability of the segmentation algorithms in connection with this type of application.
- Now I will explain you how to perform the task. Prior to each trial, you will look at the center of the screen of the computer monitor. You may move your eyes during the presentation of the clip. You will be presented with one video clip on each trial: the segmented video under test. Each trial will last 5 seconds. It is important that the distance from the monitor to your eyes is preserved during the presentation, so try not to lean forward or backward. The indicator to your right shows the distance at which we would like you to have your eyes.
- At the end of the presentation, a question will appear on the computer monitor. The same question will be asked after every trial. Do not spend a lot of time thinking about your responses: we want to know your initial impressions. You will be asked to give a global judgment about the observed video by entering a number that is proportional to the annoyance degree perceived as a result of the imperfections in the segmentation.
- Here is how you will determine the annoyance value. I am about to show you a set of sample clips. This will give you an idea of the range of image quality that you will be seeing. Among these videos you should identify the worst segmentation and mentally assign an annoyance value of 100 to it. Then, in the experiment, you will use that video as a term of comparison to judge the other ones: for instance if the annoyance value of a defect in the segmentation is half of the worst sample clip, give it a 50; if it is 1/10th as bad, give it a 10; if it is 1.5 times as bad, give it a 150. If the defect did not annoy you at all, call it 0 (zero). I will show you the sample clips now.
- [Click on <**START**>]
- Did you see any defects? Do you want to see the segmentations again? Remember that the most annoying defect that you have seen is to have a value of 100.

- Before we start the experiment, you will have nine practice trials to be sure that you understand the task. You will respond in these trials just like you will in the main experiment. Remember that you have one question to answer: how annoying was the defects in the segmentations relative to the worst example? Use the numeric keypad to enter the annoyance value, after you finish entering the annoyance value, click on <OK > or hit <**ENTER**> on the numeric keypad and press <**NEXT**> when you are ready for the next video. We will not use the data from the practice trials, so don't be concerned if you make a mistake here.
- [Start the practice trials]
- Do you have any questions?
- You can take a break at any time by entering your answers for the most recent video, but waiting to hit <**NEXT**> until you are ready to go on. You should stop if you are confused about what to do, if you realize you have entered data incorrectly, or if you need a break. You cannot stop the video from playing or go back and fix the data from a previous clip after you hit <**NEXT**>. So if something goes wrong, for example if you enter 600 instead of 60 and hit <**NEXT**>, watch the video and then tell me what happened specifying the number of the video (it appears up in the window where you put your answer), I will go back and fix it later.
- This part of the experiment includes 30 videos and it takes approximately 5 minutes to complete: the same question will come up on the computer screen after the video is displayed. Remember that we want to know your initial impressions so you should answer soon enter a judgment about the entire clip, and not only about the last frames.
- Do you have any questions?
- [Start the experiment.]

B.5 Sample Instructions for Video Surveillance Application Annoyance Task

- This part of the trial is concerned with the quality of the segmentation in connection with the application of video-surveillance. You will see the sequences obtained using different segmentation algorithms and you should judge them giving a valuation that considers all the defects and the impairments noticed. Remember that we are not concerned with the content of the videos: we want to know how you judge any imperfections in the segmentations and how annoying they are. The videos that you will examine contain one or more objects highlighted by white or green boundaries using the segmentation algorithms. Let me show an example of the type of video that will be shown.
- [Click on <**Play**>].
- The video-surveillance systems are used in different fields: monitoring the traffic to detect incidents or jams, analysis of the human behavior to identify thefts, brawls or other dangerous situations, security of reserved zones to control the access of a non-authorized person or of abandoned objects. The segmentation can be employed by these systems to identify all the objects in the scene and then detect anomalous situations, for instance one could introduce a post-processing block for the face detection and recognition that activates an alarm if the segmented person is not authorized. Also in less sophisticated systems, where the shots are shown on the monitor and directly controlled by a human operator, the segmentation information can be useful to help him in his task through a scene representation as the one I showed to you, with the highlighted objects. Since these systems work in real time, an essential requirement for the algorithms is a low computational cost while a high precision of the segmented object contours is not necessary: what is important is that all the objects are identified and entirely cut out because an only partially detected object could generate an error in the successive phases.
- Now I will explain you how to perform the task. Prior to each trial, you will look at the center of the screen of the computer monitor. You may move your eyes during the presentation of the clip. You will be presented with one video clip on each trial: the segmented video under test. Each trial will last 5 seconds. It is important that the distance from the monitor to your eyes is preserved during the presentation, so try not to lean forward or backward. The indicator to your right shows the distance at which we would like you to have your eyes.
- At the end of the presentation, a question will appear on the computer monitor. The same question will be asked after every trial. Do not spend a lot of time thinking about your responses: we want to know your initial impressions. You will be asked to give a global judgment about the observed video by entering a number that is proportional to the annoyance degree perceived as a result of the imperfections in the segmentation.
- Here is how you will determine the annoyance value. I am about to show you a set of sample clips. This will give you an idea of the range of image quality that you will be seeing. Among these videos you should identify the worst segmentation and mentally assign an annoyance value of 100 to it. Then, in the experiment, you will use that video as a term of comparison to judge the other ones: for instance if the annoyance value of a defect in the segmentation is half of the worst sample clip, give it a 50; if it is 1/10th as bad, give it a 10; if it is 1.5 times as bad, give it a 150. If the defect did not annoy you at all, call it 0 (zero). I will show you the sample clips now.

- [Click on <**START**>]
- Did you see any defects? Do you want to see the segmentations again? Remember that the most annoying defect that you have seen is to have a value of 100.
- Before we start the experiment, you will have nine practice trials to be sure that you understand the task. You will respond in these trials just like you will in the main experiment. Remember that you have one question to answer: how annoying was the defects in the segmentations relative to the worst example? Use the numeric keypad to enter the annoyance value, after you finish entering the annoyance value, click on <OK > or hit <**ENTER**> on the numeric keypad and press <**NEXT**> when you are ready for the next video. We will not use the data from the practice trials, so don't be concerned if you make a mistake here.
- [Start the practice trials]
- Do you have any questions?
- You can take a break at any time by entering your answers for the most recent video, but waiting to hit <**NEXT**> until you are ready to go on. You should stop if you are confused about what to do, if you realize you have entered data incorrectly, or if you need a break. You cannot stop the video from playing or go back and fix the data from a previous clip after you hit <**NEXT**>. So if something goes wrong, for example if you enter 600 instead of 60 and hit <**NEXT**>, watch the video and then tell me what happened specifying the number of the video (it appears up in the window where you put your answer), I will go back and fix it later.
- This part of the experiment includes 30 videos and it takes approximately 5 minutes to complete: the same question will come up on the computer screen after the video is displayed. Remember that we want to know your initial impressions so you should answer soon enter a judgment about the entire clip, and not only about the last frames.
- Do you have any questions?
- [Start the experiment.]

B.6 Sample Instructions for Watermarked 3D Model Distortion Task

After getting the subject into position, centered in front of the screen and at the correct distance (about 0.4 cm), the following instructions are read:

- “This test concerns the evaluation of the *distortions or impariments* introduced by watermarking algorithms on the surfaces of *3D models*.

  What is a 3D model? A 3D model is a collection of data that represent a 3D shape. 3D models are used in particular in entertainment industries, for examples movies and video games use a lot of 3D models.

  What is watermarking? Digital watermarking is a technology used to embed information inside a digital media. Imagine you want to associate some information with a digital media, i.e. the name of the owner of an image to the image itself. A watermarking algorithm, specific for images, can process the image and embed this information inside the image data itself. So, this information can be eliminated only modifying the watermarked image. In order to embed the data watermarking algorithms modify some properties of the digital media producing always some distortions in the watermarked media. The purpose of the test is the evaluation of the distortions introduced by watermarking algorithms on 3D models. Usually these distortions are visible on the model surfaces. So, the test is very simple: you interact with some models and you have to indicate if you see or not a certain kind of distortions. Obviously, I will show to you these distortions so you can understand what you have to evaluate.

- During the test you interact with the models and you have to evaluate these impairments. In particular you will indicate whether you detect any distortion or impariment.

- For those 3D models you detect a distortion you will indicate *how much* you perceive such distortion by entering a number that is proportional to its distortion value.

- Additionally you will have to indicate the part of the models *where* the distortions are more evident (this task is present only for Experiment I).

- Here I will show you the models without any distortion. The test includes four models. You have to imagine these models like *statues*. In particular I will show to you a model called ”Venus” that represent the head of a statue of Venus, a mythological feline with wings called ”Feline”, a ”Horse” and, finally a ”Bunny” .

- [Show originals]

- Are you able to remember these models? Do you want to see it again?

- Before we start the experiment, you will see how the typical distortions introduced by watermarking process look like. In few words the roughness of the surface of the model is increased in some way. We have to recognize this roughness, so it is important that you remember, for each model, the roughness of its parts. Another important aspect that you have to consider during your evaluation is that the distortions introduced by the algorithm is uniform on the surface.

- [Show the watermarked models]

- Have you understood how these distortions look like? Do you have any questions?

