Visual saliency extraction from compressed streams

Marwa Ammar

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

Spécialité : Informatique et Télécommunications

Ecole doctorale : Informatique, Télécommunications et Electronique de Paris

Présentée par

Marwa AMMAR

Pour obtenir le grade de

DOCTEUR DE TELECOM SUDPARIS

Visual saliency extraction from compressed streams

Soutenue le 15 juin 2017

devant le jury composé de :

Pr. Patrick GALLINARI, Université Pierre et Marie Curie	**Président**
Pr. Jenny BENOIS-PINEAU, Université de Bordeaux	**Rapporteur**
MdC. HDR Claude DELPHA, Université ParisSud	**Rapporteur**
Pr. Faouzi GHORBEL, ENSI Tunis	**Examinateur**
Dr. Matei MANCAS, Université de Mons	**Examinateur**
Pr. Patrick LE CALLET, Université de Nantes	**Invité**
MdC HDR. Mihai MITREA, IMT-TSP	**Directeur de thèse**

Thèse n° 2017TELE0012

To my sons, my husband and my parents

This thesis becomes a reality with the support and help of many people to whom I would like to express my sincere thanks and acknowledgment.

My deep gratitude goes to my thesis director, HDR Mihai Mitrea for his warm welcome when I first stepped to the ARTEMIS department at the IMT – Telecom SudParis. I would like to express my appreciation for his trust and for seeing in me a future PhD. I would also like to thank him for granting me with the chance of starting the research work on the novel and exciting topic of visual saliency in the compressed stream as well for his valuable guidance, timely suggestions and support throughout not only my thesis but my engineering and masters internships as well.

My deep gratitude also goes to the distinguished members of my defense committee, and particularly to the two reviewers, Prof. Jenny Benois-Pineau and Prof. Claude Delpha, for their precious feedback and enriching comments that contributed to the final version of this manuscript.

I am thankful to Ecole Nationale des Sciences de l'Informatique, for the sound education I received there during my engineering and masters programs.

My colleagues Marwen Hasnaoui and Ismail Boujelbane deserve a special mention: I thank them for helping me with their watermarking and software skills and passion as well as for their availability during this thesis.

I would like to thank Mrs. Evelyne Taroni for her proactive attitude and valuable administrative help during my engineering, master and PhD internships at the ARTEMIS department.

In addition, I like to thank the entire ARTEMIS team, former and present members that I have met and particularly Rania Bensaied who helped me with the subjective evaluation experiments.

I am mostly fortunate to have the opportunity to acknowledge gratitude to the people who mean the most to me. My parents Mohamed and Leila, who raised me, taught me, and supported me all throughout my life: their selfless love, care, pain and sacrifices shaped my life.

I like to deeply thank my brother Yassine, my sister Siwar and my nephew Hassan for their motivational discussions and emotional support.

I am extremely thankful to my family in law for loving and encouraging me during my thesis.

For my friends Mehdi, Azza and Ola and for all those who have touched my life in any way since I started this thesis, I am grateful for all they have done.

Last but not the least, I owe thanks to a very special person, my husband Anis, for his continuous and unfailing love, support and understanding during the pursuit of my PhD degree. He was always around at times I thought that it would be impossible to continue, he helped me to keep things in perspective and that made the completion of thesis possible. I appreciate my little son Anas, for abiding my ignorance and for the patience he showed during my thesis writing. Words would never say how grateful I am to both of them. I consider myself the luckiest in the world to have such a lovely and caring family, standing beside me with their love and unconditional support.

Table of Contents

RESUME 15

ABSTRACT 25

I. INTRODUCTION 35

I.1. Saliency context 37
I.1.1. Biological basis for visual perception 37
I.1.2. Image processing oriented vision modeling 38

I.2. Watermarking context 43

I.3. Video coding & redundancy 45

I.4. Conclusion 46

II. STATE OF THE ART 49

II.1. Bottom-up visual saliency models 51
II.1.1. Image saliency map 51
II.1.2. Video saliency map 58
II.1.3. Conclusion 64

II.2. Visual saliency as a watermarking optimization tool 68

II.3. Direct compressed video stream processing 72

III. SALIENCY EXTRACTION FROM MPEG-4 AVC STREAM 77

III.1. MPEG-4 AVC saliency map computation 79
III.1.1. MPEG-4 AVC elementary saliency maps 79
III.1.2. Elementary saliency maps post-processing 83
III.1.3. Elementary saliency map pooling 84

III.2. Experimental results 85
III.2.1. Ground truth validation 86
III.2.2. Applicative validation 97

III.3. Discussion on the results 100

III.4. Conclusion **108**

IV. SALIENCY EXTRACTION FROM HEVC STREAM 109

IV.1. HEVC saliency map computation **111**
IV.1.1. HEVC elementary saliency maps 112
IV.1.2. Elementary saliency map post-processing 115
IV.1.3. Saliency maps pooling 115

IV.2. Experimental results **115**
IV.2.1. Ground truth validation 116
IV.2.2. Applicative validation 124

IV.3. Discussion on the results **126**

IV.4. Conclusion **132**

V. CONCLUSION AND FUTURE WORK 133

V.1. Conclusion **134**
V.1.1. Saliency vs. Compression 134
V.1.2. Saliency vs. Watermarking 136

V.2. Future works **137**

VI. APPENDIXES 139

A Fusing formula investigation **140**
A.1. MPEG-4 AVC fusing formula validation 142
A.2. HEVC fusing formula validation 147
A.3. Conclusion 151

B. MPEG-4 AVC basics **152**
B.1. Structure 152
B.2. Encoding 153

C. HEVC basics **158**
C.1. Structure 158
C.2. Encoding 159
C.3 How HEVC is different? 161

D. Tables of the experimental results **162**
D.1 MPEG-4 AVC saliency map validation 162
D.2 HEVC saliency map validation 165
D.3 Conclusion 168

E. Graphics of the experimental results 171

REFERENCES 173

LIST OF PUBLICATIONS 181

LIST OF ACRONYMS 183

List of figures

Figure 0-1: Evolution du contenu multimédia. 17

Figure 0-2: Le temps moyen (en heure) passé en regardant un contenu télé/vidéo dans le monde durant la deuxième trimestre de 2016 [WEB01]. 18

Figure 0-3: Le trafic internet du consommateur 2015-2019 [WEB02]. 18

Figure 1: Multimedia content evolution. 27

Figure 2: Average daily time (in hours) spent on viewing TV/video content worldwide during the second quarter 2016 [WEB01]. 28

Figure 3: Consumer Internet traffic 2015-2019 [WEB02]. 28

Figure I-1: Human eye anatomy. 37

Figure I-2: Visual saliency features. 42

Figure I-3: General scheme of watermarking approach. 44

Figure I-4: MPEG-4 AVC/HEVC compression chain. 45

Figure II-1: Domains of bottom-up saliency detection models; in blue: studies related to still images; in green: studies related to videos. P, T, Q, E stand for Prediction, Transformation, Quantification and Encoding, respectively. 51

Figure II-2: Synopsis of Itti's model [ITT98]: the saliency map is obtained by a multi-scale extraction model consisting on three feature extraction, normalization and fusion of the elementary maps. 52

Figure II-3: Saliency extraction based on the Shannon's self-information [BRU05]: the visual saliency is determined by a sparse representation of the image statistics, learned from the prior knowledge of the brain. 53

Figure II-4: Computation steps of Harel's model [HAR06]: the saliency is determined by extracting features, normalising, then fusing the elementary maps. 54

Figure II-5: Flowchart of the biologically inspired model advanced in [LEM06] 54

Figure II-6: Saliency map computation flowchart: extracting visual saliency by exploiting the singularities in the spectral residual. 55

Figure II-7: A context aware saliency model: the saliency is enhanced by using multiple scale filtering and visual coherency rules [GOF10]. 55

Figure II-8: Principle of the saliency approach [MUR11]: the saliency is obtained according to a biologically inspired representation based on predicting color appearance. 56

Figure II-9: Soft image abstraction and decomposition into perceptually homogenous regions [CHE13]: the saliency map is extracted by considering both appearance similarity and spatial overlap. 57

Figure II-10: Saliency map computation steps [FAN12]: the saliency map is obtained, in the transformed domain of the JPEG compression, through a so-called coherent normalized-based fusion. 57

Figure II-11: Workflow of the saliency model [ZHA06]: the saliency map is obtained through a dynamic fusion of the static and the temporal attention model. 58

Figure II-12: Flowchart of the proposed model [LEM07]: the saliency map is the result of a weighted average operation of achromatic and two chromatic saliency maps. 59

Figure II-13: Incremental coding length model's different steps [HOU08]: the saliency extraction model is based on the incremental coding length of each feature. 60

Figure II-14: Illustration of image/video saliency detection model [SEO09]: the saliency map is obtained by applying the self resemblance indicating the likelihood of saliency in a given location. 61

Figure II-15: Saliency computation graph [MAR09]: the attention model was computed on two parallel ways: the static way and the dynamic way. 62

Figure II-16: Multiresolution spatiotemporal saliency detection model based on the phase spectrum of quaternion Fourier transform (PQFT) [GUO10]. 62

Figure II-17: Flowchart of the saliency computation model [FAN14]: the visual saliency is extracted from the transformed domain of the MPEG-4 ASP. 64

Figure II-18: Principle of a watermark embedding scheme based on saliency map. 68

Figure II-19: Video quality evolution. 72

Figure III-1: Saliency map computation in a GOP. 79

Figure III-2: Orientation saliency: the central block into a 5x5 block neighborhood is not salient when its "orientation" is identical with its neighbors (see the left side of the figure); conversely, if the block orientation differs from its neighbors, the block is salient (see the right side of the figure). 82

Figure III-3: Motion saliency: the motion amplitude over all the P frames in the GOP is summed-up. 83

Figure III-4: Features map normalization. 84

Figure III-5: MPEG-4 AVC saliency map (on the left) vs. density fixation map (on the right). 87

Figure III-6: KLD between saliency map and density fixation map. 88

Figure III-7: AUC between saliency map and density fixation map. 91

Figure III-8: Saliency map behavior at human fixation locations (in red + signs) vs. saliency map behavior at random locations (in blue x signs). 92

Figure III-9: KLD between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence). 93

Figure III-10: AUC between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence). 93

Figure III-11: KLD between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence). 95

Figure III-12: AUC between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence). 97

Figure III-13: Illustrations of saliency maps computed with different models. 107

Figure IV-1: Difference between HEVC and MPEG-4 AVC block composition. 112

Figure IV-2: KLD between saliency map and density fixation map. 117

Figure IV-3: AUC between saliency map and density fixation map. 119

Figure IV-4: KLD between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence). 121

Figure IV-5: AUC between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence). 122

Figure IV-6: KLD between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence). 122

Figure IV-7: AUC between saliency maps at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence). 124

Figure IV-8: Illustrations of saliency maps computed with different models. 131

List of tables

Tableau 0-1: Extraction de la saillance visuelle à partir du domaine vidéo compressé: contraintes, défis, limitations et contributions..... 24

Table 1: Visual saliency extraction from video compressed domain: constraints, challenges, current limitations and contributions. 34

Table II-1: State of the art synopsis of saliency detection models. 66

Table II-2: State-of-the-art of the watermark embedding scheme based on saliency map. 71

Table II-3: State of the art of the compressed stream application. 75

Table III-1: Assessment of the model performance in predicting visual saliency. 87

Table III-2: KLD gains between Skewness-max, Combined-avg and Addition-avg and the state of the art methods [CHE13] [SEO09] [GOF12]. 89

Table III-3: KLD sensitivity gains between Skewness-max, Combined-avg and Addition-avg and the state of the art methods [CHE13] [SEO09] [GOF12]. 90

Table III-4: AUC values between saliency map and density fixation map with different binarization thresholds. 91

Table III-5: AUC sensitivity gains between Skewness-max and Combined-avg and the state-of-the-art methods [CHE13][SEO09][GOF12]. 94

Table III-6: AUC values between saliency map at fixation locations and saliency map at random locations with different binarization thresholds (N=100 trials)..... 94

Table III-7: KLD gains between Multiplication-avg and Static-avg and the three state of the art methods [CHE13][SEO09][GOF12]. 96

Table III-8: KLD sensitivity gains between Multiplication-avg and Static-avg and the three state of the art methods [CHE13][SEO09][GOF12]..... 96

Table III-9: Objective quality evaluation of the transparency when alternatively considering random selection and “Skewness-max” saliency map based selection..... 99

Table III-10: MOS gain between the QIM method with random selection and saliency map “Skewness-max” based selection..... 100

Table III-11: Ground truth validation results..... 100

Table III-12: Computational complexity comparison between our method and the three state of the art models considered in our study. 104

Table III-13: Computational time per processed frame of our method and the three state of the art models considered in our study. 104

Table IV-1: KLD gains between all the combination of HEVC saliency maps and the state of the art methods [CHE13] [SEO09] [GOF12] and MPEG-4 AVC saliency map..... 118

Table IV-2: KLD sensitivity gains between all considered HEVC saliency map combinations and the state of the art methods [CHE13] [SEO09] [GOF12] and MPEG-4 AVC saliency map. 118

Table IV-3: AUC gains between all the combinations of HEVC saliency maps and the state of the art methods [CHE13] [SEO09] [GOF12] and MPEG-4 AVC saliency map. 119

Table IV-4: AUC sensitivity gains between Combined-avg, Addition-avg and Static-avg and the state of the art methods [CHE13] [SEO09] [GOF12] and MPEG-4 AVC saliency map. 120

Table IV-5: KLD gains between Multiplication-avg and Static-avg and the state of the art methods [CHE13] [SEO09] [GOF12] and MPEG-4 AVC saliency map. 123

Table IV-6: KLD sensitivity gains between Multiplication-avg and Static-avg and the state of the art methods [CHE13] [SEO09] [GOF12] and MPEG-4 AVC saliency map. 123

Table IV-7: Objective quality evaluation of the transparency when alternatively considering random selection and "Combined-avg" saliency map based selection. 125

Table IV-8: MOS gain between the watermarking method with random selection and saliency map "Combined-avg" based selection. 126

Table IV-9: Ground truth validation results 127

Table V-1: Comparison of the results of KLD and AUC between saliency maps and fixation maps. 135

Table V-2: Comparison of the results of KLD and AUC between saliency maps at fixation locations and saliency maps at random locations (N=100 trials for each frame in the video sequence). 136

Résumé

Le contexte

Dans dix ans, allez-vous lire ce rapport de thèse ou le regarder en tant que vidéo? Que vont capter vos yeux en premier lieu?

En 2020, 82% du trafic sur internet sera conquis par les vidéos…

Au début des années 1980, les ordinateurs ont émergé dans les entreprises, les écoles et les maisons.

À la fin des années 1980 et au cours des années 1990, les scientifiques ont commencé à imaginer comment les ordinateurs pouvaient être exploités comme jamais auparavant. Ils ont considéré le multimédia comme un moyen d'utiliser les ordinateurs d'une manière personnelle, unique, en fournissant des informations non seulement en utilisant du texte, mais aussi des images, du son, de la vidéo et des graphiques 3D.

Au fil des années, les technologies et les applications multimédias ont progressivement conquis notre vie, faisant partie aujourd'hui de notre routine professionnelle et personnelle, Figure 0-1. Des encyclopédies aux livres de cuisine et de la simulation scientifique aux jeux FIFA, le contenu multimédia devient notre référence et, qu'on l'accepte ou non, notre premier repère dans les activités sociales professionnelles et personnelles.

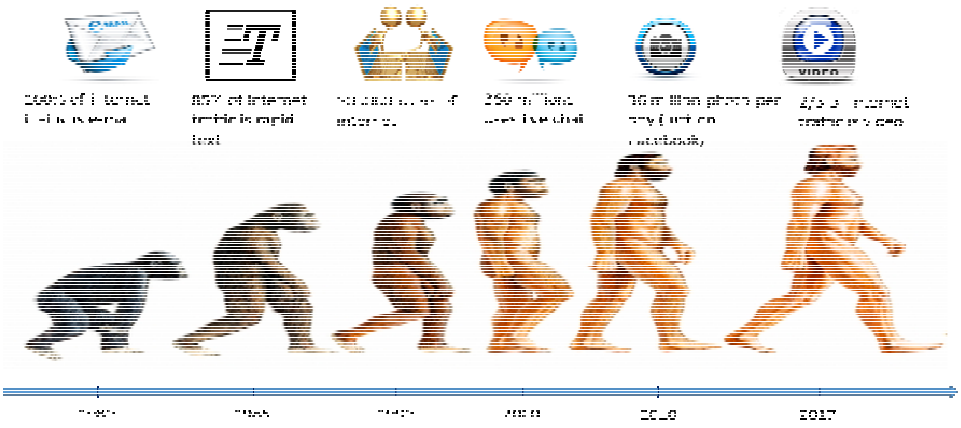


Figure 0-1: Evolution du contenu multimédia.

De nos jours, grâce aux dispositifs abordables (capture, traitement et stockage) et à l'ubiquité de l'accès (très) haut débit, une quantité massive de contenu vidéo générée par l'utilisateur est produite et distribuée instantanément. Au moment de la rédaction du présent document, 2,5 exabytes de données vidéo (soit environ 90 ans de vidéos HD) sont produites chaque jour. Figure 0-2 montre la durée moyenne, dans le monde, que passe un utilisateur à regarder une vidéo sur internet ou devant la télé. Par exemple, en France, 4,1 heures par jour sont consacrées à regarder du contenu vidéo !

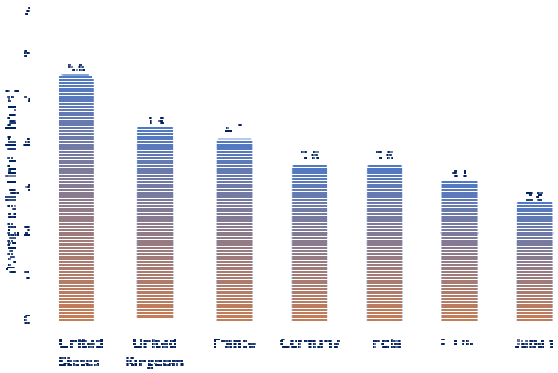


Figure 0-2: Le temps moyen (en heure) passé en regardant un contenu télé/vidéo dans le monde durant la deuxième trimestre de 2016 [WEB01].

L'enregistrement de toutes les visualisations et toutes les inscriptions des utilisateurs des réseaux sociaux montre des statistiques très intéressantes sur la tendance de l'utilisation vidéo, [WEB02]. Chaque jour, les utilisateurs de Snapchat regardent 6 milliards de vidéos alors que les utilisateurs de YouTube passent 46 000 ans à regarder des vidéos. Le contenu 'How-to' lié à la cuisine et à la nourriture sur YouTube est incroyablement populaire, avec 419 millions de vues, tandis que 68% des mères de la génération millénaire ont déclaré avoir regardé aussi des vidéos pendant la préparation des repas [WEB03]. Aux États-Unis, plus que 155 millions de personnes jouent à des jeux vidéo malgré la différence et la variété de leurs âges, sexes et statuts socioéconomiques.

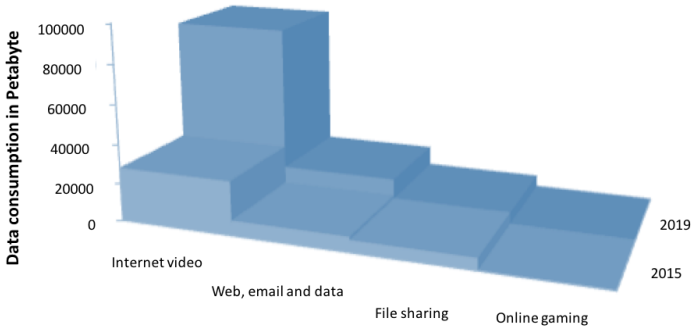


Figure 0-3: Le trafic internet du consommateur 2015-2019 [WEB02].

La Figure 0-3 montre que l'internaute a une préférence remarquable pour regarder la vidéo plutôt que de consommer tout autre contenu multimédia. La suprématie du contenu vidéo sur le trafic internet sera renforcée dans un proche avenir: en 2020, 82% du trafic sur internet sera conquis par les vidéos [WEB04].

Le monde contient trop d'information visuelle pour arriver à la percevoir spontanément …

En raison de sa taille et de sa complexité, la production, la distribution et l'utilisation des vidéos a augmenté le besoin et la nécessité des études et des recherches scientifiques qui traitent la relation entre les contenus numériques et le mécanisme visuel humain.

Il y a une énorme différence entre l'image affichée sur un dispositif et l'image que notre cerveau perçoit. Il existe, par exemple, une différence entre la luminance d'un pixel sur un écran d'ordinateur et son impact visuel. La vision dépend non seulement de la perception des objets, mais aussi d'autres facteurs visuels, cognitifs et sémantiques.

Le système visuel humain (SVH) a la capacité remarquable d'être attiré automatiquement par des régions saillantes. Les bases théoriques de la modélisation de la saillance visuelle ont été établies, il y a 35 ans, par Treisman [TRE80] qui a proposé la théorie d'intégration du système visuel humain : dans tout contenu visuel, certaines régions sont saillantes grâce à la différence entre leurs caractéristiques (intensité, couleur, texture, et mouvement) et les caractéristiques de leurs voisinages.

Peu de temps après, Koch [KOC85] a mis en œuvre un mécanisme de sélectivité, stimulant l'attention humaine : dans n'importe quel contenu visuel, les régions qui stimulent les nerfs de la vision sont d'abord choisies et traitées, puis le reste de la scène est interprété.

Dans le traitement de l'image et de la vidéo, le mécanisme complexe de l'attention visuelle est généralement présenté par une carte dite *carte de saillance*. Une carte de saillance est généralement définie comme une carte topographique 2D représentant les régions d'une image/vidéo sur laquelle le système visuel humain se focalisera spontanément.

Les objectifs

Cette thèse vise à offrir un cadre méthodologique et expérimental complet pour traiter la possibilité d'extraire les régions saillantes directement à partir des flux vidéo compressés (MPEG-4 AVC et HEVC), avec des opérations de décodage minimales.

Notez que l'extraction de la saillance visuelle à partir du domaine compressé est à priori une contradiction conceptuelle. D'une part, comme suggéré par Treisman [TRE80], la saillance est donnée pas des singularités visuelles dans le contenu vidéo. D'autre part, afin d'éliminer la redondance visuelle, les flux compressés ne sont plus censés présenter des singularités. Par conséquence, la thèse étudie si la saillance peut être extraite directement à partir du flux compressé ou, au contraire, des opérations complexes de décodage et de pré/post-traitement sont nécessaires pour ce faire.

La thèse vise également à étudier le gain pratique de l'extraction de la saillance visuelle du domaine compressé. A cet égard, on a traité le cas particulier du tatouage numérique robuste des contenus vidéo. On s'attend que la saillance visuelle acte comme un outil d'optimisation, ce qui permet d'améliorer la transparence (pour une quantité de données insérées et une robustesse contre les attaques prescrites) tout en diminuant la complexité globale de calcul. Cependant, la preuve du concept est encore attendue.

L'état de l'art: limitations et contraintes:

La thèse porte sur les limitations et les contraintes liées au cadre méthodologique de l'extraction de la saillance visuelle à partir du domaine compressé, à sa validation par rapport à la vérité terrain ainsi que sa validation applicative.

Tout d'abord, il faut noter que plusieurs études, concernant les images fixes et la vidéo, ont déjà considéré des cartes de saillance afin d'améliorer les performances d'une grande variété d'applications telles que le traitement des scènes rapides, la prédiction des vidéos surveillances et la détection/reconnaissance d'objets… Ces études couvrent une large étendue d'outils méthodologiques, de la décomposition pyramidale dyadique gaussienne aux modèles inspirés par la biologie. Cependant, malgré leur vaste spectre méthodologique, les modèles existants extraient les régions saillantes à partir du domaine des pixels. D'après notre connaissance, au début de cette thèse, aucun modèle d'extraction dans le domaine compressé n'a été signalé dans la littérature.

Deuxièmement, d'un point de vue évaluation quantitative, les études de la littérature considèrent différentes bases de données, de différentes tailles (par exemple, de 8 images fixes à 50 séquences vidéo jusqu'à 25 min) et / ou pertinence (cartes de densité de fixation, les emplacements du saccade, …). La confrontation de la carte de saillance obtenue à la vérité terrain est étudiée en considérant des types particuliers de mesures, comme les métriques basées sur la distribution (par exemple, la divergence de Kullback Leibler, le coefficient de corrélation linéaire, la similitude, …) et les métriques basées sur la localisation (surface sous la courbe, selon différentes implémentations). Par conséquence, assurer une évaluation objective et une comparaison entre les modèles les plus modernes reste un défi.

Enfin, les particularités du SVH sont déjà déployées avec succès en tant qu'outil d'optimisation de tatouage, comme par exemple l'adaptation perceptive au contenu (*preceptual shaping*), le masquage perceptuel, les mesures de qualité inspirées par la biologie. Malgré que la saillance visuelle ait déjà prouvé son efficacité dans le domaine compressé, aucune application de tatouage utilisant la carte de saillance comme outil d'optimisation n'a été présentée avant le début de cette thèse.

Les contributions

La thèse présente les contributions suivantes.

Cadre méthodologique de l'extraction de la saillance visuelle à partir du flux compressé

La détection automatique de la saillance visuelle est un domaine de recherche particulier. Son arrière-plan fondamental (neurobiologique) est représenté par les travaux de Treisman avançant la théorie de l'intégration pour le système visuel humain et par ceux de Koch *et al.* mettant en évidence un mécanisme de sélectivité temporelle de l'attention humaine. D'un point de vue méthodologique, toutes les études publiées dans la littérature suivent une approche expérimentale inhérente: certaines hypothèses sur la façon dont les caractéristiques neurobiologiques peuvent être (automatiquement)

calculées à partir du contenu visuel sont d'abord formulées puis validées par des expériences. On peut donner ainsi comme exemple l'étude d'Itti [ITT98] qui a été cité, selon scholar google, environ 7000 fois.

Dans ce cadre, la contribution de la thèse n'est pas de proposer une nouvelle approche, mais à contrario, de démontrer méthodologiquement la possibilité de lier les éléments de syntaxe des flux MPEG-4 AVC et HEVC à la représentation mathématique originale d'Itti. Il est ainsi mis en évidence que les normes de compression les plus efficaces aujourd'hui (MPEG-4 AVC et HEVC) conservent toujours dans leurs éléments de syntaxe les singularités visuelles auxquelles le SVH est adapté.

Afin de calculer la carte de saillance directement à partir des flux compressés MPEG-4 AVC / HEVC, les principes de conservation de l'énergie et de la maximisation du gradient sont conjointement adaptés aux particularités du SVH et de la syntaxe du flux MPEG. Dans ce cas, les caractéristiques statiques et de mouvement sont d'abord extraites des trames *I* et respectivement *P*. Trois caractéristiques statiques sont considérées. L'intensité est calculée à partir des coefficients *luma* résiduels, la couleur est calculée à partir des coefficients *chroma* résiduels tandis que l'orientation est donnée par la variation (gradient) des modes de prédiction intra-directionnelle. Le mouvement est considéré comme l'énergie des vecteurs de mouvement. Deuxièmement, nous calculons les cartes de saillance individuelles pour les quatre caractéristiques mentionnées ci-dessus (intensité, couleur, orientation et mouvement). Les cartes de saillance sont obtenues à partir des cartes de caractéristiques après trois étapes incrémentales : la détection des *outliers*, le filtrage moyenneur avec le noyau de la taille de la fovéa et la normalisation dans l'intervalle [0, 1].

Enfin, nous obtenons une carte de saillance statique en fusionnant les cartes d'intensité, de couleur et d'orientation. La carte de saillance globale est obtenue en regroupant la carte statique et celle de mouvement selon 48 combinaisons différentes de techniques de fusion.

Confrontation de la carte de saillance extraite directement à partir du flux compressé à la vérité terrain

Comme nous l'avons déjà expliqué, chaque modèle d'extraction de la saillance visuelle doit être validé par une évaluation quantitative.

De ce point de vue, la principale contribution de la thèse consiste à définir un *test-bed* générique permettant une validation objective et une analyse comparative.

Le *test-bed* défini dans cette thèse est caractérisé par trois propriétés principales: (1) il permet d'évaluer les différences entre la vérité terrain et la carte de saillance par différents critères, (2) il comprend différentes typologies de mesures et (3) il assure une pertinence statistique aux évaluations quantitatives.

En conséquent, ce *test-bed* est structuré à trois niveaux, selon les critères d'évaluations et selon les mesures et les corpus utilisés, respectivement.

Tout d'abord, plusieurs critères d'évaluation peuvent être pris en considération. La P*récision* (définie comme la ressemblance entre la carte de saillance et la carte de fixation) et la *Discriminance* (définie comme la différence entre le comportement de la carte de saillance dans les zones de fixations et les régions aléatoires) des modèles de saillance sont considérés.

Deuxièmement, pour chaque type d'évaluation, plusieurs mesures peuvent être considérées. Notre évaluation est basée sur deux mesures de deux types différents: la KLD (divergence de Kullback Leibler), basée sur la distribution statistique des valeurs [KUL51][KUL68] et l'AUC (surface sous la courbe) qui est une mesure basée sur la localisation des valeurs.

Deux corpus sont considérés: (1) le corpus dit *de référence* organisé par [WEB05] à IRCCyN et (2) le corpus dit *d'étude comparative* organisé par [WEB06] au CRCNS. Ces deux corpus sont sélectionnés selon leurs compositions (diversité du contenu et disponibilité de la vérité terrain en format compressé), leurs représentativités pour la communauté de la saillance visuelle ainsi que leurs tailles. Une attention particulière est accordée à la pertinence statistique des résultats présentés dans la thèse. À cet égard, nous considérons:

- Pour les deux critères d'évaluation, la *Précision* et la *Discriminance,* chaque valeur de KLD et d'AUC est présenté avec sa moyenne, ses valeurs minimales et maximales, et l'intervalle de confiance à 95% correspondant.
- Pour l'évaluation de la Discriminance, chaque expérience (c'est-à-dire pour chaque trame dans chaque séquence vidéo) est répétée 100 fois (c'est-à-dire pour 100 ensembles de localisation aléatoire). La valeur finale est moyennée sur toutes ces configurations et toutes les trames dans la séquence vidéo;
- Pour l'étude de la Précision et de la Discriminance, on a analysé la sensibilité des mesures KLD et AUC par rapport au caractère aléatoire du contenu vidéo constituant le corpus utilisé.

Ce *test-bed* a été considéré pour comparer notre méthode d'extraction de la carte de saillance MPEG-4 AVC contre trois méthodes de l'état de l'art. La carte de saillance HEVC a été comparée à son tour contre les mêmes trois méthodes de l'état de l'art ainsi que contre la carte de saillance MPEG-4 AVC. Les trois méthodes de l'état de l'art ont été choisies selon les critères suivants: la représentativité dans l'état de l'art, la possibilité d'une comparaison équitable et la complémentarité méthodologique.

Pour illustration, les résultats de la confrontation de notre carte de saillance MPEG-4 AVC par rapport à la vérité terrain montrent des gains relatifs en KLD entre 60% et 164% et en AUC entre 17% et 21% contre les trois modèles de l'état de l'art. Pour la carte de saillance HEVC, les gains en KLD se situent entre 0,01 et 0,4 tandis que les gains en AUC se situent entre 0,01 et 0,22 contre les mêmes modèles de l'état de l'art.

Validation applicative dans une méthode de tatouage robuste

Nous étudions les avantages de l'extraction de la carte de saillance directement à partir du flux compressé lors du déploiement d'une application de tatouage robuste. En fait, **en utilisant le modèle d'extraction de la saillance visuelle à partir des flux MPEG-4 AVC / HEVC comme guide pour**

sélectionner les régions dans lesquelles la marque est insérée, des gains de transparence (pour une quantité de données insérées et une robustesse prédéfinies) sont obtenus. La validation applicative révèle des gains de transparence allant jusqu'à 10 dB en PSNR pour les cartes de saillance MPEG-4 AVC et jusqu'à 3dB en PSNR pour les cartes de saillance HEVC (pour une quantité de données insérées et une robustesse bien définies).

En plus de sa pertinence applicative, ces résultats peuvent également être considérés comme une première étape vers une validation à posteriori de l'hypothèse de Koch : la saillance à court terme et le masquage perceptuel à long terme peuvent être considérés d'une manière complémentaire afin d'accroitre la qualité visuelle.

Comme conclusion générale, la thèse démontre que bien les normes MPEG-4 AVC et HEVC ne dépendent pas explicitement de tout principe de saillance visuelle, ses éléments syntaxiques préservent cette propriété.

La structure de la thèse

Afin d'offrir une vision méthodologique et expérimentale complète de la possibilité d'extraire les régions saillantes directement à partir des flux compressés vidéo (MPEG-4 AVC et HEVC), cette thèse est structurée comme suit.

Le chapitre I couvre les aspects introductifs et se compose de trois parties principales, liées à la saillance visuelle, au tatouage et au codage vidéo, respectivement.

Le chapitre II est consacré à l'analyse de l'état de l'art. Il est divisé en trois parties principales. Le chapitre II.1 traite les méthodes d'extraction de la saillance visuelle *bottom-up* et est structurée en deux niveaux : image contre vidéo et pixel contre domaine compressé. Le chapitre II.2 donne un bref aperçu sur la relation méthodologique entre les applications de tatouage et la saillance visuelle. Le chapitre II.3 concerne les applications traitant directement le domaine vidéo compressé.

Le chapitre III présente le cadre méthodologique et expérimental de l'extraction de la saillance visuelle à partir du flux compressé MPEG-4 AVC. Le chapitre VI est structuré de la même manière que le chapitre III et présente le cadre méthodologique et expérimental pour l'extraction de la saillance visuelle à partir du flux compressé HEVC.

Le dernier chapitre est consacré aux conclusions et aux perspectives.

La thèse contient cinq annexes. L'annexe A est consacrée à l'étude de la technique de fusion pour les modèles d'extraction MPEG-4 AVC et HEVC. L'annexe B donne un aperçu sur la norme MPEG-4 AVC. L'annexe C identifie les principaux éléments de nouveauté pour la norme HEVC. L'annexe D détaille les valeurs numériques des résultats données dans les chapitres III, IV et V. L'annexe E présente sous forme de graphiques les résultats présentés dans les tableaux du chapitre III.

Tableau 0-1: Extraction de la saillance visuelle à partir du domaine vidéo compressé: contraintes, défis, limitations et contributions

Contraintes	Défis	Limitations	Contributions
Extraction de la saillance visuelle	• L'extraction de la saillance visuelle à partir des flux compressés: MPEG-4 AVC et HEVC	• Les caractéristiques de la saillance visuelle sont extraites à partir des pixels	• Spécifier un formalisme reliant le système visuel humain aux caractéristiques élémentaires des éléments de syntaxe des flux MPEG-4 AVC et HEVC • Définir des stratégies de normalisation pour les cartes obtenues • Etudier la fusion des cartes statiques et dynamiques pour obtenir une carte de saillance du flux compressé
Evaluation des performances	• Confrontation à la vérité terrain: *Précision et Discriminance*	• Données limitées • Procédures d'évaluation variables	• Spécifier un *test-bed* cohérent et unitaire permettant la confrontation des cartes de saillance à la vérité terrain: ▪ Les critères d'évaluation : • Précision : La ressemblance entre la carte de saillance et la carte de fixation • Discriminance : La différence entre le comportement de la carte de saillance dans les régions de fixation et les endroits aléatoires ▪ Typologie des mesures : • Une métrique basée sur la distribution: le KLD implémenté en fonction de la théorie de l'information de Kullback Leibler [KUL51], [KUL68] • Une métrique basée sur l'emplacement: AUC ▪ Des corpus différents : • Le corpus de référence organisé par IRCCyN [WEB05] • Le corpus de l'analyse comparative organisé by Itti [WEB06] ▪ Pertinence statistique : • Précision et Discriminance : valeurs expérimentales présentées aves leurs moyennes, min, max et intervalle de confiance à 95%. • Discriminance: Processus de calcul de la moyenne supplémentaire sur les testes aléatoires répétées; • Précision et Discriminance: Évaluation de la sensibilité des mesures par rapport au caractère aléatoire du contenu visuel.
Intégration dans l'application de tatouage	• Garder les caractéristiques de l'application tout en diminuant le coût de calcul.	• Pas de validation d'une carte de saillance dans une application dans le domaine compressé	• Démontrer la possibilité d'intégration de la carte de saillance du flux compressé dans une application de tatouage pour guider l'insertion de la marque. • Améliorer de la transparence de la méthode de tatouage, à une robustesse et une quantité de données préservées, tout en réduisant le coût de calcul.

Abstract

Context

Ten years from now on, would you be still reading this thesis manuscript or watching it as a video? What would your eyes first pick-up from it?

By 2020, 82% of the world's Internet traffic will be video…

Early 1980s, computers became relevant in enterprises, schools and homes.

Late 1980s and during the 1990s, scientists started imagining how computers could be used as never before. They considered multimedia as a way to utilize computers in a unique personal way, by delivering information not only using text but pictures, audio, video and 3D graphics, as well.

Over the years, multimedia technologies and applications have gradually conquered our lives, becoming part of our intimate, professional and personal routine, Figure 1. From encyclopedias to cookbooks and from scientific simulation to FIFA gaming, the multimedia content becomes our reference and, accepting it or not, our first ground in professional and personal social activities.

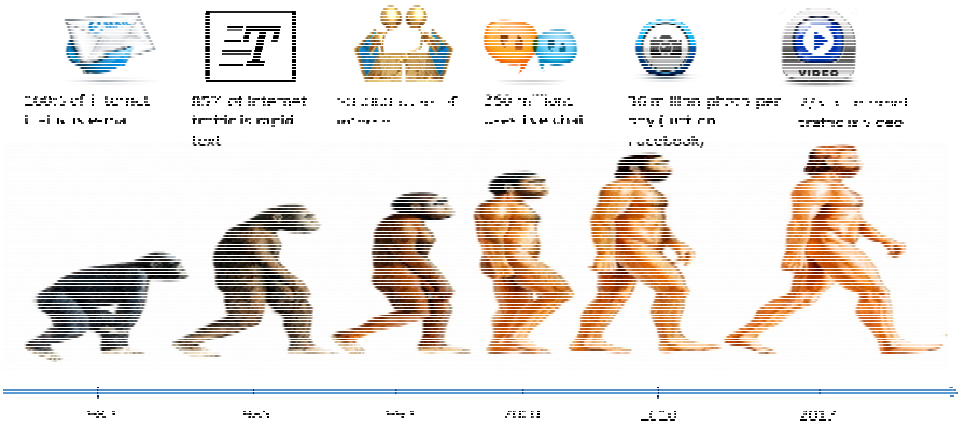


Figure 1: Multimedia content evolution.

Nowadays, thanks to the affordable devices (capturing, processing and storage) and to the ubiquity of broadband access, massive amount of user-generated video content is instantaneously produced and distributed. At the time of writing, 2.5 Exabyte of video data (that is, about 90 years of HD videos) are produced every day. Figure 2 shows the worldwide average (per user) daily time spent watching TV and Internet video content, sorted by country; the figures are reported by Statista [WEB01] and correspond to the second quarter of 2016. Just for illustration, in France, 4.1 hours a day are spent for watching video content!

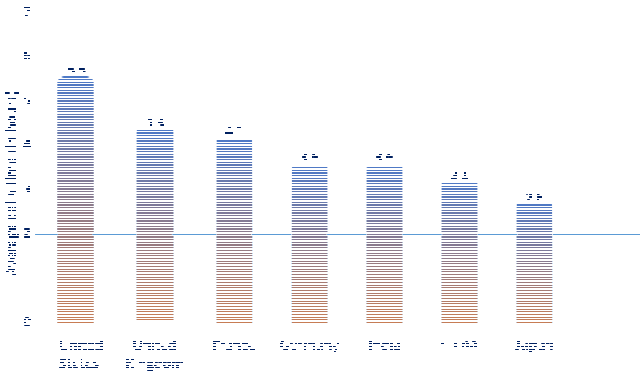


Figure 2: Average daily time (in hours) spent on viewing TV/video content worldwide during the second quarter 2016 [WEB01].

Recording every view and every sign up of social media users come across with a very interesting statistics about the tendency of the video usage, [WEB02] and Figure 3. Every day, Snapchat users watch 6 billion videos while YouTube users spend 46000 years watching videos. “How-to” content related to food on YouTube is incredibly popular, with 419 million views while 68% of millennial moms said that they also watch videos while cooking [WEB03]. In US, over 155 million people with a large variety of backgrounds, ages, genders and socioeconomic statues are playing video games.

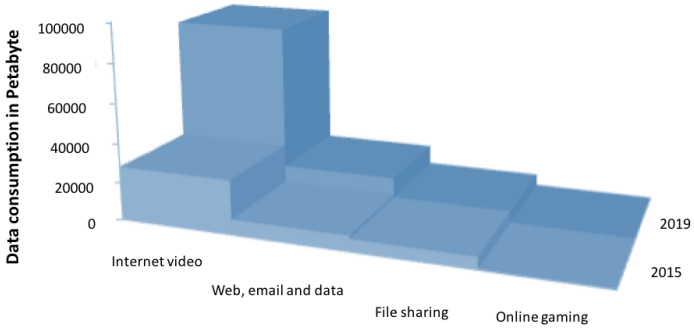


Figure 3: Consumer Internet traffic 2015-2019 [WEB02].

Figure 3 shows that the Internet user has a remarkable preference to watch video than consuming any other multimedia content. The video content supremacy over the Internet traffic will be reinforced in the near future: by 2020, 82% of the world's Internet traffic will be video [WEB04].

The world contains too much visual information to be perceived at once…

Because of its size and complexity, video content production, distribution and usage increases the need for research studies connecting the digital representation to the inner human visual mechanisms.

There is a tremendous difference between the image displayed on a device and the image our brain actually perceive. For instance, there is a difference between the luminance of a pixel on a computer

screen and its perceived impact. Vision depends not only on the ability to perceive objects (i.e., evaluated by the ratio between their size and the distance between the eye and the screen), but also on other visual, cognitive and semantic factors.

The human visual system (HVS) has the remarkable ability to automatically attend to salient regions. It can be considered that the theoretical ground for visual saliency modeling was established some 35 years ago by Treisman [TRE80] who advanced the integration theory for the human visual system: in any visual content, some regions are salient (appealing) because of the discrepancy between their features (intensity, color, texture, motion) and the features of their surrounding areas. Soon afterwards, Koch [KOC85] brought to light a time selectivity mechanism in the human attention: in any visual content, the regions that stimulate the vision nerves are firstly picked and processed, and then the rest of the scene is interpreted. In image/video processing, the complex visual saliency mechanism is generally abstracted to a so-called *saliency map*. In its broadest acceptation, a saliency map is a 2D topographic map representing the regions of an image/video on which the human visual system will spontaneously focus.

Objectives

The present thesis aims at offering a comprehensive methodological and experimental view about the possibility of extracting the salient regions directly from video compressed streams (namely MPEG-4 AVC and HEVC), with minimal decoding operations.

Note that saliency extraction from compressed domain is a priori a conceptual contradiction. On the one hand, as suggested by Treisman [TRE80], saliency is given by visual singularities in the video content. On the other hand, in order to eliminate the visual redundancy, the compressed streams are no longer expected to feature singularities. Consequently, the thesis studies weather the visual saliency can be directly bridged to stream syntax elements or, on the contrarily, complex decoding and post-processing operations are required to do so.

The thesis also aims at studying the practical benefit of the compressed domain saliency extraction. In this respect, the particular case of robust video watermarking is targeted: the saliency is expected to act as an optimization tool, allowing the transparency to be increased (for prescribed quantity of inserted information and robustness against attacks) while decreasing the overall computational complexity. However, the underlying proof of concepts is still missing and there is no a priori hint about the extent of such a behavior.

State-of-the-art limitations and constraints

The thesis deals with three-folded limitations and constraints related to the methodological framework for the compressed-domain saliency map extraction, to its ground-truth validation and to its applicative integration.

First, note that several incremental studies, from still images to uncompressed video, already considered saliency maps in order to improve the performance of a large variety of applications such as processing

of rapid scenes, selective video encoding, prediction of video surveillance, rate control, and object recognition to mention but a few. Those studies cover a large area of methodological tools, from dyadic Gaussian pyramid decomposition to biologically inspired models. However, despite their wide methodological range, the existing methods still extract the salient areas from the video pixel domain. To the best of our knowledge, at the beginning of this thesis, no saliency extraction model working on video encoded domain was reported in the literature.

Secondly, from the quantitative assessment point of view, the studies reported in the literature consider different databases, of different sizes (e.g. from 8 still images to 50 video clips summing-up to 25 min) and/or relevance (density fixation maps, saccade locations, …). The matching of the obtained saliency map to the ground truth is investigated by considering particular types of measures, like the distribution-based metrics (e.g. Kullback-Leibler Divergence, Linear Correlation Coefficient, Similarity, …) and location-based metrics (Area Under Curve, according to different implementations). Consequently, ensuring objective evaluation and comparison among and between state-of-the-art methods still remains a challenge.

Finally, the HVS peculiarities are already successfully deployed as an optimization tool in watermarking: perceptual shaping, perceptual masking, bio-inspired quality metrics stand just for some examples in this respect. Under this framework, while visual saliency already proved its effectiveness in the uncompressed domain, no study related to the possibility of using compressed domain saliency in watermarking was reported before this thesis started.

Contributions

The thesis presents the following incremental contributions.

Methodological framework for stream-based saliency extraction

The automatic visual saliency detection is a particular research field. Its fundamental (neuro-biological) background is represented by the early works of Treisman, advancing the integration theory for the human visual system and by Koch *et al.* who brought to light a time selectivity mechanism in the human attention. From the methodological point of view, all the studies published in the literature follow an inherent experimental approach: some hypotheses about how these neuro-biological characteristics can be (automatically) computed from the visual content are first formulated and then demonstrated through experiments. Maybe the most relevant example is the seminal work of Itti [ITT98], which was cited, according to scholar Google, about 7000 times

Under this framework, the thesis contribution is not to propose yet another arbitrary hypothesis, but *a contrario*, to methodologically demonstrate the possibility of linking MPEG-4 AVC and HEVC stream syntax elements to the Itti's original mathematical representation. It is thus brought to light that the most efficient to-date compression standards (MPEG-4 AVC and HEVC) still preserves in their syntax elements the visual singularities the HVS system is matched to.

In order to compute the saliency map directly in the MPEG-4 AVC/HEVC encoded domains, energy preserving and gradient maximization principles are jointly matched to the HVS and MPEG stream syntax

peculiarities. In this respect, static and the motion feature are first extracted from the *I* and *P* frames, respectively. Three static features are considered: the intensity computed from the residual luma coefficients, the color computed from the residual chroma coefficients and the orientation given by the variation (gradient) of the intra directional prediction modes. The motion feature is considered to be the energy of the motion vectors. Second, we compute individual saliency maps for the four above-mentioned features (intensity, color, orientation and motion). The saliency maps are obtained from feature maps following four incremental steps: outliers' detection, average filtering with fovea size kernel, and normalization within the [0, 1] dynamic range. Finally, we obtain a static saliency map by fusing the intensity, color and orientation maps. The global saliency map is obtained by pooling the static and the motion maps according to 48 different combinations of fusion techniques.

Ground-truth validation for stream-based saliency extraction

As explained above, any saliency extraction methodological framework must be demonstrated through quantitative evaluation. **From this point of view, the main thesis contribution consists in defining a generic test-bed allowing an objective quantitative evaluation/benchmarking.**

Any saliency test-bed should be able to ensure objective evaluation of the results, i.e. to be able to accommodate any saliency map methodology, be it from the state of the art or newly advanced.

The test-bed defined in the present thesis is characterized by three main properties: (1) it allows the assessment of the differences between the ground-truth and the saliency-map based results by different criteria, (2) it includes different measure typologies and (3) it grants statistical relevance for the quantitative evaluations.

Consequently, the test-bed is structured at three nested levels, according to the evaluation criteria and to the actual measures and corpora, respectively.

First, several evaluation criteria can be considered. Both *Precision* (defined as the closeness between the saliency map and the fixation map) and *Discriminance* (defined as the difference between the behavior of the saliency map in fixation locations and in random locations) of the saliency models are considered.

Secondly, for any type of evaluation, several measures can be considered. Our assessment is based on two measures of two different types (the KLD, a distribution based metric based on Kullback's Information theory [KUL51], [KUL68] and the AUC, a location based metric according to the Borji's implementation [WEB07]).

