Union internationale des télécommunications

UIT-T

SECTEUR DE LA NORMALISATION
DES TÉLÉCOMMUNICATIONS
DE L'UIT

J.144

(03/2004)

SÉRIE J: RÉSEAUX CÂBLÉS ET TRANSMISSION DES SIGNAUX RADIOPHONIQUES, TÉLÉVISUELS ET AUTRES SIGNAUX MULTIMÉDIAS

Mesure de la qualité de service

Techniques de mesure objective de la qualité vidéo perçue pour la télévision numérique par câble en présence d'un signal de référence complet

Recommandation UIT-T J.144

Recommandation UIT-T J.144

Techniques de mesure objective de la qualité vidéo perçue pour la télévision numérique par câble en présence d'un signal de référence complet

Résumé

La présente Recommandation contient des lignes directrices relatives au choix d'un équipement approprié de mesure de la qualité vidéo perçue à utiliser dans les applications de télévision numérique par câble lorsqu'on peut utiliser la méthode de mesure avec référence complète. Les données de tests de validation ne comportent pas d'erreurs de canaux. La présente Recommandation définit quatre modèles de calcul objectif dont il a été montré qu'ils étaient de meilleurs outils de mesure automatique que la valeur crête du rapport signal/bruit de crête (PSNR, *peak signal to noise ratio*) pour évaluer la qualité d'une séquence vidéo diffusée.

La présente Recommandation révisée propose dans la partie normative quatre méthodes de calcul objectif permettant d'évaluer la qualité vidéo perçue.

Source

La Recommandation UIT-T J.144 a été approuvée le 15 mars 2004 par la Commission d'études 9 (2001-2004) de l'UIT-T selon la procédure définie dans la Recommandation UIT-T A.8.

AVANT-PROPOS

L'Union internationale des télécommunications (UIT) est une institution spécialisée des Nations Unies dans le domaine des télécommunications et des technologies de l'information et de la communication (ICT). Le Secteur de la normalisation des télécommunications (UIT-T) est un organe permanent de l'UIT. Il est chargé de l'étude des questions techniques, d'exploitation et de tarification, et émet à ce sujet des Recommandations en vue de la normalisation des télécommunications à l'échelle mondiale.

L'Assemblée mondiale de normalisation des télécommunications (AMNT), qui se réunit tous les quatre ans, détermine les thèmes d'étude à traiter par les Commissions d'études de l'UIT-T, lesquelles élaborent en retour des Recommandations sur ces thèmes.

L'approbation des Recommandations par les Membres de l'UIT-T s'effectue selon la procédure définie dans la Résolution 1 de l'AMNT.

Dans certains secteurs des technologies de l'information qui correspondent à la sphère de compétence de l'UIT-T, les normes nécessaires se préparent en collaboration avec l'ISO et la CEI.

NOTE

Dans la présente Recommandation, l'expression "Administration" est utilisée pour désigner de façon abrégée aussi bien une administration de télécommunications qu'une exploitation reconnue.

Le respect de cette Recommandation se fait à titre volontaire. Cependant, il se peut que la Recommandation contienne certaines dispositions obligatoires (pour assurer, par exemple, l'interopérabilité et l'applicabilité) et considère que la Recommandation est respectée lorsque toutes ces dispositions sont observées. Le futur d'obligation et les autres moyens d'expression de l'obligation comme le verbe "devoir" ainsi que leurs formes négatives servent à énoncer des prescriptions. L'utilisation de ces formes ne signifie pas qu'il est obligatoire de respecter la Recommandation.

DROITS DE PROPRIÉTÉ INTELLECTUELLE

L'UIT attire l'attention sur la possibilité que l'application ou la mise en œuvre de la présente Recommandation puisse donner lieu à l'utilisation d'un droit de propriété intellectuelle. L'UIT ne prend pas position en ce qui concerne l'existence, la validité ou l'applicabilité des droits de propriété intellectuelle, qu'ils soient revendiqués par un membre de l'UIT ou par une tierce partie étrangère à la procédure d'élaboration des Recommandations.

A la date d'approbation de la présente Recommandation, l'UIT avait été avisée de l'existence d'une propriété intellectuelle protégée par des brevets à acquérir pour mettre en œuvre la présente Recommandation. Toutefois, comme il ne s'agit peut-être pas de renseignements les plus récents, il est vivement recommandé aux développeurs de consulter la base de données des brevets du TSB sous http://www.itu.int/ITU-T/ipr/.

TABLE DES MATIÈRES

Introduction

La télévision numérique donne lieu à de nouvelles considérations en termes de qualité de service, avec des relations complexes entre les mesures objectives de paramètres et la qualité subjective de l'image. S'il est souhaitable d'avoir une bonne corrélation entre les mesures objectives et l'évaluation subjective de la qualité afin d'obtenir une qualité de service optimale dans l'exploitation des systèmes de télévision par câble, il faut bien voir que les mesures objectives ne constituent pas un substitut aux évaluations subjectives de la qualité.

Les évaluations de qualité subjective sont des procédures soigneusement élaborées qui ont pour but de déterminer l'opinion moyenne de spectateurs au sujet de séquences vidéo pour une application donnée. Les résultats de ce type d'évaluation sont très utiles dans la conception des systèmes et les tests d'évaluation des performances. L'évaluation de la qualité subjective pour une application différente dans d'autres conditions donnera toujours des résultats révélateurs, même si les notes d'opinion pour le même ensemble de séquences seront sans doute différentes. Les mesures objectives sont destinées à une large gamme d'applications produisant des résultats identiques pour un même ensemble de séquences vidéo. Le choix des séquences vidéo qu'il convient d'utiliser et l'interprétation des mesures objectives qui en résultent sont quelques-uns des facteurs que l'on peut faire varier pour une application donnée.

Les mesures objectives et les évaluations subjectives de la qualité sont donc complémentaires plutôt qu'interchangeables. Si les évaluations subjectives répondent à des besoins liés à la recherche, les mesures objectives sont nécessaires dans la spécification des équipements ainsi que la surveillance et la mesure quotidiennes des performances des systèmes.

La convention terminologique suivante a été adoptée pour les besoins de la présente Recommandation:

– le terme "mesure objective" désigne la détermination de la qualité ou de la dégradation d'images de type "programme de télévision" présentées à un groupe d'évaluateurs pendant des séances de visionnement;

– le terme "mesure perceptuelle objective" désigne la mesure des performances d'une chaîne de programme par l'emploi d'images de type "programme de télévision" et de méthodes de mesure objective (au moyen d'instruments) pour obtenir une indication approchant la note qui aurait été obtenue au moyen d'une évaluation subjective;

– le terme "mesure de signal" désigne la mesure des performances d'une chaîne de programme par l'emploi de signaux d'essai et de méthodes de mesure objectives (au moyen d'instruments).

Dans la présente Recommandation, les termes "mesure objective" et "mesure perceptuelle" peuvent être utilisés indifféremment pour désigner une mesure perceptuelle objective.

Il existe trois méthodes de base pour réaliser ces mesures:

- FR – Méthode applicable lorsqu'on dispose du signal vidéo de référence complet; c'est une méthode à deux extrémités qui fait l'objet de la présente Recommandation.
- RR – Méthode applicable lorsqu'on ne dispose que d'informations de référence vidéo réduites; c'est aussi une méthode à deux extrémités, avec référence réduite, qui fait l'objet d'une Recommandation distincte (à l'étude).
- NR – Méthode applicable lorsqu'on ne dispose d'aucun signal vidéo de référence ni d'aucune information associée; c'est une méthode à une seule extrémité qui fait l'objet d'une Recommandation distincte (à l'étude).

Les trois méthodes ont des applications différentes et elles offrent des degrés différents de précision de mesure, exprimés en termes de corrélation avec les résultats d'évaluation subjective.

Recommandation UIT-T J.144

Techniques de mesure objective de la qualité vidéo perçue pour la télévision numérique par câble en présence d'un signal de référence complet

1 Domaine d'application

La présente Recommandation contient des lignes directrices relatives au choix d'un équipement approprié de mesure de la qualité vidéo perçue à utiliser dans les applications de télévision numérique par câble lorsqu'on peut utiliser la méthode de mesure avec référence complète.

Cette méthode est destinée à être utilisée lorsque le signal vidéo de référence non dégradé est disponible directement au point de mesure, par exemple en cas de mesures sur un seul équipement ou sur une chaîne en laboratoire ou dans un environnement fermé tel qu'une tête de réseau de télévision par câble. Les méthodes d'estimation sont basées sur le traitement d'une séquence vidéo à composante numérique à 8 bits telle qu'elle est définie dans la Rec. UIT-R BT.601-5[1]. Le codeur peut utiliser diverses méthodes de compression (par exemple MPEG, H.263, etc.). Les modèles proposés dans la présente Recommandation peuvent être utilisés pour évaluer un codec (combinaison codeur/décodeur) ou une concaténation de diverses méthodes de compression et dispositifs d'archivage de mémoire. Le calcul des estimateurs de qualité objectifs décrit dans la présente Recommandation aura peut-être tenu compte des dégradations dues aux erreurs (erreurs sur les bits, perte de paquets) mais on ne dispose pas actuellement de résultats d'essai indépendants permettant de valider l'utilisation des estimateurs pour des systèmes présentant des dégradations dues à des erreurs. Le matériel d'essai de validation ne contenait pas d'erreurs sur les canaux. Il comportait des dégradations de codage, avec des rapports de compression de 768 kbit/s – 5 Mbit/s.

1.1 Application

La présente Recommandation donne des estimations de la qualité vidéo pour différentes classes de télévision (TV0-TV3), et pour la classe vidéo multimédia (MM4) définie dans l'Annexe B/P.911. Les applications des modèles d'estimation décrits dans la présente Recommandation sont notamment les suivantes:

1) évaluation du codec, spécification du codec, essai d'homologation, contenu de la précision limitée décrite ci-dessous;
2) contrôle de la qualité pendant le service, éventuellement en temps réel, à la source;
3) télécontrôle de la qualité au point de destination lorsqu'on dispose d'une copie de la source;
4) mesures de qualité d'un système d'archivage ou de transmission qui utilise des techniques de compression ou de décompression vidéo, par passage unique ou concaténation de telles techniques.

1.2 Limitations

Les modèles d'estimation décrits dans la présente Recommandation ne peuvent être utilisés pour remplacer les essais subjectifs. Les valeurs de corrélation entre deux essais subjectifs conçus et exécutés avec soin (par exemple dans deux laboratoires différents) se situent normalement dans la fourchette 0,92 à 0,97. La présente Recommandation ne donne pas de moyens permettant de

[1] Cela n'exclut pas l'implémentation de la méthode de mesure pour des systèmes vidéo unidirectionnels qui utilisent une entrée vidéo et des sorties vidéo composites. Les spécifications de la conversion entre le domaine composite et le domaine à composante n'entrent pas dans le cadre de la présente Recommandation. Par exemple, la norme SMPTE 170M spécifie une méthode pour effectuer cette conversion dans le cas d'un système NTSC.

quantifier d'éventuelles erreurs d'estimation. Les utilisateurs de la présente Recommandation devraient comparer les résultats des évaluations subjectives et objectives disponibles pour avoir une idée de la fourchette des erreurs d'estimation des indices de qualité vidéo.

Les performances prévues des modèles d'estimation ne sont pas actuellement validées pour des systèmes vidéo comportant des dégradations dues à des erreurs sur les canaux de transmission.

2 Références

La présente Recommandation se réfère à certaines dispositions des Recommandations UIT-T et textes suivants qui, de ce fait, en sont partie intégrante. Les versions indiquées étaient en vigueur au moment de la publication de la présente Recommandation. Toute Recommandation ou tout texte étant sujet à révision, les utilisateurs de la présente Recommandation sont invités à se reporter, si possible, aux versions les plus récentes des références normatives suivantes. La liste des Recommandations de l'UIT-T en vigueur est régulièrement publiée. La référence à un document figurant dans la présente Recommandation ne donne pas à ce document, en tant que tel, le statut d'une Recommandation.

2.1 Références normatives

– Recommandation UIT-R BT.601-5 (1995), *Paramètres de codage en studio de la télévision numérique pour des formats standards d'image 4:3 (normalisé) et 16:9 (écran panoramique).*

2.2 Références informatives

– Recommandation UIT-T J.140 (1998), *Evaluation subjective de la qualité de l'image dans les systèmes de télévision numérique par câble*.

– Recommandation UIT-T J.143 (2000), *Prescriptions d'utilisateur relatives aux mesures objectives de la qualité vidéo perçue en télévision numérique par câble*.

– Recommandation UIT-T P.910 (1996), *Méthodes subjectives d'évaluation de la qualité vidéographique pour les applications multimédias*.

– Document de référence UIT-T (2004), *Objective perceptual assesment of video quality: Full reference television*.

– Recommandation UIT-T P.911 (1998), *Méthodes d'évaluation subjective de la qualité audiovisuelle pour applications multimédias.*

– U. S. Standards Committee T1* Technical Report T1.TR.73-2001, *Video normalization methods applicable to objective video quality metrics utilizing a full reference technique.*

– Recommandation UIT-T BT.500-11 (2002), *Méthode d'évaluation subjective de la qualité des images de télévision*.

3 Termes, définitions et acronymes

La présente Recommandation définit les termes suivants:

3.1 évaluation subjective: détermination de la qualité ou de la dégradation d'images de type "programme de télévision" présentées à un groupe d'évaluateurs pendant des séances de visionnement.

---

* Les normes T1 sont maintenues par l'ATIS depuis novembre 2003.

3.2 **mesure perceptuelle objective**: la mesure des performances d'une chaîne de programme par l'emploi d'images de type "programme de télévision" et de méthodes de mesure objective (au moyen d'instruments) pour obtenir une indication approchant la note qui aurait été obtenue au moyen d'une évaluation subjective.

3.3 **mesure du signal**: le terme "mesure de signal" désigne la mesure des performances d'une chaîne de programme par l'emploi de signaux d'essai et de méthodes de mesure objectives (au moyen d'instruments).

3.4 **ANOVA (*analysis of variance*)**: analyse de variance.

3.5 **FRTV (*full reference television*)**: télévision avec image de référence complète.

3.6 **DSCQS (*double stimulus continuous quality scale*)**: échelle de qualité continue à double stimulus.

3.7 **proposant**: organisation ou entreprise qui propose un modèle de qualité vidéo pour les essais de validation, en vue de son intégration éventuelle à une Recommandation UIT.

4 Besoins de l'utilisateur

Les besoins de l'utilisateur concernant des méthodes de mesure de la qualité vidéo perçue sont formulés dans la Rec. UIT-T J.143.

5 Description de la méthode de mesure avec référence complète

La méthode de mesure aux deux extrémités avec référence complète, servant à mesurer de façon objective la qualité vidéo perçue, permet d'évaluer la performance de systèmes en établissant une comparaison entre le signal vidéo d'entrée non distordu, ou de référence, à l'entrée du système et le signal dégradé à la sortie du système (Figure 1).

La Figure 1 montre un exemple d'application de la méthode avec référence complète pour tester un codec en laboratoire.

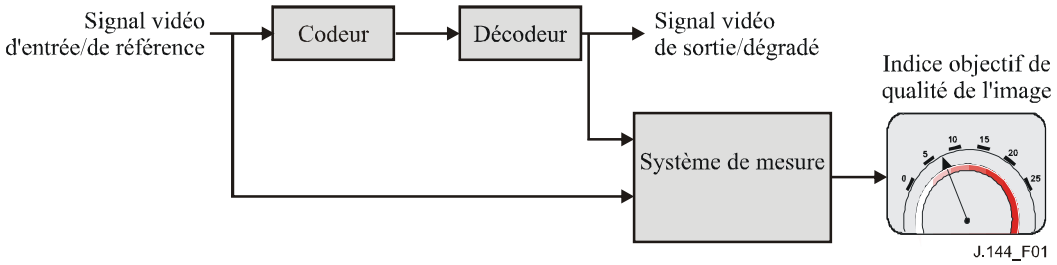


Figure 1 – Application de la méthode de mesure de la qualité perçue avec référence complète pour tester un codec en laboratoire

La comparaison entre le signal d'entrée et le signal de sortie peut nécessiter un processus d'alignement spatial et temporel pour compenser les éventuels déplacements d'image verticaux ou horizontaux ou les éventuels recadrages. Elle peut aussi nécessiter la correction des éventuels décalages et des éventuelles différences de gain dans les canaux de luminance et de chrominance. On calcule alors l'indice objectif de qualité de l'image, généralement en appliquant un modèle de perception de la vision humaine.

La normalisation désigne l'alignement et l'ajustement du gain. Cette opération est nécessaire puisque la plupart des méthodes avec référence complète compare les images traitées et les images de référence effectivement pixel par pixel. Le calcul de la valeur de crête du rapport signal sur bruit (PSNR, *peak signal to noise ratio*) en constitue un exemple. On élimine uniquement les

changements statiques stationnaires de la vidéo, alors que les changements dynamiques dus aux processus de compression et de décompression sont mesurés dans le cadre du calcul de l'indice PQR. Un examen détaillé des raisons justifiant la normalisation figure dans le document intitulé U. S. Standards Committee T1* Rapport technique T1.TR.73-2001, "*Video Normalization Methods Applicable to Objective Video Quality Metrics Utilizing a Full Reference Technique*". Les mesures de la qualité vidéo décrites aux Annexes A à D font état des méthodes de normalisation correspondantes. Le recours à d'autres méthodes de normalisation est possible en ce qui concerne les mesures de la qualité vidéo décrite aux Annexes A à D ainsi qu'aux Appendices I et II, à condition qu'elles offrent la précision de normalisation requise.

Comme l'outil de diagnostic est fondé sur un modèle de la vision humaine et non sur la mesure d'artéfacts de codage particuliers, il est en principe valable aussi bien pour les systèmes analogiques que pour les systèmes numériques. Il est aussi valable en principe pour les chaînes dans lesquelles des systèmes analogiques et des systèmes numériques sont mélangés ou dans lesquelles des systèmes de compression numérique sont concaténés.

La Figure 2 montre un exemple d'application de la méthode avec référence complète pour tester une chaîne de transmission.

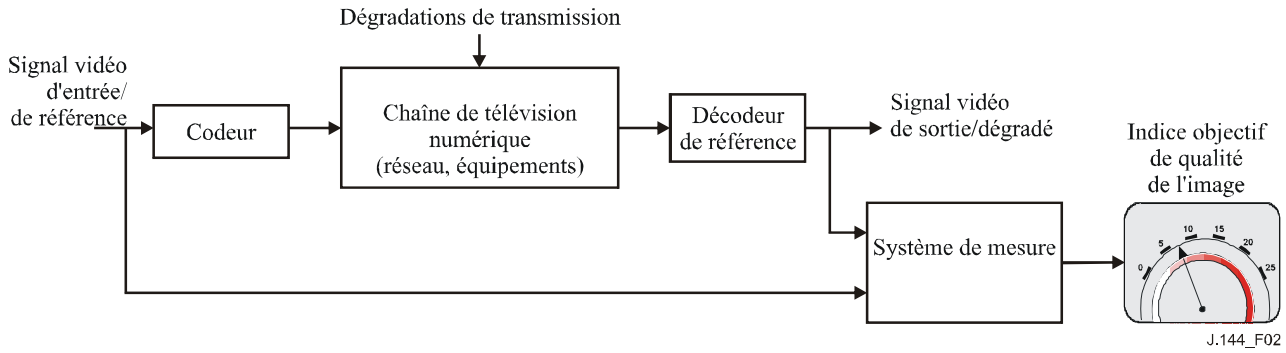


Figure 2 – Application de la méthode de mesure de la qualité perçue avec référence complète pour tester une chaîne de transmission

Dans ce cas, un décodeur de référence est alimenté depuis divers points dans la chaîne de transmission; le décodeur peut par exemple être situé en un point du réseau comme sur la Figure 2, ou directement à la sortie du codeur comme sur la Figure 1. Si la chaîne de transmission numérique est transparente, la mesure de l'indice objectif de qualité de l'image à la source est égale à la mesure en n'importe quel point ultérieur dans la chaîne.

Il est généralement accepté que la méthode avec référence complète offre la meilleure précision en ce qui concerne les mesures de la qualité d'image perçue. La méthode s'est avérée pouvoir offrir une forte corrélation avec les évaluations subjectives faites conformément aux méthodes DSCQS spécifiées dans la Rec. UIT-R BT.500-11.

6 Conclusions du Groupe d'experts sur la qualité vidéo (VQEG)

Un groupe informel, le Groupe d'experts sur la qualité vidéo (VQEG, *video quality expert group*), fait des études sur les mesures de la qualité vidéo perçue et fait rapport à la Commission d'études 9 de l'UIT-T et à la Commission d'études 6 de l'UIT-R. La première tâche du VQEG a consisté à évaluer la performance d'algorithmes proposés de mesure de la qualité vidéo perçue.

* Les normes T1 sont maintenues par l'ATIS depuis novembre 2003.

Le VQEG a publié un projet de rapport final détaillé sur la première phase de ses travaux en mars 2000. Le VQEG a publié en août 2003 un projet final de la phase II de l'essai Télévision avec image de référence complète (FRTV).

Il est conseillé aux lecteurs d'étudier ce rapport pour avoir une vue d'ensemble des travaux du VQEG jusqu'à cette date. Le but était d'évaluer les modèles des proposants en termes de:

- précision des prédictions (capacité du modèle à prédire la qualité subjective);
- monotonicité des prédictions (degré de concordance des prédictions du modèle avec le classement des indices subjectifs de qualité);
- cohérence des prédictions (degré de maintien de la précision des prédictions du modèle sur l'ensemble des séquences de test vidéo et des systèmes vidéo, c'est-à-dire robustesse de la réponse par rapport à une diversité de dégradations vidéo).

La Phase I de l'essai de validation FRTV des mesures VQEG n'a pas fourni suffisamment de données pour identifier une méthode à préconiser en matière d'évaluation objective de la qualité subjective des images vidéo. La Phase II de l'essai de validation FRTV des mesures VQEG a fourni des résultats d'après lesquels quatre méthodes répondent aux conditions requises pour être intégrées à la partie normative de la présente Recommandation.

Au cours de la Phase II, les procédures de la Rec. UIT-R BT.500-11 relatives à la méthode à double stimulus utilisant une échelle de qualité continue (DSCQS, *double stimulus continuous quality scale*) on été suivies à la lettre lors de l'évaluation subjective. Les plans de tests subjectifs et objectifs comprenaient des procédures d'analyse de la validation des notes subjectives et quatre modèles de mesure pour la comparaison de données objectives avec les résultats subjectifs. Parmi les autres analyses statistiques figurait une analyse ANOVA et l'application du test F.

Sur la base des données actuellement disponibles, quatre méthodes peuvent être recommandées actuellement à l'UIT. Il s'agit:

Annexe A – British Telecom (Royaume Uni, VQEG Proposant D);

Annexe B – Yonsei University/SK Telecom/Radio Research Laboratory (République de Corée, VQEG Proposant E);

Annexe C – CPqD (République fédérale du Brésil, VQEG Proposant F);

Annexe D – NTIA (Etats-Unis d'Amérique, VQEG Proposant H).

Les descriptions techniques de ces modèles figurent aux Annexes A à D respectivement. Il convient de signaler que l'ordre des annexes est purement arbitraire et ne revêt aucun caractère indicatif quant aux performances escomptées. Par exemple, dans l'absolu le modèle NTIA s'est caractérisé par la meilleure corrélation avec les indices subjectifs de l'essai 525 lignes.

Les Tableaux 1 et 2 donnent des indications sur les résultats des modèles au terme de l'essai FRTV Phase II du VQEG. Pour les données 525 lignes, les résultats statistiques font apparaître de meilleurs résultats des modèles NTIA et BT par comparaison aux autres, ces deux modèles étant statistiquement équivalents entre eux. En ce qui concerne les données 625 lignes, trois modèles (CPqD, NTIA, Yonsei/SKT/RRL) sont statistiquement équivalents entre eux et d'un point de vue statistique supérieur à l'autre modèle. On notera par ailleurs que seul le modèle NTIA a donné statistiquement des résultats excellents aux deux essais.

Tableau 1 – Récapitulation indicative des résultats des modèles suite au test VQEG Phase II FRTV (données 525 lignes).

Modèle de mesure	BT	Yonsei/ SKT/RRL	CPqD	NTIA	PSNR (Note)
Annexe	A	B	C	D	
Corrélation de Pearson	0,937	0,857	0,835	0,938	0,804
Erreur quadratique	0,075	0,110	0,117	0,074	0,127
NOTE – Les valeurs PSNR mentionnées sont tirées du rapport final des essais Phase II du VQEG. Elles ont été calculées par Yonsei.					

Tableau 2 – Récapitulation indicative des résultats des modèles suite au test VQEG Phase II FRTV (données 625 lignes).

Modèle de mesure	BT	Yonsei/ SKT/RRL	CPqD	NTIA	PSNR
Annexe	A	B	C	D	
Corrélation de Pearson	0,779	0,870	0,898	0,886	0,733
Erreur quadratique	0,113	0,089	0,079	0,083	0,122

7 Conclusions

7.1 Avis général aux fins de la Recommandation

Lorsque l'on effectue des mesures de la qualité vidéo perceptuelle au moyen de la méthode des conditions de référence complètes décrite dans la présente Recommandation, les exploitants devraient tout d'abord analyser comment leurs applications et leurs besoins d'utilisateurs spécifiques se traduisent en termes de caractéristiques et de performances des équipements de mesure.

Il faut notamment tenir compte des aspects suivants:

- coût d'acquisition des équipements de mesure de la qualité perçue;
- service après-vente du vendeur;
- facilité de fonctionnement;
- fiabilité;
- prescriptions de taille, poids, puissance;
- vitesse de mesure en temps réel et pas en temps réel;
- fonctionnement en ligne (en service);
- précision, monotonicité et cohérence des prédictions.

Les quatre méthodes normatives sont recommandées en raison de leur forte corrélation avec les résultats subjectifs des essais FRTV Phase II du VQEG. Toutefois, jusqu'à ce que l'utilisation de ces méthodes soit possible grâce à la disponibilité commerciale des équipements d'essai correspondants, le recours aux méthodes Tektronix/Sarnoff et KDDI est recommandé. Les Appendices I et II spécifient deux modèles de mesure de la qualité vidéo héritée. Ils ont été validés au cours de la Phase I des essais VQEG, et sont inclus ci-après en raison de leur utilisation dans un parc important d'instruments installés de mesure de la qualité vidéo. Les modèles de mesure de la qualité vidéo testés lors des essais de la Phase I du VQEG n'ont pas été jugés suffisamment précis pour être inclus à titre normatif dans une Recommandation de l'UIT. Il convient toutefois de signaler que la Phase I des essais du VQEG s'est déroulée dans un large éventail de conditions expérimentales, souvent avec de très faibles altérations, par comparaison aux essais de la Phase II. Dans ce dernier cas, les

conditions expérimentales comportaient l'utilisation de la transformée discrète en cosinus (par exemple, MPEG-2 et H.263) avec des débits binaires de 768 kbit/s à 5 Mbit/s.

7.2 Modèles de mesure objective de la qualité vidéo – Evolution vers de futures révisions

Pour qu'un modèle devienne normatif, il doit être vérifié par un organe indépendant ouvert (tel que le VQEG) qui se chargera de l'évaluation technique en respectant les directives et critères de performances établies par la Commission d'études 9. Le but de celle-ci est de recommander éventuellement une seule méthode de référence complète normative pour la télévision par câble.

Annexe A

British Telecommunications plc

Description fonctionnelle du modèle de qualité vidéo avec une image de référence complète

A.1 Introduction

L'outil d'évaluation automatique de la qualité vidéo avec une image de référence complète de BT (BTFR, *BT full-reference*) permet d'avoir des prévisions de la qualité vidéo qui sont représentatives des jugements de qualité de l'être humain. Cet outil de mesure objective simule numériquement les caractéristiques du système visuel humain (HVS, *human visual system*) pour donner des prévisions précises de la qualité vidéo et constitue une alternative viable aux évaluations subjectives classiques qui sont coûteuses et chronophages.

Une implémentation logicielle du modèle a été intégrée dans les tests VQEG2 et les résultats correspondants ont été présentés dans un rapport sur les essais [A-1].

A.2 Modèle d'image de référence complète de BT

L'algorithme BTFR effectue une détection suivie d'une intégration (voir Figure A.1). Par détection, on entend le calcul d'un ensemble de paramètres du détecteur perceptuellement significatifs à partir de la séquence vidéo non déformée (de référence) et de la séquence vidéo déformée (dégradée). Ces paramètres constituent alors les données d'entrée pour l'intégrateur qui donne une estimation de la qualité vidéo perçue avec une pondération appropriée. Le choix des détecteurs et des facteurs de pondération est fonction de caractéristiques de masquage spatial et temporel connues du système visuel humain et déterminé par étalonnage.

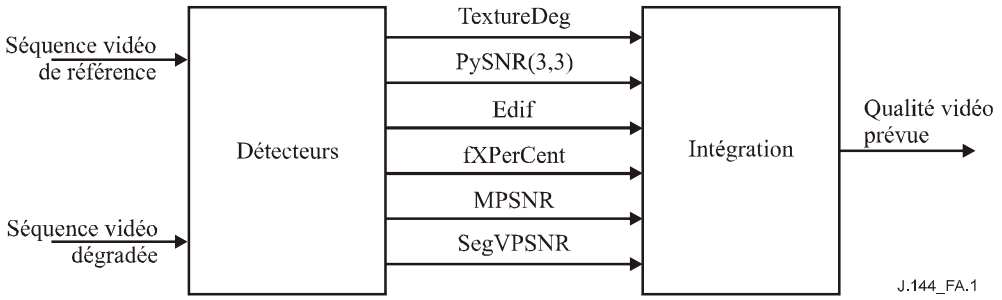


Figure A.1 – Modèle d'évaluation de la qualité vidéo avec une image de référence complète

Le modèle accepte des séquences vidéo d'entrée de type 625 (720 × 576) entrelacées à 50 trames/s et 525 (720 × 486) entrelacées à 59,94 trames/s en format YUV422.

A.3 Détecteurs

Le module de détection de l'algorithme BTFR effectue un certain nombre de mesures fréquentielles dans le domaine temporel et le domaine spatial à partir des séquences d'entrée formatées YUV (voir la Figure A.2).

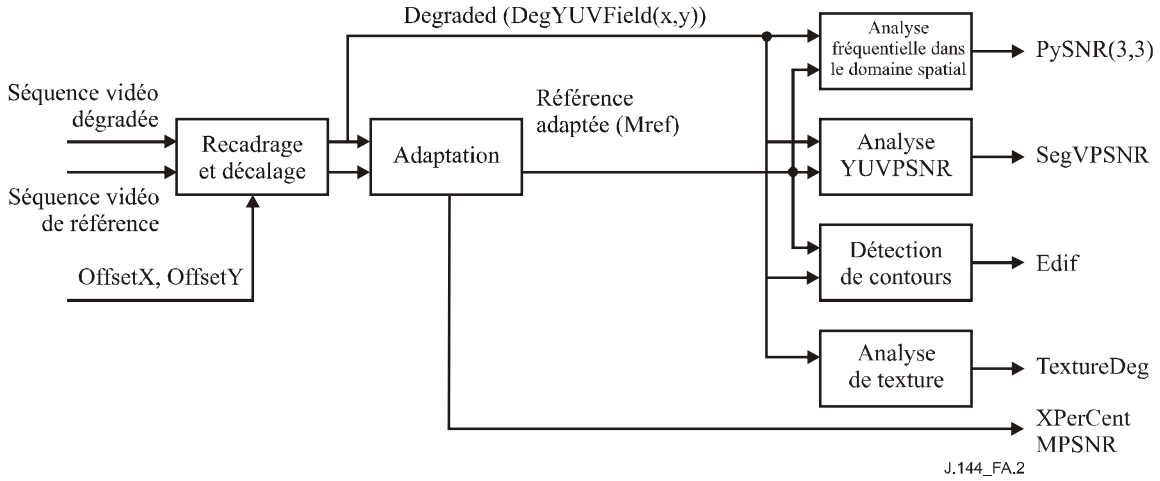


Figure A.2 – Détection

A.3.1 Conversion des séquences d'entrée

Tout d'abord, les séquences d'entrée sont converties du format entrelacé YUV422 à un format désentrelacé de bloc YUV444 de sorte que chaque trame successive est représentée par des tableaux Ref*Y*, Ref*U* et Ref*V*:

$$\text{Re}\, fY(x,y) \quad x = 0..X-1, \quad y = 0..Y-1 \tag{A.3.1-1}$$

$$\text{Re}\, fU(x,y) \quad x = 0..X-1, \quad y = 0..Y-1 \tag{A.3.1-2}$$

$$\text{Re}\, fV(x,y) \quad x = 0..X-1, \quad y = 0..Y-1 \tag{A.3.1-3}$$

où X est le nombre de pixels horizontaux dans une trame et Y le nombre de pixels verticaux. Pour une séquence d'entrée YUV422, chaque valeur de U et chaque valeur de V doivent être répétées pour obtenir les équations A.3.1-2 et A.3.1-3 avec une résolution complète.

A.3.2 Recadrage et décalage

Cette routine recadre, avec décalage, la séquence d'entrée dégradée et recadre, sans décalage, la séquence d'entrée de référence. Les paramètres de décalage OffsetX et OffsetY, déterminés extérieurement, définissent de combien de pixels horizontaux et verticaux la séquence est décalée par rapport à la séquence de référence. L'origine de l'image est située dans le coin supérieur gauche, avec un déplacement positif horizontal vers la droite et vertical vers le bas. Une valeur du paramètre XOffset de 2 indique que les trames dégradées sont décalées vers la droite de 2 pixels et une valeur du paramètre YOffset de 2 indique un décalage vers le bas de 2 pixels. Pour une trame d'entrée avec des valeurs YUV archivées en format YUV444 (voir § A.3.1) dans des tableaux InYField, InUField et InVField, la séquence de sortie recadrée et décalée est calculée selon les équations A.3.2-1 à A.3.2-17.

$$XStart = -OffsetX \tag{A.3.2-1}$$

$$si\ (XStart < C_X)\ alors \quad XStart = C_X \tag{A.3.2-2}$$

$$XEnd = X - 1 - OffsetX \tag{A.3.2-3}$$

$$si\ (XEnd > X - C_X - 1)\ alors \quad XEnd = X - C_X - 1 \tag{A.3.2-4}$$

$$YStart = -OffsetY \tag{A.3.2-5}$$

$$si\ (YStart < C_Y)\ alors \quad XStart = C_Y \tag{A.3.2-6}$$

$$YEnd = Y - 1 - OffsetY \quad \text{(A.3.2-7)}$$

$$si\ (YEnd > Y - C_y - 1)\ \ alors\ YEnd = Y - C_y - 1 \quad \text{(A.3.2-8)}$$

X et *Y* donnent respectivement la dimension de trame horizontale et la dimension de trame verticale et C_x et C_y le nombre de pixels à recadrer depuis la gauche et la droite ainsi que le haut et le bas.

Pour des séquences à 625 lignes,

$$X = 720, \qquad Y = 288, \qquad C_x = 30, \qquad C_y = 10 \quad \text{(A.3.2-9)}$$

Pour des séquences à 525 lignes,

$$X = 720, \qquad Y = 243, \qquad C_x = 30, \qquad C_y = 10 \quad \text{(A.3.2-10)}$$

Xstart, Xend, Ystart et Yend définissent maintenant la région de chaque trame qui sera copiée. Les pixels situés en dehors de cette région sont initialisés selon les équations A.3.2-11 et A.3.2-12, dans lesquelles YField, UField et VField sont les tableaux de pixels de sortie XxY contenant respectivement les valeurs *Y*, *U* et *V*.

Les barres verticales à gauche et à droite de la trame sont initialisées selon:

$$YField(x, y) = 0 \quad x = 0..XStart - 1, XEnd + 1..X - 1 \quad y = 0..Y - 1 \quad \text{(A.3.2-11)}$$

$$UField(x, y) = VField(x, y) = 128 \quad x = 0..XStart - 1, XEnd + 1..X - 1 \quad y = 0..Y - 1 \quad \text{(A.3.2-12)}$$

Les barres horizontales en haut et en bas de la trame sont initialisées selon:

$$YField\ (x, y) = 0 \quad x = XStart\ ..XEnd\ , \quad y = 0..YStart\ - 1, YEnd\ + 1..Y - 1 \quad \text{(A.3.2-13)}$$

$$UField(x, y) = VField(x, y) = 128 \quad x = XStart..XEnd \quad y = 0..YStart - 1, YEnd + 1..Y - 1 \quad \text{(A.3.2-14)}$$

Enfin, les valeurs des pixels sont copiées selon:

$$YField(x, y) = InYField(x + OffsetX, y + OffsetY) \quad x = XStart..XEnd \quad y = YStart..YEnd \quad \text{(A.3.2-15)}$$

$$UField(x, y) = InUField(x + OffsetX, y + OffsetY) \quad x = XStart..XEnd \quad y = YStart..YEnd \quad \text{(A.3.2-16)}$$

$$VField(x, y) = InVField(x + XOffset, y + YOffset) \quad x = XStart..XEnd \quad y = YStart..YEnd \quad \text{(A.3.2-17)}$$

Pour la séquence d'entrée dégradée, le recadrage et le décalage génèrent des tableaux de trame de sortie DegYField, DegUField et DegVField tandis que le recadrage sans décalage pour la séquence de référence génère RefYField, RefUField et RefVfield. Ces tableaux bidimensionnels XxY servent de données d'entrée pour les routines de détection décrites ci-après.

A.3.3 Adaptation

Le processus d'adaptation produit des signaux destinés à être utilisés dans d'autres procédures de détection, ainsi que des paramètres de détection destinés à être utilisés dans la procédure d'intégration. Pour les signaux d'adaptation, on cherche, pour de petits blocs dans chaque trame dégradée, dans une mémoire tampon de trames de référence voisines les trames qui correspondent le mieux. On obtient ainsi une séquence, la séquence de référence adaptée, destinée à être utilisée en lieu et place de la séquence de référence dans certains des modules de détection.

L'analyse d'adaptation est réalisée sur des blocs de pixels 9×9 des tableaux d'intensité RefYField et DegYField. Si l'on ajoute la dimension nombre de trames aux tableaux d'intensité, le pixel (*Px*, *Py*) de la trame de référence *N* peut être représenté comme suit:

$$\text{Re}\ f\ (N, Px, Py) = \text{Re}\ fYField(Px, Py)\ \ \text{trame}\ \ N \quad \text{(A.3.3-1)}$$

Un bloc de pixels 9 × 9 avec un pixel central (*Px*, *Py*) dans la *N*ième trame peut être représenté comme suit:

$$Block\ \mathrm{Re}\ f(N, Px, Py) = \mathrm{Re}\ f(n, x, y) \quad x = Px - 4..Px + 4, \quad y = Py - 4..Py + 4 \qquad \text{(A.3.3-2)}$$

Deg(*n, x, y*) et BlockDeg(*n, x, y*) peuvent être définis de la même manière.

Pour BlockDeg(*N, Px, Py*), on calcule une erreur d'adaptation minimale *E*(*N, Px, Py*) en cherchant les trames de référence voisines selon l'équation:

$$E(N, Px, Py) = MIN((1/81)\sum_{j=-4}^{4}\sum_{k=-4}^{4}(Deg(N, Px + j, Py + k) - \mathrm{Re}\ f(n, x + j, y + k))^2) \qquad \text{(A.3.3-3)}$$

avec

$$n = N - 4, ..., N + 5$$
$$x = Px - 4, Px, ..., Px + 4$$
$$y = Py - 4, Py, ..., Py + 4$$

où *N* est l'indice de la trame dégradée contenant le bloc dégradé qui fait l'objet de l'adaptation.

Si l'équation A.3.3-3 permet de déterminer que la meilleure correspondance avec BlockDeg(*N, Px, Py*) est BlockRef(n_m, x_m, y_m), alors un tableau de référence adaptée MRef est mis à jour selon:

$$M\ \mathrm{Re}\ f(N, Px + j, Py + k) = \mathrm{Re}\ f(n_m, x_m + j, y_m + k) \quad j = -4...4, k = -4...4 \qquad \text{(A.3.3-4)}$$

Le processus d'adaptation de recherche de la meilleure correspondance pour un bloc dégradé suivie de la copie du bloc résultant dans le tableau de référence adapté est répété pour l'ensemble de la zone d'analyse souhaitée. Cette zone d'analyse est définie par les points centraux de blocs *Px*() et *Py*() selon:

$$Px(h) = 16 + 8 \times h \quad h = 0..Qx - 1 \qquad \text{(A.3.3-5)}$$

et

$$Py(v) = 16 + 8 \times v \quad v = 0..Qy - 1 \qquad \text{(A.3.3-6)}$$

où *Qx* et *Qy* définissent le nombre de blocs d'analyse horizontaux et verticaux. Puisque le processus d'adaptation repose sur des blocs de pixels 9 × 9, les blocs d'adaptation voisins se chevauchent d'un pixel. Mref est mis à jour selon les équations A.3.3-4, A.3.3-5 et A.3.3-6, de telle sorte que les régions qui se chevauchent dans Mref sont remplacées par les résultats des calculs suivants.

L'analyse d'adaptation de la *N*ième trame produit donc une séquence de référence adaptée décrite par:

$$BlockM\ \mathrm{Re}\ f(N, Px(h), Py(v)) \qquad h = 0..Qx - 1, \qquad v = 0..Qy - 1 \qquad \text{(A.3.3-7)}$$

et un ensemble de valeurs d'erreur pour la meilleure correspondance:

$$E(N, Px(h), Py(v)) \qquad h = 0..Qx - 1, \qquad v = 0..Qy - 1 \qquad \text{(A.3.3-8)}$$

Un ensemble de tableaux de décalage *MatT*, *MatX* et *MatY* peuvent être définis de façon à:

$$BlockM\ \mathrm{Re}\ f(N, Px(h), Py(v))) = Block\ \mathrm{Re}\ f(MatT(h, v), MatX(h, v), MatY(h, v)) \qquad h = 0..Qx - 1, \qquad v = 0..Qy - 1 \qquad \text{(A.3.3-9)}$$

Les paramètres d'adaptation pour des séquences de radiodiffusion à 625 lignes ou 525 lignes sont donnés dans le Tableau A.1.

Tableau A.1 – Paramètres de recherche pour la procédure d'adaptation

Paramètre	625	525
Qx	87	87
Qy	33	28

La zone d'analyse définie par les équations A.3.3-6 et A.3.3-7 ne couvre pas l'ensemble de la trame. *M*Ref doit donc être initialisé selon l'équation A.3.3-9 de façon à pouvoir être utilisé ailleurs sans restriction.

$$M \text{ Re } f(x,y) = 0 \qquad x = 0..X-1, \qquad y = 0..Y-1 \tag{A.3.3-10}$$

A.3.3.1 Statistiques d'adaptation

Les statistiques d'adaptation horizontales sont calculées à partir du processus d'adaptation et destinées à être utilisées dans le processus d'intégration. La meilleure correspondance pour chaque bloc d'analyse, déterminée selon l'équation A.3.3-3, est utilisée dans la construction de l'histogramme *histX* pour chaque trame selon:

$$\begin{aligned} histX\left(MatX(h,v) - Px(h) + 4\right) = histX\left(MatX(h,v) - Px(h) + 4\right) + 1 \\ h = 0..Qx-1, \quad v = 0..Qy-1 \end{aligned} \tag{A.3.3.1-1}$$

où le tableau histX est initialisé à zéro pour chaque trame. L'histogramme est ensuite utilisé pour déterminer la mesure fXPerCent selon:

$$fXPerCent = 100 \times Max(histX(i)) / \sum_{j=0}^{8} histX(j) \qquad i = 0..8 \tag{A.3.3.1-2}$$

Pour chaque trame, la mesure *fXPerCent* donne la proportion (%) de blocs adaptés qui interviennent dans la crête de l'histogramme d'adaptation.

A.3.3.2 Rapport PSNR adapté

L'erreur minimale, *E*(), pour chaque bloc adapté est utilisé pour calculer un signal adapté au rapport de bruit selon:

$$\begin{aligned} si\left(\sum_{h=0}^{Qx-1}\sum_{v=0}^{Qy-1} E(N, Px(h), Py(v))\right) > 0 \qquad alors \\ MPSNR = 10\log_{10}\left(Qx \times Qy \times 255^2 / \sum_{h=0}^{Qx-1}\sum_{v=0}^{Qy-1} E(N, Px(h), Py(v))\right) \end{aligned} \tag{A.3.3.2-1}$$

$$si\left(\sum_{h=0}^{Qx-1}\sum_{v=0}^{Qy-1} E(N, Px(h), \ Py(v))\right) = 0 \quad alors \quad MPSNR = 10\log_{10}(255^2) \tag{A.3.3.2-2}$$

A.3.3.3 Vecteurs d'adaptation

Le vecteur horizontal, le vecteur vertical et le vecteur retard sont archivés en vue d'une utilisation future selon:

$$SyncT(h,v) = MatT(h,v) - N \qquad h = 0..Qx-1, \qquad v = 0..Qy-1 \tag{A.3.3.3-1}$$

$$SyncX(h,v) = MatX(h,v) - Px(h) \qquad h = 0..Qx-1, \qquad v = 0..Qy-1 \tag{A.3.3.3-2}$$

$$SyncY(h,v) = MatY(h,v) - Py(h) \qquad h = 0..Qx-1, \qquad v = 0..Qy-1 \tag{A.3.3.3-3}$$

A.3.4 Analyse fréquentielle dans le domaine spatial

Le détecteur fréquentiel dans le domaine spatial est basé sur une transformation "pyramidale" des séquences de référence dégradée et adaptée (voir Figure A.3). Tout d'abord, chaque séquence est transformée pour donner un tableau pyramidal de référence et un tableau pyramidal dégradé. Ensuite, les différences entre les tableaux pyramidaux sont calculées à l'aide d'une mesure de l'erreur quadratique moyenne et les résultats en sortie sont présentés sous forme d'un rapport signal/bruit pyramidal.

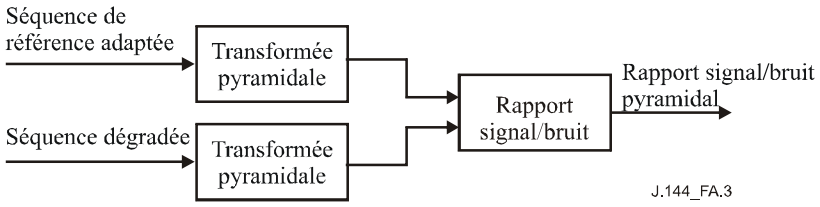


Figure A.3 – Analyse fréquentielle dans le domaine spatial

A.3.4.1 Transformée pyramidale

Tout d'abord, la trame d'entrée *F* est copiée dans un tableau pyramidal *P* selon:

$$P(x,y) = F(x,y) \qquad x = 0..X-1, \quad y = 0..Y-1 \qquad \text{(A.3.4.1-1)}$$

Ce tableau pyramidal est ensuite mis à jour par analyse horizontale et verticale en trois étapes (étape=0..2). L'analyse horizontale Hpy(stage) est définie par les équations A.3.4.1-2 à A.3.4.1-6.

Tout d'abord, il est fait une copie temporaire de l'ensemble du tableau pyramidal:

$$PTemp(x,y) = P(x,y) \qquad x = 0..X-1, \quad y = 0..Y-1 \qquad \text{(A.3.4.1-2)}$$

Ensuite les limites *x* et *y* sont calculées selon:

$$Tx = X / 2^{(stage+1)} \qquad \text{(A.3.4.1-3)}$$

$$Ty = Y / 2^{stage} \qquad \text{(A.3.4.1-4)}$$

Les moyennes et les différences des paires horizontales d'éléments du tableau temporaire sont ensuite utilisées pour mettre à jour le tableau pyramidal selon:

$$P(x,y) = 0.5 \times (PTemp(2x,y) + PTemp(2x+1,y)) \quad x = 0..Tx-1 \quad y = 0..Ty-1 \qquad \text{(A.3.4.1-5)}$$

$$P(x+Tx,y) = PTemp(2x,y) - PTemp(2x+1,y) \quad x = 0..Tx-1 \quad y = 0..Ty-1 \qquad \text{(A.3.4.1-6)}$$

L'analyse verticale Vpy(stage) est définie par les équations A.3.4.1-7 à A.3.4.1-11.

$$PTemp(x,y) = P(x,y) \qquad x = 0..X-1, \quad y = 0..Y-1 \qquad \text{(A.3.4.1-7)}$$

$$Tx = X / 2^{stage} \qquad \text{(A.3.4.1-8)}$$

$$Ty = Y / 2^{(stage+1)} \qquad \text{(A.3.4.1-9)}$$

Les moyennes et les différences des paires verticales d'éléments du tableau temporaire sont ensuite utilisées pour mettre à jour le tableau pyramidal selon:

$$P(x,y) = 0.5 \times (PTemp(x,2y) + PTemp(x,2y+1)) \quad x = 0..Tx-1, \quad y = 0..Ty-1 \qquad \text{(A.3.4.1-10)}$$

$$P(x,y+Ty) = PTemp(x,2y) - PTemp(x,2y+1) \quad x = 0..Tx-1 \quad y = 0..Ty-1 \qquad \text{(A.3.4.1-11)}$$

Pour l'étape 0, l'analyse horizontale Hpy(0) suivie de l'analyse verticale Vpy(0) met à jour l'ensemble du tableau pyramidal avec les quatre quadrants Q(étape, 0..3) structurés comme suit (Figure A.4):

Q(0,0)	Q(0,1)
Q(0,2)	Q(0,3)

Q(0,0) = moyenne de blocs de 4
Q(0,1) = différence horizontale de blocs de 4
Q(0,2) = différence verticale de blocs de 4
Q(0,3) = différence diagonale de blocs de 4

Figure A.4 – Représentation en quadrants de la sortie de l'analyse, étape 0

L'analyse étape 1 est ensuite réalisée sur Q(0,0) pour obtenir les résultats Q(1,0..3) qui sont archivés dans la pyramide selon la Figure A.5:

<table>
<tr><td>Q(1,0)</td><td>Q(1,1)</td><td rowspan="2">Q(0,1)</td></tr>
<tr><td>Q(1,2)</td><td>Q(1,3)</td></tr>
<tr><td colspan="2">Q(0,2)</td><td>Q(0,3)</td></tr>
</table>

Figure A.5 – Représentation en quadrants de la sortie de l'analyse, étape 1

L'analyse, étape 2 traite Q(1,0) et le remplace par Q(2,0..3).

A l'issue des trois stades de l'analyse, le tableau pyramidal résultant comporte un total de 10 blocs de résultats. Trois blocs Q(0,1..3) proviennent de l'analyse des pixels 2×2, étape 0, trois Q(1,1..3) de l'analyse des pixels 4×4, étape 1 et 4 Q(2,0..3) de l'analyse des pixels 8×8, étape 2.

L'analyse en trois étapes de la séquence de référence adaptée et de la séquence dégradée produit les tableaux pyramidaux Pref et Pdeg. Les différences entre ces tableaux sont ensuite mesurées dans le module SNR pyramidal.

A.3.4.2 Rapport SNR pyramidal

On mesure l'erreur quadratique entre le tableau pyramidal de référence et le tableau pyramidal dégradé sur les quadrants 1 à 3 des étapes 0 à 2 selon:

$$E(s,q) = (1/(XY)^2) \sum_{x=x1(s,q)}^{x2(s,q)-1} \sum_{y=y1(s,q)}^{y2(s,q)-1} (\mathrm{Pr}ef(x,y) - P\deg(x,y))^2 \quad s = 0..2 \quad q = 1..3 \qquad \text{(A.3.4.2-1)}$$

où, $x1$, $x2$, $y1$ et $y2$ définissent les limites horizontales et verticales des quadrants dans les tableaux pyramidaux et sont calculés selon:

$$x1(s,1) = X / 2^{(s+1)} \quad x2(s,1) = 2 \times x1(s,1) \quad y1(s,1) = 0 \quad y2(s,1) = Y / 2^{(s+1)} \qquad \text{(A.3.4.2-2)}$$

$$x1(s,2) = 0 \quad x2(s,2) = X / 2^{(s+1)} \quad y1(s,2) = Y / 2^{(s+1)} \quad y2(s,2) = 2 \times y1(s,2) \qquad \text{(A.3.4.2-3)}$$

$$x1(s,3) = X / 2^{(s+1)} \quad x2(s,3) = 2 \times x1(s,3) \quad y1(s,3) = Y / 2^{(s+1)} \quad y2(s,3) = 2 \times y1(s,3) \qquad \text{(A.3.4.2-4)}$$

Les résultats de l'équation (A.3.4.2-1) sont ensuite utilisés pour mesurer le rapport PSNR pyramidal pour chaque quadrant de chaque trame selon:

$$\begin{aligned} &si\ (E > 0.0) \quad PySNR(s,q) = 10.0 \times \log_{10}(255^2 / E(s,q)) \\ &sinon \quad SNR = 10.0 \times \log_{10}(255^2 \times (XY)^2) \end{aligned} \qquad \text{(A.3.4.2-5)}$$

où le nombre d'étapes s=0..2 et le nombre de cadrans pour chaque étape q=1..3.

A.3.5 Analyse de la texture

On mesure la texture de la séquence dégradée en enregistrant le nombre de points de transition du signal d'intensité sur les lignes horizontales de l'image, selon les équations A.3.5-1 à A.3.5-6.

Pour chaque trame, un compteur de points de transition est tout d'abord initialisé selon l'équation A.3.5-1.

$$sum = 0 \tag{A.3.5-1}$$

Puis, chaque ligne y=0..Y-1, est traitée pour x=0..X-2 selon:

$$last\_pos = 0, \qquad last\_neg = 0 \tag{A.3.5-2}$$

$$dif(x) = P(x, y) - P(x+1, y) \tag{A.3.5-3}$$

$$si\,((dif(x) < 0)\,AND\,(last\_neg < last\_pos))\;alors\;sum = sum + 1 \tag{A.3.5-4}$$

$$si\,((dif(x) > 0)\,AND\,(last\_neg > last\_pos))\;alors\;sum = sum + 1 \tag{A.3.5-5}$$

$$si\,(dif(x) > 0)\;alors\;last\_pos = x \tag{A.3.5-6}$$

$$si\,(dif(x) < 0)\;alors\;last\_neg = x \tag{A.3.5-7}$$

Quand toutes les lignes d'une trame ont été traitées, le compteur, *sum*, contiendra le nombre de points de transition du signal d'intensité horizontal. Ce nombre est ensuite utilisé pour calculer un paramètre de texture pour chaque trame selon:

$$TextureDeg = sum \times 100 / XY \tag{A.3.5-8}$$

A.3.6 Analyse des contours

Chaque trame de la séquence dégradée et de la séquence de référence adaptée subit séparément une routine de détection des bords pour produire des représentations correspondantes des bords de la trame, lesquelles sont ensuite comparées dans une procédure d'adaptation de blocs pour établir les paramètres de détection (voir Figure A.6).

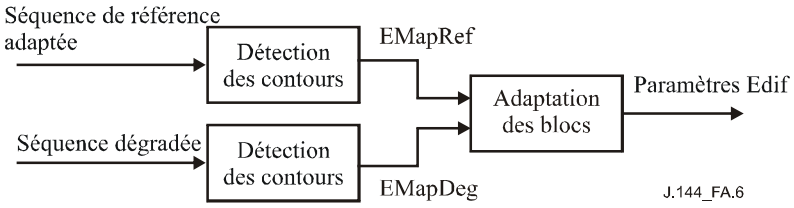


Figure A.6 – Analyse des contours

A.3.6.1 Détection des contours

On a utilisé un détecteur de contours de Canny [A-2] pour déterminer les représentations des contours, mais d'autres techniques analogues de détection de contours bords peuvent également être utilisées. Les représentations des contours résultantes, *EMapRef* et *EMapDeg*, sont des représentations de pixels où un contour est indiqué par un 1 et l'absence de contour par un 0,

Pour la détection d'un contour ou pixel (*x, y*):

$$EMap(x, y) = 1 \qquad x = 0..X-1, \qquad y = 0..Y-1 \tag{A.3.6.1-1}$$

Pour la détection d'une absence de contour ou pixel (*x, y*):

$$EMap(x, y) = 0 \qquad x = 0..X-1, \qquad y = 0..Y-1 \tag{A.3.6.1-2}$$

A.3.6.2 Différenciation des contours

La procédure de différenciation des contours permet de mesurer les différences entre les représentations des contours pour la trame dégradée et la trame de référence adaptée correspondante. L'analyse est effectuée dans *NxM* blocs de pixels ne se chevauchant pas selon les équations A.3.6.2-1 à A.3.6.2-5.

Tout d'abord, on calcule le nombre de pixels marqués par un bord dans chaque bloc d'analyse où *Bh* et *Bv* définissent le nombre de blocs ne se chevauchant pas à analyser dans les directions horizontale et verticale et *X*1 et *Y*1 définissent les décalages par rapport au bord de la trame.

$$bref(x,y) = \sum_{i=i1}^{i2} \sum_{j=j1}^{j2} EMap \,\mathrm{Re}\, f(Nx + X1 + i, My + Y1 + j) \quad x = 0..Bh-1, y = 0..Bv-1 \qquad \text{(A.3.6.2-1)}$$

$$BDeg(x,y) = \sum_{i=i1}^{i2} \sum_{j=j1}^{j2} EMapDeg(Nx + X1 + i, My + Y1 + j) \quad x = 0..Bh-1, y = 0..Bv-1 \qquad \text{(A.3.6.2-2)}$$

Les limites de sommation sont déterminées selon:

$$i1 = -(N \quad div \quad 2) \qquad i2 = (N-1) \quad div \quad 2 \qquad \text{(A.3.6.2-3)}$$

$$j1 = -(M \quad div \quad 2) \qquad j2 = (M-1) \quad div \quad 2 \qquad \text{(A.3.6.2-4)}$$

où l'opérateur "div" représente une division par un nombre entier.

Ensuite, on effectue une mesure des différences sur la totalité de la trame selon:

$$EDif = (1/(N \times M \times Bh \times Bv)) \times (\sum_{x=0}^{Bh-1} \sum_{y=0}^{Bv-1} (B \,\mathrm{Re}\, f(x,y) - BDeg(x,y))^{Q})^{1/Q} \qquad \text{(A.3.6.2-5)}$$

Pour des trames de pixel 720×288 pour une séquence vidéo à 625 lignes:

$$N = 4, \quad X1 = 6, \quad Bh = 178, \qquad M = 4, \quad Y1 = 10, \quad Bv = 69 \quad Q = 3 \qquad \text{(A.3.6.2-6)}$$

Pour des trames de pixel 720×243 pour une séquence vidéo à 525 lignes:

$$N = 4, \quad X1 = 6, \quad Bh = 178, \qquad M = 4, \quad Y1 = 10, \quad Bv = 58, \quad Q = 3 \qquad \text{(A.3.6.2-7)}$$

A.3.7 Analyse du rapport PSNR adapté

Un rapport signal/bruit adapté est calculé pour les valeurs du pixel *V* en utilisant les vecteurs d'adaptation définis dans les équations A.3.3.3-1 à A.3.3.3-3. Pour chaque ensemble de vecteurs d'adaptation une mesure d'erreur, *VE*, est calculée selon:

$$VE(h,v) = (1/81) \sum_{i=-4}^{4} \sum_{j=-4}^{4} (DegV(N, Px(h) + i, Py(h) + j) - \mathrm{Re}\, fVField(N + SyncT(h,v), Px(h) + SyncX(h,v) + i, Py(v) + SyncY(h,v) + j))^2 \qquad \text{(A.3.7-1)}$$

On calcule alors une mesure du rapport PSNR segmentaire pour la trame selon:

$$SegVPSNR = (1/Qx \times Qy) \sum_{h=0}^{Qx-1} \sum_{v=0}^{Qy-1} 10.0 \times \log_{10}(255^2/(VE(h,v)+1)) \qquad \text{(A.3.7-2)}$$

A.4 Intégration

La procédure d'intégration nécessite tout d'abord une pondération temporelle des paramètres de détection trame par trame selon l'équation A.4-1:

$$AvD(k) = (1/N) \times \sum_{n=0}^{N-1} D(k,n) \qquad k = 0..5 \tag{A.4-1}$$

où N est le nombre total de trames des séquences testées et $D(k, n)$ est le paramètre de détection k pour la trame n.