- In this phase, you will learn how to interact with the 3D model. You interact with the models by using the mouse. To rotate the model push the left button and moves the mouse. When you want to stop to rotate the model release the left button. To zoom the model push the right button and move the mouse ahead or back to zoom in or to zoom out. When you want to stop to zoom the model release the mouse button.
- [Interaction trial - Try to rotate the model. Try to zoom it. You can move the model left/right/up and down with the arrow keys. Try.].
- (Only for Experiment I) Remember that you have to decide even *where* the distortions are more evident. This is very important, so take in account, while you are interacting with the model, that you have to decide how much you perceive the distortion and where these are more evident. To indicate the part of the model that presents the most evident distortions we have to use this viewfinder, this sight [activate selection mode]. You can move the models as usual. The rectangle can be resized by using the keys '**A**', '**D**', '**X**' and '**W**'. Pay attention. You have to move the model to select the part, not the selection rectangle. Press <**ENTER**>, on the keyboard, to confirm your selection. Now, try to indicate some parts on this model. [Interaction trial - selection]
- Have you understood? Do you have any questions?
- Now I will show to you some cases that help you to evaluate numerically *how much* you perceive the distortions. You have to choose the worst case and mentally assign a value of 10 to it. This will give you an idea of the distortions that you will be seeing. The distortions could be present or not on the model. So, you are assigning a score of 10 to the most evident distortions. If the perception of the distortions during the test is half of the worst examples you chose, give it 5; if it is $1/10^{\text{th}}$ as bad give it 1, if it is 1.5 times as bad, give it 15. This is important, you can give score higher than 10. Remember that the question is *how much you perceive, how much you notice such kind of distortions.* I will show these examples now.
- [Show worst cases]
- Before we start the experiment you will have six practice trials to be sure that you understand the task. You will respond in these trials just like you will in the main experiment. The questions appear on the screen. So, you have to provide three answers (two for the Experiment II) after the interaction with the model. The first question is: did you notice any distortion? You can answer <**YES**> or <**NO**> at this question. Then, in case of positive answer you have to give a score to indicate *how much* the distortion is evident. You use the numeric keypad to enter the perception value. And finally, the third question, (only for Experiment I) *where* you noticed the distortions. To answer to this question you have to select the part of the model with the most evident distortions.
- Do you have any question? Do you want to repeat it?
- [Practice Trials]
- 40 models will be shown to you during the test (48 in Experiment II). This takes about 20 minutes (24 minutes for Experiment II) plus the time you need to indicate where the distortions are more evident (only for Experiment I). So, the test will takes about 30 minutes.
- Before to start the test I would like to give you the following practical recommendations:

1. In case of input error, please tell me what you want to do and I correct your answer at the end of the test.
2. You can take a break at any time by entering your answers (score and part) for the most recent models, but waiting to hit <**ENTER**> until you are ready to go on.
3. Finally, at the end of the test I will ask to you few questions.

- Do you have any question before to start?
- [Start the experiment]
- [Interview]

1. What is your feeling with the models? I mean, have you experienced any problem to identify the distortions on a specific model and why?
2. How would you describe the distortions that you saw?
3. Have you comments or remarks about the tests?

C

3D Model Rendering

The term *rendering* indicates the set of methods and algorithms used to generate a two-dimensional image starting from a scene described by geometric primitives. The result of the rendering process, i.e. the final rendered image, depends on the rendering techniques used and on the visual properties of the rendered scene. Nowadays, several rendering techniques exist. One first important distinction subdivides these techniques into two categories: the *real-time rendering* (RTR) methods and the *off-line rendering* ones. For real-time rendering we intend all of those techniques capable of generating the images quickly enough to allow *interaction*. The sense of interactivity with the 3D scene is constrained to the rate at which the images are displayed, measured in *frames per seconds (fps)*. The typical frame rate of video games is around 50 fps.

In contrast to real-time rendering, off-line rendering techniques process the 3D scene without the aim of interaction. In this case, the accent is posed on the creation of realistic images. Some popular off-line rendering methods include *Ray Tracing*.

Ray Tracing is one of the most used rendering methods to produce photo-realistic images. In ray tracing, a ray of light is traced in a *backward* direction, i.e. the ray starts from the eye of the observer and is traced through a pixel in the image plane into the scene determining what it hits. The pixel is then set to the color values returned by the ray. This basic idea is repeated as much as necessary to sample the entire image plain and to produce the final image (see Figure C.1).

In its basic form, ray tracing can be described by the following algorithm:

- For each pixel of the images:
    1. Construct a ray from the viewpoint
    2. For each objects in the scene
        - 2.1. Find intersection with the ray
        - 2.2. Keep the closest intersection point
    3. Shade the point the ray hits

Ray tracing could be computationally very expensive. This basic ray tracing scheme could be modified to take into account sophisticated photorealistic visual effects, such as soft shadows or caustics. Such extensions will be discussed in Section C.5.

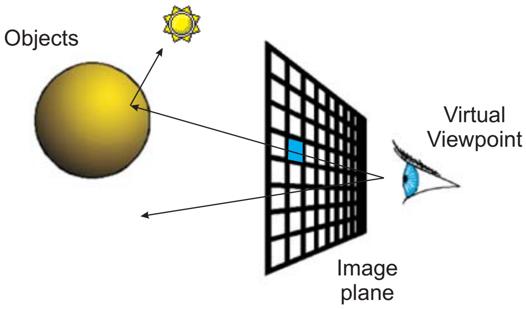


Figure C.1: Ray Tracing schematization.

C.1 Illumination Models

A *lighting model* uses a mathematical description of the interaction between the light incident on a 3D object and its surface. Such models range from simple to very complex ones. The complexity of the model varies depending on the desired level of simulation of physical interactions. Typically, real-time rendering systems require to find the right trade-off between level of realism, that requires a high computational effort, and an acceptable frame rate. On the contrary, off-line rendering systems may use sophisticated lighting models.

Before describing some classical lighting models we introduce some basic notions about light-material interaction. Figure C.2 summarizes what happens when photons hit a generic surface. Part of the incident light is reflected, part is absorbed and part is transmitted:

$$I_i = I_r + I_t + I_s + I_a \tag{C.1}$$

where I_i is the incident light, I_r is the light reflected by the material, I_t is the light transmitted, I_s is the light scattered and I_a is the light absorbed by the material. The *reflected light* can be described by considering two different effects, the *diffusion reflection* and the *specular reflection*. The diffusion reflection is responsible of the color of the objects. A yellow object, for example, when illuminated with a white light, reflects the yellow component of the light. The colored reflected light is due to the diffuse reflection. A perfect diffusive surface scatters light uniformly in all directions, hence the diffuse light does not depend on the observer's position.

In the **Lambertian** lighting model, the intensity of the diffuse light can be computed by *Lambert's Law* [171]:

$$I_d = I_i K_d \cos(\theta) \tag{C.2}$$

where K_d is the diffusion constant of the object, θ is the angle between the surface normals at the considered point and a line connecting such point with the light sources. In the following we indicate the direction of such line with the versor $\vec{L}$. The maximum light received by the surface is when the surface normal is parallel to the direction of the incident light, i.e. the surface is perpendicular to the incident light.

The specular reflection depends on the degree of glossiness of a surface. A matte surface has no specular effect but high diffusive behavior, a perfect glossy surface is a mirror. The light reflected from a mirror leaves the surface with the same angle of incidence, computed with respect to the surface normals. Hence, the amount of specular light seen by the viewer depends on the viewer's position. Simplifying the real phenomenon, we can say with a good degree of approximation that the color of the specular light is similar to the incident light, i.e. the highlight* color of an object illuminated with a white light is white.

---

*Highlight is the name of the area over which the specular reflection is seen.

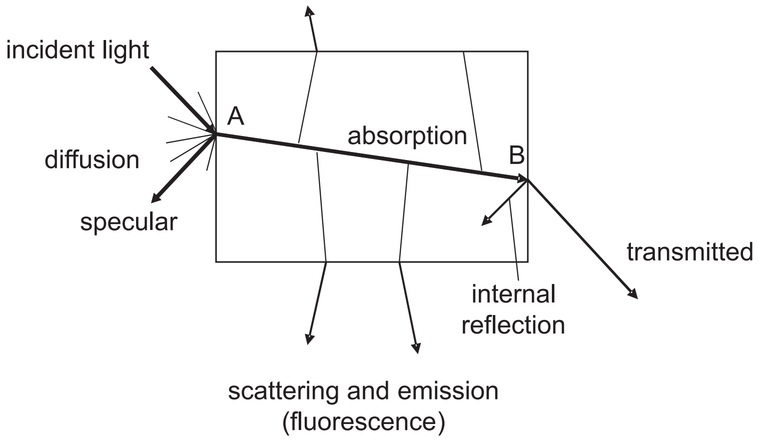
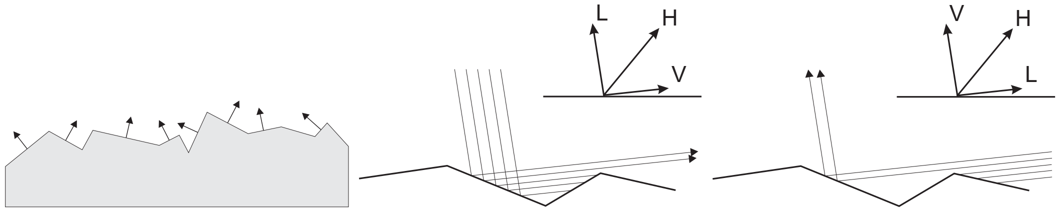


Figure C.2: Light-material interaction.



Microfacets model.

Masking effect of the microfacets.

Shadowing effects of the microfacets.

Figure C.3: Microfacets model of a reflecting surface.

The **Phong** lighting model [122] is the "standard" model used in Computer Graphics. This simple model has been developed by Bui-Tuong Phong in 1975 on the basis of empirical observations. This model provides a good trade-off between accuracy and complexity. Phong described the reflected light using three terms:

$$I_{\texttt{Phong}} = I_{\texttt{amb}} K_a + I_i K_d (\vec{N} \cdot \vec{L}) + I_i K_s (\vec{R} \cdot \vec{V})^n \qquad \text{(C.3)}$$

where K_a is the ambient constant and K_s the specular constant; $\vec{N}$ is the surface normal, $\vec{L}$ is the vector that represents the direction of the incident light, $\vec{L}$ is the vector that represents the direction of the reflected light, $\vec{V}$ is the vector that describes the direction of the line connecting the viewer with the considered point (see Figure C.4). The first term of the Eq.(C.3) is the *ambient term*; this term models the light which the object receives from the surrounding environment. The second term models the diffusion component of the reflected light. This term follows the Lambertian law (Eq. (C.2)). In fact it is proportional to the dot product between the (normalized) vectors $\vec{N}$ and $\vec{L}$. If $\vec{N} \cdot \vec{L} < 0$ the point receives no light. The third term models the specular light, in fact it is proportional to the light reflected in the direction of the viewer. The coefficient n depends on the surface's material.

