Méthode d'estimation locale de l'impact des dégradations H.264 sur la qualité vidéo subjective en télévision haute définition et cumul en une estimation globale

Stéphane Péchard[1] Patrick Le Callet[1] Dominique Barba[1]

[1] Université de Nantes – IRCCyN, équipe IVC
Polytech'Nantes, rue Christian Pauc, 44306 Nantes

{stephane.pechard, patrick.lecallet, dominique.barba}@univ-nantes.fr

Résumé

Cet article propose une méthode d'évaluation de la qualité subjective de séquences vidéo par partie et un modèle de cumul des qualités locales en une note de qualité globale. Les techniques existantes évaluent l'impact sur la qualité de classes de dégradations, ce qui les rend non robustes au changement de contenu. La méthode proposée inverse le problème de classification de manière à éviter cet écueil. La séquence est évaluée par partie et globalement, afin d'obtenir des notes de qualité pour la séquence et pour chaque partie. Une relation de cumul entre les notes locales et globale pourrait permettre de fragmenter la problématique de l'évaluation de qualité par zones de contenu homogène. Nous montrons ici qu'une telle relation est possible.

Mots clefs

Qualité vidéo, H.264, TVHD, tests subjectifs.

1 Introduction

Il existe plusieurs études concernant l'évaluation de qualité subjective des dégradations de codage vidéo [1, 2]. La majorité d'entre elles considère l'influence de plusieurs dégradations de codage. Cependant, l'aspect fortement temporel de l'évaluation de la qualité vidéo est souvent sous-estimé. Farias [2] synthétise de telles dégradations (flou, ringing, effet de bloc, bruit) dans le but de les appliquer indépendamment ou par combinaison sur des régions isolées de la séquence. Wolff [1] utilise des séquences dégradées par H.264. Ses observateurs ont alors deux tâches : évaluer la gêne globale sur la séquence et mesurer la force de chaque type de dégradation. Ces deux approches sont indépendantes des spécificités du contenu.

H.264 est considéré ici comme produisant des dégradations (dues à la quantification) qui peuvent conduire à différentes gênes perçues suivant la région spatio-temporelle où elles apparaissent. En fait, la perception des dégradations dépend fortement du contenu local. Par exemple, appliquer la même erreur de quantification donne une dégradation particulièrement visible dans une zone homogène. Cependant, cette même erreur peut être partiellement ou totalement masquée dans une zone texturée.

L'approche proposée consiste à dégrader indépendamment des régions spatio-temporelles cohérentes en termes de contenu. Au lieu de définir des classes de dégradations, il est plus pertinent de définir des classes de contenu. Des dégradations de codage réelles sont utilisées afin de refléter l'usage réel. Les séquences dégradées partiellement ou entièrement sont évaluées par des tests subjectifs. À partir des résultats, une relation entre les notes de qualité partielle et la note de qualité globale de la séquence peut être obtenue. Plusieurs combinaisons sont proposées et discutées dans cet article.

2 La méthode proposée

Le système visuel humain a différentes perceptions des dégradations suivant le contenu spatio-temporel dans lequel elles apparaissent. Par exemple, une même dégradation dans une zone homogène sera perçue plus facilement que dans une zone de textures. Ainsi plusieurs classes de contenu ont été définies afin d'étudier indépendamment leur impact. Ces cinq classes sont : des zones homogènes de faible luminance (C_1), des zones homogènes de forte luminance (C_2), des zones de textures fines (C_3), des contours (C_4) et des zones de textures fortes (C_5).

Chaque classe correspond à une certaine activité spatiale du contenu, et donc à un certain impact des dégradations dues au codage H.264 sur la perception de la qualité par les observateurs. L'approche présentée consiste en plusieurs étapes. Tout d'abord, chaque séquence est segmentée spatio-temporellement. Puis chaque segment est classifié selon une partition définie. Des séquences dégradées suivant cette classification sont ensuite générées et évaluées par des observateurs. Cette évaluation permet d'obtenir une note de qualité pour chaque segment de la séquence. Enfin, les notes locales des classes sont confrontées à la note donnée par les observateurs à la séquence entièrement dégradée.