Les paramètres de détection pondérés, $AvD(k)$, sont ensuite combinés pour donner une note de qualité prévue PDMOS, pour la séquence de trame N selon:

$$PDMOS = Offset + \sum_{k=0}^{5} AvD(k) \times W(k) \tag{A.4-2}$$

Les Tableaux A.2 et A.3 donnent les paramètres de l'intégrateur respectivement pour les séquences à 625 lignes et celles à 525 lignes.

Tableau A.2 – Paramètres d'intégration pour un système de vidéodiffusion à 625 lignes

K	Nom du paramètre	W
0	TextureDeg	–0,68
1	PySNR(3,3)	–0,57
2	EDif	+58913,294
3	fXPerCent	–0,208
4	MPSNR	–0,928
5	SegVPSNR	–1,529
Décalage	+176,486	
N	400	

Tableau A.3 – Paramètres d'intégration pour un système de vidéodiffusion à 525 lignes

K	Nom du paramètre	W
0	TextureDeg	+0,043
1	PySNR(3,3)	–2,118
2	EDif	+60865,164
3	fXPerCent	–0,361
4	MPSNR	+1,104
5	SegVPSNR	–1,264
Décalage	+260,773	
N	480	

A.5 Alignement

Le modèle FR nécessite un bon fonctionnement de l'alignement spatial et temporel. Le modèle intègre un alignement inhérent et peut prendre en charge des décalages spatiaux entre la séquence

de référence et la séquence dégradée de ±4 pixels et des décalages temporels de ±4 trames. Les décalages spatiaux ou temporels au-delà de ces limites ne sont pas pris en charge par le modèle et il faudra un module d'alignement distinct pour s'assurer que la séquence de référence et la séquence dégradée sont correctement alignées.

A.6 Références

[A-1] Document de référence UIT-T (2004), *Objective perceptual assessment of video quality: Full reference television.*

[A-2] J. CANNY: A computational approach to edge detection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8(6): pp. 679-698, 1986.

A.7 Données objectives et subjectives

NOTE – Les fichiers vidéo qui suivent ont été traités de manière à donner le résultat indiqué. Les noms de fichiers sont donnés à titre d'information uniquement. Les fichiers ne sont pas disponibles à l'échelle publique. Pour plus d'informations, voir document de référence UIT-T.

Données objectives et subjectives pour un système à 525 lignes.

Nom du fichier	SRC	HRC	Note subjective moyenne brute	Note prévue par le modèle sur la base des données brutes	Note subjective moyenne corrigée	Note prévue par le modèle sur la base de données corrigées
V2src01_hrc01_525.yuv	1	1	–38,30757576	44,945049	0,5402368	0,69526
V2src01_hrc02_525.yuv	1	2	–39,56212121	38,646271	0,5483205	0,58989
V2src01_hrc03_525.yuv	1	3	–25,9469697	32,855755	0,4024097	0,50419
V2src01_hrc04_525.yuv	1	4	–17,24090909	21,062775	0,3063528	0,36089
V2src02_hrc01_525.yuv	2	1	–35,23636364	31,260744	0,5025558	0,48242
V2src02_hrc02_525.yuv	2	2	–18,01818182	18,732758	0,3113346	0,33715
V2src02_hrc03_525.yuv	2	3	–6,284848485	8,914509	0,1881739	0,25161
V2src02_hrc04_525.yuv	2	4	–6,983333333	4,16663	0,1907347	0,21776
V2src03_hrc01_525.yuv	3	1	–31,96515152	22,348713	0,4682724	0,37461
V2src03_hrc02_525.yuv	3	2	–17,47727273	10,44728	0,3088831	0,26352
V2src03_hrc03_525.yuv	3	3	–1,104545455	2,494911	0,1300389	0,20688
V2src03_hrc04_525.yuv	3	4	–1,171212121	0	0,1293293	0,19158
V2src04_hrc05_525.yuv	4	5	–50,64090909	40,82526	0,6742005	0,6249
V2src04_hrc06_525.yuv	4	6	–28,05454545	32,552322	0,4250873	0,49999
V2src04_hrc07_525.yuv	4	7	–23,87575758	25,286598	0,3762656	0,40764
V2src04_hrc08_525.yuv	4	8	–16,60757576	19,86405	0,2972294	0,3485
V2src05_hrc05_525.yuv	5	5	–31,86969697	30,812616	0,4682559	0,47645
V2src05_hrc06_525.yuv	5	6	–18,56515152	21,413895	0,3203024	0,3646
V2src05_hrc07_525.yuv	5	7	–8,154545455	15,446437	0,2071702	0,306
V2src05_hrc08_525.yuv	5	8	–4,006060606	10,836051	0,1652752	0,26662
V2src06_hrc05_525.yuv	6	5	–41,63181818	37,342789	0,5690291	0,56967
V2src06_hrc06_525.yuv	6	6	–29,48787879	26,660055	0,4370961	0,42391
V2src06_hrc07_525.yuv	6	7	–22,25909091	20,878248	0,3591788	0,35896

Nom du fichier	SRC	HRC	Note subjective moyenne brute	Note prévue par le modèle sur la base des données brutes	Note subjective moyenne corrigée	Note prévue par le modèle sur la base de données corrigées
V2src06_hrc08_525.yuv	6	8	–12,03181818	16,896168	0,2482169	0,31941
V2src07_hrc05_525.yuv	7	5	–23,89545455	19,086998	0,3796362	0,34067
V2src07_hrc06_525.yuv	7	6	–10,15606061	10,69402	0,2276934	0,26548
V2src07_hrc07_525.yuv	7	7	–4,240909091	4,896546	0,1644409	0,22267
V2src07_hrc08_525.yuv	7	8	–5,98030303	1,555055	0,1819566	0,20099
V2src08_hrc09_525.yuv	8	9	–76,2	52,094177	0,9513387	0,83024
V2src08_hrc10_525.yuv	8	10	–61,34545455	47,395226	0,789748	0,7397
V2src08_hrc11_525.yuv	8	11	–66,02575758	52,457584	0,8405916	0,83753
V2src08_hrc12_525.yuv	8	12	–37,20454545	37,931854	0,5221555	0,57874
V2src08_hrc13_525.yuv	8	13	–31,23030303	30,95985	0,4572049	0,4784
V2src08_hrc14_525.yuv	8	14	–31,26818182	33,293602	0,4614104	0,51031
V2src09_hrc09_525.yuv	9	9	–64,42878788	54,414772	0,8262912	0,87746
V2src09_hrc10_525.yuv	9	10	–49,92878788	36,080425	0,660339	0,55061
V2src09_hrc11_525.yuv	9	11	–53,73181818	46,338791	0,7100111	0,72031
V2src09_hrc12_525.yuv	9	12	–34,36969697	23,21393	0,4921708	0,38409
V2src09_hrc13_525.yuv	9	13	–22,85454545	16,955978	0,3656559	0,31998
V2src09_hrc14_525.yuv	9	14	–16,41666667	13,694396	0,2960957	0,29046
V2src10_hrc09_525.yuv	10	9	–72,11212121	48,179104	0,9084171	0,75433
V2src10_hrc10_525.yuv	10	10	–43,11666667	30,703861	0,5908784	0,475
V2src10_hrc11_525.yuv	10	11	–56,11969697	52,63887	0,7302376	0,84118
V2src10_hrc12_525.yuv	10	12	–19,55909091	21,95225	0,3345703	0,37033
V2src10_hrc13_525.yuv	10	13	–12,34393939	16,23988	0,2565459	0,31328
V2src10_hrc14_525.yuv	10	14	–16,05	23,201355	0,2953144	0,38395
V2src11_hrc09_525.yuv	11	9	–50,40454545	36,394535	0,6675853	0,55531
V2src11_hrc10_525.yuv	11	10	–54,26212121	37,812542	0,7054929	0,5769
V2src11_hrc11_525.yuv	11	11	–41,73636364	44,128036	0,5761193	0,68087
V2src11_hrc12_525.yuv	11	12	–19,03939394	14,619688	0,32761	0,29857
V2src11_hrc13_525.yuv	11	13	–17,72121212	14,12041	0,310495	0,29417
V2src11_hrc14_525.yuv	11	14	–19,4969697	14,927424	0,331051	0,30132
V2src12_hrc09_525.yuv	12	9	–61,35	40,051254	0,7883371	0,61229
V2src12_hrc10_525.yuv	12	10	–46,84545455	31,128973	0,6295301	0,48066
V2src12_hrc11_525.yuv	12	11	–51,80151515	41,77285	0,6809288	0,6406
V2src12_hrc12_525.yuv	12	12	–22,51969697	20,868282	0,3651402	0,35886
V2src12_hrc13_525.yuv	12	13	–14,17878788	15,040992	0,2714356	0,30234
V2src12_hrc14_525.yuv	12	14	–14,6030303	13,521517	0,2782449	0,28896
V2src13_hrc09_525.yuv	13	9	–55,25	38,691498	0,7211194	0,5906
V2src13_hrc10_525.yuv	13	10	–39,55	33,054504	0,5545722	0,50696
V2src13_hrc11_525.yuv	13	11	–40,03939394	45,9454	0,5525494	0,71318

Nom du fichier	SRC	HRC	Note subjective moyenne brute	Note prévue par le modèle sur la base des données brutes	Note subjective moyenne corrigée	Note prévue par le modèle sur la base de données corrigées
V2src13_hrc12_525.yuv	13	12	–14	16,631002	0,2708744	0,31692
V2src13_hrc13_525.yuv	13	13	–14,33181818	15,113959	0,27549	0,30299
V2src13_hrc14_525.yuv	13	14	–14,31969697	16,611286	0,2733771	0,31674

Données objectives et subjectives pour un système à 625 lignes

Nom du fichier	SRC	HRC	Note subjective moyenne brute	Note prévue par le modèle sur la base des données brutes	Note subjective moyenne corrigée	Note prévue par le modèle sur la base de données corrigées
V2src1_hrc2_625.yuv	1	2	38,85185185	31,764214	0,59461	0,47326
V2src1_hrc3_625.yuv	1	3	42,07407407	21,868561	0,64436	0,36062
V2src1_hrc4_625.yuv	1	4	23,77777778	12,195552	0,40804	0,27239
V2src1_hrc6_625.yuv	1	6	18,14814815	9,169512	0,34109	0,24887
V2src1_hrc8_625.yuv	1	8	12,92592593	6,738072	0,2677	0,23128
V2src1_hrc10_625.yuv	1	10	11,88888889	2,553883	0,26878	0,20356
V2src2_hrc2_625.yuv	2	2	33,51851852	31,492788	0,54173	0,46985
V2src2_hrc3_625.yuv	2	3	46,48148148	31,1313	0,70995	0,46535
V2src2_hrc4_625.yuv	2	4	13,33333333	20,241726	0,27443	0,34432
V2src2_hrc6_625.yuv	2	6	8,814814815	17,39045	0,22715	0,31721
V2src2_hrc8_625.yuv	2	8	7,074074074	14,914576	0,21133	0,29513
V2src2_hrc10_625.yuv	2	10	3,407407407	7,352309	0,16647	0,23562
V2src3_hrc2_625.yuv	3	2	48,07407407	38,852715	0,73314	0,56845
V2src3_hrc3_625.yuv	3	3	50,66666667	38,244621	0,76167	0,55982
V2src3_hrc4_625.yuv	3	4	32,11111111	27,733229	0,49848	0,42454
V2src3_hrc6_625.yuv	3	6	22,33333333	24,80323	0,38613	0,39159
V2src3_hrc8_625.yuv	3	8	16,33333333	23,296747	0,34574	0,37544
V2src3_hrc10_625.yuv	3	10	11,96296296	16,33028	0,26701	0,30759
V2src4_hrc2_625.yuv	4	2	36,14814815	42,041592	0,58528	0,61514
V2src4_hrc3_625.yuv	4	3	55,03703704	49,283836	0,90446	0,72942
V2src4_hrc4_625.yuv	4	4	39,7037037	38,322186	0,62361	0,56091
V2src4_hrc6_625.yuv	4	6	38,03703704	36,863457	0,61143	0,54053
V2src4_hrc8_625.yuv	4	8	24,40740741	32,46579	0,43329	0,48214
V2src4_hrc10_625.yuv	4	10	12,88888889	25,918123	0,26548	0,40388
V2src5_hrc2_625.yuv	5	2	38,62962963	38,95779	0,61973	0,56995
V2src5_hrc3_625.yuv	5	3	44,18518519	40,076313	0,68987	0,58609
V2src5_hrc4_625.yuv	5	4	24,66666667	23,166002	0,41648	0,37406
V2src5_hrc6_625.yuv	5	6	23,62962963	20,592213	0,4218	0,34778

Nom du fichier	SRC	HRC	Note subjective moyenne brute	Note prévue par le modèle sur la base des données brutes	Note subjective moyenne corrigée	Note prévue par le modèle sur la base de données corrigées
V2src5_hrc8_625.yuv	5	8	12,40740741	13,763152	0,27543	0,28531
V2src5_hrc10_625.yuv	5	10	7,37037037	8,418313	0,2022	0,24332
V2src6_hrc2_625.yuv	6	2	22,48148148	33,810165	0,38852	0,49949
V2src6_hrc3_625.yuv	6	3	27,07407407	25,004984	0,44457	0,39379
V2src6_hrc4_625.yuv	6	4	13,18518519	20,889347	0,27983	0,35074
V2src6_hrc6_625.yuv	6	6	14,44444444	17,418222	0,28106	0,31747
V2src6_hrc8_625.yuv	6	8	8,740740741	15,486559	0,23726	0,30011
V2src6_hrc10_625.yuv	6	10	5,518518519	11,509192	0,17793	0,2669
V2src7_hrc4_625.yuv	7	4	39,25925926	45,231079	0,59953	0,66412
V2src7_hrc6_625.yuv	7	6	33,85185185	43,131519	0,55093	0,63163
V2src7_hrc9_625.yuv	7	9	27,07407407	39,506535	0,45163	0,57784
V2src7_hrc10_625.yuv	7	10	19,25925926	34,418381	0,35617	0,50749
V2src8_hrc4_625.yuv	8	4	15,85185185	40,408993	0,32528	0,59095
V2src8_hrc6_625.yuv	8	6	17,03703704	38,552574	0,32727	0,56418
V2src8_hrc9_625.yuv	8	9	14,85185185	35,577034	0,30303	0,52297
V2src8_hrc10_625.yuv	8	10	11,48148148	30,278536	0,26366	0,45484
V2src9_hrc4_625.yuv	9	4	28,96296296	30,515778	0,47656	0,45775
V2src9_hrc6_625.yuv	9	6	30,51851852	26,971027	0,49924	0,41577
V2src9_hrc9_625.yuv	9	9	19,66666667	23,351355	0,39101	0,37601
V2src9_hrc10_625.yuv	9	10	20,92592593	17,856861	0,37122	0,32152
V2src10_hrc4_625.yuv	10	4	40,33333333	43,640377	0,70492	0,63942
V2src10_hrc6_625.yuv	10	6	37,33333333	40,552502	0,58218	0,59305
V2src10_hrc9_625.yuv	10	9	30,92592593	36,747391	0,49711	0,53893
V2src10_hrc10_625.yuv	10	10	21,2962963	30,161013	0,37854	0,45341
V2src11_hrc1_625.yuv	11	1	50,25925926	55,909908	0,79919	0,84263
V2src11_hrc5_625.yuv	11	5	35,51851852	44,049999	0,59256	0,64572
V2src11_hrc7_625.yuv	11	7	18,7037037	26,877754	0,34337	0,4147
V2src11_hrc10_625.yuv	11	10	15,07407407	23,420477	0,30567	0,37674
V2src12_hrc1_625.yuv	12	1	36,33333333	43,837097	0,61418	0,64244
V2src12_hrc5_625.yuv	12	5	38,44444444	40,349903	0,6661	0,59008
V2src12_hrc7_625.yuv	12	7	31,11111111	37,254383	0,53242	0,54594
V2src12_hrc10_625.yuv	12	10	26,14814815	28,953564	0,44737	0,43887
V2src13_hrc1_625.yuv	13	1	43,7037037	38,333649	0,74225	0,56108
V2src13_hrc5_625.yuv	13	5	43,2962963	34,290554	0,66799	0,5058
V2src13_hrc7_625.yuv	13	7	25,2962963	26,990025	0,42065	0,41598
V2src13_hrc10_625.yuv	13	10	15,88888889	20,181463	0,33381	0,34373

Annexe B

Yonsei University/SK Telecom/Radio Research Laboratory

Description fonctionnelle du modèle de qualité vidéo avec une image de référence complète

B.1 Introduction

Depuis toujours, on utilise pour évaluer la qualité vidéo un certain nombre d'évaluateurs qui évaluent subjectivement la qualité vidéo. L'évaluation peut être faite avec ou sans séquence vidéo de référence. Dans une évaluation avec séquence de référence, on montre aux évaluateurs deux séquences vidéo: la séquence vidéo de référence (source) et la séquence vidéo traitée qui sera comparée avec la séquence vidéo source. En comparant les deux séquences vidéo, les évaluateurs attribuent des notes subjectives à chacune d'elles. Par conséquent, on parle souvent de test subjectif de qualité vidéo. Le test subjectif est considéré comme la méthode la plus précise étant donné qu'il reflète la perception de l'homme, mais il comporte plusieurs limitations. Tout d'abord il suppose la présence d'un certain nombre d'évaluateurs. Il est donc chronophage et coûteux. Par conséquent, il ne peut être fait en temps réel. On s'est donc beaucoup intéressé à l'élaboration de méthodes objectives de mesure de la qualité vidéo. Un critère important pour une méthode objective de mesure de la qualité vidéo est que cette méthode donne des résultats cohérents pour toute une série de séquences vidéo qui ne sont pas utilisées au stade de la conception. Dans cette optique, on a élaboré un modèle facile à implémenter, suffisamment rapide pour des implémentations en temps réel et résistant à toute une série de dégradations vidéo. Ce modèle est un produit élaboré conjointement par Yonsei University, SK Telecom et Radio Research Laboratory.

B.2 Mesure objective de la qualité vidéo basée sur la dégradation des contours

B.2.1 Rapport PSNR basé sur la dégradation des contours (EPSNR, *edge PSNR*)

Le modèle de mesure objective de la qualité vidéo est une méthode avec une image de référence complète. En d'autres termes, on suppose qu'une séquence vidéo de référence est fournie. En analysant comment les êtres humains perçoivent la qualité vidéo, on observe que le système visuel humain est sensible aux dégradations autour des contours. En d'autres termes lorsque les zones des contours d'une séquence vidéo sont floues, les évaluateurs ont tendance à donner à cette séquence de mauvaises notes même si l'erreur quadratique moyenne globale est faible. On observe en outre que les algorithmes de compression vidéo ont tendance à produire davantage de défauts (artéfacts) autour des zones des contours. Sur la base de cette observation, le modèle fournit une méthode de mesure objective de la qualité vidéo qui permet de mesurer les dégradations autour des contours. Dans ce modèle on applique tout d'abord un algorithme de détection des bords à la séquence vidéo source pour localiser les zones des bords. Ensuite on mesure la dégradation de ces zones des bords en calculant l'erreur quadratique moyenne. A partir de cette erreur on calcule le rapport EPSNR, rapport que l'on utilise comme mesure de la qualité vidéo après post-traitement.

Dans le modèle, il faut tout d'abord appliquer un algorithme de détection de contours pour localiser les régions des contours. On peut utiliser n'importe quel algorithme de détection de contours même s'il peut y avoir des différences minimes dans les résultats. Par exemple, on peut utiliser n'importe quel opérateur gradient pour localiser les régions des contours. Un certain nombre d'opérateurs gradient ont été proposés. Dans de nombreux algorithmes de détection de contours, on calcule tout d'abord à l'aide d'opérateur gradient l'image du gradient horizontal $g_{horizontal}(m,n)$ et l'image du gradient vertical $g_{vertical}(m,n)$. On peut ensuite calculer l'image du gradient d'amplitude $g(m,n)$ comme suit:

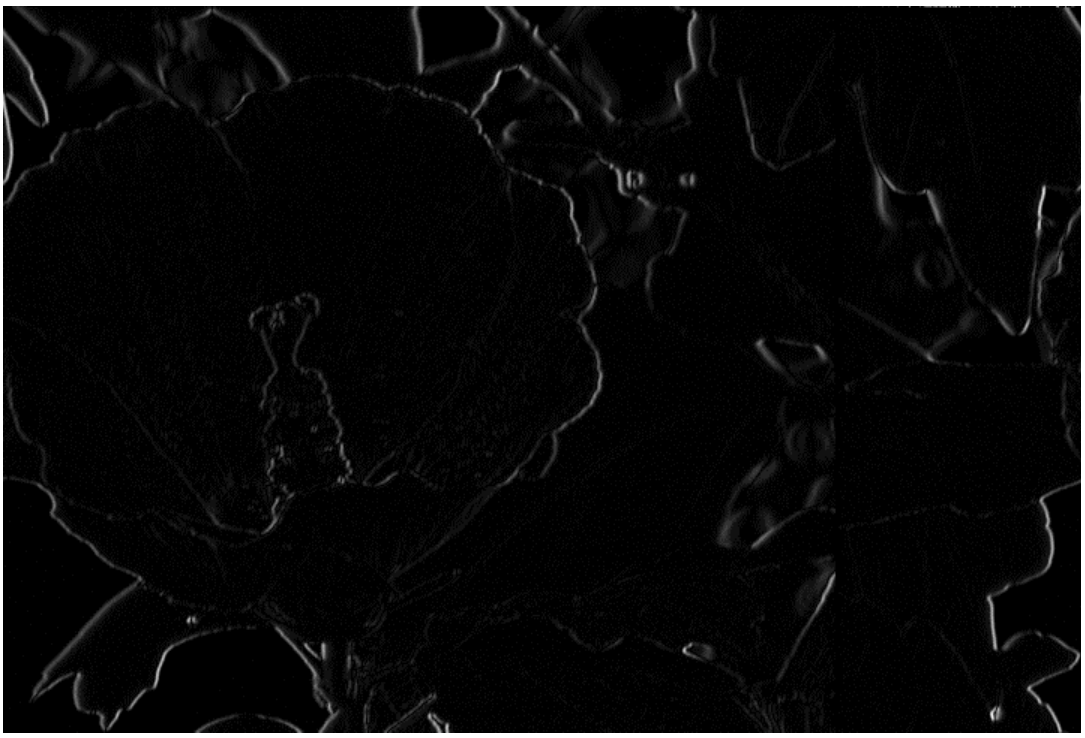
$$g(m,n) = \left|g_{horizontal}(m,n)\right| + \left|g_{vertical}(m,n)\right|$$

Enfin, on applique un seuillage à l'image du gradient d'amplitude $g(m,n)$ pour trouver les régions des contours. En d'autres termes, les pixels dont les gradients d'amplitude dépassent une valeur seuil sont considérés comme étant les régions des contours.

Les Figures B.1 à B.5 illustrent cette procédure. La Figure B.1 montre une image source. La Figure B.2 montre une image du gradient horizontal $g_{horizontal}(m,n)$, laquelle est obtenue par application d'un opérateur gradient horizontal à l'image source de la Figure B.1. La Figure B.3 montre une image du gradient vertical $g_{vertical}(m,n)$, laquelle est obtenue par application d'un opérateur gradient vertical à l'image source de la Figure B.1. La Figure B.4 montre l'image du gradient d'amplitude (image des contours) et la Figure B.5 l'image binaire des contours (image de masquage), lesquelles sont obtenues par application d'un seuillage à l'image du gradient d'amplitude de la Figure B.4.

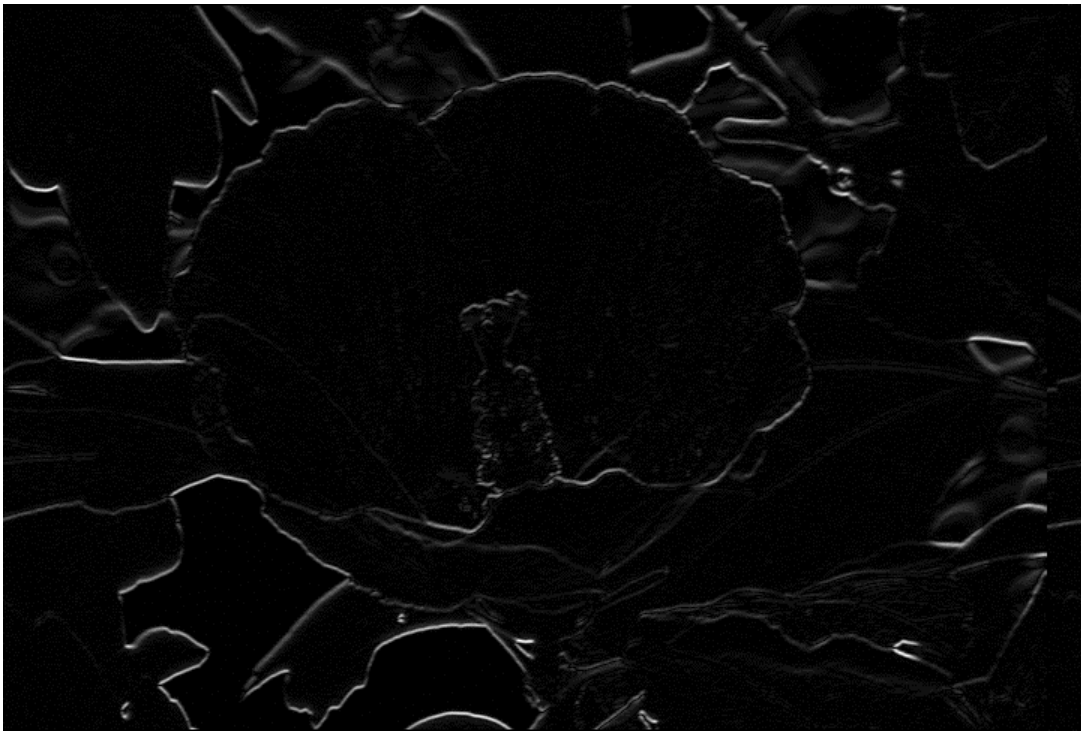
1683-07

Figure B.1 – Image source (image d'origine)



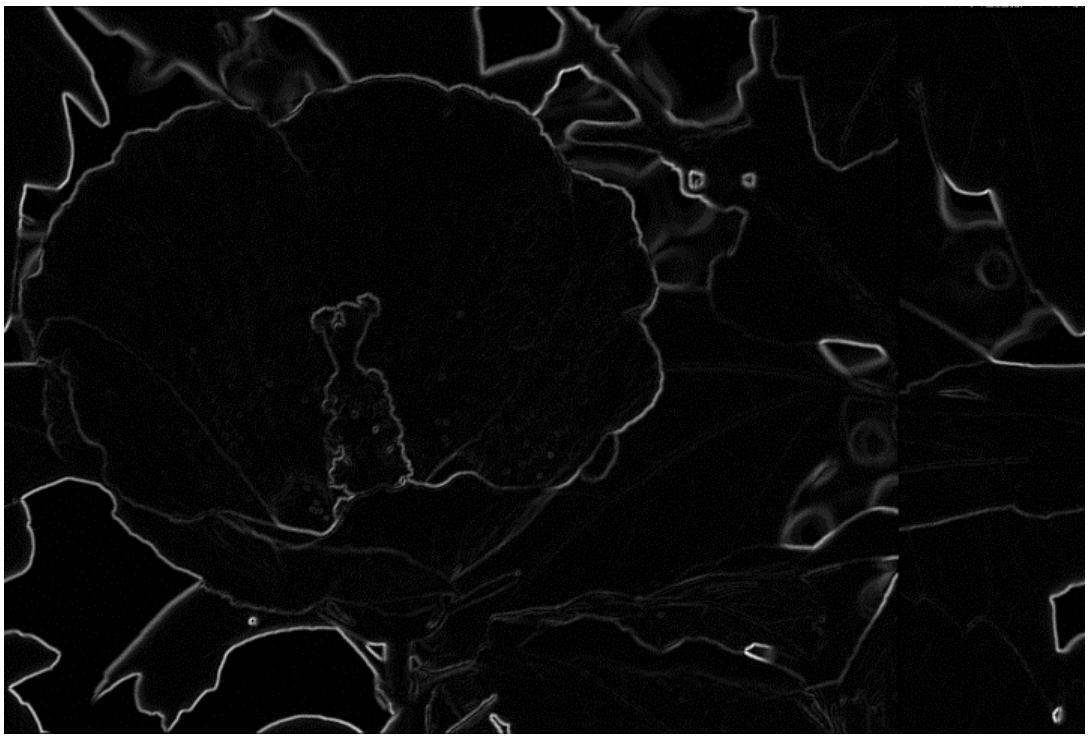
1683-08

Figure B.2 – Image du gradient horizontal, laquelle est obtenue par application d'un opérateur gradient horizontal à l'image source de la Figure B.1



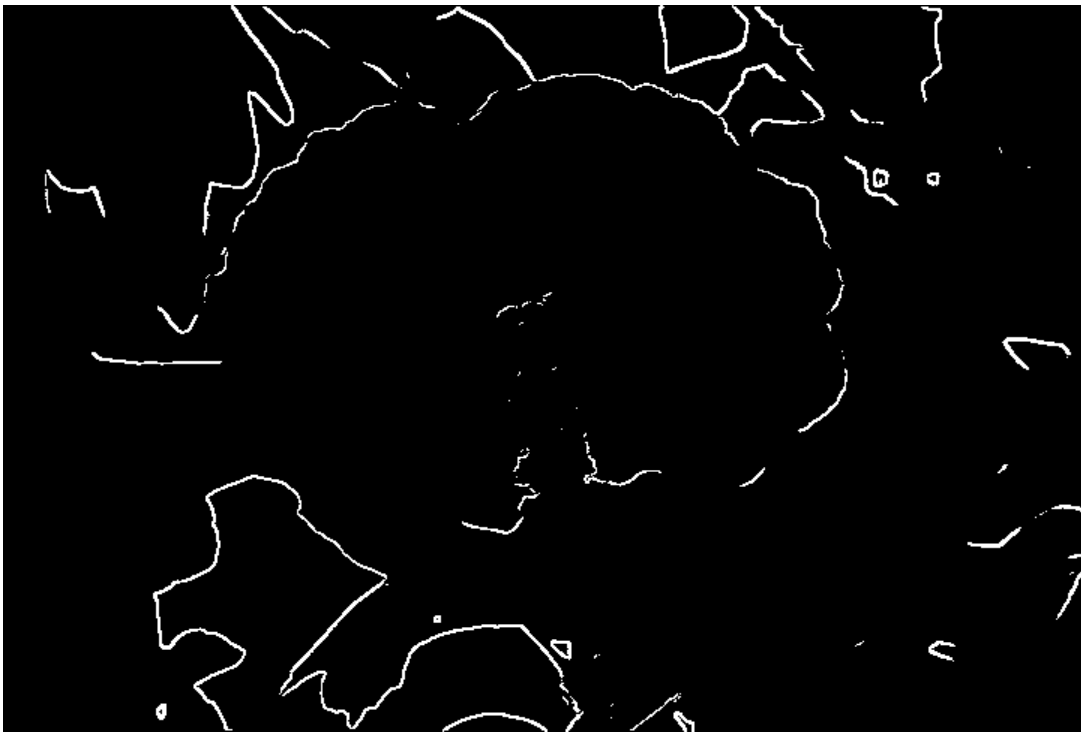
1683-09

Figure B.3 – Image du gradient vertical, laquelle est obtenue par application d'un opérateur gradient vertical à l'image source de la Figure B.1



1683-10

Figure B.4 – Image du gradient d'amplitude

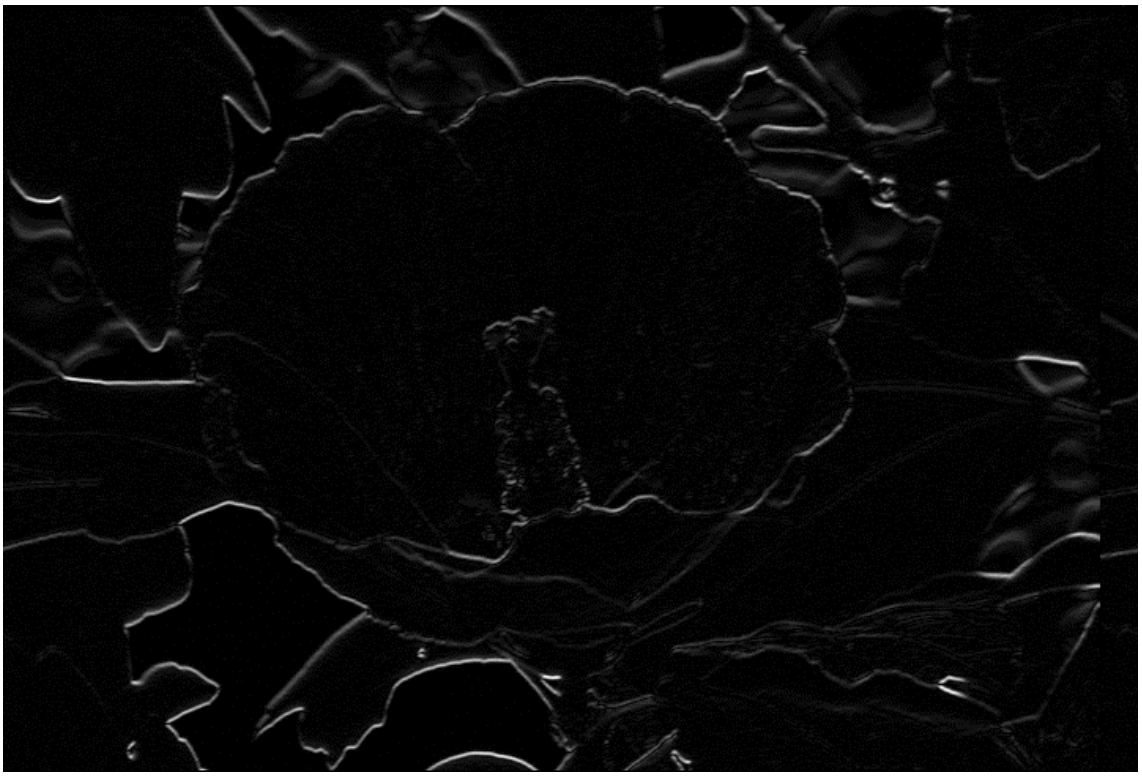


1683-11

Figure B.5 – Image binaire des contours (image masque) obtenue par application d'une opération de seuillage à l'image du gradient d'amplitude de la Figure B.4

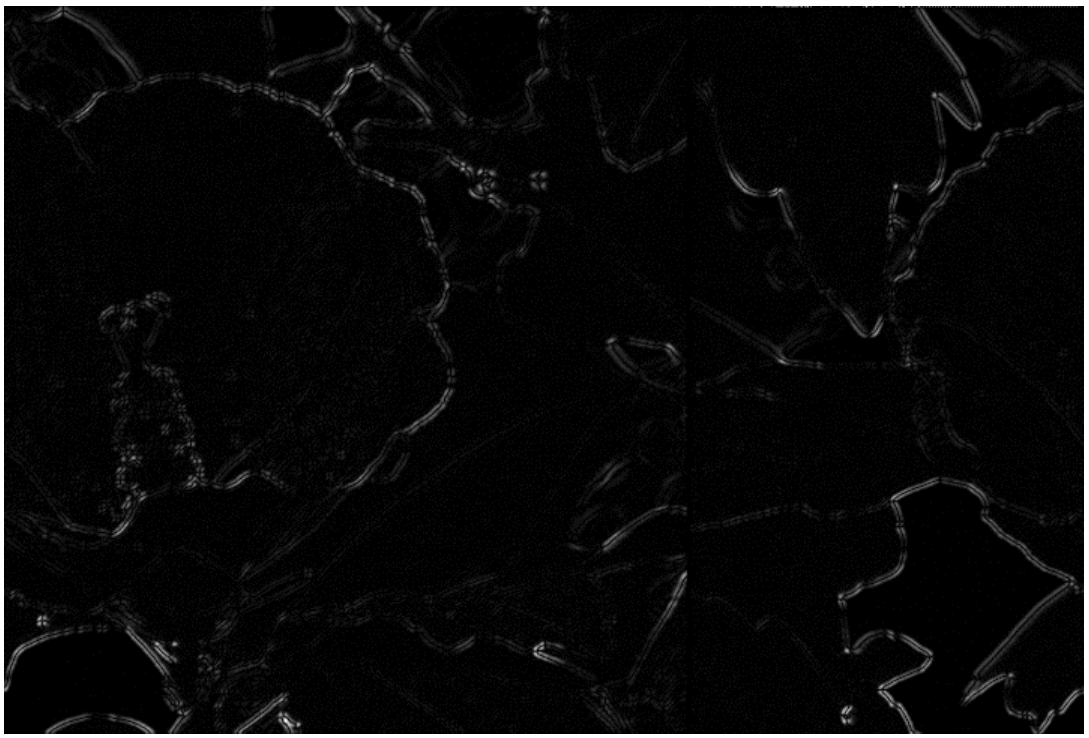
On peut également utiliser une procédure modifiée pour localiser les régions des contours, par exemple, en appliquant tout d'abord un opérateur gradient vertical à l'image source, ce qui donne

l'image du gradient vertical; on applique ensuite un opérateur gradient horizontal à l'image du gradient vertical, ce qui donne une image du gradient successif modifié (image du gradient horizontal et du gradient vertical). Enfin, on peut appliquer un seuillage à l'image du gradient successif modifié pour trouver les régions des contours. En d'autres termes, les pixels de l'image du gradient successif modifié qui dépassent une valeur seuil sont considérés comme étant les zones des contours. Les Figures B.6 à B.9 illustrent la procédure modifiée. La Figure B.6 montre une image du gradient vertical $g_{vertical}(m,n)$, laquelle est obtenue par application d'un opérateur gradient vertical à l'image source de la Figure B.1. La Figure B.7 montre une image du gradient successif modifié (image du gradient horizontal et du gradient vertical), laquelle est obtenue par application d'un opérateur gradient horizontal à l'image du gradient vertical de la Figure B.6. La Figure B.8 montre l'image binaire des contours (image masque) obtenue par application d'un seuillage à l'image du gradient successif modifié de la Figure B.7.



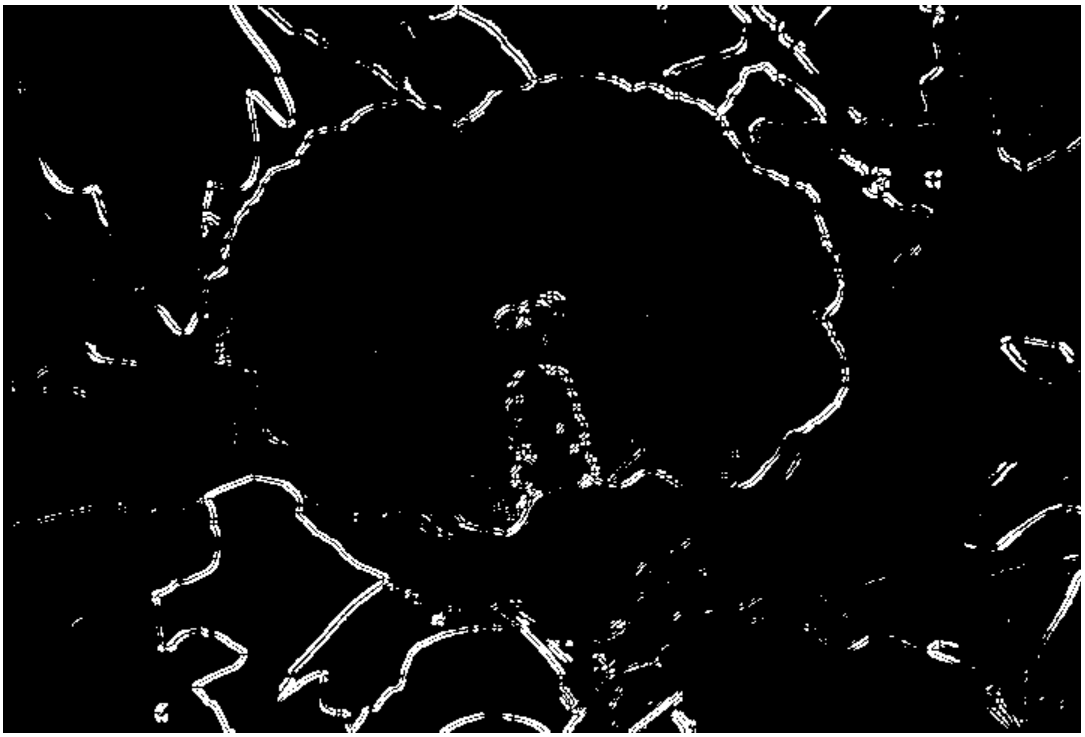
1683-12

Figure B.6 – Image du gradient vertical, laquelle est obtenue par application d'un opérateur gradient vertical à l'image source de la Figure B.1



1683-13

Figure B.7 – Image des gradients successifs modifiée (image des gradients horizontal et vertical), laquelle est obtenue par application d'un opérateur gradient horizontal à l'image du gradient vertical de la Figure B.6



1683-14

Figure B.8 – Image binaire des contours (image masque), obtenue par application d'un seuil à l'image des gradients successifs modifiée de la Figure B.7

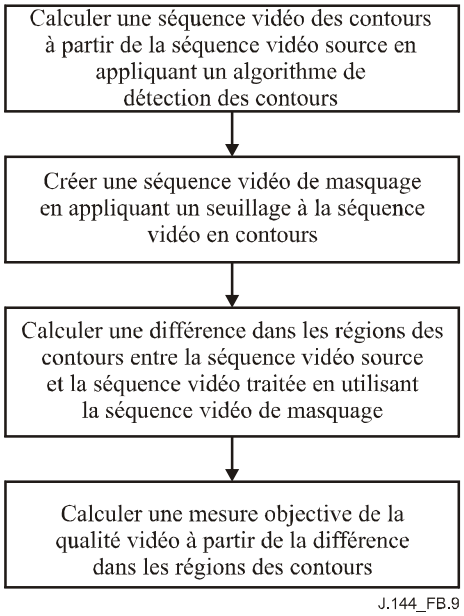


Figure B.9 – Schéma fonctionnel d'un rapport EPSNR

On notera que les deux méthodes peuvent être considérées comme un algorithme de détection de contours. On peut choisir n'importe quel algorithme de détection de contours selon la nature des séquences vidéo et des algorithmes de compression. Toutefois, certaines méthodes peuvent donner de meilleurs résultats que d'autres.

Ainsi, dans le modèle, on applique tout d'abord un opérateur de détection de contours; ce qui permet d'obtenir des images des contours (Figures B.4 et B.7). Ensuite, on crée une image de masquage (image binaire des contours) en appliquant un seuillage à l'image des contours (Figures B.5 et B.8). En d'autres termes, les pixels de l'image des contours dont la valeur est inférieure au seuil t_e sont mis à zéro et les pixels dont la valeur est égale ou supérieure à ce seuil sont positionnés à une valeur autre que zéro. Les Figures 5 et 8 donnent des exemples d'images de masquage. On notera que cet algorithme de détection des contours est appliqué à l'image source. On peut appliquer l'algorithme de détection des contours aux images traitées mais il est plus exact de l'appliquer aux images source. Etant donné qu'une séquence vidéo peut être considérée comme une séquence d'images ou de trames, la procédure susmentionnée peut être appliquée à chaque image ou à chaque trame de séquence vidéo. Etant donné que le modèle peut être utilisé pour des séquences vidéo composées de trames ou d'images, on utilisera le terme "d'image" pour parler indifféremment de trame ou d'image.

Ensuite, on calcule les différences entre la séquence vidéo source et la séquence vidéo traitée correspondant aux pixels ayant une valeur autre que zéro de l'image de masquage. En d'autres termes, l'erreur quadratique des régions des contours de la *l*ième trame est calculée comme suit:

$$se_e^l = \sum_{i=1}^{M}\sum_{j=1}^{N}\{S^l(i,j) - P^l(i,j)\}^2 \; if \left|R^l(i,j)\right| \neq 0 \qquad \text{(B-1)}$$

où $S^l(i,j)$ est la *l*ième image de la séquence vidéo source, $P^l(i,j)$ est la *l*ième image de la séquence vidéo traitée, $R^l(i,j)$ est la *l*ième image de la séquence vidéo de masquage, M est le nombre de rangées et N le nombre de colonnes. Lorsque le modèle est implémenté, on peut sauter la génération de la séquence vidéo de masquage. En fait, sans créer la séquence vidéo de masquage, l'erreur quadratique des régions des contours de la *l*ième image est calculée comme suit:

$$se_e^l = \sum_{i=1}^{M}\sum_{j=1}^{N}\{S^l(i,j) - P^l(i,j)\}^2 \; if \left|Q^l(i,j)\right| \geq t_e \qquad \text{(B-2)}$$

où $Q^l(i,j)$ est la *l*ième image de la séquence vidéo des contours et t_e est un seuil. L'erreur quadratique moyenne est utilisée à l'équation B-1 pour calculer la différence entre la séquence vidéo source et la séquence vidéo traitée mais on peut utiliser tout autre type de différence. Par exemple, on peut également utiliser la différence absolue. Dans le modèle soumis aux tests VQEG Phase II, t_e a été mis à 260 et l'algorithme de détection des contours modifié a été utilisé avec l'opérateur de Sobel.

Cette procédure est répétée pour l'ensemble des séquences vidéo et l'erreur quadratique moyenne des contours est calculée comme suit:

$$mse_e = \frac{1}{K}\sum_{l=1}^{L} se_e^l \tag{B-3}$$

où L est le nombre d'images (trames ou images) et K est le nombre total de pixels des contours. Enfin, le rapport PSNR des zones des contours est calculé comme suit:

$$EPSNR = 10\log_{10}\left(\frac{P^2}{mse_e}\right) \tag{B-4}$$

où P est la valeur crête des pixels. Dans le modèle, ce rapport PSNR pour les contours (EPSNR) est utilisé comme note objective de base de la qualité vidéo. La Figure B.9 donne un schéma fonctionnel de calcul du rapport EPSNR.

B.2.2 Postajustements

B.2.2.1 Désaccentuation d'un rapport EPSNR élevé

Lorsque le rapport EPSNR a une valeur supérieure à 35, il surestime, semble-t-il, la qualité perceptuelle. On utilise par conséquent la mise à l'échelle linéaire par paliers suivante:

$$EPSNR = \begin{cases} EPSNR & si\ 0 \le EPSNR \le 35 \\ EPSNR \times 0.9 & si\ 35 \le EPSNR \le 40 \\ EPSNR \times 0.8 & si\ EPSNR > 40 \end{cases} \tag{B-5}$$

B.2.2.2 Prise en considération de contours floutés

On observe que lorsque les contours sont très flous dans des séquences vidéo de qualité médiocre, les évaluateurs ont tendance à donner des notes subjectives médiocres. En d'autres termes, si les régions des contours de la séquence vidéo traitée sont nettement plus petites que celles de la séquence vidéo source, les évaluateurs donnent de moins bonnes notes. Par ailleurs, on observe que certaines séquences vidéo ont un très petit nombre de pixels ayant des composantes haute fréquence. En d'autres termes, le nombre de pixels des régions des contours est très faible. Pour tenir compte de ces problèmes, les régions des contours de la séquence vidéo source et de la séquence vidéo traitée sont calculées et le rapport EPSNR est modifié comme suit:

$$MEPSNR = \begin{cases} EPSNR - 60 \times \left(0.1225 - \left(\frac{EP_{common}}{EP_{src}}\right)^2\right) & si\ EPSNR < 25\ et\ \frac{EP_{common}}{EP_{src}} < 0.35\ et \frac{EP_{hrc}}{EP_{src}} < 0.13 \\ EPSNR & dans\ les\ autres\ cas \end{cases} \tag{B-6}$$

où:

EP_{src}: nombre total des pixels des contours dans la séquence vidéo (source) SRC

EP_{hrc}: nombre total de pixels des contours dans les séquences vidéo traitées (HRC)

EP_{common}: nombre total de pixels des contours communs dans les séquences vidéo SRC et HRC (c'est-à-dire pixels des contours apparaissant au même endroit)

MEPSNR: EPSNR modifié.

Pour certaines séquences vidéo, EP_{src} peut être très faible. Dans le cas le plus défavorable, EP_{src} peut être nul (image effacée ou image à très faible fréquence), ce qui entraîne une erreur de division par 0. Afin d'éviter ce type de cas, il est suggéré à la modification suivante: Si EP_{src} est inférieur à 10 000 pixels (environ 10 000/240 = 41,7 pixels par trame pour des séquences vidéo 525 lignes de 8 secondes et d'environ 10 000/200 = 50 pixels par trame pour des séquences vidéo 625 lignes de 8 secondes), l'utilisateur peut réduire le seuil t_e dans l'équation B-2 de 20 jusqu'à ce que EP_{src} soit supérieur ou égal à 10 000 pixels. Si EP_{src} est inférieur à 10 000 pixels, on ne procède pas au postajustement à l'aide de l'équation B-6. Dans ce cas, on calcule le rapport EPSNR en utilisant $t_e = 60$. Si cette option est retenue, l'utilisateur peut supprimer la condition $EP_{hrc}/EP_{src} < 0{,}13$ dans l'équation B-6.

B.2.2.3 Mise à l'échelle

Ensuite, les notes objectives sont remises à l'échelle de façon à être comprises entre 0 (non distinguable de séquence vidéo d'origine) et 1.

$$VQM = 1 - MEPSNR \times 0.02 \tag{B-7}$$

Cette mesure VQM est utilisée comme la note objective du modèle.

B.2.3 Précision d'alignement

La précision d'alignement recommandée pour le modèle est une précision d'un demi-pixel dans les séquences vidéo entrelacées, ce qui équivaut à une précision d'un quart de pixel dans un format vidéo progressif. L'interpolation spline cubique [B-2], voire mieux, est fortement recommandée pour calculer les valeurs des sous-pixels.

B.2.4 Schéma fonctionnel du modèle

La Figure B.10 donne le schéma fonctionnel complet du modèle. D'autre part la Figure B.11 indique un schéma fonctionnel modifié, qui évite l'erreur de la division par zéro.

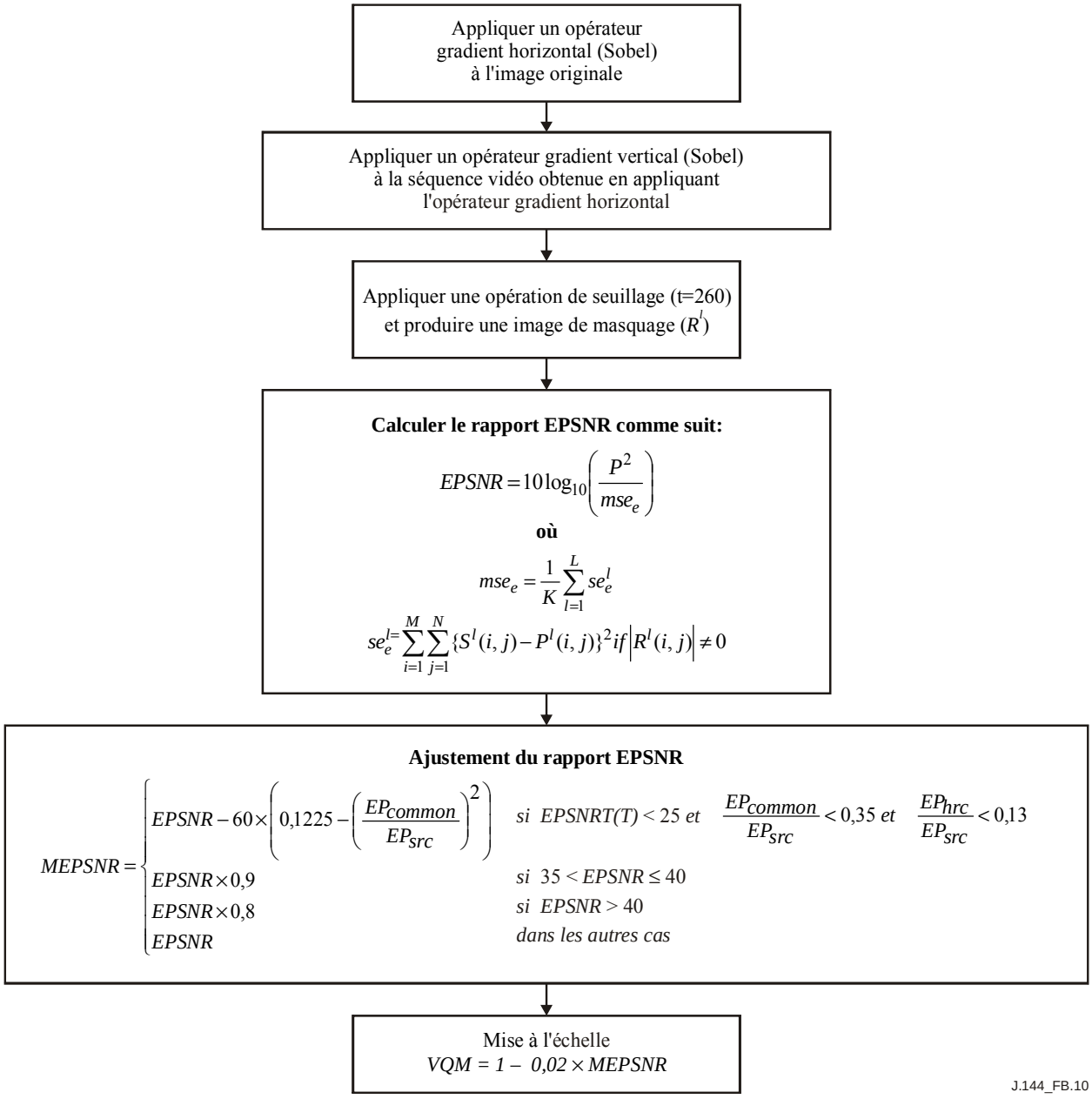


Figure B.10 – Schéma fonctionnel complet du modèle (le modèle utilise *P*=255)

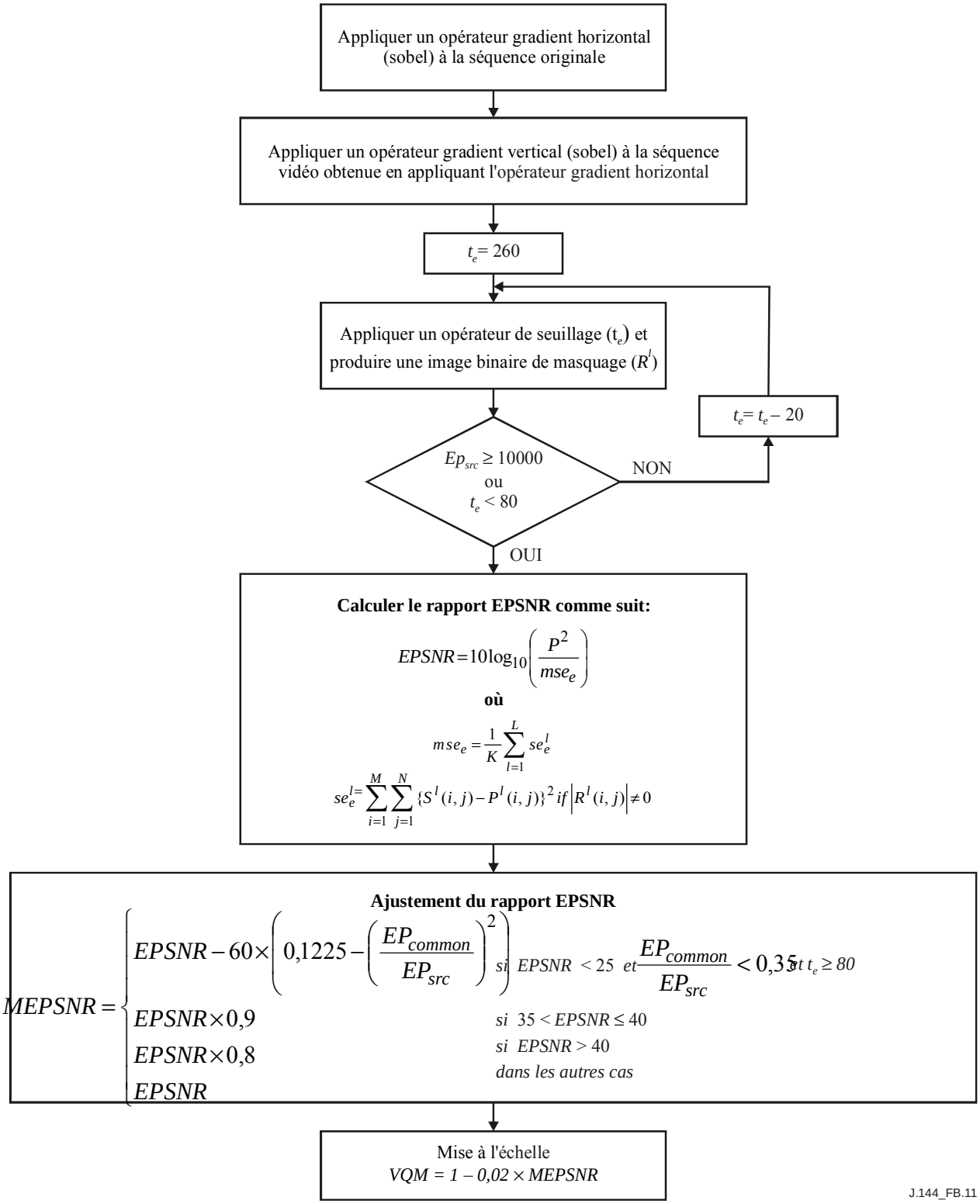


Figure B.11 – Schéma fonctionnel modifié, évitant l'erreur de la division par zéro (le modèle utilise *P*=255)

B.3 Alignement

B.3.1 Alignement vidéo

L'obtention de la meilleure adaptation entre deux séquences vidéo exige un alignement. Pour évaluer la qualité vidéo, il faut déterminer l'importance du décalage spatial ou temporel de la vidéo

traitée. Si l'on représente le décalage par le vecteur de déplacement $D = [d_1, d_2, d_3]^T$, l'erreur quadratique moyenne entre une séquence vidéo d'origine et une séquence vidéo traitée décalée sous l'effet de *D* est calculée comme suit:

$$MSE(d_1, d_2, d_3) = \frac{1}{LMN} \sum_{l}^{L} \sum_{m}^{M} \sum_{n}^{N} (U(m,n,l) - V(m + d_1, n + d_2, l + d_3))^2 \qquad \text{(B-8)}$$

où *U* et *V* représente les séquences vidéo. La meilleure estimation du vecteur de déplacement assurant l'adaptation optimale est obtenue en réduisant au minimum l'erreur quadratique moyenne:

$$\hat{D} = \arg \min_{(d_1, d_2, d_3)} MSE(d_1, d_2, d_3) \qquad \text{(B-9)}$$

La précision des composantes verticales et horizontales du vecteur de déplacement peut être de l'ordre d'un pixel ou d'une fraction de pixel. S'il s'agit d'une fraction de pixel, il faut utiliser une technique d'interpolation de type bilinéaire ou spline cubique, par exemple. En règle générale, la précision de la composante temporelle du vecteur de déplacement est d'une trame vidéo pour une séquence vidéo dans un format vidéo progressif, tel qu'indiqué à la Figure B.12. Pour une séquence vidéo entrelacée, le décalage est d'une image. 1/50 de seconde pour les vidéo entrelacées à 50 Hz et 1/60 sec pour les vidéo entrelacée à 60 Hz tel qu'indiqué à la Figure B.13. Pour le format entrelacé, il faut construire une trame complète à partir de chaque image de façon à déterminer le déplacement spatial des séquences vidéo entrelacées.