The **Cook and Torrance** [21] lighting model includes energy conservation within the incident, the reflected light and the color change by the specular highlight. The first reflection model based on physical considerations to model the specular light was proposed by James Blinn [12]. Instead of providing an empirical formulation like Phong's model, Blinn's model was based on a surface model introduced by Torrance and Sparrow (1967) [149]. Such surface model assumes that a surface is composed of a collection of microfacets, each of them behaving like a mirror, as in Fig. C.3. The

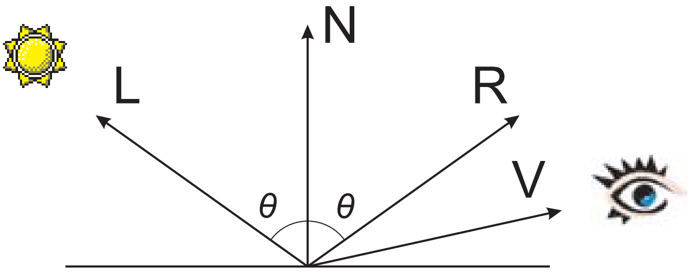


Figure C.4: Phong model schematization.

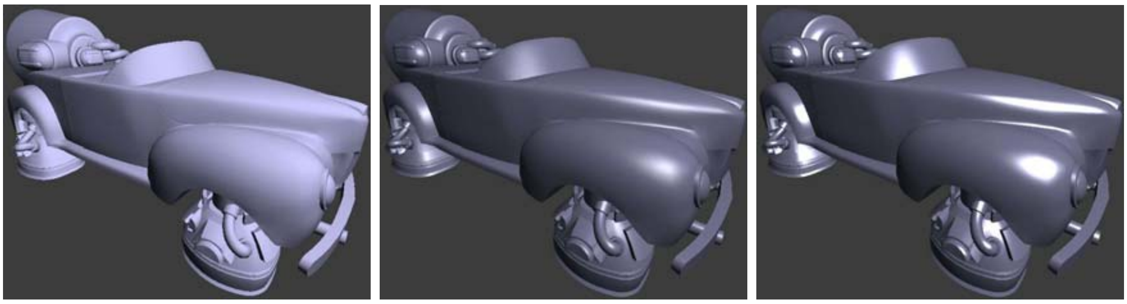

Lambertian model Phong model Cook and Torrance model

Figure C.5: Lighting models (from [83]).

distribution of the directions of the microfacets determines the specular component of the light. Cook and Torrance enhanced Blinn's model by introducing two new physical aspects in it: energy conservation between the incident and the reflected light and the change of the color within the specular highlight. A geometric attenuation factor in Cook and Torrace lighting model takes into account the masking and shadowing effect of the microfacets (Figure C.3).

Figure C.5 shows an example of application of the Lambertian, the Phong and the Cook and Torrance models. The Lambertian model is based on Lambert's Law ((C.2)) and takes into account only the diffuse component of the reflected light.

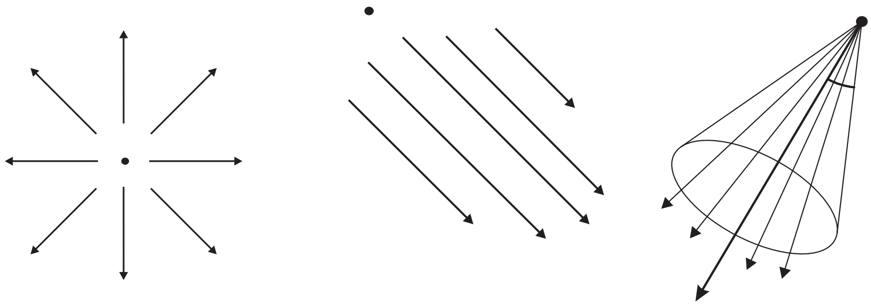
C.2 Light Sources

The first step to simulate lighting is to model the light sources. Typically, simple lighting models approximate the light sources as a point source since volumetric lights are complex to simulate. The three kinds of light sources commonly used are the directional light, the point light and the spot light. Figure C.6 summarizes these three kinds of light sources.

The *point light* models a light source as a point that emanates photons uniformly in all directions. This kind of light source is characterized completely by its position and by the color of the emitted light.

The *directional light* can be seen as a point light source moved at infinity. This light is completely defined by the light direction. Hence, the direction of the light received by the model is constant over all its surface. This kind of light can be used, for example, to model the light emitted from the sun in outdoor scenes.

The *spot light* simulates a cone of light. This kind of light requires several parameters to be defined: the position, a vector indicating the direction of the light and a cut-off angle, which is the half of the angle of the spotlight cone. Finally, to control the attenuation of the light within the



(Point light) (Directional light) (Spot light)

Figure C.6: Different kind of light sources.

cone, another parameter called *spot exponent* might be used. The spot exponent modulates the concentration of the light distribution in the central part of the cone.

Usually the color of the light emitted by the light source is defined by a set of three values representing the RGB components of the color. Another important property of the light emitted by a light source is its intensity. The intensity of the light decreases with the square of the distance from the light source. Often, in real-time rendering systems this physical property is not taken into account, and the intensity of the light received by the objects is assumed independent from the distance to the source.

C.3 Basic Shading Techniques

The term *shading* is the process of performing lighting computations, on the basis of the chosen lighting models, and determining the colors of the pixels. The three basic shading techniques are the flat, the Gouraud and the Phong* shading [171]. Referring to polygonal meshes, these techniques correspond to computing the light per-face, per-vertex and per-pixel, respectively.

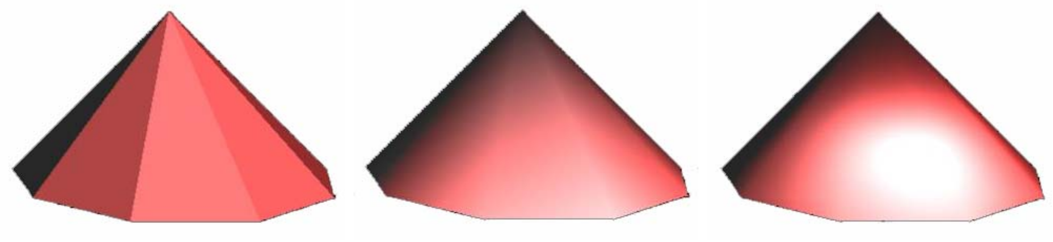
In *flat shading*, the light is computed for each triangle using the face normal. Hence, the effect of shading is highly dependent on the level of detail, i.e. the number of faces and of the objects that are rendered. It may be useful when the visualization purpose is to well-distinguish the faces that compose the model.

In the *Gouraud shading* the light is computed for each vertex using the vertex normals. The values of light over each face are interpolated. This method provides better visualization of the curved surface than the flat shading, i.e. the curve looks smoother and realistic. Some problems of the Gouraud shading include missing highlights and failure to capture spot light effects [53, 57].

In the *Phong shading*, the per-pixel lighting is computed by interpolating the vertex normals instead of the color of the vertices as in the Gouraud shading. The Phong shading is rarely used since it is computationally expensive and the Gouraud shading can provide the same visual results if the triangles of the model are smaller than a pixel.

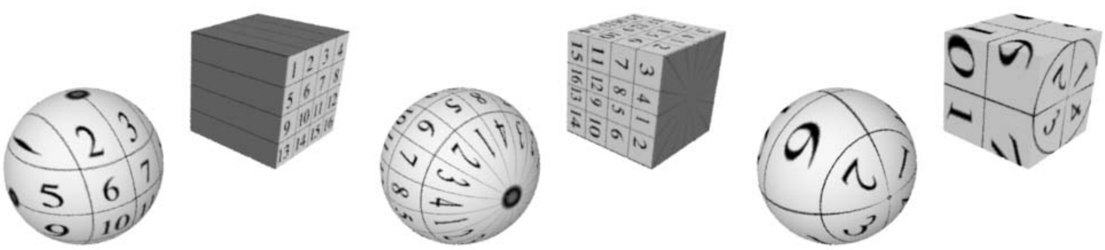
---

*The Phong shading must not be confused with the Phong lighting model. The latter is a reflection model while the former one is a way to interpolate the values of light obtained from a generic lighting model.



(Flat shading) (Gouraud shading) (Phong Shading)

Figure C.7: Basic shading techniques for polygonal meshes.



Planar Mapping Cylindrical Mapping Spherical Mapping

Figure C.8: Different texture projections.

C.4 Texturing

Texturing is a process of locally modifying the appearance of a surface by using images or repeating motifs. In other words, the surface properties are modulated by particular images called *textures.* For example, if we want to render a brick wall, the image of a brick wall can be spread over planar polygons giving the impression of the geometric details of the wall, even if they are not present in the geometric data of the wall. This kind of texturing is called *image texture mapping.* Another example could be the modulation of the surface transparency to simulate particular objects, such as clouds. Alternatively, the modulation of the surface's properties can be achieved by using bi- or three-dimensional functions instead of using textures. In the following we give a brief description of some texturing methods such as *image texture mapping*, *bump mapping* and *gloss mapping.*

Projector functions are used to convert a three-dimensional point in space into texture coordinates. The most commonly used projector functions include spherical, cylindrical and planar projections [10, 70]. In Figure C.8 several projection functions are shown.

In *image texturing*, one or more images are applied on the model's surface. The main problem is the association between the texture coordinates and the surface, in other words, to parameterize the surface [60, 78].

Bump mapping techniques use texture information to modulate the surface normals. The actual shape of the surface remains the same, but thanks to bump mapping the surface is rendered as if it were a different shape with more details [1, 13]. One way to represent bumps is to use a *heightfield* to modify the surface normal's direction (see Figure C.9). Each monochrome texture value represents a height.

Gloss mapping of texture is used to simulate non-uniformity in shiny surfaces. More specifically, a *gloss map* is a texture that modulates the contribution of the specular component over the surface. The idea at the base of gloss mapping is that the material properties can be encoded with textures, instead of using per-vertex values. Figure C.10 shows an example of gloss mapping.