Two different corpora are considered and further referred to as: (1) the *reference* corpus organized in by [WEB05] at IRCCyN and (2) the *cross-checking* corpus organized in by [WEB06] at CRCNS. These two corpora are selected thanks to their composition (content diversity and ground-truth availability in compressed format), they representativeness for the saliency community as well as their size. A particular attention is paid to the statistical relevance of the results reported in the thesis. In this respect, we consider:

- for both the *Precision* and the *Discriminance* assessment, all the KLD and AUC values reported in the present thesis are presented by their average, min, max and 95% confidence limits;
- for the *Discriminance* assessment, each experiment (i.e. for each frame in each video sequence)

is repeated 100 times (i.e. for 100 different random location sets) then averaged over all these configurations and all frames in the video sequence;

- for both the *Precision* and the *Discriminance* investigation, the sensitivity of the KLD and AUC measures with respect to the randomness of the video content representing the processed corpus is analyzed.

This test-bed was considered in order to benchmark the MPEG-4 AVC saliency map against three state-of-the-art methods; the HEVC saliency map was benchmarked against the same three state-of-the-art methods and MPEG-4 AVC saliency map. The three state-of-the-art methods were selected according to the following criteria: representatively in the state of the art, the possibility of fair comparison, and the methodological complementarity.

Just for illustration, the ground truth results of the MPEG-4 AVC saliency maps exhibit relative gains in KLD between 60% and 164% and in AUC between 17% and 21% against three models of the state-of-the-art. For the HEVC saliency maps gains in KLD were between 0.01 and 0.40 and in AUC between 0.01 and 0.22 against the same three models of the state-of-the-art.

Applicative validation for robust watermarking

We investigate the benefits of extracting saliency map directly from the compressed stream when designing robust watermarking applications. Actually, **by using the MPEG-4 AVC/HEVC saliency model as a guide in selecting the regions in which the watermark is inserted, gains in transparency (for prescribed data payload and robustness properties) are obtained.**

The applicative validation brings to light transparency gains up to 10dB in PSNR (for prescribed data payload and robustness properties) for the MPEG-4 AVC saliency maps and up to 3dB in PSNR (for prescribed data payload and robustness properties) for the HEVC saliency maps.

Besides its applicative relevance, these results can be also considered as a first step towards an a posteriori validation of the Koch hypothesis: short-time saliency and long-term perceptual masking can be complementary considered in order to increase the visual quality.

As an overall conclusion, the thesis demonstrates that although the MPEG-4 AVC and the HEVC standards do not explicitly rely on any visual saliency principle, its stream syntax elements preserve this property.

Thesis structure

In order to offer a comprehensive methodological and experimental view about the possibility of extracting the salient regions directly from video compressed streams (namely MPEG-4 AVC and HEVC), this thesis is structured as follow.

Chapter I covers the Introduction aspects and is composed of three main parts, related to visual saliency, watermarking and its properties and video coding and redundancies, respectively.

Chapter II is devoted to the state-of-the-art analysis. It is divided into three main parts. Chapter II.1 deals with bottom-up visual saliency extraction and is structured according to a nested, dichotomy: image vs. video and pixel vs. compressed domain. Chapter II.2 gives as an overview about the methodological relationship between watermarking applications and visual saliency. Chapter II.3 relates to the application processing directly the compressed video domain.

Chapter III introduces the methodological and experimental visual saliency extraction directly from the MPEG-4 AVC compressed stream syntax elements. Chapter IV is paired-structured with Chapter III and presents our methodological and experimental results on visual saliency extraction from the HEVC compressed stream syntax elements.

The last Chapter is devoted to concluding remarks and perspectives

The thesis contains five appendixes. Appendix A is devoted to the fusion technique investigation for both MPEG-4 AVC and HEVC visual saliency extraction models. Appendix B gives an overview about the MPEG-4 AVC standard. Appendix C shows the novelty of the HEVC and the principle differences with respect to its predecessor. Appendix D details the numerical experimental values reported in Chapters III, IV and V. Appendix E represents as plots (graphics) the main applicative results of the objective quality evaluation in Chapter III.

Table 1: Visual saliency extraction from video compressed domain: constraints, challenges, current limitations and contributions.

Constraint	Challenge	Current limitations	Contributions
Saliency extraction	• Visual saliency extraction from the compressed stream syntax elements (MPEG-4 AVC and HEVC)	• Visual saliency features are extracted from the uncompressed stream	• Specifying a formalism connecting the human visual system to elementary features of the MPEG-4 AVC and HEVC streams syntax elements • Defining normalization strategies for the obtained maps • Studying the pooling of the static and the dynamic saliency maps into a final compressed stream saliency map
Performance evaluations	• Confrontation to the ground truth: *Precision and Discriminance*	• Limited data sets • Variable and un-coherent evaluation procedures	• Specifying a coherent, unitary test-bed allowing the confrontation of the compressed stream saliency maps to the ground truth: ▪ Evaluation criteria: • *Precision*: the closeness between the saliency map and the fixation map • *Discriminance:* the difference between the behavior of the saliency map in fixation locations and in random locations ▪ Typology of measures: • A distribution based metric: the KLD implemented based on Kullback's Information theory [KUL51], [KUL68] • A location based metric: the AUC implementation made available by Borji [WEB09] ▪ Different corpora: • The *reference* corpus organized by IRCCyN [WEB05] • The *cross-checking* corpus organized by Itti [WEB06] ▪ Statistical relevance • *Precision* and *Discriminance:* experimental values reported alongside with their average, min, max and 95% confidence limits; • *Discriminance:* additional averaging process over repeated random test configurations; • *Precision* and *Discriminance:* assessment of the sensitivity of the measures with the randomness of the visual content.
Applicative integration (watermarking)	• Preserving the application characteristics at a low computational cost	• No saliency validation for compressed domain applications	• Proof of concepts for the integration of the compressed stream saliency map into a watermarking application to guide the watermark insertion • Improving the transparency of the watermarking method, at preserved robustness and data payload properties, while reducing the computational cost

I. Introduction

The present thesis is placed at the confluence of visual saliency, watermarking and video compression. Consequently, the present chapter introduces the basic concepts related to these three realms and identifies two a priori mutual contradictions among and between their concepts.

The first contradiction corresponds to the saliency extraction from the compressed stream. On the one hand, saliency is given by visual singularities in the video content. On the other hand, in order to eliminate the visual redundancy, the compressed streams are no longer expected to feature singularities.

The second contradiction corresponds to saliency guided watermark insertion in the compressed stream. On the one hand, watermarking algorithms consist on inserting the watermark in the imperceptible features of the video. On the other hand, lossy compression schemes try to remove as much as possible the imperceptible data of video.

The thesis will subsequently be structured around these two contradictions.

By its very objective (visual saliency extraction from compressed stream and its subsequent usage in watermarking applications), the present thesis is placed at the confluence of visual saliency, watermarking and video compression. Consequently, the present section will introduce the basic concepts related to these three realms and will state the conceptual relationship among and between them.

I.1. Saliency context

The Human Visual System (HVS) allows us to see, organize and interpret our environment thanks to the complementarities between its major sensory organ (the eye) and the central nervous system (the brain). The eye receives physical stimuli in the form of light and sends those stimuli as bio-electrical signals to the brain, which interprets them as images [WEB09].

I.1.1. Biological basis for visual perception

The human eye is one of the most complicated structures on earth [WEB10]. In order to allow our advanced visual capabilities, it integrates many components, structured on three major layers [WEB08], Figure I-1:

- the sclera, which maintains, protects, and supports the shape of the eye and includes the cornea;
- the choroid, which provides oxygen and nourishment to the eye and includes the pupil, iris, and lens;
- the retina, which allows us to pack images together and includes cones and rods.

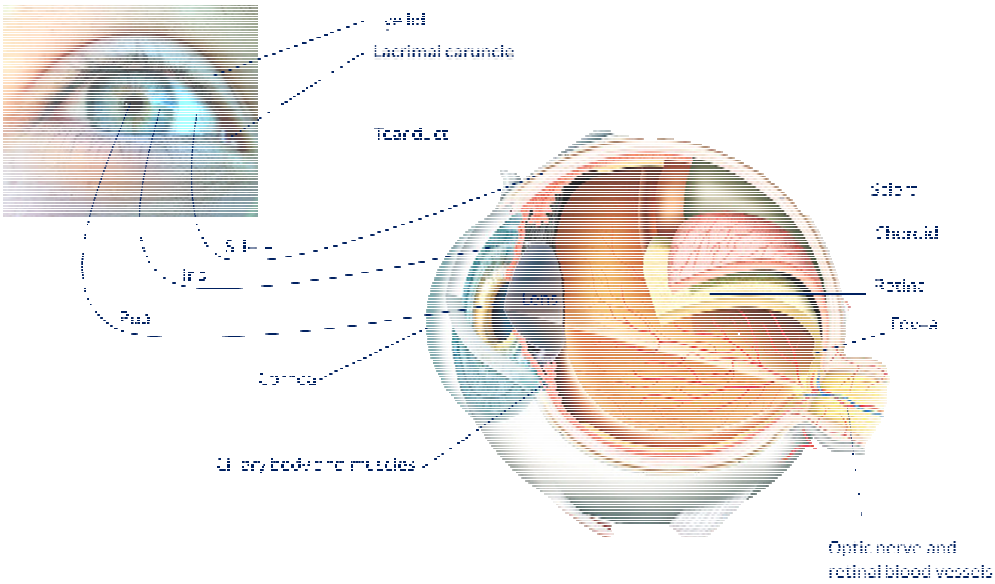


Figure I-1: Human eye anatomy.

The information perceived by the retina is subsequently converted as nerve signals and conducted to the brain by the optic nerves. Then, the visual cortex analyses the received stimulus and develops visual perception.

It is commonly accepted that human vision is neurobiologically based on four different physical realms [TRE80]. First, the rods in retina are sensitive to intensity of the light radiations. Secondly, the cones in retina are sensitive to color contrast (the differences in the wave length corresponding to the spatially adjacent areas). Thirdly, the cortical selective neurons are sensitive to luminance contrast along different orientations (i.e. the difference in the luminance corresponding to the angular directions in a given area). Finally, the magnocellular and koniocellular pathways are sensitive to temporal differences and mainly involved in motion analysis.

However, vision depends not only on the ability to perceive objects assessed by the ratio between their size and the distance between the eye and the screen, but also on other visual, cognitive or semantic factors.

I.1.2. Image processing oriented vision modeling

Modeling the visual perception has gradually become a major issue. Take the example of a high quality video that needs to be distributed and transferred through the Internet. To provide both a smaller version for bandwidth and keep appealing visual quality, the HVS peculiarities should be exploited. In this respect, perceptual masking and saliency maps are two different approaches commonly in use in image/video processing.

Perceptual masking

Perceptual masking is a neurobiological phenomenon occurring when the perception of one stimulus (a spatial frequency, temporal pattern, color composition … etc.) is affected by the presence of another stimulus, called a mask [BEL10].

In image processing, perceptual masking describes the interaction between multiple stimuli; in this respect, the perceptual characteristics of human eye are modeled by three filters denoted by T, L and C and representing the susceptibility artifacts, the luminance perception, and contrast perception, respectively.

The perceptual mask was obtained by first sub-sampling the Noorkami [NOO05] matrix and further adapted to take into consideration the amendments introduced in the compressed stream integer DCT transformation. A value in the matrix represents the visibility threshold, i.e. the maximal value of a distortion added on a pixel (classical) DCT coefficient which is still transparent (imperceptible) for a human observer.

Initially, in order to estimate the behavior of these filters, Peterson [PET93] proposed quantization masking matrix of luminance and color components, depending on the viewing conditions. Subsequently,

an improvement of this model was made by Watson [WAT97] which redefines quantization thresholds taking into consideration the local luminance and the contrast by setting a specific threshold to each one.

Sensitivity to artifacts (T)

The T filter is the sensitivity of the human vision to the artifacts. This filter is defined as the perception of distortions from a well determined threshold.

In each domain and according to each study [WAT97][PET93][AHU92][BEL10], a table has been defined as a filter of the sensitivity to artifacts. This table is defined as a function of some parameters such as image resolution and the distance between the observer and the image. Each value in this table represents the smallest value of the DCT coefficient in a perceptible block (without any noise). Thus, the smaller the value is, the more sensible is our eye to a given frequency.

Luminance perception (L)

The L filter is the luminance perception. It consists of the object perception compared to the luminance average of the entire image [WAT97].

The luminance masking means that, if the average intensity of a block is brighter, a DCT coefficient can be changed by a larger quantity before being noticed. The most brilliant region in a given image can absorb more variation without being noticeable.

Contrast perception (C)

The C filter is the contract perception. It is the perception of an object relative to another object.

The contrast masking, which means the reduction of the visibility of change in a frequency due to the energy present therein, results in a masking thresholds. The final thresholds estimate the amounts by which the individual terms of the DCT block can be changed before resulting in a JND (Just Noticeable Distortion) [WAT97].

Perceptual masking and compressed stream

Thanks to both its methodological and applicative interest, the topic of adapting the perceptual masking to the compressed stream particularities has been of continuous interest during the last two decades.

The study in [WAT97] reports on a masking matrix derived for compression domains based on the classical 8x8 DCT (e.g. JPEG or MPEG-2). This model served as basis for a large variety of compression and watermarking-oriented optimization studies [VER96], [CAB11].

Belhaj *et al.* [BEL10] comes across with a new perceptual mask matched to the MPEG-4 AVC stream; in this respect, the basic [WAT97] model is adapted so as to take into account the three main AVC peculiarities related to the DCT computation: (1) it is no longer applied to 8x8 blocks but to 4x4 blocks; (2) it is computed in integers, and (3) it is no longer applied to pixels but to inter/intra prediction errors.

This model was integrated under a watermarking framework. It points to significant improvement in both transparency (e.g. a gain of 3 dB) and data payload (e.g. a gain of 50%) with respect to the state of the art masking models.

Visual saliency

In its broadest acceptation, a saliency map is a 2D topographic map representing the regions in an image/video on which the human visual system will spontaneously focus.

Actually, the concept of saliency map was introduced by Koch and Ullman [KOC85], as a topographic map representing conspicuousness (salient) locations in the scene. According to Le Callet and Niebur [LEC13], a saliency map is a topographic map of the visual field whose scalar value is the saliency at the respective location.

The saliency property principally and typically arises from contrasts between items (objects, structures, patterns, pixels, etc.) and their neighborhood; additionally, it can also be voluntarily directed to objects of current importance to the observer. The study in [LEC13] defines two different dichotomies of saliency computational models: overt *vs*. covert and bottom-up *vs.* top-down.

Overt vs. covert visual attention

The human visual system is generally attracted by the most relevant areas in a visual scene. This generates a series of fixations called “overt attention”. Using an eye tracker, we can follow the movement of the human eye and draw a “scan path”. By analyzing the details of a given “scan path”, we can have information about the state of the human mind [LEC13].

However, the human eye can also focus in regions other than the center of gaze. As mentioned in [LEC13], it has been discovered that humans are able to fix their attention on peripheral locations, e.g. a car driver fixates the road while simultaneously and covertly monitoring road signs and lights appearing in the retinal periphery. Since this redirection of attention is not immediately noticeable, it is referred to as covert attention.

Bottom-up vs. Top-down

The top-down mechanisms relate to a recognition process influenced by some prior knowledge about the content. Actually, the same visual scene is always differently perceived by different observers. The perception depends on the observer motivation, psychology, and expectations (what they are actually looking for). The personal emotions and history of each observer make the development of a detailed “top-down” model very difficult. The work in [BUS15] explores the “center bias” hypothesis, its limits and underlying proposals. A geometrical cue is considered in case when the central-bias hypothesis does not hold. The proposed visual saliency models are trained based on eye fixations of observers and incorporated into spatio-temporal saliency models. The experimental results are promising: they highlight the necessity of a non-centered geometric saliency cue.

Conversely, the bottom-up mechanism relates to a perception process for automatically detecting saliency, with no prior semantic knowledge about it. The basis of many saliency attention models dates back to Treisman and Glades [TRE80] [TRE88], where the basic visual features and their combination so as to drive the human attention were identified. Koch and Ullman [KOC85] proposed a feed-forward model to fuse these features and introduced the concept of a saliency map (a topographic map that represents conspicuousness locations in the scene).

The first complete implementation and verification of the Koch and Ullman's model was proposed by Itti *et al*. [ITT98]. Since then, a huge variety of approaches with different assumptions for attention modeling has been proposed and has been evaluated against different datasets: according to scholar Google, the Itti's study was cited about 7000 times!

Bottom-up saliency maps are generally based on four different visual characteristics. First, in the spatial domain, three features are to be considered: intensity, color and orientation. Secondly, in the temporal domain, the saliency extracted at the frame level is complemented by the motion information.

Intensity

The human visual system is often attracted by regions with intensity lighter than others. For example, in Figure I-2-a, our vision is first directed to the center which is the lightest region.

Color

The human eye has an extreme low sensitivity to light with wavelengths less than 390 nm and greater than 720 nm [BLA03]. In [ITT98], it is brought to light that the elementary colors are represented in cortex according to a so-called color double-opponent system. In the center of their receptive fields, neurons are excited by one color (e.g., red) and inhibited by another (e.g., green), Figure I-2-b, while the opposite is true in the surrounding areas. Such spatial and chromatic opponency exists for the red/green and yellow/blue color pairs (and, similarly, for their complementary green/red and blue/yellow color pairs).

Orientation

Retinal input is processed in parallel by multiscale low-level feature maps, which detect local spatial discontinuities using simulated center-surround neurons. In fact, there are four neuronal features sensitive to four orientations (0°,45°,90° and 180°) [ITT04]. In Figure I-2-c, we can remark that our vision is attracted by the regions of discontinuity between vertical and horizontal directions.

Motion

When watching videos, human eyes tend to concentrate on moving objects and to ignore the static ones. Actually, HVS is sensitive to regions having the highest motion energy [ZHI09]. In Figure I-2-d, which is extracted from a video sequence, our visual system fixe the fly and try to follow it and somehow overlook the background.

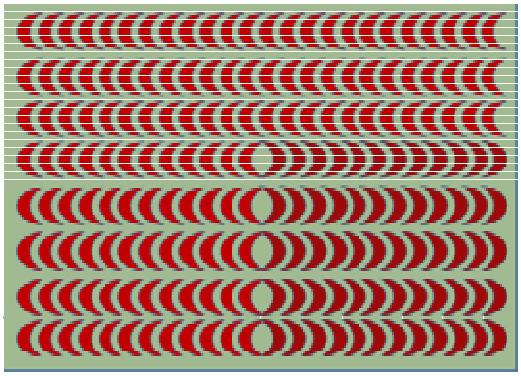
The motion perception is a sophisticated mechanism, implicitly including the time variance. It is also influenced by interactions between the bottom-up and top-down attentions. Just for illustration,

consider the example in which a human looks after regions corresponding to wild animals in a given scene (the target); in such a case, an unexpected, sudden appearance of a non-animal object (the distractor) may inadvertently draw the attention of the subject. In general, in the top-down saliency, motion perception is influenced by the interaction between targets and distractors, especially when both of them have a multimodal distribution and/or significant overlaps exist between them. We cannot speak about distractor in the bottom-up models since we are just extracting the a-priori attractive regions in a video content.

Smooth pursuit eye movements allow the HVS to closely follow a moving object. Pursuit eye movements are initiated within 90-150 ms, while typical latencies for voluntary saccades are in the order of 200-250 ms. While for top-down saliency model the pursuit eye movements is an explicit research topic, for bottom-up models it acts implicitly.

a) Intensity contrast

b) Color contrast



c) Orientation discontinuity

d) Motion contrast

Figure I-2: Visual saliency features.

Perceptual masking vs. Visual saliency

Generally, the visual saliency and the perceptual masking are considered as two different, quite unrelated approaches. This can be justified by the a priori conceptual contradiction in their very principles. On the one hand, perceptual masking related to objects/regions which are somewhat neglected by the HVS. On the other hand, saliency map highlights the object/regions to which the human eye will spontaneously look at.

However, the early Koch work brings to light the saliency is an intrinsic time related behavior; consequently, when considering a longer analysis period, we can expect some synergies between saliency and masking to be established.

To the best of our knowledge, the first studies combining visual saliency and perceptual masking are the study in [AMM14] (see Chapters III in the present thesis) and the study in [CAO15]. The main contribution of [CAO15] consists in choosing the least salient and sensitive regions for HVS to embed the secret data. Experimental results demonstrate that such an approach outperforms in terms of quantity of inserted information and/or image quality four existing steganographic approaches.

From the methodological point of view, the present thesis relates to the overt, bottom-up visual saliency extraction from the compressed stream. However, in the watermarking applicative perspectives, saliency / perceptual masking synergies will be also investigated.

I.2. Watermarking context

Digital watermarking can be defined as the process of imperceptibly embedding a pattern of information into a cover digital content (image, audio, video, *etc.*) [COX02] [MIT07], see Figure I-3. The insertion of the mark is always controlled by some secret information referred to as a key. While the key should be kept secret (*i.e.* known only by the owner), the embedded information and even the embedding method can be public. Once watermarked, the watermarked data can be transmitted and/or stored in a hostile environment, *i.e.* in an environment where changes attempting to remove the watermark are likely to occur. The subsequent mark detection can be used in a wide area of applications such as intellectual property right preservation, content integrity verification, piracy tracking or broadcast monitoring.

From the functional point of view, any watermarking procedure is evaluated according at least three essential properties, namely transparency, robustness and data payload:

- The **data payload** is the quantity of information that is inserted into the host document. It should be high enough so as to allow the owner to be identified (e.g. 64 bits would correspond to an ISBN number). Additional data could bring information about the document buyer, vendor, date and time of purchase, etc.
- The **transparency** refers to the imperceptibility of the watermark in the document. This may signify either that the user is not distributed by the artifacts induced by the watermark in the host document or that the user cannot identify any difference between the marked and the unmarked document. From the conceptual point of view, the transparency property relates to the possibility of exploiting the visual redundancy existing in the host data so as to hide messages.
- The **robustness** refers to the ability to detect the watermark after applying some signal operations on the marked document, such as spatial filtering and loss compression scanning, etc. The copyright protection requires very high robustness, as attacks are very likely to appear. As a limit case, the mark would withstand any attack that does not render the document unusable. The robustness is generally assessed by the probability of error at the detection.

A good watermarking system must reach the trade-off between a large data payload, a good transparency and a strong robustness. In our work, we are particularly interested in the transparency of digital watermarking while studying the human visual system and exploiting the saliency map.

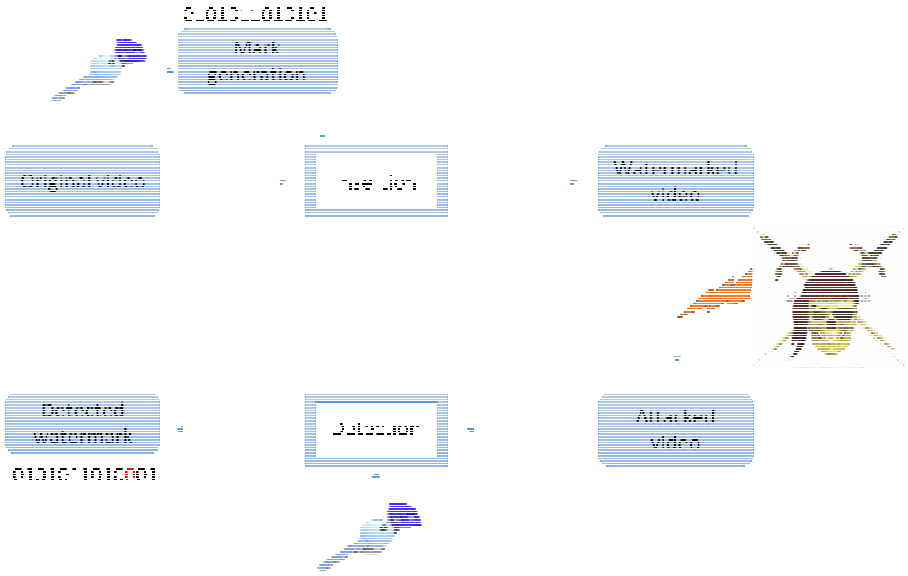


Figure I-3: General scheme of watermarking approach.

The watermarking schemas are commonly divided into two main classes, namely Spread Spectrum (SS) and Side Information (SI).

The SS systems have been already deployed in telecommunication applications (*e.g.* CDMA), by providing a preferment solution for very low power signal transmission over noisy channel [COX97]. Consequently, an SS based watermarking method spreads the mark across the host signal by creating redundancy, requiring a much larger bandwidth than strictly necessary. In practice this approach remains robust against attacks, while offering limited data payload [MIT07].

The SI principle [SHA58], [EGG03], [CHE98] stipulates that a given noise channel known at the transmitter and unknown at the receiver would not decrease the channel capacity (the maximum amount of information which can be theoretically transmitted). Thus, the original document should no longer be considered as a constraint for to the watermark detection. Consequently, the side information watermarking is a priori optimal from the data payload point of view (under fixed transparency and robustness constraints). However, in practice, the methods following this approach feature very weak robustness in spite of a very high quantity of embedded information.

As it can be intuitively deduced, under the watermarking framework, the saliency principles are expected to be used as a transparency optimization enabler (for fixed data payload and robustness

constraints). The principle is to use saliency as guide for increasing the transparency, i.e. decreasing the impact of the artifacts perceived by the HVS.

I.3. Video coding & redundancy

During the last three decades, image and video coding has never stopped evolving: from MPEG-1 (of particular interest for video CD) and MPEG-2 (considered for video DVD), to MPEG-4 AVC (a.k.a. H.264) and the latest HEVC (a.k.a H.265), each generation of compression standard increased by a factor of at least 2 the bandwidth reduction for a constant video quality [RIC03], [SUL12].

The most generic representation for an encoder is given by a four-step chain, Figure I-4: Prediction P, Transformation T, Quantization Q and arithmetic (entropic) coding E.

The Prediction is designed so as to eliminate the spatial (intra-prediction) and temporal (inter-prediction) redundancy. The Transformation is meant to represent data as uncorrelated (separated into components with a minimum interdependence) and compacted (energy concentration on a small number of coefficients) information. Quantization is then applied and some of the information is lost. The final phase of the compression chain is the entropy coding (lossless). Of course, differences exist among and between the ways in which each and every video encoder implements the four above-mentioned operations.

However, any codec is meant to remove both the visual redundancy (i.e. to process the original video content so as to remove visual insignificant information) and data redundancy (in the sense of the Shannon's first theorem). Consequently, the compressed stream syntax elements are expected to be uniformly distributed (or, at least, their first/second order statistics) and to avoid any singularities.

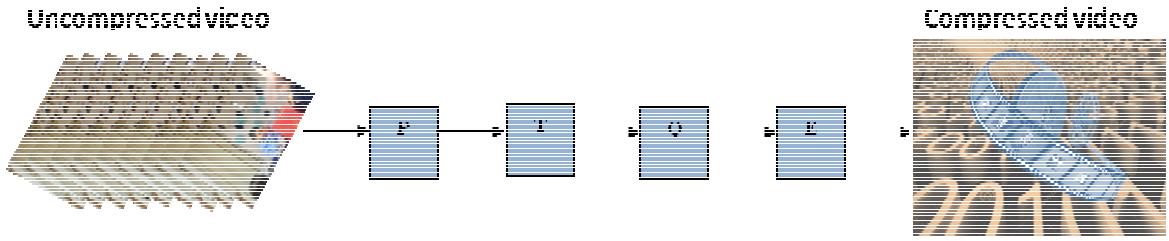


Figure I-4: MPEG-4 AVC/HEVC compression chain.

The MPEG-4 AVC/HEVC video sequences are structured into Groups of Pictures (GOP). A GOP is constructed by an *I* (Intra frame) and by a number of successive *P* and *B* frames (Predicted and Bidirectional predicted, respectively). The *I* frame describes a full image coded independently, containing only references to itself. The unidirectional predicted frames *P* use one or more previously encoded frames (of *I* and *P* types) as reference for picture encoding/decoding. The bidirectional predicted frames *B* consider in their computation both forward and backward reference frames, be they of *I*, *P* or *B* types. Details related to the MPEG-4 AVC/HEVC stream syntax can be found in Appendix B/C.

I.4. Conclusion

The aim of the present Introductive section is to bring to light the basic concepts underlying the present thesis, namely visual saliency, watermarking and compressed streams.

First, a saliency map is a topographically arranged map that highlights regions of interest (singularities) in a corresponding visual scene. It represents the conspicuity at every location in the visual field by a scalar quantity, based on the spatio-temporal distribution of saliency. For still images, the static saliency map is composed of three feature maps: intensity map, color map and orientation map. These three maps correspond to different physical realms. The intensity map corresponds to the sensibility of the retina to the intensity of the light. The color map is related to the sensibility to the colors composing in each image (r, g, and b). The orientation map is given by the four orientations (0, 45°, 90°, 135°) for which neuronal sensitive features exist in the human visual system. For the video, the static saliency map should be combined with a motion saliency map, in order to take into consideration, the sensibility of the human eye to the moving regions.

Secondly, digital watermarking can be defined as the process of imperceptibly and persistently embedding a pattern of information into a cover digital content (image, audio, video, *etc.*). A good watermarking system must reach the trade-off between a large data payload, a good transparency and a strong robustness. In other words, we are interested in trading the visual redundancy existing in the host data for persistently hiding the watermark.

Finally, the goal of any video compression standard is to eliminate the video redundancy. Both the visual redundancy (i.e. to process the original video content so as to remove visual insignificant information) and data redundancy (in the sense of the Shannon's first theorem) are concerned by the encoding schemes.

These three main characteristics above bring to light that the present thesis should face two a priori conceptual contradictions among and between visual saliency, watermarking and compressed streams.

The first contradiction corresponds to the saliency extraction from the compressed stream. On the one hand, saliency is given by visual singularities in the video content. On the other hand, in order to eliminate the visual redundancy, the compressed streams are no longer expected to feature singularities. The second contradiction corresponds to watermark insertion in the compressed stream. On the one hand, watermarking algorithms consists on inserting the watermark in the imperceptible (non-salient) features of the video. On the other hand, lossy compression schemes try to remove as much as possible the imperceptible data of video.

Consequently, the thesis first studies weather the visual saliency can be directly bridged to stream syntax elements or, on the contrarily, complex decoding and post-processing operations are required to do so.

The thesis also aims at studying the practical benefit of the compressed domain saliency extraction, for the particular case of video watermarking. The saliency is expected to act as an optimization tool, allowing the transparency to be increased (for prescribed quantity of inserted information and robustness against attacks) while decreasing the overall computational complexity. However, the

underlying proof of concepts is still missing and there is no a priori hint about the extent of such a behavior.

II. State of the art

This chapter is structured into three parts, related to the visual saliency extraction, to the visual saliency as a watermarking optimization tool and to the direct compressed video stream processing, respectively.

This three-folded state of the art analysis brings to light that:

- *Automatic visual saliency detection is as a particular research field. Its fundamental (neuro-biological) background is represented by the early works of Treisman et al., advancing the integration theory for the human visual system and by Koch et al. who brought to light a time selectivity mechanism in the human attention. From the methodological point of view, all the studies published in the literature follow an inherent experimental approach: some hypotheses about how these neuro-biological characteristics can be (automatically) computed from the visual content are first formulated and then demonstrated through experiments. In this respect, maybe the most relevant example is the seminal work of Itti [ITT98]. While the large majority of studies generally converge in the type of the main methodological steps (extracting individual intensity, color, orientation and motion maps and subsequently fusion them at spatial and spatio-temporal levels), lot of divergences still remains in their definition, assessment (ground-truth vs. applicative, objective vs. subjective evaluation, composition of corpora, type of measures, etc.). Moreover, no study related to the saliency extraction in the compressed domain, i.e. in-between the Quantization and Entropic coding steps has been identified.*
- *While the relationship between saliency and watermarking shows different promising results and exploring the ROI (regions of interest) can be benefic for each of the main watermarking properties, no study on the trade-off between watermark embedding and the visual saliency extraction in compressed domain has been identified.*
- *Today, image/video processing directly in the compressed stream becomes more a necessity rather than an option: just for example, fingerprinting, image retargeting and detecting moving object can benefit from such an approach. However, the integration of visual saliency extraction directly from compressed domain in such applications is not yet studied.*

Consequently, in this thesis, we take the challenge of extracting the saliency map in the compressed domain in order to guide the watermark insertion in a compressed stream watermarking application (both MPEG-4 AVC and HEVC), with minimal decoding operations.

This Chapter is structured according to the three main research fields underlying the thesis: visual saliency extraction, the usage of saliency as an optimization tool in watermarking and the direct compressed stream processing applications.

II.1. Bottom-up visual saliency models

As defined in the Introduction section, a saliency map is a 2D topographic map representing the regions in an image/video on which the human visual system will spontaneously focus.

Under this framework, the present thesis belongs to the overt, bottom-up (see [LEC13] and the Introduction chapter) saliency research field, which is already covered by about 20 years of very rich and heterogeneous scientific publications. As an exhaustive state of the art study becomes today practically impossible, we limit ourselves to 18 publications, staring from the seminal Itti's work in 1998: [ITT98], [BRU05], [HAR06], [LEM06], [HOU07], [GOF10], [MUR11], [CHE13], [ITT05], [ZHA06], [LEM07], [HOU08], [SEO09], [MAR09], [GUO10], [GOF12], [FAN12], and [FAN14].

The presentation is structured at several incremental levels: image versus video and pixel-based versus compressed-based saliency models, as illustrated in Figure II–1.

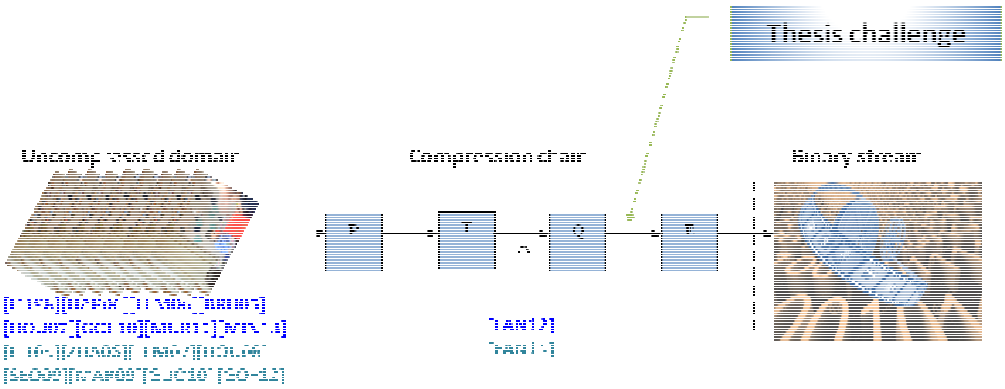


Figure II-1: Domains of bottom-up saliency detection models; in blue: studies related to still images; in green: studies related to videos. P, T, Q, E stand for Prediction, Transformation, Quantification and Encoding, respectively.

II.1.1. Image saliency map

As a general direction in the state-of-the-art studies, the saliency map comprises spatial (static 2D) and temporal (motion) information. The spatial model is computed at the frame level as the image saliency map. The temporal model is based on the motion (difference between successive frames).

In order to extract still image saliency, Itti *et al.* [ITT98] consider 9 image scales obtained through a dyadic Gaussian pyramid decomposition. First, visual features related to intensity, color, and orientation are extracted at the multiple image scales while taking into account the center-surround differences between a center (finer scale) and a surround (coarser scale). Secondly, three saliency maps corresponding to the above three features are created by a strategy based on iterative localized interactions combination. Finally, these three maps are averaged so as to generate the still image saliency map. The general architecture of this model was illustrated by authors in Figure II-2, where they represented its different computation steps. The experiments consider 258 images. The spatial frequency content (SFC) is computed on two cases: (1) on the locations detected as salient and (2) on the whole image. It is shown that the ratio of the SFC computed on the salient location to the average SFC belongs to the (1.6; 2.5) interval (according to the level of the decomposition).

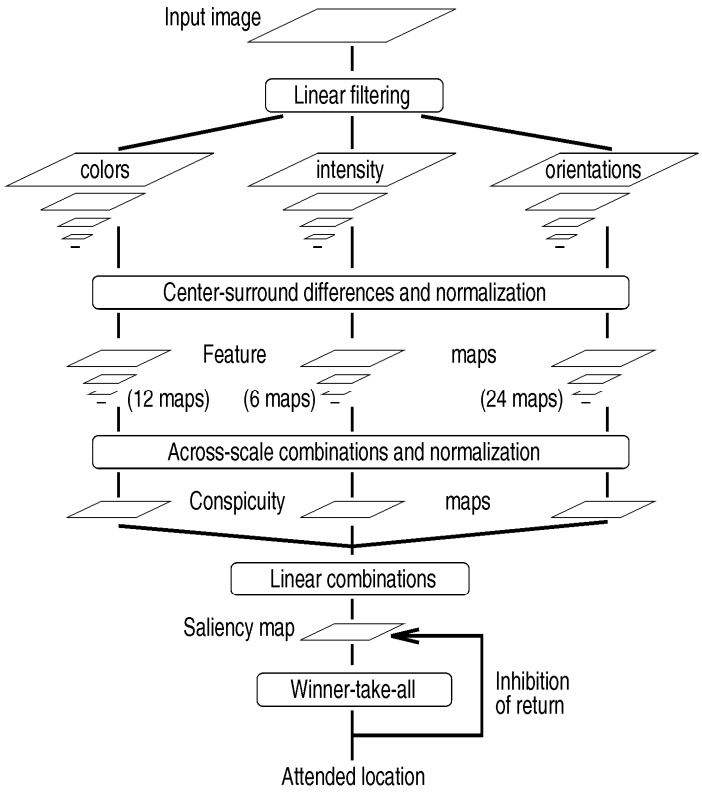


Figure II-2: Synopsis of Itti's model [ITT98]: the saliency map is obtained by a multi-scale extraction model consisting on three feature extraction, normalization and fusion of the elementary maps.

Bruce and Tsotsos [BRU05] opt to determine saliency by quantifying the Shannon's self-information[1] of each local image patch. The principle is to consider the visual saliency is determined by a sparse

[1] A probabilistic theory allowing to quantify the average information content of a set of messages, whose computer coding satisfies a precise statistical distribution [RIC45].

representation of the image statistics a priori learned by the brain. The first step in saliency computation consists in dividing the original image in 7x7 RGB patches and in performing the related ICA (Independent Component Analysis). For a given image, an estimate of the distribution of each basis coefficient is learned across the entire image through non-parametric density estimation. The probability of observing the RGB values corresponding to a patch centered at any image location is then evaluated by independently considering the likelihood of each corresponding basis coefficient. Figure II-3 shows the framework of this model. A validation based on comparison with Itti's model [ITT98] reveals the efficacy of this model with an AUC=0.7288.

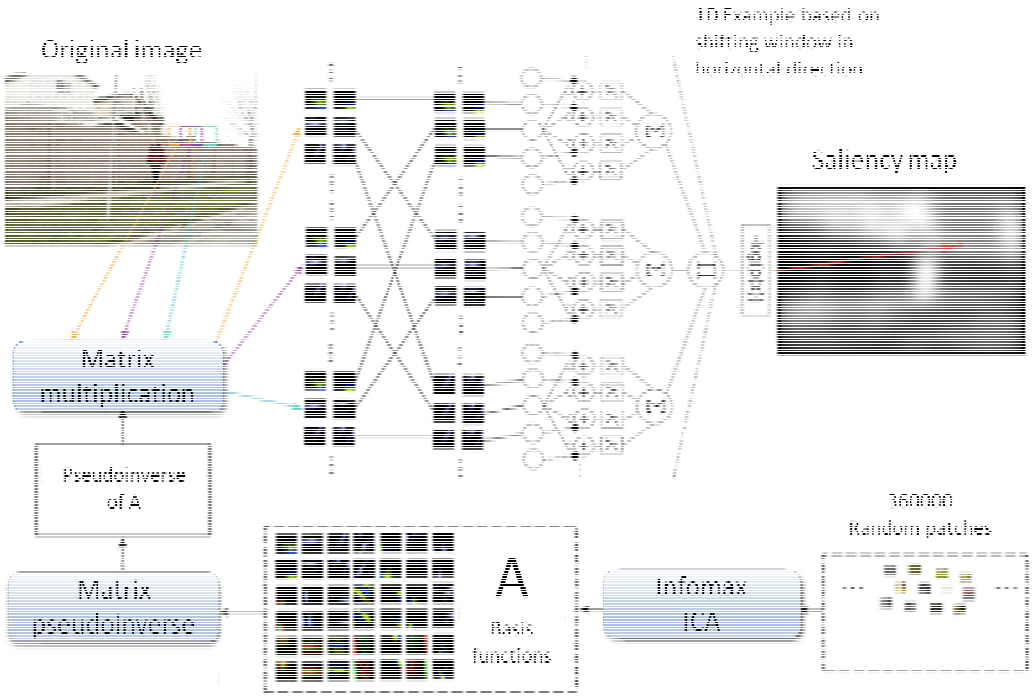


Figure II-3: Saliency extraction based on the Shannon's self-information [BRU05]: the visual saliency is determined by a sparse representation of the image statistics, learned from the prior knowledge of the brain.

The Harel's model [HAR06] is based on a three step approach, Figure II-4. First, elementary feature maps are extracted by common linear/non-linear image filtering techniques. Secondly, for each feature, an activation map is computed so as to detect the locations with unusual (singular) behavior (i.e. inhomogeneous locations). Thirdly, the elementary maps are pooled according to the activation maps, through a Markovian-based weighted summation. The experimental results are obtained on 108 images and correspond to calculate the ROC area between the human fixation and the saliency map for each graph used to activate and normalize maps; it is shown that the obtained values validates all the end-to-end algorithms by a value varying between 0.96 and 0.98.

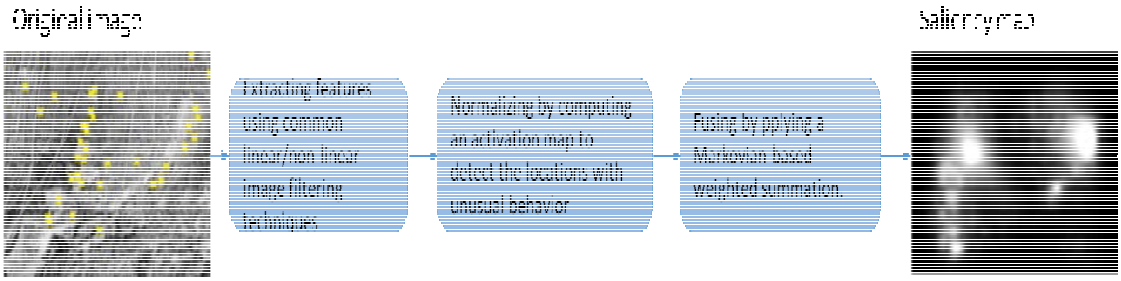


Figure II-4: Computation steps of Harel's model [HAR06]: the saliency is determined by extracting features, normalising, then fusing the elementary maps.

In [LEM06], Le Meur *et al.* design a biologically inspired model which automatically detect the most relevant parts of the picture based on different HVS properties such us contrast sensitivity functions, perceptual decomposition, visual masking, and center-surround interactions. These relevant regions are subsequently independently normalized to a common scale and combined based on a coherent psycho-visual space. An illustration of the computation steps is made in Figure II-5. The experimental results, obtained on 10 images, evaluate two objective metrics: the linear correlation coefficient (CC) and the Kullback Leibler divergence (KLD) between the human fixation and the saliency map; average values CC=0.71 and KLD=0.46 are obtained.

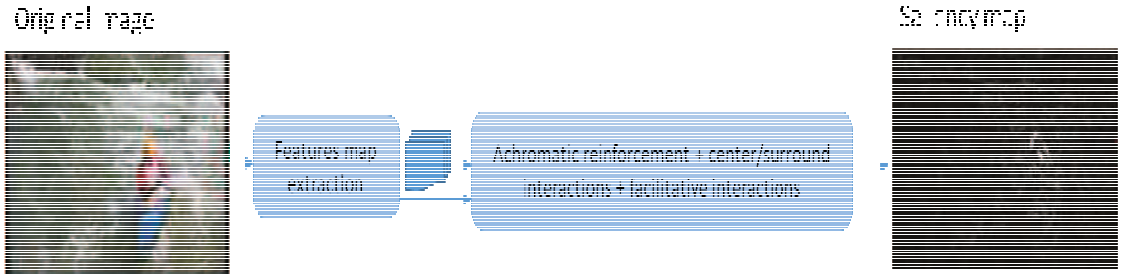


Figure II-5: Flowchart of the biologically inspired model advanced in [LEM06].

Hou and Zhang [HOU07] present a method for the natural images saliency detection by analyzing the log-spectrum of each input image, Figure II-6. The principle is to consider that the singularities in the image (i.e. the salient locations) are given by spectral residuals (computed on a log-spectrum basis). Consequently, the spectral residuals are first extracted, and then subjected to an Inverse Fourier Transform and to some post-processing operations (like Gaussian filtering, thresholding, etc). The experiments consider 4 naïve observers which compare 62 natural images to their related saliency maps. The subjective Hit Rate and the False Alarm Rate are computed and compared to the values reported by [ITT98]. The advanced method outperforms [ITT98] on both Hit Rate and the False Alarm Rate. Additionally, a significant increase in the processing speed (by a factor of 15) is reported.

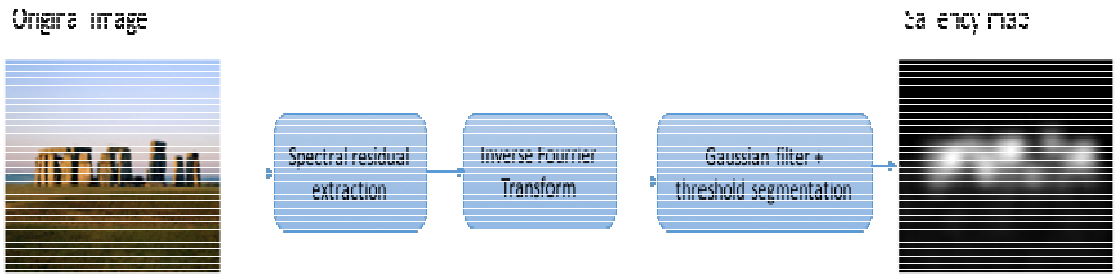


Figure II-6: Saliency map computation flowchart: extracting visual saliency by exploiting the singularities in the spectral residual.

Goferman *et al.* [GOF10] define a new type of saliency, context aware saliency, which aims at detecting the image regions that represent the scene. This model is based on four principles: (1) Local low-level considerations, including factors such as contrast and color, (2) Global considerations, which suppress frequently occurring features, while maintaining features that deviate from the norm, (3) Visual organization rules, which state that visual forms may possess one or several centers of gravity about which the form is organized and (4) High-level factors, such as human faces. The algorithm of this detection model consists on establishing synergies among and between all these principles, Figure II-7. First, a single-scale local-global saliency is defined according to the principles (1)-(3). Then, the saliency is enhanced by using multiple scale filtering and visual coherency rules. Finally, principle (4) is implemented as a post-processing operation. This approach is evaluated on used the database provided by [HOU07], by calculating the AUC: it is thus proved that two other state of the art models [HOU07] [WAL06] are outperformed. The method is also validated under the image retargeting applicative framework.

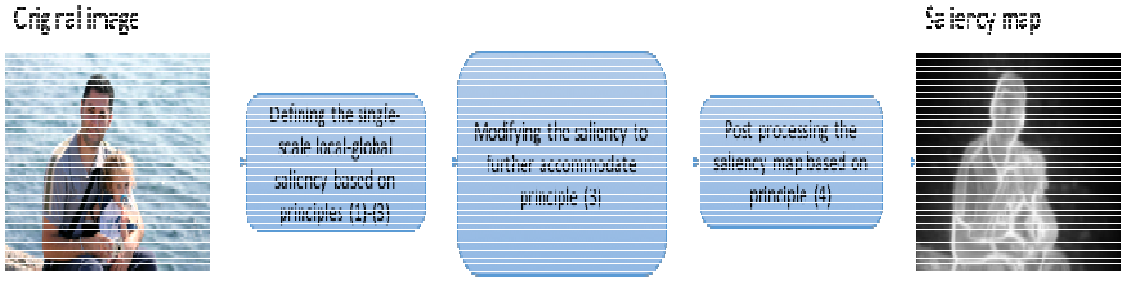


Figure II-7: A context aware saliency model: the saliency is enhanced by using multiple scale filtering and visual coherency rules [GOF10].

Murray *et al.* [MUR11] exploit low-level, biologically inspired representation predicting color appearance phenomena, see Figure II-8. First, the basic color-opponent and luminance channels are modeled by Gabor-based wavelet multi-scale decomposition. Secondly, the inhibition mechanism is modeled by filters whose parameters are estimated through a Gaussian Mixture strategy. Finally, the integration of the information extracted at different scales is achieved by a non-linear formula based on the Extended Contrast Sensitivity Function [MUL85]. The experiments are performed on two ground truth data-bases [BRU05] [JUD09] and consist in computing the KLD and AUC; the obtained values are KLD=0.426,

AUC=0.701 and KLD=0.278, AUC=0.664, respectively. This model outperforms 5 state-of-the-art models [BRU05], [SEO09], [ZHA08], [ITT98], and [GAO08].