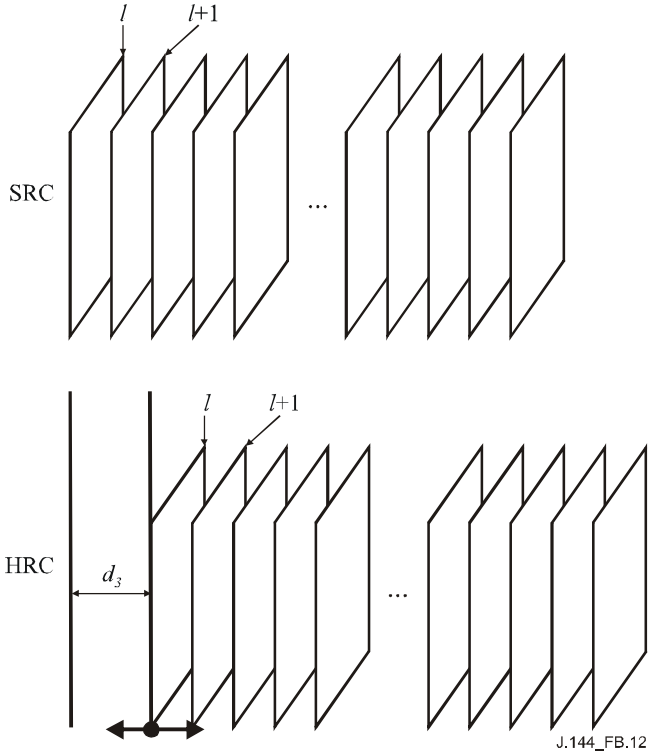


Figure B.12 – Alignement temporel des séquences vidéo de format progressif. Unité de d_3 une trame vidéo

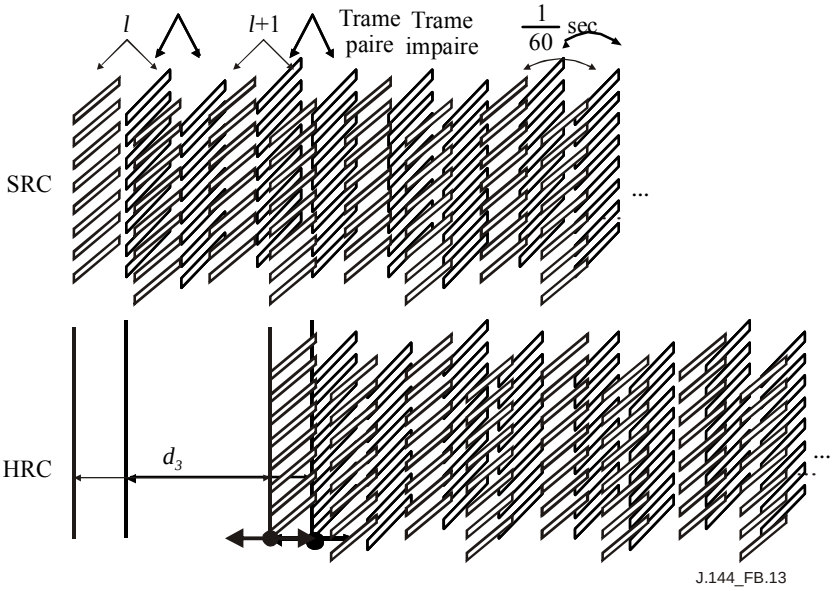


Figure B.13 – Alignement temporel des séquences vidéo entrelacées. Unité d_3: une image (1/60 sec pour les séquences vidéo entrelacées à 60 Hz)

B.3.2 Alignement vidéo d'après la région d'intérêt

Normalement, la détermination de la meilleure adaptation prend un temps de traitement très long à l'aide des équations B-8 et B-9. S'il faut un alignement précis, toutes les trames d'une séquence vidéo peuvent servir à la détermination de l'erreur quadratique moyenne. Toutefois, il faudrait un temps de traitement considérable et cette opération n'est pas nécessairement réalisable en temps réel. Dans le but de réduire le temps de traitement, avec le modèle en question, un nombre restreint de sous-régions (régions d'intérêt) sont choisies dans la séquence vidéo. Ensuite le modèle détermine la meilleure adaptation entre deux séquences vidéo en calculant l'erreur quadratique moyenne dans les régions d'intérêt (ROI, *regions of interest)*:

$$MSE_{\mathrm{ROI}}(d_1,d_2,d_3)=\frac{1}{K}\sum_{(m,n,l)\in ROI}(U(m,n,l)-V(m+d_1,n+d_2,l+d_3))^2 \qquad \text{(B-10)}$$

où K représente le nombre de pixels dans la région d'intérêt.

On constate que les zones des séquences vidéo dont l'évolution est rapide fournissent de précieuses informations utiles à l'alignement de l'image. Par conséquent, le modèle situe les zones de ce type et les utilise pour les besoins de l'alignement vidéo. Afin d'identifier les scènes comportant une évolution rapide, l'erreur quadratique moyenne de trame (*(fMSE)* de la $l^{\text{ème}}$ trame d'une séquence vidéo d'origine est calculée comme suit:

$$fMSE(l)=\frac{1}{MN}\sum_{m}^{M}\sum_{n}^{N}(U(m,n,l)-U(m,n,l+1))^2 \qquad \text{(B-11)}$$

Après avoir calculé *fMSE*(*l*) pour toute les trames de la séquence vidéo d'origine, le modèle choisi cinq trames de référence dont les valeurs *fMSE* sont maximales. Ces trames peuvent être considérées comme celles dont l'évolution est la plus rapide dans le sens temporel. De plus, le modèle introduit une restriction supplémentaire selon laquelle les intervalles entre les trames de référence doivent être supérieurs à un certain délai. A cet effet, le modèle divise uniformément la totalité de la séquence vidéo en cinq sous-séquences. Dans chacune des sous-séquences, le modèle choisit une trame pour laquelle la valeur *fMSE* est la plus élevée à partir de l'équation B-11.

Parmi les cinq trames choisies, le modèle choisit celle dont l'erreur quadratique moyenne est la plus élevée. Ensuite, il applique une fonction de transformation par ondelettes 2-D à la dite trame, les zones dotées de coefficient de haute fréquence prépondérante sont considérées comme celles dont l'évolution est rapide, dans le domaine spatial. La transformée par ondelettes 2-D est obtenue en appliquant deux transformations distinctes par ondelettes 1-D dans le sens horizontal et dans le sens vertical. La transformée par ondelettes 1-D est calculée comme suit:

$$y_L^{(1)}[n] = \sum_k x[k] h_0[2n-k] \tag{B-12}$$

$$y_H^{(1)}[n] = \sum_k x[k] h_1[2n-k] \tag{B-13}$$

où $x[n]$ est un signal d'entrée 1-D et $h_0[n]$, $h_1[n]$ sont les réponses impulsionnelles des filtres d'analyse passe bas et passe haut (filtres Harr). Les termes $y_L^{(1)}[n]$, $y_H^{(1)}[n]$ représentent les signaux de sortie des filtres. Le modèle applique à plusieurs reprises la transformée par ondelettes à la sous-bande haute fréquence selon la formule:

$$y_H^{(l+1)}[n] = \sum_k y_H^{(l)} h_1[2n-k] \tag{B-14}$$

La Figure B.14 représente la transformée récursive de la sous-bande haute fréquence dans le cas de l'image 2-D. Le modèle choisit ensuite douze coefficient de valeur maximale dans la sous-bande de fréquence la plus élevée et les sous-régions correspondantes sont retenues en tant que régions d'intérêt tel qu'indiqué à la Figure B.14. Le modèle utilise une décomposition à 4 niveaux. Ainsi, la dimension des régions d'intérêt de l'image d'origine est 16×16 at$(2^4 m, 2^4 n)$. De cette façon, le modèle choisit 12 régions d'intérêt 16×16 dont les fréquences spatiales sont élevées.

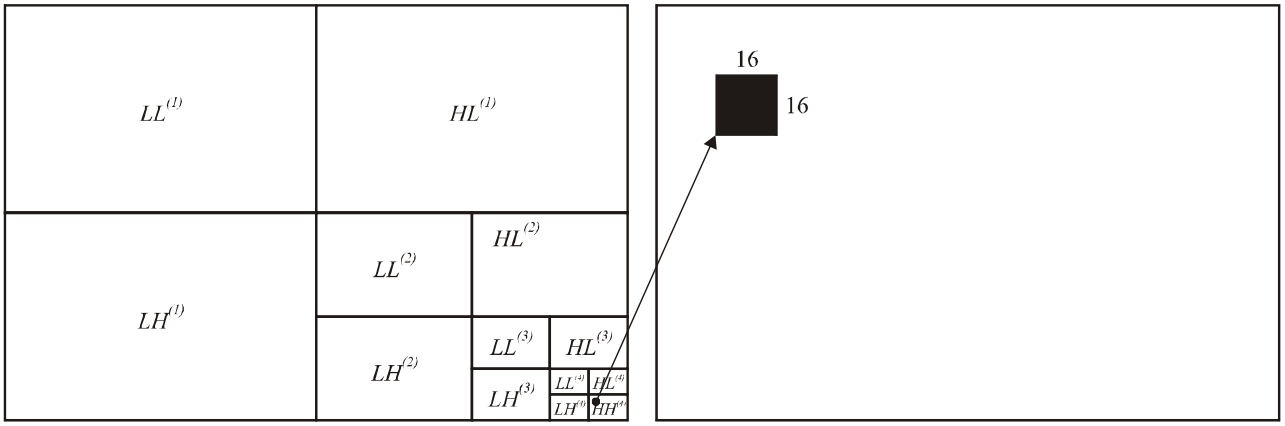


Figure B.14 – Trame transformée par ondelettes et région d'intérêt utilisée pour l'alignement des séquences vidéo rapides

Afin de choisir les régions d'intérêt dont les fréquences temporelles sont élevées, la trame dont la valeur *fMSE* est la plus élevée est divisée en un certain nombre de blocs 16×16 (Figure B.15). Ensuite, le modèle choisit les douze blocs dont le différence absolue d'un bloc à l'autre est la plus élevée. La différence absolue d'un bloc à l'autre (*ABD*) est calculée comme suit:

$$ABD = \frac{1}{256} \sum_{(m,n) \in k-th\,block} \left| U(m,n,l) - U(m,n,l+1) \right| \tag{B-15}$$

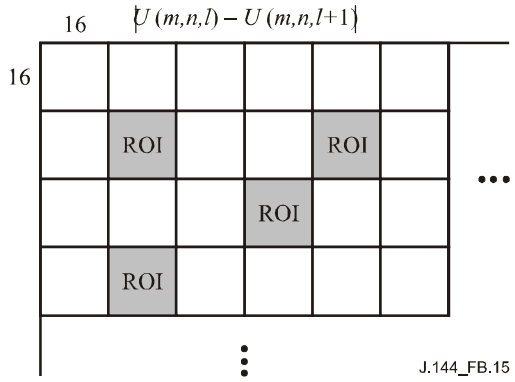


Figure B.15 – Le modèle choisit 12 blocs dont la valeur ABD d'un bloc à l'autre est la plus élevée, dans la trame dont l'erreur quadratique moyenne est la plus élevée

Le modèle choisit donc 12 blocs dont les fréquences spatiales sont élevées et 12 blocs dont la différence absolue *ABD* est maximale dans la trame dont l'erreur quadratique moyenne de trame est la plus élevée. Ces 24 blocs sont utilisés en tant que région d'intérêt (ROI). Il convient de signaler que les blocs (24×4=96 blocs) des quatre trames restantes qui se trouvent au mêmes emplacements sont également utilisés en tant que région d'intérêt. Par conséquent, l'ensemble des 120 blocs dont la taille est de 16×16 font office de régions d'intérêt et l'équation B-10 permet de déterminer le vecteur de déplacement optimal. Il convient en outre d'observer que les 24 blocs de la trame dont l'erreur quadratique moyenne de trame est la plus élevée ont une pondération double par comparaison aux blocs des quatre trames restantes. Au cours du processus d'alignement, l'interpolation spline cubique permet donc de réaliser un processus d'alignement à un quart de pixel.

B.4 Conclusion

Un nouveau modèle de mesure objective de la qualité vidéo basé sur la dégradation des contours est proposé. Ce modèle est extrêmement rapide. Une fois la représentation binaire générée, le modèle est plusieurs fois plus rapide que le rapport PSNR classique, d'où une amélioration importante. Par conséquent, le modèle convient bien pour les applications qui nécessitent une évaluation de la qualité vidéo en temps réel.

B.5 Références

[B-1] Document de référence UIT-T (2004), *Objective perceptual assessment of video quality: Full reference television.*

Tableau B.1 – Matrice VQM à 525 lignes (données brutes)[2]

SRC (Image)	HRC=1		HRC=2		HRC=3		HRC=4		HRC=5		HRC=6		HRC=7		HRC=8		HRC=9		HRC=10		HRC=11		HRC=12		HRC=13		HRC=14	
1	1	0,679	4	0,525	7	0,512	10	0,419																				
2	2	0,431	5	0,365	8	0,313	11	0,342																				
3	3	0,558	6	0,452	9	0,340	12	0,305																				
4									13	0,668	17	0,581	21	0,556	25	0,535												
5									14	0,543	18	0,485	22	0,443	26	0,410												
6									15	0,631	19	0,477	23	0,441	27	0,411												
7									16	0,467	20	0,415	24	0,376	28	0,346												
8																	29	0,787	35	0,734	41	0,740	47	0,551	53	0,520	59	0,537
9																	30	0,848	36	0,559	42	0,723	48	0,495	54	0,462	60	0,465
10																	31	0,552	37	0,449	43	0,542	49	0,352	55	0,308	61	0,377
11																	32	0,610	38	0,628	44	0,633	50	0,475	56	0,471	62	0,498
12																	33	0,576	39	0,539	45	0,577	51	0,470	57	0,436	63	0,448
13																	34	0,554	40	0,569	46	0,517	52	0,399	58	0,382	64	0,412

[2] Une fois le modèle soumis, des erreurs d'alignement et d'opérateur ont été relevées. Les données incriminées présentées dans la présente annexe sont les mêmes que celles qui figurent dans le Rapport final des tests VQEG Phase II. Par conséquent, lorsque la méthode décrite dans la présente annexe est correctement implémentée, l'utilisateur peut obtenir des données objectives différentes de celles figurant dans le Tableau B.1.

Tableau B.2 – Matrice VQM à 625 lignes (données brutes)[3]

SRC (Image)	HRC=1		HRC=2		HRC=3		HRC=4		HRC=5		HRC=6		HRC=7		HRC=8		HRC=9		HRC=10	
1			4	0,612	10	0,531	16	0,452			29	0,434			42	0,436			52	0,382
2			5	0,544	11	0,540	17	0,451			30	0,437			43	0,440			53	0,363
3			6	0,572	12	0,571	18	0,497			31	0,479			44	0,478			54	0,418
4			7	0,601	13	0,656	19	0,557			32	0,547			45	0,526			55	0,472
5			8	0,603	14	0,621	20	0,500			33	0,492			46	0,444			56	0,390
6			9	0,591	15	0,520	21	0,483			34	0,469			47	0,461			57	0,423
7							22	0,576			35	0,555					48	0,531	58	0,501
8							23	0,512			36	0,500					49	0,482	59	0,457
9							24	0,507			37	0,487					50	0,468	60	0,436
10							25	0,610			38	0,594					51	0,575	61	0,540
11	1	0,753							26	0,594			39	0,508					62	0,485
12	2	0,643							27	0,556			40	0,550					63	0,496
13	3	0,669							28	0,524			41	0,481					64	0,441

[3] Une fois le modèle soumis, des erreurs d'alignement et d'opérateur ont été relevées. Les données incriminées présentées dans la présente annexe sont les mêmes que celles qui figurent dans le Rapport final des tests VQEG Phase II. Par conséquent, lorsque la méthode décrite dans la présente annexe est correctement implémentée, l'utilisateur peut obtenir des données objectives différentes de celles figurant dans le Tableau B.2.

Tableau B.3 – Matrice VQM à 525 lignes (données corrigées)[4]

SRC (Image)	HRC=1		HRC=2		HRC=3		HRC=4		HRC=5		HRC=6		HRC=7		HRC=8		HRC=9		HRC=10		HRC=11		HRC=12		HRC=13		HRC=14	
1	1	0,727	4	0,490	7	0,467	10	0,304																				
2	2	0,324	5	0,224	8	0,162	11	0,195																				
3	3	0,549	6	0,359	9	0,192	12	0,153																				
4									13	0,715	17	0,588	21	0,546	25	0,509												
5									14	0,523	18	0,418	22	0,344	26	0,289												
6									15	0,665	19	0,404	23	0,341	27	0,292												
7									16	0,386	20	0,298	24	0,239	28	0,199												
8																	29	0,823	35	0,783	41	0,788	47	0,537	53	0,481	59	0,512
9																	30	0,854	36	0,550	42	0,774	48	0,435	54	0,377	60	0,383
10																	31	0,539	37	0,354	43	0,520	49	0,206	55	0,156	61	0,241
11																	32	0,634	38	0,662	44	0,669	50	0,400	56	0,393	62	0,442
12																	33	0,580	39	0,515	45	0,581	51	0,390	57	0,332	63	0,353
13																	34	0,542	40	0,568	46	0,476	52	0,273	58	0,247	64	0,293

[4] Une fois le modèle soumis, des erreurs d'alignement et d'opérateur ont été relevées. Les données incriminées présentées dans la présente annexe sont les mêmes que celles qui figurent dans le Rapport final des tests VQEG Phase II. Par conséquent, lorsque la méthode décrite dans la présente annexe est correctement implémentée, l'utilisateur peut obtenir des données objectives différentes de celles figurant dans le Tableau B.3.

Tableau B.4 – Matrice VQM à 625 lignes (données corrigées)[5]

SRC (Image)	HRC=1		HRC=2		HRC=3		HRC=4		HRC=5		HRC=6		HRC=7		HRC=8		HRC=9		HRC=10	
1			4	0,625	10	0,429	16	0,204			29	0,164			42	0,169			52	0,082
2			5	0,467	11	0,454	17	0,202			30	0,170			43	0,177			53	0,062
3			6	0,542	12	0,538	18	0,327			31	0,275			44	0,272			54	0,134
4			7	0,605	13	0,686	19	0,502			32	0,475			45	0,414			55	0,255
5			8	0,609	14	0,641	20	0,335			33	0,312			46	0,185			56	0,091
6			9	0,586	15	0,395	21	0,284			34	0,248			47	0,226			57	0,143
7							22	0,551			35	0,496					48	0,430	58	0,339
8							23	0,371			36	0,335					49	0,283	59	0,217
9							24	0,356			37	0,298					50	0,243	60	0,168
10							25	0,623			38	0,590					51	0,549	61	0,455
11	1	0,741							26	0,592			39	0,359					62	0,290
12	2	0,672							27	0,499			40	0,482					63	0,322
13	3	0,698							28	0,406			41	0,279					64	0,179

[5] Une fois le modèle soumis, des erreurs d'alignement et d'opérateur ont été relevées. Les données incriminées présentées dans la présente annexe sont les mêmes que celles qui figurent dans le Rapport final des tests VQEG Phase II. Par conséquent, lorsque la méthode décrite dans la présente annexe est correctement implémentée, l'utilisateur peut obtenir des données objectives différentes de celles figurant dans le Tableau B.4.

Annexe C

Telecommunications Research and development Center (CPqD)

Description technique de l'évaluation d'image fondée sur la segmentation (IES)

C.1 Introduction

La présente annexe donne une description générale de l'algorithme d'évaluation d'image fondée sur la segmentation (IES, *image evaluation based on segmentation*) proposé par le Telecommunications Research and Development Center (CPqD). Cet algorithme constitue une méthode d'évaluation de la qualité vidéo par des mesures objectives de l'altération des images, calculées en zones de plans, de contours et de texture, selon un processus de segmentation des images. Il prévoit l'opinion d'un spectateur – note moyenne d'opinion (*MOS, mean opinion score*) et représente une alternative efficace et pertinente aux évaluations subjectives classiques actuellement employées [C-1] et [C-2] qui sont coûteuses et chronophages.

Le logiciel prototype CPqD-IES a été soumis aux tests phase II du VQEG [C-3]; les données objectives brutes mises à l'échelle qui ont été obtenues sont reproduites aux Tableaux C.2 et C.3.

C.2 Description générale du système IES

La Figure C.1 donne une description générale de l'application de l'algorithme CPqD-IES à des scènes naturelles. Chaque scène naturelle est représentée par une scène originale (de référence) ***O*** et une scène dégradée ***I*** qui résulte de l'application d'un codec à la scène ***O***. Des corrections de décalage et de gain sont appliquées à la scène ***I*** pour créer une scène dégradée corrigée ***I´***, de telle sorte que chaque trame *f* de ***I´*** correspond à la trame de référence *f* de ***O*** pour $f = 1, 2, ..., n$ (voir § C.3).

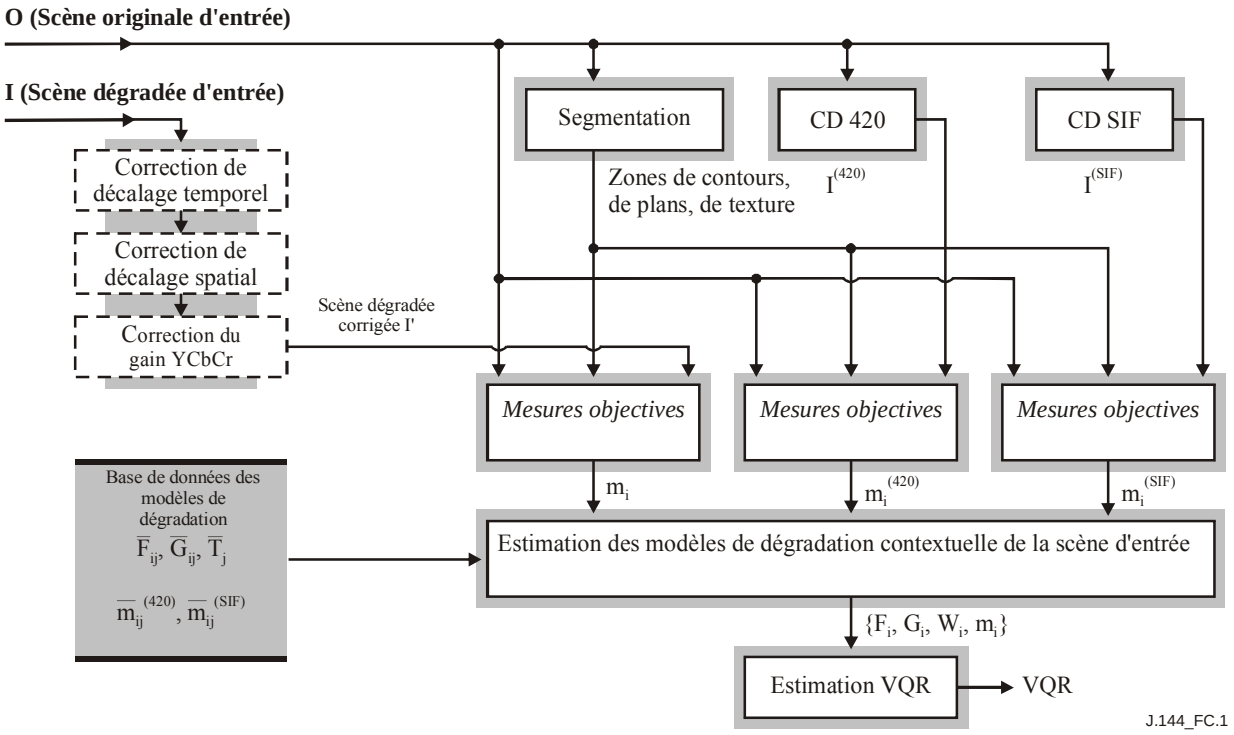


Figure C.1 – Description générale de l'application de l'algorithme CPqD-IES

Les scènes d'entrée ***I*** et ***O*** pour le système IES ont un format YCbCr4:2:2 conformément à la Rec. UIT-R BT.601-5 [C-4].

La composante Y de chaque trame d'image *f* de ***O*** est segmentée en trois catégories: texture, contours et plans (§ C.4). Une mesure objective est calculée à partir de la différence entre les trames correspondantes de ***O*** et ***I´*** pour chacune de ces zones et pour chaque composante d'image Y, Cb et Cr, formant ainsi un ensemble de 9 mesures objectives $\{m1, m2, ..., m9\}$ pour chaque trame d'image *f* (§ C.5). Pour chaque mesure objective m_i, $i = 1, 2, ..., 9$, on obtient un niveau de dégradation contextuelle L_i basé sur son modèle d'estimation de la dégradation qui est donné par:

$$L_i = 100 / \left[1 + \left(\frac{F_i}{m_i} \right)^{G_i} \right] \qquad \text{(C-1)}$$

où F_i et G_i sont deux paramètres calculés à partir d'une base de données de modèles de dégradation (§ C.7), de l'attribut spatial *S* et de l'attribut temporel *T* (§ C-5) et des mesures objectives $m_i^{(420)}$ et $m_i^{(SIF)}$ pour la trame *f* résultant des opérations des codecs CD420 et CDSIF appliqués à ***O*** (§ C.7). Les deux codecs d'altération de référence, CD420 (codeur/décodeur MPEG-2 4:2:0) et CDSIF (codeur/décodeur MPEG-1 SIF) sont entièrement fondés sur les routines tirées directement de MPEG2 [C-5] et MPEG1 [C-6], disponibles sur le site www.mpeg.org/MPEG/MSSG. Dans la version actuellement utilisée de l'algorithme CPqD-IES, ces routines fonctionnent en mode intra avec un pas de quantification fixe égal à 16. Il importe de noter que les codecs CD420 et CDSIF n'introduisent pas de différence de décalage ou de gain par rapport à ***O***.

L'indice de qualité vidéo VQR_f de la trame *f* est obtenu par combinaison linéaire des niveaux de dégradation contextuelle L_i, $i = 1, 2, ..., 9$, comme suit:

$$VQR_f = \sum_{i=1}^{9} W_i . L_i \qquad \text{(C-2)}$$

où W_i est le facteur de pondération du niveau de dégradation L_i pour cette scène naturelle particulière, lequel est calculé tel qu'indiqué au § C.7.

Maintenant, la séquence de valeurs VQR_1, VQR_2, ... VQR_n est transformée par un filtre médian de taille 3 en une autre séquence $VQR´_1$, $VQR´_2$, ... $VQR´_n$ en ne calculant pas la valeur médiane dans le voisinage immédiat de VQR_1 et VQR_n. Pendant le filtrage médian, l'algorithme évite la répétition de deux valeurs médianes consécutives, c'est-à-dire que si la valeur médiane $VQR´_{f-1}$ calculée à 1 unité près de VQR_f est égale à la valeur médiane $VQR´_{f-2}$ calculée dans le voisinage immédiat de $VQR´_{f-1}$, l'algorithme choisit $VQR´_{f-1}$ comme valeur minimale calculée entre VQR_{f-1}, VQR_f, et VQR_{f+1}. Cet algorithme peut être décrit comme suit:

1) pour chaque ***f*** compris entre 2 et ***n – 1***;
2) calculer *med*, la valeur médiane entre VQR_{f-1}, VQR_f, VQR_{f+1};
3) si $med = VQR´_{f-2}$ alors
4) calculer $VQR´_{f-1}$ comme la valeur minimale entre VQR_{f-1}, VQR_f, VQR_{f+1};
5) sinon
6) $VQR´_{f-1} \leftarrow med$.

L'indice de qualité vidéo final *VQR* est alors la moyenne des valeurs $VQR_f´$.

$$VQR = \frac{1}{n-2} . \sum_{f=1}^{n-2} VQR´_f \qquad \text{(C-3)}$$

Les équations C-1 et C-2 et l'algorithme ci-dessus décrivent le processus permettant d'estimer la valeur *VQR* à partir des modèles de dégradation contextuelle $\{F_i, G_i, W_i\}$ et des mesures

objectives m_i, i = 1, 2, ..., 9. Les paragraphes suivants terminent la description de la méthode en présentant les détails à l'intérieur des blocs restants de la Figure C.1.

C.3 Correction du décalage et du gain

C.3.1 Décalage temporel

Le décalage temporel *dt* est un entier compris entre –2 et 2. Les valeurs positives de *dt* signifient que la scène dégradée ***I*** est retardée par rapport à scène originale ***O***, tandis que les valeurs négatives ont une signification inverse, la scène originale ***O*** étant alors retardée par rapport à la scène dégradée ***I***. Les scènes d'entrée présentant des décalages temporaires situés en dehors de cette fourchette ne sont pas prises en considération. Supposons que $\boldsymbol{I_{dt}}$ est la scène dégradée ***I*** avec un déplacement de *dt* trames. Un coefficient de dissemblance est calculé entre la scène originale ***O*** et chaque scène déplacée $\boldsymbol{I_{dt}}$. Le déplacement comportant le plus petit coefficient de dissemblance est utilisé comme décalage temporel et le résultat $\boldsymbol{I_{dt}}$ est ensuite la scène dégradée ***I*** déplacée de ce décalage en vue du prochain calcul. Le coefficient de dissemblance entre la scène ***O*** et la scène $\boldsymbol{I_{dt}}$ est obtenu comme suit, où *n* est le nombre de trames situées dans l'intersection temporelle entre elles:

1) $\xi_T \leftarrow 0$;
2) pour chaque *f* de 1 à *n*;
3) calculer S_b;
4) calculer S_b';
5) calculer D_b;
6) calculer μ, valeur moyenne des pixels en *Db*;
7) $\xi_T \leftarrow \xi_T + (\mu/n)$;
8) retour ξ_T (coefficient de dissemblance entre ***O*** et $\boldsymbol{I_{dt}}$).

où:

S_b = amplitude du gradient de Sobel [C-7] de la composante *Y* de la *f*ième trame de ***O***;

S_b' = amplitude du gradient de Sobel de la composante *Y* de la *f*ième trame de $\boldsymbol{I_{dt}}$;

D_b = la différence absolue au niveau des pixels entre S_b et S_b'.

C.3.2 Décalage spatial

Le décalage spatial (d_x, d_y) est l'un des déplacements horizontaux et verticaux suivants d_x = –6, –5, ... 6 et d_y = –6, –5, ..., 6. Les valeurs négatives de déplacement signifient qu'une trame de la scène dégradée ***I*** est décalée par rapport à une trame de la scène originale ***O,*** vers la gauche et vers le haut. Les valeurs positives de déplacement signifient qu'une trame de la scène dégradée ***I*** est décalée par rapport à une trame de la scène originale ***O,*** vers la droite et vers le bas. Les scènes d'entrée présentant des décalages spatiaux situés en dehors de cette fourchette ne sont pas prises en considération.

Soit $\boldsymbol{I_{dx,\ dy}}$ scène dégradée $\boldsymbol{I_{dt}}$ avec toutes les trames déplacées de (d_x, d_y) pixels. On calcule un coefficient de dissemblance entre ***O*** et $\boldsymbol{I_{dx,\ dy}}$. Le déplacement spatial présentant la plus faible dissemblance est utilisé comme décalage spatial et le résultat $\boldsymbol{I_{dx,\ dy}}$ est alors $\boldsymbol{I_{dt}}$ déplacé de ce décalage, utilisé pour la correction de gain.

La dissemblance entre ***O*** et $\boldsymbol{I_{dx,\ dy}}$ est décrite ci-après:

1) $\xi_S \leftarrow 0$; $c \leftarrow 0$;
2) pour chaque *f* de 1 à *n*;
3) pour *x* de *x*0 à (*x*0 + *w*/4);

4) pour y de $y0$ à $(y0 + h/4)$;

5) $\xi_S \leftarrow \xi_S + |Y(4x,4y) - Y'(4x + dx, 4y + dy)| +$
$+ |Cb(4x,4y) - Cb'(4x + dx, 4y + dy)| +$
$+ |Cr(4x,4y) - Cr'(4x + dx, 4y + dy)|$;

6) $c \leftarrow c + 3$;

7) $\xi_S \leftarrow \xi_S / c$;

8) retour ξ_S (coefficient de dissemblance entre $\boldsymbol{O}$ et $\boldsymbol{I}_{dx, dy}$);

où:

$w \times h =$	dimensions (colonnes x lignes) de la zone d'intersection entre $\boldsymbol{O}$ et $\boldsymbol{I}_{dx, dy}$;
$Y(x, y)$, $Cb(x, y)$, $Cr(x, y)$	sont les valeurs dans les composantes d'image d'une trame f de $\boldsymbol{O}$ pour un pixel (x, y);
$Y'(x + dx, y + dy)$, $Cb'(x + dx, y + dy)$, $Cr'(x + dx, y + dy)$	sont les valeurs dans les composantes d'image d'une trame f de $\boldsymbol{I}_{dx, dy}$ pour un pixel $(x + dx, y + dy)$.

C.3.3 Gain

Le gain d'amplitude entre $\boldsymbol{O}$ et $\boldsymbol{I}_{dx, dy}$ est calculé séparément pour chaque composante d'image *Y*, *Cb* et *Cr*. L'algorithme calcule la moyenne des gains sur l'ensemble des n trames et corrige chaque composante d'image en conséquence. Le résultat $\boldsymbol{I}'$ est la scène dégradée qui est utilisée pour tous les calculs ultérieurs. Le gain d'amplitude entre une composante d'image C' de la trame f de $\boldsymbol{I}_{dx, dy}$ par rapport à la même composante C de la trame f de $\boldsymbol{O}$ est obtenu en floutant les deux images C' et C au moyen d'un filtre gaussien [C-7] de noyau

$$\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}$$

et en calculant le rapport entre la somme de leurs valeurs de pixels dans les images floutées. Un seul de chacun des 16 pixels est pris en considération (en balayant les images à composante floutées par incréments horizontal et vertical de 4 pixels, comme dans l'algorithme de calcul de ξ_s présenté au § C.3.2).

C.4 Segmentation de l'image

Au départ, l'algorithme de segmentation classe chaque pixel de la composante *Y* d'une trame donnée f de la scène originale $\boldsymbol{O}$ dans la région des plans et ou une autre région. L'algorithme applique également à *Y* un détecteur de contours et la région des contours est définie par les contours qui sont situés dans les limites de la région du plan. La région de texture est composée par les pixels restants de l'image *Y* (voir Figure C.2).

La segmentation est calculée sur chaque trame de la composante *Y* à partir de la scène originale d'entrée $\boldsymbol{O}$. Pour les composantes *Cb* et *Cr*, les régions segmentées sont obtenues en suréchantillonnant la position des pixels de la composante *Y* par un facteur 2 dans le sens horizontal.

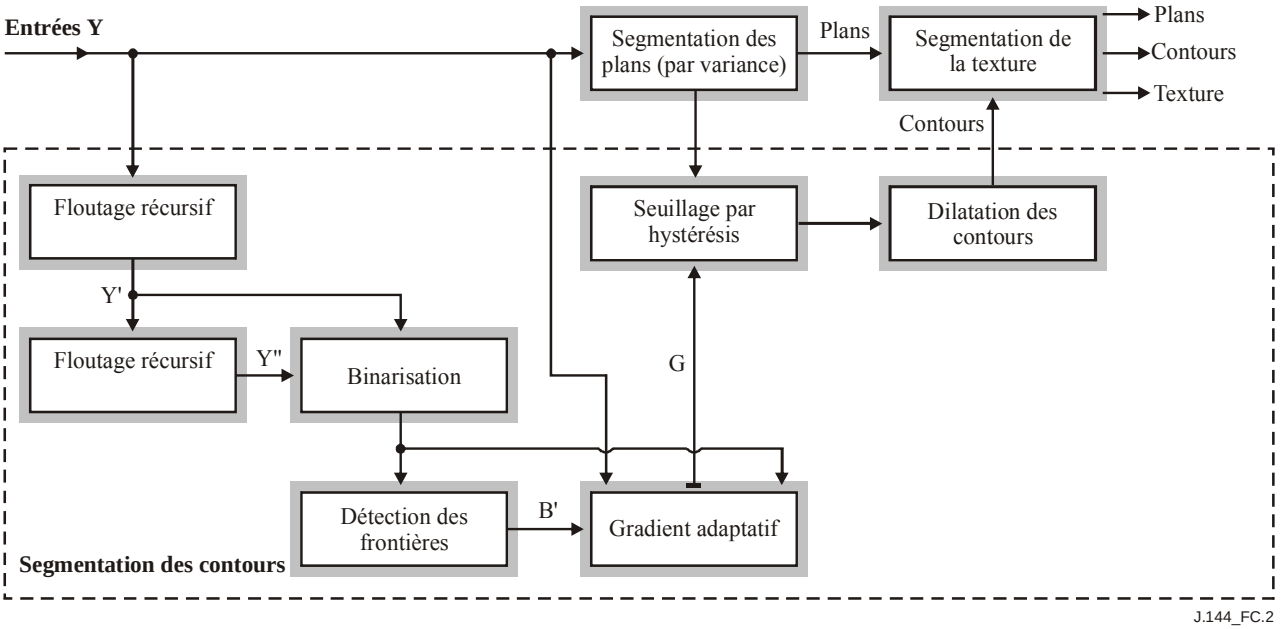


Figure C.2 – Schéma fonctionnel du processus de segmentation

C.4.1 Régions du plan

La variance de la brillance de chaque pixel de la composante *Y* est calculée dans un voisinage de 5 × 5 pixels de part et d'autre du pixel considéré. Une opération de seuillage est appliquée à la variance de l'image de sorte que les pixels présentant une valeur de variance inférieure à 25^2 sont classés comme appartenant à la région du plan. Il résulte de ce processus que de petites composantes de pixels sont classées, à tort, dans la zone de texture. Un filtre médian de 3 × 3 est appliqué pour supprimer ces petites composantes. Enfin, l'image binaire de la région du plan est dilatée par un élément structurant circulaire d'un diamètre de 11 pixels [C-7]. Cette opération correspond à l'application de la relation suivante à une image binaire d'entrée *A*, afin d'obtenir une image binaire agrandie *A''*

$$A'(x,y) = \max\{A(x',y')\} \ \textit{for all pixels}\ (x',y') \in N_{81}(x,y) \qquad \text{(C-4)}$$

où $N_{81}(x,y)$ désigne l'ensemble des 81 pixels voisins immédiats du pixel *(x, y).*

C.4.2 Régions des contours

Un filtrage récursif est appliqué à *Y*, créant une première image floue *Y´*, puis à *Y´* pour créer une deuxième image floue *Y´´*. Chaque filtrage récursif se compose de quatre grilles appliquées à l'image d'entrée. Cet algorithme est décrit ci-après pour la composante d'image *Y* de la seule trame de la scène d'entrée ***O***.

1) *Pour y* variant de 0 à $(h - 1)$
2) *Pour x* variant de 0 à $(w - 2)$
3) $Y(x + 1, y) \leftarrow Y(x, y) + 0{,}7.[Y(x + 1, y) - Y(x, y)]$;
4) *Pour y* variant de 0 à $(h - 1)$
5) *Pour x* variant de $(w - 1)$ à 1
6) $Y(x - 1, y) \leftarrow Y(x, y) + 0{,}7.[Y(x - 1, y) - Y(x, y)]$;
7) *Pour x* variant de 0 à $(w - 1)$
8) *Pour y* variant de 0 à $(h - 2)$
9) $Y(x, y + 1) \leftarrow Y(x, y) + 0{,}7.[Y(x, y + 1) - Y(x, y)]$;

10) *Pour x* variant de 0 à (w – 1)

11) *Pour y* variant de (h – 1) à 1

12) $Y(x, y-1) \leftarrow Y(x, y) + 0.7.[Y(x, y-1) - Y(x, y)]$;

13) Sauvegarder image Y en image Y'

où:

$Y(x, y)$ = brillance du pixel (x, y)

h = nombre de lignes de Y

w = nombre de colonnes de Y.

La seconde application de l'algorithme créera Y''. Une image binaire B est créée à partir de Y' et Y'':

$$B(x,y)=\begin{cases}1, & \text{si } Y'(x,y)\geq Y''(x,y),\\ 0, & \text{sinon}\end{cases} \tag{C.5}$$

Après cela, l'algorithme identifie les pixels frontières des zones de B ayant une valeur de pixel de 1 en créant une seconde image binaire B':

$$B'(x,y)=\begin{cases}1, & \text{si } B(x,y)=1 \text{ et } B(x',y')=0 \text{ pour tout pixel } (x',y')\in N_9(x,y)\\ 0, & \text{sinon}\end{cases} \tag{C-6}$$

où $N_9(x, y)$ est l'ensemble des 9 pixels voisins immédiats de (x, y).

Un filtre à gradient adaptatif est appliqué à Y limité aux pixels où $B'(x, y)=1$:

$$G(x,y)=\begin{cases}|\mu_1-\mu_0|, & \text{si } B'(x,y)=1,\\ 0, & \text{sinon,}\end{cases} \tag{C.7}$$

où:

μ_1 = valeur moyenne de $Y(x',y')$, pour tout $(x',y')\in N_9(x,y)$ de sorte que $B(x',y')=1$

μ_0 = valeur moyenne de $Y(x',y')$, pour tout $(x',y')\in N_9(x,y)$ de sorte que $B(x',y')=0$.

A noter que l'algorithme utilise B en lieu et place de B' pour calculer les valeurs moyennes μ_1 et μ_0.

Un seuillage par hystérésis [C-8] est appliqué à G limité aux pixels dont on a établi au § C.4.1 qu'ils appartiennent à la zone des plans. Le seuil inférieur est de 30 et le seuil supérieur de 40. L'algorithme identifie tout d'abord comme appartenant à la zone des contours tous les pixels (x, y) de G, de sorte que $G(x, y) > 40$, puis applique un algorithme à zone croissante le long des lignes de G en utilisant ces pixels comme éléments de départ et en restreignant la croissance aux pixels appartenant à la même ligne pour lesquels $G(x, y) > 30$. Toutes les composantes 4 connexes comportant moins de 6 pixels sont éliminées de ce résultat. L'image binaire finale est dilatée par un élément structurant circulaire d'un diamètre de 5 pixels, c'est-à-dire utilisant l'ensemble $N_{13}(x, y)$ des 13 pixels voisins immédiats du pixel (x, y), comme d'après l'équation C-4, qui ignore la restriction à la zone des plans. Les pixels de valeur 1 dans cette dilatation sont classés comme appartenant à la zone des contours.

C.4.3 Régions de texture

La région de texture se compose des pixels de Y qui ont été classés comme n'appartenant ni à la région des contours ni à la région du plan.

C.5 Mesures objectives

Soit S_b l'image d'amplitude du gradient de Sobel [C-7] calculée pour une composante donnée (*Y, Cb ou Cr*) d'une trame donnée *f* de la scène originale ***O***, et S'_b, l'image d'amplitude du gradient de Sobel pour la même composante de la trame *f* de la scène dégradée ***I´***. L'image D_b de la différence absolue au niveau des pixels entre S_b et S'_b est calculée et la zone R de pixels de l'image D_b qui appartient à un contexte donné (plan, contours ou texture) est prise en considération. La différence absolue de Sobel (ASD, *absolute Sobel difference*) pour cette composante d'image et ce contexte est définie comme étant la moyenne des valeurs de pixels de l'image D_b restreinte à $\Re$.

Cette procédure donne un ensemble de neuf mesures objectives $\{m_1, m_2, ..., m_9\}$ pour chaque trame d'image *f, f* = 1, 2, ..., *n*, tenant compte de l'ensemble des trois contextes et des trois composantes d'image.

Le même processus est appliqué pour créer des mesures objectives $\{m_1^{(420)}, m_2^{(420)}, ..., m_9^{(420)}\}$ et $\{m_1^{(SIF)}, m_2^{(SIF)}, ..., m_9^{(SIF)}\}$, pour la trame *f*, avec un fonctionnement des codecs MPEG-2 4:2:0 et MPEG-1 SIF sur la scène ***O*** (Figure C.1). Ces mesures servent de références avec les attributs spatial *S* et temporel *T* pour déterminer le modèle de dégradation contextuelle pour ***I´*** (§ C.7). L'attribut temporel *T* est la valeur moyenne de la différence absolue au niveau des pixels entre les segmentations des trames *f* et *f–1*, normalisée dans l'intervalle [0, 1]. L'attribut spatial *S* est défini comme étant le rapport $m_7^{(SIF)}/m_7^{(420)}$, normalisé dans l'intervalle [0,1], où $m_7^{(SIF)}$ et $m_7^{(420)}$ sont les différences ASD correspondantes pour la zone plans de la composante *Y* de la trame *f*.

C.6 Base de données des modèles de dégradation

L'algorithme CPqD-IES utilise une base de données de modèles de dégradation pour des scènes différentes de la scène de référence ***O*** pour évaluer l'indice de qualité vidéo de ***I´***. Cette base de données regroupe des informations sur douze scènes à 60 Hz illustrant divers degrés de mouvement (scènes dynamiques ou statiques), de nature (scènes réelles ou scènes synthétiques), et de contexte (quantité de pixels de texture, de plan ou de contour). Cette base de données a été créée comme suit.

Les valeurs moyennes des mesures objectives, $\left\{\overline{m}_{1,j}^{(420)}, \overline{m}_{2,j}^{(420)}, ..., \overline{m}_{9,j}^{(420)}\right\}$, et $\left\{\overline{m}_{1,j}^{(SIF)}, \overline{m}_{2,j}^{(SIF)}, ..., \overline{m}_{9,j}^{(SIF)}\right\}$ ont été calculées pour les trames de chaque scène *j, j* = 1, 2, ..., 12.

Les valeurs de $\overline{T}_j = \{27.01, 25.33, 45.54, 36.40, 32.02, 12.63, 28.38, 10.19, 0.01, 7.26, 7.60, 14.27\}$ ont été calculées comme étant la moyenne des attributs temporels (voir le § C.5) pour chaque trame de la portion finale de chaque scène *j*.

Toutes les scènes dégradées de la base de données ont elles aussi fait l'objet d'une évaluation subjective, ce qui donne un niveau de dégradation subjective SL_j, normalisé dans l'intervalle entre [0% et 100%] pour chaque scène *j*.

Chaque mesure objective $\overline{m}_{i,j}$, i = 1, 2, ..., 9 et j = 1, 2, ..., 12, est rattachée à un niveau de dégradation contextuelle $\overline{L}_{i,j}$, selon l'équation C-1. Les valeurs de $F_{i,j}$ et $G_{i,j}$ dans l'équation C-1 ont été calculées pour chaque scène *j* en minimisant l'espérance de l'erreur quadratique moyenne $E\left[\left(\overline{SL}_j - \overline{L}_{i,j}\right)^2\right]$.

A l'issue du processus, la base de données des modèles de dégradation comprend 5 ensembles $\overline{F}_{i,j}, \overline{G}_{i,j}, \overline{T}_j, \overline{m}_{i,j}^{(420)}, \overline{m}_{i,j}^{(SIF)}$, i = 1, 2, ..., 9 de paramètres pour chaque scène *j, j* = 1, 2, ..., 12.

Le Tableau C.1 contient les valeurs de $\overline{F}_{i,j}, \overline{G}_{i,j}, \overline{m}_{i,j}^{(420)}, \overline{m}_{i,j}^{(SIF)}$, où Y, Cb et Cr désignent les composantes d'une trame et les suffixes P, E et T correspondent respectivement aux zones de plan, de contour et de texture.

Tableau C.1 – Mesures de dégradation concernant 12 scènes de la base de données: $\overline{m}_{i,j}^{(420)}$, $\overline{m}_{i,j}^{(SIF)}$, $\overline{F}_{i,j}$, $\overline{G}_{i,j}$

Scène j	$\overline{m}_{1,j}^{(420)}$ **YP**	$\overline{m}_{2,j}^{(420)}$ **CbP**	$\overline{m}_{3,j}^{(420)}$ **CrP**	$\overline{m}_{4,j}^{(420)}$ **YE**	$\overline{m}_{5,j}^{(420)}$ **CbE**	$\overline{m}_{6,j}^{(420)}$ **CrE**	$\overline{m}_{7,j}^{(420)}$ **YT**	$\overline{m}_{8,j}^{(420)}$ **CbT**	$\overline{m}_{9,j}^{(420)}$ **CrT**
1	10,89	7,08	7,41	22,73	24,21	24,68	22,93	20,94	19,25
2	11,69	5,82	5,12	20,06	15,80	12,22	21,55	12,41	9,76
3	8,14	5,36	4,32	13,77	12,54	11,21	14,43	10,89	10,43
4	14,18	6,04	5,44	21,89	15,57	12,19	21,50	12,55	10,80
5	6,87	5,44	4,50	19,42	17,24	14,40	20,36	19,18	16,98
6	8,96	4,31	4,17	16,67	8,08	10,56	15,58	6,38	6,15
7	14,25	8,10	6,99	22,69	17,65	18,62	21,66	16,80	16,57
8	7,06	3,36	3,92	21,40	17,45	22,31	22,44	30,12	22,21
9	8,80	6,61	7,19	20,65	17,24	13,81	21,02	15,94	12,51
10	16,04	15,92	10,28	19,58	20,93	12,59	19,57	25,38	14,39
11	6,70	4,41	4,98	17,48	11,97	13,50	17,42	15,27	16,44
12	12,10	7,09	8,14	21,49	25,22	22,38	21,58	19,83	19,18
Scène j	$\overline{m}_{1,j}^{(SIF)}$ **YP**	$\overline{m}_{2,j}^{(SIF)}$ **CbP**	$\overline{m}_{3,j}^{(SIF)}$ **CrP**	$\overline{m}_{4,j}^{(SIF)}$ **YE**	$\overline{m}_{5,j}^{(SIF)}$ **CbE**	$\overline{m}_{6,j}^{(SIF)}$ **CrE**	$\overline{m}_{7,j}^{(SIF)}$ **YT**	$\overline{m}_{8,j}^{(SIF)}$ **CbT**	$\overline{m}_{9,j}^{(SIF)}$ **CrT**
1	21,65	9,88	10,97	83,35	35,03	35,40	72,68	28,46	27,22
2	17,85	6,97	5,90	68,56	20,50	15,08	53,11	15,88	11,67
3	12,57	7,05	5,20	32,24	17,38	14,42	32,80	13,54	12,48
4	21,77	7,17	6,20	75,52	20,10	15,20	61,95	15,88	13,17
5	11,79	6,07	5,24	88,69	22,84	19,79	84,11	25,62	23,44
6	11,85	4,63	4,49	43,21	9,83	13,09	26,98	7,04	6,79
7	21,35	8,62	8,13	110,79	20,89	23,41	88,45	20,36	21,23
8	9,97	3,82	4,48	70,29	22,30	30,25	72,19	40,05	28,45
9	18,27	7,98	8,20	61,46	21,79	16,82	54,04	19,81	15,33
10	27,18	22,33	12,85	42,67	30,69	15,35	39,09	36,27	19,24
11	8,66	4,70	5,73	51,38	14,96	18,01	42,91	19,17	21,01
12	13,99	10,33	11,11	76,35	43,82	35,30	62,17	31,69	31,35

Tableau C.1 – Mesures de dégradation concernant 12 scènes de la base de données: $\overline{m}_{i,j}^{(420)}$, $\overline{m}_{i,j}^{(SIF)}$, $\overline{F}_{i,j}$, $\overline{G}_{i,j}$

Scène j	$\overline{F}_{1,j}$ **YP**	$\overline{F}_{2,j}$ **CbP**	$\overline{F}_{3,j}$ **CrP**	$\overline{F}_{4,j}$ **YE**	$\overline{F}_{5,j}$ **CbE**	$\overline{F}_{6,j}$ **CrE**	$\overline{F}_{7,j}$ **YT**	$\overline{F}_{8,j}$ **CbT**	$\overline{F}_{9,j}$ **CrT**
1	19,67	9,60	10,34	64,69	35,11	34,99	62,74	30,48	28,92
2	16,80	7,02	5,80	50,94	20,85	15,31	50,94	20,85	15,31
3	16,25	8,43	5,97	48,49	19,54	15,89	50,17	15,57	13,89
4	20,59	7,04	6,01	52,95	20,70	15,58	49,29	16,91	13,90
5	10,64	6,03	5,39	58,91	23,39	19,79	60,79	27,51	24,66
6	11,01	4,48	4,36	29,93	9,27	12,43	24,51	6,84	6,58
7	20,56	8,49	7,91	69,41	20,68	23,91	60,70	20,06	22,14
8	10,18	3,88	4,52	58,20	22,37	30,64	61,92	43,33	31,30
9	24,49	8,92	9,02	70,80	24,17	19,21	63,14	23,06	18,00
10	22,55	20,91	12,45	32,29	29,62	15,01	32,45	36,43	19,03
11	8,03	4,68	5,61	32,73	14,48	16,94	33,55	19,15	20,70
12	13,04	9,30	10,01	44,95	40,64	32,98	45,45	30,93	30,62
Scène j	$\overline{G}_{1,j}$ **YP**	$\overline{G}_{2,j}$ **CbP**	$\overline{G}_{3,j}$ **CrP**	$\overline{G}_{4,j}$ **YE**	$\overline{G}_{5,j}$ **CbE**	$\overline{G}_{6,j}$ **CrE**	$\overline{G}_{7,j}$ **YT**	$\overline{G}_{8,j}$ **CbT**	$\overline{G}_{9,j}$ **CrT**
1	1,85	4,17	3,80	1,27	4,38	4,61	1,63	5,36	4,99
2	3,52	8,17	10,04	1,50	8,26	8,80	2,36	8,63	10,54
3	3,69	6,52	9,67	2,09	6,73	7,65	2,35	8,25	9,32
4	2,84	10,70	8,95	1,15	5,56	5,24	1,31	5,23	5,06
5	5,25	14,15	12,98	2,00	10,02	8,41	3,37	13,05	12,69
6	4,07	19,74	11,09	1,26	10,68	7,10	1,63	11,86	7,54
7	4,42	8,98	8,96	1,81	9,17	6,63	2,08	7,32	6,29
8	2,19	8,71	10,86	1,69	8,74	8,73	2,64	8,97	9,20
9	3,11	6,45	7,26	2,33	5,69	8,05	2,69	5,66	5,83
10	6,49	15,92	10,79	2,02	9,03	8,06	2,57	7,88	7,96
11	5,50	4,71	5,78	1,50	2,27	3,80	1,78	3,70	3,86
12	13,04	9,30	10,01	44,95	40,64	32,98	45,45	30,93	30,62

C.7 Estimation des modèles de dégradation

Les modèles de dégradation contextuels pour une trame f de $\boldsymbol{I}'$ se composent des paramètres $\{F_i, G_i, W_i\}$ des équations C-1 et C-2, i = 1, 2, ..., 9. Le présent paragraphe décrit comment calculer ces paramètres en utilisant les scènes dégradées $I^{(420)}$ et $I^{(SIF)}$ comme référence.

C.7.1 Calcul de W_i

Les distances locales contextuelles $D_{i,j}$ entre une trame f des scènes dégradées $I^{(420)}$ et $I^{(SIF)}$, et chaque scène j de la base de données sont définies comme suit:

$$D_{i,j} = \frac{1}{2}.\left(\left|L_{i,j}^{(420)} - \overline{L}_{i,j}^{(420)}\right| + \left|L_{i,j}^{(SIF)} - \overline{L}_{i,j}^{(SIF)}\right|\right) \qquad \text{(C-8)}$$

où:

$$\begin{cases} \overline{L}_{i,j}^{(420)} = 100 / \left[1 + \left(\overline{F}_{i,j} / \overline{m}_i^{(420)} \right)^{\overline{G}_{i,j}} \right] \\ \overline{L}_{i,j}^{(SIF)} = 100 / \left[1 + \left(\overline{F}_{i,j} / \overline{m}_i^{(SIF)} \right)^{\overline{G}_{i,j}} \right] \\ L_{i,j}^{(420)} = 100 / \left[1 + \left(\overline{F}_{i,j} / m_i^{(420)} \right)^{\overline{G}_{i,j}} \right] \\ L_{i,j}^{(SIF)} = 100 / \left[1 + \left(\overline{F}_{i,j} / m_i^{(SIF)} \right)^{\overline{G}_{i,j}} \right] \end{cases} \quad \text{(C-9)}$$

L'algorithme trouve l'ensemble Ω des six scènes les plus proches de la base de données sur la base de la distance $D_{i,j}$ et définit $W_{i,j}$ comme:

$$a_k = \begin{cases} 1, & \text{si (scene } k) \in \Omega, \\ 0, & \text{sinon.} \end{cases} \quad \text{(C-10)}$$

$$W_{i,j} = \frac{a_j . D_{i,j}^{-1}}{\sum_{k=1}^{12} a_k . D_{i,k}^{-1}} \quad \text{(C-11)}$$

Soit i = {1, 2, ..., 9} ≡ {(*plane, Y*), (*plane, Cb*), (*plane, Cr*), (*edge, Y*), (*edge, Cb*), (*edge, Cr*), (*texture, Y*), (*texture, Cb*), (*texture, Cr*)}, où (*plane, C*), (*edge, C*) et (*texture, C*) représentent les régions texture, contours et plan de la composante d'image *C, C = Y, Cb, Cr*.

Soit *u = texture*, *edge*, *plane* et *v = Y*, *Cb*, *Cr*, les valeurs W_i, i = 1, 2, ..., 9, sont calculées comme suit:

$$E_i = \sum_{j=1}^{12} D_{i,j} . W_{i,j}$$

$$\kappa_{u,v} = \begin{cases} 1 & \text{si } v = Y_i, \\ 1/2 & \text{sinon} \end{cases}$$

$$\tau = \sum_u \left[\frac{1}{E_{u,Y}} + \frac{1}{2} . \left(\frac{1}{E_{u,Cb}} + \frac{1}{E_{u,Cr}} \right) \right]$$

$$W_i = \frac{\kappa_i}{\tau} . \frac{1}{E_i} \quad \text{(C-12)}$$

C.7.2 Calcul de F_i et G_i

Les niveaux de dégradation contextuels $L_i^{(420)}$ et $L_i^{(SIF)}$ de la trame *f* pour *CD420* et *CDSIF* sont calculés comme suit:

$$L_i^{(420)} = \frac{1}{\gamma} . \sum_{j=1}^{12} W_{i,j} . L_{i,j}^{(420)} \quad \text{(C-13)}$$

$$L_i^{(SIF)} = \frac{1}{\gamma} . \sum_{j=1}^{12} W_{i,j} . L_{i,j}^{(SIF)} \quad \text{(C-14)}$$

où γ est un facteur limité à [1/2, 2], qui est calculé à partir des distances vectorielles D_j entre les attributs spatial et temporel, (S_j, T_j) et $(\bar{S}_j, \bar{T}_j)$, de la scène d'entrée et de chaque scène de la base de données, respectivement. Les attributs spatiaux $\bar{S}_j$ de la base de données de modèles de dégradation sont calculés (voir § C.5) directement d'après $\bar{}_{mi,j}(420)$, $\bar{}_{mi,j}(SIF)$, i=1, 2,..., 9 et j=1, 2,..., 12.

$$D_j = \left(S - \bar{S}_j\right)^2 + \left(T - \bar{T}_j\right)^2 \tag{C-15}$$

$$w_j = \frac{D_j^{-1}}{\sum_{k=1}^{12} D_k^{-1}}$$

$$a = \sum_{j=1}^{12} w_j \cdot \left[\frac{\bar{S}_j \cdot \bar{T}_j}{2} + (1 - \bar{T}_j^2) \cdot \left(1 - \frac{\bar{S}_j^2}{2}\right) \right] \tag{C-16}$$

$$b = \frac{S \cdot T}{2} + (1 - T^2) \cdot \left(1 - \frac{S^2}{2}\right)$$

$$\gamma = 1 + a - b$$

Les paramètres F_i et G_i sont enfin obtenus en résolvant le système d'équations ci-après:

$$L_i^{(420)} = 100 / \left[1 + \left(\frac{F_i}{m_i^{(420)}} \right)^{G_i} \right] \tag{C-17}$$

$$L_i^{(SIF)} = 100 / \left[1 + \left(\frac{F_i}{m_i^{(SIF)}} \right)^{G_i} \right] \tag{C-18}$$

C.8 Références

[C-1] Recommandation UIT-R BT.500-11 (2002), *Méthodologie d'évaluation subjective de la qualité des images de télévision*.