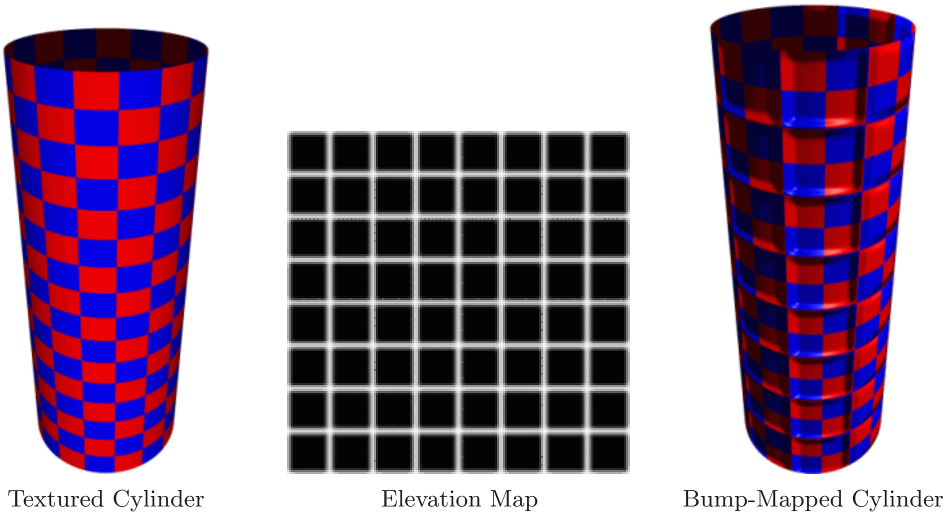

Textured Cylinder
Elevation Map
Bump-Mapped Cylinder

Figure C.9: An example of Bump Mapping.

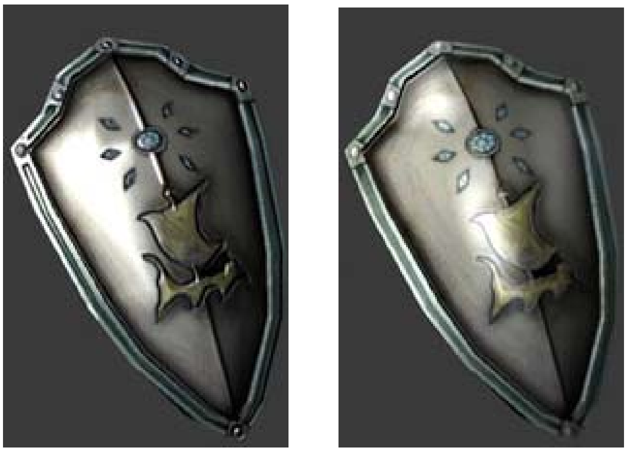
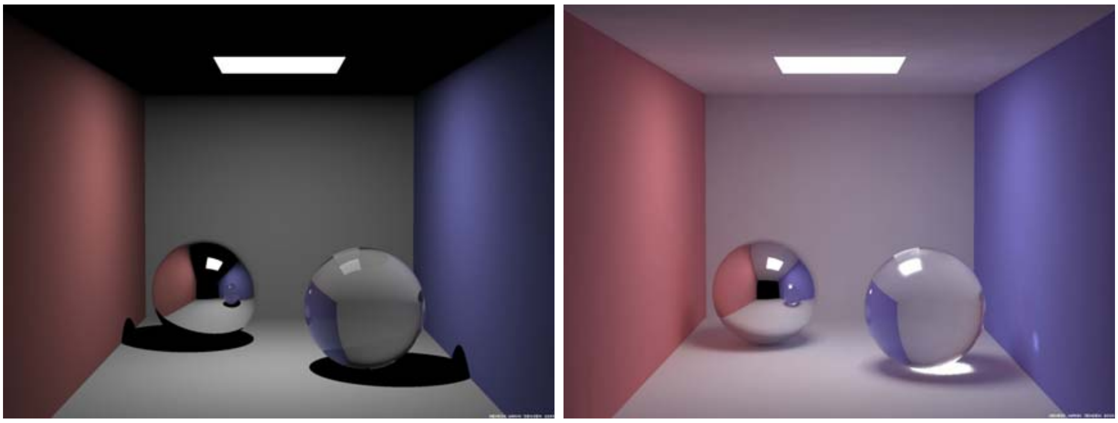


Figure C.10: Gloss mapping. The shininess of the shield on the right is modulated in order to appear more realistic.



Ray Tracing Photon Mapping

Figure C.11: Global Illumination Effects. (Left) Ray Tracing can be used to reproduce "hard" shadows and reflections. (Right) Photon Mapping can simulate any kind of visual effects produced by illumination in real scenes.

C.5 Global Illumination

The concept of *global illumination* is at the base of the generation of synthetic images with an outstanding level of realism. The illumination models presented so far are not able to capture a lot of sophisticated visual effects that are present in real scenes. In fact, those models are *local*, i.e. they do not take into account the contributions of the indirect light, i.e. the light coming from the other objects in the scene. In this sense, Ray Tracing is a first step in the direction of global evaluation of the illumination. The basic Ray Tracing algorithm proposed above can be easily modified to properly render shadows and reflections. More complex algorithms such as Path Tracing or Photon Mapping are necessary to take into account many of the global illumination visual effects present in the real scene. These effects (see Figure C.11) are:

- *Indirect light.* The light received by an object is not only the light emitted from the light sources but also the light reflected (or diffuse) by other objects. In Figure C.11 we notice that the light reaches the roof of the box thanks to the contributions of the indirect light.
- *Soft shadows.* The contour of the shadows in real images is smooth due to the fact that real light sources are volumetric.
- *Color bleeding.* This particular effect of the indirect light regards the fact that the color of an object is influenced by the color of the near objects. Such visual effect depends on the light diffused by the near objects. In Figure C.11 it is possible to see that the the roof is red near the red wall.
- *Caustics.* The caustics are regions of the scene where the reflected light is concentrated. An example is the light concentrated around the base of a glass. Such effect requires global illumination techniques to be properly simulated.

In simple words, the goal of global illumination rendering techniques is to trace all photons through a scene in order to simulate all the global illumination visual effects that characterize a real scene.

The rendering conditions presented in this Appendix are schematized in Fig. C.12.

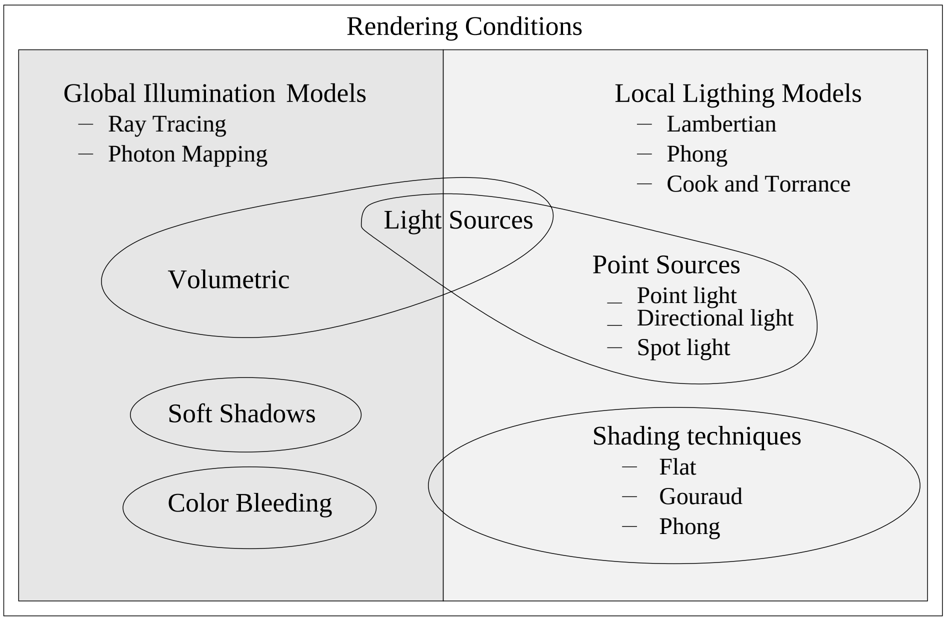


Figure C.12: Rendering conditions schematization.

Bibliography

[1] T. Akenine-Möller, E. Haines (2002). *Real-Time Rendering (second edition).* AK Peters.

[2] F. Alonso, M. Algorri, F. Flores-Mangas (2004). Composite index for the quantitative evaluation of image segmentation results. In *26th Annual International Conference of the IEEE EMBS*, pp. 1794–1797.

[3] artlive (1999). Architecture and authoring tools prototype for living images and new video experiments. `http://www.tele.ucl.ac.be/{PROJECTS}/art.live/`.

[4] N. Aspert, D. Santa-Cruz, T. Ebrahimi (2002). Mesh: Measuring error between surfaces using the hausdorff distance. In *Proceedings of the IEEE International Conference on Multimedia and Expo 2002 (ICME)*, vol. I, pp. 705–708.

[5] M. Bach et al. (2005). Comparison and validation of tissue modelization and statistical classification methods in t1-weighted mr brain images. *IEEE Transactions on Medical Imaging* p. in press.

[6] P. G. Barten (1990). Evaluation of subjective image quality with the square-root integral method. *Journal of Optical Society of America* **7**(10):2024–2031.

[7] F. Bartolini, M. Barni, V. Cappellini, A. Piva (1998). Mask building for perceptually hiding frequency embedded watermarks. In *Proceedings of the 5th IEEE International Conference on Image Processing, ICIP98*, vol. I, pp. 450–454, Chicago, IL, USA.

[8] O. Benedens (1999). Two high capacity methods for embedding public watermarks into 3D polygonal models. In *Proceedings of the Multimedia and Security-Workshop at ACM Multimedia 99*, pp. 95–99, Orlando, Florida.

[9] O. Benedens (1999). Watermarking of 3D polygon based models with robustness against mesh simplification. In *Proceedings of SPIE: Security and Watermarking of Multimedia Contents*, vol. 3657, pp. 329–340.

[10] E. A. Bier, K. R. Sloan (1986). Two-part texture mapping. *IEEE Computer Graphics and Applications* **6**(9):40–53.

[11] J. Black, T. Ellis, P. Rosin (2003). A novel method for video tracking performance evaluation. In *The Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 125–132.

[12] J. F. Blinn (1977). Models of light reflection for computer synthesized pictures. In *Proceedings of the 4th annual conference on Computer graphics and interactive techniques*, pp. 192–198, ACM Press, San Jose, California.

[13] J. F. Blinn (1978). Simulation of wrinkled surfaces. In *SIGGRAPH '78: Proceedings of the 5th annual conference on Computer graphics and interactive techniques*, pp. 286–292, ACM Press.

[14] M. R. Bolin, G. W. Meyer (1998). A perceptually based adaptive sampling algorithm. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pp. 299–309, ACM Press.