2.1 Segmentation

À partir de la séquence originale au format 1080i non compressée, des volumes spatio-temporels élémentaires sont créés. Pour cela, la séquence est découpée en « tronçons »

de cinq images, soit dix trames. Pour les deux ensembles de cinq trames de même parité, chaque bloc de la trame centrale subit une estimation de mouvement en utilisant les deux trames précédentes et les deux suivantes comme le montre la figure 1.

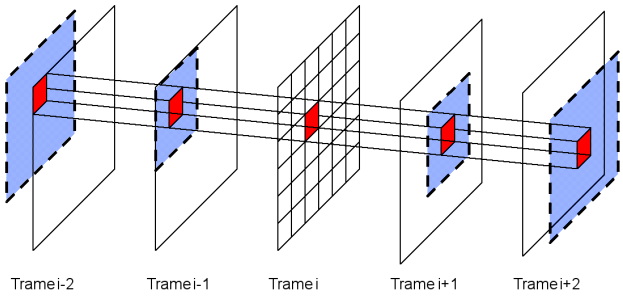


Figure 1 – *Création d'un tube spatio-temporel le long de cinq trames.*

Spatialement, les fenêtres de recherche sont définies afin de pouvoir contenir le plus grand mouvement de la séquence. Les différences entre blocs sont évaluées par l'erreur moyenne quadratique (MSE) sur les trois composantes YUV. Le vecteur de mouvement retenu est celui minimisant cette MSE. Pour chaque bloc de l'image, le vecteur de mouvement est pris comme la moyenne des deux vecteurs issus de la trame paire et de la trame impaire correspondantes. Le suivi temporel de chaque bloc définit un « tube » spatio-temporel élémentaire. Celui-ci est orienté selon le mouvement local. Ce concept de tube tridimensionnel a été introduit par Wolf et Pinson [3] pour une métrique objective de qualité vidéo. Dans l'approche de Wolf, les tubes sont fixes dans le temps, alors qu'ici ils sont orientés le long du mouvement local. Ainsi, les tubes temporels sont cohérents en termes de mouvement et d'activité spatiale. Cette estimation de mouvement est réalisée sur une représentation multi-résolution afin de réduire la complexité calculatoire. Elle est d'abord calculée à la plus faible résolution, puis le vecteur obtenu est affiné en tenant compte de la résolution supérieure et ainsi de suite.

Enfin, les tubes sont fusionnés temporellement de proche en proche pour former des volumes spatio-temporels le long de la séquence entière. Cette fusion consiste à assigner la même classe aux tubes qui se chevauchent comme le montre la figure 2. Des portions d'image non classées peuvent apparaitre entre les tubes. Elles sont fusionnées avec la classe existante la plus proche spatialement. Ainsi, chaque pixel de la séquence source appartient à une classe unique.

2.2 Classification

La seconde étape de la segmentation consiste à fusionner les volumes spatio-temporels en classes de contenu homogène. Ceci permet de suivre temporellement des objets tout le long de la séquence. Cette étape utilise quatre gradients spatiaux (ΔH, ΔV, ΔD_{45° et ΔD_{135°) calculés

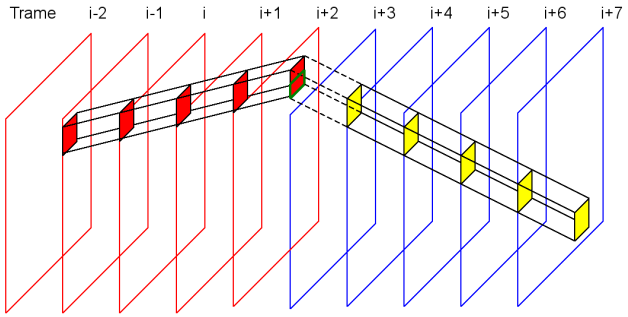


Figure 2 – *Fusion des tubes se chevauchant.*

pour chaque bloc B de l'image I en chaque pixel de coordonnées (m, n) :