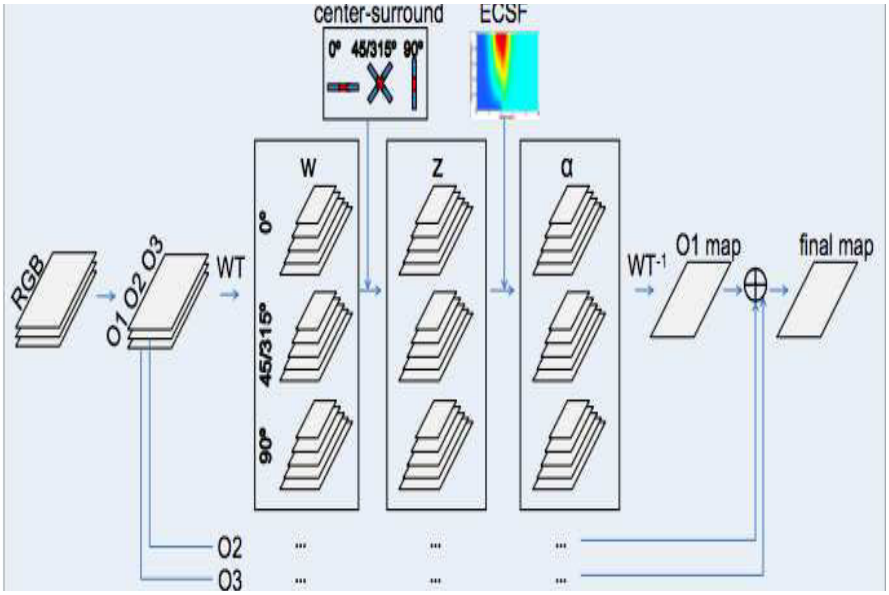


Figure II-8: Principle of the saliency approach [MUR11]: the saliency is obtained according to a biologically inspired representation based on predicting color appearance.

Cheng *et al.* [CHE13] present a global components representation which decomposes the image into large scale perceptually homogeneous elements. The representation considers both appearance similarity and spatial overlap, leading to a decomposition that better approximates the semantic regions (Figure II-9) in images and that can be used for reliable global saliency cues estimation. The nature of the hierarchical indexing mechanism of these representations allows efficient global saliency cue estimation, with complexity linear in the number of image pixels, resulting in high quality full resolution saliency maps. Experimental results on a public available dataset (1000 images) show that their salient object region detection results are 25% better than the previous best results (compared against 17 alternative state of the-art-methods [ITT98], [MA03], [HOU07], [GOF10], [HAR06], [PER12], [CHE11], [MUR11], [SEO09], [ZHA08], [RAH10], [BRU09], [DUA11], [ZHA06], [ACH08], [ACH09], and [ACH10]), in terms of Mean Absolute Error (MAE), while also being faster.

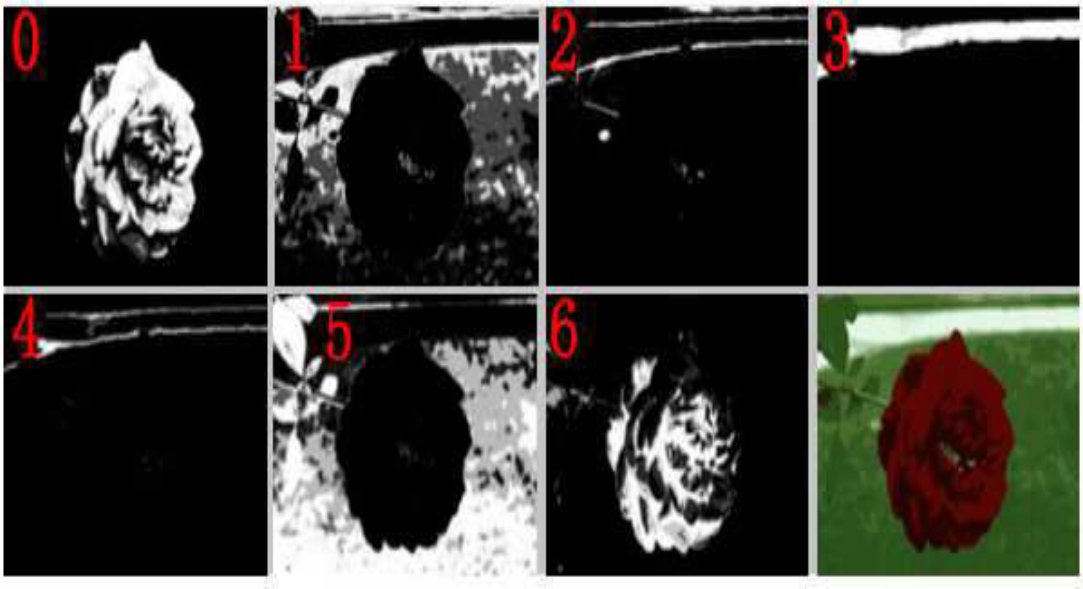


Figure II-9: Soft image abstraction and decomposition into perceptually homogenous regions [CHE13]: the saliency map is extracted by considering both appearance similarity and spatial overlap.

In order to extract the saliency maps, Fang *et al.* [FAN12] no longer consider pixel representation of the image but a transformed domain related to the JPEG compression. The model is presented in Figure II-10. The features (intensity, color, texture) are directly extracted from the 8×8 JPEG discrete cosine transform (DCT). In order to extract the intensity and color maps, the JPEG native YCrCb transformed color space is translated into the RGB transformed color space, and then, the intensity and color features are extracted according to the Itti's principles (the Y channel represents the luminance component, while Cr and Cb represent the chroma components). The texture feature is given by the AC coefficient in YCrCb color space. The global saliency map is obtained through a so-called coherent normalized-based fusion method, i.e. through a weighted addition of the elementary maps. The experimental results are obtained on 1000 images and correspond to the AUC (Area Under the ROC Curve) between the human fixation and the saliency maps; an average AUC value of 0.93 is obtained and shown to be larger than the values corresponding to three other state of the art studies [HOU07], [ACH09], and [ITT98].

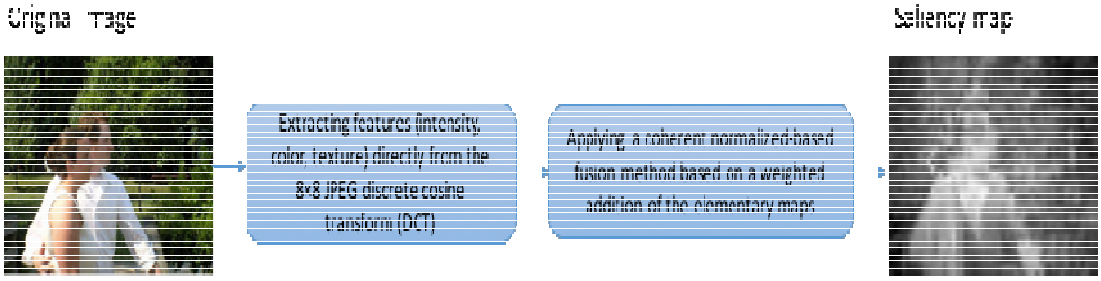


Figure II-10: Saliency map computation steps [FAN12]: the saliency map is obtained, in the transformed domain of the JPEG compression, through a so-called coherent normalized-based fusion.

II.1.2. Video saliency map

As a general direction in the state-of-the-art studies, the spatial (static 2D) saliency extracted at the frame level is complemented with temporal (motion) information.

Rather than being directly focused on visual saliency in video, Itti *et al.* [ITT05] deal with a broader concept, namely the surprise. First, the study provides a formal mathematical model for the surprise elicited by a visual stimulus or event. In this respect, a Bayesian framework is considered. The background information of an observer is represented by its prior probability distribution over a given model. Starting from this prior distribution of beliefs, the fundamental effect of a new data observation D on the observer is to change the prior distribution in the posterior distribution via Bayes theory. The new data observation D carries no surprise if the posterior distribution is identical to the prior one. Conversely, D is surprising if the posterior distribution differs from the prior distribution. The same data may carry different amount of surprise for different observers, or even for the same observer taken at different times. Secondly, the surprise is connected to the visual saliency through experiments considering both TV and video games content. It is thus brought to light that more than 72% of human saliency is connected to the surprise.

Zhai *et al.* [ZHA06] design an attention detection model, Figure II-11, highlighting regions that jointly correspond to interesting objects and actions. The static map is computed based on the color contrast (extracted at the color histogram level) while the motion map is computed based on the motion contrast between successive frames. These two elementary maps are pooled through a dynamic averaging technique (the temporal attention is dominant over the spatial attention when large motion contrast exists and vice versa). The experimental results are obtained on 9 video sequences and correspond to subjective evaluations: a panel of 5 observers watches these 9 videos together with their saliency maps. They assessed the concordance between the saliency map and their own intuition about saliency, by granting three quality marks: Good, Poor and Failed. The results show that the Good label is the most voted (with an average of 0.77) while the Failed label is granted with a frequency of 0.08.

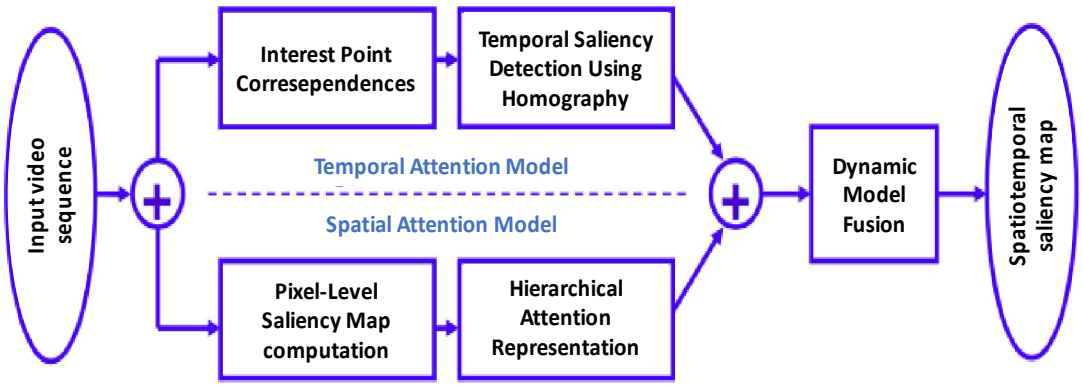


Figure II-11: Workflow of the saliency model [ZHA06]: the saliency map is obtained through a dynamic fusion of the static and the temporal attention model.

Le Meur *et al.* [LEM07] consider the use of center surround filters (CSF) in order to obtain one achromatic and two chromatic saliency maps which are subsequently pooled by a weighted average operation to obtain the spatial saliency map. In Figure II-12, we modified the flowchart presented in [LEM07] in order to obtain a simple illustration of the proposed model. The temporal map is calculated as the predicted relative motion (the relative motion weighted by its median value). The spatio-temporal saliency map is obtained as a weighted summation and product of the individual maps. The experimental results are obtained on 7 video sequences; the CC, the KLD, the cumulative probability and the ROC curve between the saliency map and the density fixation map are calculated. It is shown that regardless of the considered metric, the proposed model shows significant improvement over the selected benchmarking models. Just for illustration, CC=0.41, KLD=19.21.

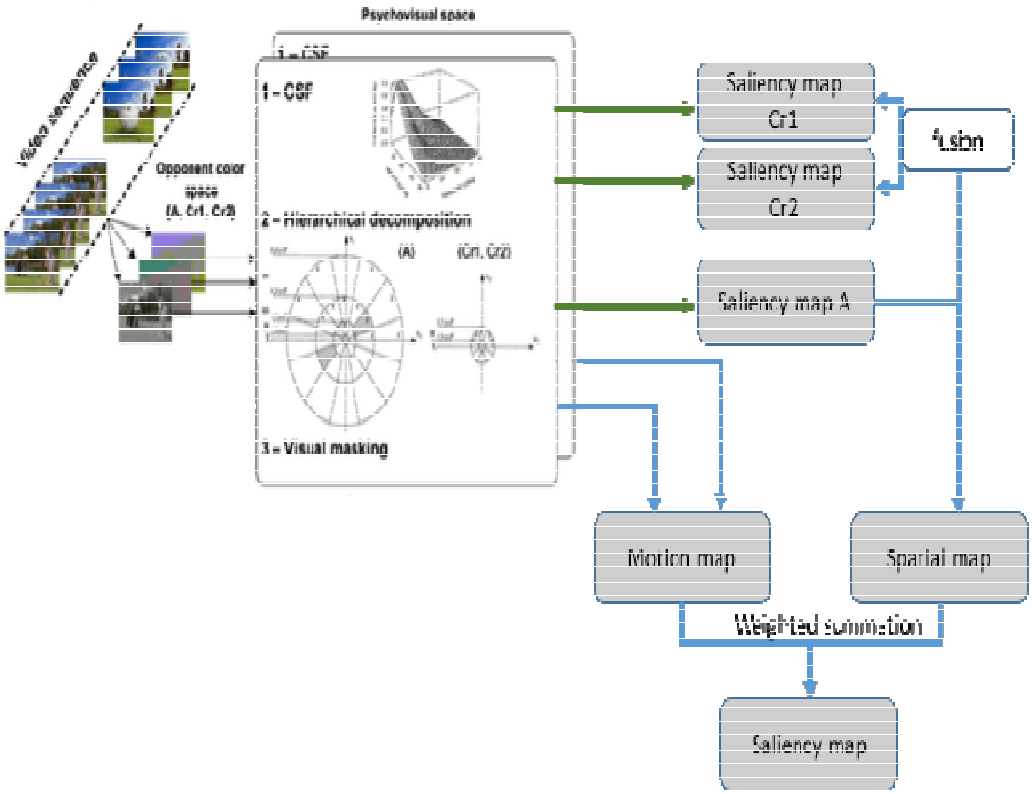


Figure II-12: Flowchart of the proposed model [LEM07]: the saliency map is the result of a weighted average operation of achromatic and two chromatic saliency maps.

Motivated by the sparse coding strategy discovered in primary visual cortex, Hou *et al.* [HOU08] represent in its model, Figure II-13, an image patch as a linear combination of sparse coding basis functions, which are referred to as features. The activity ratio of a feature (static or dynamic) is its average response to image patches over time and space. Each feature is then evaluated according to its Incremental Coding Length (ICL) which is defined as the ensemble's entropy gain during the activity

increment of the feature. According to the general principle of predictive coding, the energy is distributed to features according to their ICL contribution. Finally, the global saliency is obtained by summing up the activity of all features at that region. The experimental results are obtained on 120 images and 1 video. The differences between the ground truth and the obtained saliency map are expressed by computing the AUC for still images and the KLD for video sequences; average values of 0.79 and 0.54 are obtained, respectively (which outperform three other state of the art saliency detection models [ITT98], [BRU05], and [GAO07]).

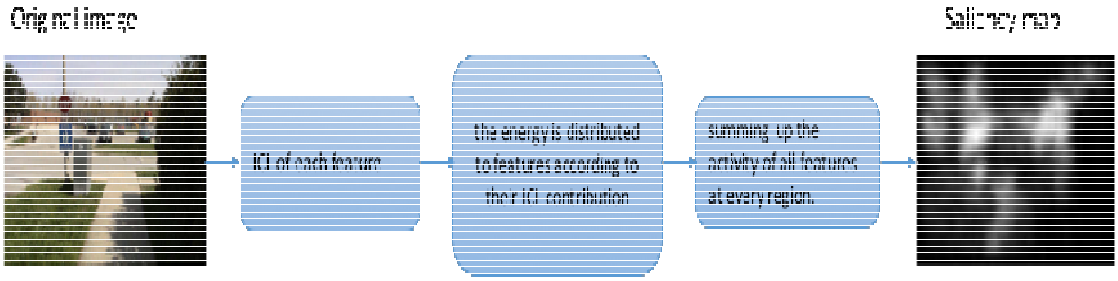


Figure II-13: Incremental coding length model's different steps [HOU08]: the saliency extraction model is based on the incremental coding length of each feature.

Seo *et al.* [SEO09] present a two-folded study on saliency detection, see Figure II-14. First, the local regression kernels are used as features which capture the underlying local structure of the exceeding data. Second, nonparametric kernel density estimation is considered for such features. The final result is a so-called "self-resemblance" saliency measure, i.e. a measure indicating the likelihood of saliency in a given location. The experimental results are obtained on the corpus from [BRU05]: the saliency maps are compared to the ground truth by calculating KLD and AUC. The average AUC value is 0.67 and the average KLD value is 0.34, which outperform 4 other state of the art models [ITT05], [ZHA09], [BRU05], and [ZHA08].

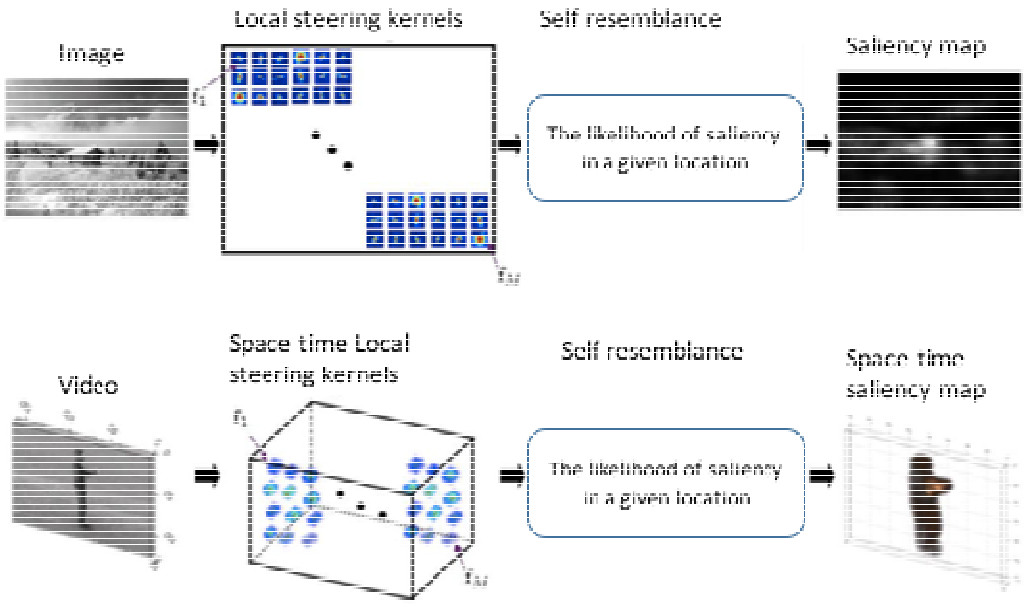


Figure II-14: Illustration of image/video saliency detection model [SEO09]: the saliency map is obtained by applying the self resemblance indicating the likelihood of saliency in a given location.

Marat *et al.* [MAR09] propose a video summarization based on a visual attention model (Figure II-15). The attention model was computed on two parallel ways: (1) the static way highlights objects based on textured and contrasted regions in each frame. The static saliency map is normalized and obtained after applying a retinal filter, a Gabor filter then a temporal filter (2) the dynamic way that gives information about moving objects. The dynamic saliency map is normalized and obtained after applying the temporal filter to the motion difference frame. This summarization method has been tested on three videos of different length and content. A harmonic average between Precision and Recall rates is used as evaluation measure and referred to the F1 score. The F1 value of this method outperforms the results of the random summary and the summary selecting one frame in the middle of each shot.

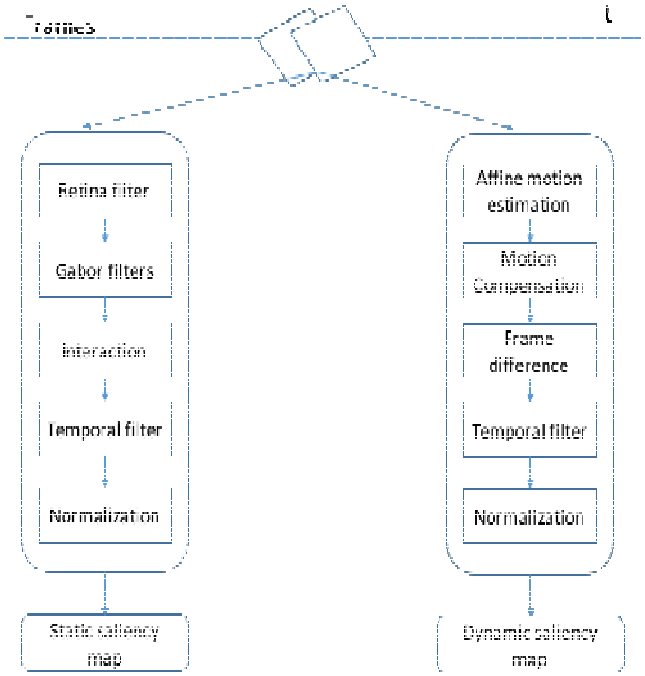


Figure II-15: Saliency computation graph [MAR09]: the attention model was computed on two parallel ways: the static way and the dynamic way.

Guo *et al.* [GUO10] propose a multiresolution spatiotemporal saliency detection model based on the Phase spectrum of Quaternion Fourier Transform (PQFT). Each frame is considered as a composition of three components (intensity – the average of r, g and b channels, color – the difference between color pairs (red/green, blue/yellow), and motion – difference between successive frames) and a quaternion representation is associated to it. The final spatiotemporal saliency map is obtained by processing these components and by fusing them according to a QFT formula. Figure II-16 illustrates the different computation steps of this model. The experimental results are obtained on 100 natural images and 1 video (988 frames): the average AUC (between the human fixation and the saliency map) value is 0.83, which outperforms 4 state of the art models [ITT98], [ITT00], [HOU07], and [HAR06].

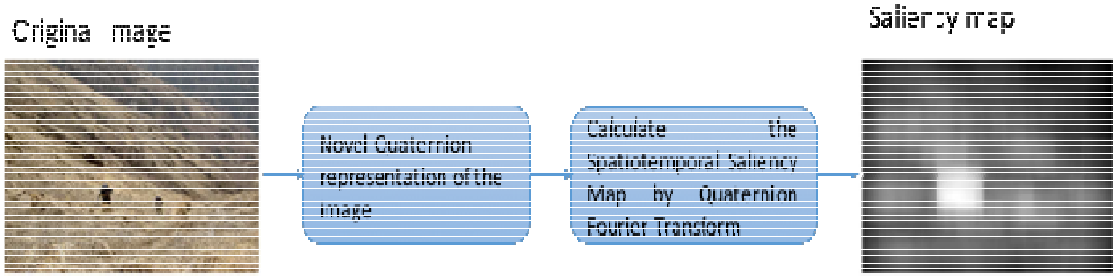


Figure II-16: Multiresolution spatiotemporal saliency detection model based on the phase spectrum of quaternion Fourier transform (PQFT) [GUO10].

In [GOF12], Goferman *et al.* propose an extension of the work in [GOF10] and calculated the saliency from a video content based on the context aware approach. This model follows four principles of human visual attention (Figure II-7), which are: (1) Local low-level considerations, including factors such as contrast and color. (2) Global considerations, which suppress frequently occurring features while maintaining features that deviate from the norm. (3) Visual organization rules, which state that visual forms may possess one or several centers of gravity about which the form is organized (the salient pixels should be grouped together and not spread all over the image). (4) High-level factors, such as priors on the salient object location and object detection (implemented as post processing operations). This model was qualitatively and quantitatively evaluated. The qualitative evaluation is done on 12 images with different scenes and it proves that the context aware method can always detect the salient objects according to the context of the image. The quantitative evaluation consists on comparing the ROC curves on two different benchmarks presented in [HOU07], [JUD09]. The experimental results show that this method outperforms state of the art methods [ACH09], [GUO08], [HAR06], [HOU07], [ITT98], [JUD09], and [RAH10].

Fang *et al.* [FAN14] propose a saliency detection model in MPEG-4 ASP [WEB11]. This model uses DCT coefficients of unpredicted frames (*I* frames) to get static features and predicted (*P* and *B* frames) to get motion information, see Figure II-17. YCrCb color space is used in MPEG-4 ASP video bit stream. The AC coefficients represent texture information for image blocks. The motion vectors are then extracted to get the motion feature. The combination of the static and the motion features is then applied based on a dynamic fusion. The experimental results are obtained on 50 video sequences and correspond to calculate the KLD and the AUC between the saliency map and the fixation map at saccade locations; it is shown that this model is validated by a KLD=1.828 and AUC=0.93.

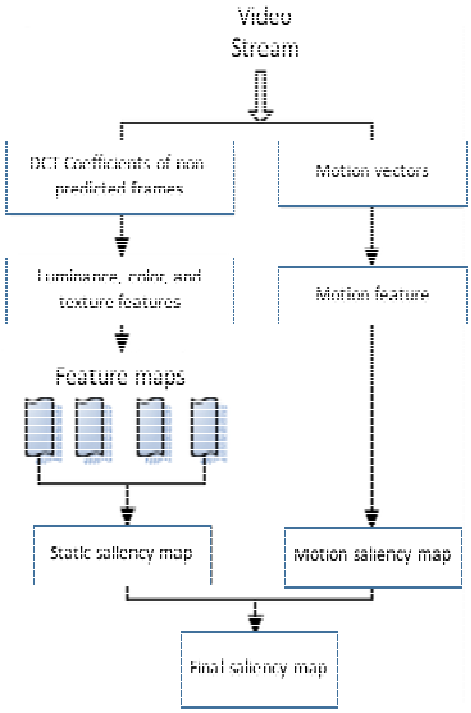


Figure II-17: Flowchart of the saliency computation model [FAN14]: the visual saliency is extracted from the transformed domain of the MPEG-4 ASP.

II.1.3. Conclusion

Based on 18 directly investigated studies (and on 25 additional studies to which these 18 refer to), the present state-of-the-art analysis can be synoptically presented in Table II.1. It brings to light a large variety of approaches for bridging human visual system and automatic saliency computation. While they generally converge in the type of the main methodological steps (extracting individual intensity, color, orientation and motion maps and subsequently fusion them at spatial and spatio-temporal levels), lot of divergences still remains in their definition, assessment (ground-truth vs. applicative, objective vs. subjective evaluation, composition of corpora, type of measures, etc.). Note that some top-down saliency studies consider in addition to the spatial and temporal saliency a third cue; for instance, Boujut *et al.* [BOU12] propose a fusion of spatial, temporal and geometric cues.

The state of the art analysis identifies automatic visual saliency detection as a particular research field. Its fundamental (neuro-biological) background is represented by the early works of Treisman *et al.*, advancing the integration theory for the human visual system and by Koch *et al.* who brought to light a time selectivity mechanism in the human attention. From the methodological point of view, all the studies published in the literature follow an inherent experimental approach: some hypotheses about how these neuro-biological characteristics can be (automatically) computed from the visual content are

first formulated and then demonstrated through experiments. In this respect, maybe the most relevant example is the seminal work of Itti [ITT98].

Moreover, we could not find any study related to the saliency extraction in the compressed domain, i.e. in-between the Q and E steps represented in Figure II-1.

Consequently, in order to address the conceptual contradiction between saliency and compressed streams, the present thesis should offer a comprehensive methodological and experimental view about the possibility of extracting the saliency regions directly from the compressed domain (both MPEG-4 AVC and HEVC), with minimal decoding operations.

Table II-1: State of the art synopsis of saliency detection models.

Model	Saliency detection / pooling	Validation	Results
Uncompressed image methods			
[ITT98]	Center-surround Gaussian differences /Average pooling	Ground truth: - 258 images - SFC	SFC(salient locations)>SFC(average)
[BRU05]	Quantifying the self-information of each local image patch / Gaussian filter	Ground truth: - 3600 natural images - ROC curve	ROC[TSO06] >ROC[ITT98] AROC=0.7288
[HAR06]	Graph-based model / Markovian-based weighted summation	Ground truth: - 108 images - AUC	0.96< AUC <0.98
[LEM06]	Center-surround interactions / weighted addition	Ground truth: - 10 images - CC and KLD	CC=0.71 KLD=0.46
[HOU07]	The spectral residual of a log-spectrum of an image/Gaussian filter	- 62 natural images - 4 naïve subjects - comparison with [ITT98] calculating the HitRate and the FalseAlarmRate and the computational coast in seconds	-HR[HOU07] >= HR[ITT98] -FAR[HOU07] <= FAR[ITT98] -lower computational coast (4.041s<61.621s)
[GOF10]	Context aware detection / post-processing based on the fourth principle	Ground truth [HOU07] - 62 images - ROC curves Applicative validation: - Image retargeting and summarization	ROC curves [GOF10] > ROC curves [HOU07] ROC curves [GOF10] > ROC curves [WAL06]
[MUR11]	Low-level video representation that predicts color appearance phenomena/inverse wavelet transform	Ground truth [BRU05] - 120 color images - 20 different subjects - KLD and AUC Ground truth [JUD09] - 1003 images - 15 subjects - KLD and AUC	 KLD=0.426 AUC=0.701 KLD=0.278 AUC=0.664
[CHE13]	Color contrast and color spatial distribution / Pooling based on compactness	Ground truth: - 1000 images - MAE	MAE decreased by 25.2%
Compressed image methods (JPEG)			
[FAN12]	Extracting intensity, color, and orientation from DCT coefficients / Weighted summation	Ground truth: - 1000 images - AUC	AUC= 0.93

Table II-1 (continuing): State of the art synopsis of saliency detection models.

Uncompressed videos methods			
[ITT05]	Detecting the low level surprising event in the video	Ground truth: - [WEB06] - KL scores	KL= 0.241
[ZHA06]	Contrast based features extraction / dynamic averaging technique	- 9 video sequences - 5 assessors votes on the correctness of the detection	*Good*=0.77 *Poor*=0.15 *Failed*= 0.08
[LEM07]	The center surrounds filters and the relative motion / weighted average	Ground truth: - 7 video sequences - CC, KLD, and ROC curves	CC=0.41 KLD=19.21
[HOU08]	Incremental Coding Length (ICL) based saliency model / weighted summation	Ground truth: - 1 video sequence and 120 still images - KLD and AUC	KLD= 0.54 AUC= 0.79
[SEO09]	Regression kernel / self-resemblance	Ground truth: - corpus [BRU05] - KLD and AUC	KLD=0.34 AUC=0.67
[MAR09]	Two parallel ways (static biologically inspired and dynamic highlights moving objects) / parallel saliency maps	Applicative validation: - three videos - harmonic average between *precision* and recall F1.	F1 (MAR09) > F1(random summary) > F1 (one frame selection at the middle of each shot)
[GUO10]	Phase based saliency model detection / QFT formula	Ground truth: - 1 video (988 frames) and 100 still images - AUC	AUC= 0.83
[GOF12]	Context aware detection / fusion based on centers of gravity	Ground truth: - corpus [HOU07][JUD09] - ROC curve	ROC curve (context aware) > ROC curve (State of the art methods)
Compressed video methods (MPEG-4 ASP)			
[FAN14]	Extracting intensity, color, and orientation from DCT coefficients, motion from motion vector / Dynamic pooling	Ground truth: - corpus [WEB06] - KLD and AUC	KLD=1.82 AUC=0.93

II.2. Visual saliency as a watermarking optimization tool

By its very nature, under the watermarking framework, the visual saliency is related to the concept of transparency: a priori, saliency maps are expected to act as an optimization tool for selecting the locations for mark insertion, Figure II-18. For prescribed levels of robustness and data payload, inserting the mark into salient regions is expected to result into a lower transparency and, conversely, inserting the mark into non-salient regions is expected to increase the transparency. Of course, this general expectation can be extended for other watermarking properties. For instance, for prescribed transparency and data-payload constraints, inserting the watermark in salient regions is expected to ameliorate robustness. Similarly, for prescribed transparency and robustness constraints, inserting the watermark in salient regions is expected to increase the data payload.

However, there is no a priori hint about the extent to which saliency can be benefic for watermarking. For solving this issue, several research studies are already reported [SUR09], [NIU11], [TIA11], [LI12], [AGA13], [CHE15], [WAN15], [BHO16], and [GAW16].

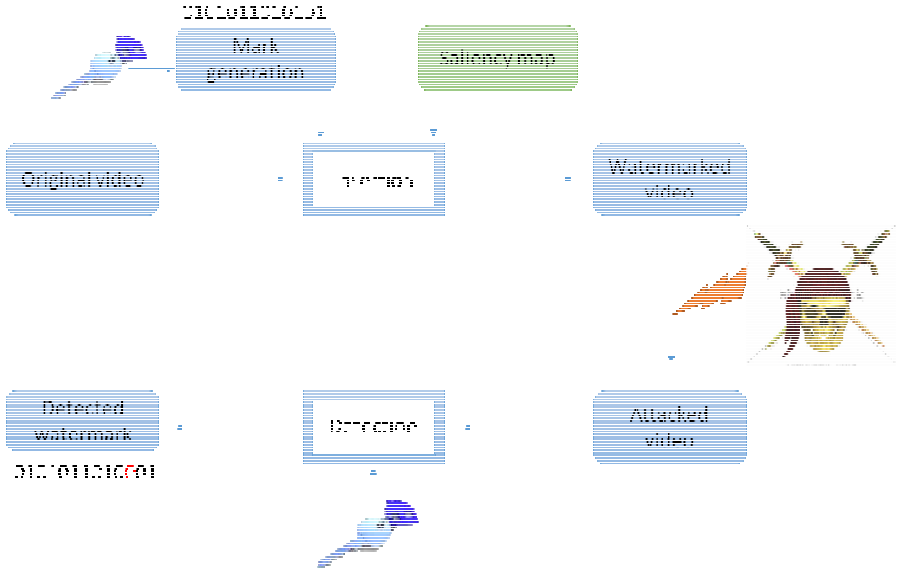


Figure II-18: Principle of a watermark embedding scheme based on saliency map.

Sur *et al*. [SUR09] propose a new spatial domain adaptive image watermarking scheme. First, the Itti's saliency model [ITT98] is used so as to determine the salient locations. Then, the least salient pixels from those regions are replaced by watermarked pixels; the watermarking method itself is based on the LSB technique. The experimental results mainly investigate the transparency property, expressed through an HVS-related objective measure, namely the Watson's Total Perceptual Error (TPE): gains by factors between 1.5 and 4 (according to the data payload) are obtained.

The study in [NIU11] considers a two-folded HVS approach for increasing the transparency of the SS (spread spectrum) techniques in the DCT domain. The mark is inserted into non-salient regions detected according to the [HOU07] saliency model. However, prior to the insertion, the AWGN (additive white Gaussian Noise) represented the mark is modulated according to JND (Just Noticeable Distortion) profiles. This allows shaping lower injected-watermark energy into more sensitive regions and higher energy into the less perceptually significant regions in the image. The experimental results are illustrated through one images showing perceptual improvement with respect to the original JND-based spread-spectrum method.

Tian *et al.* [TIA11] propose an integrated visual saliency-based watermarking approach, which can be used for both synchronous image authentication and copyright protection. First, the regions of interest (ROI) are extracted according to a proto-object model and the copyright information is embedded therein as the robust watermark. Secondly, the edge map of the most salient ROI is embedded into the LL sub-band of the wavelet-decomposed watermarked image as the fragile watermark. The experiments show the efficiency of the method in terms of transparency (evaluated through the PSNR). The robustness experiments concerns a restricted class of attacks (white noise addition, median filtering and the JPEG compression) and show that the advanced method outperforms [MOH08]. The fragility and the efficiency to detect and locate tampering attacks are also investigated.

In order to verify the integrity of face (biometric) images, Li *et al.* [LI12] define a multi-level authentication watermarking scheme based on He *et al.* [HE06]. Biometric data related to the face images are considered as watermarks to be inserted into the same image. The face images are segmented into regions of interest (ROI) and regions of background (ROB) based on salient region detection. The watermark is adaptively embedded into the biometric images based on detection results. The saliency map is computed according to the method presented in [MAL90]. The analysis of the perceptual quality is validated by a PSNR = 33.13 dB. In order to evaluate the performance of the proposed multi-level authentication watermarking scheme, an analysis on the tamper detection probability inspired by Yu [YU07] is conducted. When face images suffer from malicious tamper, the extracted watermarks can be used to recover the damaged biometric data and reconstruct face images. Even if the tamper ratio is up to 0.4, the re- covered face image can be used for verification.

Agarwal *et al*. [AGA13] introduce an algorithm that embeds information into visually interesting areas within the host image. The watermarking algorithm consists on inserting in non-salient regions of the blue component (as the change in blue component is the least perceptible to human visual system). The saliency map is generated based on the Graph-Based Visual Saliency (GBVS). The advanced method performs a 3-Level Selective DWT on the blue component of RGB cover image. The paper shows the result of the watermarking schema on four RGB images. The experimental results are structured at three levels. First, it is shown that the watermark remains imperceptible even after increasing the data payload: for a data payload of 1024 bytes, the PSNR=41.3. Secondly, the robustness against three types of attacks (namely Gaussian blurring, JPEG compression, and median filtering) is evaluated by computing the correlation between the inserted and the recovered watermarks. It is thus stated that the advanced method outperforms the studies in [TIA11] and [MOH08]. Finally, it is shown that for prescribed BER (Bit Error Rate) and PSNR values, the advanced model increases the value of payload.

Chen *et al.* [CHE15] advance a method embedding the watermark into the DC (Direct Component) component of the DCT, according to a JND adaptive strategy. The saliency map is obtained by applying a JND fusion on the static and the dynamic saliency map. The motion saliency map is computed by applying the motion JND and the static saliency map is obtained according to [ITT98]. Experimental results demonstrate the effectiveness of this method: by keeping the same data payload and the same robustness, the transparency is ameliorated by 3 dB.

Wan *et al.* [WAN15] propose a visual saliency based logarithmic STDM (Spread Transform Dither Modulation) watermarking scheme. The watermark is embedded into a sub-set of non-salient DCT coefficients. The visual saliency is determined based on the energy of the DCT features of luminance and texture. By investigating the BER results under different attacks, the method robustness against AWGN addition, JPEG compression and S&P (Salt and Pepper) noise is proved. The results show the method has statistically significant better outcomes in terms of the VS-based IQA metric. The robustness is improved by at most 5%.

Bhowmik *et al.* [BHO16] also adapt the strength of the watermark according to the salient / non-salient feature of the DWT coefficients bearing that watermark. A low complexity wavelet domain visual attention model is proposed. It uses all detail coefficients across all wavelet scales for center-surround differencing and normalization. Subsequently, it fusses 3 orientation features in a non-separable manner to obtain the final saliency map. The performance evaluation shows up to 25% and 40% improvement against JPEG2000 compression and common filtering attacks, respectively.

Gawish *et al.* [GAW16] report on a saliency guided watermarking approach. A weighted sum between the non-saliency and heterogeneity-brightness maps generates a map locating the best (in the perceptual sense) places to hide the watermark. The DCT middle frequency coefficients of the top candidates of the watermarking map are then used for bearing the data. Experiments shows that this method outperforms the Harris-Laplace based method [ZHA12] in terms of transparency (an increase of 0.5 dB in PSNR) and robustness (a decrease of 0.1 in (NHS) Normalized Hamming Similarity) over different attacks.

As a conclusion, this concise state-of-the-art study (see Table II-2) on the relationship between saliency and watermarking shows different promising results. For instance, guiding the insertion of the watermark by the saliency map offers significant improvements. Moreover, the investigated models bring to light that exploring the ROI can be benefic for each of the three main watermarking properties: robustness ([TIA11], [LI12], [AGA13], [WAN15], [BHO16], and [GAW16]) transparency ([SUR09], [NIU11], [TIA11], [LI12], [CHE15], [WAN15], and [GAW16]) and data payload [AGA13].

By analyzing the 9 state-of-the-art studies we can notice that the trade-off between watermark embedding and the visual saliency extraction is not yet reached in the compressed domain, i.e. in-between the Q and E steps represented in Figure II-1. Thus, to guide a compressed stream watermarking application we should extract saliency directly in the compressed stream syntax elements in order to avoid decoding/re-encoding operations.

Table II-2: State-of-the-art of the watermark embedding scheme based on saliency map.

Reference	Watermarking schema	Visual saliency model	Benefits
[SUR09]	LSB (lowest significant bit)	[ITT98]	Gains in TPE by factors between 1.5 and 4 (according to the data payload)
[NIU11]	SS in the DCT domain	[HOU07]	Subjective amelioration
[TIA11]	Inserting robust watermark into DCT of ROI and the fragile watermark into LL sub-band	Proto-object model	Transparency: PSNR >= 42 Fragileness and efficiency: Preserving authentication while detecting tampering Robustness: outperforms the [MOH08] when resisting the white noise, median filter and the JPEG compression attacks.
[LI12]	Embedding watermark in biometric images	[MAL90]	PSNR = 33.13 dB. A super performance at detection probabilities and false detection probabilities. Even if the tamper ratio is up to 0.4, the recovered face image can be used for verification.
[AGA13]	Inserting the watermark on non salient regions of the blue component	Graph Based Visual Saliency (GBVS)	Outperforms [TIA11] and [MOH08] in term of robustness against no attack, Gaussian blur, JPEG compression, and median filter and proved that their method For a prescribed BER and PSNR the model increases the value of payload.
[CHE15]	Watermark insertion in the DC coefficient	[ITT98] and motion JND	Increasing the PSNR by 3 dB
[WAN15]	Inserting the watermark in the host vector of the DCT coefficients	Extracting features from DCT coefficients	Statistically significant better outcomes in terms of the VS-based IQA metric. The robustness is improved by at most 5%.
[BHO16]	Inserting the watermark in the wavelet domain	Low complexity wavelet domain model	up to 25% and 40% improvement against JPEG2000 compression and common filtering attacks
[GAW16]	Inserting watermark in natural images	Feature redundancy	decreasing robustness by 0.1 in NHS and increase transparency by 0.5 dB in PSNR

II.3. Direct compressed video stream processing

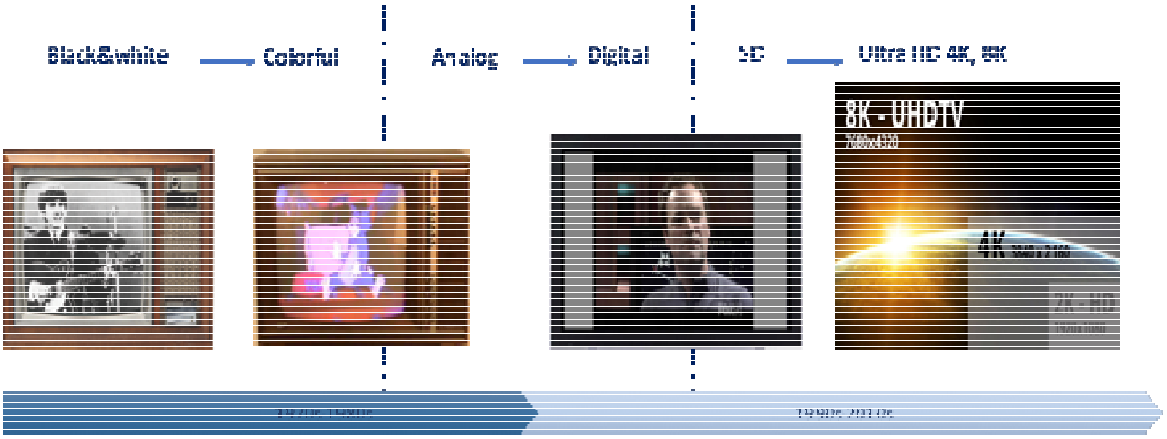


Figure II-19: Video quality evolution.

Nowadays, the video quality improves in parallel with the increase of the quantity of generated video content, thus stressing the urge for better, more sophisticated compression standards.

While compression reduces the storage and network costs, it intrinsically increases the cost of subsequent processing of the visual content: applying traditional, pixel-oriented image processing algorithms would require the a priori decompression of data and, in some cases, even the a posteriori re-compression of the processed data. The overhead of such an approach would range somewhere in-between 1 and 20. Just for example, the study in [HAS14] reports that for an MPEG-4 AVC semi-fragile video watermarking method, more than 94% of the total processing time is required by the video encoding/decoding operations while the watermarking itself covers only 6% from the total time!

In order to circumvent such an issue, several research studies took the challenge of processing the visual content directly in the compressed stream format; we shall illustrate the principles of such approaches by considering 9 studies, namely [KRA05], [THI06], [MAN08], [POP09], [ZHO10], [BEL10], [FAN12], [AMO12], and [OGA15], which will be presented in chronological order.

The study in [KRA05] addresses the problem of constructing a super-resolution (SR) mosaic from MPEG compressed video stream; such a mosaic can be used as a tool for increasing the image quality / resolution, without decomposing the content. The method consists in the use of color information only from *I* frames and motion information only from *P* frames. The main contribution of this paper is minimizing the decoding overhead (i.e. to decode as less data as possible) while improving the visual quality of initial DC-resolution mosaics. Experimental results show that the SR mosaics thus obtained are visually better than other methods and the first results are promising. A discussion on the impact of the main parameter of the reconstruction method is also presented.

Thiemert *et al.* [THI06] advance a semi-fragile watermarking system devoted to the MPEG-1/2 video sequences. The mark computation is based on the properties of the entropy computed at the 8x8 block levels. The mark is embedded by enforcing prescribed relationship between the DCT coefficients of some

blocks. The experiments are run on one sequence (whose length is not précised) encoded at 1125 kbps. The method proved both robustness (against JPEG compression with QF=50) and fragility against temporal (with 2 frame accuracy) and spatial (with a non-assessed accuracy) content changing.

Manerba *et al.* [MAN08] present a method for foreground object extraction following a "rough indexing" paradigm. This method combines motion masks with the morphological color segmentation operated at DC coefficients of MPEG1,2 compressed stream. In this respect, each group of picture (GOP) is first analyzed and, based on color and motion information (extracted from the *I* and *P* frames, respectively), foreground objects are extracted. Secondly, a post-processing step is performed so as to refine the result and to correct the errors due to the low-resolution approach. Results proved that the percentage of the object detection varies from a video sequence to another from 0 to 100%. The object extraction computation time also depends on the video sequence (0.08s to 0.43s).

Poppe *et al.* [POP09] introduce a method to detect moving objects in H.264/AVC compressed video surveillance sequences. However, motion vectors are created from a coding perspective and additional complexity is needed to clean the noisy field. Hence, an alternative approach is presented, based on the size (in bits) of the blocks and transform coefficients used within the video stream. The system is restricted to the syntax level and achieves high execution speeds, up to 20 times faster than the state-of-the-art (at that time) studies. Finally, the influence of different encoder settings is investigated to show the robustness of their system.

Belhaj *et al.* [BEL10] introduce a binary spread transform based QIM for MPEG-4 AVC stream watermarking. By combining QIM principles, spread transform, a perceptual shaping mechanism, and an information-theory driven selection criterion, they achieved a good transparency and robustness against transcoding and geometric attacks. By advancing the *m*-QIM theoretical framework, [HAS10] extends the QIM watermark principle beyond the binary case. In this respect, the research was structured at two levels: (1) extending the insertion rule from the binary to *m*-ary case and (2) computing the optimal detection rule, in sense of average probability error minimization under the condition of Gaussian noise constraints. Thus, the size of the inserted mark is increased by a factor $\log_2 m$ (for prescribed transparency and robustness constraints).

Zhou *et al.* [ZHO10] advance an application of digital fingerprinting[2] directly in the MPEG-2 compressed video stream. Fingerprints are embedded into each I-frame of the video, by means of data repetition technique so as to ensure accurate extraction of fingerprint. First, the fingerprint is generated according to two-tier structure based on error correcting code and spread spectrum. Second, the fingerprint is embedded during decoding. The algorithm selects the I-frame in the video for embedding to enhance the robustness of the fingerprint. Finally, the extraction step of the fingerprint is described as easy and effective since the data repeating technology is adopted in the embedding algorithm. The embedding method satisfies the requirements of invisibility and real-time quite well. In term of invisibility (PSNR>=35 dB) while in term of real-time (0.1s gain in the Average Running Time compared to other method).

In order to extract the saliency maps, Fang *et al.* [FAN12] no longer consider pixel representation of the

[2] In this study, the term 'fingerprinting' also encompasses a multiple-bit watermarking technique.

image but a transformed domain related to the JPEG compression. He proposes an image retargeting algorithm to resize images, based on the extracted saliency information from the compressed domain. Thanks to the directly derived saliency information, the proposed image retargeting algorithm effectively preserves the objects of attention and removes the less appealing regions. The statistical results for 500 retargeted images show that the mean opinion score of images retargeted according to [FAN12], namely 3.708, is higher than those according to three state-of-the-art algorithms [RUB08], [WOL07] and [REN09], which were reported to be 3.278, 3.348, and 3.424, respectively.