[15] M. Borsotti, P. Campadelli, R. Schettini (1998). Quantitative evaluation of color image segmentation results. *Pattern Recognition Letters* **19**:741–747.

[16] M. Carli et al. (2005). Quality assessment using data hiding on perceptually important areas. In *International Conference on Image Processing, ICIP'05*, IEEE, IEEE.

[17] A. Cavallaro, E. Drelie, T. Ebrahimi (2002). Objective evaluation of segmentation quality using spatio-temporal context. In *Proc. IEEE International Conference on Image Processing, Rochester(NY),22-25 September 2002*, pp. 301–304.

[18] A. Cavallaro, F. Ziliani (2000). *Image Analysis for Advanced Video Surveillance*. Multimedia Video-Based Surveillance Systems, G.L.Foresti, P.Mahonen, C.S.Regazzoni (Eds.), Kluwer Academic Publisher, Boston, chapter 2.3.

[19] F. Cayre, B.Macq (2003). Data hiding on 3-D triangle meshes. *IEEE Signal Processing* **51**(4):939–949.

[20] F. Cayre et al. (2003). Application of spectral decomposition to compression and watermarking of 3D triangle mesh geometry. *Image Communications - Special issue on Image Security* **18**:309–319.

[21] R. L. Cook, K. E. Torrance (1982). A reflectance model for computer graphics. *ACM Transactions on Graphics* **1**(1):7–24.

[22] P. Correia, F. Pereira (2000). Estimation of video object's relevance. In IEEE (ed.), *Proc European Signal Processing Conf. - EUSIPCO*, pp. 925–928.

[23] P. Correia, F. Pereira (2003). Objective evaluation of video segmentation quality. *IEEE Transaction on Image Processing* **12**:186–200.

[24] P. L. Correia, F. Pereira (2004). Classification of video segmentation application scenarios. *IEEE Trans. on Circuits and Systems for Video Tech.* **14**:735–741.

[25] M. Corsini, E. Drelie, T. Ebrahimi (2004). *Watermarked 3D Object Quality Assessment*. Technical Report ITS-2004.029., EPFL, Lausanne, Switzerland.

[26] I. Cox, M. L. Miller (1997). A review of watermarking and the importance of perceptual modeling. In *Proceedings of SPIE: Vol. 3016. Human Vision and Electronic Imaging II*, pp. 92–99.

[27] I. J. Cox, J. Kilian, F. T. Leighton, T. Shamoon (1997). Secure spread spectrum watermarking for multimedia. *International Conference on Image Processing, ICIP* **6**(12):1673–1687.

[28] R. Cucchiara, C. Grana, M. Piccardi, A. Prati (2003). Detecting moving objects, ghosts and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **25**(10):1337–1342.

[29] S. Daly (1993). The visible difference predictor; an algorithm for the assessment of image fidelity. In A. B. Watson (ed.), *Digital Images and Human Vision*, MIT Press, Cambridge, MA.

[30] H. de Ridder (1992). Minkowski-metrics as a combination rule for digital-image-coding impairments. In *Human Vision, Visual Processing, and Digital Display III; Bernice E. Rogowitz; Ed.*, vol. 1666, pp. 16–26.

[31] J. F. Delaigle, C. D. Vleeschouwer, B. Macq (1998). Watermarking algorithm based on a human visual model. *Signal Processing* **66**(3):319–336.

[32] D. Demigny, T. Kamle (1997). A discrete expression of canny's criteria for step edge detector performances evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**(11):1199–1211.

[33] M. Desbrun, M. Meyer, P. Schröder, A. H. Barr (1999). Implicit fairing of irregular meshes using diffusion and curvature flow. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 317–324, ACM Press/Addison-Wesley Publishing Co.

[34] A. Dobson (1990). *An Introduction to Generalized Linear Models.* Champan and Hall CRC.

[35] E. Drelie, E. Salvador, T. Ebrahimi (2003). Intuitive strategy for parameter setting in video segmentation. In *Visual Communications and Image Processing 2003*, vol. 5150 of *Proc. of SPIE*, pp. 998–1008, SPIE, SPIE.

[36] E. Drelie, D. Tomasic, T. Ebrahimi (2005). Which colors best catch your eyes: a subjective study of color saliency. In *First International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, Arizona, USA.*

[37] E. Drelie et al. (2004). Annoyance of spatio-temporal artifacts in segmentation quality assessment. In *International Conference on Image Processing*, IEEE, IEEE.

[38] E. Drelie et al. (2004). Towards perceptually driven segmentation evaluation metrics. In *CVPR 2004 Workshop (Perceptual Organization in Computer Vision)*, IEEE, IEEE.

[39] R. Dugad, K. Ratakonda, N. Ahuja (1998). A new wavelet-based scheme for watermarking images. In *International Conference on Image Processing, ICIP*, vol. 2, pp. 419–423, Chicago, IL.

[40] M. P. Eckert, A. P. Bradley (1998). Perceptual quality metrics applied to still image compression. *Signal Processing* **70**.

[41] P. G. Engeldrum (1999). Image quality modeling: Where are we? In *IS&T's 1999 PICS Conference*, pp. 251–255.

[42] P. G. Engeldrum (2000). *Psychometric Scaling: A Tool for Imaging Systems Development.* Imcotek Press, Winchester, MA.

[43] C. Erdem, B. Sankur (2000). Performance evaluation metrics for object-based video segmentation. In *Proc. X European Signal Processing Conference, Tampere, Finland*, vol. 2, pp. 917–920.

[44] C. Erdem, B. Sankur, A. M. Tekalp (2004). Performance measures for video object segmentation and tracking. *IEEE Transactions on Image Processing* **13**(7):937–951.

[45] A. M. Eskicioglu (2000). Toward a perceptual video quality metric. In *Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 6 of *Proc. of IEEE*, pp. 1907–1910, IEEE, IEEE.

[46] M. Everingham, H. Müller, B. Thomas (2001). Evaluating image segmentation algorithms using monotonic hulls in fitness/cost space. In T. Cootes, C. Taylor (eds.), *Proceedings of the 12th British Machine Vision Conference (BMVC2001)*, pp. 363–372, BMVA.

[47] M. D. Fairchild (1997). *Color Appearance Models.* Addison-Wesley, Boston.

[48] M. Farias (2004). *No-Reference and Reduced Reference Video Quality Metrics: New Contributions.* Ph.D. thesis, University of Santa Barbara, California.

[49] N. L. Fernandez-Garcia et al. (2004). Characterization of empirical discrepancy evaluation measures. *Pattern Recogn. Lett.* **25**(1):35–47.

[50] J. A. Ferwerda, P. Shirley, S. N. Pattanaik, D. P. Greenberg (1997). A model of visual masking for computer graphics. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pp. 143–152, ACM Press/Addison-Wesley Publishing Co.

[51] C. for AM Comparisons (1996). Compare your segmentation algorithm to the cost 211 quat analysis model. `http://www.iva.cs.tut.fi/COST211/Call/Call.htm`.

[52] A. R. J. François, G. G. Medioni (1999). Adaptive color background modeling for real-time segmentation of video streams. In *International on Imaging Science, System, and Technology*, pp. 227–232.

[53] A. S. Glassner (1995). *Principles of Digital Image Synthesis*, vol. 2. Morgan Kauffman Publisher Inc.

[54] E. B. Goldstein (1996). *Sensation and Perception.* Internazional Thomson Publishing Company, 4th Edition.

[55] A. B. Goumeidane et al. (2003). New discrepancy measures for segmentation evaluation. In *IEEE International Conference on Image Processing (ICIP'03) , Barcelona, Spain*, vol. 2, pp. 411–414.

[56] V. Q. E. Group (2000). Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000. `http://www.vqeg.org`.

[57] R. Hall (1989). *Illumination and Color in Computer Generated Imagery.* Springer-Verlag.

[58] R. Hamberg, H. de Ridder (1999). Time-varying image quality: Modeling the relation between instantaneous and overall quality. *SMPTE Journal* **108**:802–811.

[59] T. Harte, A. Bors (2002). Watermarking 3D models. In *Proceedings of IEEE International Conference on Image Processing 2002*, vol. III, pp. 661–664, Rochester, NY, USA.

[60] P. Heckbert (1989). *Fundamentals of Texture Mapping and Image Warping.* Master's thesis, University of California, Berkeley.

[61] W. Hopinks (2001). A new view of statistics. `http://www.sportsci.org/resource/stats/modelsdetail.html`.

[62] T. Horprasert, D. Harwood, L. S. Davis (1999). A statistical approach for real-time robust background substraction and shadow detection. In *IEEE ICCV Frame Rate Workshop*.

[63] Y. Inazumi, Y. Horita, K. Kotani, T. Murai (1999). Quality evaluation method considering time transition of coded quality. In *International Conference on Image Processing*, vol. 4, pp. 338–342.

[64] ITU (1995). *Studio Encoding Parameters of Digital Television for Standard 4:3 and Widescreen 16:9 Aspect Ratio*. ITU-R Recommendation BT.601, Geneva.

[65] ITU (1996). *Subjective Video Quality Assessment Methods for Multimedia Applications Recommendation P.910*. International Telecommunication Union, Geneva, Switzerland.

[66] ITU (2002). *Methodology for Subjective Assessment of the Quality of Television Pictures Recommendation BT.500-11*. International Telecommunication Union, Geneva, Switzerland.

[67] S. Jabri, Z. Duric, H. Wechsler, A. Rosenfeld (2000). Detection and location of people in video images using adaptive fusion of color and edge information. In *International Conference on Pattern Recognition (ICPR)*, pp. 627–630.

[68] T. R. Jones, F. Durand, M. Desbrun (2003). Non-iterative, feature-preserving mesh smoothing. *ACM Transactions on Graphics* **22**(3):943–949.

[69] S. Kanai, H. Date, T. Kishinami (1998). Digital watermarking for 3D polygons using multiresolution wavelet decomposition. In *Proceedings of the Sixth IFIP WG 5.2 International Workshop on Geometric Modeling: Fundamentals and Applications (GEO-6)*, pp. 296–307, Tokyo, Japan.

[70] K. Kershaw (1992). *A Generalized Texture-Mapping Pipeline*. Master's thesis, Cornell University, New York.

[71] C. Kim, J. Hwang (2002). Fast and automatic video object segmentation and tracking for content-based applications. *IEEE Transactions on Circuits and Systems for Video Technology* **12**(2).