[C-2] Recommandation UIT-R BT.802-1 (1994), *Images et séquences d'essai pour l'évaluation subjective des codecs numériques véhiculant des signaux produits conformément à la Recommandation UIT-R BT.601*.

[C-3] Document de référence UIT-T (2004), *Objective perceptual assessment of video quality: Full reference television*.

[C-4] Recommandation UIT-R BT.601-5 (1995), *Paramètres de codage en studio de la télévision numérique pour des formats standards d'image 4:3 (normalisé) et 16:9 (écran panoramique)*.

[C-5] Recommandation UIT-T H.262 (2000), *Technologies de l'information – Codage générique des images animées et du son associé: données vidéo*.

[C-6] ISO/CEI 11172-1:1993, *Technologies de l'information – Codage de l'image animée et du son associé pour les supports de stockage numérique jusqu'à environ 1,5 Mbit/s – Partie 1: Systèmes*.

[C-7] Gonzalez, R.C. and Woods, R.E. (1992), *Digital Image Processing*, Addison-Wesley.

[C-8] Trucco, E. and Verri A. (1998), *Introductory Techniques for 3-D Computer Vision*, Prentice-Hall.

C.9 Résultats objectifs des essais, Phase II du VQEG

Tableau C.2 – Matrice de données objectives brutes 625/60

SRC	HRC									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1		0,6343	0,5083	0,287		0,2461		0,1951		0,1548
2		0,5483	0,5966	0,3649		0,3185		0,2668		0,1597
3		0,5998	0,6299	0,4551		0,3927		0,3428		0,2553
4		0,6055	0,8159	0,5684		0,5397		0,4158		0,309
5		0,6483	0,7268	0,4358		0,418		0,2874		0,1898
6		0,6146	0,4908	0,3671		0,3139		0,2562		0,2107
7				0,5865		0,5536			0,4841	0,3917
8				0,5023		0,457			0,3949	0,3158
9				0,4563		0,3927			0,3399	0,2667
10				0,7036		0,6511			0,6025	0,5083
11	0,8124				0,6374		0,3205			0,3221
12	0,7015				0,547		0,4997			0,3922
13	0,709	0,5098					0,4199			0,3298

Tableau C.3 – Matrice de données objectives brutes 525/60

SRC	HRC													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	0,5472	0,3698	0,3429	0,1918										
2	0,5075	0,226	0,1028	0,0789										
3	0,3549	0,127	0,058	0,0339										
4					0,6062	0,419	0,36	0,3108						
5					0,4444	0,2957	0,2152	0,1635						
6					0,6098*	0,3462	0,2546	0,1967						
7					0,2404	0,135	0,0864	0,0609						
8									0,8666	0,7554	0,6944	0,7048	0,6685	0,494
9									0,8896	0,7134	0,6204	0,6504	0,6246	0,2326
10									0,8776	0,6419	0,4788	0,6392	0,6237	0,1571
11									0,8623	0,7207	0,5719	0,5619	0,5796	0,3012
12									0,8262	0,6193	0,5139	0,5391	0,4946	0,1992
13									0,8223	0,5609	0,3454	0,437	0,4246	0,215

* La valeur SRC = 6, HRC = 5 a été tirée de l'analyse car elle dépassait les critères d'alignement temporel du plan d'essai VQEG.

Annexe D

National Telecommunications and Information Administration (NTIA)

Description technique du modèle de mesure de la qualité vidéo (VQM, *video quality metric*)

La présente annexe contient une description fonctionnelle complète du modèle VQM de la NTIA et des techniques d'étalonnage qui lui sont associées. Les algorithmes d'étalonnage décrits dans la présente annexe sont suffisants pour garantir un fonctionnement correct du dispositif d'évaluation de la qualité vidéo de la NTIA. Ils présentent généralement une précision d'alignement spatial de plus ou moins 1/2 pixel et une précision d'alignement temporel de plus ou moins une trame entrelacée.

D.1 Introduction

La présente annexe contient une description technique complète du modèle général de la NTIA (*National Telecommunications and Information Administration*) et des techniques d'étalonnage qui lui sont associées (par exemple évaluation et correction de l'alignement spatial, de l'alignement temporel et des erreurs de gain/décalage). Le modèle général correspond au modèle H dans les essais de télévision avec image de référence complète de Phase II du VQEG. Il était conçu pour être un modèle VQM universel pour des systèmes vidéo avec une très large plage de niveaux de qualité et de débits binaires. De nombreux essais subjectifs et objectifs ont été effectués afin de vérifier les performances du modèle général avant de le soumettre aux essais de Phase II du VQEG, lesquels ont uniquement porté sur l'évaluation des performances du modèle général pour des systèmes vidéo MPEG-2 et H.263. Mais le modèle général devrait fonctionner correctement pour de nombreux autres types de systèmes de codage et de transmission.

Les algorithmes d'étalonnage décrits dans la présente annexe sont suffisants pour garantir un fonctionnement correct du dispositif d'évaluation de la qualité vidéo. Ils présentent généralement une précision d'alignement spatial de plus ou moins 1/2 pixel et une précision d'alignement temporel de plus ou moins une trame entrelacée.

NTIA a indiqué sa volonté de fournir à toutes les parties intéressées un logiciel qui implémente le modèle général et les techniques associées d'étalonnage automatique. Les parties intéressées peuvent le trouver à l'adresse: www.its.bldrdoc.gov/n3/video/vqmsoftware.htm.

Clause de non-garantie: l'UIT ne sera en aucun cas tenue responsable de dommages quelconques (y compris, à titre non limitatif, dommages pour manque à gagner, l'interruption d'exploitation, la perte d'information, ou toute autre perte pécuniaire) résultant de ou en relation avec l'utilisation ou l'impossibilité d'utiliser le logiciel identifié. L'UIT décline toute garantie, explicite ou implicite, y compris et sans s'y limiter, les garanties de commerciabilité ou d'adaptation à un besoin particulier.

D.2 Références

D.2.1 Références normatives

– Recommandation UIT-R BT.601-5 (1995), *Paramètres de codage en studio de la télévision numérique pour des formats standards d'image 4:3 (normalisé) et 16:9 (écran panoramique)*.

D.3 Définitions

D.3.1 4:2:2: format d'échantillonnage d'image Y, Cb, Cr pour lequel les plans de chrominance (Cb et Cr) sont échantillonnés horizontalement à une fréquence qui vaut la moitié de la fréquence d'échantillonnage du plan de luminance (Y). Voir la Rec. UIT-R BT.601-5 (§ D.2).

D.3.2 information temporelle absolue (ATI, *absolute temporal information*): caractéristique déduite de la valeur absolue des images d'information temporelle qui sont calculées comme étant la différence entre deux images successives d'un clip vidéo. La caractéristique ATI quantifie la quantité de mouvement présente dans une scène vidéo. Le § D.7.5 contient la définition mathématique précise.

D.3.3 big YUV: format de fichier binaire utilisé pour stocker les clips qui ont été échantillonnés conformément à la Rec. UIT-R BT.601-5. Dans ce format, toutes les images vidéo d'une scène sont stockées dans un seul grand fichier binaire, dans lequel chaque image est échantillonnée conformément à la Rec. UIT-R BT.601-5. *Y* représente l'information de canal de luminance, *U* représente le canal de différence de couleur bleue (c'est-à-dire C_B dans la Rec. UIT-R BT.601-5) et V représente le canal de différence de couleur rouge (c'est-à-dire C_R dans la Rec. UIT-R BT.601-5). L'ordre des pixels dans le fichier binaire est le même que celui qui est spécifié dans le document 125M de la SMPTE [D-7]. La spécification complète du format de fichier Big YUV figure au § D. 5 et les routines logicielles permettant de lire et d'afficher des fichiers au format Big YUV sont données dans le document [D-14].

D.3.4 clip: représentation numérique d'une scène qui est stockée sur support informatique.

D.3.5 qualité VQM d'un clip: qualité VQM d'un seul clip vidéo traité.

D.3.6 chrominance (C, C_B, C_R): partie du signal vidéo qui achemine avant tout l'information de couleur (C), qui peut de plus être séparée en un signal de différence de couleur bleue (C_B) et un signal de différence de couleur rouge (C_R).

D.3.7 codec: abréviation pour codeur/décodeur ou compresseur/décompresseur.

D.3.8 format intermédiaire commun (CIF, *common intermediate format*): structure d'échantillonnage vidéo utilisée en visioconférence, pour laquelle le canal de luminance est échantillonné à 352 pixels par 288 lignes [D-2].

D.3.9 caractéristique: grandeur associée à – ou extraite d' – une sous-région spatio-temporelle d'un flux vidéo (d'origine ou traité).

D.3.10 trame: la moitié d'une image, contenant toutes les lignes impaires ou toutes les lignes paires.

D.3.11 image: une image de télévision complète.

D.3.12 images par seconde (FPS, *frames per second*): nombre d'images d'origine par seconde transmises par le système vidéo testé. Par exemple, un système vidéo NTSC transmet environ 30 FPS.

D.3.13 gain: facteur multiplicatif appliqué par le circuit fictif de référence (HRC) à tous les pixels d'un plan d'image donné (par exemple luminance, chrominance). Le gain du signal de luminance est généralement appelé contraste.

D.3.14 modèle général: modèle de mesure de la qualité vidéo, ou modèle VQM, qui fait l'objet de la présente annexe (§ D.9). Ce modèle a été soumis aux essais de phase II réalisés par le Groupe d'experts en qualité vidéo (VQEG). Le rapport final du VQEG sur la phase II décrit les performances du modèle général (voir le document [D-15], modèle H).

D.3.15 H.261: désigne la Rec. UIT-T H.261 [D-2].

D.3.16 circuit fictif de référence (HRC, *hypothetical reference circuit*): système vidéo testé, par exemple un codec ou un système de transmission vidéonumérique.

D.3.17 séquence vidéo d'entrée: séquence vidéo avant traitement ou distorsion par un circuit fictif de référence (voir la Figure D.1). On parle aussi de séquence vidéo d'origine.

D.3.18 unité (IRE, *institute for radio engineers*): unité de tension couramment utilisée pour mesurer les signaux vidéo. Une IRE vaut 1/140 de volt.

D.3.19 union internationale des télécommunications (UIT): organisation internationale du système des Nations Unies où le secteur public et le secteur privé coordonnent les réseaux et services mondiaux de télécommunications. L'UIT inclut le Secteur des radiocommunications (UIT-R), le Secteur de la normalisation des télécommunications (UIT-T) et le Secteur du développement des télécommunications (UIT-D).

D.3.20 luminance (Y): partie du signal vidéo qui achemine avant tout l'information de luminance (c'est-à-dire la partie en noir et blanc de l'image).

D.3.21 note moyenne d'opinion (MOS, *mean opinion score*): appréciation subjective moyenne de la qualité d'un clip vidéo traité attribuée par un groupe d'observateurs.

D.3.22 groupe d'experts pour les images animées (MPEG, *moving picture experts group*): groupe de travail de l'ISO/CEI chargé d'élaborer des normes pour la représentation codée des séquences audio et vidéo numériques (par exemple MPEG-1, MPEG-2, MPEG-4).

D.3.23 système (NTSC, *national television systems committee*): système couleur de vidéo composite analogique à 525 lignes [D-8].

D.3.24 décalage ou décalage de niveau: facteur additif appliqué par le circuit fictif de référence à tous les pixels d'un plan d'image donné (par exemple luminance, chrominance). Le décalage du signal de luminance est généralement appelé brillance.

D.3.25 région d'intérêt d'origine (OROI, *original region of interest*): région d'intérêt (ROI) extraite de la séquence vidéo d'origine, spécifiée en coordonnées de rectangle.

D.3.26 séquence vidéo d'origine: séquence vidéo avant traitement ou distorsion par un circuit fictif de référence (voir la Figure D.1). On parle aussi de séquence vidéo d'entrée puisque c'est la séquence vidéo qui entre dans le système de transmission vidéo numérique.

D.3.27 région valable d'origine (OVR, *original valid region*): région valable d'un clip vidéo d'origine, spécifiée en coordonnées de rectangle.

D.3.28 séquence vidéo de sortie: séquence vidéo qui a été traitée ou distordue par un circuit fictif de référence (voir la Figure D1). On parle aussi de séquence vidéo traitée.

D.3.29 surbalayage: partie du flux vidéo qu'on ne peut généralement pas voir sur un écran de télévision standard.

D.3.30 système (PAL, *phase-altering line*): système couleur de vidéo composite analogique à 625 lignes.

D.3.31 paramètre: mesure de la distorsion vidéo résultant de la comparaison de deux flux parallèles de caractéristiques, l'un des flux provenant de la séquence vidéo d'origine et l'autre étant le flux correspondant provenant de la séquence vidéo traitée.

D.3.32 région d'intérêt traitée (PROI, *processed region of interest*): région d'intérêt (ROI) extraite de la séquence vidéo traitée et dont les décalages spatiaux dus au circuit fictif de référence ont été corrigés, spécifiée en coordonnées de rectangle.

D.3.33 séquence vidéo traitée: séquence vidéo qui a été traitée ou distordue par un circuit fictif de référence (voir la Figure D.1). On parle aussi de séquence vidéo de sortie puisque c'est la séquence de sortie du système de transmission vidéo numérique.

D.3.34 région valable traitée (PVR, *processed valid region*): région valable d'un clip vidéo traité provenant d'un circuit fictif de référence, spécifiée en coordonnées de rectangle. La région PVR est toujours spécifiée par rapport à la séquence vidéo d'origine, il faut donc corriger les décalages spatiaux de la séquence vidéo dus au circuit fictif de référence avant de calculer la région PVR. Ainsi, la région PVR est toujours contenue dans la région valable d'origine (OVR). La région comprise entre la région PVR et la région OVR est la partie de la séquence vidéo qui a été supprimée ou altérée par le circuit fictif de référence.

D.3.35 format de production: grille d'image qui représente le format maximal possible de l'image pour un système standard donné. Le format de production représente le format souhaitable pour l'acquisition, la génération et le traitement de l'image, avant suppression. Pour les séquences vidéo échantillonnées selon la Rec. UIT-R BT.601-5, le format de production est de 720 pixels × 486 lignes pour les systèmes à 525 lignes et de 720 pixels × 576 lignes pour les systèmes à 625 lignes [D-9].

D.3.36 quart de format intermédiaire commun (QCIF, *quarter common intermediate format*): structure d'échantillonnage vidéo utilisée en visioconférence, pour laquelle le canal de luminance est échantillonné à 176 pixels par 144 lignes [D-2].

D.3.37 Recommandation UIT-R BT.601-5: norme (voir le § D. 2) commune d'échantillonnage vidéo sur 8 bits selon laquelle le canal de luminance (Y) est échantillonné à 13,5 MHz et les canaux de différence de couleur bleue et rouge (C_B et C_R) sont échantillonnés à 6,75 MHz. Pour plus d'informations, on se reportera au § D.5.

D.3.38 coordonnées de rectangle: sous-région d'image de forme rectangulaire qui est entièrement contenue dans le format de production et qui est spécifiée par quatre coordonnées (haut, gauche, bas, droite). La numérotation, qui commence à zéro, est telle que le coin (haut, gauche) de l'image échantillonnée a pour coordonnées (0, 0). Voir le § D.5.3.

D.3.39 référence réduite: méthode de mesure de la qualité vidéo qui utilise des caractéristiques de faible largeur de bande extraites des flux vidéo d'origine et traité, par opposition à une méthode fondée sur l'image de référence complète pour laquelle il faut connaître entièrement les flux vidéo d'origine et traité [D-3]. Les méthodes fondées sur une référence réduite présentent des avantages quant à la surveillance de qualité de bout en bout en service étant donné que les informations de référence réduite sont transmises facilement sur les réseaux de télécommunications du monde entier.

D.3.40 resynchronisation de trame: processus consistant à réordonner, dans une image vidéo, deux trames entrelacées échantillonnées consécutivement d'une séquence vidéo traitée. La resynchronisation de trame est nécessaire lorsque des circuits fictifs de référence ne conservent pas l'ordre standard des trames entrelacées (par exemple une trame NTSC de type 1 sort sous forme de trame NTSC de type 2 et inversement). Voir le § D.6.1.2.

D.3.41 région d'intérêt (ROI, *region of interest*): grille d'image (spécifiée en coordonnées de rectangle) utilisée pour désigner une sous-région particulière d'une trame ou d'une image vidéo. Voir aussi SROI.

D.3.42 scène: séquence d'images vidéo.

D.3.43 information spatiale (SI, *spatial information*): caractéristique fondée sur des statistiques qui sont extraites des gradients spatiaux (c'est-à-dire des contours) d'une image ou d'une scène vidéo. La référence [D-4] contient une définition de l'information spatiale fondée sur des statistiques extraites d'images auxquelles on a appliqué des filtres de Sobel 3 × 3 [D-6] tandis que le § D.7.2.2 contient une définition de l'information spatiale fondée sur des statistiques extraites d'images auxquelles on a appliqué des filtres de souligné des contours de taille beaucoup plus grande (13 × 13) (voir la Figure D.11).

D.3.44 région d'intérêt spatiale (SROI, *spatial region of interest*): grille d'image particulière (spécifiée en coordonnées de rectangle) utilisée pour calculer la qualité VQM d'un clip vidéo. La région SROI est un sous-ensemble rectangulaire entièrement compris dans la région valable traitée. Pour les séquences vidéo échantillonnées selon la Rec. UIT-R BT.601-5, la région SROI recommandée est de 672 pixels × 448 lignes pour les systèmes à 525 lignes et de 672 pixels × 544 lignes pour les systèmes à 625 lignes, centrée à l'intérieur du format de production. Cette région SROI recommandée correspond approximativement à la partie de l'image vidéo que l'on peut voir sur un écran, à l'exclusion de la zone de surbalayage. Voir aussi ROI.

D.3.45 alignement spatial: processus utilisé pour évaluer et corriger les décalages spatiaux de la séquence vidéo traitée par rapport à la séquence vidéo d'origine.

D.3.46 sous-région spatio-temporelle (S-T): bloc de pixels d'image d'un flux vidéo d'origine ou traité qui inclut une dimension verticale (nombre de lignes), une dimension horizontale (nombre de colonnes) et une dimension temporelle (nombre d'images). Voir la Figure D.9.

D.3.47 société des ingénieurs en images animées et télévision (SMPTE, *Society of Motion Picture and Television Engineers*): importante pour les industriels travaillant dans le domaine des images animées et de la télévision, cette société se charge de développer la théorie et les applications dans le domaine des images animées, y compris les films, la télévision, la vidéo, l'imagerie sur ordinateur et les télécommunications. Les industriels attendent de la SMPTE qu'elle élabore des normes, des lignes directrices en matière d'ingénierie et des pratiques recommandées qui doivent ensuite être suivies par les professionnels respectifs sur le terrain.

D.3.48 information temporelle (TI, *temporal information*): caractéristique fondée sur des statistiques qui sont extraites des gradients temporels (c'est-à-dire du mouvement) d'une scène vidéo. La référence [D-4] et le § D.7.5 contiennent des définitions de l'information temporelle fondée sur des statistiques extraites de simples différences entre images.

D.3.49 région d'intérêt temporelle (TROI, *temporal region of interest*): segment temporel, séquence ou sous-ensemble particulier d'images qui est utilisé pour calculer la qualité VQM d'un clip. La région TROI est un segment contigu d'images qui est entièrement contenu dans la région valable temporelle. La région TROI maximale correspond au segment temporel entièrement aligné et contient toutes les images alignées temporellement de la région TVR. Si une resynchronisation de trame est requise, elle s'applique toujours au clip traité, mais pas au clip d'origine.

D.3.50 alignement temporel: processus utilisé pour évaluer et corriger le décalage temporel (c'est-à-dire le retard vidéo) de la séquence vidéo traitée par rapport à la séquence vidéo d'origine (voir le § D.6.4.1).

D.3.51 région valable temporelle (TVR, *temporal valid region*): segment temporel, séquence ou sous-ensemble maximal d'images vidéo pouvant être utilisé pour l'étalonnage et le calcul de la qualité VQM. Les images situées en dehors de ce segment temporel seront toujours considérées comme non valables.

D.3.52 incertitude (*U, uncertainty*): évaluation de l'erreur d'alignement temporel (plus ou moins), compte tenu de la valeur la plus probable du retard vidéo dû au circuit fictif de référence. Voir le § D.6.4.

D.3.53 région valable (VR, *valid region*): partie rectangulaire d'une grille d'image (spécifiée en coordonnées de rectangle) qui n'est ni supprimée ni altérée par le traitement. La région valable est un sous-ensemble du format de production du système vidéo standard considéré et n'inclut que les pixels d'image qui contiennent une information d'image qui n'a été ni supprimée ni altérée. Voir région valable d'origine et région valable traitée.

D.3.54 groupe d'experts en qualité vidéo (VQEG, *video quality experts group*): groupe d'experts internationaux en qualité vidéo qui réalisent des essais de validation de méthodes objectives de mesure de la qualité vidéo. Les résultats du VQEG sont transmis à l'Union internationale des

télécommunications (UIT) et peuvent servir de base à des recommandations internationales sur la mesure de la qualité vidéo.

D.3.55 mesure de la qualité vidéo, modèle de mesure de la qualité vidéo, qualité VQM (VQM, *video quality metric, model, or measurement*): mesure globale de la dégradation de la qualité vidéo (voir qualité VQM d'un clip, modèle général). La qualité VQM est un nombre unique dont la plage nominale est comprise entre zéro et un, zéro correspondant à aucune dégradation perçue et un à la dégradation maximale perçue.

D.4 Aperçu général du calcul de la qualité VQM

La présente annexe contient une description complète du modèle général et des algorithmes d'étalonnage qui lui sont associés. La méthode de mesure objective automatisée considérée ici donne des résultats proches des appréciations globales (notes moyennes d'opinion) de la qualité vidéo numérique attribuées par des groupes d'observateurs [D-1]. La Figure D.1 donne un diagramme d'ensemble des processus requis pour calculer la qualité VQM selon le modèle général. Ces processus comprennent l'échantillonnage des flux vidéo d'origine et traité (§ D.5), l'étalonnage de ces flux (§ D.6), l'extraction de caractéristiques fondées sur la perception (§ D.7), le calcul de paramètres de qualité vidéo (§ D.8) et le calcul de la qualité VQM selon le modèle général (§ D.9). Le modèle général mesure les modifications perçues de la qualité résultant de distorsions dues à n'importe quel composant du système de transmission vidéo numérique (par exemple codeur, canal numérique, décodeur).

La méthode de mesure décrite ici utilise des paramètres de référence réduite de largeur de bande élevée [D-3]. Ces paramètres sont fondés sur des caractéristiques extraites de régions spatio-temporelles (S-T) de la séquence vidéo (voir le § D.7.1.1). La méthode de mesure présentée ici peut donc aussi être utilisée pour surveiller la qualité vidéo en service lorsqu'un canal de données auxiliaires est disponible pour transmettre les caractéristiques extraites entre la source et la destination d'un circuit fictif de référence (voir la Figure D.1).

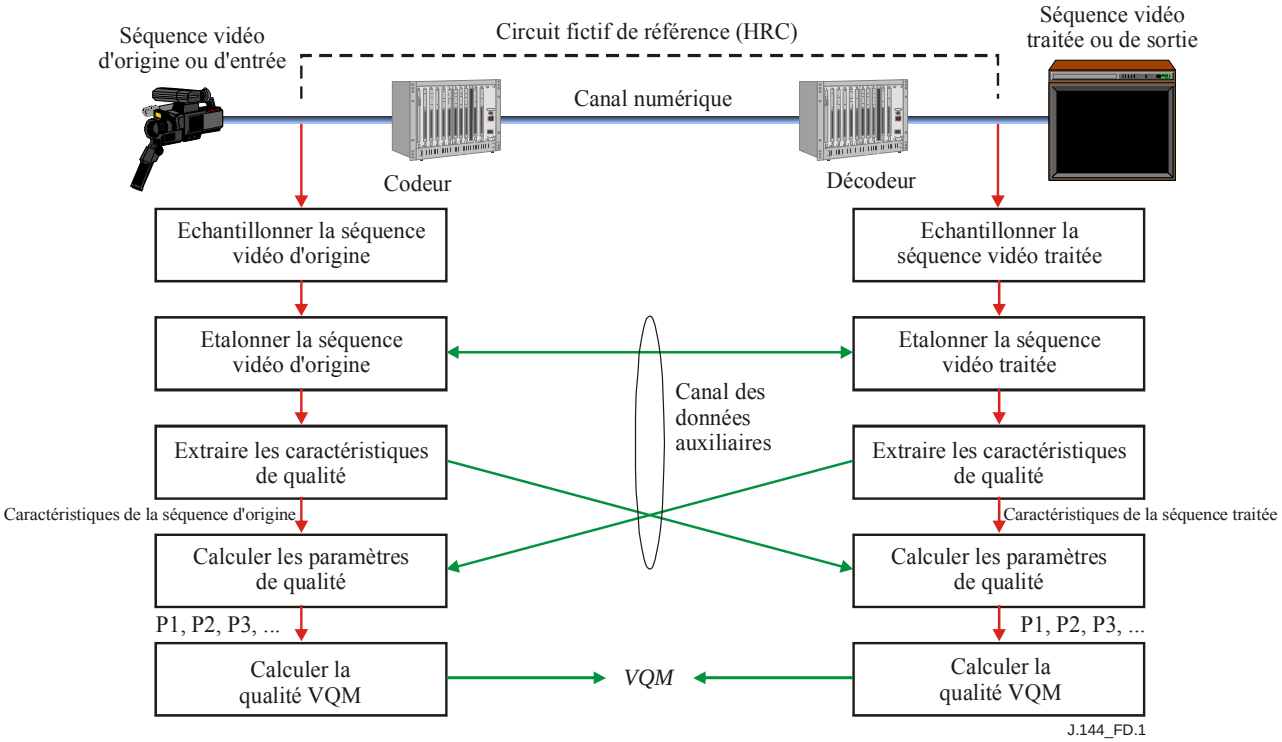


Figure D.1 – Etapes nécessaires pour calculer la qualité VQM

D.5 Echantillonnage

Pour les algorithmes informatiques exposés dans la présente annexe, on suppose que les flux vidéo d'origine et traité sont disponibles sous forme de représentations numériques stockées sur support informatique (on parle de clip dans la présente annexe). Si le flux vidéo est en format analogique, l'une des normes d'échantillonnage numérique les plus largement utilisées est la Rec. UIT-R BT.601-5 (§ D.2). Un flux vidéo composite (par exemple NTSC ou PAL) doit d'abord être converti en flux vidéo en composantes contenant les trois signaux suivants: luminance (Y), différence de couleur bleue (C_B) et différence de couleur rouge (C_R). L'échantillonnage selon la Rec. UIT-R BT.601-5 est souvent appelé échantillonnage 4:2:2 car la fréquence d'échantillonnage du canal Y est le double de la fréquence d'échantillonnage des canaux C_B et C_R. La Rec. UIT-R BT.601-5 spécifie une fréquence d'échantillonnage de 13,5 MHz pour le canal Y, qui produit 720 échantillons Y par ligne vidéo. Etant donné que dans le système NTSC à 525 lignes, les informations d'image sont contenues dans 486 lignes, l'image vidéo Y complète échantillonnée selon la Rec. UIT-R BT.601-5 sera de 720 pixels par 486 lignes. De même, lorsqu'un flux vidéo PAL à 625 lignes est échantillonné selon la Rec. UIT-R BT.601-5, l'image vidéo Y sera de 720 pixels par 576 lignes. Si on utilise 8 bits pour échantillonner de manière uniforme le signal Y, la Rec. UIT-R BT.601-5 spécifie que la valeur d'échantillonnage du noir de référence (c'est-à-dire 7,5 IRE) est "16" et que celle du blanc de référence (c'est-à-dire 100 IRE) est "235". Ainsi, une marge de travail est prévue pour les signaux vidéo qui dépassent les niveaux du noir et du blanc de référence avant écrêtage par le convertisseur analogique-numérique. Chacun des canaux de chrominance (C_B et C_R) est échantillonné à 6,75 MHz et le premier couple d'échantillons de chrominance (C_B, C_R) est associé au premier échantillon de luminance Y, le deuxième couple d'échantillons de chrominance est associé au troisième échantillon de luminance, etc. Comme les canaux de chrominance sont bipolaires, la valeur d'échantillonnage du signal nul est "128".

D.5.1 Indexation temporelle des images figurant dans les fichiers vidéo d'origine et traité

Une image de luminance de flux vidéo échantillonnée selon la Rec. UIT-R BT.601-5 sera désignée par $\boldsymbol{Y}(t)$. La variable t est utilisée ici comme indice pour les images échantillonnées figurant dans les fichiers Big YUV d'origine et traité; elle ne désigne pas le temps véritable. Si le fichier Big YUV contient N images, comme indiqué sur la Figure D.2, $t = 0$ désigne la première image qui a été échantillonnée et $t = (N - 1)$ désigne la dernière image qui a été échantillonnée.

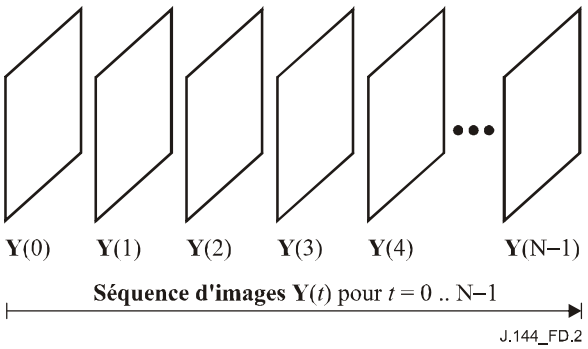


Figure D.2 – Indexation temporelle des images figurant dans les fichiers Big YUV

Tous les algorithmes décrits ici fonctionnent sur la base de couples de fichiers échantillonnés, chaque couple comprenant un fichier pour la séquence vidéo d'origine et un fichier pour la séquence vidéo traitée associée. Pour éviter toute confusion, on suppose que les deux fichiers d'un couple ont la même longueur. Par ailleurs, on suppose au départ que la première image du fichier d'origine est alignée temporellement avec la première image du fichier traité, avec plus ou moins une certaine incertitude temporelle.

Pour les implémentations en service et en temps réel, cette hypothèse d'incertitude bilatérale peut être remplacée par une hypothèse d'incertitude unilatérale, découlant de la causalité. Par exemple, une image traitée apparaissant à l'instant $t = n$ doit provenir d'images d'origine apparues à l'instant t = n ou antérieurement.

L'hypothèse susmentionnée concernant les fichiers vidéo d'origine et traité (à savoir que les premières images sont alignées) équivaut à choisir la valeur la plus probable du retard dû au circuit fictif de référence présenté sur la Figure D.1. Par conséquent, l'incertitude restante quant à l'évaluation du retard vidéo sera de plus ou moins $\boldsymbol{U}$.

D.5.2 Indexation spatiale des images des flux vidéo d'origine et traité

Le système de coordonnées utilisé pour les images de luminance échantillonnées est présenté sur la Figure D.3. Les coordonnées horizontale et verticale du coin en haut à gauche des images de luminance sont définies comme valant (v = 0, h = 0), où la valeur de la coordonnée sur l'axe horizontal (h) croît vers la droite et la valeur de la coordonnée sur l'axe vertical (v) croît vers le bas. La coordonnée sur l'axe horizontal est comprise entre 0 et le nombre de pixels d'une ligne moins un. La coordonnée sur l'axe vertical est comprise entre 0 et le nombre de lignes moins un, le nombre de lignes étant le nombre de lignes d'une image pour les systèmes à balayage progressif et soit le nombre de lignes d'une trame soit le nombre de lignes d'une image pour les systèmes à balayage avec entrelacement. L'amplitude du pixel de $\boldsymbol{Y}(t)$ échantillonné correspondant à la ligne i (v = i) et à la colonne j (h = j) et à l'instant t est désignée par $Y(i, j, t)$.

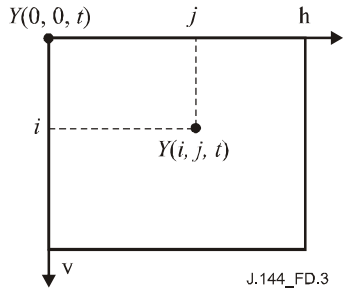


Figure D.3 – Système de coordonnées utilisé pour les images Y de luminance échantillonnées

Un clip vidéo échantillonné selon la Rec. UIT-R BT.601-5 est stocké dans un fichier de format "Big YUV", Y désignant l'information de luminance selon la Rec. UIT-R BT.601-5, U l'information de différence de couleur bleue (c'est-à-dire C_B dans la Rec. UIT-R BT.601-5) et V l'information de différence de couleur rouge (c'est-à-dire C_R dans la Rec. UIT-R BT.601-5). Avec le format de fichier Big YUV, toutes les images sont stockées séquentiellement dans un seul grand fichier binaire continu. Les pixels d'image sont stockés séquentiellement par ligne de balayage vidéo sous forme d'octets dans l'ordre suivant: C_{B0}, Y_0, C_{R0}, Y_1, C_{B2}, Y_2, C_{R2}, Y_3, etc., l'indice numérique désignant le numéro du pixel (on doit procéder à une duplication de pixel ou à une interpolation entre pixels pour déterminer les échantillons de chrominance C_B et C_R associés à Y_1, Y_3, …). Cet ordre des octets est équivalent à celui qui est spécifié dans le document 125M de la SMPTE [D-7].

D.5.3 Spécification de sous-régions rectangulaires

On utilise des sous-régions rectangulaires d'une image échantillonnée pour contrôler le calcul de la qualité VQM. Par exemple, on peut calculer la qualité VQM sur la région valable de l'image échantillonnée ou sur une région d'intérêt spatiale spécifiée par l'utilisateur qui est plus petite que la région valable. Pour spécifier des sous-régions rectangulaires, on utilise les coordonnées de rectangle définies par les quatre grandeurs suivantes: haut, gauche, bas et droite. La Figure D.4 illustre la spécification d'une sous-région rectangulaire d'une image vidéo échantillonnée. Les pixels

rouges de l'image sont inclus dans la sous-région mais les pixels noirs de l'image en sont exclus. Pour le calcul de la qualité VQM, une image est souvent subdivisée en un grand nombre de sous-régions plus petites contiguës. La définition d'une sous-région rectangulaire présentée sur la Figure D.4 permet de définir la grille utilisée pour afficher ces sous-régions contiguës et les fonctions mathématiques utilisées pour extraire les caractéristiques de chacune de ces sous-régions.

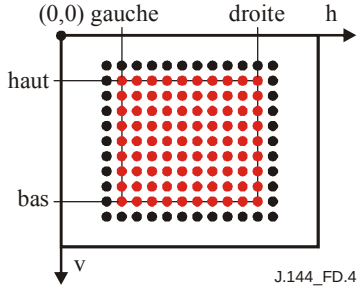


Figure D.4 – Coordonnées de rectangle pour la spécification de sous-régions d'une image

D.5.4 Considérations relatives aux séquences vidéo de plus de 10 secondes

Pour les mesures de la qualité vidéo dont il est question dans la présente annexe, on s'est fondé sur les résultats d'essais subjectifs relatifs à des clips vidéo de 8 à 10 secondes. Lorsque la séquence est plus longue, il convient de la subdiviser en segments vidéo plus courts, chaque segment étant supposé avoir ses propres attributs d'étalonnage et de qualité. La méthode consistant à subdiviser le flux vidéo en segments se chevauchant et à traiter chaque segment indépendamment des autres permet d'émuler des évaluations continues de la qualité pour les longues séquences vidéo au moyen des techniques de mesure VQM présentées ici.

D.6 Etalonnage

Quatre étapes sont nécessaires pour étalonner correctement les séquences vidéo échantillonnées en vue de l'extraction des caractéristiques. Ces étapes sont les suivantes:

1) évaluation de l'alignement spatial et correction;
2) évaluation de la région valable afin de limiter l'extraction des caractéristiques aux pixels qui contiennent l'information d'image;
3) évaluation du gain et du décalage de niveau (généralement appelés contraste et brillance) et correction;
4) évaluation de l'alignement temporel et correction.

L'étape 2 doit être appliquée aux flux vidéo d'origine et traité. Les étapes 1, 3 et 4 doivent être appliquées au flux vidéo traité. Généralement, l'alignement spatial, le gain et le décalage de niveau sont constants pour un système vidéo donné et ces grandeurs n'ont donc à être calculées qu'une seule fois. Toutefois, il est courant que la région valable et l'alignement temporel changent en fonction du contenu de la scène. Par exemple, une scène au format plein écran et une scène au format boîte aux lettres auront des régions valables différentes; les systèmes de visioconférence présentent souvent des retards vidéo variables qui dépendent du contenu de la scène (par exemple une scène dans laquelle on voit la tête d'une personne qui parle et une scène d'une épreuve sportive). En plus des techniques d'étalonnage présentées ici, le lecteur souhaitera peut-être aussi examiner d'autres méthodes d'alignement spatial et temporel [D-5].

Le fait de procéder à un étalonnage avant l'extraction des caractéristiques implique que les décalages horizontal et vertical de l'image, les décalages temporels du flux vidéo résultant de retards vidéo non nuls et les modifications du contraste et de la brillance d'image comprises dans la plage dynamique de l'unité d'échantillonnage vidéo n'auront pas d'incidence sur la qualité VQM. Ces

grandeurs liées à l'étalonnage peuvent avoir une grande incidence sur la qualité globale perçue (par exemple des images à faible contraste issues d'un système vidéo avec un gain de 0,3), mais la philosophie adoptée ici consiste à séparer les informations liées à l'étalonnage de la qualité VQM. De bonnes pratiques techniques permettent généralement d'ajuster les décalages spatiaux, les régions valables, les gains et les décalages de niveau; les décalages temporels fournissent des informations importantes sur la qualité lors de l'évaluation de systèmes vidéo bidirectionnels ou interactifs.

Pour toutes les caractéristiques et tous les paramètres de qualité vidéo (§ D.7 et 8), on suppose qu'un seul retard vidéo est supprimé pour l'alignement temporel de la séquence vidéo traitée (retard vidéo constant). Certains systèmes vidéo ou circuits fictifs de référence appliquent un retard différent à chaque image traitée (retard vidéo variable). Dans la présente annexe, on considère que tous les systèmes vidéo ont un retard vidéo constant. Les variations par rapport à ce retard sont considérées comme des dégradations qui sont mesurées par les caractéristiques et les paramètres. Cette approche semble conduire à de meilleures corrélations avec la note subjective que les mesures de la qualité vidéo fondées sur des séquences vidéo traitées dont le retard vidéo variable a été supprimé. Lorsqu'une séquence vidéo est longue (voir le § D.5.4), il convient de la subdiviser en segments vidéo plus courts, chaque segment ayant son propre retard vidéo constant, ce qui autorise une certaine variation du retard en fonction du temps. Il est possible d'obtenir une évaluation plus continue des variations du retard en subdivisant la séquence en segments temporels se chevauchant.

Si le circuit fictif de référence testé réduit ou agrandit la taille de l'image (par exemple zoom), il faudrait inclure, dans le processus d'étalonnage, une étape additionnelle visant à évaluer et à supprimer cette réduction ou cet agrandissement spatial. Cette étape n'entre pas dans le cadre de la présente annexe.

D.6.1 Alignement spatial

D.6.1.1 Aperçu général

Le processus d'alignement spatial détermine les décalages spatiaux horizontal et vertical d'une image vidéo traitée par rapport à l'image vidéo d'origine. Un décalage horizontal positif correspond à une image traitée qui a été déplacée vers la droite par un certain nombre de pixels. Un décalage vertical positif correspond à une image traitée qui a été déplacée vers le bas par un certain nombre de lignes. Ainsi, pour l'alignement spatial d'une image vidéo avec balayage à entrelacement, il faut tenir compte de trois grandeurs: le décalage horizontal en nombre de pixels, le décalage vertical de la trame une en nombre de lignes de trame et le décalage vertical de la trame deux en nombre de lignes de trame. Pour l'alignement spatial d'une image vidéo avec balayage progressif, il faut tenir compte de deux grandeurs: le décalage horizontal et le décalage vertical en nombre de lignes d'image. L'algorithme d'alignement spatial est précis au pixel près pour les décalages horizontaux et à la ligne près pour les décalages verticaux. Une fois que l'alignement spatial a été calculé, le décalage spatial est supprimé du flux vidéo traité (par exemple une image traitée qui a été décalée vers le bas est redécalée vers le haut). En cas de balayage à entrelacement, le processus peut inclure une resynchronisation de trame du flux vidéo traité découlant de la comparaison des décalages verticaux des trames une et deux.

Dans le cas du balayage à entrelacement, toutes les opérations s'appliquent à chaque trame séparément; dans le cas du balayage progressif, toutes les opérations s'appliquent à l'image entière. Dans un souci de simplicité, l'algorithme d'alignement spatial sera d'abord entièrement décrit dans le cas du balayage à entrelacement, car c'est le cas le plus compliqué. Les modifications à apporter dans le cas du balayage progressif sont présentées au § D.6.1.6.

L'alignement spatial doit être déterminé avant la région valable traitée (PVR, *processed valid region*), le gain et le décalage de niveau ainsi que l'alignement temporel. Plus précisément, pour calculer chacune de ces grandeurs, il faut comparer le contenu vidéo d'origine et le contenu vidéo traité qui a été aligné spatialement. Si le flux vidéo traité a été décalé spatialement par rapport au

flux vidéo d'origine et que ce décalage spatial n'a pas été corrigé, les évaluations seraient mauvaises car elles seraient fondées sur des contenus vidéo non analogues. Malheureusement, on ne peut pas déterminer correctement l'alignement spatial si on ne connaît pas la PVR, le gain et le décalage de niveau ainsi que l'alignement temporel. L'interdépendance de ces grandeurs cause un problème de mesure du type "de la poule et de l'œuf". Pour pouvoir calculer l'alignement spatial d'une trame traitée, il faut connaître la PVR, le gain et le décalage de niveau ainsi que la trame d'origine lui correspondant le mieux. Toutefois, il est impossible de déterminer ces grandeurs si le décalage spatial n'est pas connu. Une recherche entièrement exhaustive couvrant toutes les variables nécessiterait un nombre considérable de calculs en cas de grosses incertitudes concernant les grandeurs ci-dessus.

La solution présentée ici consiste à procéder à une recherche itérative afin de trouver la trame d'origine correspondant le mieux à chaque trame traitée. Cette recherche inclut une mise à jour itérative des évaluations de PVR, de gain et de décalage de niveau ainsi que d'alignement temporel. Toutefois, pour certaines trames traitées, l'algorithme d'alignement spatial peut échouer. Généralement, lorsque l'alignement spatial n'est pas évalué correctement pour une trame traitée, l'ambiguïté est due aux caractéristiques de la scène. Considérons, par exemple, une scène avec balayage à entrelacement créée numériquement contenant un panoramique vers la gauche. Comme le panoramique a été généré par ordinateur, cette scène pourrait comporter un panoramique horizontal d'exactement un pixel à chaque trame. Du point de vue de l'algorithme de recherche de l'alignement spatial, il serait impossible de faire une différence entre l'alignement spatial correct calculé par rapport à la trame d'origine correspondante, et un décalage horizontal de deux pixels calculé par rapport à la trame qui précède de deux trames la trame d'origine correspondante. Considérons un autre exemple dans lequel une image est entièrement constituée de lignes verticales noires et blanches numériquement parfaites. Comme l'image ne contient pas de ligne horizontale, le décalage vertical est complètement ambigu. Comme le motif de lignes verticales se répète, le décalage horizontal est ambigu, deux décalages horizontaux ou davantage étant tout aussi probables.

Par conséquent, il convient d'appliquer l'algorithme de recherche itérative à une séquence de trames traitées. Les évaluations individuelles de décalage spatial de plusieurs trames traitées peuvent alors servir à produire une évaluation plus robuste. Les évaluations de décalage spatial de plusieurs séquences ou scènes peuvent ensuite être combinées afin de produire une évaluation encore plus robuste pour le circuit fictif de référence testé, dans l'hypothèse où le décalage spatial est constant pour toutes les scènes qui passent par ce circuit.

D.6.1.2 Questions relatives à l'entrelacement

L'alignement spatial vertical est plus complexe pour un flux vidéo avec balayage à entrelacement que pour un flux vidéo avec balayage progressif, car le processus d'alignement spatial doit faire la différence entre la trame une et la trame deux. Trois conditions de décalage vertical doivent être différenciées afin d'obtenir l'alignement vertical correct pour les systèmes avec balayage à entrelacement: le décalage vertical de la trame une est égal au décalage vertical de la trame deux, le décalage vertical de la trame une est inférieur de un au décalage vertical de la trame deux, le décalage vertical est tout autre.

Certains circuits fictifs de référence décalent de manière identique la trame une et la trame deux; dans ce cas, le décalage vertical de la trame une est égal au décalage vertical de la trame deux. Pour les circuits fictifs de référence qui ne répètent pas les trames ou les images (c'est-à-dire les circuits fictifs de référence qui émettent au plein débit d'images du système vidéo), cette condition signifie que ce qui était une trame une dans le flux vidéo d'origine est également une trame une dans le flux vidéo traité et ce qui était une trame deux dans le flux d'origine est également une trame deux dans le flux traité.

D'autres circuits fictifs de référence procèdent à une resynchronisation de trame du flux vidéo, décalant l'image échantillonnée par un nombre impair de lignes d'image. La trame une de la

séquence d'origine devient la trame deux de la séquence traitée et la trame deux de la séquence d'origine devient la trame une de l'image suivante. Visuellement, le flux vidéo affiché semble correct car l'être humain ne peut pas percevoir un décalage d'image du flux vidéo correspondant à une ligne.

Comme indiqué sur la Figure D.5, la trame une commence à la ligne d'image une et contient toutes les lignes d'image impaires. La trame deux commence à la ligne d'image zéro (ligne d'image la plus en haut) et contient toutes les lignes d'image paires. Pour les systèmes NTSC, la trame une est la première dans le temps et la trame deux la deuxième. Pour les systèmes PAL, la trame deux est la première dans le temps et la trame une la deuxième.

Une resynchronisation de trame a lieu lorsque la première trame devient la deuxième et la deuxième devient la première de l'image suivante (retard d'une trame) ou lorsque la deuxième trame devient la première et la première de l'image suivante devient la deuxième de l'image en cours (avance d'une trame). Par exemple, lorsque la trame d'origine NTSC deux devient la trame une de l'image NTSC suivante, la ligne du haut de la trame qui était la ligne d'image 0 de la trame d'origine deux devient la ligne d'image 1 de la trame traitée une. Selon le numérotage des lignes de trame, la ligne du haut reste la ligne de trame 0; ainsi, la trame traitée une présente un décalage vertical nul (car les décalages verticaux sont mesurés pour chaque trame au moyen des lignes de trame). Lorsque la trame NTSC d'origine une devient la trame deux de la même image, la ligne du haut de la trame qui était la ligne d'image 1 de la trame d'origine une devient la ligne d'image 2 de la trame traitée deux. Selon le numérotage des lignes de trame, la ligne du haut qui était la ligne de trame 0 devient la ligne de trame 1; ainsi, la trame traitée deux présente un décalage vertical d'une ligne de trame. La règle générale applicable à la fois au système NTSC et au système PAL est la suivante: lorsque le décalage vertical de la trame deux (en nombre de lignes de trame) est supérieur de un au décalage vertical de la trame une (en nombre de lignes de trame), une resynchronisation de trame a eu lieu.

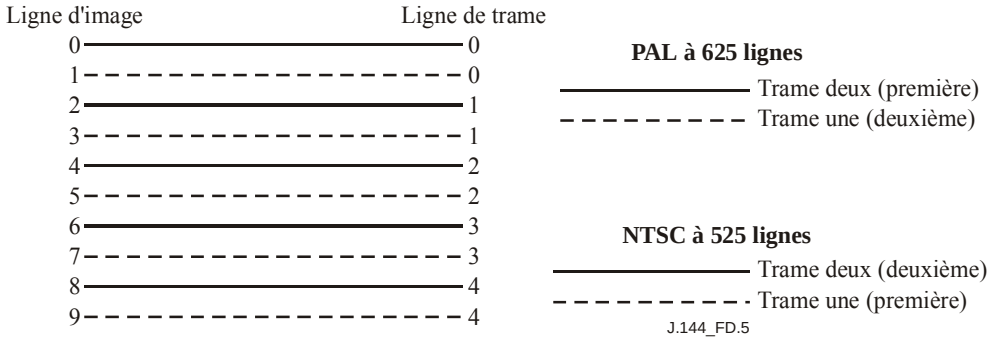


Figure D.5 – Diagramme illustrant le numérotage des trames entrelacées et des lignes d'image/de trame

Si le décalage vertical de la trame deux est différent de celui de la trame une et ne vaut pas non plus un de plus que celui de la trame une, le circuit fictif de référence a altéré l'échantillonnage spatial des deux trames entrelacées et la scène vidéo résultante apparaîtra comme montant et descendant brusquement. Une telle dégradation est évidente et gênante pour l'observateur et, de fait, se produit rarement dans la pratique car le concepteur de circuits fictifs de référence découvre et corrige l'erreur. Par conséquent, l'alignement spatial repose la plupart du temps sur deux schémas courants. Dans les systèmes sans resynchronisation de trame, le décalage vertical de la trame une est égal au décalage vertical de la trame deux; dans les systèmes avec resynchronisation de trame, le décalage vertical de la trame une plus un est égal au décalage vertical de la trame deux.

En outre, il est à noter que l'alignement spatial inclut certaines informations d'alignement temporel, notamment la question de savoir s'il y a eu resynchronisation de trame ou non. Le processus d'alignement temporel peut ne pas être capable de détecter une resynchronisation de trame, mais

même s'il le peut, la resynchronisation de trame est inhérente au processus d'alignement spatial. L'alignement spatial doit donc être capable de déterminer si la trame traitée considérée correspond le mieux à une trame d'origine une ou deux. L'alignement spatial pour chaque trame ne peut être calculé correctement que lorsque la trame traitée est comparée avec la trame d'origine dont elle est issue. Mis à part la question de la resynchronisation de trame, l'utilisation de la mauvaise trame d'origine (trame une/trame deux) peut entraîner des imprécisions quant à l'alignement spatial en raison des différences intrinsèques de contenu spatial dans les deux trames entrelacées.

D.6.1.3 Variables d'entrée requises par l'algorithme d'alignement spatial

Le présent paragraphe contient la liste des variables d'entrée requises par l'algorithme d'alignement spatial. Ce sont notamment la plage des décalages spatiaux et la plage des trames d'origine sur lesquelles la recherche doit porter. Si ces plages sont trop grandes, la vitesse de convergence de l'algorithme de recherche itérative utilisé pour trouver le décalage spatial risque d'être lente et la probabilité pour que l'alignement spatial pour des scènes au contenu répétitif soit erroné sera élevée (par exemple quelqu'un qui fait un signe de la main). Inversement, si ces plages sont trop petites, l'algorithme de recherche se heurtera aux limites des plages de recherche et les repoussera *lentement* au cours des itérations successives. Cette intelligence de recherche intégrée est utile si l'utilisateur fait une faible erreur d'évaluation des incertitudes de la recherche, mais risque d'augmenter considérablement le temps d'exécution si l'utilisateur fait une forte erreur d'évaluation. Par ailleurs, l'algorithme de recherche risque de ne pas trouver le décalage spatial correct dans ce cas.

D.6.1.3.1 Plage prévue des décalages spatiaux

La plage prévue des décalages spatiaux pour des flux vidéo à 525 lignes et à 625 lignes échantillonnés conformément à la Rec. UIT-R BT.601-5 est de ±20 pixels horizontalement et de ±12 lignes de *trame* verticalement. Elle a été déterminée empiriquement sur la base du traitement de données vidéo issues de centaines de circuits fictifs de référence. La plage prévue des décalages spatiaux pour des flux vidéo échantillonnés conformément à d'autres formats plus petits que ceux de la Rec. UIT-R BT.601-5 (par exemple CIF) est supposée être moitié moins grande que la plage observée pour les systèmes à 525 lignes et à 625 lignes. L'algorithme de recherche devrait fonctionner correctement – quoiqu'un peu plus lentement – lorsque la trame traitée présente des décalages spatiaux non compris dans la plage prévue des décalages spatiaux. Cela est dû au fait que l'algorithme de recherche élargira la recherche au-delà de la plage prévue des décalages spatiaux lorsque c'est justifié. Toutefois, le résultat de la détermination de l'alignement spatial correct risque d'être signalé comme étant un échec si les excursions dépassent 50% de la plage prévue.

D.6.1.3.2 Incertitude temporelle

L'utilisateur doit aussi spécifier l'incertitude quant à l'alignement temporel, c'est-à-dire la plage des trames d'origine à examiner pour chaque trame traitée. Cette incertitude temporelle est exprimée sous la forme d'un certain nombre de trames avant et après l'alignement temporel par défaut. Si les séquences vidéo d'origine et traitée sont stockées sous forme de fichiers, un alignement temporel par défaut raisonnable consiste à supposer que la première trame d'un fichier est alignée avec la première trame de l'autre fichier. L'incertitude temporelle qui est spécifiée devrait être suffisamment grande pour inclure l'alignement temporel réel. Une incertitude de plus ou moins une seconde (30 images dans le cas NTSC à 525 lignes; 25 images dans le cas PAL à 625 lignes) devrait suffire pour la plupart des systèmes vidéo. Une incertitude temporelle plus grande pourra être nécessaire pour les circuits fictifs de référence présentant de longs retards vidéo. L'algorithme de recherche pourra envisager des alignements temporels qui sortent de la plage d'incertitude spécifiée lorsque c'est justifié (par exemple lorsque la trame d'origine la plus éloignée est choisie comme correspondant au meilleur alignement temporel).

D.6.1.3.3 Evaluation de la région valable traitée (PVR)

L'évaluation de la région valable traitée (PVR) consiste à spécifier la partie de l'image traitée qui n'a été ni supprimée ni altérée par le traitement, en supposant qu'il n'y a pas eu de décalage spatial (car le décalage spatial n'a pas encore été mesuré). L'évaluation de la PVR peut être déterminée empiriquement, mais une évaluation de la PVR qui est spécifiée par l'utilisateur et qui exclut la zone de surbalayage constitue un bon choix. Dans la plupart des cas, cela permet de ne pas utiliser les parties vidéo non valables dans l'algorithme d'alignement spatial. Concernant les flux vidéo NTSC à 525 lignes échantillonnés conformément à la Rec. UIT-R BT.601-5, la zone de surbalayage couvre environ 18 lignes d'image en haut et en bas de l'image et 22 pixels à gauche et à droite de l'image. Concernant les flux vidéo PAL à 625 lignes échantillonnés conformément à la Rec. UIT-R BT.601-5, la zone de surbalayage couvre environ 14 lignes d'image en haut et en bas de l'image et 22 pixels à gauche et à droite de l'image. Pour les autres formats d'image (par exemple CIF), il convient de choisir une PVR par défaut raisonnable.

D.6.1.4 Sous-algorithmes utilisés par l'algorithme d'alignement spatial

L'algorithme d'alignement spatial utilise un certain nombre de sous-algorithmes – notamment pour évaluer le gain et le décalage de niveau – et des formules permettant de déterminer la trame d'origine qui correspond le mieux à une trame traitée donnée. Ces sous-algorithmes ont été conçus pour être efficaces sur le plan du calcul, étant donné qu'ils doivent être exécutés de nombreuses fois dans le cadre de l'algorithme de recherche itérative.

D.6.1.4.1 Région d'intérêt (ROI) utilisée par tous les calculs

Toutes les comparaisons de trame opérées par l'algorithme se font entre des versions décalées spatialement d'une ROI extraite du flux vidéo traité (afin de compenser les décalages spatiaux introduits par le circuit fictif de référence) et la ROI correspondante extraite du flux vidéo d'origine. Toute ROI extraite du flux vidéo traité et décalée spatialement sera appelée PROI (ROI traitée) et la ROI correspondante extraite du flux vidéo d'origine sera appelée OROI (ROI d'origine). Les coordonnées de rectangle qui spécifient la OROI sont fixes tout au long de l'algorithme et sont choisies de manière à avoir la plus grande OROI possible qui satisfait aux deux conditions suivantes:

– la OROI doit correspondre à une PROI qui est située dans la région valable traitée (PVR) pour tous les décalages spatiaux possibles qui sont examinés;
– la OROI est centrée dans l'image d'origine.

D.6.1.4.2 Gain et décalage de niveau

L'algorithme qui suit sert à évaluer le gain du flux vidéo traité. On corrige le décalage spatial de la trame traitée examinée en utilisant l'évaluation courante du décalage spatial. Après cette correction, on choisit une PROI qui correspond à la OROI fixe déterminée au § D.6.1.4.1. On calcule ensuite l'écart type des valeurs des pixels de luminance (Y) de cette PROI et l'écart type des valeurs des pixels de luminance (Y) de la OROI. On évalue alors le gain comme étant l'écart type associé à la PROI divisé par l'écart type associé à la OROI.

A mesure que l'on se rapproche du décalage spatio-temporel correct au cours des itérations successives de l'algorithme, la fiabilité de cette évaluation du gain est renforcée. On peut utiliser un gain de 1 (c'est-à-dire aucune correction du gain) pendant les premiers cycles d'itération. Le calcul de gain décrit ci-dessus est sensible aux dégradations présentes dans le flux vidéo traité (par exemple flou). Toutefois, pour l'alignement spatial, cette évaluation du gain est utile car elle permet au flux vidéo traité de ressembler le plus possible au flux vidéo d'origine. Pour supprimer le gain de la trame traitée, la valeur de chaque pixel de luminance de la trame traitée est divisée par le gain.

Il n'est pas nécessaire de déterminer ou de corriger le décalage de niveau, car les décalages de niveau n'ont pas d'incidence sur les critères de recherche de l'algorithme d'alignement spatial (voir le § D.6.1.4.3).

D.6.1.4.3 Formules utilisées pour comparer la PROI avec la OROI

Après avoir corrigé le gain[6] dans la PROI (§ D.6.1.4.2), on utilise l'écart type de l'image de différence (OROI-PROI) pour choisir un décalage spatial et un décalage temporel parmi différentes valeurs. On utilise l'évaluation de gain associée à la meilleure correspondance précédente pour corriger le gain de la PROI. Pour déterminer un décalage spatial parmi plusieurs valeurs (le décalage temporel étant maintenu constant), on calcule l'écart type de l'image de différence (OROI-PROI) pour plusieurs PROI générées avec différents décalages spatiaux. Pour une trame traitée donnée, on choisit la combinaison de décalages spatial et temporel qui produit l'écart type le plus petit (c'est-à-dire la plus grande annulation par rapport à la trame d'origine) comme correspondant à la meilleure correspondance.