$$\Delta H(m,n) = |I(m,n+1) - I(m,n)| \quad (1)$$

$$\Delta V(m,n) = |I(m+1,n) - I(m,n)| \quad (2)$$

pour lesquels nous calculons les moyennes sur le bloc B :

$$\overline{\Delta H} = \sum_{\substack{0 \le m \le M \\ 0 \le n \le N}} \Delta H(m,n) \quad (3)$$

$$\overline{\Delta V} = \sum_{\substack{0 \le m \le M \\ 0 \le n \le N}} \Delta V(m,n) \quad (4)$$

avec $M \times N$ les dimensions du bloc B.
Chaque bloc peut être placé dans l'espace $P = (\overline{\Delta H}, \overline{\Delta V})$ où sont définies les classes d'activité. La figure 3 présente cet espace. Ainsi, un bloc des zones C_1 et C_2 appartiennent

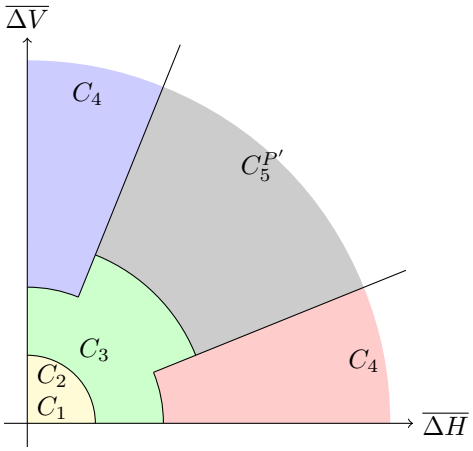


Figure 3 – *Espace* $(\overline{\Delta H}, \overline{\Delta V})$ *permettant le classement des blocs.*

à une zone homogène. Un bloc de C_3 appartient à une zone de textures fines. Un bloc de C_4 appartient à une zone de contour. $C_5^{P'}$ est la zone indeterminée contenant les blocs pour lesquels l'ambiguïté concernant l'orientation spatiale n'est pas levée. Pour ces blocs, les activités spatiales diagonales sont calculées :

$$\Delta D_{45}(m,n) = |I(m+1,n-1) - I(m,n)| \quad (5)$$

$$\Delta D_{135}(m,n) = |I(m+1,n+1) - I(m,n)| \quad (6)$$

ainsi que leurs moyennes sur le bloc B :

$$\overline{\Delta D_{45}} = \sum_{\substack{0 \le m \le M \\ 0 \le n \le N}} \Delta D_{45}(m,n) \quad (7)$$

$$\overline{\Delta D_{135}} = \sum_{\substack{0 \le m \le M \\ 0 \le n \le N}} \Delta D_{135}(m,n) \quad (8)$$

Comme précédemment, chaque bloc B est placé dans l'espace $P' = (\overline{\Delta D_{45}}, \overline{\Delta D_{135}})$, formé de manière identique à P. Dans ce nouveau plan, un bloc des zones C'_1 et C'_2 appartiennent à une zone homogène. Un bloc de C'_3 appartient à une zone de textures fines. Un bloc de C'_4 appartient à une zone de contour. Un bloc de C'_5 appartient à une zone de textures fortes.
La fusion des tubes suivant cette classification forment les cinq classes de la séquence.

2.3 Génération des séquences

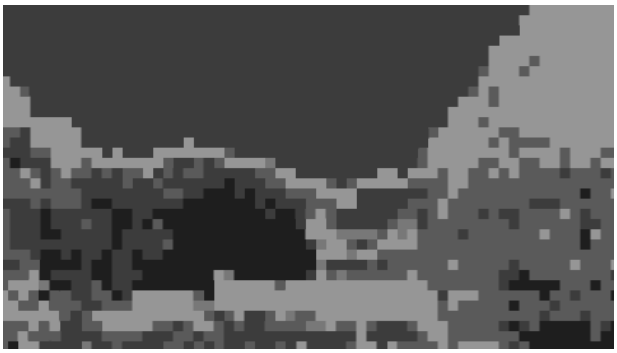
Des séquences dégradées sont générées à partir de la séquence décompressée originale, des séquences dégradées par H.264 et de la classification. Le codage H.264 est effectué par le codeur de référence [4]. Plusieurs débits sont générés afin de couvrir une gamme de qualité significative. Les débits utilisées sont présentées dans le tableau 1. Les séquences ont été fournies par la chaine de télévision suèdoise SVT.