[72] D. King, J. Rossignac (1999). Optimal bit allocation in compressed 3D models. *Comput. Geom. Theory Appl.* **14**(1-3):91–118.

[73] L. Kobbelt (1997). Discrete fairing. In *Proceedings of the Seventh IMA Conference on the Mathematics of Surfaces '97*, pp. 101–131.

[74] R. Koenen (1999). Mpeg-4: Multimedia for our time. *IEEE Spectrum* **36**:26–33.

[75] D. Kundur, D. Hatzinakos (1997). A robust digital watermarking method using wavelet-based fusion. In *Proceedings of IEEE International Conference on Image Processing '97*, vol. 1, pp. 544–547, Santa Barbara, CA.

[76] M. Kutter (1999). *Digital Image Watermarking: Hiding Information in Images*. Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, Switzerland.

[77] M. Kutter, S. Winkler (2002). A vision-based masking model for spread-spectrum image watermarking. *IEEE Transaction on Image Processing* **11**:16–25.

[78] W. Lance (1983). Pyramidal parametrics. *Computer Graphics* **7**(3):1–11.

[79] A. Lang, J. Dittmann, other contributors (1997). Audio Stirmark. `http://amsl-smb.cs.uni-magdeburg.de/smfa/main.php`.

[80] G. E. Legge, J. M. Foley (1980). Contrast masking in human vision. *Journal of Optical Society of America* **70**(12):1458–1471.

[81] E. L. Lehmann, H. J. D'Abrera (1998). *Nonparametrics: Statistical Methods Based on Ranks*. Prentice Hall, Englewood Cliffs, NJ.

[82] P. Lindstrom, G. Turk (2000). Image-driven simplification. *ACM Transaction on Graphics* **19**(3):204–241.

[83] F. Losasso (2005). Surface reflections model. `http://courses.dce.harvard.edu/~cscie236/papers/Surface%20Reflection%20Models.pdf`.

[84] J. Lubin (1993). The use of psychophysical data and models in the analysis of display system performance. In A. B. Watson (ed.), *Digital Images and Human Vision*, MIT Press, Cambridge, Massachusetts.

[85] J. Lubin (1995). A visual discrimination model for imaging system design and evaluation. In E. Peli (ed.), *Vision Models for Target Detection and Recognition*, pp. 245–283, World Scientific, New Jersey.

[86] X. Marichal, P. Villegas (2000). Objective evaluation of segmentation masks in video sequences. In *Proc. Of X European Signal Processing Conference, Tampere, Finland*, pp. 2139–2196.

[87] X. Marichal et al. (2002). The ART.LIVE architecture for mixed reality. In *Proc. of Virtual Reality International Conference (VRIC)*, pp. 19–21, Laval, France.

[88] J. B. Martens (2002). Multidimensional modeling of image quality. In *Proceedings of the IEEE*, vol. 90, pp. 133 – 153.

[89] J. B. Martens, L. Meesters (1998). Image dissimilarity. *Signal Processing* **70**(3):155–176.

[90] D. Martin, C. Fowlkes (2003). The berkeley segmentation dataset and benchmark. `http://www.cs.berkeley.edu/projects/vision/grouping/segbench/`.

[91] D. Martin, C. Fowlkes, D. Tal, J. Malik (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In IEEE (ed.), *Proceedings of the Eighth International Conference On Computer Vision (ICCV-01), July 7-14, 2001, Vancouver, British Columbia, Canada*, vol. 2, pp. 416–425.

[92] S. E. Maxwell, H. Delaney (1999). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

[93] McCane (1997). "on the evaluation of image segmentation algorithms". In *Proceedings of Digital Image Computing: Techniques and Applications*, pp. 455–460.

[94] S. J. McKenna et al. (2000). Tracking groups of people. *Computer Vision and Image Understanding* **80**:42–56.

[95] K. McKoen et al. (2000). Evaluation of segmentation methods for surveillance applications. In *EUSIPCO*, pp. 1045–1048.

[96] H. M.D., S. S., S. T., B. K.W. (1997). A robust visual method for assessing the relative performance of edge detection algorithms. *Transactions on Pattern Analyis and Machine Intelligence* **19**:1338–1359.

[97] R. Mech, F. Marques (2001). Objective evaluation criteria for 2d-shape estimation results of moving objects. In *Workshop on Image Analysis for Multimedia Interactive Services*, Tampere, Finland.

[98] Merriam-Webster (2005). Merriam-webster dictionary on-line. `http://www.m-w.com/`.

[99] V. Mezaris, I. Kompatsiaris, M. G. Strintzis (2003). Still image objective segmentation evaluation using ground truth. In J. P. B. Kovar, M. Vlcek (eds.), *5th COST 276 Workshop 2003*, pp. 9–14.

[100] P. Milgram, H. Colquhoun (1999). A taxonomy of real and virtual world display integration. In Y. Otha, H. Tamura (eds.), *Mixed Reality, Merging Real and Virtual Worlds*, Ohmsha-Springer.

[101] S. S. Monaci G, Menegaz G, K. K (2004). Chromatic contrast detection in spatial chromatic noise. *Visual Neuroscience* **21**(3):291–294.

[102] M. Moore (2002). *Psychophysical Measurement and Prediction of Digital Video Quality*. Ph.D. thesis, University of Santa Barbara, California.

[103] K. Myszkowski, P. Rokita, T. Tawara (1999). Perceptually-informed accelerated rendering of high quality walkthrough sequences. In *Proceedings of the Tenth Eurographics Workshop on Rendering*, pp. 5–18, Granada, Spain.

[104] J. Nascimento, J. S. Marques (2004). New performance evaluation metrics for object detection algorithms. In *6th International Workshop on Performance Evaluation for tracking and Surveillance (PETS, ECCV), Prague, May 2004*.

[105] M. R. M. Nijenhuis, F. J. J. Blommaert (1996). A perceptual error measure for sampled and interpolated complex colour images. *Displays* **17**:27–36.

[106] F. Oberti, E. Stringa, G. Vernazza (2001). Performance evaluation criterion for characterizing video surveillance systems. *Real Time Imaging* **7**:457–471.

[107] R. Ohbuchi, H. Masuda, M. Aono (1997). Watermaking three-dimensional polygonal models. In *Proceedings of the fifth ACM international conference on Multimedia*, pp. 261–272, ACM Press, Seattle, Washington, United States.

[108] R. Ohbuchi, A. Mukaiyama, S. Takahashi (2002). A frequency-domain approach to watermarking 3D shapes. In *Proceedings of EUROGRAPHICS 2002*, Saarbrucken, Germany.

[109] R. Ohbuchi, S. Takahashi, T. Miyazawa, A. Mukaiyama (2001). Watermarking 3D polygonal meshes in the mesh spectral domain. In *Proceedings of Graphics interface 2001*, pp. 9–17, Canadian Information Processing Society, Ottawa, Ontario, Canada.

[110] R. J. Oliveira, P. C. Ribeiro, J. S. Marques, J. M. Lemos (2004). A video system for urban surveillance: Function integration and evaluation. In *International Workshop on Image Analysis for Multimedia Interactive Systems, 2004*.

[111] Optimark (2001). Benchmark for digital watermarking techniques. `http://poseidon.csd.auth.gr/optimark/`.

[112] W. Osberger (1999). *Perceptual Vision Models for Picture Qaulity Assessment and Compression Applications*. Ph.D. thesis, Queensland, University of Technology, Brisbane, Australia.

[113] R. S. P. Cignoni, C. Rocchini (1998). Metro: measuring error on simplified surfaces. *Computer Graphics Forum* **17**(2):167–174.

[114] D. L. Page et al. (2003). Shape analysis algorithm based on information theory. In *Proceedings of the International Conference on Image Processing, ICIP2003*, vol. 1, pp. 229–232.

[115] N. R. Pal, S. P. Pal (1993). A review on image segmentation techniques. *Pattern Recognition* **26**:1277–1293.

[116] Y. Pan, I. Cheng, A. Basu (2003). Perceptual quality metric for qualitative 3D scene evaluation. In *Proceedings of International Conference on Image Processing 2003*, vol. 3, pp. 169–172, Barcelona, Spain.

[117] S. Pereira, other contributors (2001). Checkmark. `http://watermarking.unige.ch/checkmark/index.html`.

[118] S. Pereira et al. (2001). Second generation benchmarking and application oriented evaluation. In *Proceedings of Information Hiding Workshop*, Pittsburgh, PA, USA.

[119] F. Petitcolas, other contributors (1997). Stirmark benchmark 4.0. `http://www.petitcolas.net/fabien/watermarking/stirmark/`.

[120] F. A. P. Petitcolas (2000). Watermarking schemes evaluation. *IEEE Multimedia Signal Processing* **17**:58–64.

[121] F. A. P. Petitcolas, R. J. Anderson, M. G. Kuhn (1998). Attacks on copyright marking systems. In D. Aucsmith (ed.), *Proceedings of the 2nd International Workshop on Information Hiding, IH98*, pp. 219–239., Springer-Verlag, Portland, Oregon, USA.

[122] B. T. Phong (1975). Illumination for computer generated pictures. *Communications of the ACM* **18**(6):311–317.

[123] R. Piroddi, T. Vlachos (to appear 1st quarter 2006). A framework for single-stimulus quality assessment of segmented video. *EURASIP Journal on Applied Signal Processing* .

[124] C. I. Podilchuk, W. Zeng (1998). Image-adaptive watermarking using visual models. *IEEE Journal on Selected Areas in Communications* **16**:525–539.

[125] E. C. Project (2001). Caviar test case scenarios. `http://homepages.inf.ed.ac.uk/rbf/CAVIAR`.

[126] M. Project (1998). Digital michelangelo project. `http://graphics.stanford.edu/projects/mich/`.

[127] M. Ramasubramanian, S. N. Pattanaik, D. P. Greenberg (1999). A perceptually based physical error metric for realistic image synthesis. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 73–82, ACM Press/Addison-Wesley Publishing Co.

[128] M. Reddy (1997). *Perceptually Modulated Level of Detail for Virtual Environments.* Ph.D. thesis, University of Edinburgh.

[129] H. D. Ridder (1991). Subjective evaluation of scale-space image. *SPIE Human Vision, Visual Processing,. and Digital Display II* **1453**:31–42.