Amon *et al.* [AMO12] present a method for compressed domain stitching of HEVC streams, with applications to video conferencing. The methodological approach considers three incremental levels, namely pixel, syntax elements, and entropy coding. The results show gains in terms of quality of resulted video content (between 0.5 dB and 0.8 dB with respect to the method in the pixel domain), in compression efficiency (evaluated as a PSNR-bitrate function) and computational complexity (in the sense that the operation involved in the advance method are less complex than a complete encoding/decoding chain).

Ogawa and Ohtake [OGA15] propose a watermarking method for HEVC/H.265 video streams that embeds information while encoding the video. After quantizing, the quantized data is divided into two parts: common and distinct. The quantized values in the common part are encoded using the arithmetic coding CABAC (Entropy Coding). The quantized value in the distinct part is changed according to the information bit. After the change of the quantized values, the values are encoded using CABAC. Thus, a modified HEVC elementary stream is generated. Authors state that it is possible to embed information into a compressed stream using this method without degrading the content and with an appropriate robustness that meets the requirements of the users. There is no discussion on the quality of the watermarking.

To conclude with, the huge amount of the visual content stored and transmitted in a compressed stream bring to the light that image/video processing directly in the compressed stream becomes more a necessity rather than an option. The analysis of the 9 state-of-the-art compressed stream application studies brings to light that proceeding directly in the compressed stream offers the possibility of a gain in complexity and computational cost while preserving or even improving the application properties.

Consequently, in this thesis, we take the challenge of extracting the saliency map in the compressed domain in order to guide the watermark insertion in a compressed stream watermarking application (both MPEG-4 AVC and HEVC), with minimal decoding operations.

Table II-3: State of the art of the compressed stream application.

Reference	Application	Compressed domain
[KRA05]	Super-resolution (SR) mosaic	MPEG
[THI06]	Watermarking	MPEG1/2
[MAN08]	Foreground object extraction	MPEG1/2
[POP09]	Detecting moving object	MPEG-4 AVC
[ZHO10]	Fingerprinting	MPEG-2
[BEL10]	Watermarking	MPEG-4 AVC
[FAN12]	Image retargeting	JPEG
[AMO12]	Compressed domain stitching of streams coded	HEVC
[OGA15]	Watermarking	HEVC

III. Saliency extraction from MPEG-4 AVC stream

By bridging uncompressed-domain saliency detection and MPEG-4 AVC compression principles, the present thesis advances a methodological framework for extracting the saliency maps directly from the stream syntax elements. In this respect, inside each GOP, the intensity, color, orientation and motion elementary saliency maps are related to the energy of the luma coefficients, to the energy of chroma coefficients, to the gradient of the prediction modes and to the amplitude of the motion vectors, respectively. The experiments consider both ground-truth and applicative evaluations. The ground-truth benchmarking investigates the relation between the predicted MPEG-4 AVC saliency map and the actual human saliency, captured by eye-tracking devices. The applicative validation is carried out by integrating the MPEG-4 AVC saliency map into a robust watermarking application.

III.1. MPEG-4 AVC saliency map computation

In this chapter, we extract the visual saliency map directly from the MPEG-4 AVC compressed stream. We first follow the Itti's [ITT98] basic principles according to which visual saliency can be obtained by combining three elementary static saliency maps (intensity, color, orientation); we complete then this static saliency map with a motion saliency map, [BOR13].

For each GOP (see Figure III-1), the static saliency map is computed from the *I* frame. The intensity and color maps are extracted from the residual MPEG-4 AVC luma and chroma coefficients, respectively, while the orientation map is computed based on the intra prediction modes. The motion map is generated based on the motion vectors from the *P* frames.

The computing of each map as well as their post-processing and pooling are detailed in the following sub-sections.

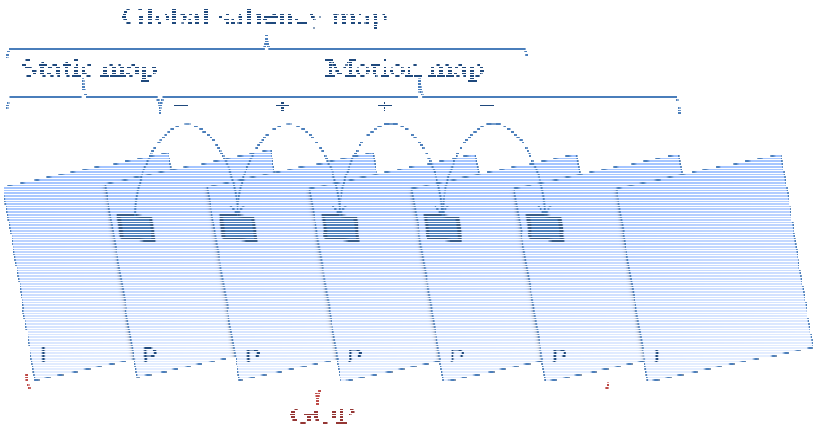


Figure III-1: Saliency map computation in a GOP.

III.1.1. MPEG-4 AVC elementary saliency maps

Intensity map

As explained in Chapter II.1, according to Itti, visual neurons are most sensitive in a small region of the visual space (the center) while stimuli presented in a border, weaker antagonistic region concentric with the center (the surround) inhibit the neural response. In order to model this human vision behavior for uncompressed image saliency extraction, Itti considers a dyadic Gaussian pyramid decomposition to compute the center-surround differences. When considering now a compressed video stream, an analysis of the syntax elements brings to light that the differences between some stimuli in the image and their neighborhood is represented by the intra prediction syntax elements. Hence, we shall start our study on static visual saliency by considering the intra prediction related syntax elements.

I frames are encoded according to Intra prediction modes which exploit the spatial redundancy to enhance the compression efficiency. For each 4×4 pixel block *X*, the prediction mode minimizing the rate-distortion cost is selected and is deployed so as to compute the corresponding prediction block *P* from the neighboring blocks. Consider an *R* residual block (the difference between the current block *X* and the predicted block *P*):

$$R = X - P \tag{III-1}$$

At the pixel level, the *R* blocks are represented by one luminance and two chrominance values. These values are subsequently DCT transformed and then quantified, thus obtaining the so-called luma (*Y*) and chroma (Cr, Cb) MPEG-4 AVC channels.

For each 4×4 DCT transformed and quantified *R* block, we define the intensity saliency map M_i according to (III-2):

$$M_i(k) = \sum_{u=1}^{4}\sum_{v=1}^{4} {Y_{k,u,v}}^2 \tag{III-2}$$

where *k* is the block index in the frame, *u* and *v* are the coefficient coordinates in the *k* block and *Y* is the luma residual coefficient.

According to (III-2), a luminance energy value is attached to each block: the larger this M_i value, the more salient the *k* block.

Color map

In order to define the color saliency map, we shall keep the same conceptual approach as for the intensity (i.e. associating saliency to the regions with high energy color components) and we shall take into account the human visual system peculiarities related to the color perception.

In [ITT98], it is brought to light that the elementary colors are represented in cortex according to a so-called color double-opponent system. In the center of their receptive fields, neurons are excited by one color (e.g., red) and inhibited by another (e.g., green), while the converse is true in the surrounding areas. Such spatial and chromatic opponency exists for the red/green and yellow/blue color pairs (and, of course, for their complementary green/red and blue/yellow color pairs).

Consequently, the MPEG-4 AVC color saliency map will be based on the energy featured by the composition of red/green and yellow/blue opponent pairs, as follows.

We first convert the color information extracted from the (*Y,Cr,Cb*) MPEG-4 AVC DCT and quantified color space into the transformed and quantified (*r,g,b*) space:

$$\boldsymbol{r = Y + 1.402(Cr - 128)}$$

$$\boldsymbol{g = Y - 0.34414(Cr - 128) + 0.71414(Cb - 128)}$$

$$\boldsymbol{b = Y + 1.772(Cb - 128)}$$

Secondly, through analogy with [ITT98], the two opponent color pairs *RG* (Red/Green) and *BY* (Blue/Yellow) are computed for each *(u,v)* coefficient in the macroblock:

$$\boldsymbol{RG_{u,v} = (Red_{u,v} + Green_{u,v})/2}$$

$$\boldsymbol{BY_{u,v} = (Blue_{u,v} + Yellow_{u,v})/2}$$

where

$$\boldsymbol{Red = r - (g + b)/2}$$

$$\boldsymbol{Green = g - (r + b)/2}$$

$$\boldsymbol{Blue = b - (g + r)/2}$$

$$\boldsymbol{Yellow = \frac{r + g}{2} - \frac{|r - g|}{2} - b}$$

Finally, we compute the color saliency map M_c as the sum of the energy in the double color-opponent red/green and blue/yellow spaces:

$$M_C(k) = \sum_{u=1}^{4} \sum_{v=1}^{4} RG_{k,u,v}{}^2 + BY_{k,u,v}{}^2 \qquad \text{(III-3)}$$

where *k* is the block index in the frame, while *u* and *v* are the coefficient coordinates in the *k* block. According to (III-3), a color energy value is assigned to each block: the larger this M_c value, the more salient the *k* block.

Orientation map

The MPEG-4 AVC standard offers 13 directional intra prediction modes. For each current block, a directional prediction mode which minimizes the bit rate distortion cost is selected to perform the prediction.

According to the intra MPEG-4 AVC paradigm, the prediction modes reflect the orientation of the corresponding block with respect to its neighborhood blocks. Hence, we shall compute the orientation map by analyzing the heterogeneity among the intra prediction modes inside the *I* frame.

The building of the orientation map starts by extracting values of prediction modes since each intra prediction mode gives us information about the orientation of a given block; then, the obtained orientation for each bock will be compared with those obtained for a set of neighboring blocks: blocks which feature the same direction as their neighborhood are considered as non-salient (see Figure III-2 left) while blocks with different orientation modes from their neighborhood are considered as salient (see Figure III-2 right).

The M_o orientation map is computed according to:

$$\boldsymbol{M_o(k) = 1 - \frac{Card(\{\ O_l = O_k, l \in V\})}{Card\ (V)}} \qquad \text{(III-4)}$$

where *k* is the block index in the frame, *V* is the *k* block neighborhood and *l* is the block index belonging to *V*; *Card* is the cardinality (number of elements) in the considered set.

According to (III-4), a gradient measure of the prediction mode discontinuity is associated to each block: the larger the M_o value, the more salient the *k* block.

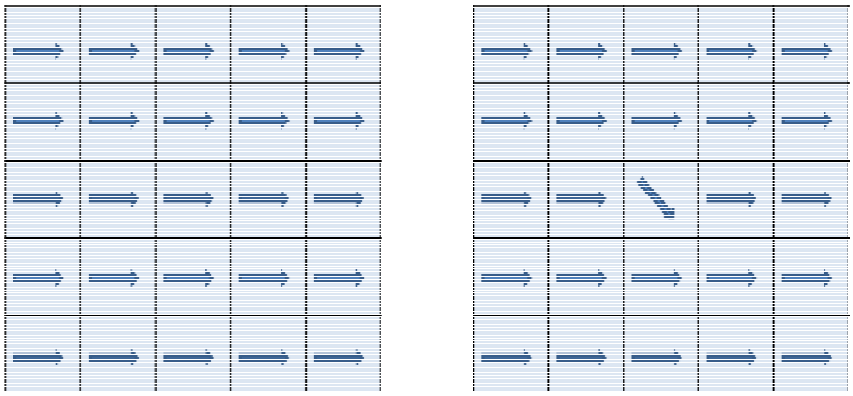


Figure III-2: Orientation saliency: the central block into a 5x5 block neighborhood is not salient when its "orientation" is identical with its neighbors (see the left side of the figure); conversely, if the block orientation differs from its neighbors, the block is salient (see the right side of the figure).

Motion map

Inside the GOP, the motion information is encoded in the *P* frames: the motion vector difference indicates the difference between the motion vector of the current block and the motion vector of a nearby block.

For each GOP, we define the motion saliency map as the global motion amplitude, computed by summing the motion amplitude over all the *P* frames in the GOP (see Figure III-3) at the same corresponding block position:

$$M_m(k) = \sum_{P \in GOP} \sqrt{MVDx_k{}^2 + MVDy_k{}^2} \qquad \text{(III-5)}$$

where ($MVDx_k$, $MVDy_k$) denote horizontal and vertical components of the motion vectors difference of the block *k*, and M_m represents the global motion amplitude among the *P* frames of a GOP; the larger this M_m value, the more salient the *k* block position.

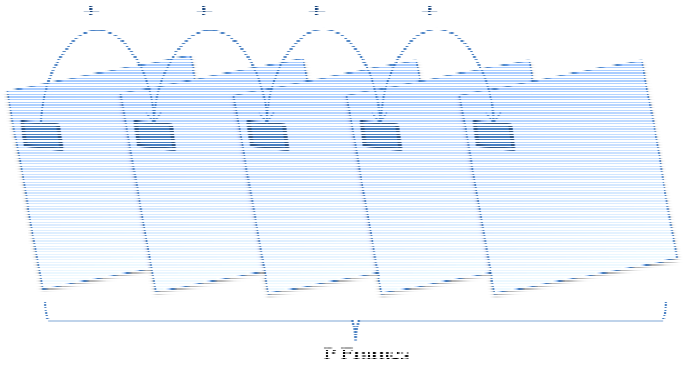


Figure III-3: Motion saliency: the motion amplitude over all the P frames in the GOP is summed-up.

III.1.2. Elementary saliency maps post-processing

The obtained saliency map corresponding to each feature is now to be normalized to the same dynamic range. This is achieved on each individual map, by a three steps approach, Figure III-4.

First, outlier detection is performed: the 5% largest and 5% lowest values are eliminated. Then, the remaining values are mapped to the [0 1] interval through an affine transform. Finally, an average filtering, with the window size equal to the fovea area is applied.

Note that the very definition of the orientation map makes these post-processing operations meaningless in its case.

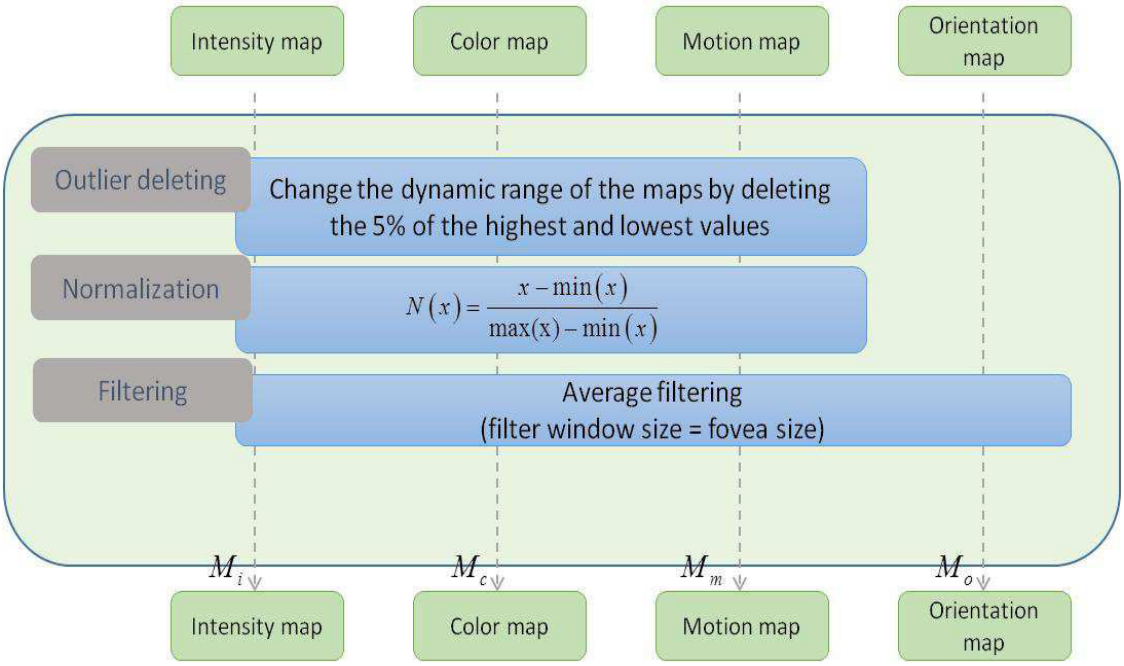


Figure III-4: Features map normalization.

III.1.3. Elementary saliency map pooling

The MPEG-4 AVC saliency map is the fusion of the static and the dynamic map. The static saliency map is in its turn a combination of intensity, color and orientation features maps. Despite the particular way in which all these elementary maps are computed, the fusion technique allowing their combination plays a critical role in the final result and makes the object of a research challenge of the studies in [AMM15], [MUD13], [MAR09].

In our study, the pooling takes place at two levels: static (i.e. pooling intensity, color and orientation maps in order to obtain the static map) and dynamic (i.e. pooling static and motion maps in order to obtain the final saliency map). In order to decide on the pooling formulas for our saliency maps, we considered two criteria. On the one hand, according to the state-of-the-art studies [ITT98], [HAR06], the most often considered static fusion formula is the average. Considering the dynamic fusion, weighted averages between static and motion maps are also very popular. Consequently, we included in our study the following pooling formulas:

$$M_s = \frac{1}{3}(M_i + M_C + M_o)$$

$$M_D = \alpha M_s + \beta M_m + \gamma (M_s \times M_m)$$

where M_D is the final MPEG-4 AVC saliency map. By changing α, β, γ values we obtain several static-dynamic fusing formulas, defined over the same average static fusion. In our study, we considered:

• α=β=γ=1, which is the combination of the addition and the multiplication static-dynamic fusion technique; the corresponding MPEG-4 AVC saliency map will be further referred to as *Combined-avg* (where avg represents the average static pooling technique);

• α=β=0, γ=1, which corresponds to the multiplication static-dynamic fusion technique; this map will be further referred to as *Multiplication-avg*;

• α=β=1, γ=0, which corresponds to an additive static dynamic fusion; this map will be further referred to as *Addition-avg*;

• α=1, β=γ=0, which corresponds to static saliency map; the corresponding map will be further referred to as *Static-avg;*

• α=0, β=1, γ=0, which corresponds to motion saliency map; the corresponding map will be further referred to as *Motion*.

On the other hand, according to the fusing formula investigation [AMM15] detailed in Appendix A, where 48 different pooling combinations (6 static pooling formula and, for each of them, 8 dynamic pooling) were investigated, the most accurate combination (in the sense of KLD and AUC computed on a ground truth database of 80 sec) is Skewness (defined as the third moment on the distribution of the map [MAR09]) static-dynamic fusion over the maximum static fusion. Consequently, we shall also include this pooling formula in our study and we shall further refer it as *Skewness-max*.

III.2. Experimental results

We will evaluate the performances of 6 alternative ways of combining the elementary maps described above: we will retain the elected spatio-temporal saliency map in the first level, resulted from the study of the fusing formula (see Appendix A.1) where 48 fusion formulas are performed: six different fusion techniques for static features and eight fusion formulas over the static and motion saliency maps. The performances of these 48 MPEG-4 AVC saliency maps are discussed by comparing them to the ground truth represented by the density fixation maps captured by the Eye Tracker on eight video sequences at the IRCCyN premises [WEB05]. The comparison to the density fixation maps is completed by using two objective measures: the KLD (Kullback Leibler Divergence, assessing the differences between the distributions of the two investigated entities) and the AUC (Area Under Curve, assessing the differences between the two entities at given locations). In addition, we will add some fusion technique generally used in the state of the art model then we will precede two different validations: the ground truth validation and the applicative validation.

In our study, we extract the saliency map only from *I* and *P* frames. We did not consider *B* frames in our experimental study because such frames may not be present in some compressed streams (e.g. the streams encoded with the Baseline profile). Nevertheless, our method can be applied to any MPEG-4 AVC video configuration, be it with or without *B* frames. Moreover, if the video compressed stream contains *B* frames, only *I* frames and *P* frames will be considered to extract static and dynamic saliency,

respectively. It is not necessary to compute the saliency from *B* frames. As the saliency prediction mostly relates to the fixation locations (including pursuit) and keeping in mind that usual human fixation duration is between 100 ms and 200 ms, we do not need to process each and every frame in a video sequence (e.g: for a frame rate of 25 fps, each frame comes every 40 ms).

III.2.1. Ground truth validation

Test-bed

Our experiments are structured at two nested levels, according to the evaluation criteria and to the actual measures and corpora, respectively Table III-1.

First, several evaluation criteria can be considered. We shall consider both the *Precision* (defined as the closeness between the saliency map and the fixation map) and the *Discriminance* (defined as the difference between the behavior of the saliency map in fixation locations and in random locations) of the saliency models.

Secondly, for each evaluation criteria, several measures can be considered. Our assessment is based on two measures of two different types (the KLD and AUC). We implemented the KLD based on [KUL51] [KUL68] while we used the AUC implementation available on Internet [WEB07].

Note that in order to ensure the statistical relevance for the KLD and AUC values, we compute the average values (both over the GOP in an individual video sequence and over all the processed video sequences), the related standard deviations, 95% confidence limits and minimal/maximal values. This way, the ratio between the average value and the standard deviation (the so-called signal to noise value [FRY65], [WAL89]) can be estimated (point estimation) in order to assess the sensitivity of the KLD and AUC with respect to the randomness of the processed visual content: the bigger the signal to noise ratio, the less sensitive the corresponding measure with respect to the visual content variability.

Two different corpora are considered and further referred to as: (1) the *reference* corpus organized in [WEB05] and (2) the *cross-checking* corpus organized in [WEB06].

The *reference* corpus is a public database organized by IRCCyN [WEB05]. It contains 8 video sequences of 10 seconds each one. For each video, the eye-tracker data are extracted for 30 observers. The distance between observers and the display is 3m. The resolution of the display is 1920×1080 with 50 Hz frame rate. Based on those results, a density fixation map is calculated for each video. In our experiments, these videos are encoded in MPEG-4 AVC Baseline Profile (no *B* frames, CAVLC entropy encoder) at 512 kb/s. The GOP size is set to 5 and the frame size is set to 576×720. The MPEG-4 AVC reference software (version JM86) is completed with software tools allowing the parsing MPEG-4 AVC syntax elements and their subsequent usage, under syntax preserving constraints.

The *cross-checking* corpus includes 50 various types of video clips, summing-up to over 25 minutes. The human saliency is represented by the saccade data captured by an eye-tracker (240-Hz infrared-video-based) from eight observers. In our experiments, we applied the same encoding operations as in the case of the *reference* corpus.

While the choice of corpora in the test-bed is always a crucial issue in image/video processing, it becomes of an upmost importance in visual saliency studies. By its very principles, any bottom-up model is a model solely depending on the visual content. In order to grant generality for our results, we considered two types of criteria when choosing our corpora:

- we used two public corpora, already considered in a large variety of publications;
- we strengthened our results by an in-depth statistical analysis:
  - we defined and computed a sensitivity measure in order to compare the dependency of the saliency model with the randomness of the content in the processed corpus,
  - we computed the minimal, maximal and the 95% confidence limits for the two investigated measures (KLD and AUC).

Table III-1: Assessment of the model performance in predicting visual saliency.

Ground truth validation: concordance between the computed saliency map and human visual saliency	
Precision: similarity with ground truth (cf. Chapter III.2.1.2)	***Discriminance***: difference with respect to random locations (cf. Chapter III.2.1.3)
Measures: KLD, AUC *Corpus*: *reference*	*Measures*: KLD, AUC *Corpus*: *reference*, *cross-checking*

During our experiments, we benchmark our MPEG-4 AVC saliency map against three state of the art methods, namely: Ming Cheng *et al*. [CHE13], Hae Seo *et al.* [SEO09] and Stas Goferman [GOF12], whose MATLAB codes are available for downloading.

Precision

In this experiment, we compare the computed saliency maps to the density fixation maps captured from the human observers (cf. illustration in Figure III-5); the *reference* corpus [WEB05] will be processed.

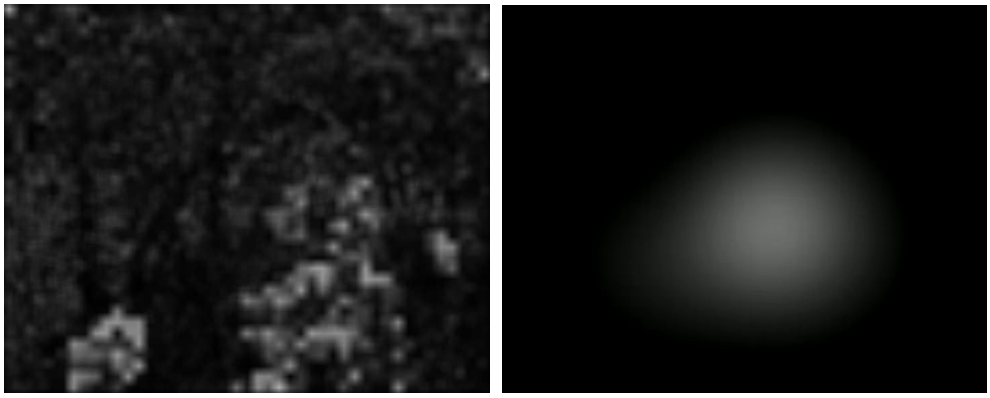


Figure III-5: MPEG-4 AVC saliency map (on the left) vs. density fixation map (on the right).

The KLD and AUC values are reported in Figure III-6, Figure III-7 and Table III-4, respectively. In such an experiment, the lower the KLD value, the better the *Precision*; conversely, the larger the AUC value, the better the *Precision*.

In Figure III-6, the abscissa corresponds to nine saliency maps: the six MPEG-4 AVC maps introduced in Chapter III.1.3 (namely the *Skewness-max*, *Combined-avg*, *Multiplication-avg*, *Addition-avg*, *Static-avg*, and *Motion*) and the three investigated state of the art methods. The coordinate corresponds to the average KLD values (averaged both over the GOP in an individual video sequence and over all the processed video sequences), plotted in black squares. These average values are presented alongside with their upper and lower 95% confidence limits (plotted in red and green lines) as well as with their minimal and maximal values (over all the frames in the corpus), plotted in purple and blue stars.

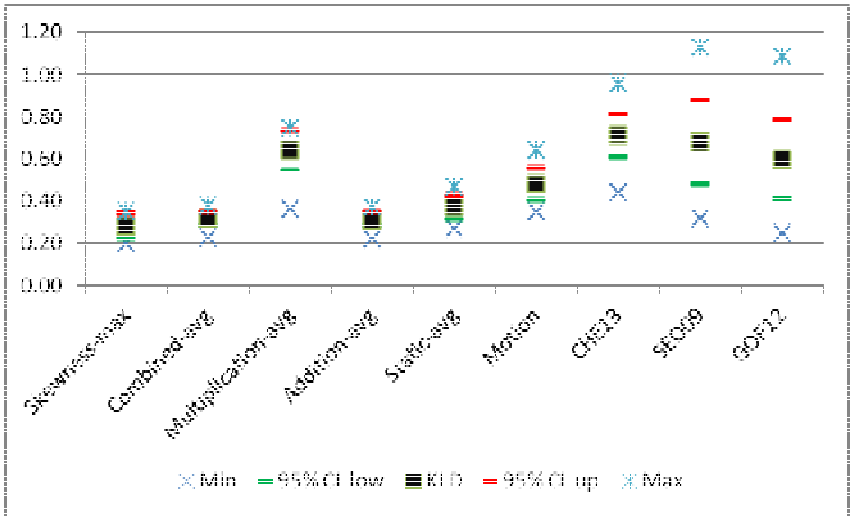


Figure III-6: KLD between saliency map and density fixation map.

The average values reported in Figure III-6 show that the lower KLD values correspond to MPEG-4 AVC saliency maps: *Skewness-max*, *Combined-avg* and *Addition-avg*. This amelioration over the state of the art methods is statistical relevant: the confidence limits for the *Skewness-max*, *Combined-avg* and *Addition-avg* do not overlap with the confidence limits corresponding to the three investigated state of the art methods.

The gain over the state of the art methods can be assessed by defining the coefficient ϱ:

$$\varrho_{MiMj} = \frac{KLD_{Mj} - KLD_{Mi}}{KLD_{Mj}} \qquad \text{(III-6)}$$

where M_i stands for an MPEG-4 AVC saliency maps (e.g. *Skewness-max*, *Combined-avg* and *Addition-avg*) while M_j stands for a state of the art saliency map. A positive ϱ_{MiMj} value means that the M_i map outperforms (in the KLD sense) the M_j map.

The quantitative results are presented in Table III-2, where the columns correspond to the same MPEG-4 AVC saliency map while the rows to the same state of the art method. It can be noticed that the best results are provided by the *Skewness-max* which outperforms the three considered state of the art methods [CHE13][SEO09][GOF12] by relative gains of 0.6, 0.58 and 0.53, respectively.

Table III-2: KLD gains between Skewness-max, Combined-avg and Addition-avg and the state of the art methods [CHE13] [SEO09] [GOF12].

	Skewness-max	*Combined-avg*	*Addition-avg*
[CHE13]	0.60	0.28	0.37
[SEO09]	0.58	0.52	0.50
[GOF12]	0.53	0.39	0.31

Figure III-6 also brings to light that the confidence limits corresponding to MPEG-4 AVC predicted saliency maps are narrower than the ones corresponding to the three investigated state of the art methods. Consequently, the KLD computation seems less sensitive to the randomness of the processed visual content in the MPEG-4 AVC domain. In order to objectively assess this behavior, we followed the principles in [FRY65], [WAL89] (also see the discussion in Chapter III.2.2.1), and we defined the coefficient ζ_{KLD} based on the signal-to-noise ratio for the random variable modeling the KLD computation:

$$\zeta_{KLD,MiMj} = \frac{KLD_{Mi}}{KLD_{Mj}} \cdot \frac{\sigma_{KLD,Mj}}{\sigma_{KLD,Mi}} \qquad \text{(III-7)}$$

where M_i stands for an MPEG-4 AVC saliency maps, M_j stands for a state of the art saliency map, and σ represent the standard deviation in the KLD computation. The larger the ζ_{KLD} coefficient, the less sensitive is the KLD on the randomness of the processed visual content.

The values corresponding to the *Skewness-max*, *Combined-avg* and *Addition-avg* predicted maps and to the three state of the art methods are presented in Table III-3 and show relative gains between 1.43 (corresponding to the *Combined-avg* / [CHE13] comparison) and 6.12 (corresponding to the *Skewness-max* / [GOF12] comparison).

Table III-3: KLD sensitivity gains between Skewness-max, Combined-avg and Addition-avg and the state of the art methods [CHE13] [SEO09] [GOF12].

	Skewness-max	*Combined-avg*	*Addition-avg*
[CHE13]	2.79	1.43	1.46
[SEO09]	5.81	2.91	2.97
[GOF12]	6.12	3.02	3.12

Figure III-7 is structured in the same way as Figure III-6: the abscissa corresponds to the nine investigated saliency maps while the ordinate to the AUC average/confidence limits/extreme values. In Figure III-7, the AUC study is carried out by considering a binarization threshold of max/2 (where max is the maximum value of the density fixation map).

The experimental results reported in Figure III-7 show that the *Skewness-max* outperforms all the other 9 investigated saliency maps; here again, the results are statically relevant (in the sense of the confidence limits).

The gain over the state of the art methods can be assessed by defining the coefficient η:

$$\eta_{MiMj} = \frac{AUC_{Mi} - AUC_{Mj}}{AUC_{Mj}} \tag{III-8}$$

where M_i stands for *Skewness-max* saliency map while M_j stands for any of the three state of the art saliency maps. A positive η_{MiMj} value means that the M_i map outperforms (in the AUC sense) the M_j map. When comparing the *Skewness-max* to the three state of the art methods [CHE13], [SEO09], and [GOF12] on the basis of the η coefficient, the following values are obtained 0.21, 0.18, and 0.17, respectively.

The sensitivity of the AUC with the randomness of the processed visual content was evaluated at the same way as in the KLD case, by defining the ζ_{AUC} coefficient:

$$\zeta_{AUC,MiMj} = \frac{AUC_{Mi}}{AUC_{Mj}} \cdot \frac{\sigma_{AUC,Mj}}{\sigma_{AUC,Mi}} \tag{III-9}$$

where M_i stands for an MPEG-4 AVC saliency maps, M_j stands for a state of the art saliency map; σ represent the standard deviation in the AUC computation. The larger the ζ_{AUC} coefficient, the less sensitive the AUC on the randomness of the processed visual content is. When computing the ζ_{AUC} coefficient between *Skewness-max* and the three state of the art methods, relative gains by factors of 33.70, 29.83 and 3.22 are thus obtained.

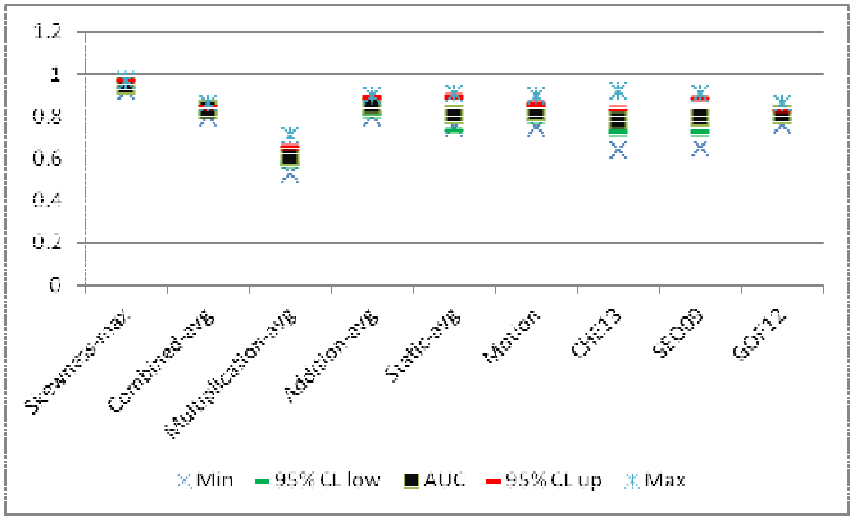


Figure III-7: AUC between saliency map and density fixation map.

Our study also investigates the impact of the choice of the binarization threshold in the AUC average values, see Table III-4. In this respect, 5 additional threshold values are considered, namely the percentiles of 90%, 80%, 70%, 60% and 50%. By combining the results presented in Table III-4 and Figure III-7, it can be stated that the binarization threshold of max/2 reaches maximal AUC values for all the nine investigated saliency maps (in the statistical relevance sense).

Table III-4: AUC values between saliency map and density fixation map with different binarization thresholds.

	90%	80%	70%	60%	50%	max/2
Skewness-max	0.89	0.86	0.83	0.81	0.75	0.95
Combined-avg	0.81	0.84	0.83	0.82	0.78	0.83
Multiplication-avg	0.63	0.68	0.67	0.66	0.59	0.61
Addition-avg	0.84	0.87	0.86	0.86	0.81	0.85
Static-avg	0.84	0.85	0.85	0.85	0.80	0.81
Motion	0.68	0.74	0.72	0.73	0.69	0.82
[CHE13]	0.64	0.69	0.69	0.67	0.69	0.78
[SEO09]	0.75	0.79	0.81	0.82	0.72	0.80
[GOF12]	0.80	0.82	0.79	0.81	0.79	0.81

Discriminance

In this sub-section, we investigate the usefulness of the saliency maps, i.e. its ability to discriminate between human fixation locations and random locations in a video content. In other words, we

investigate how selecting locations in the image according to a saliency map is better than selecting the same number of locations on a random basis, see Figure III-8.

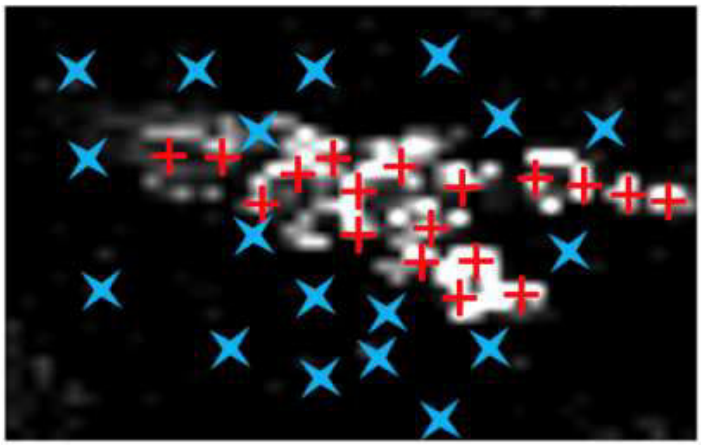


Figure III-8: Saliency map behavior at human fixation locations (in red + signs) vs. saliency map behavior at random locations (in blue x signs).

In this respect, the same two measures (KLD and AUC) are computed. This time, the larger the KLD measure, the better the *Discriminance* is (i.e. the more different the saliency selected from the randomly chosen locations). However, the AUC interpretation will be the same as in the previous experiment: the larger the AUC, the better the *Discriminance*. Actually, what changes now is the computation of the true positive and the false positive rates included in the AUC definition.

For each frame in the video sequence, we considered N=100 trials; hence, the statistical description (average/confidence limits/min/max) are this time computed over both all the frames and, for each frame, over all trials.

Two corpora will be alternatively considered, namely the *reference* corpus and the *cross-checking* corpus.

Reference results

The experimental results obtained on the *reference* corpus are presented in Figure III-9, Figure III-10 and Table III-6.

Figure III-9 shows the KLD values between the saliency map in fixation-selected locations and random selected locations. The abscissa corresponds to the same nine investigated saliency maps (cf. Figure III-6). The ordinate presents the average values, the lower and upper 95% confidence limits as well as their minimal and maximal values. The *Multiplication-avg* and the state of the art methods [CHE13] and [SEO09] give the best results. Although some differences in the average values exist (*Multiplication-avg* providing the best result), these differences are not statistical relevant (the confidence limits for *Multiplication-avg* and the state of the art methods [CHE13] and [SEO09] overlap).

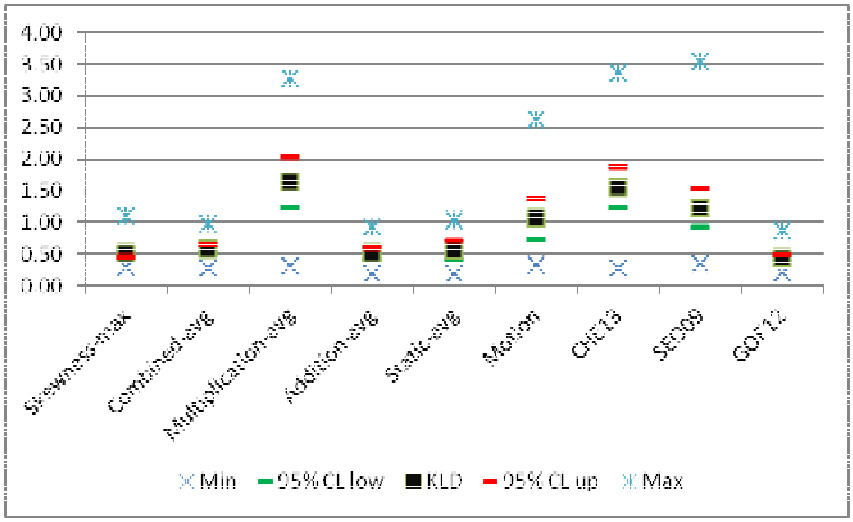


Figure III-9: KLD between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence).

We also investigated the sensitivity of the KLD with the randomness of the processed visual content, by considering the ζ_{KLD} coefficient, Eq. (III-7), between the *Multiplication-avg* and the three state of the art methods; relative gains of 1.78, 2.31 and 1.90 are thus obtained.

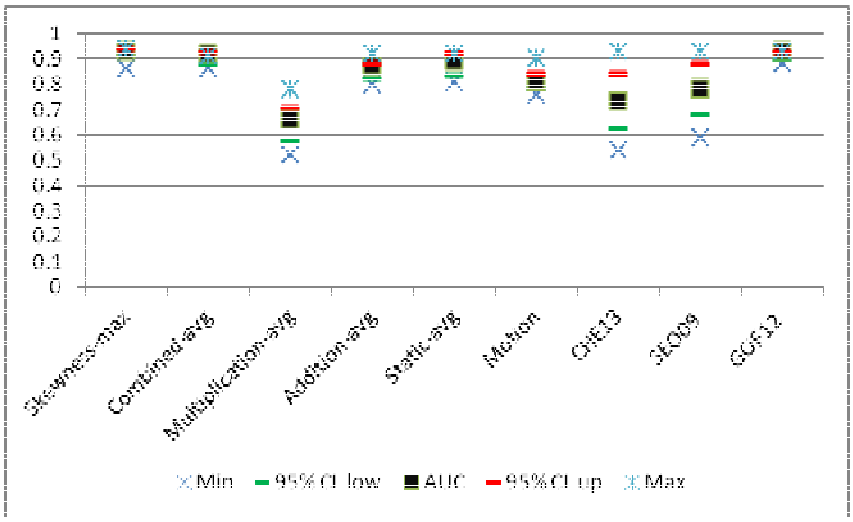


Figure III-10: AUC between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence).

Figure III-10 presents the AUC values corresponding to the saliency map in fixation-selected locations and random selected locations. The same experimental conditions as in Figure III-9 are retained: nine saliency maps and N=100 random trials for each frame. The binarization threshold was max/2 (we implicitly assumed the generality of the results in table III-4). According to the values plotted in Figure III-

10, the best saliency maps are *Skewness-max*, *Combined-avg* and the state of the art method [GOF12]: they feature the largest average AUC value and their confidence limits do not overlap with other investigated saliency maps.

The sensitivity of the AUC measure with the randomness of the visual content was investigated by computing the ζ_{AUC} coefficient, Eq. (III-9), among and between the two outperformers in the MPEG-4 AVC domain (*Skewness-max* and *Combined-avg*) and the three investigated state of the art methods. The results filled-in Table III-5 show relative gains between 1.06 (corresponding to the *Skewness-max* / [GOF12] comparison) and 2.02 (corresponding to the *Combined-avg* / [CHE13] comparison).

Table III-5: AUC sensitivity gains between Skewness-max and Combined-avg and the state-of-the-art methods [CHE13][SEO09][GOF12].

	Skewness-max	Combined-avg
[CHE13]	1.59	2.02
[SEO09]	1.38	1.76
[GOF12]	1.06	1.34

Table III-6: AUC values between saliency map at fixation locations and saliency map at random locations with different binarization thresholds (N=100 trials).

	90%	80%	70%	60%	50%	max/2
Skewness-max	0.87	0.85	0.83	0.81	0.79	0.93
Combined-avg	0.91	0.90	0.89	0.86	0.87	0.92
Multiplication-avg	0.65	0.64	0.63	0.59	0.58	0.66
Addition-avg	0.91	0.90	0.88	0.86	0.86	0.87
Static-avg	0.88	0.87	0.86	0.84	0.84	0.89
Motion	0.81	0.79	0.76	0.74	0.73	0.75
[CHE13]	0.78	0.77	0.76	0.74	0.76	0.73
[SEO09]	0.89	0.86	0.81	0.78	0.78	0.78
[GOF12]	0.92	0.91	0.89	0.87	0.86	0.93

Table III-6 investigates the impact of the choice of the binarization threshold in the AUC average values; in this respect, we kept the same 6 threshold values as in Table III-4, namely the percentile of 90%, 80%, 70%, 60%, 50% and max/2. Although the general tendency is the same as in Table III-4, the values reported in Table III-6 show a larger dependency of the AUC values on the binarization thresholds:

- *Skewness-max*, *Combined-avg*, *Multiplication-avg*, *Static-avg*, and [GOF12] have the largest AUC values for max/2.
- *Addition-avg*, *Motion*, [CHE13] and [SEO09] give the best results for the threshold of 90%.

However, the overall conclusion is the same; the best results in statistical sense are provided by *Skewness-max*, *Combined-avg* and the state of the art study [GOF12].

Cross-checking results

The results reported, previously in Chapters III.2.1.2 and *Reference results* in III.2.1.3 are obtained out of processing the so-called *reference* corpus. They brought to light that, according to both of the evaluation criteria (*Precision* and *Discriminance*) and to the considered measure (KLD or AUC), the MPEG-4 AVC saliency extraction outperforms or, is at least as good as, the state of the art methods. The quantitative results are obtained with statistical relevance (in the sense of the confidence limits). However, they are a priori dependent on the investigated video corpus; consequently, we resumed our experimental work on another publicly available corpus [WEB06], referred to in our study as the *cross-checking* corpus.

Besides its composition, this corpus also differs from the *reference* corpus in the type of the recorded human visual attention: while the *reference* corpus comes across with density fixation maps, the *cross-checking* provides the saccade locations. Consequently, we can only resume our study on *Discriminance* and not the one on *Precision*.

Except from the corpus, all the other experimental conditions as considered in the *Reference results* in Chapter III.2.1.3 are kept:

- the same nine saliency extraction models;
- the same KLD and AUC (with max/2 binarization threshold) with N=100 random trials;
- the same statistical entities: average value, lower/upper 95% confidence limits, and the minimal and maximal values;
- the same interpretation : the larger the KLD and AUC, the better the *Discriminance*.

The results are reported in Figures III-11 and III-12.

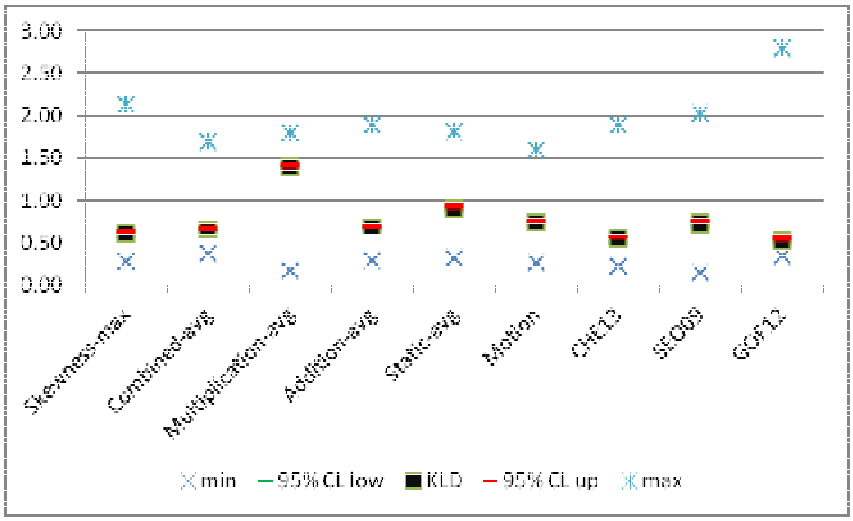


Figure III-11: KLD between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence).

According to KLD values in Figure III-11, the best results (in a statistical relevant sense) are featured by *Multiplication-avg* and *Static-avg*. The gains over the three state of the art methods, computed according to the ϑ coefficient, Eq. (III-6), are presented in Table III-7. The KLD sensitivity with respect to the randomness of the visual content was analyzed by computing the ζ_{KLD}, Eq. (III-7), among and between *Multiplication-avg* and *Static-avg* and the three state of the art methods. The experimental results reported in Table III-8 demonstrate relative gains between 1.18 (corresponding to the *Static-avg* / [CHE13] comparison) and 2.06 (corresponding to the *Multiplication-avg* / [GOF12] comparison).

Table III-7: KLD gains between Multiplication-avg and Static-avg and the three state of the art methods [CHE13][SEO09][GOF12].

	Multiplication-avg	***Static-avg***
[CHE13]	1.54	0.71
[SEO09]	0.91	0.25
[GOF12]	1.64	0.76

Table III-8: KLD sensitivity gains between Multiplication-avg and Static-avg and the three state of the art methods [CHE13][SEO09][GOF12].

	Multiplication-avg	***Static-avg***
[CHE13]	1.66	1.18
[SEO09]	1.75	1.24
[GOF12]	2.06	1.47

According to the AUC values reported in Figure III-12, the best (statistically significant) results are provided by *Skewness-max*; it outperforms the three state of the art methods by η, Eq.(III-8) , gains of 0.04, 0.17, 0.17. When computing the ζ_{AUC} coefficient, Eq. (III-9), between *Skewness-max* and the three state of the art methods, relative gains by factors of 1.34, 1.63 and 1.38 are thus obtained.