Séquence	Débits (Mbps)
(a) Above Marathon	5 ; 8 ; 10
(b) Captain	1 ; 3 ; 5
(c) Dance in the Woods	3 ; 5 ; 6
(d) Duck Fly	4 ; 6 ; 8
(e) Fountain Man	1 ; 5
(f) Group Disorder	2 ; 4
(g) Rendezvous	6 ; 8
(h) Ulriksdals	1 ; 4

Tableau 1 – *Débits pour chaque séquence (en Mbps).*

Les zones d'une séquence dégradée correspondant à une classe sont insérées dans la séquence originale. Ce processus crée une séquence par classe et par débit dont une partie spatio-temporellement cohérente est dégradée. Les étapes sont présentées sur la figure 4 avec seulement le quart de la première image de la séquence *Above Marathon*. L'image (a) présente la séquence originale non dégradée. L'image (b) montre les différentes classes. Pour plus de visibilité, les valeurs de luminance Y d'un pixel de cette image sont obtenues par la formule $Y = i \times 30$ avec i l'index de la classe. La classe C_1 seule est dégradée dans l'image (c). Ceci est visible dans les arbres au centre de l'image.

(a) Image originale.



(b) Classification d'une image.

C_1 ■ ; C_2 ■ ; C_3 ■ ; C_4 ■ ; C_5 ■

(c) Image partiellement dégradée (seulement C_1).

Figure 4 – *Étapes de création des séquences dégradées.*

2.4 Évaluation subjective de qualité

Des tests d'évaluation subjective de la qualité des séquences générées ont été effectués. Leur but est de mesurer individuellement l'impact de chaque classe sur la qualité perçue. Suivant les recommandations internationales pour les conditions de test [5], ces évaluations ont utilisé au moins 15 observateurs valides. L'écran utilisé est un LCD Philips de résolution 1920×1080. Les séquences TVHD 1080i non compressées sont lu par un lecteur V1-UHD de Doremi.

3 Résultats de la segmentation

Dans chaque séquence, une classe occupe une certaine proportion de l'espace spatio-temporel. Cette proportion est calculée comme le ratio entre le nombre de pixels dans la classe C_i et le nombre total de pixel de la séquence. Les proportions des classes pour chaque séquence sont présentées dans le tableau 2 (en pourcentages).

Séquence	C_1	C_2	C_3	C_4	C_5
(a)	3.75	17.45	27.79	0.94	50.06
(b)	13.14	78.26	6.81	1.43	0.36
(c)	3.80	22.57	53.85	3.02	16.75
(d)	0.13	8.97	19.50	10.70	60.70
(e)	10.52	70.71	13.37	1.45	3.93
(f)	25.28	38.58	29.80	1.79	4.54
(g)	8.78	12.38	19.87	2.05	56.92
(h)	13.54	41.31	40.48	1.36	3.30

Tableau 2 – *Proportions des classes pour toutes les séquences. Les labels des séquences sont ceux du tableau 1. La classe C_1 correspond aux zones homogènes de faible luminance, C_2 aux zones homogènes de forte luminance, C_3 aux zones de textures fines, C_4 aux contours et C_5 aux zones de textures fortes.*

La classe C_1 (zones homogènes de faible luminance) a une gamme de valeurs modérée, jusqu'à 25% de l'image. La classe C_2 (zones homogènes de forte luminance) a la plus large gamme de proportions. Les séquences *Captain* et *Fountain Man* ont des proportions particulièrement fortes dues à la classification dans cette classe des zones de chute d'eau. La classe C_3 (zones de textures fines) a une forte importance, entre 6 et 54%. La classe C_4 (contours) a des proportions particulièrement faibles. À part une séquence à 10%, sa proportion est inférieure ou égale à 3%. Enfin, la classe C_5 a une large gamme de valeur, allant de quasiment zéro (0.36%) à plus de 60%.