D.6.1.5 Alignement spatial utilisant des scènes arbitraires

Pour l'alignement spatial d'une trame traitée extraite d'une scène, il faut examiner plusieurs trames d'origine et décalages spatiaux car le décalage temporel (c'est-à-dire le retard vidéo) et le décalage spatial sont tous deux inconnus. Il s'ensuit que l'algorithme de recherche est complexe et nécessite beaucoup de calculs. Par ailleurs, comme le contenu de la scène est arbitraire, il est possible que l'algorithme détermine un alignement spatial incorrect (§ D.6.1.1). Il est donc prudent de calculer l'alignement spatial de plusieurs trames traitées extraites de plusieurs scènes différentes qui sont toutes passées par le même circuit fictif de référence et de combiner les résultats afin d'obtenir une évaluation robuste du décalage spatial. Un circuit fictif de référence donné devrait avoir un seul alignement spatial constant. Si ce n'est pas le cas, des décalages spatiaux variables dans le temps seraient perçus comme une dégradation (par exemple le flux vidéo rebondirait de haut en bas et de bas en haut ainsi que d'un côté à l'autre). Le présent paragraphe décrit l'algorithme d'alignement spatial dans le cas haut-bas; pour cela, on décrit d'abord les principaux composants de l'algorithme puis leur application pour des scènes et des circuits fictifs de référence.

D.6.1.5.1 Meilleure correspondance de trame d'origine dans le temps

Pour déterminer l'alignement spatial à partir du contenu d'une scène, l'algorithme doit déterminer la trame d'origine qui correspond le mieux à la trame traitée courante. Malheureusement, il se peut que cette trame d'origine n'existe pas. Par exemple, une trame traitée peut contenir des parties de deux trames d'origine différentes car elle a pu être interpolée à partir d'autres trames traitées. L'évaluation courante de la meilleure correspondance de trame d'origine (c'est-à-dire la trame d'origine qui correspond le mieux à la trame traitée courante) est conservée à toutes les étapes de l'algorithme de recherche.

On suppose au départ que la première trame du fichier Big YUV traité est alignée avec la première trame du fichier Big YUV d'origine, plus ou moins une certaine incertitude temporelle en nombre d'images (appelée **U**). Pour chaque trame traitée qui est examinée par l'algorithme, il faut un tampon de **U** images d'origine avant et après cette trame. L'algorithme commence donc à examiner les trames traitées se trouvant à **U** images après le début du fichier, examine toutes les trames qui suivent correspondant à une certaine fréquence (appelée **F**), et s'arrête **U** images avant la fin du fichier.

---

[6] Afin de réduire la complexité des calculs, la compensation du gain peut parfois être omise. Toutefois, l'omission de la correction du gain n'est recommandée qu'au cours des premières étapes de l'algorithme de recherche itérative, dont l'objectif est de déterminer un alignement spatial approximatif (voir par exemple les § D.6.1.5.2 et 6.1.5.3).

Les résultats finals de la recherche pour la trame traitée précédente (gain, décalage vertical, décalage horizontal, décalage temporel) sont utilisés pour initialiser la recherche pour la trame traitée courante. Pour calculer la meilleure correspondance de trame d'origine pour la trame traitée courante, on suppose que le retard vidéo est constant. Par exemple, s'il a été déterminé que la meilleure correspondance pour la trame traitée **N** est la trame d'origine **M** dans les fichiers Big YUV, on suppose, au début de la recherche, que la meilleure correspondance pour la trame traitée **N+F** est la trame d'origine **M+F**.

D.6.1.5.2 Recherche large du décalage temporel

Une recherche complète parmi tous les décalages spatiaux possibles dans toute la plage d'incertitude temporelle pour chaque trame traitée nécessiterait un grand nombre de calculs. A la place, on utilise une recherche en plusieurs étapes, la première étape étant une recherche large du décalage temporel sur un ensemble très limité de décalages spatiaux, dont le but est de se rapprocher de la correspondance correcte de trame d'origine.

Dans le cadre de cette recherche large pour l'image traitée considérée, on examine la trame une de cette image (voir la Figure D.5) et on ne considère que les trames d'origine qui sont des trames unes et qui sont espacées de deux images (c'est-à-dire qui sont espacées de quatre trames) dans toute la plage correspondant à plus ou moins l'incertitude d'alignement temporel. Dans le cadre de cette recherche large, on considère les quatre décalages spatiaux suivants du flux vidéo traité: pas de décalage, huit pixels vers la gauche, huit pixels vers la droite et huit lignes de trame vers le haut (voir la Figure D.6). Sur la Figure D.6, les décalages positifs correspondent aux décalages vers le bas et vers la droite du flux vidéo traité par rapport au flux vidéo d'origine. Le décalage de "huit lignes de trame vers le bas" n'est pas envisagé car des observations empiriques ont montré que très peu de systèmes vidéo déplacent l'image vers le bas. La meilleure évaluation précédente du décalage spatial (c'est-à-dire associé à une trame traitée précédemment) est également incluse comme cinquième décalage possible lorsqu'elle est disponible. Pour déterminer la trame d'origine correspondant le mieux à la trame traitée considérée, on utilise la technique de comparaison décrite au § D.6.1.4.3. Le décalage temporel associé à la meilleure correspondance de trame d'origine devient le point de départ de l'étape suivante de l'algorithme, à savoir une recherche large du décalage spatial (§ D.6.1.5.3). Conformément au système de coordonnées de la Figure D.3, un décalage temporel positif signifie que le flux vidéo traité a été décalé dans le sens temporel positif (c'est-à-dire que le flux vidéo traité est retardé par rapport au flux vidéo d'origine). En ce qui concerne les fichiers Big YUV d'origine et traité, un décalage temporel positif signifie donc que des trames doivent être éliminées au début du fichier Big YUV traité alors qu'un décalage temporel négatif signifie que des trames doivent être éliminées au début du fichier Big YUV d'origine.

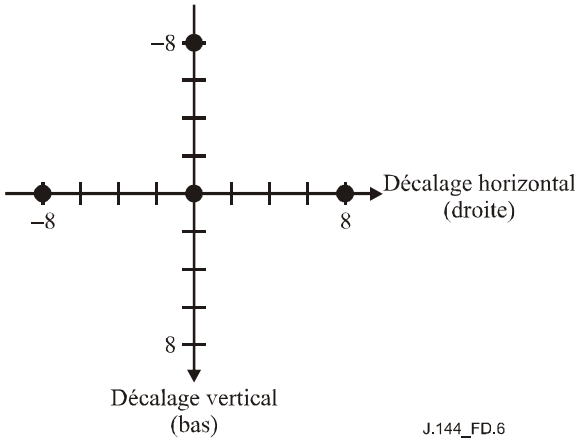


Figure D.6 – Décalages spatiaux envisagés dans le cadre de la recherche large du décalage temporel

D.6.1.5.3 Recherche large du décalage spatial

Compte tenu de l'alignement temporel déterminé par la recherche large du décalage temporel (§ D.6.1.5.2), on procède alors à une recherche large du décalage spatial sur une plage plus limitée de trames d'origine. La plage des trames d'origine qui sont considérées pour cette recherche comprend la trame d'origine de meilleure correspondance qui est une trame une (§ D.6.1.5.2) et les quatre trames d'origine les plus proches qui sont également des trames unes (trames unes des deux images qui précèdent et des deux images qui suivent la trame d'origine de meilleure correspondance). La recherche large du décalage spatial couvre la plage des décalages spatiaux donnée à la Figure D.7. Il est à noter que l'on envisage un moins grand nombre de décalages vers le bas (comme au § D.6.1.5.2), car ceux-ci sont moins fréquents dans la pratique. On applique alors la technique de comparaison décrite au § D.6.1.4.3 à l'ensemble de ces décalages spatiaux et de ces trames d'origine. Les meilleurs décalages temporel et spatial résultants servent alors d'évaluations améliorées pour l'étape suivante de l'algorithme décrite au § D.6.1.5.4.

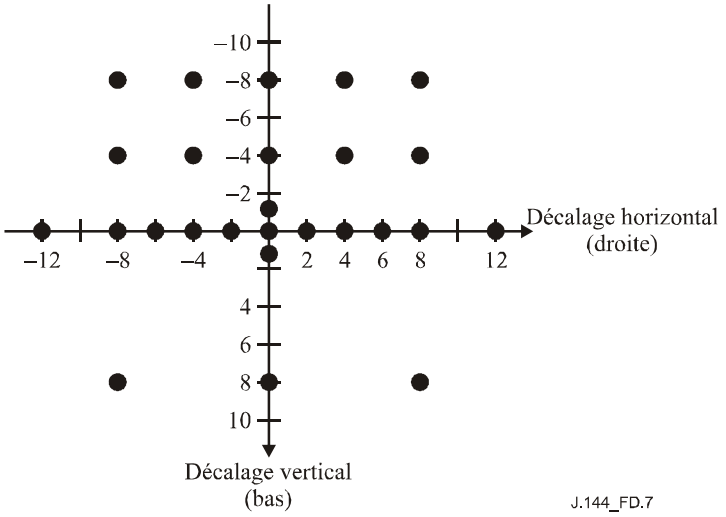


Figure D.7 – Décalages spatiaux envisagés dans le cadre de la recherche large du décalage spatial

D.6.1.5.4 Recherche fine du décalage spatio-temporel

Pour la recherche fine, on utilise un ensemble beaucoup plus petit de décalages spatiaux centrés autour de l'évaluation courante de l'alignement spatial et uniquement cinq trames centrées autour de la trame d'origine de meilleure correspondance. Ainsi, si cette trame est une trame une, on inclut dans la recherche trois trames unes et deux trames deux. Les décalages spatiaux qui sont envisagés comprennent l'évaluation courante du décalage, les huit décalages d'un pixel et/ou d'une ligne par rapport à l'évaluation courante, les huit décalages de deux pixels et/ou de deux lignes par rapport à l'évaluation courante du décalage, et le décalage nul (voir la Figure D.8). Dans l'exemple présenté sur la Figure D.8, l'évaluation courante du décalage spatial pour le flux vidéo traité est un décalage de 7 lignes de trame vers le haut et de 12 pixels vers la droite par rapport au flux vidéo d'origine. L'ensemble des décalages spatiaux présenté sur la Figure D.8 constitue un ensemble local presque complet d'alignements spatiaux proches de l'évaluation courante de l'alignement spatial. Le décalage nul est inclus comme condition de sécurité afin d'empêcher l'algorithme d'errer et de converger vers un minimum local. On applique alors avec soin la technique de comparaison décrite au § D.6.1.4.3 à l'ensemble de ces décalages spatiaux et de ces trames d'origine. Les meilleurs décalages temporel et spatial résultants servent alors d'évaluations améliorées pour l'étape suivante de l'algorithme décrite au § D.6.1.5.5.

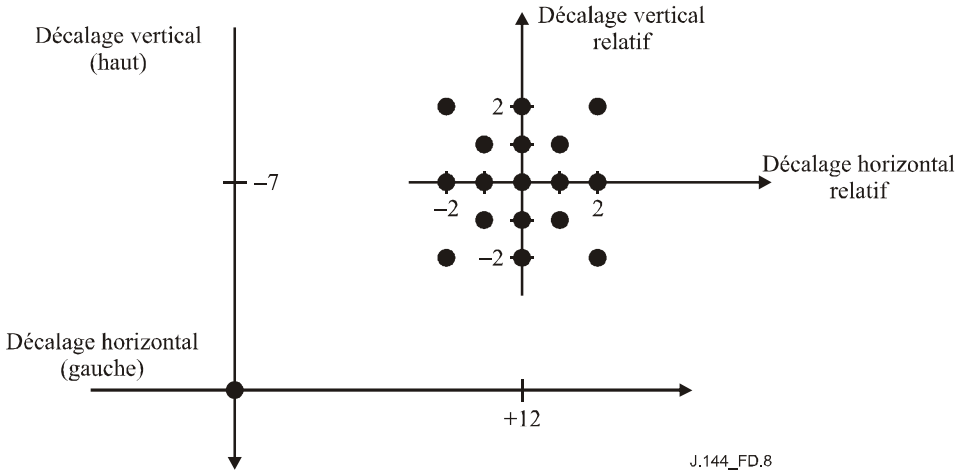


Figure D.8 – Décalages spatiaux envisagés dans le cadre de la recherche fine du décalage spatial

D.6.1.5.5 Recherches fines répétées

Lorsqu'on procède à une itération de la recherche fine décrite au § D.6.1.5.4, l'évaluation courante du décalage spatial se rapproche du décalage spatial réel ou (plus rarement) d'un faux minimum. De même, lorsqu'on procède à une telle itération, l'évaluation courante de la trame d'origine de meilleure correspondance se rapproche de la trame d'origine de meilleure correspondance réelle ou (plus rarement) d'un faux minimum. Ainsi, chaque recherche fine rapproche ces évaluations d'une valeur stable. Comme les recherches fines portent sur une zone très limitée spatialement et temporellement, elles doivent être répétées afin de s'assurer que la convergence a été atteinte. En cas d'utilisation de la compensation de gain, le gain de la trame traitée est réévalué à chaque recherche fine (voir le § D.6.1.4.2).

Les recherches fines portant sur la trame traitée (§ D.6.1.5.4) sont répétées jusqu'à ce que le meilleur décalage spatial et la trame d'origine associée à ce décalage spatial restent inchangés d'une recherche à la suivante. On cesse de répéter les recherches fines si l'algorithme alterne entre deux décalages spatiaux (par exemple un décalage horizontal de 3 puis un décalage horizontal de 4, toutes les autres grandeurs gardant les mêmes valeurs). Cette alternance apparaît lorsque la meilleure évaluation courante du décalage spatial et la trame d'origine associée à ce décalage spatial sont identiques à celles qui ont été déterminées deux itérations avant.

Parfois, les recherches répétées ne parviennent pas à converger. En l'absence de convergence au bout d'un certain nombre maximal d'itérations demandées, l'algorithme est arrêté et une condition "d'échec de la détermination du décalage" est signalée pour cette trame traitée. Ce cas particulier ne pose généralement pas de problème car de multiples trames traitées sont examinées pour chaque scène (§ D.6.1.5.6) et de multiples scènes sont examinées pour chaque circuit fictif de référence (§ D.6.1.5.7).

D.6.1.5.6 Algorithme pour une scène donnée

On commence par calculer une évaluation de base (de départ) du décalage vertical, du décalage horizontal et de l'alignement temporel sans compensation de gain comme suit. On saute les premières images du fichier Big YUV traité correspondant à l'incertitude temporelle (**U**). Une recherche large du décalage temporel est appliquée à la trame traitée suivante qui est une trame une (§ D.6.1.5.2). Il est à noter que cette recherche large porte sur les $\mathbf{U} \times 2 + 1$ premières images de la séquence vidéo d'origine afin de trouver la trame une de meilleure correspondance. On procède alors à une recherche large du décalage spatial, centrée sur cette trame d'origine de meilleure correspondance (§ D.6.1.5.3). On procède ensuite à un maximum de cinq recherches fines afin d'affiner les évaluations du décalage spatial et du décalage temporel (§ D.6.1.5.4 et 6.1.5.5). Si ces

recherches fines répétées n'aboutissent pas à un résultat stable, on élimine cette trame traitée de l'ensemble des trames considérées. On répète la procédure ci-dessus pour chaque image correspondant à une certaine fréquence (F) jusqu'à ce qu'on trouve une trame d'origine qui soit une trame une et qui produise des résultats stables. L'évaluation de base sera mise à jour régulièrement, comme décrit ci-dessous.

Les évaluations du décalage spatial sont calculées pour les deux trames d'une image du fichier Big YUV traité comme suit. En utilisant l'évaluation de base comme point de départ, on applique un maximum de trois recherches fines à la première trame traitée qui est une trame une. Si l'évaluation de base est correcte ou pratiquement correcte, les recherches fines répétées conduiront à un résultat stable. Si c'est le cas, le décalage spatial et le décalage temporel pour cette trame traitée sont stockés dans une matrice réservée au stockage des résultats relatifs aux trames unes. Si aucun résultat stable n'est trouvé, il est très probable que le décalage spatial est correct mais que l'évaluation du décalage temporel est aberrante (c'est-à-dire qu'elle est éloignée de plus de deux images du décalage temporel réel). On procède alors à une recherche large du décalage temporel qui inclut la meilleure évaluation courante du décalage spatial. Cette recherche large permet généralement de corriger l'évaluation du décalage temporel. Lorsque cette recherche est terminée, son résultat est utilisé comme point de départ et on procède à un maximum de cinq recherches fines. Si cette deuxième série de recherches fines n'aboutit pas à un résultat stable, on signale alors un échec d'alignement spatial pour l'image considérée (c'est-à-dire à la fois pour la trame une et pour la trame deux). Si cette deuxième série aboutit à un résultat stable, le décalage spatial et le décalage temporel pour cette trame sont stockés dans la matrice des trames unes. Par ailleurs, le décalage spatial et le décalage temporel utilisés comme point de départ pour la trame traitée suivante qui est une trame une sont mis à jour (autrement dit, on utilise les résultats de base pour la première trame traitée et, ensuite, on utilise le dernier résultat stable). Une fois que le décalage spatial a été évalué pour la première trame traitée qui est une trame une, on évalue le décalage spatial pour la première trame traitée qui est une trame deux. En utilisant les résultats spatiaux de la trame une comme point de départ, on applique les mêmes étapes pour trouver le décalage spatial de la trame deux (c'est-à-dire les trois recherches fines et, si nécessaire, une recherche large du décalage temporel suivie par cinq recherches fines). Si un résultat stable est trouvé pour la trame deux, on stocke le décalage vertical et le décalage horizontal de la trame deux dans une matrice différente qui est réservée au stockage des résultats pour les trames deux.

On applique la procédure décrite dans le paragraphe ci-dessus pour évaluer le décalage spatial des deux trames de chaque image correspondant à la fréquence (**F**) du fichier Big YUV qui contient la séquence vidéo traitée. On saute les premières images du fichier Big YUV traité correspondant à l'incertitude temporelle (**U**). On utilise alors cette séquence d'évaluations pour calculer une évaluation robuste du décalage spatial pour les trames unes de la scène considérée et une évaluation robuste du décalage spatial pour les trames deux de cette scène. On trie les résultats de décalage vertical de la trame une de chaque image et on retient la valeur du 50^{e} percentile comme valeur globale du décalage vertical pour les trames unes. De même, on trie les résultats de décalage vertical de la trame deux de chaque image et on retient la valeur du 50^{e} percentile comme valeur globale du décalage vertical pour les trames deux. On trie les résultats de décalage horizontal de la trame une de chaque image et on retient la valeur du 50^{e} percentile comme valeur globale du décalage horizontal. Toute différence entre le décalage horizontal des trames unes et celui des trames deux est très probablement due à un décalage horizontal sous-pixel (par exemple un décalage horizontal de 0,5 pixel). Les décalages horizontaux sous-pixel conduisent à des évaluations qui incluent les deux décalages les plus proches. L'utilisation de la valeur du 50^{e} percentile permet de choisir le décalage horizontal le plus probable, conduisant à une précision de l'alignement spatial à 0,5 pixel près[7].

---

[7] Alignement spatial à 0,5 pixel près est suffisant pour les mesures de la qualité vidéo décrites dans la présente annexe. Les techniques d'alignement spatial sous-pixel sortent du cadre de la présente annexe.

D.6.1.5.7 Algorithme pour un circuit fictif de référence donné

Si plusieurs scènes sont passées par le même circuit fictif de référence, les résultats de l'alignement spatial pour chaque scène devraient être identiques. Ainsi, le filtrage des résultats obtenus pour de multiples scènes permet d'augmenter la robustesse et la précision des mesures du décalage spatial. On peut alors utiliser les résultats globaux d'alignement spatial obtenus pour le circuit fictif de référence considéré pour procéder à une compensation pour toutes les séquences vidéo traitées par ce circuit.

D.6.1.5.8 Commentaires concernant l'algorithme

Certaines scènes vidéo ne conviennent pas vraiment pour l'évaluation de l'alignement spatial. L'algorithme décrit aura parfois pour résultat un faux minimum. D'autres fois, il errera entre plusieurs solutions et ne donnera jamais de résultat stable. C'est pourquoi il est conseillé d'examiner de multiples images d'une même scène et de déterminer la valeur médiane (c'est-à-dire de trier les résultats de la valeur la plus faible à la valeur la plus élevée et de choisir la valeur du 50^{e} percentile) de ces résultats sur plusieurs scènes. L'algorithme d'alignement spatial fondé sur des scènes est un algorithme heuristique utilisant les décalages spatiaux qui ont été observés pour un échantillon de systèmes vidéo. Ces hypothèses peuvent être incorrectes pour certains systèmes, auquel cas l'algorithme détermine un décalage spatial incorrect. Toutefois, lorsque l'algorithme donne des résultats incorrects, il a tendance à produire des décalages spatiaux qui sont incohérents d'une image à l'autre et d'une scène à l'autre (autrement dit, lorsque l'algorithme donne des résultats incorrects, il produit généralement des résultats épars). Lorsque l'algorithme a pour résultat le même décalage spatial ou des décalages spatiaux très semblables pour chaque scène, cela indique un niveau de confiance élevé. En cas de résultats épars pour les trames d'une scène donnée, cela indique un niveau de confiance faible.

D.6.1.6 Alignement spatial d'un flux vidéo avec balayage progressif

L'alignement spatial d'un flux vidéo avec balayage progressif suit le même algorithme que dans le cas d'un flux vidéo avec balayage à entrelacement, avec quelques légères modifications. L'algorithme dans le cas du balayage avec entrelacement s'applique séparément à la trame une et à la trame deux, alors que l'algorithme dans le cas du balayage progressif s'applique à l'image entière. Ainsi, il faut ignorer toutes les mentions de trame deux et, à l'exception des recherches fines, il faut doubler la plage des décalages verticaux.

La modification de la plage des décalages verticaux est particulièrement importante pour la recherche large du décalage spatial. Pour une telle recherche (§ D.6.1.5.3), il faut doubler les nombres sur l'axe vertical de la Figure D.7 (par exemple +8 devient +16 et –4 devient –8)[8]. Par ailleurs, dans le cas des images CIF et QCIF à balayage progressif, les plages de décalage horizontal et de décalage vertical utilisées pour les recherches larges sont réduites de moitié car les décalages observés avec ces formats d'image sont généralement plus petits. Par exemple, dans le cas d'images CIF, l'axe horizontal de la Figure D.7 irait de –6 à +6 pixels et l'axe vertical irait de –8 à +8 lignes d'image.

La plage utilisée pour la recherche du décalage temporel, spécifiée en nombre d'images, reste essentiellement la même. Pour la recherche large du décalage temporel décrite au § D.6.1.5.2, au lieu de comparer une trame traitée une avec une trame d'origine une sur deux, l'algorithme dans le cas du balayage progressif compare une image traitée avec une image d'origine sur deux. Concernant l'algorithme pour la mire chromatique, la recherche examine les décalages spatiaux

[8] Il existe une exception possible à ce doublement: le décalage spatial de zéro pixel horizontalement et de plus ou moins une ligne de trame verticalement peut être laissé à plus ou moins une ligne d'image verticalement. Les décalages spatiaux très proches de (zéro, zéro) sont fréquents.

entre une seule image traitée et une seule image d'origine (autrement dit il n'y a pas de recherche de décalage temporel).

La seule étape qui nécessite des modifications plus complexes est l'étape de recherche fine du § D.6.1.5.4. Dans cette étape, les décalages verticaux restent inchangés, compris entre –2 lignes d'image et +2 lignes d'image. Ainsi, les nombres représentés sur l'axe vertical de la Figure D.8 sont interprétés comme étant des nombres de lignes d'image. On peut définir la plage des décalages temporels pour cette recherche fine comme comprenant les cinq images d'origine centrées sur l'image d'origine courante, au lieu des trois images d'origine susmentionnées. Une plage de cinq images peut améliorer la vitesse et l'efficacité de la recherche fine par rapport à l'algorithme dans le cas du balayage à entrelacement, car les circuits fictifs de référence à balayage progressif ont davantage tendance à engendrer des retards vidéo plutôt que des décalages spatiaux non nuls.

Lorsqu'on examine les modifications à apporter à l'algorithme utilisé pour les systèmes vidéo à balayage progressif, il est possible de modifier de nombreux paramètres utilisés pour la recherche du décalage spatial sans compromettre l'intégrité de l'algorithme. Considérons, à titre d'exemple, les décalages spatiaux autres que zéro pixel et zéro ligne utilisés pour la recherche large du décalage temporel. Le décalage spatial de zéro pixel horizontalement et de 8 lignes de trame verticalement utilisé pour les systèmes à balayage à entrelacement peut être porté à 16 lignes d'image pour les systèmes à balayage progressif, comme recommandé plus haut, ou fixé à 8 lignes d'image, si on suppose qu'il est peu probable que des séquences vidéo à balayage progressif contiennent un décalage vertical de 16 lignes d'image. De même, un décalage spatial de zéro ligne verticalement et de 8 pixels horizontalement peut être porté à 9 ou 10 pixels horizontalement sans effets préjudiciables. Autre exemple: le nombre exact de répétitions de la recherche fine peut être augmenté ou diminué pour des applications particulières. Les valeurs exactes recommandées ici sont nettement moins élevées que dans la structure réelle de l'algorithme de recherche.

D.6.2 Région valable

Les séquences vidéo NTSC (525 lignes) et PAL (625 lignes) échantillonnées conformément à la Rec. UIT-R BT.601-5 sont susceptibles d'avoir une bordure de pixels et de lignes qui ne contient pas d'information d'image. Il est possible que la séquence vidéo d'origine saisie par la caméra ne remplisse qu'une partie de l'image telle qu'elle est définie dans la Rec. UIT-R BT.601-5. Un système vidéo numérique qui utilise une compression risque de réduire encore la zone de l'image afin de réduire le nombre de bits transmis. Si les pixels et les lignes qui ne sont pas transmis se trouvent dans la zone de surbalayage de l'image de télévision, l'utilisateur final ne devrait pas remarquer qu'il manque des lignes et des pixels. Si les pixels et les lignes qui ne sont pas transmis dépassent la zone de surbalayage, l'observateur pourra remarquer une bordure noire tout autour de l'image, car le système insérera généralement du noir dans cette zone d'image non transmise. Les systèmes vidéo (notamment ceux qui procèdent à un filtrage passe-bas) risquent de causer une avancée de la bordure noire dans la zone d'image. La plupart du temps, ces effets transitoires ont lieu à gauche et à droite de l'image mais ils peuvent aussi avoir lieu en haut ou en bas. Par ailleurs, la séquence vidéo traitée peut parfois contenir plusieurs lignes de données vidéo altérées en haut ou en bas de l'image que l'observateur ne verra pas nécessairement (les magnétoscopes VHS altèrent plusieurs lignes en bas de l'image dans la zone de surbalayage). Afin d'éviter que les zones ne contenant pas d'information d'image aient une incidence sur les mesures de la qualité VQM, il convient d'exclure ces zones de ces mesures. L'algorithme automatisé de la région valable présenté ici évalue la région valable du flux vidéo d'origine et du flux vidéo traité de sorte que, pour les calculs suivants, on ne tienne pas compte des lignes altérées en haut et en bas de l'image telle qu'elle est définie dans la Rec. UIT-R BT.601-5, des pixels de la bordure noire ou des effets transitoires où la bordure noire avance dans la zone d'image.

D.6.2.1 Algorithme principal de la région valable

Le présent paragraphe décrit l'algorithme principal de la région valable qui est appliqué à une seule image d'origine ou traitée. Cet algorithme nécessite trois arguments d'entrée: une image, une région valable maximale et l'évaluation de la région valable courante.

- **Image**. L'algorithme principal utilise l'image de luminance définie dans la Rec. UIT-R BT.601-5 associée à une seule image vidéo. Pour la mesure de la région valable d'une séquence vidéo *traitée*, tout décalage spatial imposé par le système vidéo doit avoir été supprimé de l'image de luminance avant que l'algorithme principal ne soit appliqué (voir le § D.6.1 – Alignement spatial).
- **Région valable maximale**. L'algorithme principal ne tiendra pas compte des pixels et des lignes qui se trouvent en dehors d'une région vidéo valable maximale. Cela permet à l'utilisateur de spécifier une région valable maximale qui est plus petite que la zone entière de l'image si des informations *a priori* indiquent que des pixels ou des lignes de l'image échantillonnée ont été altérés (voir le § D.6.2).
- **Région valable courante**. La région valable courante est une évaluation de la région valable qui est entièrement comprise dans la région valable maximale. Tous les pixels de la région valable courante contiennent une information vidéo valable; les pixels qui sont situés en dehors de cette région contiennent une information vidéo qui peut être soit valable soit non valable. Au départ, on prend, comme région valable courante, la plus petite zone possible située exactement au centre de l'image.

L'algorithme principal examine la zone vidéo comprise entre la région valable maximale et la région valable courante. Si certains de ces pixels contiennent une information vidéo valable, la région valable courante est élargie. L'algorithme est alors décrit en détail pour la partie gauche de l'image.

1) Calculer le niveau moyen de la colonne de pixels la plus à gauche de la région valable maximale. Cette colonne est désignée par "J–1" et la moyenne est représentée par "M_{J-1}".
2) Calculer le niveau moyen de la colonne de pixels suivante, "M_J".
3) La colonne J est déclarée comme contenant des informations vidéo non valables si elle est noire ($M_J < 20$) ou si le niveau moyen des pixels pour des colonnes successives indique une avancée de la bordure noire dans l'image valable ($M_J - 2 > M_{J-1}$). Si l'une de ces conditions est remplie, incrémenter J et répéter les étapes 2 et 3. Dans les autres cas, aller à l'étape 4.
4) Si la colonne finale J se trouve dans la région valable courante, aucune nouvelle information n'a été obtenue. Dans le cas contraire, mettre à jour la région valable courante avec J comme coordonnée de gauche.

L'algorithme permettant de déterminer le haut de l'image est analogue à celui qui est présenté ci-dessus pour la partie gauche. Pour le bas et la partie droite, J est décrémenté au lieu d'être incrémenté; à cette exception près, l'algorithme est le même. Les valeurs obtenues pour le haut, la gauche, le bas et la droite désignent le dernier pixel ou la dernière ligne valable.

Le contenu de la scène peut être tel que l'une des conditions spécifiées à l'étape 3 est remplie alors qu'elle ne devrait pas l'être. Par exemple, dans le cas d'une image qui contient du noir intentionnel dans la partie gauche (autrement dit du noir qui fait partie de la scène), l'algorithme principal conclura que la colonne vidéo valable la plus à gauche est beaucoup plus proche du milieu de l'image qu'elle ne devrait être. C'est pourquoi l'algorithme principal est appliqué à de multiples images issues d'une séquence vidéo, ce qui permet d'accroître la précision de l'évaluation de la région valable.

D.6.2.2 Application de l'algorithme principal de la région valable à une séquence vidéo

D.6.2.2.1 Séquence vidéo d'origine

L'algorithme principal est d'abord appliqué à la séquence d'images d'origine. Concernant les séquences vidéo NTSC échantillonnées conformément à la Rec. UIT-R BT.601-5 (§ D.5), il est recommandé que la région valable maximale soit telle que haut = 6, gauche = 6, bas = 482, droite = 714. Concernant les séquences vidéo PAL échantillonnées conformément à la Rec. UIT-R BT.601-5, il est recommandé que la région valable maximale soit telle que haut = 6, gauche = 16, bas = 570, droite = 704. L'algorithme principal est appliqué à la première image de la séquence vidéo et à chaque image ultérieure correspondant à une certaine fréquence. Par exemple, si la fréquence spécifiée vaut 15, l'algorithme principal examine les images de la séquence numéros 0, 15, 30, 45, etc. Une fois que toutes les images de la séquence ont été examinées, la région valable courante contient la plus grande zone valable parmi toutes les images examinées dans la séquence vidéo. Les pixels et les lignes qui sont compris entre cette région valable courante finale et la région valable maximale sont considérés comme contenant du noir ou une avancée transitoire du noir.

La région valable finale doit contenir un nombre pair de lignes et un nombre pair de pixels. Si la coordonnée du haut est impaire, elle est incrémentée de un. De même, si la coordonnée de gauche est impaire, elle est incrémentée de un. Ensuite, si la région contient un nombre impair de lignes, on décrémente la coordonnée du bas; de même, si la région contient un nombre impair de pixels (horizontalement), on décrémente la coordonnée de droite. Cela permet de simplifier le traitement chromatique des séquences vidéo échantillonnées conformément à la Rec. UIT-R BT.601-5, car la fréquence d'échantillonnage des canaux de couleur vaut la moitié de la fréquence d'échantillonnage du canal de luminance. Par ailleurs, chaque trame vidéo entrelacée contient le même nombre de lignes vidéo. Cela permet de garantir que les sous-régions spatio-temporelles (à partir desquelles les caractéristiques sont extraites) contiennent toujours des informations vidéo valables avec des contributions égales des deux trames entrelacées. La région valable résultante est retournée comme étant la région valable d'origine.

D.6.2.2.2 Séquence vidéo traitée

Pour le calcul de la région valable de la séquence vidéo traitée, on considère d'abord que la région valable maximale pour l'algorithme principal est égale à la région valable d'origine correspondante déterminée pour cette scène. On réduit ensuite la taille de cette région valable maximale en supprimant les pixels et les lignes considérés comme non valables par suite de l'alignement spatial des images vidéo traitées. L'algorithme principal est alors appliqué à la première image de la séquence vidéo traitée et à chaque image ultérieure correspondant à une certaine fréquence (si la fréquence vaut F, on utilise les images **Y**(0), **Y**(F), **Y**(2F), **Y**(3F), etc.).

Une fois que l'algorithme principal a été appliqué à la séquence vidéo traitée, la région valable déterminée par l'algorithme principal est réduite vers l'intérieur par une marge de sécurité. La marge de sécurité recommandée est de une ligne en haut et en bas et de cinq pixels à gauche et à droite. Les valeurs élevées à gauche et à droite permettent de garantir que toute avancée et tout recul du noir sont exclus de la région valable traitée.

La région valable traitée finale doit contenir un nombre pair de lignes et un nombre pair de pixels. Si la coordonnée du haut est impaire, elle est incrémentée de un. De même, si la coordonnée de gauche est impaire, elle est incrémentée de un. Ensuite, si la région contient un nombre impair de lignes, on décrémente la coordonnée du bas; de même, si la région contient un nombre impair de pixels (horizontalement), on décrémente la coordonnée de droite. La région valable résultante est retournée comme étant la région valable traitée.

D.6.2.3 Commentaires concernant l'algorithme de la région valable

Cet algorithme automatisé permet d'évaluer correctement la région valable de la plupart des scènes. En raison du très grand nombre de possibilités de contenu pour une scène, l'algorithme décrit ici est

fondé sur une approche prudente de l'évaluation de la région valable. Un examen manuel de la région valable conduirait certainement au choix d'une région plus grande. Les évaluations prudentes de la région valable conviennent mieux pour un système automatisé de mesure de la qualité vidéo, car l'élimination d'une faible quantité de contenu vidéo aura une faible incidence sur l'évaluation de la qualité et, de toute manière, ce contenu vidéo éliminé se trouvait généralement dans la zone de surbalayage de la séquence vidéo. En revanche, la prise en considération d'un contenu vidéo altéré dans les calculs de la qualité vidéo risque d'avoir une forte incidence sur l'évaluation de la qualité.

Cet algorithme ne contient pas une intelligence artificielle suffisante pour faire la distinction entre des pixels et des lignes altérés au bord d'une image et un véritable contenu de la scène. A la place, on utilise une règle empirique, selon laquelle un tel contenu vidéo non valable est généralement situé aux bords extrêmes de l'image. La spécification d'une région vidéo valable maximale prudente définissable par l'utilisateur (c'est-à-dire le point de départ de l'algorithme automatisé) permet de ne pas prendre en considération ces bords d'image éventuellement altérés.

Lorsque l'algorithme de la région valable est appliqué à une séquence vidéo qui n'est pas échantillonnée conformément à la Rec. UIT-R BT.601-5 (par exemple le format intermédiaire commun, ou CIF, utilisé par la Rec. UIT-T H.261), il est recommandé de prendre l'image entière comme région valable maximale lors de l'examen de la séquence vidéo d'origine. Dans ces cas, la séquence vidéo échantillonnée ne contient généralement pas de zone de surbalayage altérée, il est donc inutile de prendre une région valable maximale plus petite que l'image entière.

D.6.3 Gain et décalage

D.6.3.1 Algorithme principal du gain et du décalage de niveau

Le présent paragraphe expose la méthode à utiliser pour étalonner le gain et le décalage de niveau. Pour pouvoir appliquer cet algorithme, l'image d'origine et l'image traitée doivent être alignées spatialement (voir le § D.6.1). Elles doivent aussi être alignées temporellement (voir plus loin le § D.6.4). L'étalonnage du gain et du décalage de niveau peut être appliqué aux trames ou aux images, selon le cas.

Dans la méthode présentée ici, on suppose que chacun des signaux Y, C_B et C_R de la Rec. UIT-R BT.601-5 a un gain et un décalage de niveau indépendants. Cette hypothèse sera généralement suffisante pour l'étalonnage dans le cas des systèmes vidéo en composantes (par exemple Y, R-Y, B-Y). Toutefois, dans le cas des systèmes composites ou S-vidéo, il est possible d'avoir une rotation de phase des informations de chrominance car les deux composantes de chrominance sont multiplexées dans un vecteur de signal complexe comprenant une amplitude et une phase. L'algorithme présenté ici ne permet pas de procéder à un étalonnage correct dans le cas des systèmes vidéo qui introduisent une rotation de phase des informations de chrominance (par exemple l'ajustement de teinte sur un poste de télévision).

Comme indiqué précédemment, on suppose dans ce modèle d'étalonnage qu'il n'existe aucun couplage croisé entre les trois composantes vidéo. Cela étant, l'algorithme principal d'étalonnage est appliqué de manière indépendante à chacun des trois canaux: Y, C_B et C_R.

La région valable du plan d'image d'origine et celle du plan d'image traitée sont d'abord subdivisées en N sous-régions. Pour chacune des sous-régions, on calcule la valeur moyenne *origine* et la valeur moyenne *traité* (moyenne dans l'espace). On représente ensuite ces valeurs sous la forme des vecteurs colonnes à N éléments $\underline{O}$ et $\underline{P}$, respectivement:

$$\underline{O}_{N\,x1} = \begin{bmatrix} origine_1 \\ . \\ . \\ . \\ origine_N \end{bmatrix}, \underline{P}_{N\,x1} = \begin{bmatrix} traité_1 \\ . \\ . \\ . \\ traité_N \end{bmatrix}$$

Pour l'étalonnage, il faut calculer le gain (g) et le décalage de niveau (l) conformément au modèle suivant:

$$\underline{P} = g\underline{O} + l\,.$$

Comme il n'y a que deux inconnues (g et l) mais N équations (N sous-régions), il faut résoudre le système d'équations linéaires surdéterminé donné par:

$$\hat{\underline{P}} = A\begin{bmatrix} l \\ g \end{bmatrix}.$$

où A est une matrice N × 2 donnée par $A_{N\times 2}=[\underline{1} \quad \underline{O}]$, et $\underline{1}$ est un vecteur colonne à N éléments valant "1" donné par:

$$\underline{1}_{N\,x1} = \begin{bmatrix} 1_1 \\ . \\ . \\ . \\ 1_N \end{bmatrix}$$

$\hat{P}$ est l'évaluation des échantillons traités découlant de l'application du gain et du décalage de niveau aux échantillons d'origine. La solution donnée par les moindres carrés à ce problème surdéterminé (à condition que N > 2) est donnée par:

$$\begin{bmatrix} l \\ g \end{bmatrix} = \left(A^{\mathrm{T}}A\right)^{-1}A^{\mathrm{T}}P\,.$$

où l'exposant "T" désigne la transposée de la matrice et l'exposant "–1" désigne l'inverse de la matrice.

Lorsque l'algorithme principal du gain et du décalage de niveau est appliqué de manière indépendante à chacun des trois canaux, six grandeurs sont évaluées: gain Y, décalage Y, gain C_B, décalage C_B, gain C_R et décalage C_R.

D.6.3.2 Utilisation de scènes

L'algorithme de base donné au § D.6.3.1 peut être appliqué à des flux vidéo d'origine et traité sous réserve qu'ils aient été alignés spatialement et temporellement. Cette technique fondée sur les scènes subdivise l'image en blocs contigus de niveau d'intensité inconnu. Une taille de sous-région de 16 lignes × 16 pixels est recommandée pour les images (c'est-à-dire 8 lignes × 16 pixels pour une trame NTSC ou PAL Y; 8 lignes × 8 pixels pour C_B et C_R en raison du sous-échantillonnage des plans de couleur). La moyenne dans l'espace des échantillons [Y, C_B, C_R] est calculée pour chaque sous-région ou bloc d'origine et sous-région ou bloc traité correspondant, afin de former une image sous-échantillonnée spatialement. Tous les blocs choisis doivent se trouver dans la région valable traitée (PVR).

D.6.3.2.1 Alignement des images traitées

Dans un souci de simplicité, on suppose que le meilleur alignement spatial a déjà été déterminé au moyen de l'une des techniques présentées au § D.6.1. Pour pouvoir évaluer le gain et le décalage de niveau, chaque image traitée doit être alignée temporellement. L'image d'origine qui correspond le mieux à l'image traitée doit être utilisée pour le calcul du gain et du décalage de niveau. Si le retard vidéo est variable, cet alignement temporel doit être opéré pour chaque image traitée. Si le retard vidéo est constant pour la scène, il n'est nécessaire d'opérer l'alignement temporel qu'une seule fois.

Pour aligner temporellement une image traitée, on commence par créer les trames d'origine et traitée sous-échantillonnées spatialement (ou les images dans le cas du balayage progressif) comme spécifié au § D.6.3.2, après avoir corrigé le décalage spatial du flux vidéo traité. En utilisant les images Y sous-échantillonnées, on applique la fonction de recherche donnée au § D.6.1.4.3, à moins qu'on effectue cette recherche en utilisant toutes les images d'origine correspondant à l'incertitude d'alignement temporel (**U**). On utilise le meilleur alignement temporel résultant pour les trois plans d'image, Y, C_B et C_R.

D.6.3.2.2 Gain et décalage de niveau des images alignées

On utilise une solution itérative donnée par les moindres carrés avec une fonction de coût afin de réduire au minimum le poids des valeurs aberrantes dans l'ajustement. En effet, les valeurs aberrantes sont généralement dues à des distorsions et non à de simples modifications du décalage de niveau et du gain, de sorte que l'attribution d'un poids égal à ces valeurs aberrantes conduirait à une distorsion de l'ajustement.

L'algorithme suivant est appliqué séparément aux N pixels d'origine et traités correspondants issus de chacune des trois images sous-échantillonnées spatialement [Y, C_B, C_R].

1) Utiliser la solution normale donnée par les moindres carrés (voir le § D.6.3.1) pour générer l'évaluation initiale du décalage de niveau et du gain: $\begin{bmatrix} l \\ g \end{bmatrix} = \left(A^{\mathrm{T}} A\right)^{-1} A^{\mathrm{T}} \underline{P}$.

2) Générer un vecteur d'erreur ($\underline{E}$) qui est égal à la valeur absolue de la différence entre les échantillons traités réels et les échantillons traités ajustés: $\underline{E} = \left|\underline{P} - \underline{\hat{P}}\right|$.

3) Générer un vecteur de coût ($\underline{C}$) dont chaque élément est le réciproque de l'élément correspondant du vecteur d'erreur ($\underline{E}$) plus un petit epsilon (ε): $\underline{C} = \dfrac{1}{\underline{E} + \varepsilon}$. ε permet d'éviter la division par zéro et définit le poids relatif d'un point qui est sur la courbe ajustée par rapport au poids d'un point qui est en dehors de cette courbe. Il est recommandé d'utiliser une valeur de 0,1 pour ε.

4) Normaliser le vecteur de coût $\underline{C}$ (autrement dit, on divise chaque élément de $\underline{C}$ par la racine carrée de la somme des carrés de tous les éléments de $\underline{C}$).

5) Générer le vecteur de coût $\underline{C}^2$ dont chaque élément est le carré de l'élément correspondant du vecteur de coût $\underline{C}$ issu de l'étape 4).

6) Générer une matrice de coût diagonale N × N (C^2) qui contient les éléments du vecteur de coût ($\underline{C}^2$) sur la diagonale et des zéros partout ailleurs.

7) En utilisant la matrice de coût diagonale (C^2) issue de l'étape 6, procéder à un ajustement par les moindres carrés avec pondération par le coût pour déterminer l'évaluation suivante du décalage de niveau et du gain: $\begin{bmatrix} l \\ g \end{bmatrix} = \left(A^{\mathrm{T}} C^2 A\right)^{-1} A^{\mathrm{T}} C^2 \underline{P}$.

8) Répéter les étapes 2 à 7 jusqu'à ce que les évaluations du décalage de niveau et du gain convergent à la quatrième décimale près.

Ces étapes sont appliquées séparément à la trame traitée une et à la trame traitée deux, ce qui donne deux évaluations de *g* et deux évaluations de *l*. Il faut examiner séparément la trame une et la trame deux, car les trames d'origine alignées temporellement ne correspondent pas nécessairement à une même image dans la séquence vidéo d'origine. Dans le cas des systèmes vidéo à balayage progressif, les étapes ci-dessus sont appliquées à l'image traitée tout entière.

D.6.3.2.3 Evaluation du gain et du décalage de niveau pour une séquence vidéo et un circuit fictif de référence

L'algorithme décrit ci-dessus est appliqué à plusieurs couples trame d'origine – trame traitée correspondante répartis tout au long de la scène avec une certaine fréquence (dans le cas des systèmes vidéo à balayage progressif, on utilise des couples image d'origine – image traitée). On détermine alors la valeur médiane de chacun des six historiques temporels de décalages de niveau et de gains pour produire des évaluations moyennes pour la scène.

Si plusieurs scènes passent par le même circuit fictif de référence, le décalage de niveau et le gain pour chaque scène seront considérés comme identiques. Ainsi, les valeurs médianes obtenues à partir de plusieurs scènes permettent d'augmenter la robustesse et la précision des mesures de décalage de niveau et de gain. On peut alors utiliser les résultats globaux de décalage de niveau et de gain obtenus pour le circuit fictif de référence considéré pour procéder à une compensation pour tous les flux vidéo traités par ce circuit.

D.6.3.3 Application des corrections de gain et de décalage de niveau

Pour les algorithmes d'alignement temporel (voir le § D.6.4) et pour l'extraction de la plupart des caractéristiques de qualité (§ D.7), il convient de supprimer le gain calculé ici. Pour supprimer le gain et le décalage de niveau du plan Y, on applique la formule suivante à chaque pixel traité:

$$\text{Nouveau } Y(i,j,t) = [Y(i,j,t) - l] / g$$

Le gain et le décalage de niveau des plans de couleur (C_B et C_R) ne sont pas corrigés. A la place, on mesure les erreurs de chrominance perçues. Le gain et le décalage de niveau des plans d'image C_B et C_R peuvent être corrigés à des fins d'affichage.

D.6.4 Alignement temporel

Les systèmes de communication vidéo numériques modernes ont généralement besoin de plusieurs dixièmes de seconde pour traiter et transmettre le flux vidéo de la caméra au dispositif de visualisation. Des retards vidéo excessifs empêchent d'avoir une communication bidirectionnelle efficace. Les méthodes de mesure objective du retard de bout en bout pour les communications vidéo sont donc importantes pour les utilisateurs finals afin de pouvoir spécifier et comparer les services ainsi que pour les fournisseurs d'équipements/de services afin de pouvoir optimiser et mettre à jour leurs offres de produits. Le retard vidéo peut dépendre des attributs dynamiques de la scène d'origine (par exemple détail spatial, mouvement) et du système vidéo (par exemple débit binaire). A titre d'exemple, le retard vidéo risque d'être plus grand pour des scènes comportant beaucoup de mouvements que pour des scènes en comportant peu. Les mesures du retard vidéo devraient donc être faites en service afin d'être vraiment représentatives et précises. Il est nécessaire d'évaluer le retard vidéo pour pouvoir aligner temporellement les caractéristiques vidéo du flux d'origine et du flux traité (voir la Figure D.1) avant de procéder aux mesures de la qualité.

Certains systèmes de transmission vidéo peuvent fournir des informations de synchronisation temporelle (les images d'origine et traitées peuvent par exemple être étiquetées au moyen d'un certain type de système de numérotation d'image). Toutefois, la synchronisation temporelle entre le flux vidéo d'origine et le flux vidéo traité doit généralement être mesurée. Le présent paragraphe expose une technique permettant d'évaluer le retard vidéo sur la base des images vidéo d'origine et des images vidéo traitées. La technique est "fondée sur les images" en ce sens qu'elle consiste à corréler des images à plus faible résolution, sous-échantillonnées dans l'espace et extraites des flux

vidéo d'origine et traité. Cette technique fondée sur les images évalue le retard de chaque image ou de chaque trame (dans le cas des systèmes vidéo avec balayage à entrelacement). On combine ces différentes évaluations pour évaluer le retard moyen pour la séquence vidéo.

D.6.4.1 Algorithme fondé sur les images pour évaluer les décalages temporels variables entre une séquence vidéo d'origine et une séquence vidéo traitée

Le présent paragraphe décrit un algorithme d'alignement temporel fondé sur les images. Pour réduire l'influence des distorsions sur l'alignement temporel, les images sont sous-échantillonnées spatialement et normalisées de manière à avoir une variance unitaire. Cet algorithme permet d'aligner temporellement chaque image traitée séparément, en localisant l'image d'origine la plus analogue. Certaines de ces différentes mesures d'alignement temporel peuvent être incorrectes mais les erreurs ont tendance à être distribuées aléatoirement. Lorsqu'on attribue les mesures du retard issues d'une série d'images au moyen d'un système de vote, on obtient une évaluation globale du retard moyen d'une séquence vidéo relativement précise. Cet algorithme d'alignement temporel n'utilise pas les parties fixes ou pratiquement sans mouvement de la scène, car les images d'origine sont pratiquement identiques les unes aux autres.

D.6.4.1.1 Constantes utilisées par l'algorithme

BELOW_WARN: seuil utilisé lors de l'examen des corrélations afin de décider si un maximum de corrélation secondaire est suffisamment grand pour indiquer un alignement temporel ambigu. Il est recommandé d'utiliser une valeur de 0,9 pour BELOW_WARN.

BLOCK_SIZE: facteur de sous-échantillonnage, spécifié en nombre de lignes d'image verticalement et en nombre de pixels horizontalement. Il est recommandé d'utiliser une valeur de 16 pour BLOCK_SIZE.

DELTA: les maximums secondaires de la courbe de corrélation qui sont éloignés de moins de DELTA de la (meilleure) corrélation maximale sont ignorés. Il est recommandé d'utiliser une valeur de 4 pour DELTA.

HFW: la moitié de la largeur du filtre utilisé pour lisser l'histogramme des valeurs d'alignement temporel associées à chaque image. Il est recommandé d'utiliser une valeur de 3 pour HFW.

STILL_THRESHOLD: seuil utilisé pour détecter les scènes vidéo fixes (l'alignement temporel fondé sur les images ne peut pas être utilisé pour des scènes vidéo fixes). Il est recommandé d'utiliser une valeur de 0,002 pour STILL_THRESHOLD.

D.6.4.1.2 Variables d'entrée de l'algorithme

Une séquence de N images de luminance du flux vidéo d'origine: $\mathbf{Y_O}(t)$, $0 \le t < N$.[9]

Une séquence de N images de luminance du flux vidéo traité: $\mathbf{Y_P}(t)$, $0 \le t < N$.

Facteurs de gain et de décalage de niveau pour les images de luminance traitées.

Informations d'alignement spatial: décalage horizontal et décalage vertical. Dans le cas des systèmes vidéo avec balayage à entrelacement, le décalage vertical pour chaque trame permet de déterminer si le flux vidéo traité nécessite une resynchronisation de trame.

Région valable de la séquence vidéo traitée (PVR).

---

[9] Lorsqu'un flux vidéo avec balayage à entrelacement nécessite une resynchronisation de trame, la longueur des séquences d'origine et traitée doit être réduite de un afin de tenir compte de la resynchronisation de trame. La longueur du fichier sera donc ramenée à N –1 images vidéo (voir la Figure D.2).

Incertitude (U): nombre indiquant la précision de l'alignement temporel initial. On suppose au départ que le véritable alignement temporel pour $\mathbf{Y_P}(t)$ est compris entre plus ou moins (U – HFW) de $\mathbf{Y_O}(t)$, pour $0 \leq t < N$.

D.6.4.1.3 Images ou trames

L'algorithme d'alignement temporel fondé sur les images fonctionne à la fois pour les systèmes vidéo avec balayage à entrelacement et pour les systèmes vidéo avec balayage progressif. En cas de séquence vidéo avec balayage progressif, l'algorithme aligne des images. En cas de séquence vidéo avec balayage à entrelacement, l'algorithme aligne des trames. Lors de l'alignement de séquences vidéo avec balayage à entrelacement, soit des alignements d'image soit des alignements de trames resynchronisées sont considérés, mais pas les deux. Lorsque des alignements d'image sont considérés, la trame une de l'image vidéo traitée est comparée avec la trame une de l'image vidéo d'origine et la trame deux de l'image vidéo traitée est comparée avec la trame deux de l'image vidéo d'origine. Lorsque des alignements de trames resynchronisées sont considérés, la trame une de l'image vidéo traitée est comparée avec la trame deux de l'image vidéo d'origine et la trame deux de l'image vidéo traitée est comparée avec la trame une de l'image vidéo d'origine. Les valeurs d'alignement spatial fournies comme valeurs d'entrée de l'algorithme déterminent si ce sont des alignements d'image ou des alignements de trames resynchronisées qui sont considérés. Pour détecter la présence d'une resynchronisation de trame, on examine l'alignement spatial vertical pour chaque trame. Si le décalage vertical de la trame une est égal au décalage vertical de la trame deux, la séquence vidéo traitée n'a pas été soumise à une resynchronisation de trame; seuls des alignements d'image sont considérés. Si le décalage vertical de la trame deux vaut un de plus que le décalage vertical de la trame une, seuls des alignements de trames resynchronisées sont considérés. Toutes les autres combinaisons de décalages verticaux témoignent de l'existence de problèmes qu'il convient de régler avant l'alignement temporel.

D.6.4.1.4 Description de l'algorithme

1) *Etalonner les séquences vidéo*

   Il convient de corriger la séquence vidéo traitée, $\mathbf{Y_P}(t)$, en utilisant les informations d'alignement spatial et de gain-décalage données comme informations d'entrée de l'algorithme.

2) *Choisir la sous-région vidéo à utiliser*

   La sous-région d'intérêt à utiliser par l'algorithme doit être un multiple de BLOCK_SIZE et doit être comprise dans la PVR. Il convient de choisir la plus grande sous-région qui remplit ces deux conditions et qui est la plus proche du centre de l'image. L'ensemble du traitement ultérieur portera uniquement sur les informations vidéo présentes dans cette sous-région d'intérêt choisie.

3) *Sous-échantillonner spatialement les images d'origine et traitées*

   Il convient de sous-échantillonner spatialement la région d'intérêt de $\mathbf{Y_O}(t)$ et $\mathbf{Y_P}(t)$ par un facteur BLOCK_SIZE en calculant la moyenne de chaque bloc. Pour les images d'une séquence vidéo à balayage progressif, le sous-échantillonnage sera de BLOCK_SIZE horizontalement et verticalement, alors que pour les trames d'une séquence vidéo à balayage à entrelacement, le sous-échantillonnage sera de BLOCK_SIZE horizontalement et de BLOCK_SIZE/2 verticalement. A titre d'exemple, le sous-échantillonnage d'une séquence vidéo à balayage progressif par un BLOCK_SIZE de 16 prendra la moyenne de chaque bloc de 16 pixels par 16 lignes d'image, alors que le sous-échantillonnage d'une séquence vidéo à balayage à entrelacement par un BLOCK_SIZE de 16 prendra la moyenne de chaque bloc de 16 pixels par 8 lignes de trame. Ce sous-échantillonnage permet de réduire l'incidence des dégradations sur le processus d'alignement temporel.

4) *Normaliser les images sous-échantillonnées*

Il convient de normaliser chaque image sous-échantillonnée par l'écart type de cette image. On sautera cette normalisation pour toute image pour laquelle l'écart type est inférieur à un (par exemple pour les images contenant une trame de couleur uniforme)[10]. Cette normalisation permet de réduire au minimum l'influence des fluctuations du contraste et de l'énergie de chaque image sur les résultats de l'alignement temporel. Après cette étape, la séquence vidéo d'origine et la séquence vidéo traitée sont respectivement désignées par $\mathbf{S_O}(t)$ et $\mathbf{S_P}(t)$, afin d'indiquer que les images ont été sous-échantillonnées et normalisées.

5) *Comparer les images traitées avec les images d'origine*

Il convient de comparer chaque image traitée $\mathbf{S_P}(t)$ avec les images d'origine $\mathbf{S_O}(t + d)$, où les valeurs valables de d sont les suivantes: $(-U \leq d \leq +U)$ et les valeurs valables de t sont les suivantes: $(U \leq t < N - U)$. La comparaison entre une image traitée t et une image d'origine $t+d$, désignée par $\mathbf{C}_{td}$, est calculée comme étant l'écart type dans l'espace de l'image formée par la différence entre l'image d'origine $t+d$ et l'image traitée t: $\mathbf{C}_{td} = std_{\text{space}}(\mathbf{S_O}(t+d) - \mathbf{S_P}(t))$. Les comparaisons $\mathbf{C}_{td}$ permettent de corréler la $t^{\text{ième}}$ image traitée avec chaque image d'origine comprise dans une certaine plage d'incertitude d'alignement. Plus la valeur de $\mathbf{C}_{td}$ est faible, plus l'image traitée ressemble à l'image d'origine, étant donné qu'une plus grande partie de la variance d'image est annulée. La plage de t, $U \leq t < N - U$, couvre l'ensemble des images traitées pour lesquelles les images d'origine sont disponibles pour toute la plage d'incertitude d'alignement temporel.

6) *Vérifier globalement le degré de mouvement de la séquence vidéo*

Pour déterminer si la séquence contient suffisamment de mouvement, il convient de calculer la moyenne de $\mathbf{C}_{td}$ sur l'indice temporel t pour chaque d:

$$A_d = \frac{1}{N - 2 \times U} \sum_{t=U}^{N-U-1} \mathbf{C}_{td}$$

Cette sommation porte sur l'ensemble des images vidéo traitées t pour lesquelles toutes les images d'origine associées à l'incertitude considérée sont disponibles. A_d contient une valeur pour chaque décalage temporel d considéré. Si (maximum(A_d) – minimum(A_d) < STILL_THRESHOLD), la scène ne contient pas suffisamment de mouvement pour l'alignement temporel fondé sur les images. La scène entière est fixe ou quasiment fixe. Les résultats de corrélation pour les différents retards vidéo sont alors tellement analogues que toute différence sera due au hasard et non à des mesures fiables. En cas de détection d'une séquence vidéo fixe, l'utilisateur en est averti et l'algorithme prend alors fin.

7) *Aligner temporellement chaque image traitée*

Pour chaque image traitée t $(U \leq t < N - U)$, il convient de déterminer la valeur de d dans la plage d'incertitude temporelle $(-U \leq d \leq +U)$ qui minimise $\mathbf{C}_{td}$. En d'autres termes, pour chaque image traitée t, il convient de déterminer $d_{\min}(t)$ tel que $\mathbf{C}_{t\ d\min(t)} \leq \mathbf{C}_{td}$, quel que soit d. Le meilleur alignement temporel de l'image traitée t est donné par $d_{\min}(t)$. La plupart du temps, l'alignement temporel indiqué pour chaque image est correct ou très proche de l'alignement correct. Les cas où l'alignement temporel est incorrect peuvent s'expliquer par diverses raisons (distorsion d'image, erreurs, bruit, mouvement insuffisant, etc.).

[10] On saute la normalisation lorsque l'écart type est inférieur à un afin d'éviter toute amplification du bruit et d'éviter une éventuelle division par zéro pour les images qui contiennent un niveau d'intensité uniforme.

8) *Vérifier le degré de mouvement pour chaque image traitée*

Si, pour une image traitée t et pour toutes les valeurs de d ($-U \le d \le U$), maximum($\mathbf{C}_{td}$) – minimum($\mathbf{C}_{td}$) < STILL_THRESHOLD, alors $d_{min}(t)$ est indéfini pour cette image traitée t. Plus précisément, le mouvement est insuffisant autour de l'image t pour que l'alignement temporel fondé sur les images puisse fonctionner correctement.