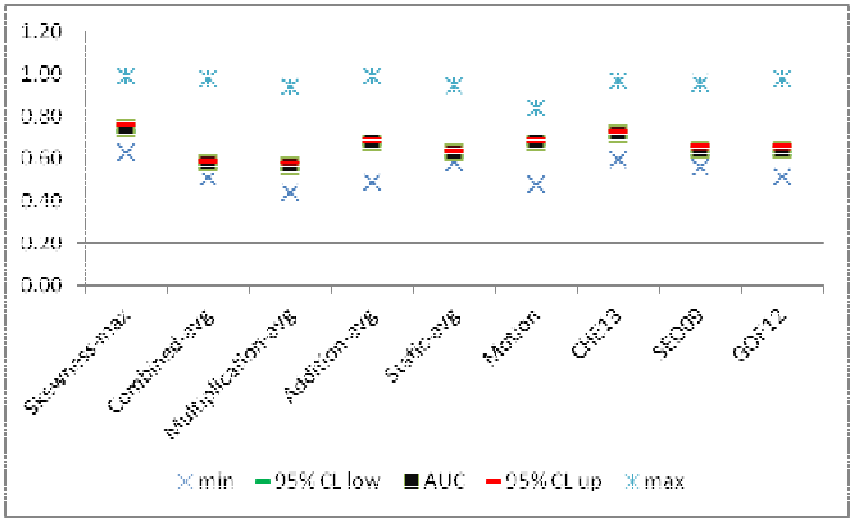


Figure III-12: AUC between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence).

III.2.2. Applicative validation

While the benchmarking of the MPEG-4 saliency model advanced in Chapter III.1 was based on the ground truth evidence, the present section will investigate the benefit of extracting the saliency directly from the video stream when deploying robust watermarking applications. Actually, by using the MPEG-4 AVC saliency model as criteria for selecting regions in which the mark is to be inserted, gains in transparency (for prescribed data payload and robustness properties) are expected.

In order to investigate the transparency, we fix the data payload (namely 30, 60, and 90 bits per *I* frame) and the robustness (namely bit error rates - BER - of 0.07, 0.03, and 0.01 against transcoding, resizing and Gaussian attacks respectively) and we evaluate the transparency for those two cases: (1) the watermarked blocks are randomly selected and (2) the watermarked blocks are selected among the blocks detected as non-salient by the best saliency map in the *Precision* sense (see Chapter III.2.1), namely the *Skewness-max* saliency map. Note that none of the state of the art saliency maps can be used for applicative benchmarking: they require decoding the MPEG-4 AVC stream in order to extract the saliency, thus slowing down the watermarking procedure.

The experimental study considers the multi-symbol quantization index modulation watermarking (m-QIM) method [HAS14] and both objective (Chapter III.2.2.1) and subjective (Chapter III.2.2.2) transparency evaluation criteria.

The watermarking corpus consists of 6 videos sequences of 20 minutes each. They were encoded with MPEG-4 AVC Baseline Profile (no *B* frames, CAVLC entropy encoder) at 512 kb/s. The GOP size is set to 8 and the frame size is set to 640×480 (according to the experiments in [HAS14]).

Objective transparency evaluation

The objective evaluation of the transparency considers three quality metrics of three different types: difference-based (PSNR), correlation based (NCC) and human psycho-visual based (DVQ).

These measures are computed at the frame level, and then averaged over all the frames of the video sequence and over all sequences in the corpus. The results are presented in Table III-9; the precision of the reported values (unit for PSNR and DVQ and 0.01 for NCC) is chosen so as to ensure the statistical significance of the results (95% confidence limits).

The analysis of the PSNR results shows that blocks selected according to our MPEG-4 AVC saliency map are more suitable for carrying the mark than random selected blocks: absolute gains of 10dB, 7dB and 3dB are obtained for the three investigated data payload (30, 60 and 90 bits/*I* frame).

The NCC values do not clearly discriminate between the random and the *Skewness-max* based selected blocks.

In order to assess the increase in the transparency according to the DVQ values, we define the relative coefficient ε:

$$\varepsilon = \frac{DVQ_{random} - DVQ_{skewness-max}}{DVQ_{skewness-max}} \quad \text{(III-10)}$$

Relative gains of 0.8, 0.68 and 0.71 are thus obtained for the three investigated data payload values.

Table III-9: Objective quality evaluation of the transparency when alternatively considering random selection and "Skewness-max" saliency map based selection.

	Data payload (bit per I frame)	*Random selection*					*Skewness-max based selection*				
		min	*95% down*	*mean*	*95% up*	*max*	*min*	*95% down*	*mean*	*95% up*	*max*
PSNR	30	34.76	50.44	51	51.56	64.07	40.32	60.53	61	61.47	68.97
	60	33.98	45.89	47	48.11	64.67	37.63	53.72	54	54.28	69.74
	90	36.08	44.08	45	45.92	62.98	36.96	47.67	48	48.33	66.93
NCC	30	0.98	0.99	1	1	1	0.98	0.99	1	1	1
	60	0.97	0.98	0.99	1	1	0.98	0.99	1	1	1
	90	0.96	0.98	0.99	1	1	0.98	0.99	0.99	1	1
DVQ	30	1280	1478	1490	1502	1753	203	292	297	302	416
	60	1520	1800	1809	1818	2064	480	559	567	575	830
	90	2030	2506	2515	2524	2780	653	699	713	727	816

Subjective transparency evaluation

The visual quality is assessed in laboratory conditions, according to a SSCQE (Single Stimulus Continuous Quality Evaluation) methodology proposed by the ITU R BT 2021. The test was conducted on a total of 30 naïve viewers. The age distribution ranges from 19 to 30 years old with an average of 23. All observers are screened for visual acuity by using Snellen chart and color vision by using Ishihara test. No outlier is identified, according to the kurtosis coefficient [TUR12]. The experiments considered a 5 level discrete grading scale.

At the beginning of the first session, 2 training presentations are introduced to stabilize the observers' opinion. The data issued from these presentations are not taken into account in the results of the test.

The MOS (Mean Opinion Score) values are presented in Table III-10; they correspond to the original video (data payload of 0 bit per *I* frame) as well as to the three investigated data payload values as in objective quality evaluation.

The values in Table III-10 show that, for a data payload of 30 bits per *I* frame, there is practically a very small difference between the scores assigned by the observers to the original content and to the content watermarked based on the *Skewness-max* saliency map; with respect to the random selection, this correspond to a MOS gain of 0.23.

Table III-10: MOS gain between the QIM method with random selection and saliency map "Skewness-max" based selection.

	Data payload (bit per I frame)	*Random selection*	*Skewness-max based selection*
MOS	0	3.38	
	30	3.11	3.34
	60	3.12	3.14
	90	2.95	2.97

When considering a data payload of 60 and 90 bit per *I* frames, the *Skewness-max* benefit becomes marginal (a MOS gain of 0.01). These results bring to light a kind of saturation behavior: for large data payloads, lots of blocks are watermarked inside the *I* frame, hence the difference between the random and saliency selection becomes less effective.

III.3. Discussion on the results

Chapter III.2.1 is devoted to ground truth validation, investigating the relation between the MPEG-4 AVC saliency map and the actual human saliency, captured by eye-tracking devices. It is based on two corpora (representing density fixation maps and saccade locations), two objective criteria called *Precision* and *Discriminance* (related to the closeness between the predicted and the real saliency maps and to the difference between the behavior of the predicted saliency map in fixation and random locations, respectively), two objective measures (the Kullback Leibler Divergence and the area under the ROC curve, respectively) and three state of the art studies (namely [CHE13], [SEO09], [GOF12]).

For both the KLD and AUC, we compute the average values (both over the GOP in an individual video sequence and over all the processed video sequences), and the related standard deviations, 95% confidence limits and minimal/maximal values. The ratio between the average value and the standard deviation (the so-called signal to noise value [FRY65], [WAL89]) was computed so as to assess the sensitivity of the KLD and AUC with respect to the randomness in the processed visual content. In order to compare the predicted MPEG-4 AVC saliency map to the state of the art methods, we define two types of coefficients, see equations (III-6) - (III-9), which are point-estimated.

The overall results are synoptically presented in Table III-11, which regroups, for each and every investigated case, the best methods (in the sense of the investigated measures and the statistical relevance).

Table III-11: Ground truth validation results

Ground truth validation: best results					
Precision		***Discriminance***			
Reference corpus		*Reference* corpus		*Cross-checking* corpus	
KLD	*AUC*	KLD	AUC	KLD	AUC
Skewness-max, Combined-avg, Addition-avg	*Skewness-max*	*Multiplication-avg,* [CHE13], [SEO09]	*Skewness-max, Combined-avg,* [GOF12]	*Multiplication-avg, Static-avg*	*Skewness-max*

For instance, the ground truth results related to *Precision* and *Discriminance*, exhibit absolute relative gains, defined according to Eq. (III-6) and Eq. (III-9), over state of the art methods:

- in KLD: between 60% (corresponding to *Precision*, the *reference* corpus and the *Skewness-max* / [CHE13] comparison) and 164% (corresponding to *Discriminance*, the *cross-checking* corpus and the *Multiplication-avg* / [GOF12] comparison),
- in AUC: between 17% (corresponding to *Precision*, the *reference* corpus and the *Skewness-max* / [GOF12] comparison) and 21% (corresponding to *Precision*, the *reference* corpus and the *Skewness-max* / [CHE13] comparison).

We also investigated the sensitivity of the measure (KLD and AUC) with respect to the randomness in the visual content. When compared to the state of the art methods, the experimental results show gains in sensitivity by factors:

- in KLD, between 1.18 (corresponding to *Discriminance*, the *cross-checking* corpus and the *Static-avg* / [CHE13] comparison) and 6.12 (corresponding to *Precision*, the *reference* corpus and the *Skewness-max* / [GOF12] comparison),
- in AUC, between 1.06 (corresponding to *Discriminance*, the *reference* corpus and the *Skewness-max* / [GOF12] comparison) and 33.7 (corresponding to *Precision*, the *reference* corpus and the *Skewness-max* / [CHE13] comparison).

All these above-reported values objectively and quantitatively demonstrate the usefulness of extracting saliency maps from the compressed domain. A closer qualitative inspection of the compressed domain saliency maps reveals an additional interesting behavior of such models. When considering bottom-up saliency models, two paths can be found in literature: (1) algorithms inspecting particular areas by maximizing local saliency on the basis of a biologically inspired ground and (2) algorithms more focused on global features, detecting saliency through transform domains. Global features should be predominant in identifying salient areas under the condition that the image contains obviously isolated foreground objects (the “pop-outs”), whereas local features are more important in an opposite situation.

Nevertheless, during the whole process of human perception, the human brain is able at the same time to combine together and to make complete global and local features. Consequently, a good bottom-up model should also be able to handle this dual behavior (local vs. global); in this respect, a qualitative analysis of our experimental results show (as illustrated in Figure III-13):

- [CHE13] succeeds in identifying all the global “pop out” objects, but lacks in precision for finer areas (e.g., Figure III-13, image (c) in the second example, the people inside the bus are considered as salient as the whole bus or as other objects in the scene);
- [SEO09] is more selective at the object level but presents an integration effect over various objects (e.g., Figure III-13, image (d) in the first example, all the players are identified as a unique, salient region);

- Compared to [CHE13] and [SEO09], [GOF12] seems both more precise and discriminative at the global object level; nevertheless, it is still not able to identify at the same time areas with different saliency sources (e.g. Figure III-13, image (e) in the third example, the players in black who are salient because of their motion, cannot be detected);
- The strength of our method seems to be achieved by its joint capacity to identify very localized salient areas (individual sub-parts from more global "pop out" objects) and to detect areas featured by different types of saliency; for instance, in Figure III-13, image (b) of the fourth example, only some details of the ducks are represented as salient while in Figure III-13, line 3, we succeeded in also detecting moving players in black.

Chapter III.2.2 relates to the applicative validation and considers the integration of the compressed-domain saliency map into a robust watermarking application: in order to increase the transparency, for a prescribed data payload and robustness, the mark is inserted into non-salient blocks, according to the predicted MPEG-4 AVC saliency map. This time, no state of the art saliency extraction method can be considered as reference for the applicative validation: as the mark is to be inserted directly in the MPEG-4 AVC stream, we can only rely on the saliency map advanced with this study. Hence, our study investigates the gains obtained when considering saliency-guided insertion with respect to blind (no saliency based) insertion.

The experiments show that the saliency prediction in the MPEG-4 AVC domain results in:

- objective study: an increase in PSNR and DVQ (up to 10dB and up 70%, respectively); the NCC measure did not exhibit a clear benefit of using saliency-guided insertion;
- subjective study: the MOS corresponding to the saliency-guided watermark insertion can approach by 0.04 the MOS corresponding to the original (un-watermarked content); a saturation mechanisms for large data payloads has also been spotted out.

However, the final advantage of any image processing method is also given by its computational complexity. Table III-12 compares the three state of the art methods investigated in Chapter III.2.1 to our MPEG-4 AVC saliency extraction method: the main operations included in both static and dynamic saliency maps are listed. An additional benefit from the MPEG-4 AVC saliency is thus brought to light: it can be achieved with a linear complexity (assuming the entropic decoding available).

In order to also provide a quantitative illustration of the practical impact of these differences in the computational complexity among the four investigated saliency methods, we also measured the computational time per processed frame. In this respect, we averaged the frame execution time over all video frames in two video sequences. We considered a PC configuration with an Intel Xeon 3.7GHz processor and with 8 GB of RAM. These values, expressed in milliseconds, are reported in Table III-13. The unit precision chosen in Table III-13 ensures that these values are statistical relevant (i.e. their 95% confidence limits variations are lower than 1). Note that in MPEG-4 AVC saliency detection case, the execution time values corresponding to the six investigated pooling formulas are identical (i.e. their differences are lower than the precision in their 95% confidence limits); consequently, in Table III-13 we reported only one value, which holds for any of the six pooling formulas we studied. We emphasize that

Table III-13 has only an illustration purpose: the codes for the four investigated methods are of two types (C/C++ and Matlab) and none of them is optimized for execution speed.

Table III-12: Computational complexity comparison between our method and the three state of the art models considered in our study.

	Spatial map	*Dynamic map*
[CHE13]	• Complete decoding of the images • Decomposing images into large scales perceptually homogenous elements using GMM	⊗
[SEO09]	• Complete decoding of the videos • Compute the local steering kernel and vectorize it into different features	• Motion vector extraction
[GOF12]	• Complete decoding of the videos • Decomposing images into patches • Multiscale saliency enhancement • K-nearest neighbor (KNN)	• Motion vector extraction
MPEG-4 AVC	• Addition and gradient of 4×4 blocks	• Motion vector difference

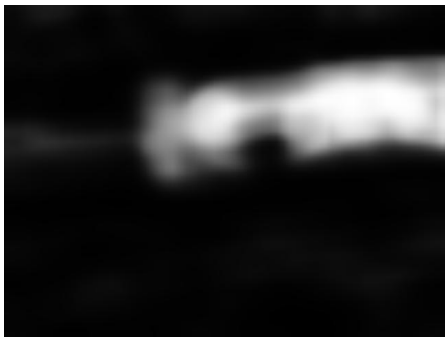
Table III-13: Computational time per processed frame of our method and the three state of the art models considered in our study.

	Computational time (in milliseconds)	*Type of code*
[CHE13]	24	C/C++
[SEO09]	1 170	Matlab
[GOF12]	35 002	Matlab
MPEG-4 AVC	9	C/C++

(a) Original image

(b) Our MPEG-4 AVC saliency map

(c) [CHE13]



(d) [SEO09]

(e) [GOF12]

(a) Original image

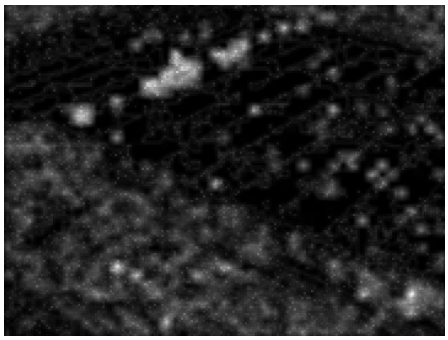
(b) Our MPEG-4 AVC saliency map

(c) [CHE13]

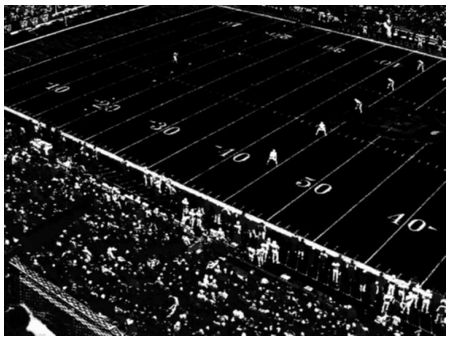
(d) [SEO09]

(e) [GOF12]

(a) Original image

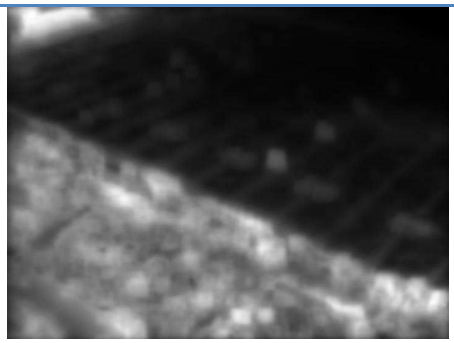


(b) Our MPEG-4 AVC saliency map



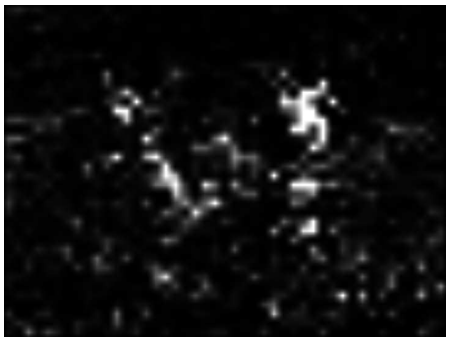
(c) [CHE13]

(d) [SEO09]



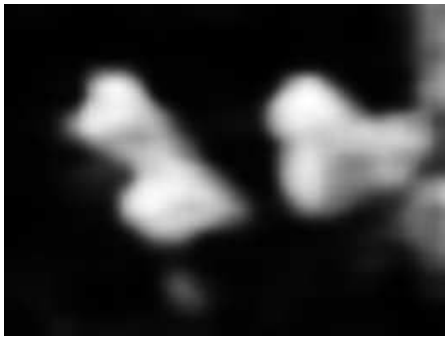
(e) [GOF12]

(a) Original image

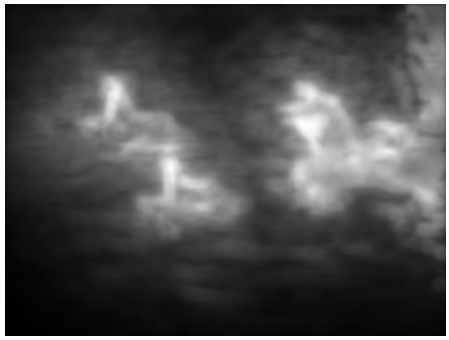


(b) Our MPEG-4 AVC saliency map

(c) [CHE13]



(d) [SEO09]



(e) [GOF12]

Figure III-13: Illustrations of saliency maps computed with different models.

III.4. Conclusion

This Chapter presents a comprehensive framework for establishing the proof of concept for saliency extraction from the MPEG-4 AVC syntax elements (before entropic coding).

From the methodological point of view, we adapt and extend the state of the art principles so as to match them to the MPEG-4 AVC stream syntax elements, thus making possible individual intensity, color, orientation, and motion maps to be defined. Several pooling formulas have been investigated.

The experimental validation takes place at two levels: ground-truth confrontation and applicative integration. The ground truth validation is based on two criteria, the so-called *Precision* (which can be useful when we aim to predict the human fixation locations) and *Discriminance* (which prove its efficiency when aiming to be as different as possible from the random locations). For each criterion, we considered two objective metrics, namely the KLD (a distance related to the statistical differences) and AUC (a measure related to the probability of error in detection). The ground truth itself is represented by two state of the art corpora, containing both fixation and saccade information. The applicative validation considers the MEPG-4 AVC saliency map as a tool guiding the mark insertion.

As an overall conclusion, the study brings to light that although the MPEG-4 AVC standard does not explicitly rely on any visual saliency principle, its stream syntax elements preserve this property. Among possible explanations for this remarkable property, one could argue a share feature between video coding and saliency. Saliency is often considered as a function of singularity (of contrast, color, orientation, motion …). On coding side, singularities are usually uncorrelated signals with their vicinities making them hard to encode and leading to more residues. Considering that there is this relationship between saliency and coding cost, a good encoder could possibly act as a winner take all approach revealing, emphasizing salient information. Mimicking such behavior in the spatial domain is not that trivial and often under considered in many approaches provided in literature.

This conclusion is supported by all our experiments, which brought to light four main benefits for the MPEG-4 AVC based saliency extraction: (1) it outperforms (or, at least, is as good as) state of the art uncompressed domain methods, (2) it allows significant gains to be obtained in watermarking transparency (for prescribed data payload and robustness), (3) it is less sensitive to the randomness in the processed visual content, and (4) it has a linear computational complexity.

IV. Saliency extraction from HEVC stream

This Chapter goes one step further and investigates whether the information related to the human visual saliency is still preserved at the level of the HEVC compressed stream. In this respect, the saliency model presented in Chapter III is reconsidered and extended so as to match the HEVC peculiarities. The same experimental test-bed as in Chapter III is considered in order to both compare the HEVC saliency to the ground-truth and to assess its applicative impact in watermarking. It is thus brought to light that the HEVC saliency model outperforms (with singular exceptions) the state-of-the-art uncompressed domain while generally being outperformed by the MPEG-4 AVC saliency model. We can thus state that, as its MPEG-4 AVC ancestor, although not designed based upon visual saliency principles, the HEVC compression standard preserves this human visual property at the level of its syntax elements.

IV.1. HEVC saliency map computation

The emerging HEVC (High Efficiency Video Coding) standard brings improvements over MPEG-4 AVC, so as to increase the compression capabilities, especially for high resolution formats [SUL12]. In this respect, HEVC offers more flexible prediction and transform block sizes, larger choice in prediction modes, more sophisticated signaling of motion vectors and more advanced interpolation filtering for motion compensation.

HEVC video sequences are structured, the same way as MPEG4-AVC, into Groups of Pictures (GOP). A GOP is composed of an *I* (intra) frame and a number of successive *P* and *B* frames (unidirectional predicted and bidirectional predicted frames, respectively).

A frame in HEVC is partitioned into coding tree units (CTUs), each of them covering a rectangular area up to 64x64 pixels depending on the encoder configuration. Each CTU is divided into coding units (CUs) that are signaled as intra or inter predicted blocks. A CU is then divided into intra or inter prediction blocks. For residual coding, a CU can be recursively partitioned into transform blocks (TB).

The HEVC saliency map definition is structured at three levels.

First, the HEVC stream syntax elements are investigated according to their a priori potentiality to be connected to the visual saliency. Note that, in this respect, the extension from MPEG-4 AVC to HEVC is not straightforward. On the one hand, HEVC allows different block sizes to be defined (see Figure IV-1); consequently the energy conservation theorem invoked in the MPEG-4 AVC intensity and color map definitions should be reconsidered and adapted to this new applicative configuration. On the other hand, both intra and inter prediction modes are changed, thus imposing a detailed investigation on the orientation and motion maps. The inter prediction modes are now structured into two classes (advanced motion vector prediction and merge modes) thus making a priori the motion saliency detection dependent on the encoding configuration.

In our work, we start from the MPEG-4 AVC saliency maps computation basic principles. Three elementary static maps are extracted (intensity, color, orientation). In order to obtain a compressed stream video saliency map, we complete the obtained elementary static saliency maps with a motion saliency map. For each GOP, we extract the saliency map only from *I* and *P* frames. The static saliency map is computed from the *I* frame. The intensity and color maps are extracted from the residual HEVC luma and chroma coefficients, respectively, while the orientation map is computed based on the intra prediction modes. The motion map is generated based on the motion vectors from the *P* frames.

For the reasons explained in Chapter III, it is not necessary to compute the saliency from *B* frames. Moreover, *B* frames are not considered in our experimental study. Nevertheless, our method can be applied to any HEVC video configuration, be it with or without *B* frames.

The computing of each map as well as their post-processing and pooling are detailed in the following sub-sections.

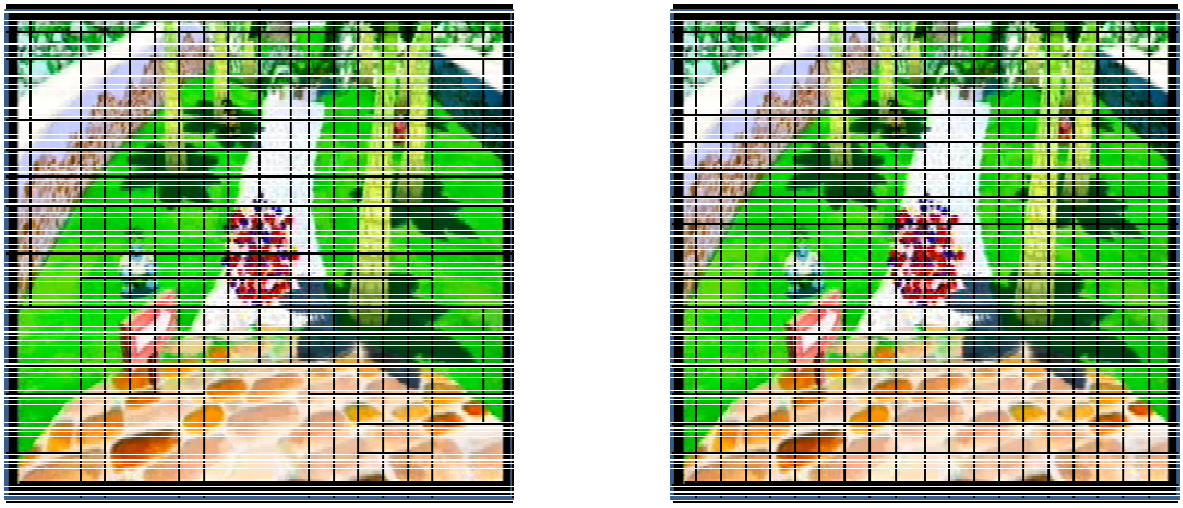


Figure IV-1: Difference between HEVC and MPEG-4 AVC block composition.

IV.1.1. HEVC elementary saliency maps

Intensity map

When defining the HEVC saliency map, we also consider that the luma residual coefficients are related to the center-surround mechanism featured by the human visual system (see Chapter III.1).

In our work, the intensity map in MPEG-4 AVC video stream is defined by computing the energy luminance for each 4x4 luma transform block. Such a technique would not be appropriate in the context of a varying transform block sizes as in HEVC, where several transform block sizes are supported: 4x4, 8x8, 16x16 and 32x32. The basic transform coding process of the prediction residual in HEVC is very similar to that of MPEG4-AVC. It is based on integer DCT basis functions, except for 4x4 luma transform blocks, in which case a DST (Discrete Sine Transform)-based transform is performed.

To compute the intensity saliency map from HEVC video stream, two steps are required. We first compute the luminance energy of the transform block (TB) and then we calculate the luminance energy of each 4×4 region inside the TB.

We extract the transformed and quantified luma coefficients for each TB directly from the compressed stream. By applying the energy conservation property between DCT or DST transformed and spatial domain, the luminance energy of a TB is computed according to:

$$M_{iTB} = \sum_{u=1}^{s}\sum_{v=1}^{s} {Y_{u,v}}^2 \qquad \text{(IV-1)}$$

where *s x s* is the size of TB, *u* and *v* are coefficient coordinates and *Y* is the luma residual coefficient.

We calculate the luminance energy of a 4×4 region inside TB as following:

$$M_i(k) = M_{iTB}/N \quad \text{(IV-2)}$$

where *k* is the 4x4 region index in the frame and *N* is the total number of 4x4 regions in TB. The intensity map will be obtained by displaying M_i; the highest values represent the salient blocks.

Color map

Thorough analogy to the way in which the intensity saliency was defined, color saliency will be based on color energy.

In the MPEG-4 AVC case, the chroma residual coefficients are first extracted. The color information (*Cr*,*Cb*) is then used to calculate the two opponent color pairs RG (Red/Green) and BY (Blue/Yellow). Finally, we compute the color saliency map as the sum of the energy in the double color-opponent RG and BY space. For the same reason as for intensity map, this technique is not appropriate with HEVC stream.

The chroma TB size of HEVC is half the luma TB size in each dimension, except when the luma TB size is 4x4, (in which case a single 4x4 chroma TB is used for the region covered by four 4x4 luma TBs).

To compute color saliency map from HEVC video stream, only chroma DC coefficients, which represent the average color of the chroma transform block TB, are extracted. First, we calculate, for each 4×4 region inside TB, a color average for each of the chroma color components *Cr* and *Cb*.

$$DC^c(k) = \sqrt{\frac{(DC^c_{TB})^2}{N}} \quad \text{(IV-3)}$$

where *k* is the 4x4 region index in the frame, *c* is the color component, DC_{TB} is the *DC* coefficient in *TB* and *N* is the total of the 4x4 regions in *TB*.

Then, based on the average color, we calculate the average opponent-color pairs RG_k and BY_k for the associated 4x4 region *k*. Finally, the color map is computed according to:

$$M_c(k) = {RG_k}^2 + {BY_k}^2 \quad \text{(IV-4)}$$

The color conspicuity map will be obtained by displaying M_c, the highest values represent the salient blocks.

Orientation map

With respect to MPEG-4 AVC, changes in the intra prediction process are introduced in HEVC, concerning both the prediction block sizes and the prediction modes. HEVC supports variable intra prediction block sizes from 64x64 down to 4x4. As MPEG-4 AVC, DC and planar mode are defined, while intra angular prediction directions are augmented from 8 to 33.

According to intra HEVC paradigm, the prediction modes reflect the orientation of the corresponding block with respect to its neighboring blocks. The orientation map will be computed by analyzing the discontinuities among the intra prediction modes of intra frame blocks: blocks which feature the same direction as their neighborhood are considered as non-salient while blocks with different orientation modes are considered as salient.

The building of the orientation map starts by analyzing the intra prediction block sizes. Large intra prediction blocks are considered as non-salient regions. In the remaining cases, values of the prediction modes are extracted; then, the obtained orientation for each 4×4 block will be compared to those obtained for a set of neighboring blocks.

The M_o orientation map is computed according to:

$$M_o(k) = \begin{cases} Card(\{O_l = O_k; \forall l \in V\}) \; if \;\; PB\ size \leq 8 \times 8 \\ 0 \quad else \end{cases} \tag{IV-5}$$

where *k* is the block index in the frame, *V* is the set of neighboring block and *l* is the block index belonging to *V*.

Motion map

In addition to the advanced motion vector prediction presented in prior standards, HEVC defines a new inter prediction mode: the merge mode, which derives the motion information from spatially and temporally neighboring blocks. Compared to MPEG-4 AVC, HEVC includes asymmetric motion partitioning and share the accuracy of motion compensation, which is in units of one quarter of the distance between luma samples.

For each GOP, we define the motion saliency map from HEVC stream as the global motion difference amplitude, computed by summing the motion amplitude over all the *P* frames in the GOP, at the same corresponding block position:

$$M_m(k) = \sum_{P \in GOP} \sqrt{{MVDx_k}^2 + {MVDy_k}^2} \tag{IV-6}$$

where $(MDVx_k, MVDy_k)$ denote horizontal and vertical components of motion vectors difference in the *P* frame block *k*, and M_m represents the global motion amplitude among the *P* frames in a GOP; the larger this $M_m$value, the more salient the block *k*.

IV.1.2. Elementary saliency map post-processing

The saliency maps obtained for each feature are now to be normalized to the same dynamic range. This is achieved by following the three same saliency map steps approach we considered for MPEG-4 AVC, Chapter III.1.2 (Figure III-4).

First, outlier detection is performed: the 5% largest and the 5% lowest values are eliminated. Then the remaining values are mapped to the [0 1] interval through an affine transform. Finally, an average filtering, with the window size equal to the fovea area is applied.

In the case of the orientation map where its values belong to [0 1], the first two post-processing operations are skipped.

IV.1.3. Saliency maps pooling

The HEVC saliency map is a fusion of the static and the dynamic saliency maps. The static saliency map is in its turn a combination of intensity, color and orientation features maps. As we have seen in Chapter III, the fusing formula has a critical role in the final result, thus the same fusing techniques are applied to obtain the HEVC saliency map.

We start our study on the HEVC saliency map fusion techniques by investigating 48 different pooling formula combinations (6 static pooling formula and, for each of them, 8 dynamic pooling) [AMM16], detailed in Appendix A.2. The most accurate combination (in the sense of KLD and AUC computed on a ground truth database of 80 sec) is *Motion-priority* static-dynamic fusion over the static maximum fusion referred to us *Motion priority-max*. For the assessment, we retain the *Motion priority-max* and we include as well the same fusion techniques investigated in Chapter III (*Combined-avg*, *Multiplication-avg*, *Addition-avg*, *Static-avg*, *Motion*).

IV.2. Experimental results

Our experiments are structured on two directions (ground truth and applicative validations). We considered the same test-bed as the MPEG-4 AVC case, on which we evaluate the performances of six alternative ways of combining the elementary maps described above: *Motion priority-max, Combined-avg*, *Multiplication-avg*, *Addition-avg*, *Static-avg*, *and Motion*.

IV.2.1. Ground truth validation

Test-bed

Through analogy with our work in Chapter III, the experiments will be structured at two nested levels, according to the evaluation criteria and to the actual measures and corpora:

- both *Precision* (the closeness between the saliency map and the fixation map) and *Discriminance* (the difference between the behavior of the saliency map in fixation locations and in random locations) are considered;
- two measures (KLD and AUC) are considered to assess the obtained saliency maps (same implementation as used in Chapter III);
- the average values (computed first over the GOPs in an individual video sequence and then over all the processed video sequences), the related standard deviations, 95% confidence limits and minimal/maximal values are computed;
- the sensitivity of the KLD and AUC with respect to the randomness in the processed visual content is evaluated;
- two different corpora are considered and further referred to as: (1) the *reference* corpus available in [WEB05] and (2) the *cross-checking* corpus available in [WEB06].

The *reference* corpus is a public database organized by IRCCyN [WEB05]. In these experiments, videos are encoded with HEVC Main Profile (no *B* frames, CABAC entropy encoder) and with a quantification parameter Q_p = 32. The GOP size is set to 5 and the frame size is set to 576×720. The HEVC reference software is completed with software tools allowing the parsing of the syntax elements and their subsequent usage, under syntax preserving constraints. The same encoding configuration is considered for the *cross-checking* corpus [WEB06].

During our experiments, we benchmark our HEVC saliency maps against the same three state of the art methods, namely: Ming Cheng *et al*. [CHE13], Hae Seo *et al.* [SEO09] and Stas Goferman [GOF12], whose MATLAB codes are available for downloading. In addition, we confront the HEVC saliency maps to the MPEG-4 AVC saliency map in each experience.

Precision

In this experiment, we compare the computed HEVC saliency maps to the density fixation maps captured from the human observers (as explained in the previous chapter). The *reference* corpus [WEB05] will be processed.

The KLD and AUC values are reported in Figure IV-2 and Figure IV-3 respectively. The lower the KLD value, the better the *Precision.* Conversely, the larger the AUC value, the better the *Precision.*

In Figure IV-2, the abscissa corresponds to ten saliency maps: the six HEVC maps previously introduced (namely the *Motion priority-max*, *Combined-avg*, *Multiplication-avg*, *Addition-avg*, *Static-avg*, and *Motion*), the three investigated state of the art methods and the retained MPEG-4 AVC saliency map. The ordinate corresponds to the average KLD values (averaged both over the GOPs in an individual video sequence and over all the processed video sequences), plotted in black squares. These average values are presented alongside with their upper and lower 95% confidence limits (plotted in red and green lines) as well as with their minimal and maximal values (over all the frames in the corpus), plotted in purple and blue stars.

The average values reported in Figure IV-2 show that the lower KLD values correspond to MPEG-4 AVC saliency maps and two of HEVC fusion technique combination saliency maps: the *Combined-avg* saliency map and the *Addition-avg* saliency map. The improvement over the state of the art methods [CHE13][SEO09] is statistically relevant: the confidence limits for the *Combined-avg* and the *Addition-avg* saliency maps do not overlap with the confidence limits corresponding to both of the investigated state of the art methods [CHE13] and [SEO09].

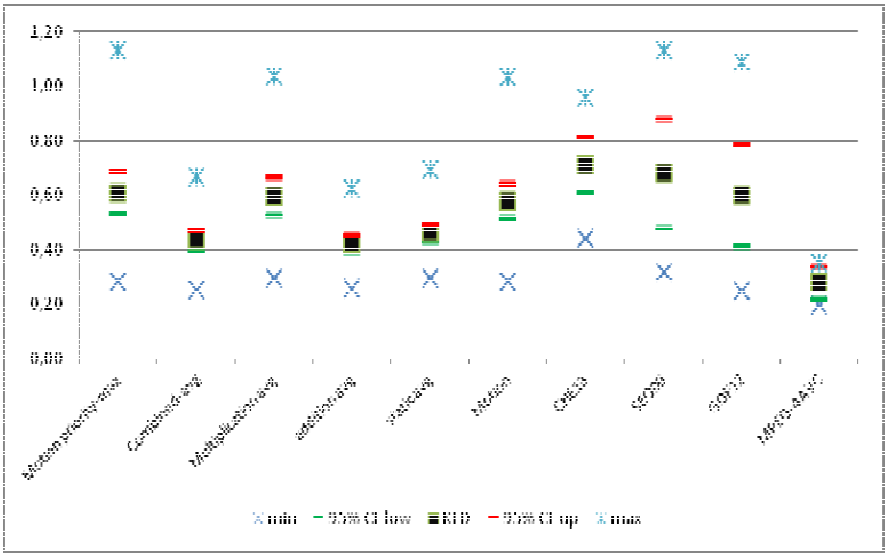


Figure IV-2: KLD between saliency map and density fixation map.

Same as in the MPEG-4 AVC saliency extraction chapter, the gain over the other saliency extraction methods (the three state of the art and the MPEG-4 AVC methods) can be assessed by computing the coefficient ϑ, Eq. (III-6), between HEVC saliency maps and the state of the art and the MPEG-4 AVC saliency maps. A positive value implies that the HEVC map outperforms (in the KLD sense) the related map.

The quantitative results are presented in Table IV-1, where the columns correspond to the same HEVC saliency maps while the rows to the same state of the art methods and MPEG-4 AVC saliency map. It shows that all the HEVC saliency maps give better results than the state of the art methods (expect

Motion priority-max against [GOF12]) but the MPEG-4 AVC saliency map outperforms all of them. The best HEVC saliency map results are provided by *Combined-avg* and *Addition-avg* which outperform the three considered state of the art methods, [CHE13], [SEO09], and [GOF12], by relative gains of 0.39, 0.36, and 0.27 and 0.40, 0.38 and 0.29, respectively.

Table IV-1: KLD gains between all the combination of HEVC saliency maps and the state of the art methods [CHE13] [SEO09] [GOF12] and MPEG-4 AVC saliency map.

	Motion priority-max	*Combined-avg*	*Multiplication avg*	*Addition avg*	*Static avg*	*Motion*
[CHE13]	0.14	0.39	0.16	0.40	0.35	0.19
[SEO09]	0.11	0.36	0.12	0.38	0.32	0.15
[GOF12]	-0.01	0.27	0.01	0.29	0.23	0.04
MPEG-4 AVC	-1.17	-0.56	-1.13	-0.51	-0.65	-1.06

Figure IV-2 brings to light that the confidence limits corresponding to HEVC predicted saliency maps are narrower than confidence limits corresponding to the three investigated state of the art methods. Consequently, the KLD computation seems less sensitive to the randomness of the processed visual content in the HEVC domain. In order to objectively assess this behavior, we calculate the ζ_{KLD}, Eq. (III-7), between the HEVC saliency maps and the state of the art saliency map. The larger the ζ_{KLD} coefficient is, the less sensitive is the KLD to the randomness of the processed visual content. The values corresponding to the different combinations of the HEVC saliency maps and the three outperformed state of the art are presented in Table IV-2 and show relative gains between 5.3 (corresponding to *Motion* / [CHE13] comparison) and 21.39 (corresponding to the *Multiplication-avg* / MPEG-4 AVC comparison).

Table IV-2: KLD sensitivity gains between all considered HEVC saliency map combinations and the state of the art methods [CHE13] [SEO09] [GOF12] and MPEG-4 AVC saliency map.

	Motion priority-max	*Combined-avg*	*Multiplication-avg*	*Addition-avg*	*Static-avg*	*Motion*
[CHE13]	6.53	7.20	8.44	8.15	5.47	5.30
[SEO09]	6.82	7.52	8.81	8.51	5.71	5.53
[GOF12]	7.73	8.52	9.98	9.64	6.47	6.27
MPEG-4 AVC	16.56	18.25	21.39	20.66	13.44	13.44

Figure IV-3 is structured the same way as Figure IV-2: the abscissa corresponds to the ten investigated saliency maps while the ordinate to the AUC average/confidence limits/extreme values. In Figure IV-3, the AUC study is carried out by considering a binarization threshold of max/2 (where max is the maximum value of the density fixation map).

The experimental results reported in Figure IV-3 show that all the HEVC saliency maps outperforms the three investigated state of the art methods while only the *Combined-avg*, the *Addition-avg* and the

Static-avg outperforms the MPEG-4 AVC saliency map; here again, the results are statically relevant against the three state of the art methods (in the sense of the confidence limits).

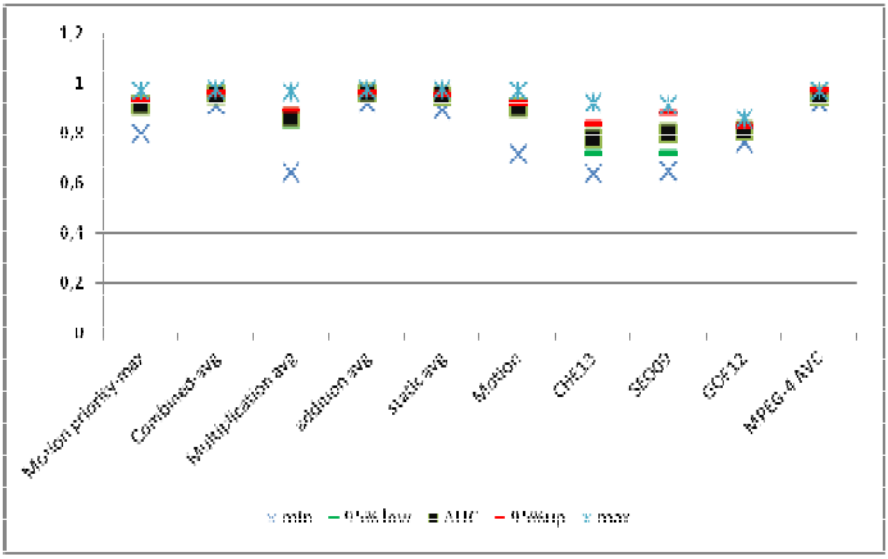


Figure IV-3: AUC between saliency map and density fixation map.

As in Chapter III, we compute the coefficient Eq. (III-8), to assess the gain in AUC of the HEVC saliency maps over the state of the art methods. Thus we obtained positive values against the state of the art methods and the MPEG-4 AVC saliency extraction model with three different HEVC saliency maps: *Combined-avg*, *Addition-avg* and *Static-avg,* see Table IV-3.

Table IV-3: AUC gains between all the combinations of HEVC saliency maps and the state of the art methods [CHE13] [SEO09] [GOF12] and MPEG-4 AVC saliency map.

	Motion priority-max	*Combined-avg*	*Multiplication-avg*	*Addition-avg*	*Static-avg*	*Motion*
[CHE13]	0.17	0.22	0.11	0.22	0.22	0.15
[SEO09]	0.14	0.20	0.08	0.20	0.19	0.12
[GOF12]	0.12	0.18	0.04	0.18	0.01	0.11
MPEG-4 AVC	-0.03	0.008	-0.09	0.01	0.001	-0.05

The sensitivity of the AUC to the randomness of the processed visual content was evaluated in the same way as in the KLD case, by calculating the ζ_{AUC} coefficient, Eq. (III-9).

When computing the ζ_{AUC} coefficient between HEVC saliency maps and the three state of the art methods and the MPEG-4 AVC saliency map, we obtain, as reported in Table IV-4, relative gains by

factors between 8.39 (corresponding to *Combined-avg* / MPEG-4 AVC comparison) and 15.12 (corresponding to *Addition-avg* / [CHE13] comparison).

Table IV-4: AUC sensitivity gains between Combined-avg, Addition-avg and Static-avg and the state of the art methods [CHE13] [SEO09] [GOF12] and MPEG-4 AVC saliency map.

	Combined-avg	*Addition-avg*	*Static-avg*
[CHE13]	12.77	15.12	15.01
[SEO09]	9.96	11.80	11.71
[GOF12]	12.30	14.56	14.45
MPEG-4 AVC	8.39	9.93	9.86

Discriminance

The effectiveness of the HEVC saliency map will be evaluated in this section by investigating its ability to discriminate between human fixation locations and random locations in a video content; in this respect:

- the KLD and AUC are computed; the same interpretation as in Chapter III.2.1 is considered, namely the larger the KLD and AUC measures are, the better is the *Discriminance*;
- 100 random trials are considered for each frame in each video sequence;
- both the *reference* and the *cross-checking* corpora are processed;
- the KLD and AUC average measures are presented alongside with the confidence limits and the related min/max values (over both all the frames and, for each frame, over all trials).

Reference results

The experimental results obtained on the *reference* corpus are presented in Figure IV-4 and Figure IV-5.

Figure IV-4 shows the KLD values between the saliency map in fixation-selected locations and random selected locations. The abscissa axis corresponds to the same ten investigated saliency maps (cf. Figure IV-2). The ordinate axis presents the average values, the lower and upper 95% confidence limits as well as their minimal and maximal values. The MPEG-4 AVC gives the best result against the three state of the art models and all the combination of the HEVC saliency map. These differences are not statistically relevant (the confidence limits for MPEG-4 AVC and the state of the art methods [SEO09] and [CHE13] overlap). The best result in HEVC saliency maps is given by the *Addition-avg* saliency map which outperforms the [GOF12] by a gain of 0.95.

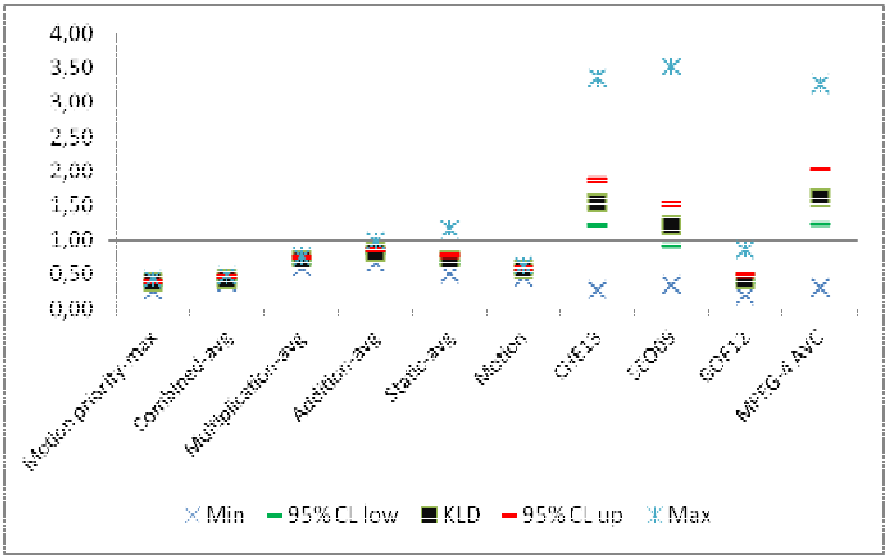


Figure IV-4: KLD between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence).

We also investigated the sensitivity of the KLD to the randomness of the processed visual content, by considering the ζ_{KLD} coefficient, Eq. (III-7), between the HEVC *Addition-avg* saliency map and the [GOF12]; relative gain of 6.84 is thus obtained.

Figure IV-5 presents the AUC values corresponding to the saliency map in fixation-selected locations and random selected locations. The same experimental conditions as in Figure IV-4 are kept (and the same as in Chapter III): ten saliency maps and N=100 random trials for each frame. The binarization threshold is max/2. According to the obtained values in Figure IV-5, the best saliency maps are *Motion priority-max*, the state of the art method in [GOF12], and the MPEG-4 AVC method: they feature the highest average AUC values.

The sensitivity of the AUC measure to the randomness of the visual content was investigated by computing the ζ_{AUC} coefficient, Eq. (III-9), among and between the *Motion priority-max* in the HEVC domain and the investigated saliency extraction methods (the state of the art methods and MPEG-4 AVC method). The results show relative gains of 0.96 against [CHE13] and 1.98 against [SEO09]).