Ces proportions sont cohérentes avec la nature des séquences. Celles-ci sont faites de contenus réalistes avec une prise de vue en extérieur. Ainsi, elles contiennent peu de contours, quelques zones homogènes (comme le ciel) et beaucoup de textures (arbres, herbe, etc.).

4 Relation entre les pertes de qualité globale et locales

4.1 Définition des pertes de qualité

La note de qualité subjective de la séquence S_j non dégradée est notée MOS(S_j) et celle de la séquence entièrement dégradée à un débit B_k MOS(S_j, B_k). De plus, une note de qualité, notée MOS(S_j, B_k, C_i), est obtenue pour chaque séquence S_j, chaque débit B_k et chaque classe C_i.
La différence entre MOS(S_j) et MOS(S_j, B_k) représente la perte de qualité globale, notée DMOS(S_j, B_k). La différence entre un MOS(S_j, B_k, C_i) partiel et MOS(S_j) est appelée ΔMOS(C_i, S_j, B_k). Elle indique la perte de qualité induite par les dégradations de la classe C_i. Chaque classe introduit une perte de qualité qui est une part de la perte de qualité globale DMOS.
La figure 5 montre une représentation de ce type de relation. Sur une échelle de qualité allant de 0 à 100, le MOS(S_j) est la note maximale obtenue et MOS(S_j, B_k) la note minimale. Chaque ΔMOS est une perte de qualité partielle résultant de l'évaluation d'une séquence partiellement dégradée.

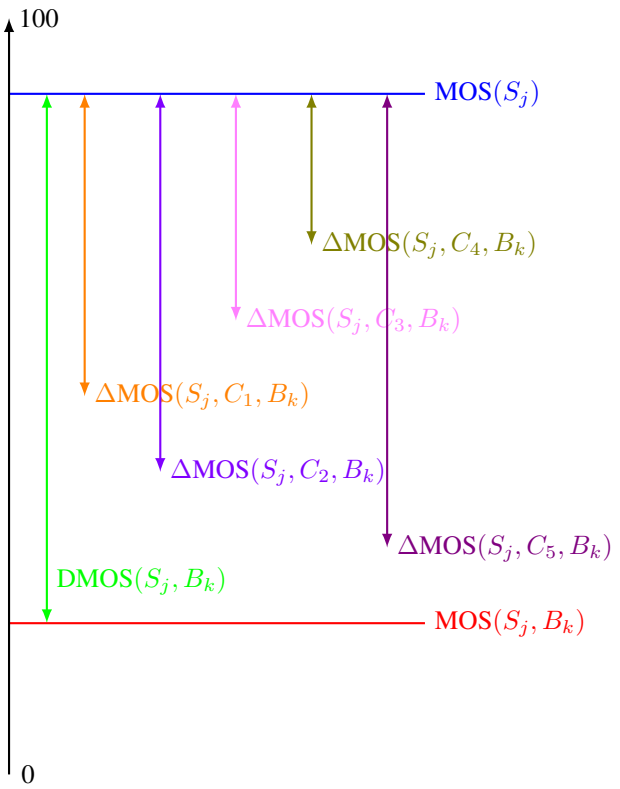


Figure 5 – *Représentation du DMOS global et des ΔMOS locaux.*

4.2 Peut-on relier les pertes locales à la perte globale ?

Une relation entre le DMOS et les différents ΔMOS des classes seraient très intéressante dans la conception d'une

Combinaison	CC	RMSE
$C_2 + C_4 + C_5$	0.9485	14.51
$C_2 + C_5$	0.9440	12.55
$C_2 + C_3 + C_4$	0.9094	21.52
$C_1 + C_2 + C_3 + C_4 + C_5$	0.9058	67.16
$C_1 + C_2 + C_4 + C_5$	0.9052	35.42
$C_2 + C_3 + C_4 + C_5$	0.9041	44.53
...	...	...
C_2	0.7664	22.40
C_3	0.7094	28.54
C_5	0.6400	35.80
C_4	0.5472	54.64
C_1	0.5349	36.42

Tableau 3 – *Combinaisons des ΔMOS de classes et leurs CC et RMSE avec le DMOS global.*

métrique objective de la qualité. Celle-ci pourrait alors évaluer la qualité globale d'une séquence à partir des qualités locales.