9) *Etablir un histogramme de tous les alignements temporels définis*

Il convient d'établir un histogramme en utilisant toutes les valeurs définies de $d_{min}(t)$ avec 2×U+1 bâtons, chaque bâton représentant un retard vidéo différent (de –U à +U). Les valeurs de $d_{min}(t)$ qui sont indéfinies (par exemple images fixes) sont ignorées dans l'établissement de l'histogramme. Cet histogramme, désigné par H_d, est l'histogramme des décalages temporels pour toutes les images traitées qui contenaient suffisamment de mouvement pour pouvoir effectuer un alignement temporel valable. Chaque bâton de l'histogramme contient le nombre d'images traitées présentant un certain retard vidéo d, où d varie de –U à +U.

10) *Lisser l'histogramme*

On lisse l'histogramme H_d en procédant à sa convolution avec un filtre passe-bas de longueur 2×HFW+1 et défini à l'indice k par:

$$F_k = \frac{0.5 + 0.5 \times \cos[\pi \times (k - \mathrm{HFW})/(1 + \mathrm{HFW})]}{\sum_{i=0}^{2 \times \mathrm{HFW}} \{0.5 + 0.5 \times \cos[\pi \times (i - \mathrm{HFW})/(1 + \mathrm{HFW})]\}}, \quad 0 \le k \le 2 \times \mathrm{HFW}$$

Concernant l'histogramme lissé SH_d résultant de cette étape, les HFW bâtons à chaque extrémité de SH_d sont considérés comme indéfinis. Cela restreint les retards vidéo qui peuvent être évalués à plus ou moins (incertitude-HFW). Le lissage de l'histogramme permet d'augmenter la robustesse des évaluations du retard vidéo.

11) *Examiner les informations de l'histogramme*

A partir de l'histogramme d'origine H_d et de l'histogramme lissé SH_d, on détermine les trois valeurs suivantes:

max_H_value: valeur maximale de H_d.

max_SH_offset: décalage d qui maximise SH_d.

max_SH_value: valeur maximale de SH_d (c'est-à-dire pour d = max_SH_offset).

On procède ensuite aux deux vérifications suivantes:

- la valeur de U était-elle suffisamment élevée? On rappelle que les HFW premiers et les HFW derniers bâtons de H_d sont ôtés dans SH_d. On examine les valeurs de H_d dans ces bâtons. Si (H_d > max_H_value × BELOW_WARN), l'incertitude d'alignement temporel est trop faible. Il faut refaire tourner l'algorithme avec une valeur de U plus grande. Les valeurs de d à examiner sont ($-U \le d < -U + HFW$) et ($U - HFW < d \le U$).
- est-ce que SH_d a un retard bien défini? On examine SH_d, sauf pour les décalages situés à moins de DELTA de max_SH_offset. Si (SH_d > max_SH_value × BELOW_WARN) pour tout retard vidéo d tel que ($-U \le d <$ max_SH_offset – DELTA) ou (max_SH_offset + DELTA $< d \le U$), l'alignement temporel est ambigu.

Si la réponse aux deux vérifications ci-dessus est positive, on choisit le retard vidéo donné par max_SH_offset comme meilleur alignement temporel moyen pour la scène.

D.6.4.1.5 Observations et conclusions

L'algorithme de mesure du retard vidéo fondé sur les images utilise des séquences vidéo d'origine et traitée sous-échantillonnées. Il permet d'aligner des séquences vidéo dans un environnement hors

service entièrement automatisé, avant qu'il ne soit procédé aux mesures de la qualité vidéo. Cet algorithme évalue l'alignement temporel pour chaque image, établit des histogrammes avec les différentes évaluations puis utilise le retard le plus couramment indiqué comme retard vidéo global – ou alignement temporel – pour la séquence d'images vidéo considérée.

Le retard indiqué à la dernière étape de l'algorithme (étape 11 du § D.6.4.1.4) peut être différent du retard qu'un observateur choisirait s'il alignait les scènes visuellement. Les observateurs ont tendance à se concentrer sur le mouvement, alignant les parties de la scène présentant beaucoup de mouvement, alors que l'algorithme fondé sur les images détermine le retard le plus fréquemment observé parmi toutes les images examinées. Les histogrammes globaux de retard peuvent servir à déterminer l'amplitude et des statistiques de tout retard vidéo variable dû au circuit fictif de référence.

D.6.4.2 Application de la correction d'alignement temporel

Pour toutes les caractéristiques de qualité, il faut que le décalage temporel calculé ici soit supprimé. Pour les décalages positifs, on supprime des images au début du fichier traité et à la fin du fichier d'origine. Pour les décalages négatifs, on supprime des images à la fin du fichier traité et au début du fichier d'origine. En cas de resynchronisation de trame de séquences vidéo à balayage à entrelacement, la séquence traitée est soumise à une resynchronisation de trame. Il convient donc de supprimer une trame au début et à la fin de la séquence vidéo traitée en plus des suppressions susmentionnées. Il faut simultanément supprimer une image au début du fichier vidéo d'origine (pour un retard de trame global de –1) ou à la fin de ce fichier (pour un retard de trame global de +1).

La correction de l'alignement temporel a pour effet de réduire le nombre d'images disponibles dans la séquence vidéo. Dans un souci de simplicité, tous les calculs ultérieurs sont fondés sur le nombre d'images vidéo disponibles une fois que toutes les corrections liées à l'étalonnage ont été appliquées.

D.7 Caractéristiques de qualité

D.7.1 Introduction

Une *caractéristique* de qualité est définie comme étant une grandeur associée à – ou extraite d' – une sous-région spatio-temporelle d'un flux vidéo (d'origine ou traité). Les flux de caractéristiques qui sont produits sont fonction de l'espace et du temps. En comparant les caractéristiques extraites d'une séquence vidéo traitée étalonnée avec les caractéristiques extraites de la séquence vidéo d'origine étalonnée, on peut calculer un ensemble de *paramètres* de qualité (§ D.8) qui donnent une indication des modifications perçues de la qualité vidéo. Le présent paragraphe décrit un ensemble de caractéristiques de qualité qui caractérisent les modifications perçues des propriétés spatiales, temporelles et de chrominance des flux vidéo. Un filtre de perception est généralement appliqué au flux vidéo afin d'accentuer certaines propriétés de la qualité vidéo perçue, comme les informations de contour. Une fois ce filtrage opéré, les caractéristiques sont extraites des sous-régions spatio-temporelles (S-T) au moyen d'une fonction mathématique (par exemple un écart type). Enfin, un seuil de perceptibilité est appliqué aux caractéristiques extraites.

Dans ce qui suit, un flux de caractéristiques d'une séquence d'origine sera désigné par $f_o(s, t)$ et le flux de caractéristiques de la séquence traitée correspondante sera désigné par $f_p(s, t)$, où s et t sont des indices qui désignent respectivement la position spatiale et la position temporelle de la région S-T dans les flux vidéo d'origine et traité étalonnés. Pour nommer les caractéristiques, qui sont décrites dans les paragraphes qui suivent, on utilise des caractères en indice, ceux-ci étant choisis de manière à indiquer ce que la caractéristique mesure. Toutes les caractéristiques concernent des images d'une séquence vidéo étalonnée (voir le § D.6); les questions relatives à l'entrelacement étant abordées au moment de l'étalonnage. Toutes les caractéristiques sont

indépendantes de la taille d'image (autrement dit la taille de la région S-T ne varie pas quand la taille d'image varie)[11].

En résumé, les étapes du calcul des caractéristiques sont les suivantes. Pour certaines caractéristiques, les étapes marquées comme étant une [option] ne seront peut-être pas nécessaires.

1) [option] Appliquer un filtre de perception.
2) Subdiviser le flux vidéo en régions S-T.
3) Extraire les caractéristiques, ou les statistiques récapitulatives, de chaque région S-T (par exemple la moyenne, l'écart type).
4) [option] Appliquer un seuil de perceptibilité.

Pour certaines caractéristiques, on peut utiliser deux filtres de perception différents ou plus.

D.7.1.1 Régions S-T

En général, les caractéristiques sont extraites des régions S-T localisées après application d'un ou de plusieurs filtres de perception aux flux vidéo d'origine et traité. Les positions des régions S-T sont telles que les flux vidéo sont subdivisés en régions S-T contiguës. Comme le flux vidéo traité a été étalonné, pour chaque région S-T de ce flux, il existe une région S-T du flux d'origine ayant la même position spatiale et la même position temporelle dans le flux vidéo. Pour extraire les caractéristiques de chaque région S-T, on calcule des statistiques récapitulatives ou on applique une certaine autre fonction mathématique sur la région d'intérêt S-T.

Chaque région S-T correspond à un bloc de pixels. La taille d'une région S-T est décrite par:

1) le nombre de pixels horizontalement;
2) le nombre de lignes d'image verticalement;
3) la dimension temporelle de la région, donnée en nombre équivalent d'images vidéo d'un système vidéo à 30 images par seconde[12].

La Figure D.9 illustre une région S-T de 8 pixels horizontaux × 8 lignes verticales × 6 images vidéo NTSC, pour un total de 384 pixels. Dans le cas d'un système vidéo à 25 images par seconde (PAL), cette même région S-T couvre 8 pixels horizontaux × 8 lignes verticales × 5 images vidéo, pour un total de 320 pixels.

Un cinquième de seconde est une dimension temporelle souhaitable, en raison de la facilité de conversion entre les fréquences d'images (un cinquième de seconde donne un nombre entier d'images vidéo pour les systèmes vidéo fonctionnant à 10, 15, 25 et 30 images par seconde). La règle générale à appliquer pour la conversion entre fréquences d'images consiste à prendre la dimension de la région S-T en nombre d'images vidéo d'un système à 30 images par seconde, à diviser par 30 puis à multiplier par la fréquence d'images du système vidéo testé. Les régions S-T qui contiennent une seule image vidéo sont supposées toujours contenir une seule image vidéo, indépendamment de la fréquence d'images.

---

[11] On suppose implicitement que le rapport entre la distance de visualisation et la hauteur d'image reste fixe (on utilise des distances de visualisation plus courtes lorsque les images sont plus petites). On trouvera au § D.9 davantage d'observations sur la distance de visualisation supposée.

[12] Dans la présente annexe, toutes les dimensions temporelles seront données en nombre équivalent d'images vidéo d'un système vidéo à 30 images par seconde. Ainsi, une dimension temporelle de 6 images (F) représente à la fois 6 images d'un système NTSC (6/30) et 5 images d'un système PAL (5/25). Par ailleurs, on utilise 30 images par seconde et 29,97 images par seconde de manière interchangeable dans la présente annexe, étant donné que cette légère différence de la fréquence d'images n'a pas d'incidence sur le calcul de la qualité VQM.

La région d'intérêt spatial (SROI, voir § D.3) englobant toutes les régions S-T est identique pour la séquence vidéo d'origine et la séquence vidéo traitée étalonnées. La SROI doit être entièrement comprise dans la PVR, éventuellement avec un tampon de pixels, comme cela est requis par les filtres de perception convolutifs. La dimension horizontale de la SROI doit être divisible par la dimension horizontale de la région S-T. De même, la dimension verticale de la région SROI doit être divisible par la dimension verticale de la région S-T. Un utilisateur peut ensuite contraindre la SROI à englober une région d'intérêt particulière, par exemple le centre de l'image vidéo.

Temporellement, la séquence vidéo d'origine et la séquence vidéo traitée étalonnées sont subdivisées en un nombre identique de régions S-T, commençant à la première image vidéo alignée temporellement. Si le nombre d'images valables disponibles n'est pas divisible par la dimension temporelle de la région S-T, on ne tient pas compte des images se trouvant à la fin du clip.

Pour certaines caractéristiques, par exemple celles qui sont présentées au § D.7.2, le bloc 8 × 8_6F permet d'obtenir une très bonne corrélation avec les évaluations subjectives. Il est toutefois à noter que la corrélation décroît *lentement* à mesure qu'on s'éloigne de la taille de région S-T optimale. Des dimensions horizontales et verticales allant jusqu'à 32 voire davantage et des dimensions temporelles allant jusqu'à 30 images donnent des résultats satisfaisants, ce qui laisse une grande liberté au concepteur du système de mesures objectives pour adapter les caractéristiques au volume de stockage ou à la largeur de bande de transmission disponible [D-12].

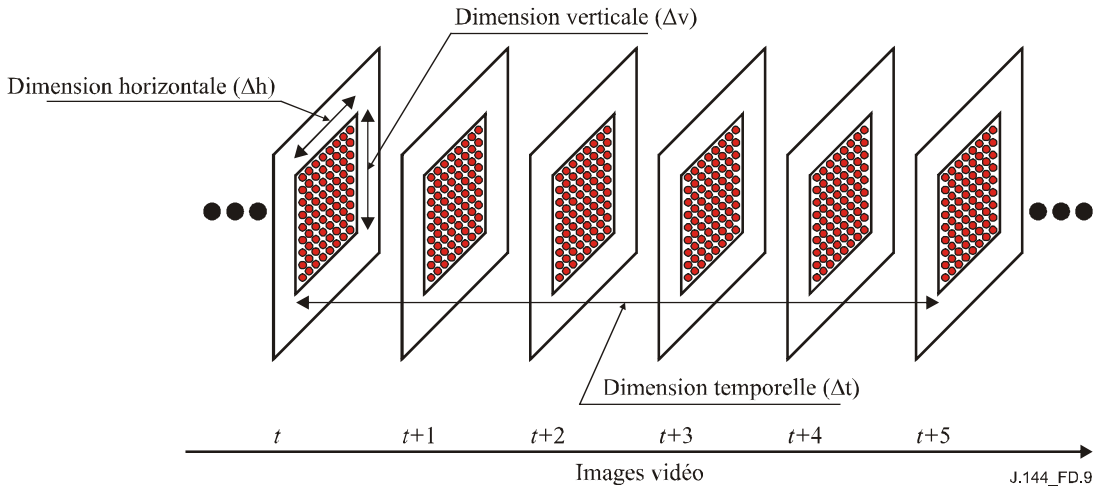


Figure D.9 – Exemple de taille de région spatio-temporelle (S-T) pour l'extraction des caractéristiques

Une fois que le flux vidéo a été subdivisé en régions S-T, l'axe temporel de la caractéristique (*t*) ne correspond plus à des images individuelles. En revanche, il contient un nombre d'échantillons égal au nombre d'images valables de la séquence vidéo étalonnée divisé par la dimension temporelle de la région S-T.

Lorsqu'on calcule simultanément deux caractéristiques ou plus, d'autres considérations deviennent importantes. Idéalement, toutes les caractéristiques devraient être calculées pour la même SROI.

D.7.2 Caractéristiques fondées sur les gradients spatiaux

Les caractéristiques déduites des gradients spatiaux peuvent servir à caractériser les distorsions perçues au niveau des contours. Par exemple, une perte générale d'informations relatives aux contours résulte d'un flou tandis qu'un excès d'informations relatives aux contours horizontaux et verticaux peut être lié à une distorsion due à une subdivision en blocs ou à un pavage. Les composantes Y du flux vidéo d'origine et du flux vidéo traité sont filtrées au moyen d'un filtre d'accentuation des contours horizontaux et d'un filtre d'accentuation des contours verticaux. Ces

flux vidéo filtrés sont ensuite subdivisés en régions spatio-temporelles (S-T) à partir desquelles on extrait les caractéristiques, ou les statistiques récapitulatives, permettant de quantifier l'activité spatiale en fonction de l'angle d'orientation. Ces caractéristiques sont ensuite coupées à l'extrémité inférieure afin d'émuler les seuils de perceptibilité. Les filtres d'accentuation des contours, la taille de la région S-T et les seuils de perceptibilité ont été choisis sur la base de flux vidéo conformes à la Rec. UIT-R BT.601-5 qui ont été évalués subjectivement à une distance de visualisation de six hauteurs d'image. La Figure D.10 présente un aperçu de l'algorithme utilisé pour extraire les caractéristiques fondées sur les gradients spatiaux.

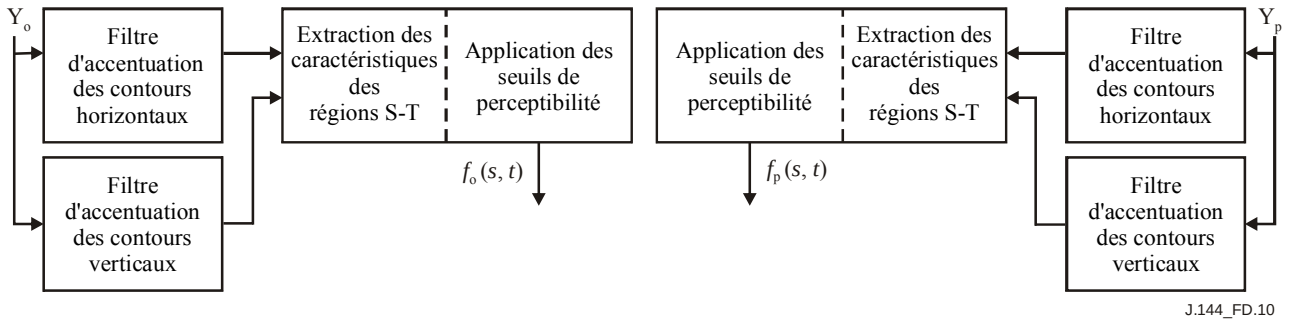


Figure D.10 – Aperçu de l'algorithme utilisé pour extraire les caractéristiques fondées sur les gradients spatiaux

D.7.2.1 Filtres d'accentuation des contours

Les *images* Y (de luminance) du flux vidéo d'origine et du flux vidéo traité sont d'abord traitées par des filtres d'accentuation des contours horizontaux et verticaux qui accentuent les contours tout en réduisant le bruit. Les deux filtres présentés sur la Figure D.11 sont appliqués séparément. Le premier (filtre de gauche) accentue les différences entre pixels horizontaux tout en procédant à un lissage vertical et le second (filtre de droite) accentue les différences entre pixels verticaux tout en procédant à un lissage horizontal.

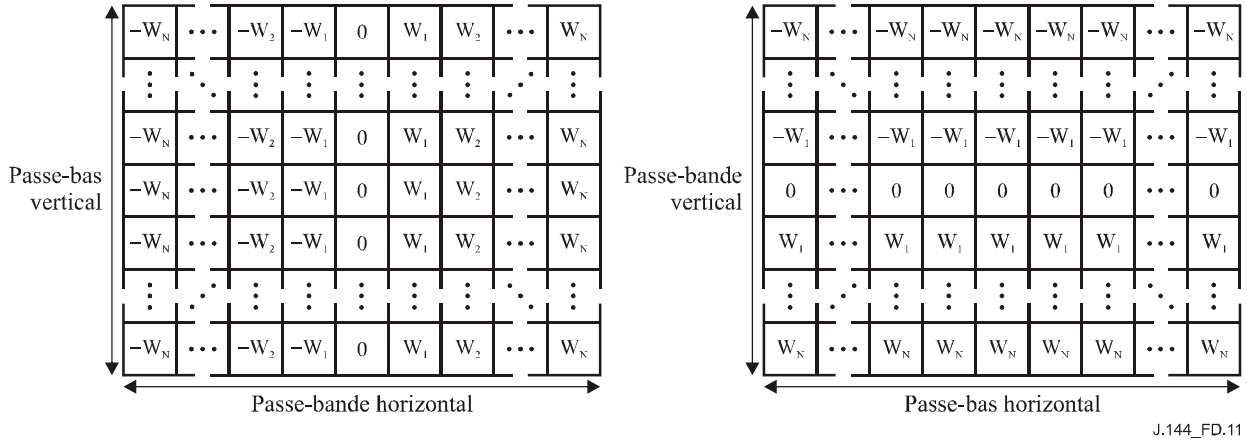


Figure D.11 – Filtres d'accentuation des contours

Les deux filtres sont transposés l'un de l'autre, ont une taille de 13 × 13, et ont pour coefficients de pondération:

$$w_x = k \times \left(\frac{x}{c}\right) \times \exp\left\{-\left(\frac{1}{2}\right)\left(\frac{x}{c}\right)^2\right\}$$

x représente le déplacement en pixels par rapport au centre du filtre (0, 1, 2, …, N), *c* est une constante fixant la largeur du filtre passe-bande et *k* est une constante de normalisation choisie de manière que chaque filtre présente le même gain qu'un véritable filtre de Sobel [6]. L'expérience a montré que le filtrage optimal en passe-bande horizontal pour une distance de visualisation égale à six hauteurs d'image était réalisé pour un filtre avec $c = 2$ présentant une réponse crête d'environ 4,5 cycles/degré. Les coefficients de pondération utilisés pour le filtre passe-bande sont les suivants:

[–.0052625, –.0173446, –.0427401, –.0768961, –.0957739, –.0696751, 0, .0696751, .0957739, .0768961, .0427401, .0173446, .0052625].

Il est à noter que les filtres de la Figure D.11 présentent une réponse passe-bas uniforme. Cette réponse a généré la meilleure évaluation de qualité et présente de plus l'avantage d'être efficace sur le plan des calculs (dans le cas du filtre de gauche de la Figure D.11 par exemple, il suffit de sommer les pixels d'une colonne et de multiplier le résultat par le coefficient de pondération).

D.7.2.2 Description des caractéristiques f_{SI13} et f_{HV13}

Le présent paragraphe décrit l'extraction de deux caractéristiques d'activité spatiale des régions S-T de flux vidéo d'origine et traité aux contours accentués comme décrits au § D.7.2.1. Ces caractéristiques servent pour la détection de dégradations spatiales telles que le flou et la subdivision en blocs. Le filtre présenté sur la Figure D.11 (à gauche) accentue les gradients spatiaux suivant la direction horizontale (H) alors que le transposé de ce filtre (à droite) accentue les gradients spatiaux suivant la direction verticale (V). On peut tracer pour chaque pixel la réponse de ces filtres H et V sur un diagramme à deux dimensions comme celui de la Figure D.12: la réponse du filtre H correspond à l'abscisse et la réponse du filtre V correspond à l'ordonnée. Pour un pixel donné de l'image repéré par sa ligne *i*, sa colonne *j* et le temps *t*, les réponses des filtres H et V seront respectivement notées $H(i, j, t)$ et $V(i, j, t)$. On peut convertir ces réponses en coordonnées polaires (R, θ) en utilisant les relations suivantes:

$$R(i, j, t) = \sqrt{H(i, j, t)^2 + V(i, j, t)^2}\ , \text{ et}$$

$$\theta(i, j, t) = \tan^{-1}\left[\frac{V(i, j, t)}{H(i, j, t)}\right]$$

La première caractéristique, qui est une mesure de l'information spatiale (SI, *spatial information*) globale, est désignée par f_{SI13} car les images ont été préalablement traitées par les filtres 13 × 13 présentés sur la Figure D.11. Cette caractéristique se calcule simplement comme l'écart type (std, *standard deviation*) sur la région S-T des échantillons *R(i, j, t)*. Elle est ensuite coupée au seuil de perceptibilité *P* (ce qui signifie que f_{SI13} est fixé à *P* si le calcul de *std* donne un résultat inférieur à *P*). On obtient ainsi:

$$f_{SI13} = \left\{ std\left[R(i, j, t)\right]\right\}\Big|_P :\ i, j, t \in \left\{\text{S} - \text{T Région}\right\}$$

Cette caractéristique est sensible aux modifications affectant la quantité globale d'activité spatiale au sein d'une région S-T donnée. Par exemple, un flou localisé entraîne une diminution de la quantité d'activité spatiale alors qu'un bruit accroît cette dernière. Le seuil *P* recommandé pour cette caractéristique est de 12.

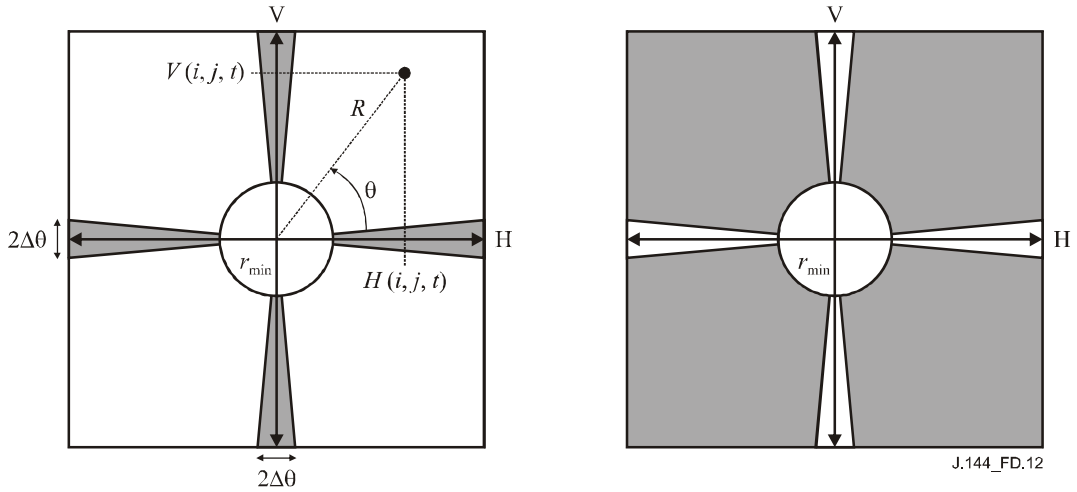


Figure D.12 – Subdivision de l'activité spatiale horizontale (H) et verticale (V) en distribution $\overline{HV}$ (à droite) et *HV* (gauche)

La seconde caractéristique, f_{HV13}, est sensible aux modifications de distribution angulaire (ou d'orientation) de l'activité spatiale. On calcule les images complémentaires avec les distributions de gradients spatiaux représentées en ombragé sur la Figure D.12. L'image avec les gradients horizontaux et verticaux, notée *HV*, contient les pixels *R(i, j, t)* correspondant à des contours horizontaux ou verticaux (les pixels correspondant à des contours diagonaux sont mis à zéro). L'image avec les gradients diagonaux, notée $\overline{HV}$, contient les pixels *R(i, j, t)* correspondant à des contours diagonaux (les pixels correspondant à des contours horizontaux ou verticaux sont mis à zéro). Les amplitudes de gradient *R(i, j, t)* inférieures à r_{min} sont mises à zéro dans les deux images pour garantir des calculs exacts de θ. On peut représenter mathématiquement les pixels de *HV* et $\overline{HV}$ de la façon suivante:

$$HV(i,j,t)=\begin{cases} R(i,j,t) \quad si\ R(i,j,t)\geq r_{\min}\ et\ m\dfrac{\pi}{2}-\Delta\theta<\theta(i,j,t)<m\dfrac{\pi}{2}+\Delta\theta \quad (m=0,1,2,3) \\ 0 \quad sinon \end{cases},$$

et

$$\overline{HV}(i,j,t)=\begin{cases} R(i,j,t) \quad si\ R(i,j,t)\geq r_{\min}\ et\ m\dfrac{\pi}{2}+\Delta\theta\leq\theta(i,j,t)\leq(m+1)\dfrac{\pi}{2}-\Delta\theta \quad (m=0,1,2,3) \\ 0 \quad sinon \end{cases}$$

avec

$$i,j,t\in\{S-T\ Région\}$$

Pour les calculs de *HV* et $\overline{HV}$ ci-dessus, il est recommandé d'utiliser une valeur de 20 pour r_{min} et une valeur de 0,225 radians pour Δθ. La caractéristique f_{HV13} pour une région S-T donnée est ensuite obtenue comme étant le rapport entre la moyenne de *HV* et la moyenne de $\overline{HV}$, ces moyennes étant coupées à leur seuil de perceptibilité *P*. On obtient ainsi:

$$f_{HV13} = \left\{mean\left[HV(i,j,t)\right]\right\}\Big|_P \Big/ \left\{mean\left[\overline{HV}(i,j,t)\right]\right\}\Big|_P .$$

Le seuil de perceptibilité P recommandé pour les moyennes de HV et $\overline{HV}$ est de 3. La caractéristique f_{HV13} est sensible aux modifications de distribution angulaire de l'activité spatiale au sein d'une région S-T donnée. Par exemple, si les contours horizontaux et verticaux sont plus flous que les contours diagonaux, la valeur de f_{HV13} du flux vidéo traité sera moins élevée que celle du flux vidéo d'origine. D'autre part, si des contours horizontaux ou verticaux erronés sont introduits (par exemple sous forme de distorsions liée à une subdivision en blocs ou à un pavage), la valeur de f_{HV13} du flux vidéo traité sera alors plus élevée que celle du flux vidéo d'origine. La caractéristique f_{HV13} fournit aussi un moyen simple pour tenir compte des variations de sensibilité du système visuel humain en fonction de l'angle d'orientation[13].

D.7.3 Caractéristiques fondées sur les informations de chrominance

Dans le présent paragraphe, on décrit une seule caractéristique qui peut être utilisée pour mesurer les distorsions des signaux de chrominance (C_B, C_R). Pour un pixel donné de l'image repéré par sa ligne i, sa colonne j et le temps t, désignons par $C_B(i, j, t)$ et $C_R(i, j, t)$ les valeurs des composantes C_B et C_R définies dans la Rec. UIT-R BT.601-5[14]. Les composantes d'un vecteur de caractéristique de chrominance à deux dimensions, f_{COHER_COLOR}, sont calculées comme étant la moyenne (*mean*) sur la région S-T des échantillons $C_B(i, j, t)$ et $C_R(i, j, t)$, respectivement, un poids de perception plus élevé étant affecté à la composante C_R:

$$\underline{f}_{COHER\_COLOR} = \left(mean\left[C_B(i,j,t)\right], W_R \times mean\left[C_R(i,j,t)\right]\right):\ i,j,t \in \{Région\ S-T\},$$
$$\text{et } W_R = 1{,}5$$

La formule ci-dessus permet de procéder à une intégration cohérente (d'où le nom f_{COHER_COLOR}) car la relation de phase entre C_B et C_R est préservée. Pour ceux qui connaissent bien les vecteurscopes, l'utilité du vecteur de caractéristique de chrominance apparaît directement à l'examen des signaux de mire chromatique. Pour les scènes générales, on peut visualiser l'utilité du vecteur de caractéristique de chrominance concernant la mesure des distorsions de la chrominance pour des blocs vidéo qui couvrent une certaine plage spatio-temporelle. Toutefois, si la taille de la région S-T est trop grande, de nombreuses couleurs risquent d'être incluses dans le calcul et l'utilité de f_{COHER_COLOR} est alors réduite. Une taille de région S-T de 8 pixels horizontaux × 8 lignes verticales × (1 à 3) images vidéo permet de générer un vecteur de caractéristique de chrominance robuste (en fait 4 pixels C_B et C_R horizontaux, étant donné que ces signaux sont sous-échantillonnés par un facteur deux horizontalement pour ce qui est de l'échantillonnage selon la Rec. UIT-R BT.601-5).

D.7.4 Caractéristiques fondées sur les informations de contraste

Les caractéristiques qui mesurent les informations de contraste localisé sont sensibles aux dégradations de la qualité telles que le flou (perte de contraste) et l'ajout de bruit (gain de contraste). On peut calculer facilement une caractéristique de contraste localisé, f_{CONT}, pour chaque région S-T à partir de l'image de luminance Y comme suit:

---

[13] Cet exposé de la caractéristique f_{HV13}, quoique globalement valable, est quelque peu simplifié. Par exemple, lorsqu'il rencontre certaines formes, le filtre f_{HV13} se comporte d'une manière qui peut être contraire à l'intuition (par exemple un coin formé par une ligne horizontale et une ligne verticale conduira à une énergie diagonale).

[14] Les corrections de gain et de décalage ne sont pas appliquées aux plans d'image C_B et C_R. Voir le § D.6.3.3.

$$f_{CONT} = \{ std[Y(i, j, t)] \}|_P : \; i, j, t \in \{ S-T \; Région \}$$

Le seuil de perceptibilité P recommandé pour la caractéristique f_{CONT} est compris entre quatre et six.

D.7.5 Caractéristiques fondées sur l'information temporelle absolue (ATI, *absolute temporal information*)

Les caractéristiques qui mesurent les distorsions du flux de mouvement sont sensibles aux dégradations de la qualité telles que l'élimination ou la répétition d'images (perte de mouvement) et l'ajout de bruit (gain de mouvement). On calcule une caractéristique d'information temporelle absolue, f_{ATI}, pour chaque région S-T. Pour cela, on commence par générer un flux de mouvement fondé sur la valeur absolue de la différence entre deux images vidéo consécutives aux instants t et $t-1$ puis on calcule l'écart type sur la région S-T. Mathématiquement, ce processus est représenté de la façon suivante:

$$f_{ATI} = \{ std \; |Y(i, j, t) - Y(i, j, t-1)| \}|_P : \; i, j, t \in \{ Région \; S-T \}$$

Le seuil de perceptibilité P recommandé pour la caractéristique f_{ATI} est compris entre un et trois.

L'utilisation d'une image précédente introduit des considérations qui vont au-delà de celles qui sont associées aux autres caractéristiques. Lors du calcul de f_{ATI} conjointement avec une autre caractéristique (par exemple $f_{CONTRAST\_ATI}$ définie au § D.7.6) ou lors de son utilisation dans un modèle (voir le § D.9), l'image supplémentaire requise a pour effet de compliquer la tâche de positionnement des régions S-T (voir le § D.7.1.1).

D.7.6 Caractéristiques fondées sur le produit croisé du contraste et de l'information temporelle absolue

La quantité de mouvement présente peut avoir une incidence sur la perceptibilité des dégradations spatiales. De même, la quantité de détails spatiaux présents peut avoir une incidence sur la perceptibilité des dégradations temporelles. Une caractéristique déduite du produit croisé de l'information de contraste et de l'information temporelle absolue permet de tenir compte partiellement de ces interactions. Cette caractéristique, désignée par $f_{CONTRAST\_ATI}$, est calculée comme étant le produit des caractéristiques définies aux § D.7.4 et 7.5[15]. Le seuil de perceptibilité recommandé $P = 3$ est appliqué séparément à chaque caractéristique (f_{CONT} et f_{ATI}) avant que leur produit croisé ne soit calculé. Les dégradations seront davantage visibles dans les régions S-T présentant un produit croisé faible que dans les régions S-T présentant un produit croisé élevé. Cela est particulièrement vrai pour les dégradations de type bruit et blocs d'erreurs.

L'image supplémentaire requise pour f_{ATI} a pour effet de compliquer légèrement $f_{CONTRAST\_ATI}$, car les régions S-T utilisées par les deux caractéristiques f_{CONT} et f_{ATI} doivent être positionnées de la même façon. Soit on n'utilise pas la première image de la séquence vidéo pour f_{ATI}, soit les régions S-T situées au début de la séquence vidéo contiennent une image de moins (par exemple, si on considère une dimension temporelle de 6 images, la première région S-T pour f_{ATI} utiliserait 5 images au lieu de 6). Pour les paramètres et modèles spécifiés ici, on suppose que c'est la seconde solution qui est utilisée.

---

[15] On utilise un produit croisé standard des caractéristiques f_{CONT} et f_{ATI} (à savoir $f_{CONT} \times f_{ATI}$) pour les caractéristiques du flux traité $f_p(s, t)$ et du flux d'origine $f_o(s, t)$ dans les fonctions de comparaison *ratio_loss* et *ratio_gain* décrites au § D.8.2.1. Toutefois, pour les fonctions de comparaison *log_loss* et *log_gain*, les caractéristiques du flux traité et du flux d'origine sont calculées de la manière suivante: $\log_{10}[f_{CONT}] \times \log_{10}[f_{ATI}]$, et les fonctions de comparaison utilisent une différence (à savoir $f_p(s, t) - f_o(s, t)$ plutôt que $\log_{10}[f_p(s, t) / f_o(s, t)]$).

D.8 Paramètres de qualité

D.8.1 Introduction

Les paramètres de qualité, qui servent à mesurer les distorsions de qualité vidéo dues aux gains et aux pertes associés aux valeurs de caractéristiques, sont d'abord calculés pour chaque région S-T. Pour cela, on compare les valeurs des caractéristiques du flux d'origine, $f_o(s, t)$, avec les valeurs des caractéristiques du flux traité correspondant, $f_p(s, t)$ (§ D.8.2). On utilise plusieurs relations fonctionnelles pour émuler le masquage visuel des dégradations pour chaque région S-T. Des fonctions de regroupement des erreurs dans l'espace et dans le temps émulent ensuite la façon dont l'être humain déduit les évaluations de qualité subjective. Le regroupement d'erreurs dans l'espace est appelé regroupement spatial (§ D.8.3) et le regroupement d'erreurs dans le temps est appelé regroupement temporel (§ D.8.4). L'application séquentielle des fonctions de regroupement spatial et de regroupement temporel au flux de paramètres de qualité S-T génère des paramètres de qualité pour le clip vidéo tout entier, dont la durée nominale est comprise entre 5 et 10 secondes. La valeur finale de chaque paramètre après regroupement temporel peut être corrigée puis coupée (§ D.8.5) et ce, afin de tenir compte des relations non linéaires entre la valeur du paramètre et la qualité perçue et de réduire encore la sensibilité du paramètre.

En résumé, les étapes du calcul des paramètres sont les suivantes. Dans certains cas, l'étape indiquée comme une [option] ne sera pas nécessaire.

1) Comparer les valeurs des caractéristiques du flux d'origine avec les valeurs des caractéristiques du flux traité.
2) Procéder au regroupement spatial.
3) Procéder au regroupement temporel.
4) [Option] appliquer une correction non linéaire et/ou procéder à une coupure.

Tous les paramètres sont conçus pour prendre uniquement des valeurs positives ou uniquement des valeurs négatives. Une valeur de paramètre de zéro indique qu'il n'y a pas de dégradation.

D.8.2 Fonctions de comparaison

La dégradation perçue au niveau de chaque région S-T est calculée au moyen de fonctions qui modélisent le masquage visuel des dégradations spatiales et temporelles. Le présent paragraphe expose les fonctions de masquage qui sont utilisées par les divers paramètres pour produire des paramètres de qualité qui sont fonction de l'espace et du temps.

D.8.2.1 Fonction de rapport et fonction de logarithme

La perte et le gain sont généralement examinés séparément, car ils produisent des effets fondamentalement différents sur la perception de la qualité (par exemple perte d'activité spatiale due au flou et gain d'activité spatiale dû au bruit ou à la subdivision en blocs). Parmi les nombreuses fonctions de comparaison qui ont été évaluées, deux formes ont produit de manière cohérente une très bonne corrélation avec les évaluations subjectives. Chacune de ces formes peut être utilisée avec les calculs de gain ou de perte pour un total de quatre fonctions de comparaison S-T de base. Les quatre formes primaires sont les suivantes:

$$ratio\_loss(s,t) = np\left\{\frac{f_p(s,t) - f_o(s,t)}{f_o(s,t)}\right\}$$

$$ratio\_gain(s,t) = pp\left\{\frac{f_p(s,t) - f_o(s,t)}{f_o(s,t)}\right\}$$

$$\log\_loss(s,t) = np\left\{\log_{10}\left[\frac{f_p(s,t)}{f_o(s,t)}\right]\right\} \text{ et}$$

$$\log\_gain(s,t) = pp\left\{\log_{10}\left[\frac{f_p(s,t)}{f_o(s,t)}\right]\right\}$$

où *pp* est l'opérateur de partie positive (autrement dit, les valeurs négatives sont remplacées par des zéros) et *np* est l'opérateur de partie négative (autrement dit, les valeurs positives sont remplacées par des zéros).

Ces fonctions de masquage visuel impliquent que la perception des dégradations est inversement proportionnelle à la quantité d'activité spatiale ou temporelle localisée qui est présente. En d'autres termes, les dégradations spatiales deviennent moins visibles à mesure que l'activité spatiale augmente (masquage spatial) et les dégradations temporelles deviennent moins visibles à mesure que l'activité temporelle augmente (masquage temporel). Les fonctions de comparaison de type rapport et logarithme ont un comportement très similaire, mais la fonction de logarithme a tendance à être légèrement plus avantageuse pour les gains tandis que la fonction de rapport a tendance à être légèrement plus avantageuse pour les pertes. La fonction de logarithme a une plage dynamique plus grande, ce qui est utile lorsque les valeurs des caractéristiques du flux traité dépassent de beaucoup les valeurs des caractéristiques du flux d'origine.

D.8.2.2 Distance euclidienne

Une autre fonction de comparaison S-T utile est la simple distance euclidienne, représentée par la longueur du vecteur de différence entre le vecteur de caractéristique du flux d'origine $f_o(s, t)$ et le vecteur de caractéristique du flux traité correspondant, $f_p(s, t)$,

$$euclid(s,t) = \left\| \underline{f}_p(s,t) - \underline{f}_o(s,t) \right\|$$

La Figure D.13 donne une illustration de la distance euclidienne pour un vecteur de caractéristique à deux dimensions extrait d'une région S-T (par exemple le vecteur de caractéristique f_{COHER_COLOR} décrit au § D.7.3), où les indices *s* et *t* désignent respectivement la position spatiale et la position temporelle de la région S-T dans le flux vidéo d'origine et le flux vidéo traité étalonnés. Le segment en pointillés sur la Figure D.13 illustre la distance euclidienne. La mesure de la distance euclidienne peut être généralisée pour des vecteurs de caractéristique qui ont un nombre arbitraire de dimensions.

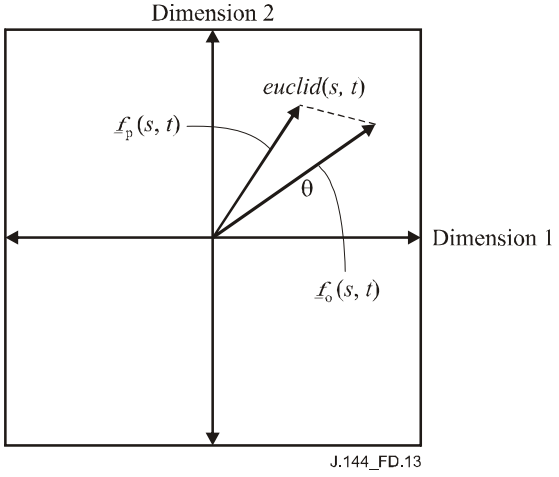


Figure D.13 – Illustration de la distance euclidienne *euclid*(*s, t*) pour un vecteur de caractéristique à deux dimensions

D.8.3 Fonctions de regroupement spatial

Les paramètres issus des régions S-T (§ D.8.2) constituent des matrices à trois dimensions: une dimension temporelle et deux dimensions spatiales (position horizontale et position verticale de la région S-T). On regroupe ensuite les dégradations issues des régions S-T ayant le même indice temporel *t* au moyen d'une fonction de regroupement spatial. Le regroupement spatial génère un historique temporel des valeurs de paramètre. Cet historique, désigné génériquement par *p*(*t*), doit ensuite être regroupé temporellement au moyen d'une fonction de regroupement temporel donnée au § D.8.4. Le Tableau D.1 présente une récapitulation des fonctions de regroupement spatial les plus couramment utilisées.

Des examens approfondis ont montré que les fonctions de regroupement spatial optimales incluent généralement un certain traitement associé au cas le plus défavorable, comme le calcul de la moyenne des 5% de distorsions les plus mauvaises observées sur l'indice spatial *s* ([D-10]-[D-13]). Cela s'explique par le fait que les dégradations localisées ont tendance à attirer l'attention de l'observateur, pour lequel la plus mauvaise partie de l'image constitue le facteur prédominant dans l'évaluation de la qualité subjective. Par exemple, la fonction de regroupement spatial "*above95%*" est calculée pour chaque indice temporel *t* pour la fonction *log_gain*(*s*,*t*) décrite au § D.8.2.1 comme étant la moyenne des 5% de valeurs positives les plus élevées sur l'indice spatial *s*[16]. Cela revient à trier les distorsions du gain de la valeur la plus faible à la valeur la plus élevée pour chaque indice temporel *t* et à calculer la moyenne des distorsions qui sont au-dessus du seuil de 95% (car plus les valeurs positives sont élevées, plus la distorsion est grande). De même, les distorsions dues à des pertes comme celles qui sont calculées par la fonction *ratio_loss*(*s*,*t*) décrite au § D.8.2.1 sont triées pour chaque indice temporel *t*, mais on calcule la moyenne des distorsions qui sont au-dessous du seuil de 5% ("*below5%*") (car les pertes sont négatives).

D.8.4 Fonctions de regroupement temporel

L'historique temporel du paramètre *p*(*t*) résultant de la fonction de regroupement spatial (voir le § D.8.3) fait ensuite l'objet d'un regroupement au moyen d'une fonction de regroupement temporel et ce, afin de produire un paramètre objectif *p* pour le clip vidéo, dont la durée nominale est comprise entre 4 et 10 secondes. Les observateurs semblent utiliser plusieurs fonctions de regroupement temporel lorsqu'ils évaluent subjectivement des clips vidéo d'une durée approximative de 10 secondes. La moyenne (*mean*) dans le temps donne une indication de la qualité moyenne qui est observée pendant la période considérée. Les niveaux à *90%* et à *10%* dans le temps donnent une indication de la qualité transitoire la plus mauvaise qui est observée respectivement pour les gains et pour les pertes (des erreurs de transmission numérique peuvent par exemple causer une perturbation de 1 à 2 secondes dans le flux vidéo traité). Après regroupement temporel, un paramètre *p* donné prend uniquement des valeurs négatives ou uniquement des valeurs positives. Le Tableau D.2 présente une récapitulation des fonctions de regroupement temporel les plus couramment utilisées.

---

[16] Il est à noter que l'indice temporel *t* ne correspond pas ici à des images individuelles (voir le § D.7.1.1). En effet, chaque valeur de *t* correspond aux régions S-T ayant la même dimension temporelle.

Tableau D.1 – Fonctions de regroupement spatial et leur définition

Fonction de regroupement spatial	Définition
below5%	Pour chaque indice temporel *t*, on trie les valeurs de paramètre de la valeur la plus faible à la valeur la plus élevée. On calcule la moyenne de toutes les valeurs de paramètre qui sont inférieures ou égales au seuil de 5%. Pour les paramètres de perte, cette fonction de regroupement spatial produit un paramètre qui donne une indication de la qualité la plus mauvaise dans l'espace.
above95%	Pour chaque indice temporel *t*, on trie les valeurs de paramètre de la valeur la plus faible à la valeur la plus élevée. On calcule la moyenne de toutes les valeurs de paramètre qui sont supérieures ou égales au seuil de 95%. Pour les paramètres de gain, cette fonction de regroupement spatial produit un paramètre qui donne une indication de la qualité la plus mauvaise dans l'espace.
mean	Pour chaque indice temporel *t*, on calcule la moyenne de toutes les valeurs de paramètre. Cette fonction de regroupement spatial produit un paramètre qui donne une indication de la qualité moyenne dans l'espace.
std	Pour chaque indice temporel *t*, on calcule l'écart type de toutes les valeurs de paramètre. Cette fonction de regroupement spatial produit un paramètre qui donne une indication des variations de la qualité dans l'espace.
below5%tail	Pour chaque indice temporel *t*, on trie les valeurs de paramètre de la valeur la plus faible à la valeur la plus élevée. On calcule la moyenne de toutes les valeurs de paramètre qui sont inférieures ou égales au seuil de 5% puis on soustrait le niveau à 5% de cette moyenne. Pour les paramètres de perte, cette fonction de regroupement spatial permet de mesurer l'étalement des plus mauvais niveaux de qualité dans l'espace. Elle est utile pour la mesure des effets des distorsions localisées spatialement sur la qualité perçue.
above99%tail	Pour chaque indice temporel *t*, on trie les valeurs de paramètre de la valeur la plus faible à la valeur la plus élevée. On calcule la moyenne de toutes les valeurs de paramètre qui sont supérieures ou égales au seuil de 99% puis on soustrait le niveau à 99% de cette moyenne. Pour les paramètres de gain, cette fonction de regroupement spatial permet de mesurer l'étalement des plus mauvais niveaux de qualité dans l'espace. Elle est utile pour la mesure des effets des distorsions localisées spatialement sur la qualité perçue.

Tableau D.2 – Fonctions de regroupement temporel et leur définition

Fonction de regroupement temporel	Définition
10%	On trie l'historique temporel des valeurs de paramètre de la valeur la plus faible à la valeur la plus élevée et on prend le seuil de 10%. Pour les paramètres de perte, cette fonction de regroupement temporel produit un paramètre qui donne une indication de la plus mauvaise qualité dans le temps. Pour les paramètres de gain, elle produit un paramètre qui donne une indication de la meilleure qualité dans le temps.
25%	On trie l'historique temporel des valeurs de paramètre de la valeur la plus faible à la valeur la plus élevée et on prend le seuil de 25%.
50%	On trie l'historique temporel des valeurs de paramètre de la valeur la plus faible à la valeur la plus élevée et on prend le seuil de 50%.
90%	On trie l'historique temporel des valeurs de paramètre de la valeur la plus faible à la valeur la plus élevée et on prend le seuil de 90%. Pour les paramètres de perte, cette fonction de regroupement temporel produit un paramètre qui donne une indication de la meilleure qualité dans le temps. Pour les paramètres de gain, elle produit un paramètre qui donne une indication de la plus mauvaise qualité dans le temps.
mean	On calcule la moyenne de l'historique temporel des valeurs de paramètre. Cette fonction produit un paramètre qui donne une indication de la qualité moyenne dans le temps.
std	On calcule l'écart type de l'historique temporel des valeurs de paramètre. Cette fonction de regroupement temporel produit un paramètre qui donne une indication des variations de la qualité dans le temps.
above90%tail	On trie l'historique temporel des valeurs de paramètre de la valeur la plus faible à la valeur la plus élevée et on calcule la moyenne de toutes les valeurs de paramètre qui sont supérieures ou égales au seuil de 90% puis on soustrait le niveau à 90% de cette moyenne. Pour les paramètres de gain, cette fonction de regroupement temporel permet de mesurer l'étalement des plus mauvais niveaux de qualité dans le temps. Elle est utile pour la mesure des effets des distorsions localisées temporellement sur la qualité perçue.

D.8.5 Application d'une correction non linéaire et coupure

On peut appliquer un facteur de correction au paramètre *p* prenant uniquement des valeurs positives ou uniquement des valeurs négatives issu du regroupement temporel (§ D.8.4) afin de tenir compte des relations non linéaires entre la valeur du paramètre et la qualité perçue. Il est préférable de supprimer les éventuelles relations non linéaires avant d'établir les modèles de qualité vidéo (§ D.9), car on utilise un algorithme linéaire fondé sur la méthode des moindres carrés pour déterminer les poids optimaux pour les paramètres. Les deux fonctions de correction non linéaire pouvant être appliquées sont la fonction racine carrée, désignée par *sqrt*, et la fonction carrée, désignée par *square*. Si la fonction *sqrt* est appliquée à un paramètre dont toutes les valeurs sont négatives, on commence par faire en sorte que ce paramètre ne prenne que des valeurs positives (on prend la valeur absolue).

Enfin, on peut appliquer une fonction de coupure désignée par *clip_T*, où *T* est le seuil de coupure, afin de réduire la sensibilité du paramètre aux faibles dégradations. La fonction de coupure remplace toute valeur du paramètre comprise entre le niveau de coupure et zéro par le niveau de

coupure puis le niveau de coupure est soustrait de la valeur résultante du paramètre. La représentation mathématique en est la suivante:

$$clip\_T(p) = \begin{cases} \max(p,T) - T & si\ p\ ne\ prend\ que\ des\ valeurs\ positives \\ \min(p,T) - T & si\ p\ \ ne\ prend\ que\ des\ valeurs\ négatives \end{cases}$$

D.8.6 Convention pour la dénomination des paramètres

Le présent paragraphe résume la convention de dénomination technique utilisée pour les paramètres de qualité vidéo. Selon cette convention, on attribue à chaque paramètre un nom très long constitué de mots d'identification (sous-noms) séparés par des soulignés. Le nom de paramètre technique résume le processus exact utilisé pour calculer le paramètre. Chaque sous-nom identifie une fonction ou une étape du processus de calcul du paramètre. Les sous-noms sont énumérés dans l'ordre dans lequel les fonctions ou les étapes se déroulent, de gauche à droite. Le Tableau D.3 récapitule les sous-noms utilisés pour créer un nom de paramètre technique, énumérés dans l'ordre susmentionné. Le § D.8.6.1 donne quelques exemples de nom de paramètre technique et des sous-noms associés issus du Tableau D.3.

Tableau D.3 – Convention utilisée pour la dénomination technique des paramètres de qualité vidéo

Sous-nom	Définition	Exemples
Couleur	Plans d'image de l'espace chromatique utilisés par le paramètre.	*Y* pour le plan d'image de luminance *color* pour les plans d'image (C_B, C_R)
Propre à la caractéristique	Ce sous-nom décrit les calculs qui rendent unique le paramètre considéré. Tous les autres sous-noms qui suivent correspondent à des processus génériques qui peuvent être utilisés par de nombreux types différents de paramètres. Le sous-nom "propre à la caractéristique" est généralement le nom de la caractéristique qui est extraite du plan "couleur" à ce stade dans le flux, autrement dit à l'emplacement de ce sous-nom. Toutefois, des informations non prises en considération par la convention de dénomination peuvent aussi être incluses ici. Par exemple, le paramètre HV applique le sous-nom "statistique de bloc" séparément aux plans d'image *HV* et $\overline{HV}$. Le rapport entre HV et $\overline{HV}$ qui en découle est spécifié par le sous-nom "propre à la caractéristique" (plutôt que d'occuper un sous-nom distinct après le sous-nom "statistique de bloc").	*si13* pour la caractéristique f_{SI13} du § D.7.2.2 *hv13_angleX.XXX_*r_{min}*YY* pour la caractéristique f_{HV13} du § D.7.2.2, où *X.XXX* est la valeur de $\Delta\theta$ et *YY* celle de r_{min} *coher_color* pour la caractéristique f_{COHER_COLOR} du § D.7.3 *cont* pour la caractéristique f_{CONT} du § D.7.4 *ati* pour la caractéristique f_{ATI} du § D.7.5 *contrast_ati* pour la caractéristique $f_{CONTRAST\_ATI}$ du § D.7.6
Décalage de bloc	Présent en cas de superposition de blocs S-T (par exemple des blocs qui se chevauchent dans le temps). Lorsque ce sous-nom est absent, les blocs sont supposés être contigus dans le temps.	*sliding*
Image complète	Présent lorsque la taille de bloc S-T contient toute la région valable de l'image. Lorsque ce sous-nom est absent, le sous-nom "taille de bloc" doit être présent.	*image*

Tableau D.3 – Convention utilisée pour la dénomination technique des paramètres de qualité vidéo

Sous-nom	Définition	Exemples
Taille de bloc	Présent lorsque l'image est subdivisée en blocs S-T (voir le § D.7.1.1). Dans un souci de cohérence, la taille de bloc est toujours indiquée en nombre de lignes d'image et en nombre de pixels d'image du plan de luminance (Y). Ainsi, pour les séquences vidéo échantillonnées selon le format 4:2:2, les blocs de couleur contiendront en réalité la moitié du nombre de pixels spécifié horizontalement. Lorsque ce sous-nom est absent, le sous-nom "image complète" doit être présent.	*8 × 8* pour les blocs comprenant 8 lignes d'image verticalement par 8 pixels d'image horizontalement *128 × 128* pour les blocs comprenant 128 lignes d'image verticalement par 128 pixels d'image horizontalement
Images de bloc	Ce sous-nom indique la dimension temporelle des blocs S-T (voir le § D.7.1.1), pour une fréquence vidéo de 30 images par seconde (fps). Par exemple, *6F* représente un cinquième de seconde, quelle que soit la fréquence d'images utilisée (ce qui correspond à 5 images pour un système à 25 fps, 3 images pour un système à 15 fps, 2 images pour un système à 10 fps).	*1F* pour une dimension temporelle d'une image *6F* pour une dimension temporelle d'un cinquième de seconde
Statistique de bloc	Ce sous-nom indique la fonction statistique qui est utilisée pour extraire la caractéristique de chaque région S-T et qui produit un nombre pour chaque bloc de pixels S-T. Ce sous-nom est présent sauf si la "taille de bloc" est égale à *1 × 1* (c'est-à-dire un pixel). Avant application de la fonction "statistique de bloc", les résultats intermédiaires contiennent des historiques temporels des images avec un nombre par pixel (images filtrées); après, les résultats intermédiaires contiennent un nombre pour chaque région S-T (images de caractéristiques). Les paramètres associés à deux plans d'image (par exemple *hv*13 et *coher_color*) appliqueront séparément la fonction "statistique de bloc" aux deux plans d'image, produisant deux images de caractéristiques.	*mean* est la moyenne des valeurs des pixels *std* est l'écart type des valeurs des pixels *rms* est la valeur quadratique moyenne des valeurs des pixels
Seuil de perceptibilité	Les valeurs produites par la fonction "statistique de bloc" peuvent être coupées à un seuil de perceptibilité *P*. Les valeurs comprises entre zéro et ce seuil sont remplacées par le seuil.	*3* pour une valeur minimale de caractéristique de 3.0 *12* pour une valeur minimale de caractéristique de 12.0
Fonction de comparaison	Il s'agit de la fonction utilisée pour comparer les caractéristiques extraites des flux d'origine et traité (voir le § D.8.2). Avant application de la fonction de comparaison, les résultats intermédiaires contiennent des historiques temporels des images de caractéristiques des flux d'origine et traité; après, les résultats intermédiaires contiennent un historique temporel des images du paramètre.	*log_gain* (voir le § D.8.2.1) *ratio_loss* (voir le § D.8.2.1) *euclid* (voir le § D.8.2.2)

Tableau D.3 – Convention utilisée pour la dénomination technique des paramètres de qualité vidéo

Sous-nom	Définition	Exemples
Fonction de regroupement spatial	Voir le § D.8.3. La fonction est appliquée à chaque image du paramètre (par exemple toutes les régions S-T ayant le même indice temporel) et produit un historique temporel des valeurs du paramètre. Avant regroupement spatial, les résultats intermédiaires sont constitués des images du paramètre contenant une valeur pour chaque bloc S-T; après, les résultats intermédiaires correspondent à un historique temporel de nombres (historique temporel du paramètre). Ce sous-nom doit être présent pour tous les paramètres à l'exception des paramètres de type "image complète".	Voir le Tableau D.1
Fonction de regroupement temporel	Voir le § D.8.4. La fonction est appliquée à l'historique temporel du paramètre et produit une seule valeur du paramètre pour toute la séquence vidéo. Après regroupement temporel, le paramètre prend uniquement des valeurs négatives ou uniquement des valeurs positives. Zéro correspond à aucune dégradation et, plus la valeur du paramètre est éloignée de zéro, plus la dégradation est forte. Ce sous-nom doit être présent pour tous les paramètres.	Voir le Tableau D.2
Fonction non linéaire	Voir le § D.8.5. L'examen des valeurs du paramètre peut indiquer qu'il convient d'appliquer une correction non linéaire au paramètre afin d'assurer une correspondance linéaire avec les données subjectives. C'est la fonction non linéaire qui procède à cette correction finale. Si la fonction *sqrt* est appliquée à un paramètre prenant uniquement des valeurs négatives, on commence par faire en sorte que le paramètre prenne uniquement des valeurs positives (on prend la valeur absolue).	*sqrt* pour la racine carrée de la valeur du paramètre issue du regroupement temporel *square* pour le carré de la valeur du paramètre issue du regroupement temporel
Fonction de coupure	Voir le § D.8.5. L'examen final des valeurs du paramètre peut indiquer qu'il est nécessaire de réduire encore la sensibilité du paramètre aux faibles dégradations (valeurs du paramètre proches de zéro). On remplace toute valeur comprise entre le niveau de coupure *T* et zéro par le niveau de coupure puis on soustrait le niveau de coupure de la valeur résultante du paramètre.	*clip_0.45* Si le paramètre ne prend que des valeurs positives, on remplace toute valeur inférieure à 0,45 par 0,45 puis on soustrait 0,45 de la valeur résultante du paramètre. Si le paramètre ne prend que des valeurs négatives, on remplace toute valeur supérieure à –0,45 par –0,45 puis on ajoute 0,45 à la valeur résultante du paramètre.