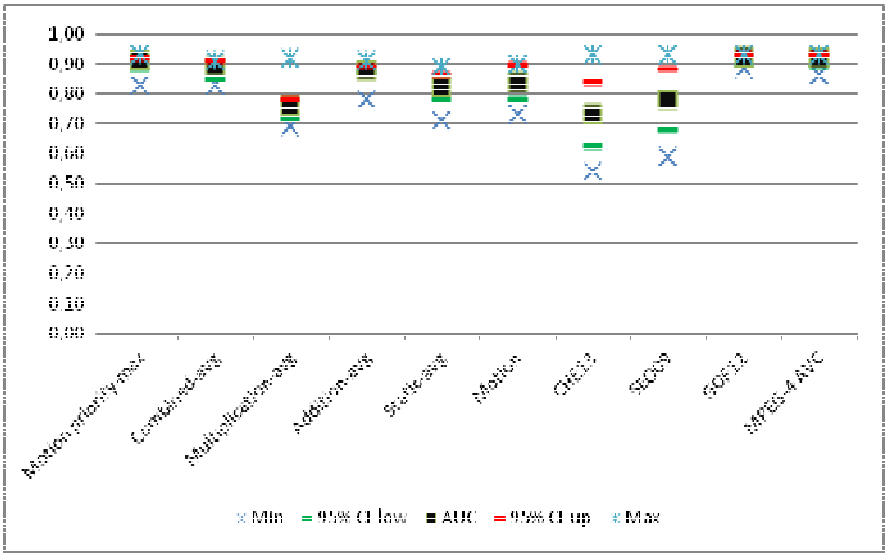


Figure IV-5: AUC between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence).

Cross-checking results

The experimental results obtained on the *reference* corpus are reported in Figures IV-6 and IV-7.

According to Figure IV-6, the best KLD result is given by the MPEG-4 AVC saliency map. The *Multiplication-avg* and the *Static-avg* feature the best results among the HEVC saliency maps. The values of the ϑ coefficient, Eq. (III-6) are presented in Table IV-5: gains are obtained only against [CHE13] and [GOF12].

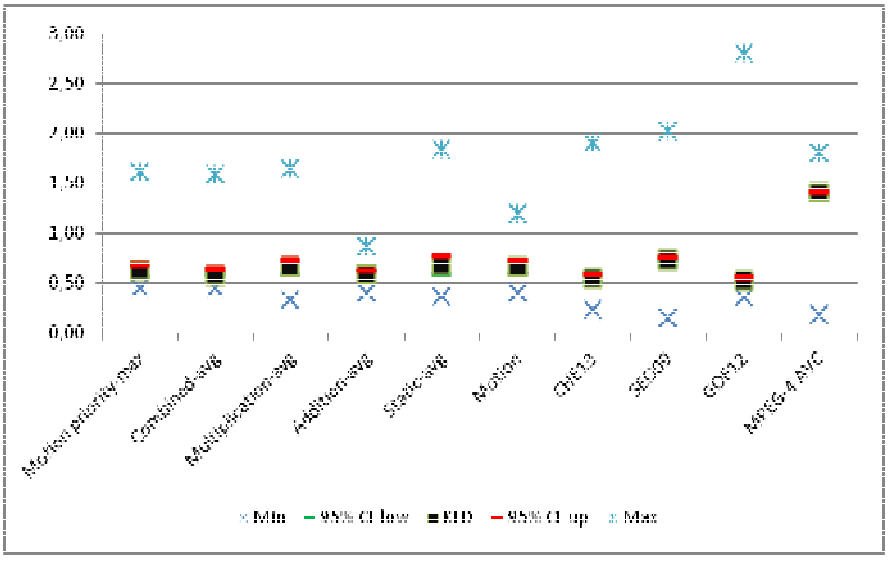


Figure IV-6: KLD between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence).

Table IV-5: KLD gains between Multiplication-avg and Static-avg and the state of the art methods [CHE13] [SEO09] [GOF12] and MPEG-4 AVC saliency map.

	Multiplication-avg	*Static-avg*
[CHE13]	0.20	0.23
[SEO09]	-0.07	-0.05
[GOF12]	0.24	0.28
MPEG-4 AVC	-0.74	-0.72

The KLD sensitivity with respect to the randomness of the visual content was analyzed by computing the ζ_{KLD} in Eq. (III-7) among and between *Multiplication-avg* and *Static-avg* and the same investigated methods. The experimental results reported in Table IV-6 demonstrate relative gains between 0.003 (corresponding to the *Static-avg* / MPEG-4 AVC comparison) and 0.75 (corresponding to the *Multiplication-avg* / [CHE13] comparison).

Table IV-6: KLD sensitivity gains between Multiplication-avg and Static-avg and the state of the art methods [CHE13] [SEO09] [GOF12] and MPEG-4 AVC saliency map.

	Multiplication-avg	*Static-avg*
[CHE13]	0.75	0.01
[SEO09]	0.002	0.003
[GOF12]	0.23	0.02
MPEG-4 AVC	0.01	0.003

According to the AUC values reported in Figure IV-7, the best (statistically significant) results are also provided by the MPEG-4 AVC saliency map; it outperforms all the compared models (HEVC saliency maps and the state of the art methods). Among the HEVC saliency maps, the best result was provided by the *Motion priority-max* which outperforms the three state of the art methods by η, Eq. (III-8), gains of 0.02, 0.1 and 0.1, respectively. Relative gains ζ_{AUC}, Eq. (III-9), of 0.47, 0.42 and 0.38 are thus obtained.

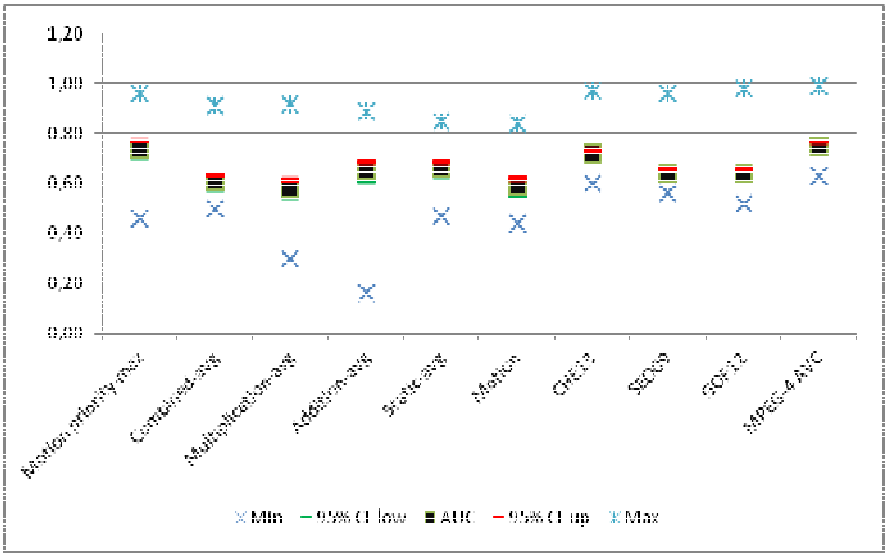


Figure IV-7: AUC between saliency maps at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence).

IV.2.2. Applicative validation

Our MPEG-4 AVC saliency method already proved its efficiency (Chapter III) in term of compressed stream saliency extraction by providing significant gains in a watermarking application. In Chapter IV.1, the HEVC saliency map is validated by a confrontation to the ground truth. However, we still need an investigation of the benefit of extracting the saliency directly from the HEVC compressed stream when deploying a watermarking application. As its predecessor, the HEVC saliency model will be used as criteria for selecting regions in which the mark is to be inserted; gains in transparency (for prescribed data payload) are expected.

In order to investigate the transparency, we fix two data payload (namely 30 and 50 bits per *I* frame) and we evaluate the transparency for those two cases: (1) the watermarked blocks are randomly selected and (2) the watermarked blocks are selected among the blocks detected as non-salient by the best saliency map in the *Precision* sense (see Chapter IV.2.1), namely the *Combined-avg* saliency map. Note that, as in Chapter III, none of the state of the art saliency maps can be used for applicative benchmarking: they require decoding the HEVC stream in order to extract the saliency, thus slowing down the watermarking procedure.

The experimental study considers a simple compressed stream watermarking application, where the mark is additively inserted in the last coefficient of a selected 16x16 transform blocks (TB). We considered here a 16x16 TB as inserting in a smaller TB size (4x4 or 8x8) will alter significantly the watermarked videos while inserting in a 32x32 TB cannot give a good evaluation of the saliency based selection method against the random selection method since those blocks are usually non salient (in HEVC compressed stream format, the 32x32 transform blocks represent homogenous regions).

The watermarking corpus discussed in Chapter III is here encoded with HEVC main Profile (no *B* frames, CABAC entropy encoder) and with Q_p=32. The GOP size is set to 5 and the frame size is set to 720×576.

Objective transparency evaluation

The objective evaluation of the transparency considers three quality metrics: the peak signal to noise ratio (PSNR) and the image fidelity (IF) as difference-based measure and the correlation quality (CQ) as a correlation based measure. These measures are computed at the frame level, averaged over all the frames of the video sequence and then over all sequences in the corpus. The results are presented in Table IV-7; the precision of the reported values (unit for PSNR and CQ and 0.01 for IF) is chosen so as to ensure the statistical significance of the results (95% confidence limits).

The analysis of the PSNR results shows that non-salient blocks selected using our HEVC saliency map are more suitable for carrying the mark than random selected blocks: absolute gains of 1.43dB and 1.69dB are obtained, respectively, for the two investigated data payload (30 and 50 bits/*I* frame).

However, the obtained CQ and IF values do not show a relevant improvement of the saliency based selection method over the random selection method.

Table IV-7: Objective quality evaluation of the transparency when alternatively considering random selection and "Combined-avg" saliency map based selection.

	Data payload (bit per I frame)	***Random selection***					***Saliency based selection***				
		min	*95% down*	*mean*	*95% up*	*max*	*min*	*95% down*	*mean*	*95% up*	*max*
PSNR	30	23.56	39.13	40.08	41.03	61.67	25.33	41.038	41.51	41.982	65.97
	50	26.78	37.09	37.83	38.57	59.73	27.45	38.81	39.52	40.23	66.34
CQ	30	187.92	198.87	201.53	204.18	216.56	189.38	199.68	201.41	203.13	217.51
	50	188.39	199.59	201.27	202.94	216.78	190.62	199.44	201.31	203.17	217.67
IF	30	0.963	0.997	0.9976	0.997	0.999	0.971	0.997	0.997	0.997	0.999
	50	0.956	0.995	0.996	0.996	0.999	0.965	0.996	0.997	0.997	0.999

Subjective transparency evaluation

The visual quality is assessed in laboratory conditions, according to the SSCQE (Single Stimulus Continuous Quality Evaluation) methodology proposed by the ITU R BT 2021. The test is conducted on a total of 25 naïve viewers. The age distribution ranges from 20 to 28 years old with an average of 25. All observers are screened for visual acuity by using Snellen chart and for color vision by using Ishihara test. No outlier is identified, according to the kurtosis coefficient [TUR12]. The experiments consider a 5 level discrete grading scale.

At the beginning of the first session, two training presentations are introduced to stabilize the observers' opinion. The data organized from these presentations are not taken into account in the final results of the test.

The MOS (Mean Opinion Score) values are presented in Table IV-8; they correspond to the original video (data payload of 0 bit per *I* frame) as well as to the three investigated data payload values as in objective quality evaluation.

The values in Table IV-8 show that the watermarking insertion based on saliency outperforms the random method. We obtained, for both 30 and 50 bits per *I* frame, a MOS value increased by 0.13 and 0.03.

Table IV-8: MOS gain between the watermarking method with random selection and saliency map "Combined-avg" based selection.

	Data payload (bit per I frame)	*Random selection*	*Saliency based selection*
MOS	0	3.79	
	30	3.63	3.79
	50	2.66	2.69

IV.3. Discussion on the results

Chapter IV is structured in the same way as in Chapter III in order to investigate whether the relation between the new compressed stream HEVC saliency map and the actual human saliency, captured by eye-tracking devices, will be the same as its predecessor MPEG-4 AVC. In this fact, the evaluation is based on:

- two corpora (representing density fixation maps and saccade locations),
- two objective criteria called Precision and Discriminance (related to the closeness between the predicted and the real saliency maps and to the difference between the behavior of the predicted saliency map in fixation and random locations, respectively),
- two objective measures (the Kullback Leibler Divergence and the area under the ROC curve)
- 3 state of the art studies (namely [CHE13], [SEO09], [GOF12]) and the MPEG-4 AVC saliency extraction model.
- For both the KLD and AUC, we compute the average values (both over the GOP in an individual video sequence and over all the processed video sequences), and the related standard deviations, 95% confidence limits and minimal/maximal values.
- Assessment of the sensitivity, using the same defined coefficient defined in Chapter III (Eq. (III-6)-Eq. (III-9)), of the KLD and AUC with respect to the randomness of the processed visual content.

The overall results are synoptically presented in Table IV-9, which regroups, for each and every investigated case, the best methods (in the sense of the investigated measures and the statistical relevance).

Table IV-9: Ground truth validation results

Ground truth validation: best results					
Precision		***Discriminance***			
Reference corpus		*Reference* corpus		*Cross-checking* corpus	
KLD	*AUC*	KLD	AUC	KLD	AUC
Combined-avg, Addition-avg, MPEG-4 AVC	*Combined-avg, Addition-avg, Static-avg*	*MPEG-4 AVC*	*Motion priority-max,* [GOF12], MPEG-4 AVC	*Multiplication -avg and the Static-avg,* MPEG-4 AVC	*Motion priority-max* and MPEG-4 AVC

For instance, the ground truth results related to *Precision* and *Discriminance*, exhibit absolute relative gains, defined according to Eq. (III-7) and Eq. (III-9), over the state of the art and the MPEG-4 AVC saliency extraction methods:

- in KLD: between 28% (corresponding to *Discriminance*, the *cross-checking* corpus and S*tatic-avg*/ [GOF12] comparison) and 40% (corresponding to *Precision*, the *reference* corpus and the *Addition-avg* / [CHE13] comparison),
- in AUC: between 2% (corresponding to *Discriminance*, the *cross-checking* corpus and the *Motion priority-max* / [GOF12] comparison) and 22% (corresponding to *Precision*, the *reference* corpus and the *Combined-avg* / [CHE13] comparison).

We also investigated the sensitivity of the KLD and AUC measures with respect to the randomness in the visual content. When compared to the state of the art methods, the experimental results show gains in related to sensitivity by:

- in KLD: between 0.01 (corresponding to *Discriminance*, the *reference* corpus and the *Static-avg* / [CHE13] comparison) and 9.98 (corresponding to *Precision*, the *reference* corpus and *Multiplication-avg* / [GOF12] comparison),
- in AUC: between 0.38 (corresponding to *Discriminance,* the *reference* corpus and *the Motion priority-max* / [GOF12] comparison) and 15.12 (corresponding to *Precision*, the *reference* corpus and *Addition-avg* / [CHE13] comparison)

All these above-reported values demonstrate, objectively and quantitatively, the usefulness of extracting saliency maps from the compressed domain.

As explained in Chapter III.3, the human brain is able at the same time to combine together and to make complete global and local features. Consequently, a good bottom-up model should also be able to handle this dual behavior (local vs. global). A qualitative analysis based on saliency models behavior was explained in Chapter IV.3 and presented by examples in Figure IV-8 (composed from four original image and for each of them the saliency maps computed according the HEVC, MPEG-4 AVC and the three state of the art methods [CHE13], [SEO09], and [GOF12]), cf. discussion in Chapter III.3.

Figure IV-8 shows that same as MPEG-4 AVC method, the HEVC method ensures identifying much localized salient areas (individual sub-parts from more global “pop out” objects) and detecting areas

featured by different types of saliency (e.g., Figure IV-8, image (b) in the fourth example, only some details of the moving persons are represented as salient while in Figure IV-8, image (b) in the third example we succeeded in detecting the face of the child in addition to the lights. Figure IV-8 can be compared to Figure III-13: we deliberately changed the original image so as to enrich the overall illustrations in the thesis.

Chapter IV.2.2 is related to the applicative validation and considers the integration of the HEVC saliency map into a robust watermarking application: in order to increase the transparency, for a prescribed data payload, the mark is inserted into non-salient blocks, according to the predicted HEVC saliency map. Hence, our study investigates the gains obtained when considering saliency-guided insertion with respect to blind (random) insertion.

The experiments show that the saliency prediction in the HEVC domain results in:

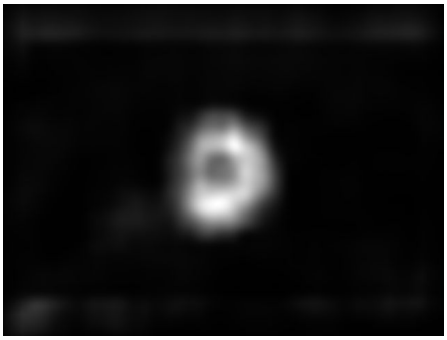
- objective study: an increase in PSNR by 1.55dB;
- subjective study: the MOS corresponding to the saliency-guided watermark insertion (30 bits per *I* frame) is equal to the MOS corresponding to the original video (un-watermarked content);

However, an important criterion and the final advantage of any image processing method is also given by its computational complexity. Compared to the models presented in Table III-12 (the three investigated state of the art and our MPEG-4 AVC saliency extraction methods), the HEVC saliency extraction algorithm uses the same main operations performed for generating static and dynamic MPEG-4 AVC saliency maps, with the difference of processing on TB with different sizes.

Moreover, we also measured the computational time of the C/C++ code of the HEVC saliency extraction model. We reported only one value, which holds for any of the six pooling formulas we studied, namely 11 milliseconds.

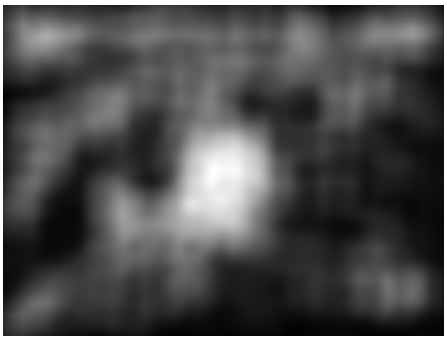
(a) Original image

(b) Our HEVC saliency map

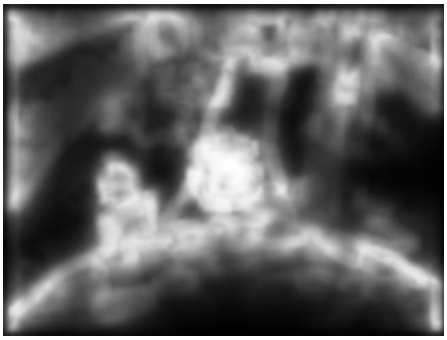


(c) Our MPEG-4 AVC saliency map

(d) [CHE13]



(e) [SEO09]

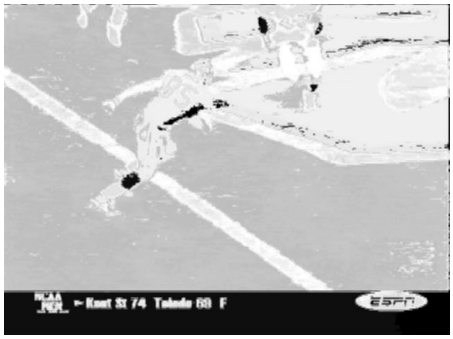


(f) [GOF12]

(a) Original image

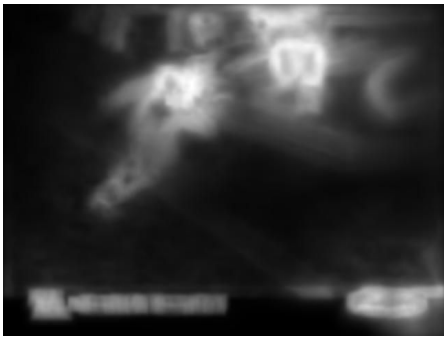
(b) Our HEVC saliency map

(c) Our MPEG-4 AVC saliency map



(d) [CHE13]

(e) [SEO09]



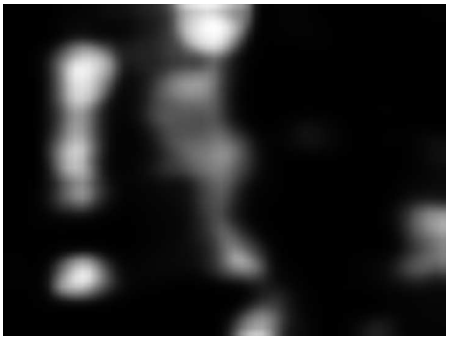
(f) [GOF12]

(a) Original image

(b) Our HEVC saliency map

(c) Our MPEG-4 AVC saliency map

(d) [CHE13]



(e) [SEO09]

(f) [GOF12]

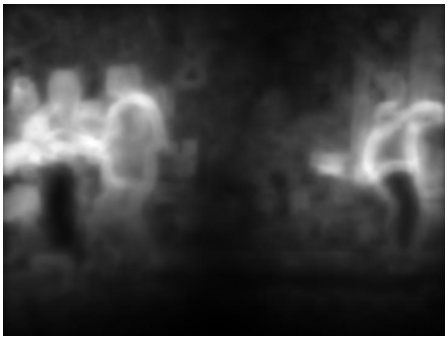
(a) Original image

(b) Our HEVC saliency map

(c) Our MPEG-4 AVC saliency map

(d) [CHE13]

(e) [SEO09]



(f) [GOF12]

Figure IV-8: Illustrations of saliency maps computed with different models.

IV.4. Conclusion

From the methodological point of view, we adapt and extend the MPEG-4 AVC saliency model principles so as to match them to the HEVC stream syntax elements, thus making possible individual intensity, color, orientation, and motion maps to be defined. Moreover, several pooling formulas have been investigated.

The experimental validation takes place under the same framework defined for MPEG-4 AVC: ground-truth confrontation and applicative integration. The ground truth validation is based on two criteria, the *Precision* and *Discriminance.* For each criterion, we considered two objective metrics, namely the KLD and AUC. The ground truth itself is represented by two state of the art corpora, the first one is featured by fixation information and the second one by saccade information. The applicative validation is an integration of the HEVC saliency map in a compressed stream watermarking framework that considers the saliency map as a tool guiding the mark insertion.

The main benefits of computing the saliency directly at the stream level are the same as in the MPEG-4 AVC case, namely, performance (confrontation to the ground truth) with respect to the state of the art methods, gains in watermarking transparency, sensitivity to the randomness in the processed visual content, and linear computational complexity.

V. Conclusion and future work

V.1. Conclusion

The present thesis aims at offering a comprehensive methodological and experimental view about the possibility of extracting the salient regions directly from video compressed streams (namely MPEG-4 AVC and HEVC), with minimal decoding operations. The peculiarities of each of these two domains were studied in Chapters III and IV, respectively: the related methodology was presented alongside with in-depth experiments (both ground truth and applicative validations) and the detailed conclusions were drawn in Chapter III.4 and IV.4, respectively.

However, as studied in the Introduction and beyond the technical anchors, the present thesis is about studying two a priori conceptual contradictions (see Chapter II). The first contradiction corresponds to the saliency extraction from the compressed stream. On the one hand, saliency is given by visual singularities in the video content. On the other hand, in order to eliminate the visual redundancy, the compressed streams are no longer expected to feature singularities. The second contradiction corresponds to saliency guided watermark insertion in the compressed stream. On the one hand, watermarking algorithms consist on inserting the watermark in the imperceptible features of the video. On the other hand, lossy compression schemes try to remove as much as possible the imperceptible data of video.

Consequently, the remaining of this Chapter will present the thesis point of view on these two contradictions.

V.1.1. Saliency vs. Compression

As an overall conclusion, the study brings to light that although the MPEG-4 AVC and HEVC standards does not explicitly rely on any visual saliency principle, their stream syntax elements preserve this property.

Among possible explanations for this remarkable property, one could argue a share feature between video coding and saliency. Saliency is often considered as a function of singularity (of contrast, color, orientation, motion …). On coding side, singularities are usually uncorrelated signals with their vicinities making them hard to encode and leading to more residues. Considering that this relationship between saliency and coding cost holds, a good encoder could possibly act as a *winner take all* approach revealing, emphasizing salient information. Mimicking such behavior in the compressed domain is not that trivial and often under-considered in many approaches provided in literature.

In order to investigate whether such a behavior is proper to MPEG-4 AVC and HEVC, we also consider the case of MPEG-4 ASP format [WEB11]. Actually, as explained in Chapter II, the study of Fang [FAN14], published during the development of the present thesis, deals with saliency extraction in the transformed domain.

We then evaluated the Fang's model under the same test-bed as the MPEG-4 AVC and HEVC. Table V-1 illustrates the KLD and AUC values, for the three state of the art methods acting in the uncompressed

domain ([CHE13], [SEO09], and [GOF12]) and the three methods acting in the compressed domains (our MPEG-4 AVC saliency model, our HEVC saliency model and the methods of Fang [FAN14]). Both *Precision* (*reference corpus*) and *Discriminance* (*reference corpus* and *cross-checking corpus*) results are presented.

Table V-1 shows that, in term of *Precision* and both KLD and AUC, the compressed stream saliency extraction models outperform the uncompressed stream models.

When considering the *Discriminance*, the results also go in the same directions but with a more nuanced tendency, Table V-2. Actually, for the *reference corpus*, the KLD values show that MPEG-4 AVC outperforms other methods while the AUC values show that both MPEG-4 AVC and the uncompressed domain model [GOF12] give the best results. However, for the *cross-checking corpus*, the KLD results bring to light the MPEG-4 AVC and [FAN14] methods as the best solutions while the AUC points to the supremacy of the three compressed-domain methods (HEVC, MPEG-4 AVC, and [FAN14]).

This investigation reinforces our results and proves that, contrarily to our expectation, the compressed domain saliency extraction models have greater performance than the uncompressed domain saliency extraction model. This behavior a posteriori demonstrates the very need and the value of our overall proof of concepts study presented in the thesis: the simple intuition is not able to a priori state whether and how saliency extraction in MPEG-4 AVC would outperform saliency extraction from pixels, from MPEG-4 ASP and from the very sophisticated HEVC compression format!

Table V-1: Comparison of the results of KLD and AUC between saliency maps and fixation maps.

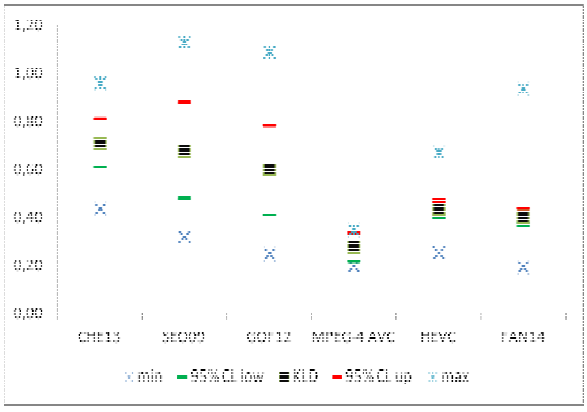
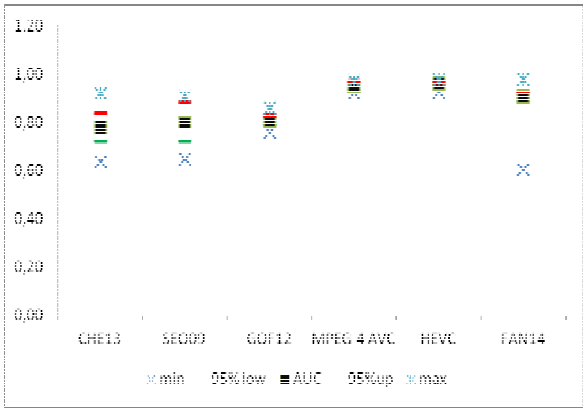
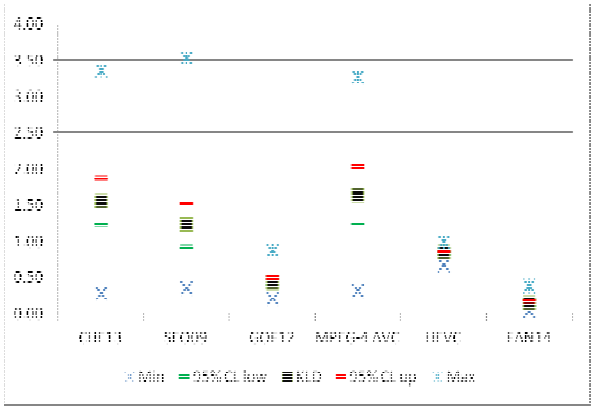
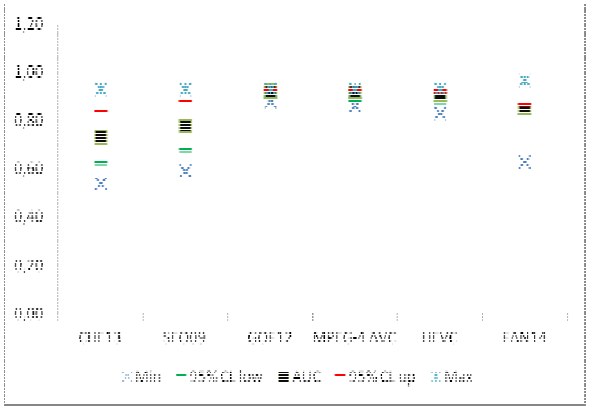
Precision	
Reference corpus	
	

Table V-2: Comparison of the results of KLD and AUC between saliency maps at fixation locations and saliency maps at random locations (N=100 trials for each frame in the video sequence).

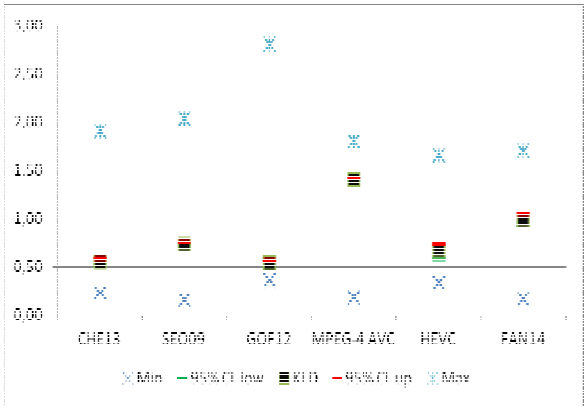
Discriminance

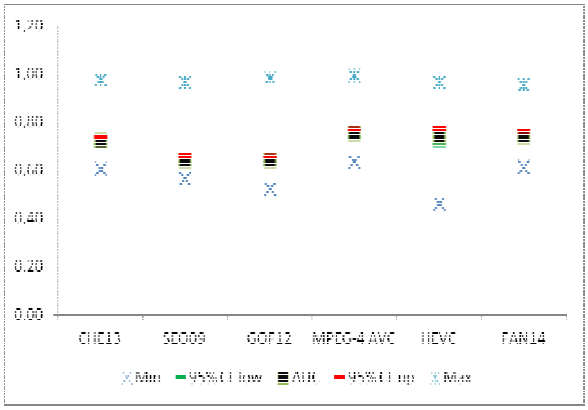
Reference corpus





Cross-checking corpus





V.1.2. Saliency vs. Watermarking

As extracting visual saliency directly from the compressed stream syntax elements is expected to have practical benefits, the thesis aims at studying the impact of integrating the compressed stream saliency maps in a compressed stream application. The particular case of video watermarking is considered and the a priori expectation is validated: saliency acts as an optimization tool, allowing the transparency to be increased (for prescribed quantity of inserted information and robustness) while decreasing the overall computational complexity. However, this result brings to light two additional behaviors which were not forecasted at the beginning of the thesis.

First, a detailed analysis of the transparency results shows that both objective and subjective transparency measures are ameliorated. ***Consequently, we can state that the saliency is not only a human visual system related optimization tool in watermarking but also a signal processing***

optimization tool: it also allows the increase of the energy of a perturbation (i.e. the mark) which corrupts an original signal, under the constraint of a prescribed difference (e.g. PSNR or NCC) between the original and the modified signals.

Secondly, note that from the watermarking point of view, the MPEG-4 AVC method is more effective than the HEVC method. However, we cannot state yet the reason of this difference. While one possible explanation would be related to the very nature of the two types of encoding standards, note that our MPEG-4 AVC watermarking experiments also included a perceptual masking step which was not considered for HEVC (to the best of our knowledge, no masking model in HEVC compressed stream yet exists). So, an alternative explanation would be that ***the coupling of the perceptual masking (a long-term psycho-visual mechanism) and saliency (a short term psycho-visual mechanism) lead to applicative watermarking synergies***. However, a true methodological and experimental study is required in order to support this affirmation.

V.2. Future works

Short-term perspectives – ameliorate the compressed domain saliency maps

The present thesis brought to light that a straightforward relation between the Itti's models and the MPEG-4 AVC and HEVC stream syntax elements exists. The corresponding experimental results demonstrated that saliency extraction in compressed domain is not only fast (linear complexity) but also closer to the ground-truth then the pixel-based models. However, several possible ways of ameliorating the MPEG-4 AVC and HEVC models still exist.

First, note that our intensity, color and motion maps are defined as energies of the stream syntax element values. Although these definitions are related to the Itti's model, future work will be devoted to investigate whether different averaging formulas can be considered instead of energy.

Secondly, we shall investigate the possibility of considering more elaborated fusion techniques among the elementary maps. In this respect, the ones based on Quaternion Fourier Transform (QFT) formula [GUO10] and the principle of self-adaptive saliency map fusion in [YAN14] will be starting points.

Mid-term perspectives – integrate compressed domain saliency maps in challenging applicative field

While the compressed domain saliency extraction already demonstrated their effectiveness in the watermarking applications, work will be devoted to deploy them for other applicative fields like video retargeting [LUO11], object segmentation [KIM14] and discovery [YAN15], video surveillance [KIM14] or decision support systems for virtual collaborative medical environments [GAN15].

Long-term perspectives – define an information theory based model for saliency detection

Although the large majority of the saliency extraction studies are based on the Itti's models, the study in [KHA15] shows a correlation between the size (in bits) of the encoded macroblock representation and its

saliency. Our study goes one step further and identifies, inside the macroblock, which syntax elements are actually connected to saliency.

These observations can be considered as the first two steps towards defining an information-theory based model for saliency. The principle of such a model would be to validate whether the classical information theory entities (and mainly the ones related to source coding) are able to accommodate the saliency computation and deployment or new entities matched to this human visual related field should be defined.

Such a model would also implicitly provide answers to the open points raised in Chapters V.1.1 and V.1.2, namely about the visual saliency as a signal processing optimization tool and the extent to which synergies can be established between perceptual masking and saliency, two complementary human visual peculiarities.

VI. Appendixes

A Fusing formula investigation

A total of 48 fusion formulas (6 for combining static features and, for each of them, 8 to combine static to dynamic features) are investigated in our study, both for MPEG-4 AVC (as reported in Chapter III) and HEVC (as reported in Chapter IV), [AMM15], [AMM16].

Static saliency map fusion formulas

We consider 6 formulas for fusing the elementary static maps: 4 weighted additions, 1 multiplication and 1 maximal, as follows.

The static saliency map can be computed as a linear combination of the intensity, the color, and the orientation normalized maps:

$$M_s = \beta_1 N(M_i) + \beta_2 N(M_C) + \beta_3 N(M_o) \quad \text{(A-1)}$$

Where β_1, β_2, and β_3 are the parameters determining respectively the weight for the intensity map M_i, color map M_c, orientation map M_o, and the normalization formula N (mentioned in Chapter III).

- **Color advantage fusion:** we consider the equation (A-1) and we define the weight of the color saliency map as the highest weight β_1=0.2, β_2=0.6, and β_3=0.2
- **Orientation advantage fusion:** we consider the equation (A-1) however we accord the highest weight to the orientation saliency map β_1=0.2, β_2=0.2, and β_3=0.6.
- **Intensity advantage fusion:** we consider the equation (A-1) and we affect the following weights to the features saliency maps β_1=0.6, β_2=0.2, and β_3=0.2
- **Mean fusion**: this fusion technique consists on considering that all the static features have the same effect on the human vision attention, thus we use equal weights for all of the elementary features saliency maps $\beta_1 = \beta_2 = \beta_3 = 1/3$.
- **Max fusion:** This is a winner takes all strategy where the maximum value between the three features maps is retained for each block:

$$M_S = max(M_i, M_c, M_o) \quad \text{(A-2)}$$

- **Multiplication fusion:** a block by block multiplication is applied. We aim at reinforcing the regions that are salient on all elementary features map and eliminating the regions that have a zero value even in only one feature map:

$$M_S = M_i \times M_c \times M_o \quad \text{(A-3)}$$

Spatio-temporal saliency map fusion formulas

Each and every time a saliency map is computed; elementary feature maps are first individually processed then fused in order to get the final map. This fusion process takes place at two levels: static (inside each frame of the video) and then dynamic, when the static components are combined with the temporal information.

However, the choice of the fusion formulas themselves is an open research topic, as testified by the large variety of choices made in the literature [ITT98], [MUD13], [MAR09], [MAR08], [LU10], and [PEN10]. Moreover, the study in [MUD13] is devoted to this topic: it discusses various ways of fusing the static and dynamic saliency maps for uncompressed video sequences, as briefly presented below. In the sequel the following notations are made: M_F is the fused saliency map, M_D is the dynamic saliency map and M_S is the static saliency map.

- **Mean fusion** [ITT98][MUD13]: this fusion technique takes the average of both static and dynamic saliency map:

$$M_F = (M_s + M_D)/2 \tag{A-4}$$

- **Maximum fusion** [MUD13][MAR09]: this is a *winner takes all* strategy, where the maximum value between the two saliency maps is taken for each location:

$$M_F = max(M_s, M_D) \tag{A-5}$$

- **Multiplication fusion** [MUD13][MAR09]: this requires an element-wise multiplication:

$$M_F = M_s \times M_D \tag{A-6}$$

- **Maximum skewness fusion** [MUD13][MAR09]: the static pathway is modulated by its maximum and the dynamic saliency map is modulated by its skewness value (defined as the third moment on the distribution of the map [MAR08]). The salient areas both in static and dynamic maps are reinforced by the product of the static map's maximum and the motion map's skewness value, as shown in the following formula:

$$M_F = \alpha M_s \times \beta M_D + \gamma(M_s + M_D) \tag{A-7}$$

  where $\alpha = max\ (M_s)$, $\beta = skewness(M_D)$ and $\gamma = \alpha\beta$.

- **Binary threshold fusion** [MUD13][LU10]: first, a binary mask M_B is generated by thresholding the static saliency map (the mean value of M_S is used as threshold). Second this M_B is used to exclude spatiotemporal inconsistent areas and to enhance the robustness of the final saliency map when the global motion parameters are not estimated properly:

$$M_F = max\ (M_s, M_D \cap M_B) \quad \text{(A-8)}$$

- **Motion priority fusion** [MUD13][PEN10]: this fusion technique relates to the cases in which the viewer attention is attracted by the motion of an object even when the static background is (as saliency map value) higher:

$$M_F = (1-\alpha)M_S + \alpha M_D \quad \text{(AVI-9)}$$

with $\alpha = \lambda e^{1-\lambda}$ *and* $\lambda = max\ (M_S)$*-mean* (M_D).

- **Dynamic weight fusion** [MUD13][XIA10]: this fusion is a dynamic fusion scheme dependent on the content of the video. The weights are determined by the ratio between the means of the static and dynamic maps for each frame:

$$M_F = \alpha M_S + (1-\alpha)M_D \quad \text{(A-10)}$$

where α *= mean* (M_D)*/ (mean* (M_S) *+ mean* (M_D)*).*

- **Scale invariant fusion** [MUD13][KIM11]: in this fusion technique, the input images are analyzed at three different scales, (32×32, 128×128 and the original image size). The three maps obtained at these scales are subsequently linearly combined into the final spatio-temporal saliency map:

$$M_F = \sum_{k=1}^{3} w_l M_F^k \quad \text{(A-11)}$$

where $M_F^k = (1-\alpha)M_D + \alpha M_S\ with\ \alpha = 0.5$ is the map at scale k and the coefficients of the linear combination are $w_1 = 0.1, w_2 = 0.3$ and $w_3 = 0.6$.

A.1.MPEG-4 AVC fusing formula validation

We consider the database organized at the IRCCyN Laboratory [WEB05] and we kept the same experimental conditions as presented in Chapter III.

The experimental results are shown in Figures A-2-A-9: for each investigated case, we report the average value of the metrics (average over the video frames) as well as the underlying 95% confidence limits.

Each of these 8 figures corresponds to one of the particular way in which the static and dynamic maps are fused (cf. equation (A-4)-(A-11)): mean fusion in figure A-2, maximum fusion in figure A-3, multiplication fusion in figure A-4, maximum Skewness fusion in figure A-5, binary threshold fusion in figure A-6, motion priority fusion in figure A-7, dynamic weight in figure A-8, and scale invariant fusion in figure A-9.

At their turn, each of these 8 figures is divided into two plots: the left one stands for the KLD while the right one corresponds to the AUC. On the one hand, that KLD is the distance between the distributions of the saliency maps and the density fixation maps corresponding to *I* frames in each GOP of the video; consequently, the lower the KLD value, the more accurate the saliency map. On the other hand, the AUC is computed between the saliency map and the density fixation map (binarized with a threshold of max/2), at the fixation locations. Consequently, the larger the AUC value, the better the saliency prediction. For each of these two metrics, and for each of the 8 static-dynamic fusing formulas, the 6 ways of fusing elementary static maps are represented from left to right: col-adv (color advantage fusion), ori_adv (orientation advantage fusion), the int_adv (intensity advantage fusion), the stat (mean fusion), the stat-max (maximum fusion), and the stat_mult (multiplication fusion). Two state-of-the-art techniques, namely SV1 [SEO09],[WEB11], and SV2 [GOF12],[WEB12], are also included in the experiments and reported on each and every plot here below.

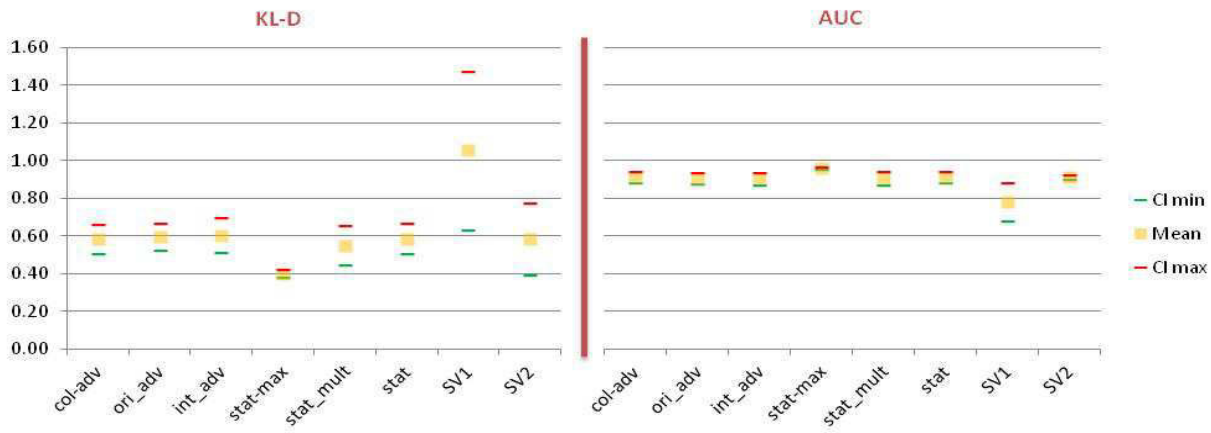


Figure A-2: Mean fusion of the static and dynamic map.

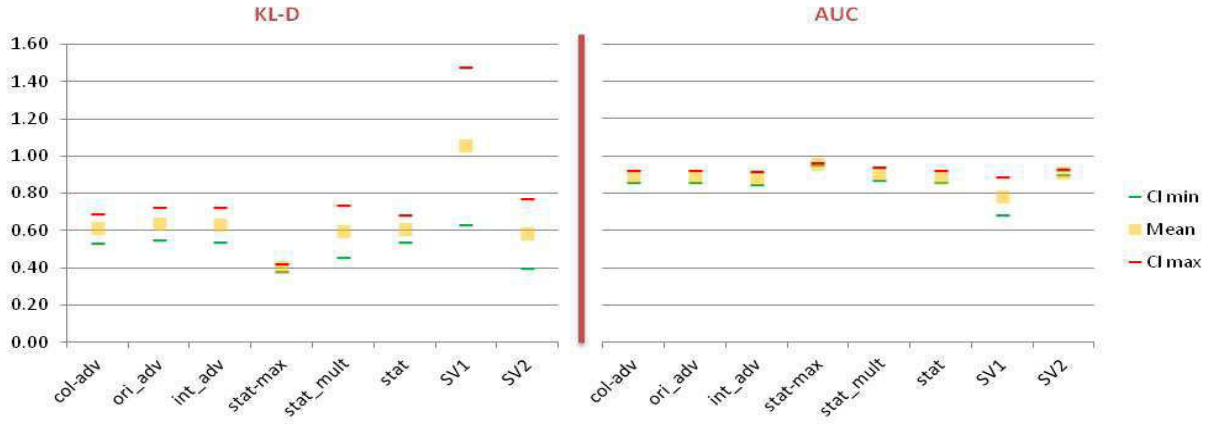


Figure A-3: Maximum fusion of the static and dynamic map.

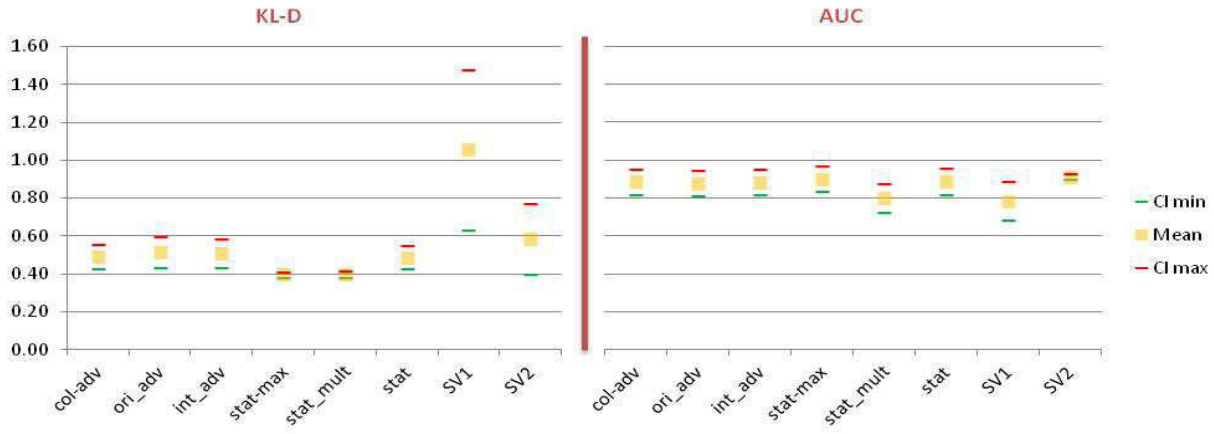


Figure A-4: Multiplication fusion of the static and dynamic map.

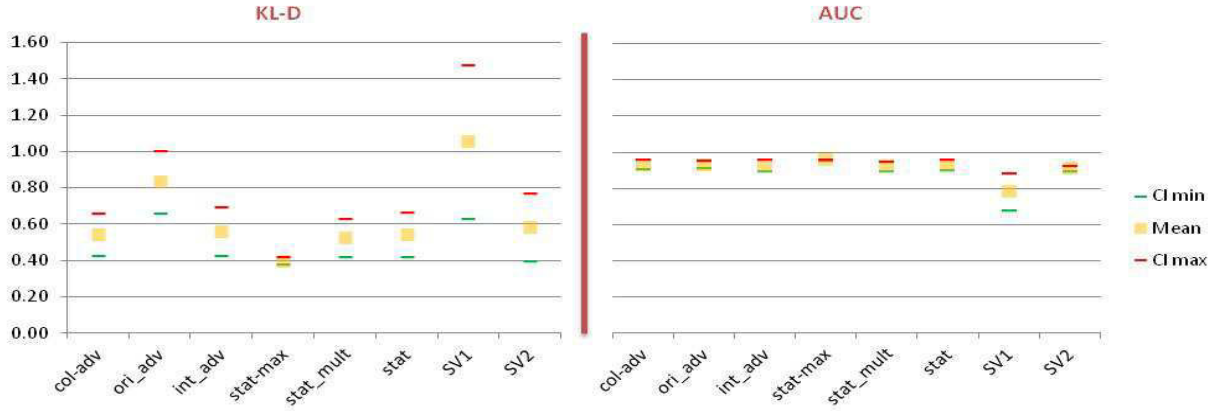


Figure A-5: Maximum Skweness fusion of the static and dynamic map.