Plusieurs opérations simples ont été testées dans le but d'établir une telle relation. Le tableau 3 présente des combinaisons basées sur la somme des ΔMOS d'une ou plusieurs classes. Pour chaque relation, le tableau fournit également le coefficient de corrélation (CC) entre cette somme et le DMOS global ainsi que l'erreur moyenne quadratique (RMSE). Pour CC < 0.9, seules les combinaisons avec une seule classe sont présentées.

4.3 Résultats

Les performances de ces combinaisons révèlent l'importance relative de chaque classe dans le processus de cumul réalisé par l'observateur moyen. Malgré sa simplicité, une telle approche permet une très bonne corrélation avec quelques classes stratégiques. La combinaison des seules zones homogènes de forte luminance (C_2) et des zones de textures fortes (C_5) obtient une très bonne corrélation et la plus faible des erreurs. Ceci s'explique notamment par la nature des séquences, celles-ci contenant en majorité ce type de zones. Malgré ses faibles proportions et ses mauvaises performances individuelles (CC=0.5472 et RMSE=54.64), la classe C_4 est présente dans cinq des six premières combinaisons. Ainsi, les dégradations présentes dans ces trois classes C_2, C_5 et C_4 sont étroitement liées à la qualité globale de la séquence.

Les combinaisons avec une seule classe obtiennent les plus mauvais résultats. Cela signifie que la qualité d'une seule classe ne suffit pas à représenter la qualité globale de la séquence. En contrepartie, la combinaison des cinq classes n'obtient pas la meilleure performance, indiquant que certaines zones ne sont pas ou peu prises en compte dans la construction du jugement de qualité des observateurs.

Notons également que les erreurs quadratiques moyennes sont assez fortes, révélant la médiocre précision de la relation. Néanmoins, ce modèle cherche à approcher le cumul et non la prédiction du DMOS global. De plus, les proportions spatiales de chaque classe ne sont pas prises en compte dans les combinaisons. Cette information permet de refléter l'importance relative de chaque classe dans la construction du jugement global.

5 Conclusion

Dans cet article, la relation entre la qualité globale d'une séquence et les qualités partielles de zones homogènes la constituant est étudiée. Pour cela, une méthode de segmentation spatio-temporelle est utilisée pour isoler ces zones homogènes et les évaluer séparément par des tests subjectifs. Plusieurs combinaisons des notes de qualité locale en une note de qualité globale ont été présentées. Elles révèlent l'importance relative de chaque classe et la possibilité de relier le jugement local au jugement global.

Loin d'être triviale, l'existence de cette relation permet la simplification du problème de l'évaluation de qualité vidéo, ainsi fragmentée par zones de contenu homogène. Dans l'optique d'un critère de qualité, celui-ci serait conçu pour prédire indépendemment la qualité de chacune des zones. Ensuite, une relation telle que celles présentées ici permettrait de cumuler les qualités locales en une note de qualité pour la séquence entière. La connaissance a priori de la classe considérée serait un atout dans la conception de la métrique.

Références

[1] Tobias Wolff, Hsin-Han Ho, John M. Foley, et Sanjit K. Mitra. H.264 coding artifacts and their relation to perceived annoyance. Dans *Proceedings of European Signal Processing Conference*, 2006.

[2] Mylène Farias. *No-reference and reduced reference video quality metrics : new contributions*. Thèse de doctorat, University of California, 2004.

[3] Stephen Wolf et Margaret H. Pinson. Spatial-temporal distortion metric for in-service quality monitoring of any digital video system. Dans *Proceedings of SPIE, Multimedia Systems and Applications II*, volume 3845, pages 266–277, 1999.

[4] Joint Video Team (JVT). H.264/Advanced Video Coding reference software version 10.2, 2006. http ://iphome.hhi.de/suehring/tml/.

[5] ITU-R BT. 500-11. Methodology for the subjective assessment of the quality of television pictures. Rapport technique, International Telecommunication Union, 2004.