[130] B. Rogowitz, H. Rushmeier (2001). Are image quality metrics adequate to evaluate the quality of geometric objects? In B. E. Rogowitz, T. N. Pappas (eds.), *Human Vision and Electronic Imaging VI*, vol. 4299, pp. 340–348, SPIE Proc.

[131] C. Rosenberger, K. Chehdi (2000). Genetic fusion: application to multi-components image segmentation. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. IEEE International Conference on*, vol. 6, pp. 2223–2226.

[132] P. L. Rosin (1999). Unimodal thresholding. In *11th Scandinavian Conference on Image Analysis, Kangerlussuaq, Grennland*, pp. 585–592.

[133] R. G. V. Schyndel, A. Z. Tirkel, C. F. Osborne (1994). A digital watermark. In *Proceedings of IEEE International Conference on Image Processing '94*, vol. 2, pp. 86–90, Austin, Texas.

[134] J. W. Senders (1997). Distribution of visual attention in static and dynamic displays. In *SPIE, Human Vision and Electronic Imaging II*, vol. 3016, pp. 186–194.

[135] A. Senior et al. (2000). Appearance models for occlusion handling. In *in 2nd IEEE Workshop Performance Evaluation of Tracking and Surveillance.*

[136] R. Shacked, D. Lischinki (2001). Automatic lighting design using a perceptual quality metric. *Computer Graphics Forum* **20**(3).

[137] C. W. Shaffrey, I. H. Jermyn, N. G. Kingsbury (2002). Psychovisual evaluation of image segmentation algorithms. In *Proceedings of ACIVS 2002 (Advanced Concepts for Intelligent Vision Systems), Ghent, Belgium, September 9-11, 2002.*

[138] J. Shen (2004). Motion detection in color image sequence and shadow elimination. In *Visual Communication and Image Processing*, pp. 731–740.

[139] K. Shoemake (1994). Arcball rotation control. In P. Heckbert (ed.), *Graphics Gems IV*, pp. 175–192, Academic Press.

[140] T. Sikora (1997). The MPEG-4 video standard verification model. *IEEE Transactions on Circuits and Systems for Video Technology,* **7**(1):19–31.

[141] T. Sikora (2001). The MPEG-7 visual standard for content description – an overview. *IEEE Transactions on Circuits and Systems for Video Technology,* **11**(6):696–702.

[142] G. W. Snedecor, W. G. Cochran (1989). *Statistical Methods.* Iowa State University, Press, Ames.

[143] M. Sonka, V. Hlavic, R. Boyle (1999). *Image Processing, Analysis and Machine Vision.* An International Thomson Publishing Company, 2nd edn.

[144] A. T1.801.03 (1996). *Digital transport of one-way video signlas - parameters for objective performance assessment.* American National Standards Institute, New Work, NY.

[145] K. T. Tan, M. Ghanbari, D. E. Pearson (1998). An objective measurement tool for mpeg video quality. *Signal Processing* **70**(2):279–294.

[146] G. Taubin (1995). A signal processing approach to fair surface design. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pp. 351–358, ACM Press.

[147] P. C. Teo, D. J. Heeger (1994). Perceptual image distortion. In *Proceedings of the first IEEE International Conference on Image Processing*, vol. 2, pp. 982–986.

[148] A. H. Tewfik, M. Swanson (1997). Data hiding for multimedia personalization, interaction and protection. *IEEE Multimedia Signal Processing* **14**(4):41–44.

[149] K. Torrance, E. Sparrow (1967). Theory for off-specular reflection from rough surfaces. *Journal of the Optical Society of America* **57**(9):1105–1114.

[150] F. Uccheddu, M. Corsini, M. Barni (2004). Wavelet-based blind watermarking of 3D models. In *Proceedings of the 2004 multimedia and security workshop on Multimedia and security*, pp. 143–154, ACM Press, Magdeburg, Germany.

[151] F. Uccheddu, M. Corsini, M. Barni, V. Cappellini (2004). A roughness-based algorithm for perceptual watermarking of 3d meshes. In *Proceedings of the Tenth International Conference on Virtual System and Multimedia*, Ogaki City, Japan.

[152] C. J. van den Branden Lambrecht (1996). *Perceptual models and architectures for video coding applications*. Ph.D. thesis, EPFL, 1015 Ecublens.

[153] A. Vetro, T. Haga, K. Sumi, H. Sun (2003). Object-based coding for long term archive of surveillance video. In *International Conference on Multimedia & Expo (ICME)*, vol. 2, pp. 417–420.

[154] P. Villegas, X. Marichal (2004). Perceptually-weighted evaluation criteria for segmentation masks in video sequences. *IEEE Transactions on Image Processing* **13**(8):1092–1103.

[155] I. . W. Vision (2005). Performance evaluation of tracking and surveillance (pets). `http://pets2005.visualsurveillance.org/`.

[156] S. Voran, S. Wolf (1992). The development and evaluation of an objective quality assessment system that emulates human viewing panels. In *International Broadcasting Convention,Amsterdam, The Netherlands*, 358.

[157] VQEG (2000). Final report from the video quality experts group on the validation of objective models of video quality assessment. `http://www.vqeg.org`.

[158] N. G. W. Müller (1999). Objective quality estimation for digital images in multimedia environment. In L. W. MacDonald (ed.), *Colour Imaging: Vision and Technology*, John Wiley and Sons.

[159] M. G. Wagner (2000). Robust watermarking of polygonal meshes. In *Proceedings of Geometric Modeling and Processing 2000. Theory and Applications.*, pp. 201–208, Hong Kong, China.

[160] H.-J. M. Wang, P.-C. Su, C.-C. J. Kuo (1998). Wavelet-based digital image watermarking. *Opt. Express* **3**(12):491–496.

[161] Z. Wang (2003). *Objective Image/Video Quality Measurement - A Literature Survey.* Tech. Rep. EE381K, Department of Electrical and Computer Engineering, University of Texas at Austin, USA,Department of Computer Science 9, University of Erlangen.

[162] Z. Wang, A. Bovik, H. R. Sheikh, E. P. Simoncelli (2004). Image quality assessment: from error visibility to structural similarity. *Image Processing* **13**(4):600–612.

[163] Z. Wang, A. C. Bovik, L. Ligang (2002). Why is image quality assessment so difficult? In *Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4 of *Proc. of IEEE*, pp. 3313–3316, IEEE.

[164] Z. Wang, H. R. Sheikh, A. C. Bovik (2003). *Objective Video Quality Assessment.* The Handbook of Video Databases: Design and Applications Furht and O. Marqure, ed., CRC Press, pp. 1041-1078.

[165] S. K. Warfield, K. H. Zou, , W. M. Wells (2004). Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. *IEEE Transaction on Medical Imaging* **23**(7):903–921.

[166] A. B. Watson (1987). Efficiency of an image code based on human vision. *Journal of Optical Society of America* **4**(12):2401–2417.

[167] A. B. Watson (1993). DCT quantization matrices visually optimized for individual images. In *Proceedings of SPIE: Vol. 1913, Human Vision, Visual Processing and Digital Display IV*, pp. 202–216.

[168] A. B. Watson (1998). Toward a perceptual video quality metric. In *Human Vision and. Electronic Imaging, (San Jose, CA)*, vol. 3299 of *Proc. of SPIE*, pp. 139–147, SPIE, SPIE.

[169] A. B. Watson, J. Hu, J. F. McGowan, J. B. Mulligan (1999). Design and performance of a digital video quality metric. In *SPIE, Human Vision and Electronic Imaging*, vol. 3644, pp. 168–174.

[170] B. Watson, A. Friedman, A. McGaffey (2000). Using naming time to evaluate quality predictors for model simplification. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 113–120, ACM Press, The Hague, The Netherlands.

[171] A. Watt (1990). *Fundamentals of three-dimensional computer graphics.* Addison-Wesley Publishing Co.

[172] E. W. Weisstein (1999). Correlation coefficient. `http://mathworld.wolfram.com/CorrelationCoefficient.html`.

[173] N. Williams et al. (2003). Perceptually guided simplification of lit, textured meshes. In *Proceedings of the 2003 symposium on Interactive 3D graphics*, pp. 113–121, ACM Press, Monterey, California.

[174] S. Winkler (1999). Issues in vision modeling for perceptual video quality assessment. *Signal Processing* **78**(2):231–252.

[175] S. Winkler (2000). *Vision models and quality metrics for image processing applications.* Ph.D. thesis, EPFL, 1015 Ecublens.

[176] S. Winkler, R. Campos (2003). Video quality evaluation for internet streaming applications. In *SPIE, Human Vision and Electronic Imaging*, vol. 5007, pp. 104–115.

[177] S. Winkler, E. Drelie, T. Ebrahimi (2003). Toward perceptual metrics for video watermark evaluation. In *Applications of Digital Image Processing*, vol. 5203 of *Proc. of SPIE*, pp. 371–378, SPIE, SPIE.

[178] R. Wolfgang, C. I. Podilchuk, E. J. Delp (1999). Perceptual watermarks for digital images and video. *IEEE Proceedings* **87**(7):1108–1126.

[179] R. B. Wolfgang, C. I. Podilchuk, E. J. D. III (1999). Perceptual watermarks for digital images and video. In P. W. Wong, E. J. D. III (eds.), *Security and Watermarking of Multimedia Contents*, vol. 3657, pp. 40–51, SPIE, San Jose, CA, USA.

[180] M. Wollborn, R. Mech (1998). Refined procedure for objective evaluation of video object generation algorithms. In *ISO/IECJTCI/SC29/WG11 M3448*, 43rd MPEG Meeting, Tokyo, Japan 1998.

[181] J.-H. Wu, S.-M. Hu, J.-G. Sun, C.-L. Tai (2001). An effective feature-preserving mesh simplification scheme based on face constriction. In *Proceedings of the 9th Pacific Conference on Computer Graphics and Applications*, p. 12, IEEE Computer Society.

[182] X.-G. Xia, C. G. Boncelet, G. R. Arce (1998). Wavelet transform based watermark for digital images. *Opt. Express* **3**(12):497–511.