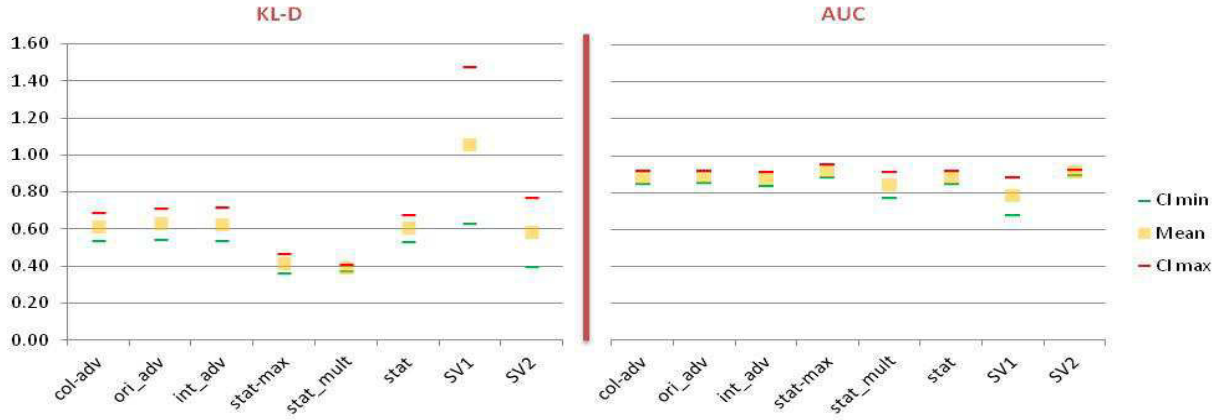


Figure A-6: Binary threshold fusion of the static and dynamic map.

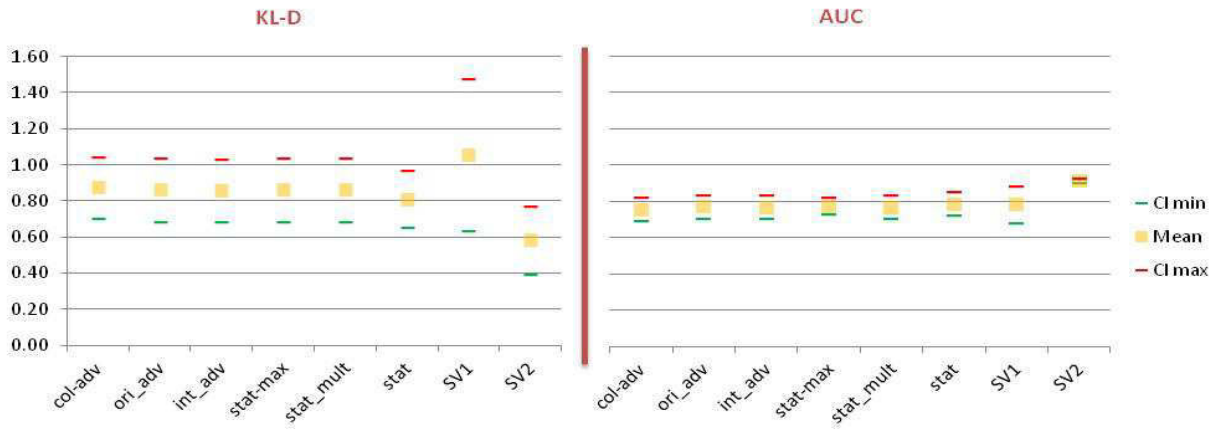


Figure A-7: Motion priority of the static and dynamic map.

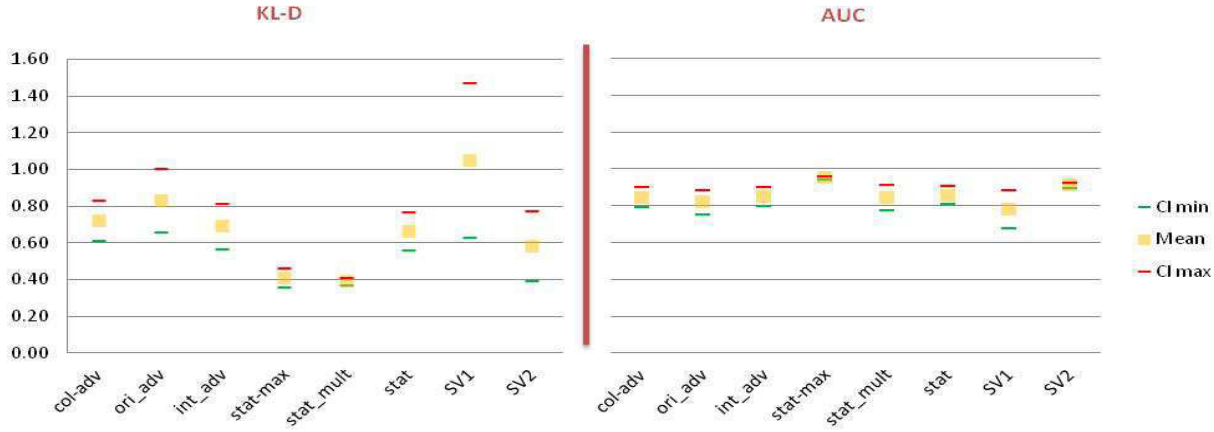


Figure A-8: Dynamic weight fusion of the static and dynamic map.

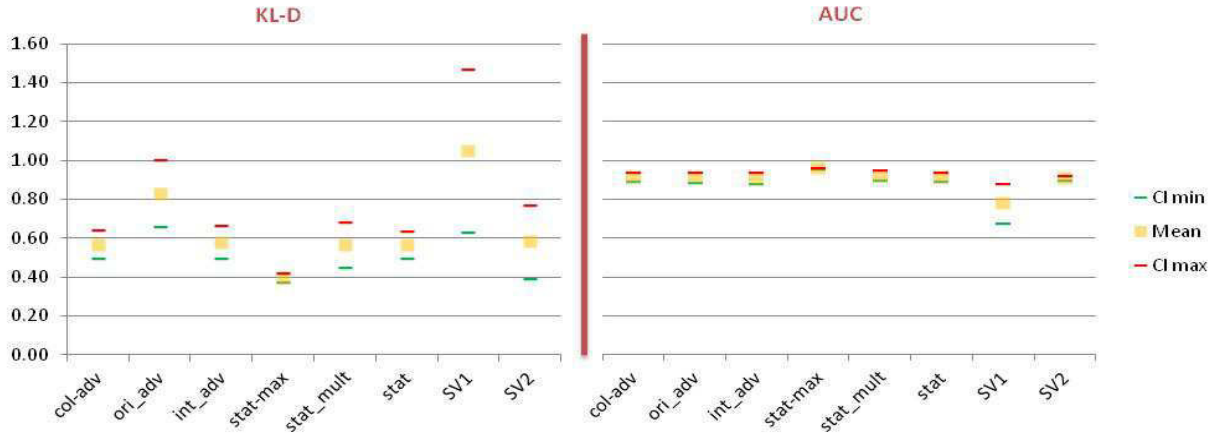


Figure A-9: Scale invariant fusion of the static and dynamic map.

By visually inspecting the values depicted in Figures A-2-A-9, a very large variability of the results with the fusing formula can be noticed. In order to allow a quantitative interpretation of the results, we

define two coefficients (ϱ and η, for KLD and AUC, respectively) expressing the relative differences between a particular investigated fusion method in the compressed domain and the state-of-the-art results:

$$\varrho_{Mij} = \frac{KLD_{Mi} - KLD_{SVj}}{KLD_{SVj}} \tag{A-12}$$

where $\mathrm{KLD_{Mi}}$ represents the KLD value of the map M_i, *i*=1, 2,…48 (the compressed domain saliency maps) and $\mathrm{KLD_{SVj}}$ is the KLD value of the maps SVj, *j* = 1,2 (the state of the art maps, presented in *SV1* and *SV2*).

$$\eta_{Mij} = \frac{AUC_{Mi} - AUC_{SVj}}{AUC_{SVj}} \tag{A-13}$$

where $\mathrm{AUC_{Mi}}$ represent the AUC value of the map M_i, i=1, 2,…48 (the compressed domain saliency maps) and $\mathrm{AUC_{SVj}}$ is the AUC value of the maps SVj, *j* = 1,2 (the state of the art map, presented in *SV1* and *SV2*).

According to these definitions, a gain with respect to the state of the art is reflected by negative ϱ and by positive η. By computing these two coefficients for each and every investigated case, we noticed that the two types of fusion (both static, the static-dynamic) have a significant impact in the results, as for example:

For a same static-dynamic technique (e.g. the mean fusion, Figure A-2), the ϱ coefficient varies between -0.62 and 0.03 while the η coefficient varies between -0.02 and 0.23, according to the static fusion formula;

Conversely, for a same static fusion formula (e.g. maximum), the ϱ coefficient varies between -0.63 and 0.48 while the η coefficient varies between -0.15 and 0.24, according to the static-dynamic fusing formula

As a general conclusion, the most accurate results (in the sense of the two objective measures, the two defined coefficients, and of the processed corpus) are provided by the Skewness static-dynamic fusion over the maximum static fusion: $\varrho_1 = -0.62; \varrho_2 = -0.22; \eta_1 = 0.05;\ \eta_2 = 0.24.$

Note that as this combination results in negative ϱ and by positive η values, we can also conclude that computing the saliency in the MPEG-4 AVC compressed domain according to the map advanced with this study and with the Skewness-maximum fusing techniques gives more accurate results than computing it in the uncompressed domain by the state-of-the-art approaches. Actually, several types of fusion technique combinations result in gains over the two investigated state-of the-art methods, for the two ϱ and η coefficients, namely: binary mask-maximum, dynamic-maximum, Skewness-orientation advantage, Skewness-intensity advantage, Skewness-maximum, Skewness-multiplication, Skewness-mean, invariant-maximum, invariant-multiplication, invariant-mean, maximum-maximum, multiplication-maximum, and mean-maximum.

A.2.HEVC fusing formula validation

All the experimental conditions are kept as described in Chapter IV.

Our experiment consists of comparing the obtained saliency maps according to different fusing formulas by calculating the distance between the saliency map and the density fixation map using two measures: the KLD and the AUC. To binarize the density fixation map, we used the threshold as the half of maximum value of the entire map.

Figures A-10-A-17 represent the result of the comparison of the obtained saliency maps with four methods of the state of the art, namely: Ming Cheng *et al*. [CHE13], Hae Seo *et al.* [SEO09], Stas Goferman [GOF12] and our previous work in MPEG-4 AVC video stream in Chapter III (referred to as AVC). In the case of the AVC method, the best result in each spatio-temporal fusion technique computed is used.

As a general tendency, Figures A-10-A-17 bring to light that saliency extraction from the HEVC stream outperforms (in both KLD and AUC sense) the three investigated uncompressed domain state-of-the-art methods. However, no sharp conclusion can be drawn when comparing the HEVC domain to AVC domain: the performances depend on both the static and spatio-temporal saliency pooling technique.

In order to quantify these behaviors we compute two coefficients ϱ_{Mij} and η_{Mij}, defined in Appendix A.1. According to these coefficients, a gain with respect to the state of the art is reflected by positive ϱ and η values.

The ϱ and η coefficients are reported in Tables 1 and 2, respectively.

Table A-1: KLD gains between HEVC spatio-temporal saliency maps and [CHE13] [SEO09] [GOF12] AVC.

	[CHE13]	[SEO09]	[GOF12]	AVC
Mean (stat_max)	0.41	0.39	0.31	-0.03
Max (stat_max)	0.39	0.37	0.28	-0.07
Multiplication (stat_mean)	0.12	0.08	-0.03	-0.58
Maximum skewness (stat_mean)	0.39	0.36	0.28	-0.07
Binary threshold (stat_max)	0.34	0.31	0.22	-0.19
Motion priority (stat_max)	0.16	0.13	0.01	0.27
Dynamic weight (stat_max)	0.41	0.39	0.31	-0.05
Scale invariant (stat_max)	0.41	0.39	0.31	-0.02

Table A-1 shows that when comparing the HEVC saliency map extracted in the HEVC domain to the three uncompressed-domain methods based on the KLD, with singular exceptions, the ϱ coefficient is larger than 0.1 (its maximal value reaching 0.41). The worst performances are provided by the (*Multiplication*, *static_mean*) pooling combination, when the Gof method outperforms by 3% the HEVC saliency

detection. When compared to the AVC saliency extraction, the pooling technique has a bigger impact in the overall performances:

- the (*Mean*, *stat-max*), (*Dynamic weight*, *stat_max*) and (*Scale invariant*, *stat_max*) combinations result in quite equal good performances, the ϱ being lower than 5%;
- the (*Max, stat_max*), (*Multiplication, stat_mean*), (*Maximum skewness, stat_mean*) and (*Binary threshold, stat_max*) combinations result in better performances for the AVC saliency map extraction;
- the (*Motion priority, stat_max*) combination ensures better performances for the HEVC saliency extraction.

A similar analysis can be performed based on the η coefficient reported in Table A-2. This time, all the figures show that HEVC saliency maps outperform the three state-of-the-art methods. The gains are ranging from 6% to 23%. Moreover, HEVC and AVC saliency extraction feature equally good performances: the absolute value of the η coefficient is always lower than 3%.

Table A-2: AUC gains between HEVC spatio-temporal saliency maps and [CHE13] [SEO09] [GOF12] AVC.

	[CHE13]	[SEO09]	[GOF12]	AVC
Mean (stat_max)	0.23	0.19	0.18	0.00
Max (stat_max)	0.22	0.19	0.18	0.00
Multiplication (stat_mean)	0.10	0.08	0.06	-0.03
Maximum skewness (stat_mean)	0.22	0.19	0.18	0.00
Binary threshold (stat_max)	0.21	0.18	0.17	0.03
Motion priority (stat_max)	0.18	0.15	0.13	-0.02
Dynamic weight (stat_max)	0.23	0.19	0.18	0.01
Scale invariant (stat_max)	0.23	0.19	0.18	0.00

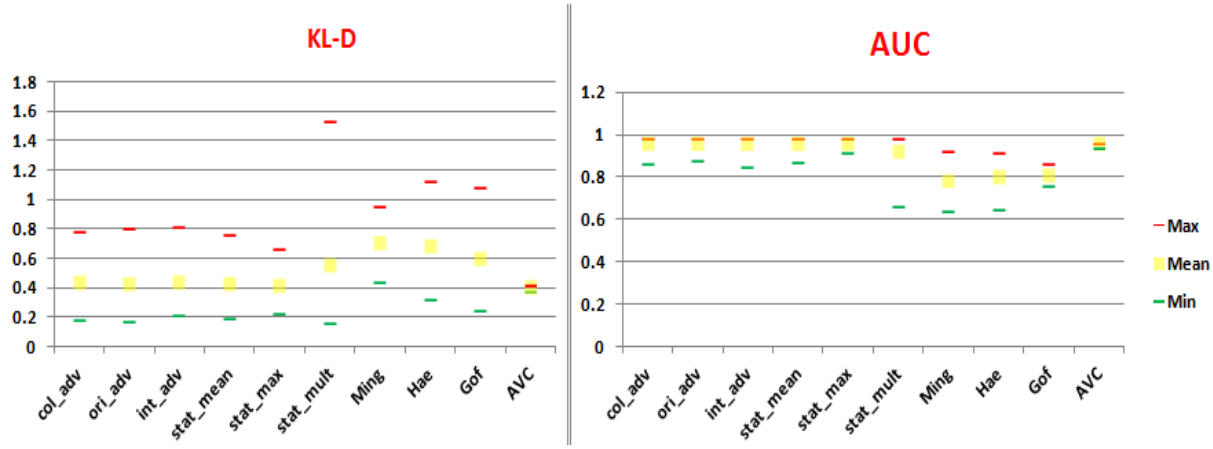


Figure A-10: Mean fusion.

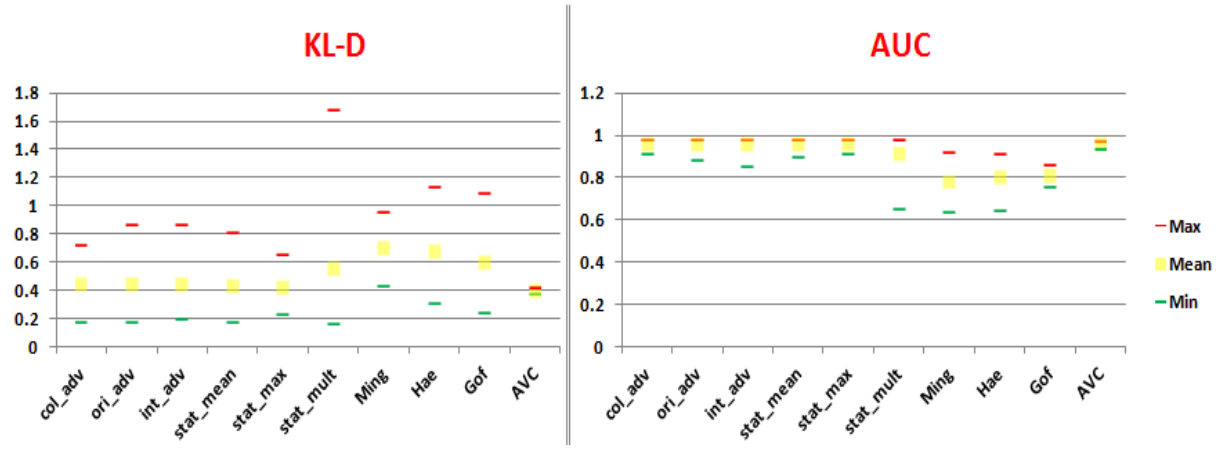


Figure A-11: Maximum fusion.

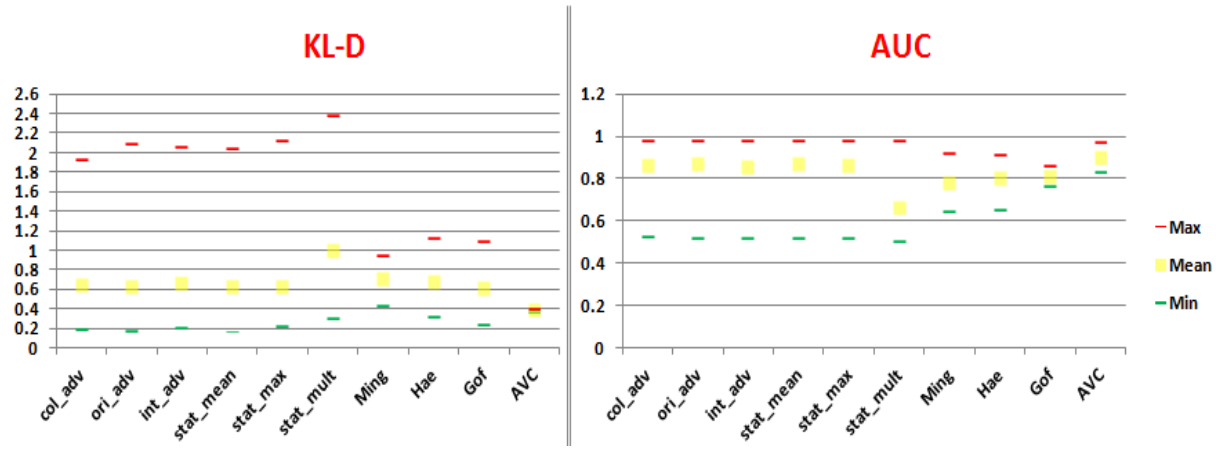


Figure A-12: Multiplication fusion.

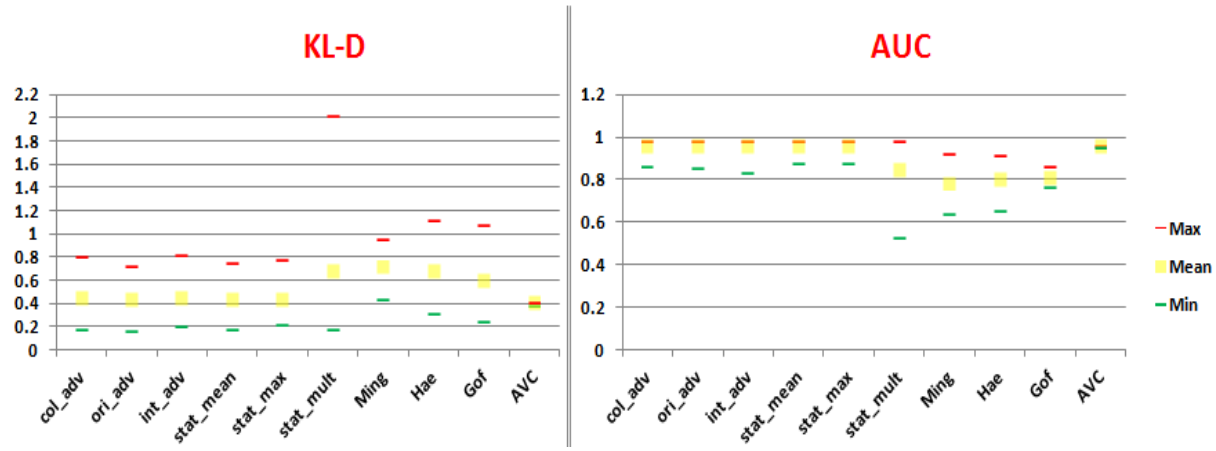


Figure A-13: Maximum Skewness fusion.

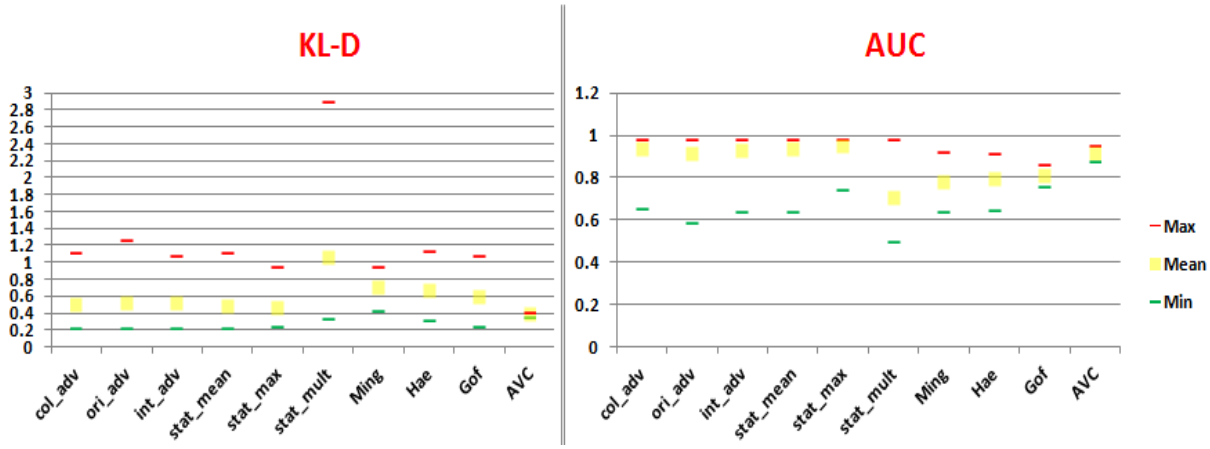


Figure A-14: Binary threshold fusion.

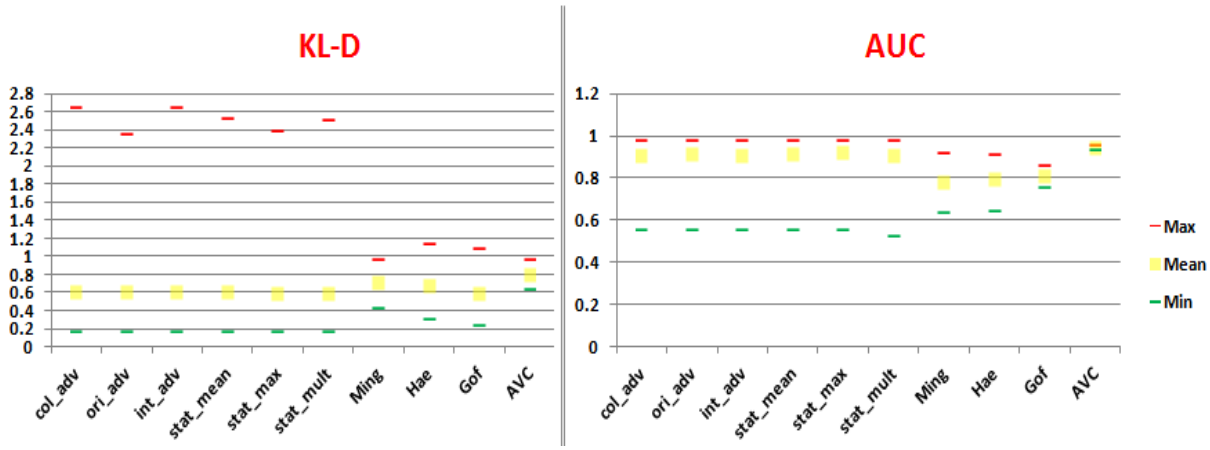


Figure A-15: Motion priority fusion.

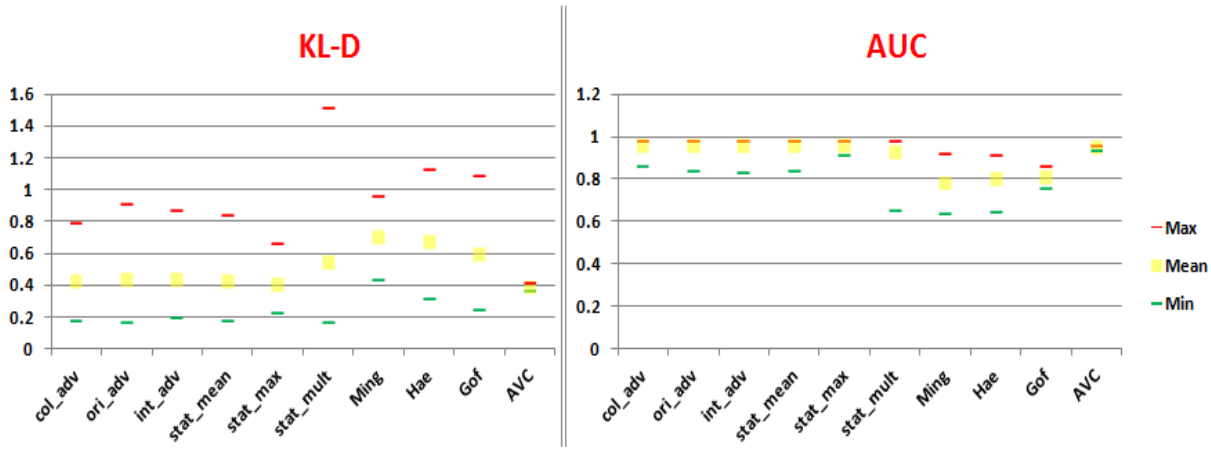


Figure A-16: Dynamic weight fusion.

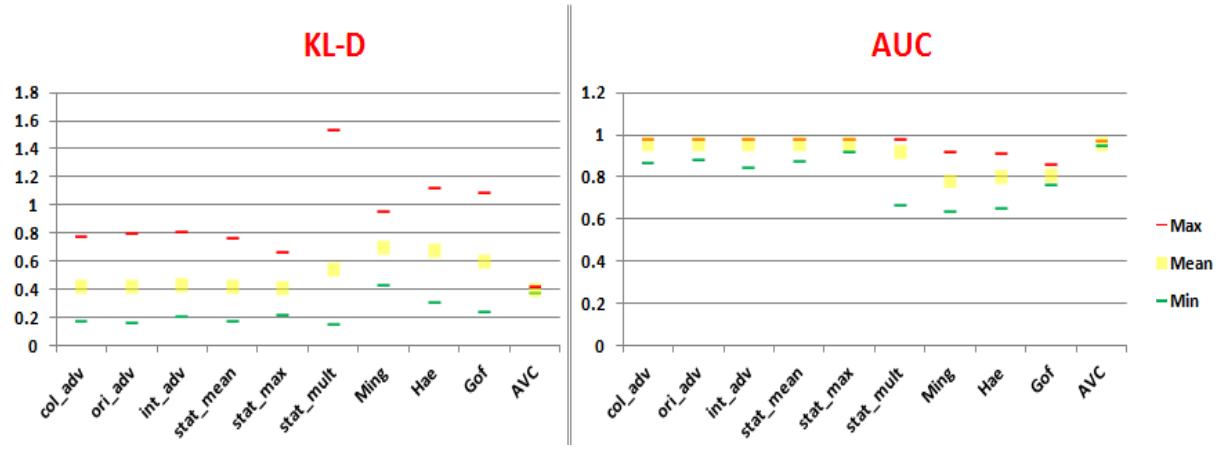


Figure A-17: Scale invariant fusion.

A.3.Conclusion

The present validation considers a detailed investigation on the static and static-dynamic fusing formula. 48 different fusing combinations are investigated and benchmarked against two state-of-the-art methods acting in the uncompressed domain. The experimental results confirm that the choice of the fusing formula is a crucial issue in the design of the saliency map: for a fixed spatio-temporal fusion, static saliency fusion can induce variation of 50 % in KLD and 20% in AUC and for a fixed static fusion, spatio-temporal fusion can induce variation of 15% in KLD and 9% in AUC.

B. MPEG-4 AVC basics

MPEG-4 AVC (Advanced Video Coding Standard) is a video coding standard, developed by the Joint Video Team (JVT), the result of collaboration between the ITU-T Coding Video Expert Group (VEG) and the ISO/IEC Moving Picture expert Group (MPEG). This standard provides substantial better video quality at the same data rates compared to previous standard (MPEG-2, MPEG-4 Part 2, H.263) with only a moderate increase of complexity [RIC03]. Used in a wide range of applications, from mobile phones to High Definition TV, it helped to revolutionize the quality of the video image operating over several types of networks and systems.

While MPEG-4 AVC standard shares common features within other existing standards, it has a number of advantages that distinguish it from previous standards [RIC03].

The following are some of the key advantages of MPEG-4 AVC standard:

- Up to 50% in bit rate saving: compared to MPEG-2 or MPEG-4 Part 2, MPEG-4 AVC allows a reduction in bit rate by up to 50% for a similar degree of encoder optimization at most bit rates.
- High quality video: MPEG-4 AVC offers consistently better video quality at the same bit rate copmpared to previous standards.
- Error resilience: MPEG-4 AVC provides necessary tools to deal with packet loss in packet networks and bit errors in wireless networks.
- Network friendliness: MPEG-4 AVC bit stream can be easily transported over different networks through the Network Adaptation Layer.

The MPEG-4 AVC standard does not defines a new encoder. However, it defines new encoding syntax elements and refines the principal encoding functions.

The purpose of this Appendix is to outline the concept of the MPEG-4 AVC encoding standard and its advantages with respect to previous standards.

B.1. Structure

The MPEG-4 AVC architecture is designed based on two main layers: The Video Coding Layer (VLC) which is constructed to efficiently represent the video contents and the Network Abstraction Layer (NAL) which encapsulates the content represented by the VCL and provides header information in an appropriate way for conveyance by a variety of transport layer or storage media [RIC03].

The VCL is structured into five layers: GOP (Group Of Picture), picture, slice, macroblock and block. Headers of each layer provide information on the encoding/decoding order for the lower layers.

A GOP consists of a number of images that can be 3 types, grouped according to a predetermined decoding order:

- **The *I* frames** correspond to independently coded images ; note that only one field *I* can be at the beginning of a GOP, as it serves as a starting point for coding *P* and *B* frames;

- **The *P* frames** are associated with motion compensated images, predicted either from an *I* or *P* or *B* frame;
- **The *B* frames** refer to any image being double (forward and backward) motion compensated.

Block partitioning

Each video image is partitioned into 16×16 macroblocks. Each macroblock consists of 16×16 luminance samples Y and of 8×8 samples for each of the two chrominance components Cb and Cr. These blocks are encoded/decoded with respect to the order described in the Figure B-1.

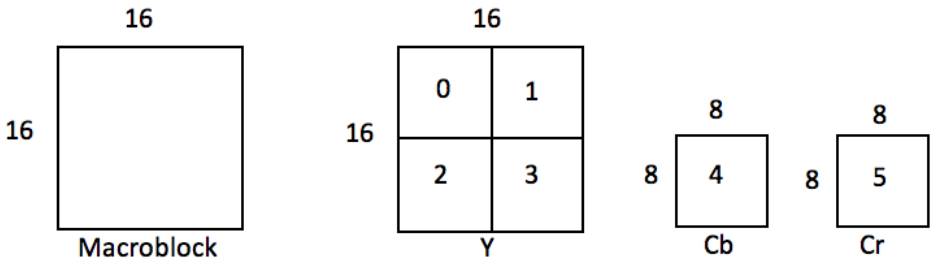


Figure B-1: Y, Cb and Cr encoding/decoding order.

B.2. Encoding

Prediction

The prediction aims at eliminating the spatial (intra prediction) and temporal (inter prediction) redundancy.

Each frame of a video sequence is processed in units of macroblock (corresponding to 16×16 pixels). Each macroblock is encoded in intra or inter mode.

For the inter prediction, the blocks are predicted from previous or following frames, by using the spatial displacement of corresponding blocks of frames specified by a motion vector. Compared to previous video coding standard which supports only 16×16 and 8×8 block sizes for motion estimation, MPEG-4 AVC supports a block sizes ranging from 16×16 to 4×4. Motion compensation for each 16×16 block can be performed according to different block sizes and shapes.

4 inter partitioning modes are initially supported: 16×16, 16×8, 8×16, and 8×8. For 8×8 partitions, an additional syntax element specifies whether it will be further partitioned into 4×8, 8×4 or 4×4 inter-prediction blocks. Figure B-2 illustrates all the partitioning modes.

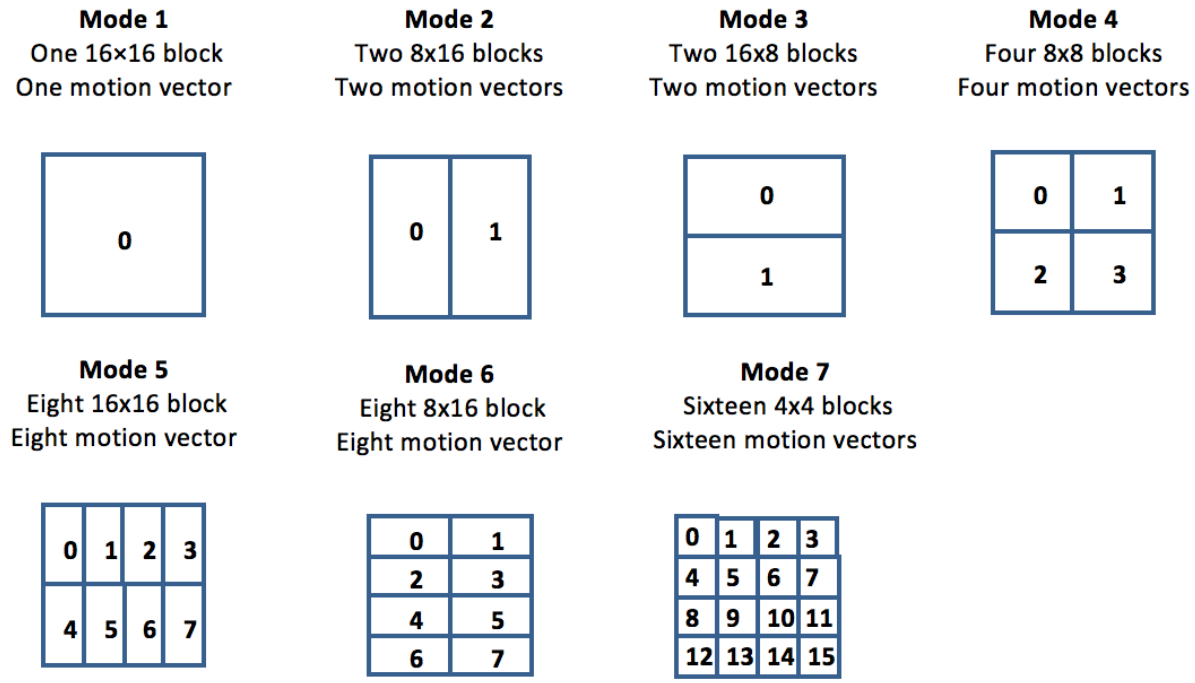


Figure B-2: Different modes of partitioning a macroblock for motion estimation in MPEG-4 AVC.

For the intra-prediction mode, the block B is constructed from samples of neighboring blocks have been previously encoded/decoded. In MPEG-4 AVC, two intra prediction block sizes are supported: 4×4 and 16×16. The 4×4 partitioning mode is well suited for encoding the textured frame area, while the intra 16×16 intra partitioning is more suited for encoding smoothed frame area.

In order to perform the intra prediction, MPEG-4 AVC offers nine modes for the prediction of 4×4 luminance blocks [RIC03], including DC prediction (Mode 2) and eight directional modes, see Figures B-3 and B-4; these figures are taken from [RIC03].

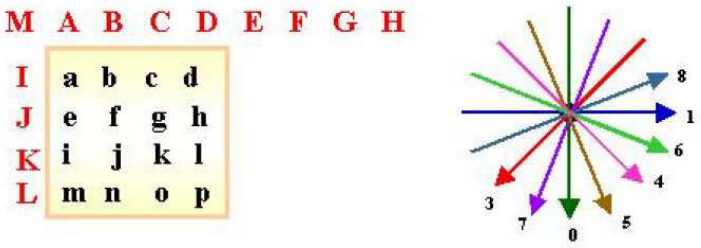


Figure B-3: Intra prediction.

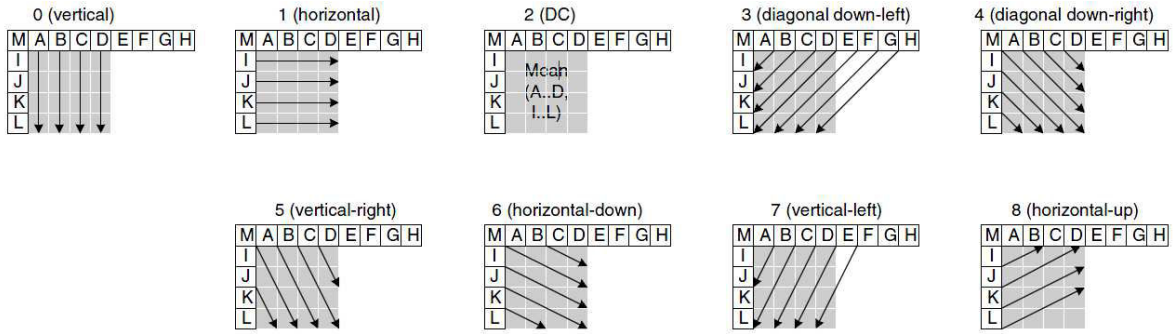


Figure B-4: Intra prediction modes for 4×4 luminance blocks [RIC03].

The predicted block is obtained by using the already encoded samples (from A to M) from neighboring blocks.

Transformation

Following the prediction, the transformation is applied with the aim of representing the data as uncorrelated (separate components with a minimum interdependence) and compacted (the energy is concentrated in a small number of frequencies) [HAL02].

Compared to previous standards which use the 8×8 Discrete Cosine Transform (DCT) as the basic transformation, MPEG-4 AVC uses three transformations depending on the type of the data to be encoded:

- An integer DCT transformation which is applied to all 4×4 blocks of luminance and chrominance components in the residual data.
- A Hadamard transformation applied to 4×4 blocks constructed of luma dc coefficients in intra macroblocks predicted according to the 16×16 mode.
- A Hadamard transformation applied to 2×2 blocks constructed of chroma dc coefficients in any macroblock.

One of the main improvements of this standard is the using of smaller 4×4 block transformation. Instead of a classical 4×4 discrete cosine transform, a separable integer transform with similar properties as a 4×4 DCT is used. The new advanced transform approaching the 4×4 DCT has several advantages:

- The core part of the transformation can be implemented using additions and shifts, resulting to less level of computation complexity.
- The precise integer specification eliminates any mismatch issues between the encoder and decoder in the inverse transform (this has been a problem with earlier standards).

Figure B-5 illustrates the way in which the data is structured and transmitted within a macroblock. If the macroblock is coded in 16×16 intra mode, then the block containing the DC coefficient of each 4×4 luma block is transmitted first. Secondly, the luma residual blocks ranging from 0 to 15 are transmitted in the order shown in Figure B-5 where the DC coefficients are set to zero. Blocks 16 and 17 containing a

2×2 array of chroma coefficients are transformed and sent. Finally, chroma residual blocks ranging from 18 to 25 (with DC coefficient set to 0) are sent.

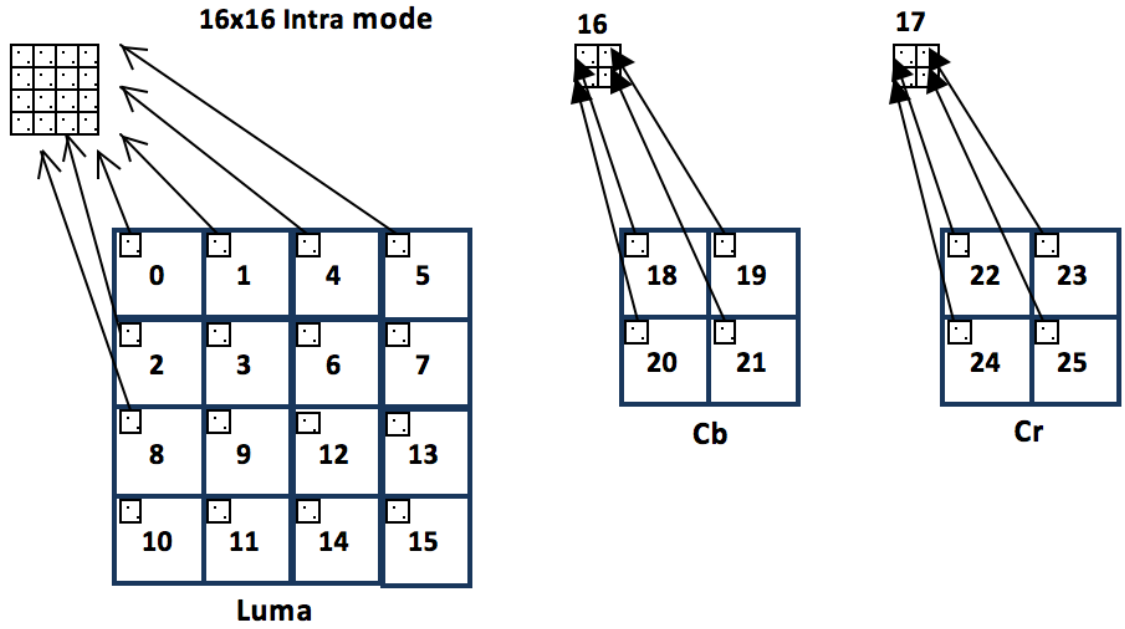


Figure B-5: Block construction for DCT and Hadamard transformations.

Quantization

The quantization phase is where the information is lost in the compression chain [HAL02]. In MPEG-4 AVC, the transformed coefficients are quantized using a scalar quantization. The basic forward quantization operation is performed as follows:

$$Z_{ij} = round(Y_{ij} / Q_{step})$$

where Y_{ij} is a coefficient of the transformed 4×4 block described above, $\mathrm{Q_{step}}$ is the quantization step and Z_{ij} is the quantized coefficient.
The MPEG-4 AVC supports a total of 52 quantization steps which are indexed by a quantization parameter Q_p as illustrated in Table B-1.

Table B-1: Quantization steps.

Q_p	0	1	2	3	4	5	6	7	8	9	10	11	12	…
Q_{step}	0.625	0.6875	0.8125	0.875	1	1.125	1.25	1.375	1.625	1.75	2	2.25	2.5	…
Q_p	…	18	…	24	…	30	…	36	…	42	…	48	…	51
Q_{step}		5		10		20		40		80		160		224

To circumvent the disadvantages of the entire division, the MPEG-4 AVC standard offers another form of quantization performing, this time right shift:

$$Z_{ij} = sign\{Y_{ij}\}\left[|Y_{ij}|A(Q_p) + f2^L \gg L\right]$$

Where f and $A(Q_p)$ are association of the quantization parameter, L is the bit length parameter for the encoding process.

Entropy coding

Entropy coding is the final phase of the MPEG-4 AVC and takes place in three stages:

- the quantized transformed coefficients are scanned in a zig-zag manner (Figure B-6) and transmitted to be encoded
- Each quantized coefficient is RL (Run-Length) encoded so as to increase the compression rate
- The bitstream is constructed according to two advanced methods of the entropy coding. The first category represents a combination of Universal Variable Length Coding (UVLC) and Context Adaptive Variable Length Coding (CAVLC) which can be used for all encoding profiles. The second method is represented by Context-Based Adaptive Binary Arithmetic Coding (CABAC) that can be used alternately with CAVLC only for main profile.

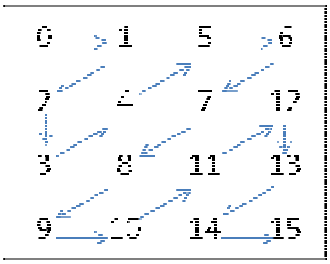


Figure B-6: Zig-zag scanning.

C. HEVC basics

The High Efficiency Video Coding (HEVC) standard is the most recent video coding standard [SUL12] developed by the Joint Collaborative Team on Video Coding (JCT-VC), a group of video coding experts from ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG).

HEVC is used in a wide range of HD videos and supports resolutions up to 8K UHDTV (8192x4320). HEVC retains the similar set of basic coding and encoding process and the high level syntax architecture used in MPEG4-AVC. However, it improved each of them by introducing new more sophisticated techniques.

Compared to the previous standard, HEVC offers larger and more flexible prediction and transform block sizes, greater flexibility in prediction modes (35 Intra prediction modes), more sophisticated signaling of modes and motion vectors and larger interpolation filter for motion compensation.

HEVC ensures a video quality identical to H.264 AVC at only half the bit rate; actually, compression gains of 30 to 60% with an average of 40% are reported, but this ratio highly varies with the content type, resolution and compression settings. The highest gain is obtained with UHD videos.

Same as the other ITU-T and ISO/IEC video coding standards, only the bit stream syntax is standardized.

C.1. Structure

The extension from MPEG-4 AVC to HEVC is not straightforward. On the one hand, HEVC allows different block sizes to be defined. On the other hand, both intra and inter prediction modes are changed.

HEVC video sequences are structured the same way as MPEG4-AVC, into Groups of Pictures (GOP). A GOP is composed of an *I* (intra) frame and a number of successive *P* and *B* frames (unidirectional predicted and bidirectional predicted, respectively). The *I* frame describes a full image coded independently by using intra prediction, containing only references to itself. The unidirectional predicted frames *P* use one or more previously encoded frames (of *I* and *P* types) as reference for picture encoding/decoding. The bidirectional predicted frames B consider in their computation both forward and backward reference frames, be they of *I*, *P* or *B* types.

A frame in HEVC is partitioned into coding tree units (CTUs), which each covers a rectangular area up to 64x64 pixels depending on the encoder configuration. Each CTU is divided into coding units (CUs) that are signaled as intra or inter predicted blocks. A CU is then divided into intra or inter prediction blocks according to its prediction mode. For residual coding, a CU can be recursively partitioned into transform blocks.

HEVC supports two modes of partitioning an intrapicture-predicted block: PART_2Nx2N and PART_NxN. The first mode indicates that the prediction block PB size is the same as the coding block CB size, while the second mode signals the splitting of the CB into four equal-sized PBs. In addition to these two mods, interpicture prediction, HEVC supports 6 types of splitting CB into two PBs.

C.2. Encoding

Prediction

In HEVC, Intra-prediction operates according to transform block sizes and pixel samples are predicted from spatially neighboring TBs by considering an intra-prediction mode. HEVC supports 35 prediction modes for luma intra-prediction: Intra_Planar prediction, Intra_DC prediction and Intra_Angular prediction which defines 33 directional orientations. For chroma intra-prediction, the mode can be signaled as horizontal, vertical, Intra_DC, Intra_Planar or the same as the luma prediction mode. This large set of intra-prediction modes finally results offers in small prediction errors. Three additional post-processing operations referred to as Reference Sample Smoothing, Boundary Value Smoothing and Reference Sample Substitution are applied

The inter-prediction in HEVC can be seen as a steady improvement and generalization of all parts known from previous coding standards. The motion vector prediction was enhanced with advanced motion vector prediction based on motion vector competition. An inter-prediction block merging technique significantly simplified the block-wise motion data signaling by inferring all motion data from already decoded blocks. When it comes to interpolation of fractional reference picture samples, high precision interpolation filter kernels with extended support improve the filtering especially in the high frequency range. The weighted prediction signaling was simplified by either applying explicitly signaled weights for each motion compensated prediction or just averaging two motion compensated predictions.