D.8.6.1 Exemples de nom de paramètre

Le présent paragraphe inclut cinq exemples de nom technique, pour lesquels la procédure de sous-dénomination donnée au Tableau D.3 est décrite pas à pas.

Y_si13_8×8_6F_std_6_ratio_loss_below5%_mean

Y signifie qu'on utilise le plan d'image de luminance. *si13* signifie que l'on filtre les images au moyen des gabarits spatiaux 13×13 du § D.7.2.1 en vue de l'extraction de la caractéristique f_{SI13} décrite au § D.7.2.2. $8 \times 8\_6F$ signifie que l'on subdivise le flux vidéo en régions S-T contenant huit lignes d'image verticalement par huit pixels horizontalement par un cinquième de seconde temporellement (c'est-à-dire 6 images NTSC, 5 images PAL). *std* signifie que l'on prend l'écart type de chaque bloc. *6* signifie que l'on applique un seuil de perceptibilité et que l'on remplace toute valeur de l'écart type inférieure à 6,0 par 6,0. *ratio_loss* signifie que l'on compare les caractéristiques des flux d'origine et traité provenant de chaque bloc au moyen de la fonction *ratio_loss*. *below5%* signifie que l'on regroupe spatialement les valeurs du paramètre pour chaque indice temporel au moyen de la fonction *below5%*. *mean* signifie que l'on regroupe temporellement l'historique temporel du paramètre au moyen de la fonction *mean*.

color_coher_color_8×8_1F_mean_euclid_std_10%_clip_0.8

color signifie que l'on utilise les plans d'image C_B et C_R *coher_color* signifie que l'on préserve la relation de phase entre les images C_B et C_R (en les traitant séparément) en vue de l'extraction de la caractéristique f_{COHER_COLOR} décrite au § D.7.3. $8 \times 8\_1F$ signifie que l'on subdivise chaque image en blocs qui font 8 lignes d'image verticalement par 4 pixels C_B et C_R horizontalement (en raison du sous-échantillonnage 4:2:2 des plans d'image C_B et C_R) par 1 image temporellement. *mean* signifie que l'on prend la valeur moyenne pour chaque bloc. *euclid* signifie que l'on calcule la distance euclidienne entre le vecteur (C_B, C_R) issu du flux d'origine et le vecteur (C_B, C_R) issu du flux traité pour chaque bloc S-T. *std* signifie que l'on utilise la fonction de regroupement spatial *std*. *10%* signifie que l'on utilise la fonction de regroupement temporel *10%*. *clip_0.8* signifie que l'on applique une coupure de la valeur finale du paramètre à 0,8 (autrement dit, on remplace toute valeur inférieure à 0,8 par 0,8 puis on soustrait 0,8).

Y_hv13_angle0.225_rmin20_8×8_6F_mean_3_ratio_loss_below5%_mean_square_clip_0.05

Y signifie que l'on utilise le plan d'image de luminance. *hv*13 signifie que l'on filtre les images Y au moyen des gabarits spatiaux 13×13 du § D.7.2.1 en vue de l'extraction de la caractéristique f_{HV13} décrite au § D.7.2.2 (autrement dit, les images *HV* et $\overline{HV}$ sont créées et traitées séparément jusqu'à l'application du seuil de perceptibilité). *angle0.225* et *rmin20* signifient que l'on utilise un $\Delta\theta$ de 0,225 radians et un r_{min} de 20 pour le calcul de la caractéristique f_{HV13}. $8 \times 8\_6F$ signifie que l'on subdivise le flux vidéo en régions S-T contenant huit lignes d'image verticalement par huit pixels horizontalement par un cinquième de seconde temporellement (c'est-à-dire 6 images NTSC, 5 images PAL). *mean* signifie que l'on prend la valeur moyenne de *HV* et $\overline{HV}$ pour chaque bloc S-T. *3* signifie que l'on applique un seuil de perceptibilité à ces moyennes et que l'on remplace toute valeur inférieure à 3,0 par 3,0. On calcule ensuite la caractéristique f_{HV13} décrite au § D.7.2.2 comme étant le rapport entre la moyenne coupée de *HV* et la moyenne coupée de $\overline{HV}$, comme spécifié dans *hv13_angle0.225_rmin20*, le sous-nom propre à la caractéristique. *ratio_loss* signifie que l'on applique la fonction de comparaison *ratio_loss* à la caractéristique f_{HV13} du flux d'origine et à la caractéristique f_{HV13} correspondante du flux traité pour chaque bloc S-T. *below5%* spécifie la fonction de regroupement spatial. *mean* spécifie la fonction de regroupement temporel. *square* spécifie la fonction non linéaire appliquée à la valeur du paramètre issue du regroupement temporel. *clip_0.05* représente la fonction de coupure pour laquelle toute valeur inférieure à 0,05 est remplacée par 0,05 puis 0,05 est soustrait de la valeur résultante (on rappelle qu'un paramètre ne prenant que des valeurs négatives devient un paramètre ne prenant que des valeurs positives après application de la fonction non linéaire *square*).

Y_contrast_ati_4×4_6F_std_3_ratio_gain_mean_10%

Y signifie que l'on utilise le plan de luminance. *contrast_ati* signifie que l'on calcule deux versions filtrées distinctes de l'image en vue de l'extraction de la caractéristique $f_{CONTRAST\_ATI}$ décrite au § D.7.6. Le premier filtre, *contrast*, prend directement en considération les plans de luminance (§ D.7.4). Le second filtre, *ati*, prend en considération les images correspondant aux différences entre les plans de luminance successifs (§ D.7.5). Les images *contrast* et *ati* sont traitées séparément jusqu'à l'application du seuil de perceptibilité. $4 \times 4\_6F$ signifie que les deux flux vidéo sont subdivisés en régions S-T contenant quatre lignes d'image verticalement par quatre pixels horizontalement par un cinquième de seconde temporellement (par exemple 6 images NTSC, 5 images PAL). Le premier bloc S-T d'images *ati* ne contiendra en réalité que 5 images et non pas 6 car une image *ati* ne peut pas être générée pour la première image de la séquence (en effet, il n'existe pas d'image antérieure dans le temps disponible). Cette exception est spécifiée dans le cadre du sous-nom propre à la caractéristique. *std* signifie que l'on calcule l'écart type pour chaque bloc. Ensuite, comme spécifié au § D.7.6, on applique un seuil de perceptibilité de 3 aux deux caractéristiques *contrast* et *ati* (on remplace toute valeur inférieure à 3 par 3,0). Ensuite, on multiplie la valeur de *contrast* avec la valeur de *ati* pour chaque bloc S-T (on trouvera dans la note de bas de page du § D.7.6 les instructions particulières sur la manière de procéder à cette multiplication) et on poursuit les calculs avec cette image de caractéristique combinée. *ratio_gain* est la fonction de comparaison utilisée pour comparer la caractéristique du flux d'origine et la caractéristique du flux traité pour chaque bloc S-T. *mean* est la fonction de regroupement spatial. *10%* est la fonction de regroupement temporel.

D.9 Modèle général

Le présent paragraphe contient une description complète du calcul de la qualité VQM selon le modèle général (désignée par VQM_G). Ce calcul est optimisé de manière à obtenir la corrélation maximale entre les mesures objectives et les mesures subjectives pour une large plage de niveaux de qualité vidéo et de débits binaires. Le modèle général comporte des paramètres objectifs pour la mesure des effets perçus d'une grande variété de dégradations telles que le flou, la distorsion due à la subdivision en blocs, les mouvements saccadés/non naturels, le bruit (à la fois dans les canaux de luminance et de chrominance) et les blocs erronés (par exemple ce que l'on peut généralement voir lorsque des erreurs de transmission numérique sont présentes). Le calcul décrit ici consiste en une combinaison linéaire de paramètres de qualité vidéo dont les conventions de dénomination sont décrites au § D.8.6. Les paramètres de qualité vidéo ont été choisis sur la base des critères d'optimisation donnés ci-dessus. Ce calcul donne des valeurs comprises entre zéro (pas de dégradation perçue) et environ un (dégradation perçue maximale). Pour pouvoir comparer les résultats avec ceux obtenus par la méthode à double stimulus utilisant une échelle de qualité continue (DSCQS), on multiplie les résultats VQM_G par 100.

La conception du modèle général repose sur des séquences vidéo conformes à la Rec. UIT-R BT.601-5 qui ont été évaluées subjectivement à une distance de visualisation de six hauteurs d'image. Lorsqu'on analyse les séquences vidéo pour différentes distances de visualisation, il faut appliquer un facteur de correction aux résultats. Plus la distance de visualisation est grande, moins les dégradations sont visibles; plus la distance de visualisation est petite, plus les dégradations sont visibles. Il convient de faire attention lorsqu'on compare les résultats pour des séquences vidéo qui sont observées à des distances de visualisation différentes.

La qualité VQM_G est donnée par une combinaison linéaire de sept paramètres. Quatre paramètres sont fondés sur des caractéristiques extraites des gradients spatiaux de la composante de luminance Y (§ D.7.2.2), deux paramètres sont fondés sur des caractéristiques extraites du vecteur formé par les deux composantes de chrominance (C_B, C_R) (voir § D.7.3) et un paramètre est fondé sur les caractéristiques de contraste et d'information temporelle absolue, toutes deux extraites de la composante de luminance Y (§ D.7.4 et D.7.5, respectivement). La qualité VQM_G est donnée par

VQM_G =

$$\{-0.2097 \times Y\_si13\_8\times8\_6F\_std\_12\_ratio\_loss\_below5\%\_10\%$$

$$+0.5969 \times Y\_hv13\_angle0.225\_rmin20\_8\times8\_6F\_mean\_3\_ratio\_loss\_below5\%\_mean\_square\_clip\_0.06$$

$$+0.2483 \times Y\_hv13\_angle0.225\_rmin20\_8\times8\_6F\_mean\_3\_log\_gain\_above95\%\_mean$$

$$+0.0192 \times color\_coher\_color\_8\times8\_1F\_mean\_euclid\_std\_10\%\_clip\_0.6$$

$$-2.3416 \times [Y\_si13\_8\times8\_6F\_std\_8\_log\_gain\_mean\_mean\_clip\_0.004 \mid^{0.14]}$$

$$+0.0431 \times Y\_contrast\_ati\_4\times4\_6F\_std\_3\_ratio\_gain\_mean\_10\%$$

$$+0.0076 \times color\_coher\_color\_8\times8\_1F\_mean\_euclid\_above99\%tail\_std\} \mid_{0.0}$$

Il est rappelé que les caractéristiques ci-dessus pour le modèle général avec une dimension temporelle de "6F" correspondent en réalité à cinq images vidéo PAL (625 lignes).

L'élévation au carré du paramètre hv_loss est nécessaire pour linéariser la réponse du paramètre par rapport aux données subjectives. Il est à noter que, comme le paramètre hv_loss devient positif après l'élévation au carré, on utilise un poids multiplicatif positif. Il est par ailleurs à noter que le paramètre hv_loss est coupé à 0,06, le paramètre color est coupé à 0,6 et le paramètre si_gain est coupé à 0,004. Le paramètre si_gain est le seul paramètre du modèle correspondant à une *amélioration* de la qualité (comme le paramètre si_gain est positif, un poids négatif conduit à des contributions négatives à la qualité VQM, autrement dit à des améliorations de la qualité). Le paramètre si_gain mesure les améliorations de la qualité qui résultent de l'accentuation des contours. Une coupure du paramètre à un seuil supérieur de 0,14 immédiatement avant la multiplication par le poids associé au paramètre empêche toute amélioration excessive de la qualité VQM de plus de 1/3 d'une unité de qualité, qui est l'amélioration maximale observée dans l'ensemble général des données subjectives (autrement dit, l'accentuation des contours opérée par un circuit fictif de référence ne permet d'améliorer la qualité que dans une faible mesure).

La qualité VQM totale (une fois que les contributions de tous les paramètres ont été ajoutées) est coupée à un seuil inférieur de 0,0 pour éviter les valeurs VQM négatives. Enfin, une fonction d'écrasement autorisant un maximum de 50% de dépassement est appliquée aux valeurs VQM supérieures à 1,0 afin de limiter les valeurs VQM qui sont associées à des séquences vidéo comportant de fortes distorsions et qui se situent en dehors de la plage des données subjectives disponibles.

Si $VQM_G > 1{,}0$, alors $VQM_G = (1 + c) \times VQM_G / (c + VQM_G)$, où $c = 0{,}5$.

Les valeurs de la qualité VQM_G calculées comme indiqué ci-dessus seront supérieures ou égales à zéro et présenteront une valeur nominale maximale de un. La valeur de la qualité VQM_G peut parfois être supérieure à un pour des scènes vidéo comportant de très fortes distorsions.

D.10 Références informatives

[D-1] Recommandation UIT-R BT.500-11, *Méthodologie d'évaluation subjective de la qualité des images de télévision*.

[D-2] Recommandation UIT-T H.261 (1993), *Codec vidéo pour services audiovisuels à p × 64 kbit/s*.

[D-3] Recommandation UIT-T J.143 (2000), *Prescriptions d'utilisateur relatives aux mesures objectives de la qualité vidéo perçue en télévision numérique par câble*.

[D-4] Recommandation UIT-T P.910 (1999), *Méthodes subjectives d'évaluation de la qualité vidéographique pour les applications multimédias*.

[D-5] Recommandation UIT-T P.931 (1998), *Mesure du temps de transmission, de la synchronisation et du débit de trames dans les communications multimédias*.

[D-6] Jain A.K., *Fundamentals of Digital Image Processing* (1989), Englewood Cliffs, NJ: Prentice-Hall Inc., pp. 348-357.

[D-7] SMPTE 125M, *Television – Component Video Signal 4:2:2 – Bit-Parallel Digital Interface*, Society of Motion Picture and Television Engineers, 595 West Hartsdale Avenue, White Plains, NY 10607.

[D-8] SMPTE 170M, *SMPTE Standard for Television – Composite Analog Video Signal – NTSC for Studio Applications*, Society of Motion Picture and Television Engineers, 595 West Hartsdale Avenue, White Plains, NY 10607.

[D-9] SMPTE Recommended Practice 187 – 1995, *Center, Aspect Ratio, and Blanking of Video Images*, Society of Motion Picture and Television Engineers, 595 West Hartsdale Avenue, White Plains, NY 10607.

[D-10] Wolf S. and Pinson M., *In-service performance metrics for MPEG-2 video systems*, in Proc. Made to Measure 98 – Measurement Techniques of the Digital Age Technical Seminar (1998), conférence technique cofinancée par l'International Academy of Broadcasting (IAB), l'UIT et la Technical University of Braunschweig (TUB), Montreux, Suisse, 12-13 novembre 1998.

[D-11] Wolf S. and Pinson M., *Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system* (1999), in Proc. SPIE International Symposium on Voice, Video, and Data Communications, Boston, MA, septembre 1999.

[D-12] Wolf S. and Pinson M., *The relationship between performance and spatial-temporal region size for reduced-reference, in-service video quality monitoring systems* (2001), in Proc. SCI/ISAS 2001 (Systematics, Cybernetics, and Informatics/Information Systems Analysis and Synthesis), juillet 2001, pp. 323-328.

[D-13] Wolf S. and Pinson M., *Video Quality Measurement Techniques* (2002), NTIA Report 02-392, juin 2002.

[D-14] Pinson M. and Wolf S., *Video Quality Measurement User's Manual* (2002), NTIA Handbook 02-1, février 2002.

[D-15] Document de référence UIT-T (2004), *Objective perceptual assessment of video quality: Full reference television*.

D.11 Données objectives brutes sur les mesures de qualité vidéo (VQM)

Le présent paragraphe expose l'ensemble des données objectives brutes de la NTIA sur les essais VQEG (FR-TV) Phase 2.

Résumé concernant les données brutes

Le modèle général élaboré par la National Telecommunications and Information Administration (NTIA) a été conçu au départ pour des valeurs en sortie sur une échelle nominale de 0 à 1 où 0 correspond à la perception d'aucune dégradation et 1 à la perception d'une dégradation maximale. Toutefois, l'exécutable binaire soumis au test VQEG FR-TV Phase II a transformé les valeurs (0, 1) du modèle général à (0, 100) dans un souci d'adaptation avec l'échelle de qualité continue à double stimulus (DSCQS). Pour les données brutes rapportées dans la présente annexe nous avons supprimé le facteur de multiplication par 100 (c'est-à-dire multiplication par 100) pour retrouver l'échelle originale (0, 1) du modèle général.

Les valeurs du modèle général calculées ici ont utilisé les 8 secondes centrales de chaque clip vidéo rejetant les 10 trames supplémentaires au début et à la fin de chaque fichier vidéo comme décrit dans le programme de tests VQEG Phase II FR-TV. Pour les routines d'étalonnage, on a utilisé une incertitude de 30 trames et une fréquence de 15 images (voir le § D.6). Par ailleurs, la région

d'intérêt spatiale (SROI) utilisée pour calculer la valeur VQM pour chaque clip a été choisie comme suit:

1) pour les systèmes vidéo à 525 lignes, on utilise une région SROI par défaut de 672 pixels × 448 lignes centrée dans l'image vidéo. Pour les systèmes vidéo à 625 lignes, on utilise une région SROI par défaut de 672 pixels × 544 lignes centrée dans l'image vidéo. Ces régions SROI par défaut peuvent être modifiées comme indiqué dans les étapes 2 et 3;

2) le modèle a besoin de 6 pixels/lignes valables supplémentaires sur tous les côtés de la région SROI susmentionnée pour que les filtres spatiaux puissent fonctionner correctement. Si la région valable traitée (PVR, calculée automatiquement selon la méthode donnée au § D.6.2) n'est pas suffisamment large pour englober la région SROI par défaut + 6 pixels/lignes (étape 1), la région SROI est alors réduite par des multiples de 8 pixels/lignes, uniquement dans la direction nécessaire (horizontale ou verticale);

3) la région SROI est toujours centrée horizontalement de façon à ce que l'échantillon gauche commence en un point d'échantillonnage identique pour la luminance/chrominance de la Rec. UIT-R BT.601-5. La région SROI est centrée verticalement de façon à ce que lorsqu'elle est subdivisée en deux trames, le même nombre de lignes est rejeté depuis le haut de chaque trame. Si la taille de la région SROI a été réduite dans l'étape 2, un centrage parfait de la région SROI dans l'image vidéo ne sera peut-être pas possible.

Données objectives brutes pour un systèmes à 525 lignes

Source #	HRC #	NTIA: modèle H
1	1	0,660 (Note)
1	2	0,347
1	3	0,286
1	4	0,178
2	1	0,449
2	2	0,246
2	3	0,119
2	4	0,061
3	1	0,321
3	2	0,167
3	3	0,076
3	4	0,049
4	5	0,396
4	6	0,280
4	7	0,222
4	8	0,183
5	5	0,329
5	6	0,217
5	7	0,159
5	8	0,115
6	5	0,542
6	6	0,266
6	7	0,189
6	8	0,139
7	5	0,258
7	6	0,161
7	7	0,108
7	8	0,076
8	9	0,911
8	10	0,717
8	11	0,721
8	12	0,526
8	13	0,424
8	14	0,311
9	9	0,827
9	10	0,453
9	11	0,512
9	12	0,264
9	13	0,188
9	14	0,124
10	9	0,666
10	10	0,250
10	11	0,375
10	12	0,129
10	13	0,078
10	14	0,153
11	9	0,513
11	10	0,534
11	11	0,407
11	12	0,161
11	13	0,148
11	14	0,159
12	9	0,600
12	10	0,410
12	11	0,471
12	12	0,244
12	13	0,171
12	14	0,114
13	9	0,537
13	10	0,425
13	11	0,346
13	12	0,215
13	13	0,188
13	14	0,169

NOTE – Pour la source 1, HRC 1, le logiciel d'étalonnage soumis au VQEG a abouti à une erreur d'alignement spatial/temporel qui a estimé de façon incorrecte la séquence vidéo traitée qui sera resynchronisée (c'est-à-dire décalée d'une trame, voir le § D.6.1.2. Pour les autres scènes de HRC 1, l'alignement spatial/temporel a été correctement estimé. Il est recommandé au § D.6.1.5.7 de soumettre les résultats de l'étalonnage à un filtrage médian pour toutes les scènes d'un HRC donné afin d'obtenir des estimations d'étalonnage plus fiables pour ce HRC. Toutefois, le programme des essais de Phase II du VQEG précisait que tous les logiciels VQM donnent une estimation de qualité unique pour chaque clip vidéo. Par conséquent, un filtrage médian des résultats d'étalonnage pour toutes les scènes d'un HRC donné n'a pas été autorisé par le programme des essais. Si le filtrage médian des résultats d'étalonnage avait été autorisé, le logiciel VQM aurait correctement aligné ce clip vidéo et la note objective brute aurait été de 0,529.

Données objectives brutes pour un système à 625 lignes

Source #	HRC #	NTIA: modèle H
1	2	0,421
1	3	0,431
1	4	0,264
1	6	0,205
1	8	0,155
1	10	0,123
2	2	0,449
2	3	0,473
2	4	0,312
2	6	0,260
2	8	0,226
2	10	0,145
3	2	0,472
3	3	0,506
3	4	0,308
3	6	0,239
3	8	0,183
3	10	0,146
4	2	0,409
4	3	0,458
4	4	0,384
4	6	0,354
4	8	0,280
4	10	0,232
5	2	0,470
5	3	0,521
5	4	0,260
5	6	0,234
5	8	0,132
5	10	0,083
6	2	0,391
6	3	0,364
6	4	0,290
6	6	0,252
6	8	0,181
6	10	0,169
7	4	0,422
7	6	0,385
7	9	0,336
7	10	0,270
8	4	0,345
8	6	0,311
8	9	0,280
8	10	0,242
9	4	0,344
9	6	0,285
9	9	0,246
9	10	0,192
10	4	0,410
10	6	0,355
10	9	0,313
10	10	0,241
11	1	0,739
11	5	0,468
11	7	0,199
11	10	0,201
12	1	0,548
12	5	0,441
12	7	0,367
12	10	0,307
13	1	0,598
13	5	0,409
13	7	0,321
13	10	0,277

Appendice I

KDDI

Système d'évaluation objective de la qualité vidéo et détermination de la performance

I.1 Domaine d'application

Depuis peu, des services de diffusion et de transmission de télévision numérique commencent à être utilisés dans la pratique. Ces services utilisent des codecs vidéo (dispositifs de codage du signal vidéo) fondés sur MPEG-2, méthode de compression de signaux vidéo numériques normalisée sur le plan international. Les codecs vidéo comprennent des codeurs, qui effectuent la compression, et des décodeurs, qui reconstituent les données vidéo compressées. Ces dispositifs suppriment les informations redondantes parmi l'énorme volume d'informations contenues dans les signaux vidéo. Cela permet de transmettre les informations efficacement en n'utilisant qu'une largeur de bande limitée.

La qualité de signaux vidéo qui ont été compressés et transmis au moyen d'un codec vidéo subit toujours une certaine dégradation. Le niveau de dégradation dépend du contenu de l'image. Généralement, il y a plus de distorsion dans les scènes très rapides (émissions sportives par exemple). Il existe aussi des variations de qualité entre les signaux de sortie produits par différents codecs. MPEG-2 est une norme internationale, mais la qualité de types particuliers de signaux vidéo compressés dépend toujours dans une certaine mesure de l'implémentation du fabricant.

Pour la transmission de télévision, notamment en TV1, TV2 et TV3 (contribution, distribution primaire et distribution secondaire), il faut s'efforcer d'obtenir une qualité invariablement élevée en contrôlant constamment la qualité des images transmises.

Pour la transmission analogique classique en modulation de fréquence, la dégradation de l'image est faible en raison du contenu de l'image ou de la modulation analogique, de sorte que la qualité est stable. Mais pour la transmission de signaux vidéo numériques compressés, la qualité de l'image varie comme décrit ci-dessus en fonction de la nature du contenu et du codec employé, et on s'attend à ce que la vérification de la qualité de ce type de signaux vidéo soit une opération très complexe.

On estime donc nécessaire de normaliser un système d'évaluation de la qualité d'image de codecs vidéo MPEG-2 principalement utilisés en TV1, TV2 et TV3. Dans ces classes, les fonctions suivantes sont considérées comme nécessaires:

- évaluation générique pour divers types de contenu vidéo prise en charge des formats vidéo analogique/numérique composite/en composantes;
- évaluation en temps réel; alignement temporel et spatial précis entre un signal d'origine et un signal en sortie de codec;
- évaluation sensible et précise des distorsions subtiles et complexes.

Cela étant, le présent appendice décrit un nouveau système d'évaluation et son implémentation sur la base des caractéristiques de la perception visuelle humaine, permettant d'obtenir des mesures très précises de la qualité vidéo.

I.2 Système d'évaluation objective de la qualité vidéo

La Figure I.1 illustre le modèle d'évaluation de la qualité d'image à trois couches tel qu'il est vu par l'œil humain. Généralement, l'œil humain ne peut pas voir en un coup d'œil une image dans sa totalité, il ne voit qu'une zone ponctuelle locale dans l'image, qui se situe autour du point de visualisation de l'œil, et reconnaît la texture ainsi que la qualité de la zone en fonction du degré et des caractéristiques du bruit mélangé dans cette texture. La totalité de l'image est observée par un déplacement du point de visualisation parmi les objets, qui sont des composantes de l'image et l'évaluation de la qualité de l'image est également faite pour la totalité de l'image simultanément. Dans ce processus, la qualité de l'image est déterminée par le bruit présent dans l'image. Pour procéder à des mesures objectives de la qualité subjective de l'image, on utilise donc les structures d'image en trois couches macro vers micro (couches objet, texture et bruit) et on propose un système de pondération du bruit de bas en haut qui utilise une certaine fonction de pondération à chaque couche compte tenu de la perception visuelle humaine (Figure I.2).

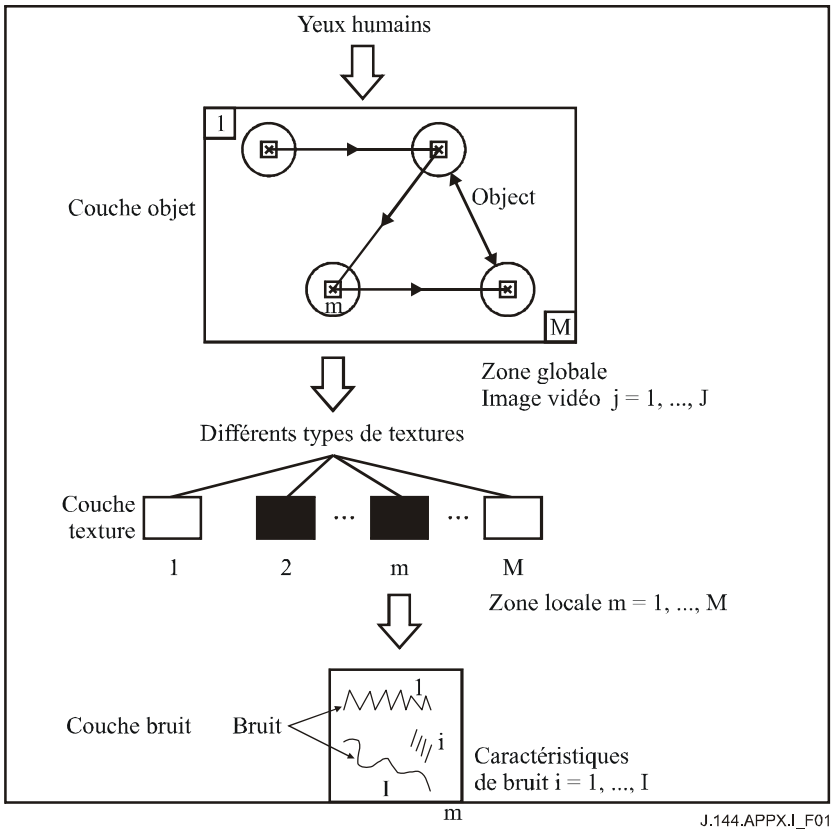


Figure I.1 – Modèle à trois couches pour le signal vidéo

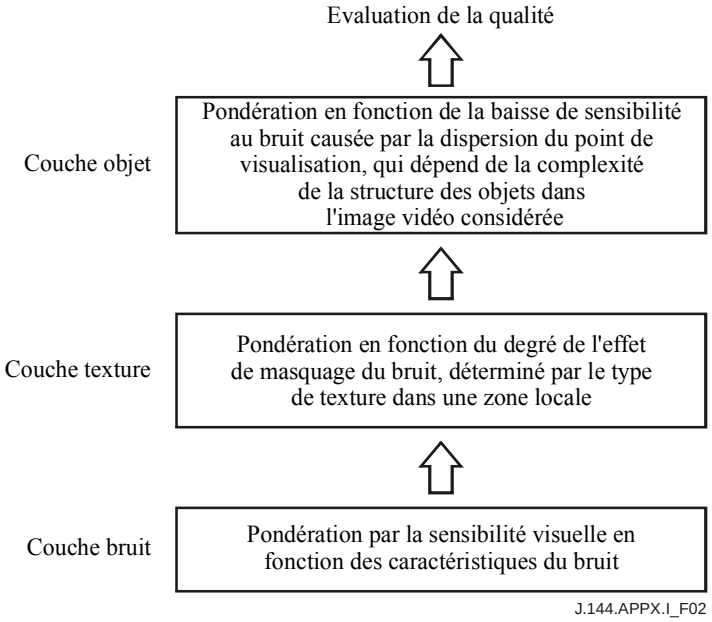


Figure I.2 – Système de pondération du bruit de bas en haut à trois couches

On commence par pondérer, au niveau de la couche bruit, le bruit commun dans un processus de compression vidéo tel que le bruit haute fréquence, le bruit basse fréquence, le bruit de chrominance, le rythme saccadé, le papillotement, etc., suivant le degré et les caractéristiques de ces bruits. Pour cette pondération, il est utile d'effectuer une conversion de fréquence pour classer ces bruits. On procède ensuite, au niveau de la couche texture, à un classement des zones ponctuelles locales en plusieurs groupes suivant le type de leur texture. Ces groupes comprennent par exemple le groupe "texture détaillée" – par exemple forêts, arbres et stade, dans lesquels le bruit est fortement masqué – et le groupe "texture uniforme" – par exemple peau humaine et ciel, dans lesquels tout bruit se reconnaît facilement. Les bruits sont donc pondérés plus ou moins suivant le type de texture. Enfin, au niveau de la couche objet, on prédit le degré de dispersion du point de visualisation en mesurant la complexité de la structure des objets dans l'image vidéo. La pondération des bruits dans l'ensemble de l'image correspond alors à la baisse de sensibilité au bruit causée par cette dispersion.

Afin d'obtenir des expressions mathématiques pour ces processus de pondération, on pose les définitions suivantes:

P(j,m,i): puissance d'un bruit *i* dans une zone locale *m* d'une image *j*;

h_i: fonction de pondération pour un bruit *i*;

C(j,m): texture d'une zone locale (*j*,*m*);

tc: fonction de pondération de bruit dans une texture *C*;

G(j): paramètre caractérisant la complexité de la structure des objets dans l'image *j*;

9 *G*: fonction de pondération de bruit dépendant du degré de dispersion du point de visualisation.

On procède alors à une sommation des bruits en allant de la couche inférieure à la couche supérieure.

Dans la couche bruit, en sommant le bruit pondéré par *hi* correspondant aux caractéristiques de bruit dans la zone locale (*j*,*m*), on calcule l'erreur $WMSE_{NL}$ comme suit:

$$WMSE_{NL}(j,m) = \frac{1}{I}\sum_{i=1}^{I} hi \cdot P(j,m,i) \qquad \text{(I-1)}$$

Ensuite, dans la couche texture, en sommant l'erreur $WMSE_{NL}(j,m)$ sur la totalité de l'image ($m = 1, \ldots, M$) pondérée par tc correspondant à la texture $C(j,m)$ dans la zone locale (j,m), on calcule l'erreur $WMSE_{TL}(j)$ comme suit:

$$WMSE_{TL}(j) = \frac{1}{M}\sum_{m=1}^{M} t_c(j,m) \cdot WMSE_{NL}(j,m) \tag{I-2}$$

Enfin, dans la couche objet, en prenant une valeur moyenne de l'erreur $WMSE_{TL}$ sur les images $j = 1, \ldots, J$ pondérée par $G(j)$ correspondant au degré de dispersion du point de visualisation, on calcule l'erreur $WMSE_{OL}$ comme suit:

$$WMSE_{OL} = \frac{1}{J}\sum_{j=1}^{J} q_G(j) \cdot WMSE_{TL}(j) \tag{I-3}$$

On convertit alors l'erreur $WMSE_{OL}$ en rapport $WSNR$ et on calcule la valeur DSCQS (méthode à double stimulus utilisant une échelle de qualité continue) (0-100%) (voir la définition figurant dans la Rec. UIT-R BT.500-11) comme suit:

$$WSNR(dB) = 10\log_{10}\frac{255^2}{WMSE} \tag{I-4}$$

$$D(\%) = f(WSNR) \tag{I-5}$$

Puissance d'un bruit de zone locale P(*j,m,i*)

On définit tout d'abord la zone locale m comme un bloc carré $m_w \times m_h$.

Supposons que la caractéristique de bruit I=1 soit un bruit pondéré sur le domaine de fréquence.

$$P(j,m,i) = \sum_{q=1}^{m_h}\sum_{p=1}^{m_w}\{X(p,q) - Y(p,q)\}^2$$

où X, Y sont respectivement les valeurs transformées des coefficients de fréquence de l'image d'origine et de l'image codée.

I.3 Implémentation

Le système est constitué de deux parties: un module de synchronisation, qui permet de comparer avec précision le signal vidéo reconstitué et le signal vidéo d'origine, et un module de calcul, qui permet de déterminer la qualité vidéo par référence à des caractéristiques de perception visuelle humaine. La Figure I.3 montre la configuration du système et le Tableau I.2 décrit les principaux paramètres. Comme l'indique le Tableau I.2, les signaux composites (NTSC) et les signaux en composantes avec échantillonnage total sont pris en charge.

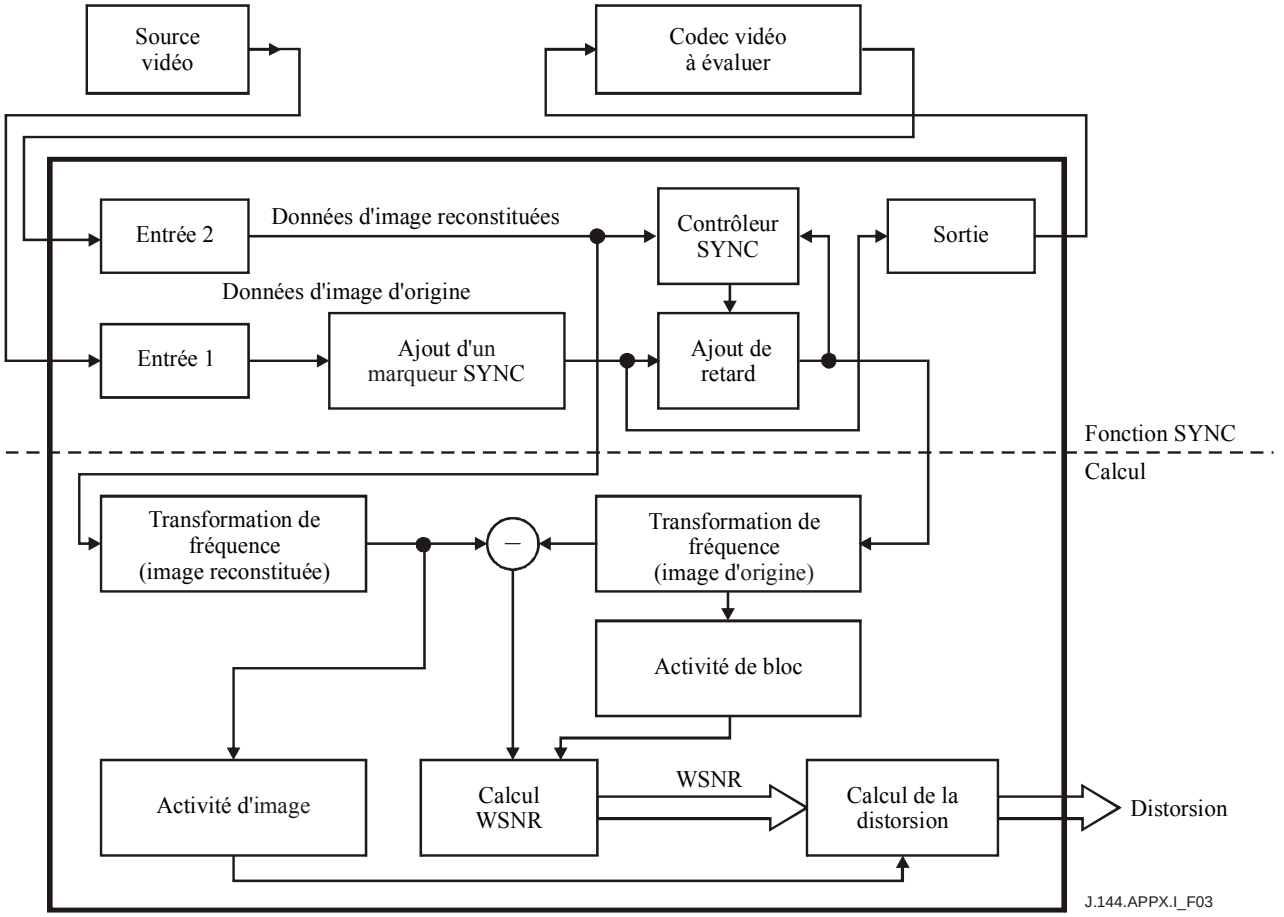


Figure I.3 – Configuration du système

I.3.1 Module de synchronisation

Le module de synchronisation, indépendant du module de calcul, est nécessaire au système d'évaluation en temps réel de la qualité video. On notera qu'il n'est pas indispensable aux calculs en différé, par exemple pour la comparaison de fichier vidéo, à des fins d'évaluation logicielle de la qualité. Ci-dessous figure la description à titre d'exemple d'une des méthodes de synchronisation.

Les signaux de télévision provenant de la source vidéo d'origine sont lus dans le système via le module d'entrée 1 et sont affectés d'un marqueur de synchronisation qui varie avec chaque image. Le marqueur, par exemple, s'apparente, à une onde sinusoïdale, dont la fréquence est modulée par le nombre de trames. Les images avec marqueurs sont ensuite envoyées au module de retard, où elles sont gardées en mémoire. En même temps, les images sont envoyées via le module de sortie au codec vidéo à évaluer. Le codec vidéo compresse les images, qui sont lues à nouveau dans le système via le module d'entrée 2 et comparées avec les images marquées qui sont mémorisées dans le module de retard du codec vidéo à évaluer. Enfin, le module de synchronisation effectue un alignement temporel (retard entre les images) et spatial (déplacement de ligne et de pixel) précis, de sorte que la dégradation de la qualité décrite ci-dessous soit aussi proche que possible de l'évaluation subjective faite par des observateurs humains.

Ces opérations assurent la synchronisation nécessaire pour l'évaluation et les marqueurs utilisés dans ces opérations sont conçus pour bien fonctionner même dans le cadre du processus conduisant à un signal très distordu (forte compression, séparation Y/C, filtrages dans un codec vidéo, etc.).

I.3.2 Module de calcul

A la différence de la vision humaine, le calcul de la qualité de l'image suit une approche de bas en haut, établissant l'ensemble à partir des diverses parties. Tout d'abord, afin d'évaluer l'effet de

variations de sensibilité dues aux fréquences spatiales du bruit, une valeur de différence (bruit) est obtenue pour les composantes fréquentielles de l'image d'origine et de l'image reconstituée. Cette valeur est insérée dans le module WSNR valeur pondérée du rapport signal sur bruit (WSNR, *weighted signal-to-noise ratio*), qui assigne différents poids de sensibilité à chaque région fréquentielle. En même temps, ce module obtient une valeur (l'activité de bloc) qui indique si chaque bloc de l'image est inactif ou actif. On applique également l'effet de masquage de bruit pour obtenir une valeur WSNR globale.

Enfin, on obtient une valeur indiquant la taille des objets constituant l'image (l'activité d'image), ce qui permet au système d'évaluer la baisse de sensibilité au bruit due à la dispersion, et on obtient le niveau de dégradation de la qualité en appliquant cette baisse au module WSNR.

Tableau I.1 – Principaux paramètres

Format de signal vidéo applicable	Signal composite NTSC Signal en composantes 525/60 Signal numérique série D1
Fréquence d'échantillonnage (entrée analogique)	14,318 MHz (NTSC) 13,5 MHz (composante Y) 6,75 MHz (composante C)
Codec applicable	Codec MPEG-1,2 Codec composite, etc.
Zone d'évaluation effective	768 pixels~480 lignes (NTSC) 720 pixels~480 lignes (composante Y) 360 pixels~480 lignes (composante C)
Analyse du signal	Transformation de Hadamard (NTSC) Transformée discrète en cosinus (composante) Autre solution: transformée de Fourier
Pondération du bruit	Sensibilité visuelle à la fréquence spatiale Effet de masquage du bruit Dispersion du point de visualisation
Résultat de l'évaluation	Evaluation de la qualité de l'image (distorsion,%) WSNR (dB) SNR (dB)
Interface du signal de commande	RS-232C

I.4 Résultats de vérification

Les résultats obtenus par le système d'évaluation proposé ont été comparés avec les résultats de tests d'évaluation subjective qui ont déjà été associés à des notes conformément à la Rec. UIT-R BT.500-11. Les cibles d'évaluation sont MPEG-2 SP@ML avec 5 Mbit/s, 7 Mbit/s et 10 Mbit/s appliqués aux signaux de test de télévision en composantes 4:2:2 de la Rec. UIT-R BT.601-5. Elles comprennent 17 données comprenant Mobile, Jardin fleuri, Leaders, etc. On a donc au total 17 données × 3 débits binaires = 51 échantillons (Tableau I.2).

Pour ces échantillons, nous avons fait un test d'évaluation subjective deux jours différents (les 23 et 24 mars 1995) dans les mêmes conditions et avec les mêmes observateurs. Le "triangle" des résultats de l'évaluation objective et des deux évaluations subjectives est représenté sur la Figure I.4.

Tableau I.2 – Liste des données de test

1	Susie
2	Onde
3	Tennis de table
4	Mobile et calendrier
5	Feuilles d'automne
6	Football
7	Tempête
8	Leaders
9	Crédits
10	Croisière
11	Bicyclette
12	Equitation
13	Fleurs d'été
14	Grande roue
15	Jardin fleuri
16	Port de Kiel 4
17	Pelotes de laine

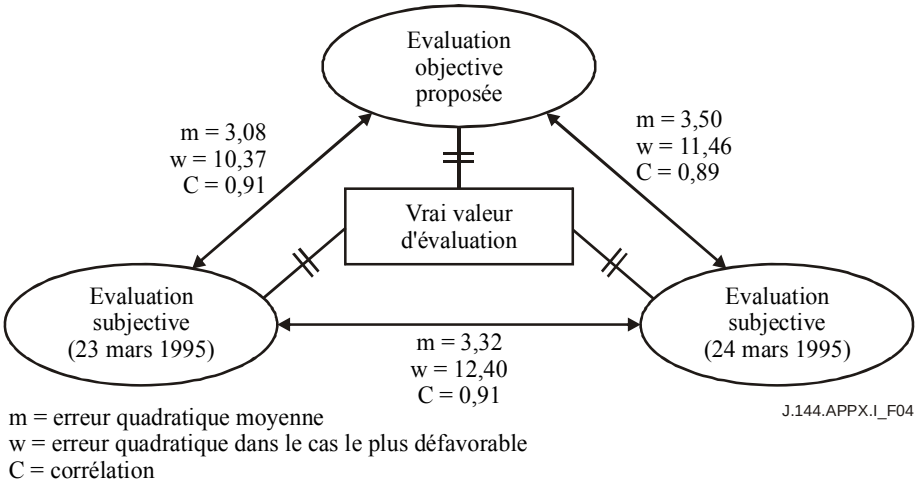


Figure I.4 – Comparaisons avec les tests d'évaluation subjective

La Figure I.4 montre que les précisions d'évaluation exprimées par l'erreur quadratique moyenne, l'erreur quadratique dans le cas le plus défavorable et la corrélation et associées aux résultats des trois évaluations sont pratiquement égales, si on se place au centre du triangle, qui correspond à la vraie valeur d'évaluation. En outre, la Figure I.5 montre des distributions des 51 échantillons pour l'évaluation objective et les deux évaluations subjectives. Dans les trois graphes, les échantillons sont distribués aléatoirement mais on peut voir une différence subtile dans chaque distribution. La distribution associée à la comparaison des évaluations subjectives des 23 et 24 mars est uniformément aléatoire tandis que dans le cas des comparaisons de l'évaluation objective et d'une évaluation subjective, on constate une inégalité dans les distributions en fonction de la plage de notes. En effet, les deux graphes associés aux évaluations des 23 et 24 mars en fonction de l'évaluation objective donnent des tracés d'échantillons avec une plus forte corrélation à 20%-40%

mais avec une plus faible corrélation à 10%-20%. Il sera procédé à un complément d'étude pour remédier à cette inégalité.

On peut donc conclure qu'il est possible d'utiliser le système proposé en complément de la Rec. UIT-R BT.500-11.

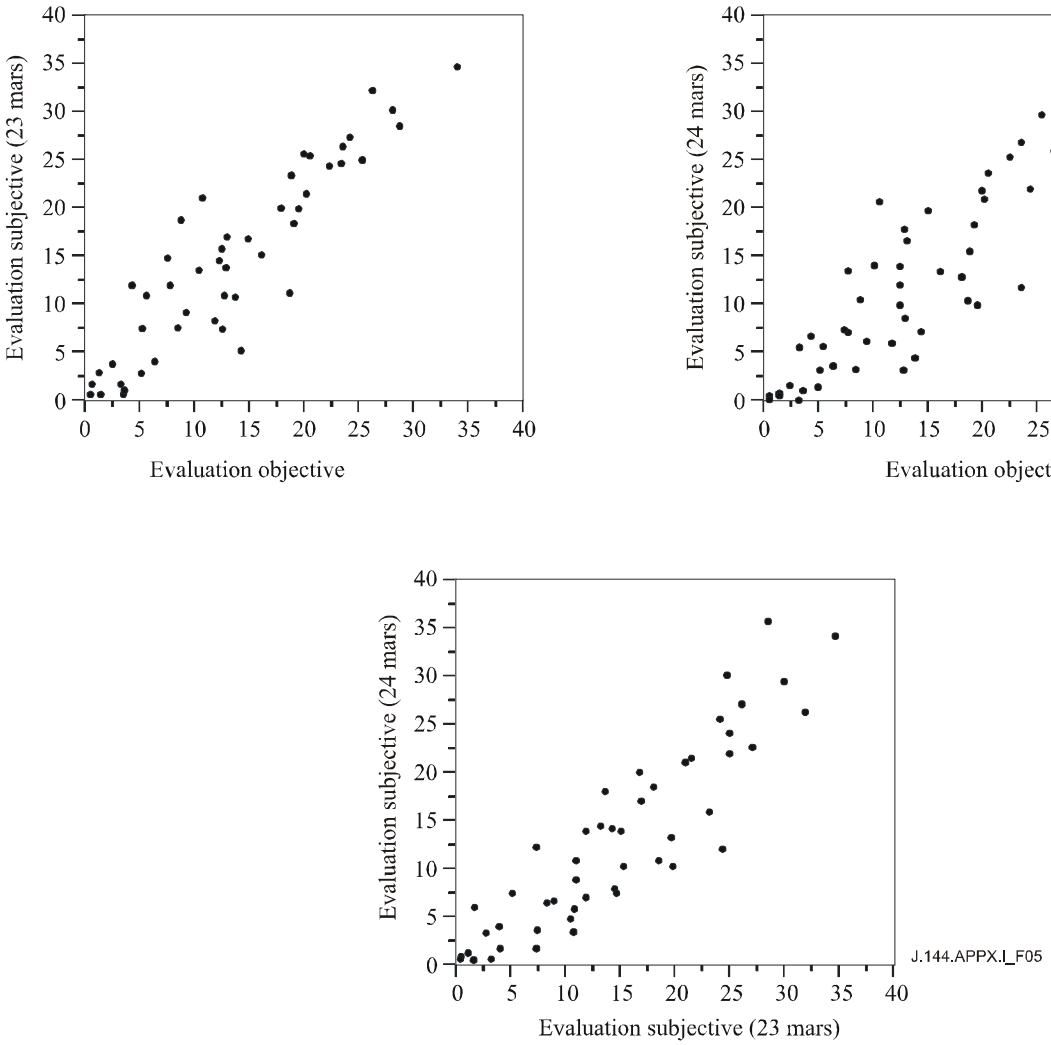


Figure I.5 – Comparaisons entre l'évaluation objective et les deux évaluations subjectives

Appendice II

Tektronix Inc. and Sarnoff Corporation

Mesure objective de la qualité vidéo perceptuelle au moyen d'une technique d'image de référence fondée sur la mesure des unités de différence tout juste perceptibles (JND)

II.1 Domaine d'application, objet et application

II.1.1 Domaine d'application

Le présent Appendice spécifie une méthode de mesure objective de la qualité vidéo perceptuelle fondée sur les unités de différence tout juste perceptibles, applicables lorsqu'on dispose du signal vidéo de référence complet. Il s'agit d'une méthode à deux extrémités connue sous le nom de méthode indice de qualité de l'image (PQR, *picture quality rating*) illustré à la Figure II.1.

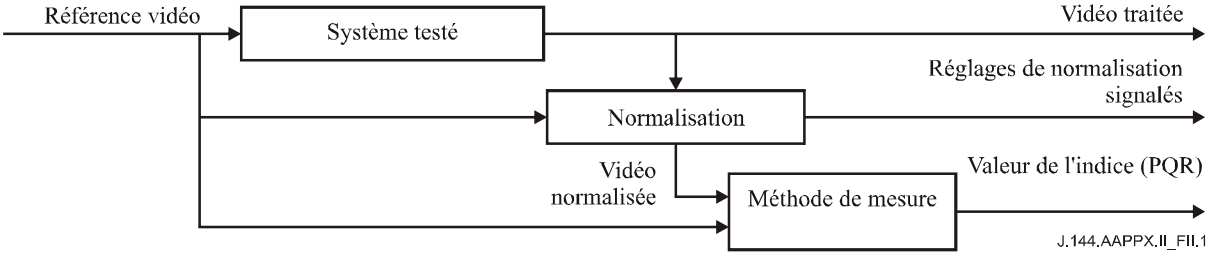


Figure II.1 – Schéma de principe du système

La méthode PQR décrite dans le présent appendice repose sur le traitement de la composante vidéo numérique à 8 bits telle qu'elle est définie dans la Rec. UIT-R BT.601-5, de façon représentative de la réponse du système de vision humaine. En raison du caractère perceptuel de la mesure, différentes méthodes de compression peuvent être prises en compte (MPEG, NTSC, PAL, etc.). En outre, le système de transmission peut comporter une concaténation des méthodes de compression ou se limiter à l'exécution d'un transfert pour les besoins de l'évaluation d'un codec (combinaison codeur/décodeur). Les résultats de la méthode PQR sont indiqués sous forme de valeur de l'indice objectif de qualité d'image (PQR).

L'application de la méthode PQR exige la normalisation du signal vidéo traité. Le présent appendice indique uniquement l'algorithme de la méthode PQR et spécifie la précision de la normalisation. Les spécifications concernant la normalisation figurent au § II.3.1.

II.1.2 Objet

Le présent appendice donne la description technique d'une méthode actuellement utilisée de mesure objective de la qualité vidéo perceptuelle. Bien que des méthodes perfectionnées puissent être mises au point à l'avenir, le présent appendice décrit une méthode de mesure de la qualité vidéo indispensable à la prise en charge de l'interconnexion et de l'interopérabilité des réseaux de télécommunication aux interfaces avec les systèmes utilisateur final, les opérateurs, les fournisseurs d'informations et de services enrichis et les équipements des locaux client.

II.1.3 Application et limites

II.1.3.1 Applications de la méthode PQR

La mise en œuvre de toute méthode fondée sur une image de référence complète implique certaines restrictions pratiques. Toutefois, la méthode PQR spécifiée dans le présent appendice n'est pas limitée aux évaluations en laboratoire. En effet, la méthode PQR est adaptée à certaines applications spécifiques telles que:

– évaluation, spécification, essai d'acceptation de codecs;

– contrôle de transmission à la source en temps réel et en service;

– évaluation de transmission à distance par rapport à une copie de l'image de référence

Lors de l'utilisation de la méthode PQR, il faut tenir compte soigneusement des limites de précision spécifiées au § II.1.3.4.

II.1.3.2 Limites

D'après la validation exposée dans le rapport final Phase I du VQEG (Document de référence UIT-T A) la méthode PQR spécifiée dans le présent appendice est adaptée aux séquences vidéo de courte durée (de 2 à 10 secondes) à une distance de visualisation de 5 hauteurs d'image (à raison de 480 lignes par hauteur d'image). Les valeurs de l'indice PQR seront utiles aux distances de visualisation plus courtes et plus longues, pour lesquelles il est admis que la perception humaine de la dégradation de l'image diminue lorsque la distance de visualisation augmente, tandis que les valeurs de l'indice PQR restent constantes. En dépit de la possibilité de modifier l'algorithme pour tenir compte de la perception humaine à d'autres distances de visualisation, ces modifications ne forment pas partie intégrante du présent appendice.

L'Appendice II.B indique une liste détaillée des facteurs d'essai, des techniques de codage et des applications concernant la précision de la méthode PQR, sur la base des données VQEG retenues.

Bien que la méthode de normalisation spécifiée au § II.3.1 puisse servir à déceler les modifications de la taille de l'image (notamment celle produite par une unité d'effets spéciaux) il n'a pas été établi qu'elle permettait de fournir l'information requise pour déterminer l'importance de la variation de taille. La méthode PQR ne permet pas d'évaluer des images dont la taille diffère de la taille d'origine de l'image introduite dans le système sous test, ou dont les décalages verticaux diffèrent d'un nombre entier de lignes.

La définition des classes de télévision figure à l'Annexe B/P.911. Le présent appendice vise à définir les mesures concernant les classes TV1, TV2 et TB3 indiquées ci-après. Ces classes se distinguent des classes vidéo multimédia dans la mesure où les codeurs assurent en permanence un débit de trame constant et un fonctionnement à temps de latence constant. Tandis que le système de compression permet de réduire le nombre de pixels (généralement uniquement dans le sens horizontal) dans le cadre du processus de codage, le signal de sortie du décodeur sera constitué d'une composante vidéo de résolution complète conformément à la Rec. UIT-R BT.601-5.

- TV 0 Classe vidéo sans perte – selon la Rec. BT.601-5, 8 bits/échantillon, utilisée pour des applications sans compression.
- TV1 – Classe vidéo utilisée pour la post-production complète avec de nombreuses couches de correction et de traitement, en transmission par réseau interne de studio. Perçue comme transparente en comparaison de la classe TV 0.
- TV 2 – Classe vidéo utilisée pour des modifications simples, des corrections peu nombreuses, des incrustations de caractères/logos, insertion de programmes et la transmission entre maillons. En radiodiffusion, il s'agira par exemple, d'une transmission de réseaux à filiale. Autres exemples: liaison descendante régionale d'un système câblé vers une tête de réseau local; système de vidéo conférence de haute qualité. Perçue comme étant presque transparente en comparaison de la classe TV 0.

- TV 3 – classe vidéo utilisée pour l'acheminement vers les foyers privés/consommateurs (sans modifications). Autres exemples: système câblé entre têtes de réseau local et terminal privé. Système de vidéo conférence de qualité moyenne à élevée. Légers défauts présents par rapport à la classe TV 2.

Ces différentes classes possèdent toutes une latence constante (mais pas nécessairement faible) dans un seul sens, ainsi qu'une variation de délai constante. La méthode PQR spécifiée dans le présent appendice n'est pas adaptée aux applications de vidéo conférence qui répètent des champs ou ne sont pas conformes aux spécifications de latence et de délai des classes vidéo. En outre, la méthode PQR est uniquement applicable aux systèmes classiques de transmission par radiodiffusion, comportant de très faibles taux d'erreurs, tels que ceux pris en compte dans les essais VQEG.

II.1.3.3 Comparaison avec l'évaluation subjective

Il est souhaitable d'avoir une bonne corrélation entre les mesures objectives et l'évaluation subjective de la qualité afin d'obtenir une qualité de service optimale, mais les mesures objectives ne sauraient se substituer directement à l'évaluation subjective de la qualité. Les évaluation de qualité subjective sont des procédures soigneusement élaborées qui ont pour but de déterminer l'opinion moyenne de spectateurs au sujet de séquences vidéo pour une application donnée. Les résultats de ce type d'évaluation sont très utiles dans la conception des systèmes et les tests d'évaluation des performances. L'évaluation de la qualité subjective pour une application différente dans d'autres conditions donnera toujours des résultats révélateurs, même si les notes d'opinion pour le même ensemble de séquences seront sans doute différentes. Les mesures objectives sont destinées à une large gamme d'applications produisant des résultats identiques pour un même ensemble de séquences vidéo. Le choix des séquences vidéo qu'il convient d'utiliser et l'interprétation des mesures objectives figurent parmi les facteurs que l'on peut faire varier pour une application donnée. Les mesures objectives et les évaluations subjectives de la qualité sont donc complémentaires plutôt qu'interchangeables. Si les évaluations subjectives répondent à des besoins liés à la recherche, les mesures objectives sont nécessaires dans la spécification des équipements ainsi que la surveillance et la mesure quotidienne des performances des systèmes.

II.1.3.4 Précision et étalonnage croisé

Pour les valeurs PQR indiquées par l'équation II-36, la précision VQM et les méthodes d'étalonnage croisé décrites en détail dans la Rec. UIT-T J.149 ont été appliquées aux données subjectives VQEG sur 525 lignes une fois supprimés les circuits HRC 15 et 16 (voir Document de référence UIT-T II.A). Ces données ont été choisies de façon à être représentatives des systèmes de télédiffusion américains. Lors de l'évaluation d'une mesure VQM, les données objectives font l'objet d'un lissage de courbe sur une échelle normalisée (0-1) afin d'obtenir la meilleure corrélation entre données subjectives et objectives. Pour le calcul de la précision, cette transformation joue également un rôle important, puisqu'elle permet d'obtenir la meilleure cohérence en termes de niveau de confiance statistique du pouvoir de résolution (précision) dans tout le domaine de variation des valeurs objectives.

La fonction de lissage intitulée Logistique II au § 4.2/J.149 est utilisée, c'est-à-dire tenue d'atteindre ou de tendre de façon asymptotique vers la valeur 0 pour des images de qualité constamment améliorée ($a = -e^{-cd}$).

$$VQM = a + \frac{b-a}{1+e^{-c(PQR-d)}}$$

La fonction logistique est également tenue d'atteindre une valeur maximale égale à 1 (b=1) dans l'intervalle normalisé 0-1 ou la valeur 100 dans l'intervalle DSCQS (voir Document de référence UIT-R BT.500-11 à l'Appendice II.A) utilisée pour le contrôle subjectif de la mesure VQM en présence d'une image de référence complète. Une fois convertie dans l'échelle commune, la

méthode PQR peut être contre-étalonnée avec n'importe quel autre modèle VQM. Compte tenu de ces contraintes, l'expression de la fonction logistique utilisée par la méthode PQR figure ci-dessous.

$$VQM = \frac{1 - e^{-c \times PQR}}{1 + e^{c(d - PQR)}}$$

Où:

c = 0,5031

d = 9,634

La Figure II.2 représente le diagramme établi à partir des données VQEG choisies et la fonction logistique ajustée obtenue pour les valeurs PQR d'origine. La Figure II.3 représente le passage des valeurs PQR d'origine aux valeurs comprises dans l'intervalle commun au moyen de la fonction logistique ajustée. Bien que la courbe de la fonction logistique ajustée puisse être calculée au-delà des points de données VQEG jusqu'à des valeurs limites adaptées au calcul des valeurs PQR et aux indices DSCQS fondés sur des données subjectives, sa validité pour les besoins des calculs de précision est limitée estime-t-on aux domaines des valeurs PQR d'origine allant de 3,5 à 10,5.