[183] L. Yang, F. Albregten, T. Lønnestad, P. Gtøttum (1995). A supervised approach to the evaluation of image segmentation methods. In *Proc. Int. Conf. on Computer Analysis of Images and Patterns, Prague, Czech Repub- lic, September 1995.*

[184] Y. Yitzhaky, E. Peli (2003). A method for objective edge detection evaluation and detector parameter selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(8):1027–1033.

[185] L. Younes (1998). Computable elastic distances between shapes. *SIAM Journal of Applied Mathematics* **58**(2):565–586.

[186] Y. J. Zhang (1996). A survey on evaluation methods for image segmentation. *Pattern Recognition* **29**:1335–1346.

[187] Y. J. Zhang (1997). Evaluation and comparison of different segmentation algorithms. *Pattern Recognition Letters* **18**:963–974.

[188] Y. J. Zhang (2001). A review of recent evaluation methods for image segmentation. In *International Symposium on Signal Processing and its Applications (ISSPA), Kuala Lumpur Malaysia, 13-16 August*, vol. 1, pp. 148 – 151.

[189] Y. J. Zhang, J. J. Gerbrands (1994). Objective and quantitative segmentation evaluation and comparison. *Signal Processing* **39**:43–54.

Index

accuracy, 17
annoyance, 9
  task, 44

benchmarking, 15
boundary effects, 13

correlation
  fit goodness, 14
  Pearson, 17
  Spearman, 18

defect, 9
digital watermarking, 19
  3D, 115
  benchmarks, 21
  capacity, 21
  robustness, 21
  video, 20
  visibility, 20
distortion, 9

edge segmentation
  objective, 58
  subjective, 36
expectation effect, 52

F test, 72
flickering
  segmentation, 52
  watermarking, 23

ground-truth, 11

image segmentation
  objective, 58
  subjective, 36
impairment, 8

Mean Opinion Score, MOS, 12
memory, 70
method, 57
  analytical, 57
  discrepancy, 58
  empirical, 58
  goodness, 58
metric, 8
  flicker, video watermarking, 26
  Hausdorff distance based, 117
  mathematical, 16
  maximum geometric, 118
  mean geometric, 118
  Minkowski, 14
  multi scale, 3D, 137
  noise, video watermarking, 26
  objective, 8
  perceptual, 16
    3D watermarking, 139
    object segmentation, 85
  smoothing based, 3D, 138
model, 16
  mechanistic, 16
  performance, 16
monitoring, 15
monotonicity, 17

N-ring, 137
noise
  segmentation, 35
  watermarking, 23
normalization, 13, 130

optimization, 15

psychometric
  definition, 7
  function, 13
    Gaussian, 13
    logistic, 13
    Weibull, 13
psychophysics, 7

quality, 7
    visual, 7

reference, 11
    full, 16
    no, 16
    reduced, 16
roughness, 135

scales, 10
    categorical, 10
    continuous, 10
    discrete, 10
    numerical, 10
segmentation
    added background, 48
    added region, 48
    application, 103
    augmented reality, 106
    border hole, 49
    compression, 105
    false negatives, 48
    false positive, 47
    general, 94
    image, 33
    inside hole, 49
    spatial errors, 46
    static background, 94
    surveillance, 105
    temporal errors, 52
    video, 34
strength, 14
    task, 45

test scenes, 11
test session, 11
tracking, 59

video segmentation
    objective, 59
    subjective, 36
viewing conditions, 11

Curriculum Vitae

Personal information

Name: Elisa Drelie Gelasca
Nationality: Italian
Date of birth: February 26, 1975
Place of birth: Trieste, Italy
Marital status: Single

Address: EPFL/STI/ITS/LTS1 ELE 232
(Batiment ELD) Station 11
1015 Lausanne, Switzerland
Phone: +41 693 26 05
Fax: +41 21 693 76 00
Email: Elisa.Drelie@epfl.ch

Objectives

My main interests lay in the study of multimedia quality assessment, color saliency, video object segmentation, watermarking.

Education

- ***December 2001 – November 2005***: Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland (EPFL)
  Ph.D student in Electrical and Eletronic Engineering.

- ***June 2001***: University of Trieste, Trieste, Italy
  Laurea (M. Sc.) in Electronic Engineering.

- ***October 2000 – April 2001***: Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland (EPFL)
  Master Thesis as *Erasmus student* at the Signal Processing Institute.

Work Experience

- ***December 2001 – present***: Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
  Research assistant:
  - *PhD Thesis:* on-going research in the field of image and video processing with emphasis on image, video and 3D object quality assessment.
  - *Teaching*: definition and supervision of semester and diploma students projects, responsible for exercises and laboratory activities for the Image and Video Processing course.
  - *Responsible* from February 2004 to present, for EPFL's contribution on segmentation system to the European Network of Excellence VISNET (Networked Audiovisual Media Technologies).
  - *Responsible* from December 2001 to September 2002, for EPFL's contribution to the European IST project *Certimark* on watermarking for protection of still pictures and low bit rate video.
    - ∗ contribution to the production of project deliverables and project documentation;
    - ∗ participation to project meetings and public project demonstrations.
  - *Organization*: in May 2003, during EPFL's Open Days, responsible with two colleagues for organizing and running the installation of an interactive immersing gaming demonstrator developed by the art.live project. The installation is currently running at the Audiorama Museum in Montreux, Switzerland.
  - *Scientific Board member* in International Conference on Visual Communications and Image Processing 2003 (VCIP), Lugano, Switzerland
  - *Seminar* August 2003, University of California, Santa Barbara, USA. Seminar's title: Objective and Subjective Evaluation of Video Object Segmentation for Multimedia Environments.
- ***July – October 2003***
  Internship: 4-month internship at the Signal Processing Laboratory of Prof. Mitra at the University of California, Santa Barbara, with special focus on subjective experiments for media quality assessment.
- ***September – November 2001***
  Consultant for Genista SA developing a metric included in product for sale (*Video* $PQoS^{TM}$) to estimate the perceptual quality in watermarked videos.
- ***October 1999 – July 2000***
  Tutor of Informatics Engineering at the University of Trieste, Italy (with special focus on Quality Assurance).

Skills

Languages

Italian:	mother tongue
English:	fluent
French:	fluent

Computer literacy

Operating systems:	LINUX, Unix, Windows
Programming languages:	Pascal, C, C++, Visual Basic, Java
Software:	Matlab, familiarity with Spice and L-Edit, LaTeX, MS Office

Personal interests

- *Basketball* and *Water–polo*: 3rd Italian league ex-player and 2nd Swiss league player.
- *Reading, Photography and Traveling.*

Publications

1. E. Drelie Gelasca, T. Ebrahimi “On Evaluating Metrics for Video Segmentation Algorithms.” accepted to *Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, USA, January 22- 24, 2006
2. E. Drelie Gelasca, M. Corsini, T. Ebrahimi, and M. Barni. “Watermarked 3D Mesh Quality Assessment” submitted to *IEEE Transactions on Multimedia*, December 2005.
3. E. Drelie Gelasca, T. Ebrahimi, M. Corsini and M. Barni. “Objective Evaluation of the Perceptual Quality of 3D Watermarking”. *IEEE International Conference on Image Processing (ICIP)*, Vol. I, pp. 241-244, Genoa, Italy, September 2005.
4. M. Carli, M. Farias, E. Drelie Gelasca, R. Tedesco and A. Neri. “Quality Assessment Using Data Hiding on Perceptually Important Areas”. *IEEE International Conference on Image Processing (ICIP)*, Vol. III, pp. 1200-1203, Genoa, Italy , September 2005.
5. M. Corsini, E. Drelie Gelasca and T. Ebrahimi. “A Multi-Scale Roughness Metric for 3D Watermarking Quality Assessment”. *Workshop on Image Analysis for Multimedia Interactive Services 2005*, April 13-15, Montreux, Switzerland., April 2005.
6. E. Drelie Gelasca, D. Tomasic and T. Ebrahimi. “Which Colors Best Catch Your Eyes: a Subjective Study of Color Saliency”. *Fisrt International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, USA, January 2005.
7. A. Neri, P. Campisi, M. Carli and E. Drelie Gelasca. “Watermarking Hiding in Video Sequences”. *First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, USA, January 2005
8. E. Drelie Gelasca, T. Ebrahimi, M. Farias, M. Carli and S. Mitra. “ Annoyance of Spatio-Temporal Artifacts in Segmentation Quality Assessment”. In Proc. of *International Conference on Image Processing*, Page(s):345 -348 Vol. 1, October 2004.
9. E. Drelie Gelasca, T. Ebrahimi, M. Farias, M. Carli and S. Mitra. “Towards Perceptually Driven Segmentation Evaluation Metrics”. *CVPR 2004 Workshop (Perceptual Organization in Computer Vision)*, Page(s):52 - 57, June 2004.
10. Elisa Drelie Gelasca, Touradj Ebrahimi, Mylene Farias and Sanjt Mitra. “Impact of Topology Changes in Video Segmentation Evaluation”. In Proc. of *Workshop on Image Analysis for Multimedia Interactive Services*, April 2004.

11. Stefan Winkler, Elisa Drelie Gelasca and Touradj Ebrahimi. “Toward perceptual metrics for video watermark evaluation”. Proc. of *SPIE, Applications of Digital Image Processing*, Vol. 5203, pp. 371-378, August 2003.

12. E. Drelie Gelasca, E. Salvador and T. Ebrahimi. “Intuitive Strategy for Parameter Setting in Video Segmentation”. Proc. of *SPIE, Visual Communications and Image Processing 2003*, Vol. 5150, p. 998-1008, Lugano, Switzerland, July 2003.

13. A. Cavallaro, E. Drelie Gelasca and T. Ebrahimi. “Objective evaluation of segmentation quality using spatio-temporal context”. Proc. of *IEEE International Conference on Image Processing*, pp. 301-304, September 2002.

14. N. Aspert, E. Drelie Gelasca, Y. Maret and T. Ebrahimi. “Steganography for Three-Dimensional Polygonal Meshes”. Proceedings of *SPIE 47th Annual Meeting, Applications of Digital Image Processing XXV*, July 2002.

15. S. Winkler, E. Drelie Gelasca and T. Ebrahimi. “Perceptual Quality Assessment for Video Watermarking”. In Watermarking Quality Evaluation Special Session at ITCC, *International Conference on Information Technology: Coding and Computing* pp.90-94, April 2002.