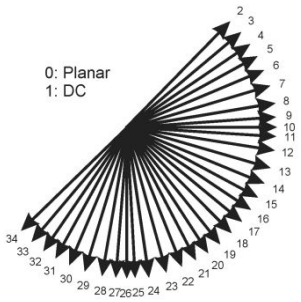


Figure C-1: Modes and directional orientations for intra prediction, cf. [SUL12].

Transformation

As in prior standards, HEVC uses transform blocks to code the prediction residual. The residual block could be is partitioned into multiple square TBs of sizes 4×4, 8×8, 16×16, and 32×32.The core transform matrices applied to residual blocks are Integer Basis functions derived from DCT basis function. Only one integer matrix for the length of 32 points is specified, and sub-sampled versions are used for other sizes. When The size of TB is 4x4, an alternative integer transform derived from a DST is applied to the luma residual blocks.

Quantization

For quantization, HEVC uses essentially the same URQ scheme controlled by a quantization parameter (QP) as in MPEG-4 AVC. The QP values rage from 0 to 51, and an increase by 6 doubles the quantization step size; hence, the mapping of QP values to step sizes is approximately logarithmic. Quantization scaling matrices are also supported.

To reduce the memory needed to store frequency-specific scaling values, only quantization matrices of sizes 4×4 and 8×8 are used. For the larger transformations of 16×16 and 32×32 sizes, an 8×8 scaling matrix is sent and is applied by sharing values within 2×2 and 4×4 coefficient groups in frequency subspaces except for values at DC (zero-frequency) positions, for which distinct values are sent and applied.

Entropy coding

HEVC specifies only one entropy coding method, CABAC (rather than two as in MPEG-4 AVC). The core algorithm of CABAC is unchanged, but its usage in the HEVC design is changed:

- Context Modeling: Appropriate selection of context modeling is known to be a key factor to improve the efficiency of CABAC coding. In HEVC, the splitting depth of the coding tree or transform tree is exploited to derive the context model indices of various syntax elements in addition to the spatially neighboring ones used in MPEG-4 AVC.
- Adaptive Coefficient Scanning: Coefficient scanning is performed in 4×4 sub-blocks for all TB sizes (i.e., using only one coefficient region for the 4×4 TB size, and using multiple 4×4 coefficient regions within larger transform blocks). Three coefficient scanning methods, diagonal up-right, horizontal, and vertical scans as shown in Figure C-2, are selected implicitly for coding the transform coefficients of 4×4 and 8×8 TB sizes in intra predicted regions. The selection of the coefficient scanning order depends on the directionalities of the intra prediction. The vertical scan is used when the prediction direction is close to horizontal and the horizontal scan is used when the prediction direction is close to vertical. For other prediction directions, the diagonal up-right scan is used. For the transform coefficients in inter picture prediction modes of all block sizes and for the transform coefficients of 16×16 or 32×32 intra picture prediction, the 4×4 diagonal up-right scan is exclusively applied to sub-blocks of transform coefficients.

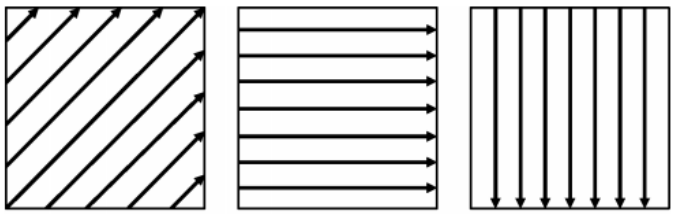


Figure C-2: Three coefficient scanning methods in HEVC: diagonal up-right scan (left), horizontal scan (middle) and vertical scan (right), cf. [SUL12].

C.3 How HEVC is different?

The main objective of HEVC is to provide essential tools to transmit the smallest amount of information required for a given level of visual quality. While HEVC inherits many concepts from MPEG-4 AVC, Table C-1 offers a synoptic view on the main differences between these two standards.

Table C-1: HEVC vs. MPEG-4 AVC

	H264/MPEG-4 AVC	H265/HEVC
Names	MPEG-4 Part 10, AVC	MPEG-H, HEVC, Part2
Approved date	2003	2013
Progression	Successor to MPEG-2	Successor to H.264/AVC
Improvements	-40-50% bit rate reduction compared with MPEG-2 Part - Available to deliver HD sources for Broadcast and Online	-40-50% bit rate reduction compared with H.264 at the same visual quality - It is likely to implement Ultra HD, 2K, 4K for Broadcast and Online
Maximal support	Up to 4k	Up to 8k
Partition sizes	Macroblock 16x16	(Large) Coding Unit 8x8 to 64x64
Partitioning	Sub-block down to 4x4	Prediction Unit Quadtree down to 4x4 square, symmetric and asymmetric (only square for intra)
Intra prediction modes	13 modes with 1/4 pixel accuracy - 9 for textured regions (4x4) - 4 for smoothed regions (16x16)	35 modes with 1/32 pixel accuracy - 33 angular modes - 1 Planar mode - 1 DC mode
Motion prediction	Spatial Median (3 block)	Advanced Motion Neighbor (3 blocks) Vector Prediction (AMVP) (Spatial + temporal)
Motion copy mode	Direct mode	Merge mode
Motion precision	½ Pixel 6-tap ¼ Pixel bi-linear	¼ Pixel for 8 tap 1/8 Pixel 4-tap chroma
Entropy coding	CABAC, CAVLC	CABAC
Filters	Deblocking filter	Deblocking filter Sample Adaptive Offset

D. Tables of the experimental results

In this appendix, we detail the main plots included in Chapter III, IV and V through detailed tables.

D.1 MPEG-4 AVC saliency map validation

Precision

Reference corpus

Table D-1: KLD between saliency map and density fixation map: corresponding to Figure III-6.

	Min	95% CL low	KLD	95% CL up	Max
Skewness-max	0.20	0.22	0.28	0.34	0.35
Combined-avg	0.23	0.29	0.32	0.35	0.38
Multiplication-avg	0.36	0.55	0.64	0.73	0.75
Addition-avg	0.22	0.29	0.31	0.35	0.37
Static-avg	0.27	0.32	0.37	0.42	0.47
Motion	0.35	0.41	0.48	0.55	0.64
CHE13	0.44	0.61	0.71	0.81	0.96
SEO09	0.32	0.48	0.68	0.88	1.13
GOF12	0.25	0.41	0.60	0.79	1.09

Table D-2: AUC between saliency map and density fixation map: corresponding to Figure III-7.

	Min	95% CL low	AUC	95% CL up	Max
Skewness-max	0.92	0.93	0.95	0.97	0.97
Combined-avg	0.8	0.81	0.83	0.84	0.86
Multiplication-avg	0.53	0.57	0.61	0.65	0.71
Addition-avg	0.8	0.81	0.85	0.89	0.9
Static-avg	0.75	0.73	0.81	0.89	0.91
Motion	0.75	0.78	0.82	0.86	0.9
CHE13	0.64	0.72	0.78	0.84	0.92
SEO09	0.65	0.72	0.8	0.88	0.91
GOF12	0.76	0.79	0.81	0.83	0.86

Discriminance

Reference corpus

Table D-3: KLD between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence: corresponding to Figure III-9.

	Min	95% CL low	KLD	95% CL up	Max
Skewness-max	0.30	0.56	0.51	0.46	1.10
Combined-avg	0.28	0.52	0.59	0.65	0.98
Multiplication-avg	0.31	1.24	1.63	2.03	3.27
Addition-avg	0.18	0.40	0.50	0.60	0.92
Static-avg	0.18	0.41	0.55	0.70	1.03
Motion	0.32	0.74	1.06	1.37	2.62
CHE13	0.28	1.23	1.55	1.87	3.36
SEO09	0.35	0.92	1.23	1.53	3.53
GOF12	0.20	0.38	0.43	0.49	0.87

Table D-4: AUC between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence: corresponding to Figure III-10.

	Min	95% CL low	AUC	95% CL up	Max
Skewness-max	0.86	0.89	0.93	0.93	0.93
Combined-avg	0.86	0.88	0.92	0.92	0.91
Multiplication-avg	0.52	0.57	0.66	0.71	0.78
Addition-avg	0.8	0.82	0.87	0.88	0.92
Static-avg	0.81	0.83	0.89	0.92	0.92
Motion	0.76	0.78	0.81	0.84	0.9
CHE13	0.54	0.62	0.73	0.84	0.93
SEO09	0.59	0.68	0.78	0.88	0.93
GOF12	0.88	0.90	0.93	0.92	0.93

Cross-checking corpus:

Table D-5: KLD between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence): corresponding to Figure III-11.

	Min	95% CL low	KLD	95% CL up	Max
Skewness-max	0.29	0.58	0.61	0.64	2.14
Combined-avg	0.38	0.64	0.66	0.68	1.70
Multiplication-avg	0.18	1.39	1.40	1.42	1.80
Addition-avg	0.30	0.67	0.69	0.71	1.90
Static-avg	0.33	0.89	0.91	0.93	1.82
Motion	0.27	0.72	0.74	0.76	1.60
CHE13	0.23	0.52	0.55	0.58	1.90
SEO09	0.15	0.71	0.73	0.75	2.03
GOF12	0.36	0.50	0.53	0.56	2.80

Table D-6: AUC between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence): corresponding to Figure III-12.

	Min	95% CL low	AUC	95% CL up	Max
Skewness-max	0.63	0.74	0.75	0.77	0.99
Combined-avg	0.51	0.57	0.58	0.59	0.98
Multiplication-avg	0.44	0.56	0.57	0.58	0.94
Addition-avg	0.49	0.67	0.68	0.69	0.99
Static-avg	0.58	0.62	0.63	0.64	0.95
Motion	0.48	0.67	0.68	0.69	0.84
CHE13	0.60	0.71	0.72	0.73	0.97
SEO09	0.56	0.62	0.64	0.66	0.96
GOF12	0.52	0.63	0.64	0.66	0.98

D.2 HEVC saliency map validation

Precision

Reference corpus

Table D-7: KLD between saliency map and density fixation map: corresponding to Figure IV-2.

	Min	95% CL low	KLD	95% CL up	Max
Motion priority-max	0.29	0.53	0.61	0.68	1.13
Combined-avg	0.25	0.40	0.44	0.47	0.67
Multiplication-avg	0.30	0.53	0.60	0.66	1.04
Addition-avg	0.26	0.39	0.42	0.46	0.62
Static-avg	0.30	0.43	0.46	0.49	0.69
Motion	0.28	0.51	0.58	0.64	1.03
CHE13	0.44	0.61	0.71	0.81	0.96
SEO09	0.32	0.48	0.68	0.88	1.13
GOF12	0.25	0.41	0.60	0.79	1.09
MPEG-4 AVC	0.20	0.22	0.28	0.34	0.35

Table D-8: AUC between saliency map and density fixation map: corresponding to Figure IV-3.

	Min	95% CL low	AUC	95% CL up	Max
Motion priority-max	0.80	0.90	0.91	0.93	0.97
Combined-avg	0.91	0.95	0.96	0.96	0.97
Multiplication-avg	0.64	0.84	0.86	0.89	0.96
Addition-avg	0.92	0.96	0.96	0.96	0.97
Static-avg	0.89	0.95	0.95	0.96	0.97
Motion	0.72	0.88	0.90	0.92	0.97
CHE13	0.64	0.72	0.78	0.84	0.92
SEO09	0.65	0.72	0.80	0.88	0.91
GOF12	0.76	0.79	0.81	0.83	0.86
MPEG-4 AVC	0.92	0.93	0.95	0.97	0.97

Discriminance

Reference corpus:

Table D-9: KLD between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence): corresponding to Figure IV-4.

	Min	95% CL low	KLD	95% CL up	Max
Motion priority-max	0.26	0.36	0.38	0.40	0.44
Combined-avg	0.39	0.42	0.45	0.48	0.49
Multiplication-avg	0.60	0.69	0.73	0.76	0.76
Addition-avg	0.68	0.82	0.84	0.86	0.99
Static-avg	0.52	0.68	0.72	0.77	1.16
Motion	0.47	0.55	0.58	0.61	0.64
CHE13	0.28	1.23	1.55	1.87	3.36
SEO09	0.35	0.92	1.23	1.53	3.53
GOF12	0.20	0.38	0.43	0.49	0.87
MPEG-4 AVC	0.31	1.24	1.63	2.03	3.27

Table D-10: AUC between saliency map at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence): corresponding to Figure IV-5.

	Min	95% CL low	AUC	95% CL up	Max
Motion priority-max	0.83	0.88	0.91	0.92	0.93
Combined-avg	0.83	0.85	0.89	0.91	0.91
Multiplication-avg	0.69	0.71	0.76	0.78	0.92
Addition-avg	0.78	0.86	0.88	0.89	0.91
Static-avg	0.71	0.78	0.82	0.86	0.89
Motion	0.73	0.78	0.84	0.90	0.90
CHE13	0.54	0.62	0.73	0.84	0.93
SEO09	0.59	0.68	0.78	0.88	0.93
GOF12	0.88	0.90	0.92	0.93	0.93
MPEG-4 AVC	0.86	0.89	0.92	0.93	0.93

Cross-checking corpus:

Table D-11: KLD between saliency maps at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence): corresponding to Figure IV-6.

	Min	95% CL low	KLD	95% CL up	Max
Motion priority-max	0.46	0.56	0.62	0.68	1.61
Combined-avg	0.45	0.59	0.58	0.65	1.59
Multiplication-avg	0.33	0.58	0.66	0.74	1.65
Addition-avg	0.41	0.62	0.58	0.62	0.87
Static-avg	0.37	0.58	0.68	0.77	1.84
Motion	0.40	0.60	0.66	0.72	1.20
CHE13	0.23	0.52	0.55	0.58	1.90
SEO09	0.15	0.71	0.73	0.75	2.03
GOF12	0.36	0.50	0.53	0.56	2.80
MPEG-4 AVC	0.18	1.39	1.40	1.42	1.80

Table D-12: AUC between saliency maps at fixation locations and saliency map at random locations (N=100 trials for each frame in the video sequence): corresponding to Figure IV-7.

	Min	95% CL low	AUC	95% CL up	Max
Motion priority-max	0.46	0.71	0.74	0.77	0.96
Combined-avg	0.50	0.58	0.61	0.64	0.91
Multiplication-avg	0.30	0.55	0.58	0.62	0.92
Addition-avg	0.16	0.61	0.65	0.69	0.89
Static-avg	0.47	0.63	0.66	0.69	0.85
Motion	0.44	0.55	0.58	0.62	0.84
CHE13	0.60	0.71	0.72	0.73	0.97
SEO09	0.56	0.62	0.64	0.66	0.96
GOF12	0.52	0.63	0.64	0.66	0.98
MPEG-4 AVC	0.63	0.74	0.75	0.77	0.99

D.3 Conclusion

Precision

Reference corpus

Table D-13: Comparison of the results of KLD between saliency maps and fixation maps: corresponding to Figure in first column in Table V-1.

	Min	95% CL low	KLD	95% CL up	Max
CHE13	0.44	0.61	0.71	0.81	0.96
SEO09	0.32	0.48	0.68	0.88	1.13
GOF12	0.25	0.41	0.60	0.79	1.09
MPEG-4 AVC	0.20	0.22	0.28	0.34	0.35
HEVC	0.25	0.40	0.44	0.47	0.67
FAN14	0.20	0.37	0.41	0.44	0.94

Table D-14: Comparison of the results of AUC between saliency maps and fixation: corresponding to Figure in second column in Table V-1.

	Min	95% CL low	AUC	95% CL up	Max
CHE13	0.64	0.72	0.78	0.84	0.92
SEO09	0.65	0.72	0.80	0.88	0.91
GOF12	0.76	0.79	0.81	0.83	0.86
MPEG-4 AVC	0.92	0.93	0.95	0.97	0.97
HEVC	0.92	0.95	0.96	0.97	0.97
FAN14	0.60	0.89	0.91	0.92	0.98

Discriminance

Reference corpus

Table D-15: Comparison of the results of KLD between saliency maps at fixation locations and saliency maps at random locations (N=100 trials for each frame in the video sequence): corresponding to Figure in first column and first line in Table V-2.

	Min	95% CL low	KLD	95% CL up	Max
CHE13	0.28	1.23	1.55	1.87	3.36
SEO09	0.35	0.92	1.23	1.53	3.53
GOF12	0.20	0.38	0.43	0.49	0.87
MPEG-4 AVC	0.31	1.24	1.63	2.03	3.27
HEVC	0.68	0.82	0.84	0.86	0.99
FAN14	0.04	0.11	0.14	0.17	0.37

Table D-16: Comparison of the results of AUC between saliency maps at fixation locations and saliency maps at random locations (N=100 trials for each frame in the video sequence): corresponding to Figure in second column and first line in Table V-2.

	Min	95% CL low	AUC	95% CL up	Max
CHE13	0.54	0.62	0.73	0.84	0.93
SEO09	0.59	0.68	0.78	0.88	0.93
GOF12	0.88	0.90	0.92	0.93	0.93
MPEG-4 AVC	0.86	0.91	0.93	0.94	0.94
HEVC	0.83	0.88	0.91	0.92	0.93
FAN14	0.63	0.83	0.85	0.87	0.97

Cross-checking corpus

Table D-17: Comparison of the results of KLD between saliency maps at fixation locations and saliency maps at random locations (N=100 trials for each frame in the video sequence): corresponding to Figure in first column and second line in Table V-2.

	Min	95% CL low	KLD	95% CL up	Max
CHE13	0.23	0.52	0.55	0.58	1.90
SEO09	0.15	0.71	0.73	0.75	2.03
GOF12	0.36	0.50	0.53	0.56	2.80
MPEG-4 AVC	0.18	1.39	1.40	1.42	1.80
HEVC	0.33	0.58	0.66	0.74	1.65
FAN14	0.16	0.91	0.98	1.05	1.70

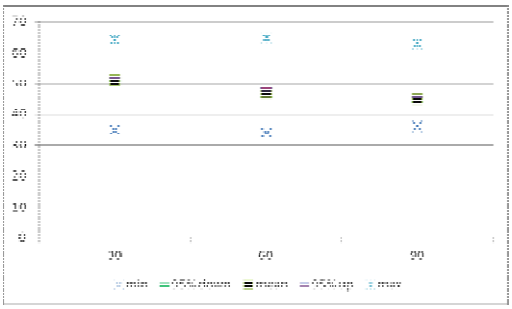
Table D-18: Comparison of the results of AUC between saliency maps at fixation locations and saliency maps at random locations (N=100 trials for each frame in the video sequence): corresponding to Figure in second column and second line in Table V-2

	Min	95% CL low	AUC	95% CL up	Max
CHE13	0.60	0.71	0.72	0.73	0.97
SEO09	0.56	0.62	0.64	0.66	0.96
GOF12	0.52	0.63	0.64	0.66	0.98
MPEG-4 AVC	0.63	0.74	0.75	0.77	0.99
HEVC	0.46	0.71	0.74	0.77	0.96
FAN14	0.61	0.72	0.74	0.76	0.95

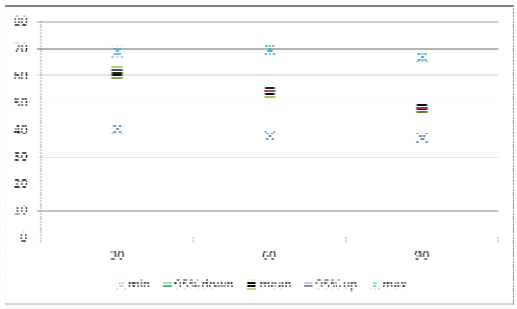
E. Graphics of the experimental results

In this appendix, we represent as plots (graphics) the main applicative results of the objective quality evaluation when alternatively considering random selection and saliency map based selection in a watermarking application. Note that these results are already presented as tables in Chapter III., included in Chapter III through some plots.

Figure E-1: PSNR results of the objective quality evaluation when alternatively considering random selection and saliency map based selection corresponding to PSNR results in Table III-9.

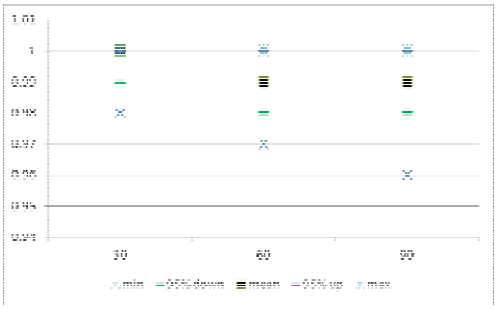


Random selection

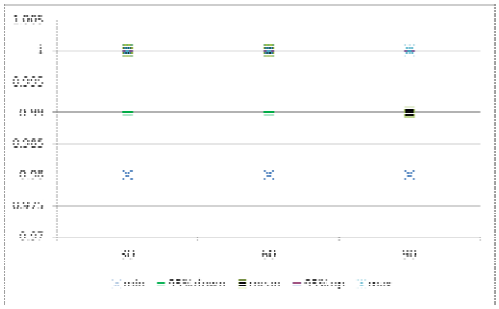


Saliency based selection

Figure E-2: NCC results of the objective quality evaluation when alternatively considering random selection and saliency map based selection corresponding to NCC results in Table III-9.

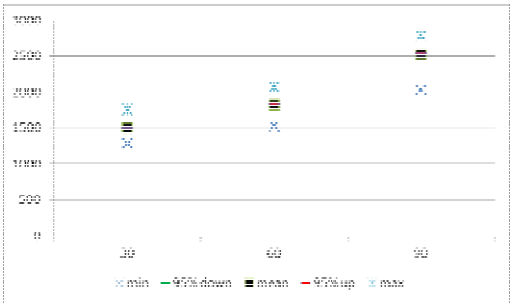


Random selection

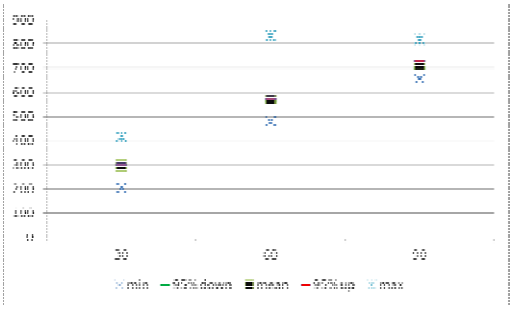


Saliency based selection

Figure E-3: DVQ results of the objective quality evaluation when alternatively considering random selection and saliency map based selection corresponding to DVQ results in Table III-9.



Random selection



Saliency based selection

References

[AHU92] Ahumada A. J., and Peterson H. A., “Luminance-model-based DCT quantization for color image compression”. In SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology. International Society for Optics and Photonics.p. 365-374 (1992).

[ACH08] Achanta R., Estrada F., Wils P., and Suüsstrunk S., ”Salient region detection and segmentation”. Computer Vision Systems, pages 66–75, (2008).

[ACH09] Achanta R., Hemami S., Estrada F., and Su¨sstrunk S., “Frequency-tuned salient region detection”. In IEEE CVPR, pages 1597–1604, (2009).

[ACH10] Achanta R. and Süsstrunk S., “Saliency detection using maximum symmetric surround”. In IEEE ICIP, (2010).

[AGA13] Agarwal C., Bose A., Maiti S., Islam N., and Sarkar S. K., “Enhanced data hiding method using DWT based on Saliency model”. In Signal Processing, Computing and Control (ISPCC), IEEE International Conference on pp. 1-6. (2013).

[AMM14] Ammar M., Mitrea M., Hasnaoui M. “MPEG-4 AVC saliency map computation”. IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, 90141A-90141A (2014).

[AMM15] Ammar M., Mitrea M., Hasnaoui M. and Callet P. L.,. “Visual saliency in MPEG-4 AVC video stream”. IS&T/SPIE Electronic Imaging International Society for Optics and Photonics, pp. 93940X–93940X. (2015).

[AMM16] Ammar M., Mitrea M., Boujelben I., and Callet P. L., “HEVC saliency map computation”. Electronic Imaging, HVEI-107 - 1-8 (2016).

[AMO12] Amon P., Sapre M., and Hutter A., “Compressed domain stitching of hevc streams for video conferencing applications”. In 19th International Packet Video Workshop (PV) pp. 36-40. IEEE. (2012).

[BEL10] Belhaj M., Mitrea M., Duta S., and Prêteux F., “MPEG-4 AVC robust video watermarking based on QIM and perceptual masking”. IEEE International Conference on Communications, pp. 477–480, Bucharest, (2010).

[BHO16] Bhowmik D., Oakes M., and Abhayaratne C., “Visual attention-based image watermarking”. IEEE Access, 4, 8002-8018. (2016).

[BLA03] Blask D. E., Dauchy R. T., Sauer L. A., Krause J. A., and Brainard G. C. “Growth and fatty acid metabolism of human breast cancer (MCF-7) xenografts in nude rats: Impact of constant light-induced nocturnal melatonin suppression”. Breast Cancer Research and Treatment 79, pp 313, (2003)

[BOR13] Borji A. and Itti L. “State-of-the-art in visual attention modeling”. IEEE transactions on pattern analysis and machine intelligence, 35(1), pp 185-207. (2013).

[BOU12] Boujut, H., Benois-Pineau, J., and Megret, R. « Fusion of multiple visual cues for visual saliency extraction from wearable camera settings with strong motion”. In European Conference on Computer Vision (pp. 436-445). Springer Berlin Heidelberg. (2012).

[BRU05] Bruce N. and Tsotsos J., “Saliency Based on Information Maximization Advances”. in Neural Information Processing Systems 18, pp. 155–162. (2005).

[BRU09] Bruce N. D., and Tsotsos J. K.,” Saliency, attention and visual search: An information theoretic approach”. Journal of vision, 9(3), 5-5. (2009).

[BUS15] Buso, V., Benois-Pineau, J., & Domenger, J. P. (2015). Geometrical cues in visual saliency models for active object recognition in egocentric videos. *Multimedia Tools and Applications*, *74*(22), 10077-10095.

[CAB11] Cabrita A. S., Pereira F. and Naccari M. "Perceptually driven coefficients pruning and quantization for the H. 264/A VC standard". In EUROCON-International Conference on Computer as a Tool (EUROCON) IEEE, pp. 1-4. (2011).

[CAO15] Cao, L., and Jung, C., "Combining Visual Saliency and Pattern Masking for Image Steganography". In Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), International Conference on pp. 320-323. IEEE. (2015).

[CHE11] Cheng M.-M, Zhang G.-X, Mitra N. J., Huang X., and Hu. S. M., "Global contrast based salient region detection". In IEEE CVPR, pages 409–416, 2011

[CHE13] Cheng M. M., Warrell J., Lin W. Y., Zheng S., Vineet V., and Crook N., "Efficient salient region detection with soft image abstraction". In Proceedings of the IEEE International Conference on Computer Vision, pp. 1529–1536.(2013)

[CHE15] Chen D., Xia S., and Lu K. "A JND-based saliency map fusion method for digital video watermarking". In Control Conference (CCC), 34th Chinese pp. 4568-4573. IEEE. (2015)

[CHE98] Chen B., and Wornell G. W., "Digital watermarking and information embedding using dither modulation". In Multimedia Signal Processing, IEEE Second Workshop on pp. 273-278. IEEE. (1998).

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

[COX97] Cox I. J., Kilian J., Leighton F. T., and Shamoon T., "Secure spread spectrum watermarking for multimedia". IEEE transactions on image processing, 6(12), 1673-1687. (1997).

[DUA11] Duan L., Wu C., Miao J., Qing L.,and Fu Y., "Visual saliency detection by spatially weighted dissimilarity". In IEEE CVPR, pages 473–480, (2011).

[EGG03] Eggers J. J., Bauml R., Tzschoppe R., and Girod B., "Scalar costa scheme for information embedding". IEEE Transactions on signal processing, 51(4), 1003-1019. (2003).

[FAN12] Fang Y., Chen Z., Lin W., and Lin C. W., "Saliency detection in the compressed domain for adaptive image retargeting". IEEE Trans.Image Processing. vol. 21, no. 9, pp. 3888–3901. (2012)

[FAN14] Fang Y., Lin W., Chen Z., Tsai C.M., and Lin C.W., "A Video Saliency Detection Model in Compressed Domain". IEEE Trans.Circuits and Systems for Video Technology, vol. 24, no. 1, pp. 27–38. (2014).

[FRY65] Fry T.C, "Probability and Its Engineering Use". D van Nostrand, Princeton (1965).

[GAN15] R-R Ganji, M. Mitrea, B. Joveski, and A. Chammem (2015, Feb.). Cross-standard user description in mobile, medical oriented virtual collaborative environments. Proc. SPIE Vol. 9411.

[GAO08] Gao D., Mahadevan V., and Vasconcelos N., "On the plausibility of the discriminant center-surround hypothesis for visual saliency". Journal of Vision, 8(7:13):1–18,(2008).

[GAW16] Gawish A., Scharfenberger C., Bi H., Wong A., Fieguth P., and Clausi D. "Robust Non-saliency Guided Watermarking". In Computer and Robot Vision (CRV), 13th Conference on IEEE, pp. 32-36 (2016).

[GOF10] Goferman S., and Zelnik -Manor L., and Tal A., "Context-aware saliency detection". In CVPR (Vol. 1, No. 2, p. 3). (2010)

[GOF12] Goferman S., and Zelnik -Manor L., and Tal A., “Context-aware saliency detection”. IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 10, pp. 1915–1926,(2012).

[GUO08] Guo C., Ma Q., and Zhang L., “Spatio-Temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform”. Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1-8, (2008).

[GUO10] Guo C. and Zhang L., “A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression”. IEEE Trans. Image Processing, vol. 19, no. 1, pp. 185–198, (2010).

[HAR06] Harel J., Koch C., and Perona P., “Graph-based visual saliency”. Adv. Neural Inf. Process. Syst., pp. 545–552, (2006).

[HAS10] Hasnaoui M., Mitrea M., Belhaj M., and Preteux F., “Visual Quality assessment for motion vector watermarking in the MPEG-4 AVC domain”. Fifth International Workshop on Video Processing and Quality Metrics , Scottsdale, U.S.A, (2010).

[HAS11] Hasnaoui M., Mitrea M., Belhaj M., and Preteux F., “MPEG-4 AVC stream watermarking by m-QIM techniques”. Multimedia on Mobile Devices; and Multimedia Content Access: Algorithms and Systems, United States. 78810L, pp.78810L. (2011).

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

[HE06] He H., Zhang J., and Tai H. M., “A wavelet-based fragile watermarking scheme for secure image authentication”. In International Workshop on Digital Watermarking. Springer Berlin Heidelberg. pp. 422-432.(2006).

[HOU07] Hou X. and Zhang L., “Saliency detection: A spectral residual approach”. Proceedings IEEE. Computer Society Conference on Computer Vision and Pattern Recognition. (2007).

[HOU08] Hou X. and Zhang L., “Dynamic visual attention: Searching for coding length increments”. Adv. Neural Inf. Process. Syst., vol. 21, no. 800, pp. 681–688. (2008).

[ITT00] Itti L. and Koch C., “A saliency-based search mechanism for overt and covert shifts of visual attention”. Vis. Res., vol. 40, pp. 1489–1506, (2000).

[ITT04] Itti L. “Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention”. IEEE TRANSACTIONS ON IMAGE PROCESSING, vol. 13, no. 10, (2004).

[ITT05] Itti L., and Baldi P. Bayesian surprise attracts human attention. Vision research, 49(10), 1295-1306. (2009).

[ITT98] Itti L. and Koch C. and Niebur E., “A model of saliency-based visual attention for rapid scene analysis”. IEEE Trans. Pattern Anal. Intell, vol. 20, no. 11, pp. 1254–1259. (1998).

[JM86] Hühring, K. “H.264 Reference Software Group” Available: www.iphome.hhide, joint model 86 (JM86).

[JUD09] Judd T., Ehinger K., Durand F., and Torralba A., “Learning to Predict Where Humans Look”. Proc. 12th IEEE Int’l Conf. Computer Vision, (2009).

[KHA15] Khatoonabadi H.S., Vasconcelos N., Bajic I.V., and Shan Y. “How many bits does it take for a stimulus to be salient”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5501-5510, (2015).

[KIM11] Kim W., Member S., Jung C., Kim C., and Member S., “Spatiotemporal Saliency Detection and Its Applications in Static and Dynamic Scenes”. IEEE Trans. Circuits Syst., vol. 21, no. 4, pp. 446–456, (2011).

[KIM14] Kim W. and Kim C., "Spatiotemporal saliency detection using textural contrast and its applications". IEEE Trans.Circuits and Systems for Video Technology, vol. 24 no. 4, pp 646-659.(2014).

[KOC85] Koch C. and, Ullman S., "Shifts in selective visual attention: towards the underlying neural circuitry". Hum Neurobiol. 1985; 4(4):219-27.(1985).

[KRA05] Kramer P., Hadar O., Benois-Pineau J., and Domenger J. P., "Super-resolution mosaicing from mpeg compressed video". In IEEE International Conference on Image Processing (Vol. 1, pp. I-893). IEEE. (2005).

[KUL51] Kullback S.and Leibler R. A., "On information and sufficiency". The Annals of Mathematical Statistics vol. 22, No. 1 pp. 79-86.(1951).

[KUL68] S. Kullback., "Information Theory and Statistics". vol. 1, no. 2. (1968)

[LEC13] Callet P.L. and Niebur E., "Visual Attention and Applications in Multimedia Technologies". Proceedings of the IEEE Institute of Electrical and Electronics Engineers. pp:2058-2067. (2013).

[LEM06] Le Meur O., Le Callet P., Barba D., and Thoreau D., "A coherent computational approach to model bottom-up visual attention". IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 5, pp. 802–817, (2006).

[LEM07] Le Meur O., Le Callet P. and Barba D., "Predicting visual fixations on video based on low-level visual features". Vision Res., vol. 47, no. 19, pp. 2483–2498, (2007).

[LI12] Li C., Wang Y., Ma B. and Zhang Z., "Tamper detection and self-recovery of biometric images using salient region-based authentication watermarking scheme". Computer Standards and Interfaces, 34(4), 367-379. (2012).

[LU10] Lu T., Yuan Z., Huang Y., Wu D., and Yu H., "Video retargeting with nonlinear spatial-temporal saliency fusion". Proc. - Int. Conf. Image Process. ICIP, pp. 1801–1804, (2010).

[LUO11] Luo Y., Yuan J., Xue, P. and Tian Q., "Saliency density maximization for efficient visual objects discovery". IEEE Trans.Circuits and Systems for Video Technology vol.21 no. 12, pp 1822-1834.(2011).

[MA03] Ma Y.F. and Zhang. H.-J. "Contrast-based image attention analysis by using fuzzy growing". In ACM Multimedia, (2003).

[MAL90] Malik J. and Perona P., "Preattentive texture discrimination with early vision mechanisms". JOSA A, 7(5), 923-932. (1990).

[MAN08] Manerba F., Benois-Pineau J., Leonardi R., and Mansencal B., "Multiple moving object detection for fast video content description in compressed domain". EURASIP Journal on Advances in Signal Processing, 2008(1), 1-15. (2007).

[MAR09] Marat S., Phuoc Ho., T., Granjon L.; Guyader N., Pellerin D., Guérin-Dugué A., "Modelling spatio-temporal saliency to predict gaze direction for short videos". Int. J. Comput. Vis 2009., vol. 82, no. 3, pp. 231–243. (2009)

[MIT07] Mitrea M., Prêteux F., and Nunez J., "Procédé de tatouage d'une séquence video". (for SFR and GET), French patent no. 05 54132 (29/12/2005), and EU patent no. 1804213 (04/07/2007).

[MOH08] Mohanty S.P., Bhargava B.K., "Invisible watermarking based on creation and robust insertion–extraction of image adaptive watermarks". ACM Transactions on Multimedia Computing, Communica- tions, and Applications (TOMCCAP) 5 (2) (2008)

[MUD13] Muddamsetty S., Sidibé D., Trémeau A. and Mériaudeau F. "A Performance Evaluation of Fusion Techniques for Spatio-Temporal Saliency Detection in Dynamic Scenes". In ICIP, 1-5 (2013).

[MUL85] Mullen. K., “The contrast sensitivity of human color-vision to red green and blue yellow chromatic gratings”. Journal of Physiology, pages 381–400, (1985).

[MUR11] Murray N., Vanrell M., Otazu X. and Parraga C. A., “Saliency estimation using a non-parametric low-level vision model”. In Computer Vision and Pattern Recognition (CVPR), IEEE Conference on (pp. 433-440). IEEE. (2011).

[NIU11] Niu Y., Kyan M., Ma L., Beghdadi A. and Krishnan S., "A visual saliency modulated just noticeable distortion profile for image watermarking". Signal Processing Conference19th European, Barcelona, pp. 2039-2043.(2011).

[NOO05] Noorkami M., Mersereau R. M., “Compressed-domain video watermarking for H.264”. Proc. ICIP, pp. 890–893, Atlanta, (2005)

[OGA15] Ogawa K. and Ohtake G., “Watermarking for HEVC/H. 265 stream”. In 2015 IEEE International Conference on Consumer Electronics (ICCE) (pp. 102-103). IEEE. (2015).

[PEN10] Peng J. and Xiao-Lin Q., “Keyframe-based video summary using visual attention clues”. IEEE Multimed, Vol. 17, pp. 64–73, (2010).

[PER12] Perazz F., Krahenbuhl P., Pritch Y., and Hornung A., “Saliency filters: Contrast based filtering for salient region detection”. IEEE CVPR, pp 733–740, (2012).

[PET 93] Peterson H.A., Ahumada A.J., and, Watson A.B., ”Improved detection model for DCT coefficient quantization”. Proc. of the SPIE Conference on Human Vision, Visual Processing and Digital Display IV, 1913, pp 191-201, (1993).

[POP09] Poppe C., De Bruyne S., Paridaens T., Lambert P., and Van de Walle R., “Moving object detection in the H. 264/AVC compressed domain for video surveillance applications”. Journal of Visual Communication and Image Representation, 20(6), pp 428-437, (2009).

[RAH10] Rahtu E., Kannala J., Salo M., and Heikkila J., “Segmenting salient objects from images and videos”. ECCV, (2010).

[REN09] Ren T., Liu Y., and Wu G., “Image retargeting based on global energy optimization”. ICME, (2009).

[RIC03] Richardson E., “H264 and MPEG-4 AVC Video compression: Video coding for next generation Multimidia”. H.264/MPEG-4 Part 10 White Paper, (2003).

[RUB08] Rubinstein M., Shamir A., and Avidan S., “Improved seam carving for video retargeting”. ACM TOG, (2008).

[SEO09] Seo H. J. and Milanfar P., “Nonparametric bottom-Up saliency detection by self-resemblance”. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp 45–52, (2009).

[SHA58] Shannon C. E., “Channels with Side Information at the Transmitter”. IBM Journal, pp 289-293, (1958).

[SUL12] Sullivan G. J., Ohm J. R., Han W. J., and Wiegand T., “Overview of the high efficiency video coding (HEVC) standard”. ,” IEEE Trans. Circuits Syst. Video Technol., Vol. 22, no. 12, pp 1649–1668, (2012).

[SUR09] Sur A., Sagar S.S., Pal R., Mitra P., and Mukhopadhyay J., “A New Image Watermarking Scheme Using Saliency Based Visual Attention Model". Annual IEEE India Conference, Gujarat, pp. 1-4, (2009).

[THI06] Thiemert S., Sahbi H., and Steinebash M., “Using entropy for image and video authentication watermarks”. Proc. SPIE on Electronic imaging: Security, Steganography and watermarking of Multimedia Contents, Vol 6072, pp 218–228, USA, (2006).

[TIA11] Tian L., Zheng N., Xue J., Li C., and Wang X., "An integrated visual saliency-based watermarking approach for synchronous image authentication and copyright protection". Signal Processing: Image Communication, pp 427-437, (2011).

[TRE80] Treisman A.M., and Gelade G., "A feature-integration theory of attention". Cogn. Psychol, Vol 12, no 1, pp 97–136, (1980).

[TRE88] Treisman A. and Gormican S., "Feature analysis in early vision: evidence from search asymmetries". Psychol Rev, Vol 95, pp 15–48, (1988).

[TUR12] TU-R BT.2021, "Subjective methods for the assessment of stereoscopic 3dtv systems". International Telecommunication Union, Geneva, Switzerland, (2012).

[VER96] Verscheure O., Basso A., El-Maliki M., and Hubaux, J. P., "Perceptual bit allocation for MPEG-2 CBR video coding". In Image Processing, Proceedings, International Conference, Vol. 1, pp 117-120, (1996).

[WAL06] Walther D., and Koch.Modeling C., "Attention to salient protoobjects". Neural Networks, pp 1-5, (2006).

[WAL89] Walpole R.E. and Myers R.H., "Probability and Statistics for Engineers and Scientists". 4th edn MacMillan Publishing, New York, (1989).

[WAN15] Wan W., Liu J., Sun J., Ge C., and Nie X., "Logarithmic STDM watermarking using visual saliency-based JND model". Electronics Letters, 51(10), pp 758-760, (2015).

[WAT97] Watson A.B. and Solomon J. A., "Model of visual contrast gain control and pattern masking". Journal of the Optical Society of America, Vol 14, no 9, pp 2379-2391, (1997).

[WEB01] https://www.statista.com/statistics/609608/internet-users-time-spent-tv-video-content/

[WEB02] https://eagleyedguide.blogspot.fr/2016/04/five-reasons-why-video-content-is.html

[WEB03] https://www.thinkwithgoogle.com/articles/millennials-eat-up-youtube-food-videos

[WEB04] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

[WEB05] ftp://.ivc.polytech.univnantes.fr/IRCCyN_IVC_Eyetracker_SD_2009_12/.

[WEB06] https://crcns.org/data-sets/eye/eye-1.

[WEB07] https://sites.google.com/site/saliencyevaluation-measures

[WEB08] http://www.detectingdesign.com/humaneye.html

[WEB09] https://www.boundless.com/psychology/textbooks/boundless-psychology-textbook/sensation-and-perception-5/sensory-processes-38/vision-the-visual-system-the-eye-and-color-vision-161-12696/

[WEB10] https://skeptoid.com/blog/2013/12/24/is-the-human-eye-irreducibly-complex/

[WEB11] ISO/IEC JTC1, Information technology Coding of audio-visual objects Part2: Visual, ISO/IEC 14492-2, (MPEG-4 Visual), Version 1: Apr.1999, Version 2: Feb. 2000, Version 3, (2004).

[WOL07] Wolf L., Guttmann M., and Cohen-Or D., "Non-homogeneous content-driven video-retargeting. In Computer Vision". ICCV 2007. IEEE 11th International Conference on, pp 1-6, (2007).

[XIA10] Xiao X., Xu C., and Rui Y., "Video based 3D reconstruction using spatio-temporal attention analysis". IEEE International Conference on Multimedia and Expo, pp 1091–1096, (2010).

[YAN15] Yang J., Zhao G., Yuan J., Shen X., Lin Z., Price B., and Brandt J., "Discovering Primary Objects in Videos by Saliency Fusion and Iterative Appearance Estimation". IEEE Trans. Circuits and Systems for Video Technology. Vol PP, no 99, pp 1, (2015).

[YU07] Yu M., He H., and Zhang J., "A digital authentication watermarking scheme for JPEG images with superior localization and security". Science in China Series F. Information Sciences, pp 491-509, (2007).

[ZHA06] Zhai Y., and Shah M., "Visual Attention Detection in Video Sequences Using Spatiotemporal Cues Categories and Subject Descriptors". Proceedings of the 14th annual ACM international conference on Multimedia. Vol 32816, pp 815–824, (2006).

[ZHA08] Zhang L., Tong M.H., Marks T.K., Shan H., and Cottrell G.W., "SUN: A Bayesian framework for saliency using natural statistics". Journal of vision, 8(7), pp 32-32, (2008).

[ZHA09] Zhang L., Tong M., and Cottrell G., "SUNDAy: Saliency using natural statistics for dynamic analysis of scenes". Submitted to International Conference on Computer Vision, (2009).

[ZHA12] Yubo Z., Hongbo B., and Haiyan Z., "A robust watermarking algorithm based on salient image features," Journal of Computational Information Sys- tems. Vol 8, no. 20, pp 8421–8426, (2012).

[ZHI09] Zhi L., Hongbo Y., Liquan S., Yongfang W., and Zhaoyang Z., "A Motion Attention Model Based Rate Control Algorithm for H.264/AVC". Eigth IEEE/ACIS International Conference on Computer and Information Science, (2009).

[ZHO10] Zhou Y., Li L., and Liu J., "A Digital Fingerprint Scheme Based on MPEG-2". In Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), Sixth International Conference on, pp 611-614. IEEE, (2010).

List of publications

Published papers

1 M. Ammar, M. Mitrea, and M. Hasnaoui, “MPEG-4 AVC saliency map computation” in IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, 90141A-90141A (2014)

2 M. Ammar, M. Mitrea, and M. Hasnaoui, P. Le Callet, “Visual saliency in MPEG-4 AVC video stream” in IS&T/SPIE Electronic Imaging International Society for Optics and Photonics, pp. 93940X–93940X (2015)

3 M. Ammar, M. Mitrea, I. Boujelben, and P. Le Callet, “HEVC saliency map computation” in Electronic Imaging, HVEI-107 - 1-8 (2016)

Oral presentations

1 M. Ammar, M. Mitrea, “Saillance visuelle pour le flux compressé MPEG-4 AVC”, GDR ISIS, Journée commune Thèmes B et D. Saillance visuelle et applications au tatouage et à la compression d'images et de vidéos (2014)

2 M. Ammar, M. Mitrea, I. Boujelben, P. Le Callet “HEVC saliency map computation”, GDR ISIS, Journée commune Thèmes B et D. Saillance visuelle et applications au tatouage et à la compression d'images et de vidéos (2016)

Technical contributions to the MEDUSA ITEA2 and MEDOLUTION ITEA3 R&D collaborative projects (under the supervision of M. Mitrea)

MEDUSA: D4.1.1 – State of the Art on secure, dependable data transfer (Dec 2013)

MEDUSA: D4.1.2 – Preliminary Design of security & transmission component (June 2014)

MEDUSA: D4.2.2 - Final release of security & transmission components (Oct 2015)

MEDOLUTION: D1.1. State of the Art Analysis (Nov. 2016)

Submitted journal paper

M. Ammar, M. Mitrea, and M. Hasnaoui, P. Le Callet, “MPEG-4 AVC stream-based saliency detection. Application to robust watermarking”

List of acronyms

ASP	Advanced Simple Profile
AUC	Area Under Curve
AVC	Advanced Video Coding
AWGN	Additive White Gaussian Noise
B frame	Bidirectional predicted frame
BER	Bit Error Rate
CABAC	Context Adaptive Binary Arithmetic Coding
Card	Cardinality
CAVLC	Context Adaptive Variable Length Coding
CC	Correlation Coefficient
CD	Compact Disc
CDMA	Code division multiple access
CRCNS	Collaborative Research in Computational Neuroscience
CSF	Center Surround Filters
CTU	Coding Tree Unit
CU	Coding Units
dB	The decibel
DC	Direct Component
DCT	Discrete Cosine Transform
DST	Discrete Sine Transform
DVD	Digital Versatile Disc
DVQ	Digital Video Quality
DWT	Discrete Wavelet Transform
E	Entropic coding
FAR	False Alarm Rate
FIFA	International Federation of Association Football
GB	GigaByte
GBVS	Graph-Based Visual Saliency
GOP	Groupe Of Picture
HD	High Definition

HEVC	High Efficiency Video Coding
HR	Hit Rate
HVS	human visual system l
Hz	Hirtz
I frame	Intra frame
ICL	Incremental Coding Length
IRCCyN	Institut de Recherche en Communications et Cypbermétique de Nanes
ISBN	International Standard Book Number
ITU	International Telecommunication Union
JND	Just Noticeable Distortion
JPEG	Joint Photographic Experts Group
KLD	Kullback-Leibler Divergence
MAE	Mean Absolute Error
Max	Maximal
Min	Minimal
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MVD	Motion Vector Differences
NCC	Normalized Cross Correlation
NHS	Normalized Hamming Similarity
P	Prediction
P frame	Predicted frame
PB	Predicted Block
PC	Personal Computer
PQFT	Phase spectrum of Quaternion Fourier Transform
PSNR	Peak Signal to Noise Ratio
Q	Quantization
QIM	Quantization Index Modulation
RAM	Random Access memory
ROB	regions of background
ROC Curve	Receiver Operating Characteristic Curve,
ROI	Region Of Interest

SD	standard-definition
SFC	Spatial Frequency Content
SI	Side Information
SR	Super resolution
SS	Spread Spectrum
SSCQE	Single Stimulus Continuous Quality Evaluation
STDM	Spread Transform Dither Modulation
T	Transformation
TB	Tranformed Block
TPE	Total Perceptual Error
TV	Television
US	United states