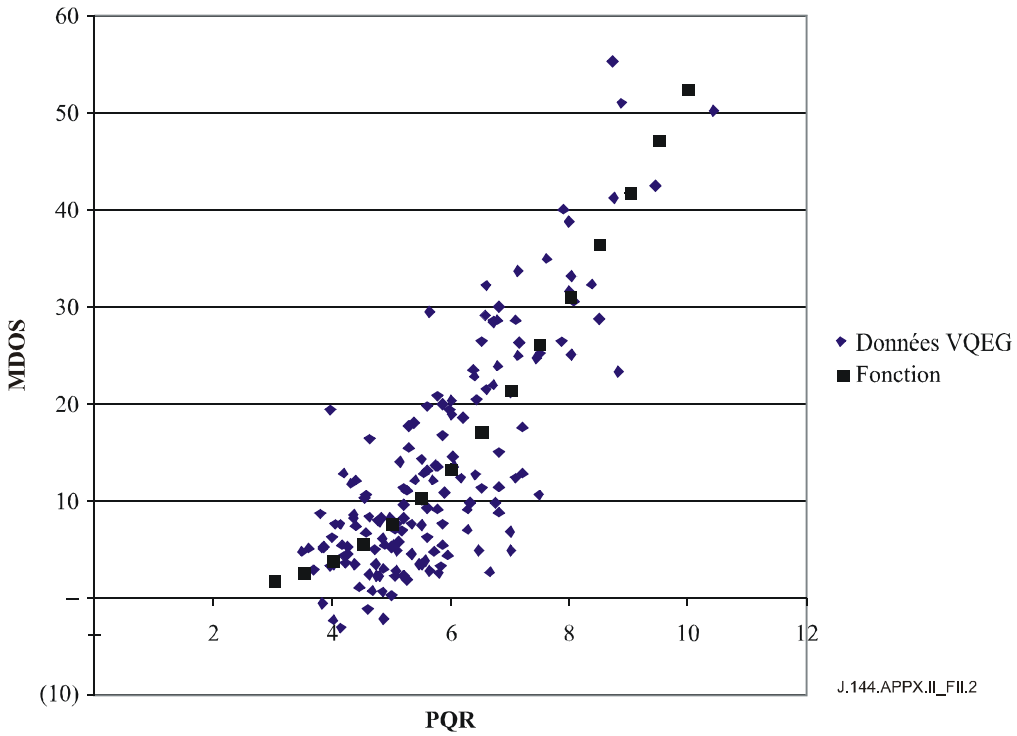


Figure II.2 – DonnéesVQEG et fonction logistique ajustée

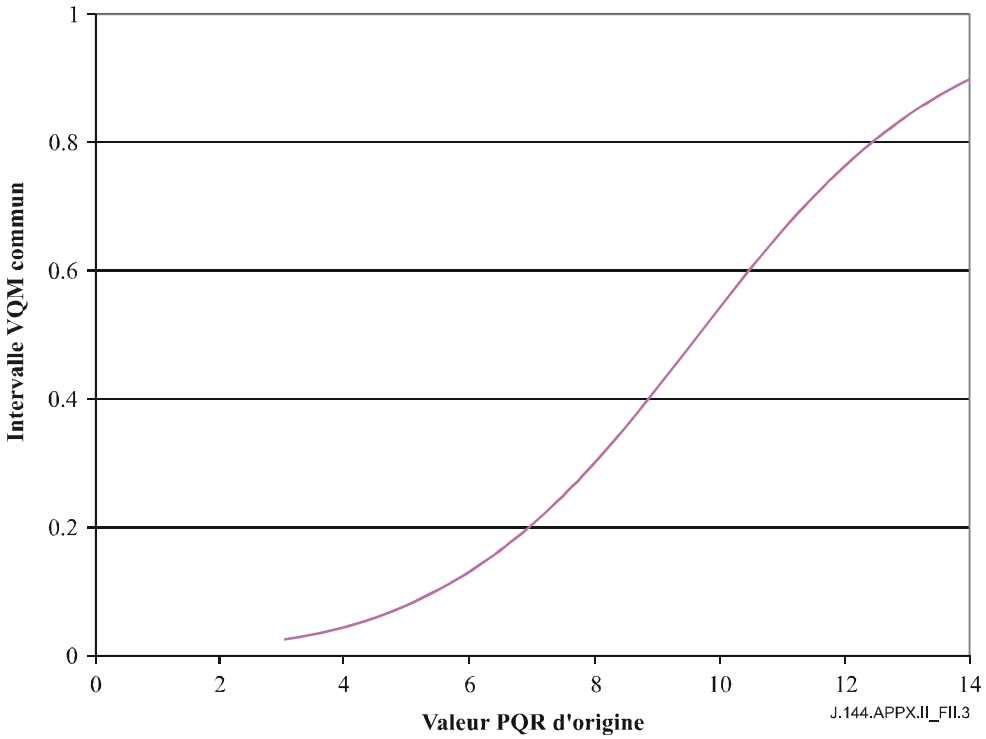


Figure II.3 – Valeurs PQR converties de l'intervalle d'origine à l'intervalle commun

La puissance de résolution désigne la différence de valeur des indices VQM obtenus, pour laquelle la mesure correspondant à la meilleure valeur VQM comporte également le meilleur indice subjectif pour un niveau de confiance déterminé. La Rec. UIT-T J.149 indique deux méthodes de calcul de la puissance de résolution d'une mesure VQM. Selon une méthode statistique sophistiquée connue sous le nom de test z on considère que la puissance de résolution n'est pas une fonction simple du niveau de confiance requis mais varie en fonction des points de données subjectives/objectives disponibles. La Figure II.4 indique les résultats du calcul effectué d'après la méthode PQR. Les valeurs delta-VQM sont indiquées dans le domaine normalisé, tandis que la puissance de résolution correspond à une valeur nominalement constante pour toutes les valeurs VQM.

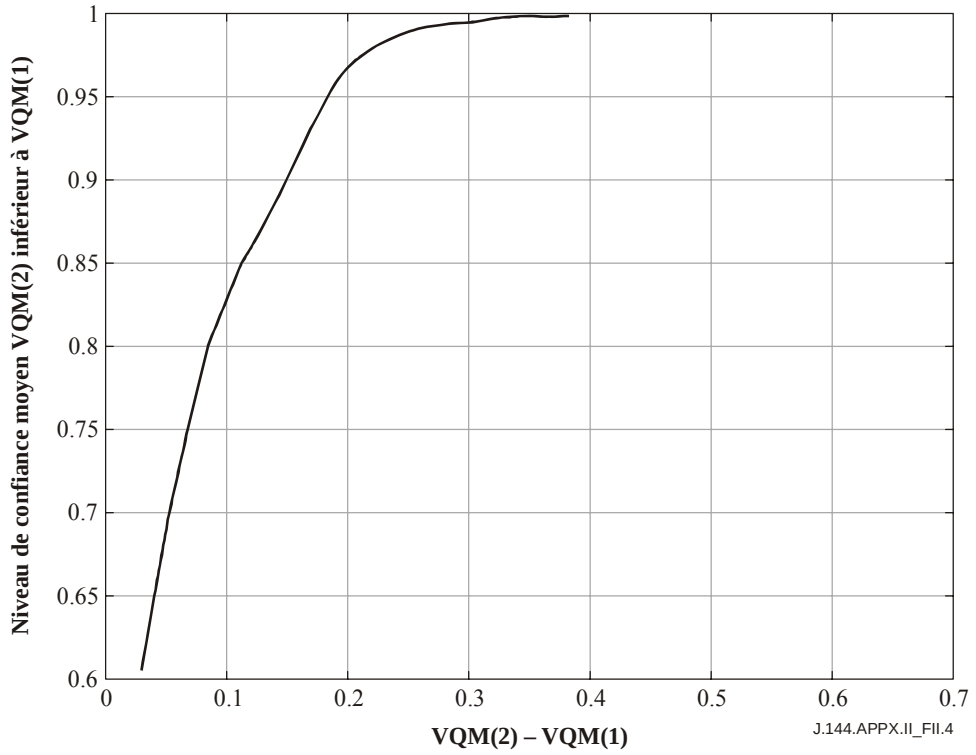


Figure II.4 – Puissance de résolution en fonction de l'intervalle normalisé VQM

D'après le diagramme, pour une différence de valeurs PQR normalisées égale à 0,1 il existe un seuil de confiance de 0,81 selon lequel la séquence dotée de la plus faible valeur PQR (qualité la plus élevée) aura le plus faible indice subjectif DSCQS (qualité la plus élevée). Le choix de la puissance de résolution à utiliser pour différentes applications est facilité par une analyse de la classification des erreurs tel qu'indiqué à l'Annexe C.

Pour effectuer un calcul en première approximation de la puissance de résolution il suffit de calculer l'erreur quadratique moyenne (RMSE, *root mean square error*) des indices subjectifs en fonction des valeurs objectives dans l'intervalle normalisé. Les différences de valeurs VQM égales à l'erreur quadratique moyenne correspondent à un seuil de confiance à 68% et celles égales à 1,96 fois l'erreur quadratique moyenne à un seuil de confiance à 95%. Bien que cette méthode donne un résultat différent de celui d'une approche plus complexe, elle est facile à comprendre et peut s'avérer très utile compte tenu des niveaux de précision observés dans les conditions réelles de fonctionnement.

$$VQM\_RMSE = 0{,}06723$$

Cette valeur correspond approximativement à la courbe plus précise représentée ci-dessous à la Figure II.4.

Seuil de confiance	Figure II.4	RMSE
68%	0,053	0,066
95%	0,187	0,132

La Figure II.4 peut être ramenée à l'échelle PQR si nécessaire, en modifiant l'échelle de l'axe des abscisses (c'est-à-dire VQM(2)–VQM(1)) au moyen de la dérivée de la fonction d'ajustement Logistic II, tel qu'indiqué au § 4.2/J.149.

$$[PQR(2)-PQR(1)]=[VQM(2)-VQM(1)]\frac{\left(1+e^{-c(PQR-d)}\right)^2}{c(b-a)e^{-c(PQR-d)}}$$

Après introduction des contraintes a = – e $^{-cd}$ et b = 0, la relation devient:

$$[PQR(2)-PQR(1)]=[VQM(2)-VQM(1)]\frac{\left(1+e^{c(d-PQR)}\right)^2}{c\left(e^{c(d-PQR)}+e^{-c(PQR)}\right)}$$

Où:

c = 0,5031

d = 9,634

On obtient ainsi une famille de courbes puisque le coefficient de modification de l'échelle des abscisses dépend de la valeur de PQR.

II.2 Références

Les normes suivantes contiennent des dispositions qui, par suite de la référence qui y est faite, constituent des dispositions valables pour le présent appendice. Au moment de la publication, les éditions indiquées étaient en vigueur. Toute Recommandation ou autre référence est sujette à révision; tous les utilisateurs du présent Appendice sont donc invités à rechercher la possibilité d'appliquer les éditions les plus récentes des Recommandations et autres références indiquées ci-après.

– Recommandation UIT-R BT.601-5 (1995), *Paramètres de codage en studio de la télévision numérique pour des formats standards d'image 4:3 (normalisé) et 16:9 (écran panoramique).*

– Recommandation UIT-T J.149 (2004), *Méthode de spécification de la précision et du contre-étalonnage des mesures de la qualité vidéo*.

– Recommandation UIT-T P.911 (Annexe B) (1998), *Méthodes d'évaluation subjective de la qualité audiovisuelle pour applications multimédias*.

II.3 Introduction

II.3.1 Normalisation

La normalisation signifie que les changements systématiques stationnaires de la vidéo entre l'entrée de référence et la sortie de la vidéo traitée sont éliminés avant d'effectuer la mesure fondée sur le système visuel humain (voir Figure II.1). La Méthode de l'indice PQR repose sur des filtres HVS qui comparent réellement pixel par pixel les images de référence et les images traitées. La méthode de normalisation spécifiée dans le document T1.TR.73-2001*, Annexe B (voir Appendice II.A) est adaptée à la méthode PQR.

Les paramètres à régler pour le processus de normalisation sont les suivants: déplacements d'image horizontaux et verticaux, modifications de gain de luminance et de couleur, modifications du niveau continu de luminance et de couleur et décalage temporel canal à canal entre composantes ou entre la luminance et la couleur. Les résultats de la méthode de mesure doivent faire état de ces changements, puisque ceux-ci sont susceptibles d'induire des variations de qualité de l'image perçue. Il est nécessaire de traiter séparément ces changements et le calcul de l'indice PQR, pour deux raisons: d'abord et surtout pour obtenir la valeur de l'indice PQR le plus précis possible, et ensuite parce qu'une telle normalisation correspond étroitement à une exploitation typique du

* Les normes T1 sont maintenues par l'ATIS depuis novembre 2003.

système pour ce qui est des paramètres de gain et de niveau continu, pour lesquels des réglages appropriés sont généralement possibles et effectués régulièrement. Si l'on considère en général que de petits déplacements d'image horizontaux ou verticaux n'altèrent pas la qualité de l'image perçue, il n'en demeure pas moins que leur présence constitue une erreur sur l'image et générera des problèmes importants pour les applications multi-générations. L'alignement temporel devant être parfait, chaque trame/image traitée est comparée à la vidéo de référence équivalente.

La vidéo traitée est normalisée trame par trame par comparaison avec la vidéo de référence ou par la mesure des signaux de test étalonnés que contient la séquence de référence. On élimine uniquement les changements statiques stationnaires de la vidéo, alors que les changements dynamiques dus aux processus de compression et de décompression sont mesurés dans le cadre du calcul de l'indice PQR. La normalisation de la vidéo traitée en préalable aux calculs de l'indice PQR, devra satisfaire aux critères de tolérance indiqués dans le Tableau II.1. La précision des valeurs de l'indice PQR après normalisation et non conformes aux critères de tolérance du Tableau II.1 sera inférieure à celle qui est spécifiée au § II.1.3.4.

Tableau II.1 – Paramètres et tolérance de normalisation

Paramètre	Tolérance de normalisation
Gain de luminance	< 0,2 dB du blanc de crête
Gain (de différence) de couleur	< 0,2 dB de l'excursion max. admise
Niveau continu de luminance	< 0,5 % du blanc de crête
Niveau continu (de différence) de couleur	< 0,5% de l'excursion max. admise
Décalage temporel canal à canal	< 2 ns
Déplacement horizontal de pixel	< 0,1 pixel
Déplacement vertical de la ligne	0 ligne (limité à des déplacements d'un nb entier de lignes)
Déplacement temporel	0 trames

II.3.2 Aperçu général de la méthode PQR

La méthode PQR permet de prévoir les indices de perception que les humains attribueront à une séquence d'images en couleur dégradée par comparaison avec la séquence équivalente non dégradée. Deux séquences d'images entrent dans le modèle, qui génère plusieurs estimations de différence, dont une mesure unique des différences perçues entre les séquences. On quantifie ces différences en unités de différences tout juste perceptibles (JND, *just-noticeable difference*).

Une séquence vidéo d'entrée est transmise dans deux canaux différents vers un observateur (qui n'apparaît pas sur la figure). Un canal est parfait (le canal de référence), tandis que l'autre (le canal testé) opère certaines distorsions de l'image. Il peut y avoir distorsion (effet secondaire de certaines mesures prises par souci d'économie) dans le codeur avant la transmission, dans le canal de transmission lui-même ou dans le processus de décodage. La case de la Figure II.8 intitulée "système sous test" correspond schématiquement à l'une quelconque de ces possibilités. L'application de la méthode PQR consiste à remplacer le dispositif de visualisation et l'observateur par le modèle de la vision humaine, qui substitue à une comparaison subjective, une comparaison des séquences de test et de référence permettant de générer une séquence de cartes JND.

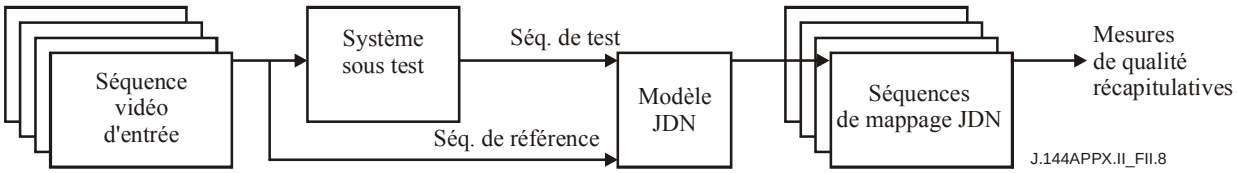


Figure II.8 – Modèle de la vision humaine pour l'évaluation du système[17]

La Figure II.9 présente un aperçu général de l'algorithme. Les entrées sont constituées de deux séquences d'images de longueur arbitraire. Pour chaque trame de chaque séquence d'entrée, il y a trois ensembles de données (étiquetés Y', C_b' et C_r' en haut de la Figure II.9), issus par exemple d'une bande D1. Les données Y, C_b et C_r sont ensuite transformées en tensions du canon électrique R', G' et B' qui donnent naissance aux valeurs de pixel affichées. Dans le modèle, des traitements ultérieurs transforment ces tensions en une image de luminance et deux images de chrominance, qui constituent les entrées des phases suivantes.

Le traitement d'entrée a pour but de transformer les signaux vidéo d'entrée en sorties optiques puis de transformer ces sorties optiques en quantités définies sur un plan psychophysique qui caractérisent séparément la luminance et la chrominance.

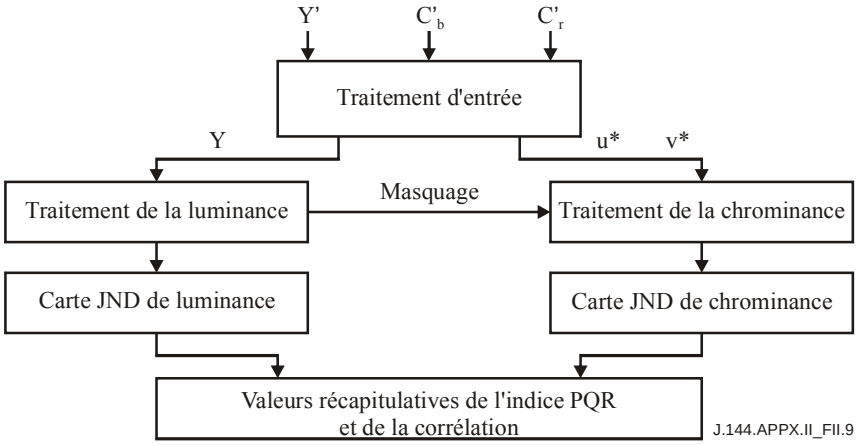


Figure II.9 – Ordinogramme du modèle de vision humaine

Deux images de luminance Y (une image de test et une image de référence), exprimées en fraction de la luminance maximale de l'affichage, constituent l'entrée du traitement de la luminance. On obtient en sortie une carte JND de luminance, dont les niveaux de gris sont proportionnels, pour un pixel donné, au nombre d'unités JND entre l'image de test et l'image de référence.

Chacune des images de chrominance u* et v* fait l'objet d'un traitement analogue, fondé sur l'espace de couleur uniforme CIE L*u*v*. Les sorties de ce traitement de u* et v* sont associées pour former la carte JND de chrominance. Les traitements de la chrominance et de la luminance dépendent tous deux des entrées provenant du canal de luminance (on parle de *masquage*), qui rendent les différences perçues plus ou moins visibles en fonction de la structure des images de luminance.

On dispose en sortie des cartes JND de luminance, de chrominance et de combinaison luminance-chrominance, ainsi qu'un nombre réduit de mesures récapitulatives issues de ces cartes.

[17] A dessein les Figures II.5 à II.7 ne sont pas utilisées.

Les valeurs récapitulatives uniques de l'indice PQR modélisent la façon dont un observateur estime globalement les distorsions affectant une séquence de test. Les cartes JND permettent une appréciation plus détaillée de l'emplacement et de l'intensité des artefacts.

II.4 Aperçu général de l'algorithme

II.4.1 Traitement d'entrée

La pile de quatre trames étiquetée Y', C_b' et C_r' qui se trouve en haut de la Figure II.10 représente un ensemble de quatre trames consécutives provenant d'une séquence d'images de test ou d'une séquence d'images de référence. La première phase du traitement transforme les données Y', C_b' et C_r' en tensions du canon électronique R', G' et B' (voir § II.5.1.1).

La seconde phase du traitement que subit chacune des images R', G' et B' se caractérise par sa non-linéarité. Cette phase modélise la transformation de R', G' et B', tensions du canon électronique, en intensités modèle (R, G, B) de l'affichage (fractions de la luminance maximale). La non-linéarité entraîne aussi un écrêtage aux luminances basses pour chaque plan de l'affichage.

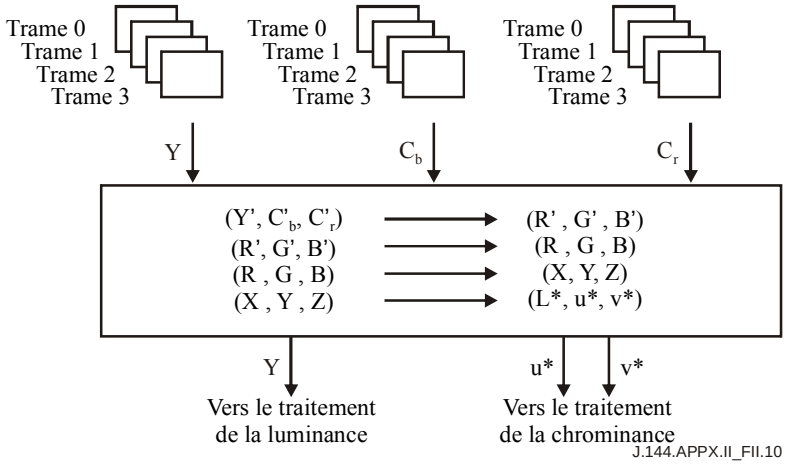


Figure II.10 – Aperçu général de la première phase du traitement

Suite à la non-linéarité on peut choisir l'une des deux options de traitement: mi-hauteur ou hauteur totale. Dans le cas de balayages avec entrelacement, les images mi-hauteur[18] sont traitées telles quelles, sans interligne vide. La modélisation hauteur totale est disponible pour les balayages progressifs (pour lesquels une trame contient une image, c'est-à-dire une image simple plutôt que deux trames entrelacées).

Ensuite, le vecteur (R,G,B) de chaque pixel de la trame fait l'objet d'une transformation linéaire (fonction de la phosphorescence de l'affichage) en coordonnées trichromatiques (X, Y, Z) CIE 1931. La composante de luminance Y de ce vecteur est dirigée vers le traitement de la luminance.

Afin de garantir (pour chaque pixel) une uniformité approximative de perception de l'espace de couleur, quant aux différences de couleur isoluminante, on mappe chaque pixel dans l'espace CIELUV, un espace de couleur uniforme répondant à une norme internationale. Les composantes

[18] Les lignes d'une image mi-hauteur correspondent à une trame, c'est-à-dire aux lignes paires ou impaires d'une image.

de chrominance u* et v* de cet espace parviennent ensuite aux phases de traitement de la chrominance du modèle[19].

II.4.2 Traitement de la luminance

Comme l'indique la Figure II.11, chaque valeur de luminance fait d'abord l'objet d'une compression non linéaire. Chaque trame de luminance est ensuite filtrée et sous-échantillonnée dans une pyramide de Gauss à quatre niveaux, afin de modéliser la décomposition que l'on observe psychophysiquement et physiologiquement pour les signaux visuels entrants dans les différentes bandes spatio-fréquentielles. Des traitements similaires (par exemple filtrage orienté) effectués à chaque niveau de la pyramide constituent l'essentiel des traitements ultérieurs.

A la suite à ce processus pyramidal, l'image de basse résolution de la pyramide fait l'objet d'un filtrage temporel et d'un calcul de contraste, alors que les images issues des trois autres niveaux subissent un filtrage spatial et un calcul de contraste. Dans tous les cas, le contraste se définit comme la différence locale des valeurs des pixels divisée par une somme locale, établie de manière judicieuse. Cela constitue au départ la définition d'une unité JND, qui est ensuite transmise aux phases ultérieures du modèle[20]. (L'étalonnage révise de manière itérative l'interprétation d'une unité JND aux phases intermédiaires du modèle.) La valeur absolue de la réponse de contraste constitue l'entrée de la phase suivante, et l'on conserve le signe algébrique que l'on rajoute juste avant la comparaison des images (calcul de la carte JND).

La phase suivante (masquage de contraste) est une opération de réglage de gain pour laquelle chaque réponse de contraste est divisée par une fonction dépendante de toutes les réponses de contraste. Cette atténuation combinée de chaque réponse par les autres réponses locales permet de modéliser les effets de "masquage" visuel tels que la baisse de sensibilité aux distorsions dans les zones "actives" de l'image. A cette étape du modèle, une structure temporelle (scintillation) est établie pour masquer les différences spatiales, et une structure spatiale permet en outre de masquer les différences temporelles. Comme on le verra ultérieurement, le masquage de luminance est aussi utilisé pour le traitement de la chrominance.

On utilise les réponses de contraste masqué (ainsi que les signes de contraste) pour générer une carte JND de luminance. Pour cela, on procède comme suit:

- séparation de chaque image en ses composantes positives et négatives (redressement à une alternance);
- sommation locale (mise en moyenne et sous-échantillonnage, pour modéliser la sommation spatiale locale que l'on observe dans les expériences psychophysiques);
- évaluation de la différence absolue entre les images, canal par canal;
- suréchantillonnage à la même résolution (qui sera moitié moindre que celle de l'image d'origine, du fait de la phase de sommation);
- évaluation sur tous les canaux de la norme Q de Minkowski.

---

[19] Le canal de luminance L* de l'espace CIELUV n'est pas utilisé dans le traitement de la luminance, mais est remplacé par une non-linéarité visuelle pour laquelle le modèle visuel a été étalonné pour une gamme de valeurs de luminance. L* est toutefois utilisé dans le traitement de la chrominance pour créer un modèle de mesure de la chrominance à peu près uniforme et familier aux ingénieurs s'occupant de la visualisation.

[20] Associer un contraste constant à 1 JND constitue une implémentation connue sous le nom de loi de Weber pour la vision.

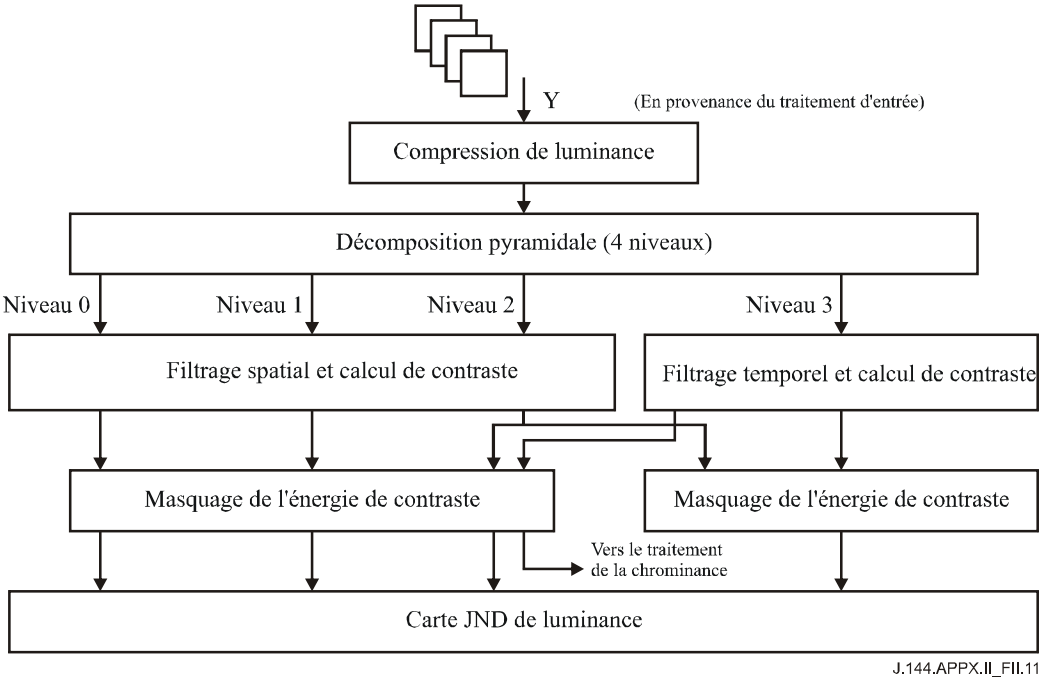


Figure II.11 – Aperçu général du traitement de la luminance

II.4.3 Traitement de la chrominance

Le traitement de la chrominance suit en grande partie le traitement de la luminance. On utilise les différences de chrominance intra-image (u* et v*) de l'espace CIELUV pour définir les seuils de détection du modèle de chrominance, tout comme le contraste (et la loi de Weber) servent à déterminer le seuil de détection du modèle de luminance. De plus, par analogie avec le modèle de luminance, les "contrastes" chromatiques définis par les différences u* et v* font l'objet d'une étape de masquage. Un transducteur non linéaire rend la discrimination d'un incrément de contraste entre deux images dépendantes de la réponse de contraste commune aux deux images.

La Figure II.12 montre qu'à l'instar du traitement de la luminance, chaque composante de chrominance u* et v* fait l'objet d'une décomposition pyramidale. Le traitement de la chrominance se compose toutefois de sept niveaux, alors que le traitement de la luminance en nécessite seulement quatre. On tient ainsi empiriquement compte du fait que les canaux de chrominance sont sensibles à des fréquences spatiales bien plus basses que les canaux de luminance et, qu'intuitivement, on peut observer des différences de couleur dans de grandes régions uniformes.

Un traitement temporel de moyennage des quatre trames d'images est effectué, qui traduit le fait que les canaux de chrominance sont intrinsèquement insensibles au scintillement.

Un noyau de Laplace filtre ensuite spatialement u* et v*. On génère ainsi une différence de couleur pour u* et v*, ce qui (par définition de l'espace de couleur uniforme) est lié, sur le plan de la mesure, aux différences de couleur tout juste perceptibles. On suppose, dans cette phase, qu'une valeur "1" signifie qu'une seule unité JND a été obtenue, par analogie au rôle joué dans le canal de luminance par le contraste fondé sur la loi de Weber (de même que pour la luminance, l'unité de chrominance 1-JND doit être réinterprétée durant l'étalonnage).

La valeur absolue de la différence de couleur pondérée est transmise (avec le signe algébrique de contraste) à la phase de masquage de contraste, qui réalise la même fonction que pour le modèle de luminance. Son fonctionnement est un peu plus simple, puisqu'elle ne reçoit en entrée que les canaux de luminance et le canal de chrominance dont la différence est évaluée. Enfin, le traitement des réponses de contraste masqué est en tout point identique à celui effectué pour le modèle de luminance (voir dernier alinéa du II.4.2).

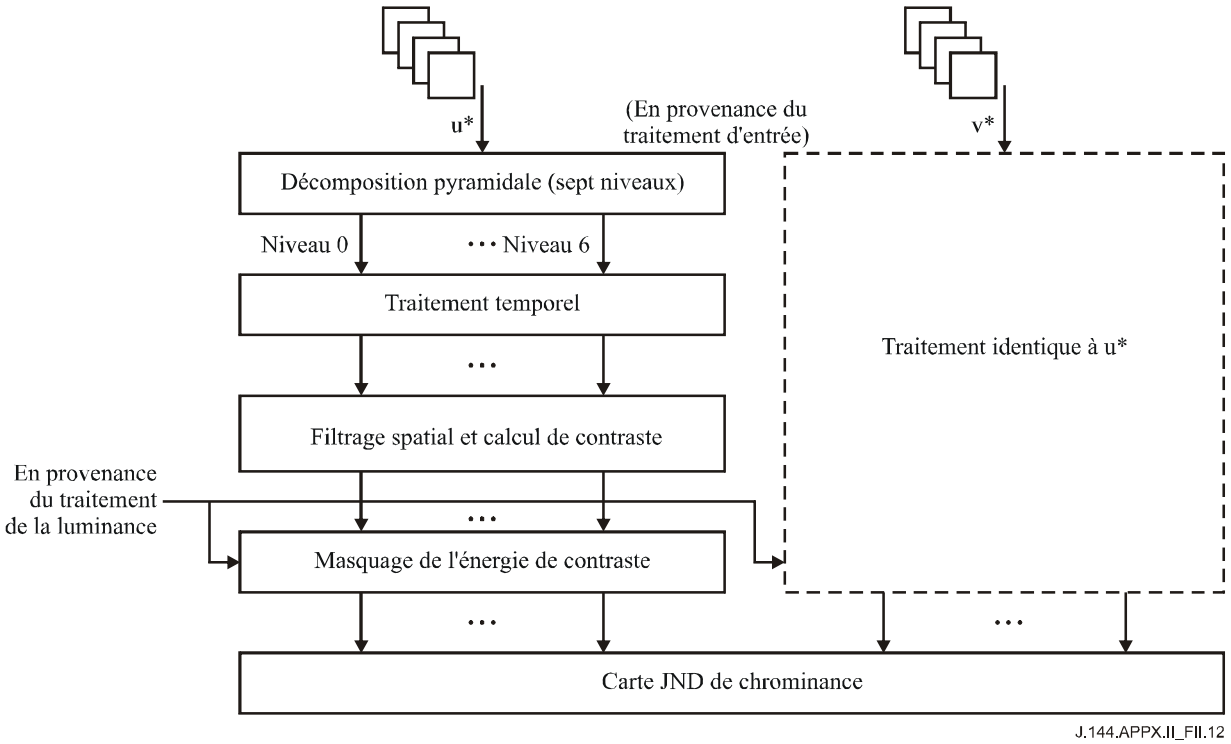


Figure II.12 – Aperçu général du traitement de la chrominance

II.4.4 Valeurs récapitulatives de sortie

Pour chaque trame de la comparaison des séquences vidéo, on combine d'abord les cartes JND de luminance et de chrominance pour former une carte JND globale. Pour calculer cette carte, on combine linéairement la somme et le maximum des valeurs des cartes de luminance et de chrominance pixel par pixel.

Chacune de ces trois cartes JND (de luminance, de chrominance et la combinaison luminance-chrominance) est ensuite réduite à un nombre unique récapitulatif, appelé valeur d'indice de qualité de l'image (PQR, *picture quality rating*), calculé à l'aide de la norme de Minkowski. Pour ce faire, on élève à la puissance Q chaque valeur de pixel d'une carte JND. L'indice PQR se calcule alors comme la racine $Q^{ème}$ de la somme normalisée de toutes ces valeurs de pixel à la puissance Q.

On calcule ensuite trois mesures de qualité uniques pour un grand nombre de trames d'une séquence vidéo (une pour la luminance, une pour la chrominance et une dernière pour la combinaison luminance-chrominance). Les valeurs d'indice PQR pour chaque trame de la séquence sont ensuite réduites à un indice PQR relatif à la séquence entière, en utilisant à nouveau une norme Q de Minkowski.

Bien que la méthode PQR soit valide pour différentes longueurs de séquences, aux fins du présent appendice, les valeurs PQR sont calculées pour 60 trames (deux secondes) vidéo. Il importe de signaler que les valeurs PQR obtenues pour des séquences d'une demie seconde au moins ne sont pas aisément comparables avec des évaluations subjectives. Cela tient à l'absence de fiabilité des évaluations subjectives pour des séquences aussi brèves.

II.5 Description détaillée de l'algorithme

L'application du modèle de mesure PQR décrit dans le présent appendice exige l'utilisation de valeurs paramétriques permettant d'étalonner l'algorithme afin de reproduire de façon approchée la réponse du système de vision humaine. Le Tableau II.2 indique les valeurs paramétriques à utiliser afin d'implémenter le modèle décrit dans le présent appendice.

Tableau II.2 – Valeurs paramétriques du modèle PQR

Type de paramètre	Symbole	Valeur	Paragraphe
Compression de luminance	m L_d	0.65 7,5 cd/m^2	II.5.2.1
Filtrage temporel 60 Hz	t_e t_l	33/64 31/64	II.5.2.3.2
Filtrage temporel 50 Hz	t_e t_l	11/16 5/16	II.5.2.3.2
Seuil de contraste de luminance (par niveau pyramidal)	w_0 w_1 w_2 w_3	1/150 1/900 1/1280 1/500	II.5.2.4
Contrastes de masquage de luminance	β a c m_f m_t m_{ft}	1.4 3/32 5/32 10/1024 50 3/64	II.5.2.5
Seuil de contraste de chrominance (par niveau pyramidal)	q_0 q_1 q_2 q_3 q_4 q_5 q_6	384 60 24 6 4 3 3	II.5.3.4
Constantes de masquage de chrominance	β_c a_c c_c m_c k	1,4 0,5 0,5 10/1024 0,7	II.5.3.5

II.5.1 Traitement d'entrée SD

Voir Figure II.10. Le traitement d'entrée transforme les signaux vidéo d'entrée Y′, C'_b et C'_r en tensions du canon électrique, puis en valeurs de luminance de trois phosphorescences et enfin, en variables psychophysiques que l'on dissocie en composantes de luminance et de chrominance. La valeur trichromatique Y, calculée au § II.5.1.3, se substitue à la "valeur d'intensité du modèle", utilisée avant incorporation du traitement de chrominance au modèle JND. De plus, les composantes de chrominance u^* et v^* sont créées pixel par pixel, conformément aux spécifications CIE d'uniformité de couleur.

II.5.1.1 (Y′, C'_b, C'_r) transformé en (R′, G′, B′)

Les étapes décrites ci-dessous correspondent à la transformation des trames images Y′, C'_b et C'_r en signaux de tension R′, G′, B′, appliqués à l'écran. En l'occurrence le caractère prime indique que les signaux d'entrée ont été précorrigés en gamma au niveau du codeur. Ces signaux après une

transformation supplémentaire, sont appliqués à un affichage cathodique[21], dont la fonction de transfert tension-intensité peut être à toutes fins pratiques, assimilée à une non-linéarité gamma.

On suppose en l'occurrence que les images numériques d'entrée présentent un format 4.2.2: la résolution complète sur le signal de luminance est corrélé avec Y′ et la demi-résolution horizontale du signal de chrominance est corrélée avec C'_b et C'_r. Les données Y′, C'_b et C'_r sont censées être enregistrées dans l'ordre spécifié par la Recommandation UIT-R BT.601-5 à savoir:

$$C'_{b0}, Y'_0, C'_{r0}, Y'_1, C'_{b1}, Y'_2, C'_{r1}, Y'_3, ..., C'_{bn/2-1}, Y'_{n-1}, C'_{rn/2-1}, Y'_{n-2}, \ldots .$$

Etape 1: introduire les tableaux Y′, C'_b, C'_r d'une trame unique. Développer ensuite les tableaux C'_b et C'_r jusqu'à obtention d'une image Y′ avec une résolution complète. Les tableaux C'_b et C'_r comportent initialement une demi-résolution horizontale, et doivent ensuite faire l'objet d'un suréchantillonnage afin de créer des trames à résolution entière. Dans un premier temps, les autres pixels C'_b, C'_r d'une même ligne sont affectés à l'élément Y'_i d'indice pair qu'ils encadrent dans le flux de données. Ensuite, la paire C'_b, C'_r a associé à l'élément Y'_i d'indice impair est calculée par mise en moyenne des deux élément horizontaux immédiatement voisins.

Etape 2: répartir les tableaux Y′, C'_b, C'_r à résolution entière en deux trames. Dans le cas de Y′, la première trame contient les lignes impaires du tableau Y′, et la deuxième les lignes paires de ce même tableau. Les tableaux C'_b et C'_r font l'objet d'un traitement identique pour obtenir les premières et deuxièmes trames C'_b et C'_r.

Etape 3: pour chaque pixel de chacune des deux trames, convertir les valeurs Y', C'_b, C'_r correspondantes pour obtenir les valeurs de tension d'entrée de canon électrique R', G', B'. Aux fins du présent appendice, les valeurs Y′, C'_b, C'_r retenues sont liées aux valeurs R′,G′,B′ par la relation suivante:

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1{,}371 \\ 1 & -0{,}336 & -0{,}698 \\ 1 & 1{,}732 & 0 \end{bmatrix} \left[\begin{Bmatrix} Y' \\ C'_b \\ C'_r \end{Bmatrix} \begin{Bmatrix} 0 \\ 128 \\ 128 \end{Bmatrix} \right] \qquad \text{(II-1)}$$

Les tableaux R′, G′ et B′ peuvent ensuite faire l'objet de l'étape suivante de l'algorithme de traitement d'entrée.

II.5.1.2 (R′, G′, B′) transformé en (R, G, B)

II.5.1.2.1 Transformation de la valeur des pixels

Calculer pour chaque pixel la fraction de la luminance maximale R correspondant à la fraction d'entrée R′. De manière analogue, calculer les luminances fractionnelles G et B à partir des tensions d'entrée G′ et B′. La luminance maximale de chaque tension d'entrée de canon électrique est supposée correspondre à la valeur d'entrée égale à 255. les relations suivantes correspondent à la transformation de (R', G', B') en (R, G, B):

$$R = \left[\frac{\max(R', t_d)}{255} \right]^{\gamma}$$

$$G = \left[\frac{\max(G', t_d)}{255} \right]^{\gamma} \qquad \text{(II-2)}$$

$$B = \left[\frac{\max(B', t_d)}{255} \right]^{\gamma}$$

[21] Voir § II.5.1.2.1 pour une description d'un modèle d'affichage cathodique.

Dans ce cas, la valeur seuil par défaut t_d est égale à 16, ce qui correspond au niveau de noir de l'écran et aux valeurs par défaut de gamma égale à 2,5. Le choix d'une valeur par défaut de t_d égale à 16 doit permettre d'obtenir pour l'écran une plage dynamique d'environ 1000:1 (c'est-à-dire $255/16)^{2,5}$).

II.5.1.2.2 Options de traitement d'image d'une hauteur entière et d'une de mi-hauteur

Le modèle PQR comporte deux options de spécifications de la représentation verticale des images (R, G, B) pour chaque trame (images progressives) et pour les trames impairs et pairs (images entrelacées).

1) *Trame*

   Images de hauteur entière contenant une image à balayage progressif.

2) *Entrelacement de mi-hauteur*

   Les images mi-hauteur sont directement transformées.

Les six premières sous-paragraphes des § II.5.2 et II.5.3 décrivent le traitement de luminance et de chrominance d'images de hauteur entière. Les sous-paragraphes II.5.2.7 et II.5.3.7 décrivent les traitements mi-hauteur.

II.5.1.3 (R, G, B) transformé en (X, Y, Z)

Calculer les coordonnées trichromatiques X, Y, Z CIE 1931 relatives à chaque pixel, compte tenu des luminances fractionnaires R, G, B. Ce calcul implique les données d'entrée suivantes en fonction du dispositif d'affichage: les coordonnées de chromatisme (x_r, y_r), (x_g, y_g), (x_b, y_b) des trois phosphorescences et la chromaticité du point blanc du moniteur (x_w, y_w).

Les chromaticités du point blanc (x_w, $y_{w)}$, = (0,3127,0,3290) correspondent aux valeurs dites d'arrière-plan D65. Le Tableau II.3 présente les différentes options de coordonnées de phosphorescence de l'affichage.

Tableau II.3 – Options de coordonnées de phosphorescence d'affichage

Source	(x_r,y_r)	(x_g,y_g)	(x_b,y_b)
ITU-R BT.709-5 (SMPTE 274M)	(0,640, 0,330)	(0,300, 0,600)	(0,150, 0,060)
SMPTE 240M	(0,630, 0,340)	(0,310, 0,595)	(0,155. 0,070)
EBU	(0,640, 0,330)	(0,290, 0,600)	(0,150, 0,060)

Compte tenu des valeurs paramétriques ci-dessus, les valeurs X, Y, Z du pixel sont obtenues à l'aide des relations suivantes:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} \frac{x_r}{y_r} Y_{or} & \frac{x_g}{y_g} Y_{og} & \frac{x_b}{y_b} Y_{ob} \\ Y_{or} & Y_{og} & Y_{ob} \\ \frac{z_r}{y_r} Y_{or} & \frac{z_g}{y_g} Y_{og} & \frac{z_b}{y_b} Y_{ob} \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \qquad \text{(II-3)}$$

Ici, $z_r = (1–x_r–y_r)$, $z_g = (1–x_g–y_g)$, $z_b = (1–x_b–y_b)$, et valeurs Y_{0r}, Y_{0g}, Y_{0b} données par

$$\begin{bmatrix} Y_{0r} \\ Y_{0g} \\ Y_{0b} \end{bmatrix} = \begin{bmatrix} \frac{x_r}{y_r} & \frac{x_g}{y_g} & \frac{x_b}{y_b} \\ 1 & 1 & 1 \\ \frac{z_r}{y_r} & \frac{z_g}{y_g} & \frac{z_b}{y_b} \end{bmatrix}^{-1} \begin{bmatrix} \frac{x_w}{y_w} \\ 1 \\ \frac{z_w}{Y_w} \end{bmatrix} \qquad \text{(II-4)}$$

avec $z_w = (1–x_w–y_w)$.

Les coordonnées trichromatiques X_n, Y_n et Z_n du point blanc des dispositifs seront également nécessaires. Elles correspondent à la chromaticité (x_w, y_w) et sont telles que pour une pleine activation de la phosphorescence (R′ = G′ = B′ = 255), Y = 1. Les valeurs trichromatiques correspondant au point blanc sont (X_n, Y_n, Z_n) = (x_w/y_w, 1, z_w/y_w).

Au stade de la dernière étape du calcul des valeurs X, Y, Z, il faut introduire une correction pour tenir compte d'une lumière ambiante créée par les rayons réfléchis par l'écran d'affichage. Cette correction se présente sous la forme

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \leftarrow \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \left(\frac{L_a}{L_{\max}} \right) \begin{bmatrix} X_n \\ Y_n \\ Z_n \end{bmatrix} \qquad \text{(II-5)}$$

Dans ce cas, deux paramètres spécifiés par l'utilisateur L_{max} et L_a sont introduits et mis à des valeurs par défaut. L_{max}, la luminance maximale de l'affichage est mis à la valeur 100 cd/m^2 ce qui correspond aux affichages fournis dans le commerce. La luminance L_a de lumière ambiante due aux rayons réfléchis est mise à la valeur 5 cd/m^2, ce qui correspond aux mesures effectuées sur les écrans dans les conditions indiquées par la Rec. UIT-R BT 500-11.

La chromaticité de la lumière ambiante est supposée identique à celle du point blanc de l'affichage. Il convient de signaler que suivant l'option de la modélisation limitée à la luminance, qui ne calcule pas le point neutre (X_n, Y_n,Z_n), la correction

$$Y \leftarrow Y + \frac{L_a}{L_{\max}} \qquad \text{(II-6)}$$

se substitue à l'équation II-5. Cette valeur est équivalente à la composante Y de l'équation II-5 parce que Y_n est toujours égal à 1. Il est à signaler que la quantité L_{max} *Y désigne la luminance de l'affichage exprimée en cd/m^2.

II.5.1.4 (X, Y, Z) transformé en (L*, u*, v*)

Transformation des valeurs X, Y, Z, pixel par pixel, conformément au système de couleur uniforme 1976 CIELUV:

$$L^* = 116\left(\frac{Y}{Y_n}\right)^{1/3} - 16 \quad pour \ \frac{Y}{Y_n} > 0{,}008856$$

$$L^* = 903{,}3\left(\frac{Y}{Y_n}\right) \quad pour \ \frac{Y}{Y_n} \le 0{,}008856 \qquad \text{(II-7)}$$

$$u^* = 13L^*(u' - u'_n) \qquad \text{(II-8)}$$

$$v^* = 13L^*(v' - v'_n) \tag{II-9}$$

avec,

$$u' = \frac{4X}{(X + 15Y + 3Z)} \tag{II-10}$$

$$v' = \frac{9Y}{(X + 15Y + 3Z)} \tag{II-11}$$

$$u'_n = \frac{4X_n}{(X_n + 15Y_n + 3Z_n)} \tag{II-12}$$

$$v'_n = \frac{9Y_n}{(X_n + 15Y_n + 3Z_n)} \tag{II-13}$$

Il est à signaler que la coordonnée L* n'intervient pas dans le calcul de la luminance. L* sert uniquement à calculer les coordonnées de chrominance u* et v*[22] Par conséquent, parmi les grandeurs ci-dessus, seules les images u* et v* sont conservées en vue de leur traitement ultérieur.

II.5.2 Traitement de la luminance

Voir Figure II.13. Dans le présent paragraphe, les images de trame de référence et les images de trame test d'entrée sont désignées I_k et I^{ref}_k (k = 0, 1, 2, 3). Les valeurs de pixel dans I_k et I^{ref}_k sont désignées respectivement à l'aide des notations $I_k(i,j)$ et $I^{ref}_k(i,j)$. Elles sont initialement égales aux valeurs trichromatiques Y, calculées au stade du traitement d'entrée. Seules les trames I_k sont examinées ci-après. Le traitement I^{ref}_k est identique. K=3 désigne l'image la plus récente d'une séquence de quatre images.

Les § II.5.2.1 à II.5.2.6 décrivent le traitement hauteur totale. Le § II.5.2.7 traite des modifications requises pour le traitement mi-hauteur.

II.5.2.1 Compression de luminance

La première étape du modèle de luminance est une non-linéarité constituée par une fonction de puissance de ralentissement, décalée par une constante. Supposons de luminance relative de la dernière trame $Y_3(i,j)$, avec 3 indice de la dernière trame. On alors

$$I_3(i,j) = [L_{\max} Y_3(i,j)]^n + L_d^m \tag{II-14}$$

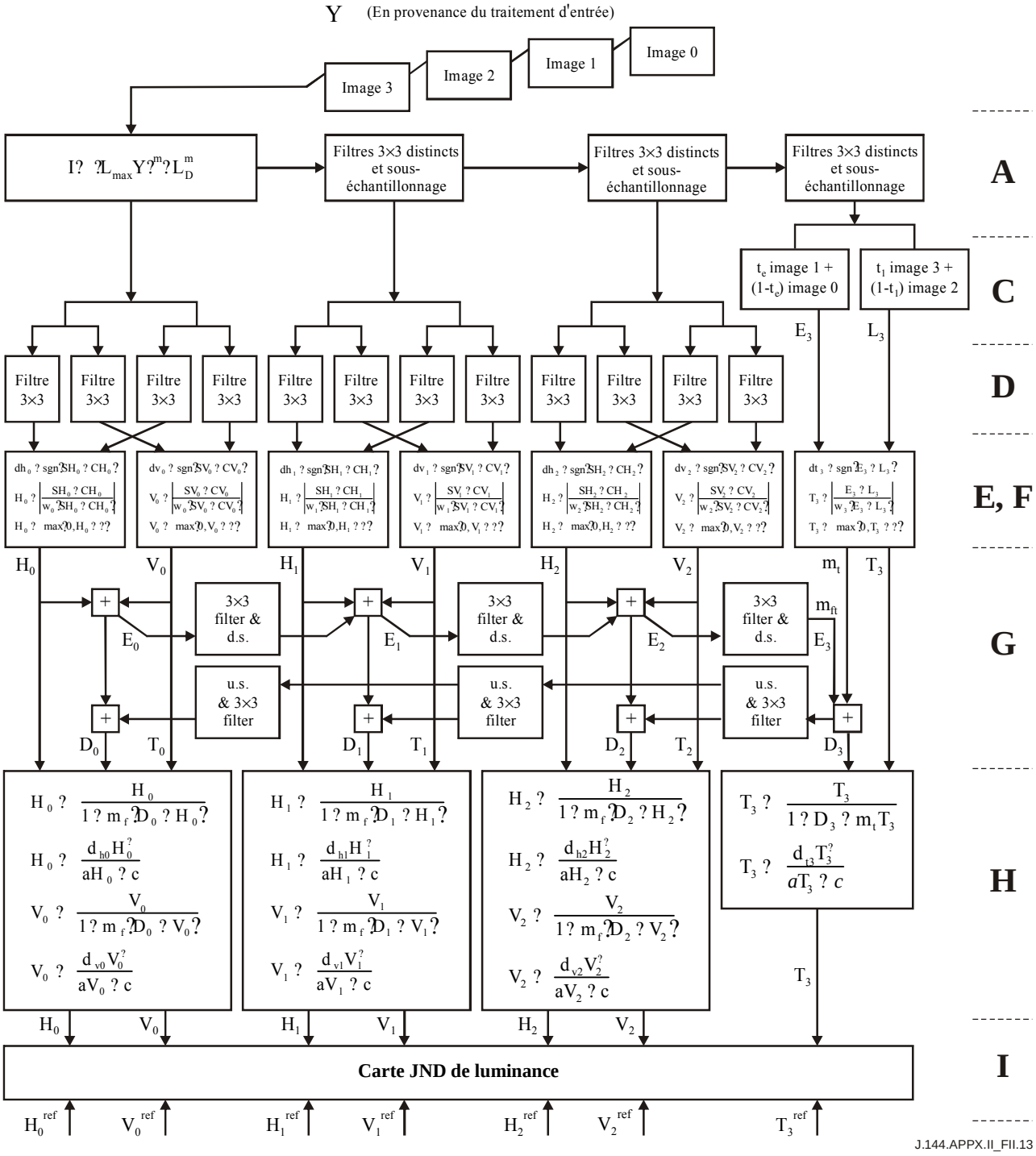
Où L_{max}, luminance maximale de l'affichage mise à la valeur 100 cd/m^2. Les valeurs L_d et m sont choisies de façon à correspondre aux données de détection du contraste à des valeurs de luminance comprises entre 0,01 et 100 ft-L.

II.5.2.2 Décomposition pyramidale de la luminance

La décomposition spatiale à quatre niveaux de résolution s'effectue par une méthode de calcul efficace connue sous le nom de transformation pyramidale, qui procède à un sous-échantillonnage et étale l'image dans un rapport 2 à chaque niveau successif de résolution de plus en plus grossière.

---

[22] Le canal de luminance L* de l'espace CIELUV n'est pas utilisé dans le traitement de la luminance, mais est remplacé par une non-linéarité visuelle pour laquelle le modèle visuel a été étalonné pour une gamme de valeurs de luminance. L* est toutefois utilisé dans le traitement de la chrominance pour créer un modèle de mesure de la chrominance à peu près uniforme et familier aux ingénieurs s'occupant de la visualisation.



3×3 filter & d.s.: Filtres 3×3 distincts et sous-échantillonnage
u.s. & 3×3 filter: Suréchantillonnage filtre de 3×3

Figure II.13 – Traitement détaillé de la luminance

L'image d'origine résolution entière est appelée niveau 0 de la pyramide, $G_0 = I_3(i,j)$. Les niveaux suivants, aux valeurs faibles de la résolution, sont obtenus par une opération intitulée REDUCE, qui fonctionne comme suit: un filtre passe-bas à 3 voies avec les coefficients de pondération (1, 2,1)/4 est appliqué à G_0, de façon séquentielle afin de produire une image floue. L'image résultante est ensuite sous-échantillonnée suivant un coefficient 2 (un pixel sur 2 est supprimé: voir pixels ombrés de la Figure II.14 ci-dessous) de façon à parvenir au niveau suivant G_1.

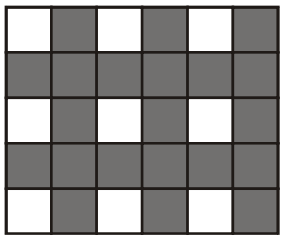


Figure II.14 – Sous-échantillonnage de l'image avec suppression des pixels gris

Si on appelle fds1() l'opération de filtrage et de sous-échantillonnage par un niveau pyramidal, il est possible de représenter par la formule suivante le processus REDUCE.

$$G_{i+1} = fds1(G_i),\ pour\ i = 1,\ 2,\ 3. \tag{II-15}$$

Le processus REDUCE est appliqué de façon récursive à chaque nouveau niveau (tel qu'indiqué dans Burt et Adelson, 1983).

A l'inverse, une opération dite EXPAND consiste à réaliser un suréchantillonnage et un filtrage à l'aide du même noyau 3×3. Cette opération est notée usf1(), et intervient dans le contexte des § II.5.2.5 et II.5.2.6.

Les noyaux de filtrage fds1 et usf1 dans chaque direction (horizontale et verticale) sont notés k_d[1,2,1] et k_u[1,2,1) respectivement avec k_d et k_u constantes choisies de façon à conserver l'uniformité de certaines trames. Pour fds1, les valeurs de la constante k_d est égale à 0,25 et pour ufs1 la constante est k_u=0,5 (en raison des zéros présents dans l'image suréchantillonnée). L'exécution de usfl en tant qu'opération de substitution, consiste à remplacer le noyau par les valeurs équivalentes obtenues par interpolation linéaire, afin de remplacer les valeurs "0". Toutefois, à des fins de simplicité, nous continuerons à appeler l'opération "filtrage suréchantillonneur".

II.5.2.3 Filtrage spatial et temporel de la luminance

Des filtres spatiaux orientés (centre et ambiance) sont appliqués aux images de niveau 0, 1 et 2 à la trame 3. Aux plus faibles niveaux de résolution (niveau 3), la première et la dernière paires de trames sont combinées de façon linéaire en images anticipées et tardives respectivement.

II.5.2.3.1 Filtrage spatial

Les filtres centraux et d'ambiance sont des filtres de 3×3 dissociables et permettent de combiner toutes les orientations: central vertical (CV), central horizontal (CH), d'ambiance vertical (SV, *surround vertical*) et d'ambiance horizontal (SH, *surround horizontal*). Les noyaux de filtrage sont définis comme suit:

$$\mathrm{CH} = \begin{bmatrix} 000 \\ 242 \\ 000 \end{bmatrix};\quad \mathrm{SH} = \begin{bmatrix} 121 \\ 000 \\ 121 \end{bmatrix};\quad \mathrm{CV} = \begin{bmatrix} 020 \\ 040 \\ 020 \end{bmatrix};\quad \mathrm{SV} = \begin{bmatrix} 101 \\ 202 \\ 101 \end{bmatrix} \tag{II-16}$$

II.5.2.3.2 Filtrage temporel

Les images anticipées et tardives de niveau 3 sont données respectivement par els relations

$$E_3 = t_e I_{3,1}(i, j) + (1 - t_e) I_{3,0}(i, j) \tag{II-17}$$

$$L_3 = t_l I_{3,3}(i, j) + (1 - t_l) I_{3,2}(i, j) \tag{II-18}$$

Les constantes t_e et t_l n'ont pas les mêmes valeurs à 50 Hz et à 60 Hz.

II.5.2.4 Calcul du contraste de luminance

Les données d'entrée sont constituée par les images de centrales et d'ambiance CV_i, CH_i, SV_i et SH_i (i=0, 1 et 2 pour les niveaux pyramidaux 0, 1 et 2) ainsi que les images anticipées et tardives E_3 et L_3(niveau 3) calculées au § 5.2.3. Le calcul du rapport de contraste utilise des formules analogues à celles du calcul du contraste Michelson $(L_{max} - L_{min})/(L_{max} + L_{min})$, qui ont permis d'obtenir une bonne modélisation de la vision. Pour les orientations dans le sens horizontal et vertical, les contrastes correspondants, pixel par pixel sont obtenus par les relations

$$\frac{(SH_i - CH_i)}{w_i(CH_i + SH_i)} \; et \; \frac{(SV_i - CV_i)}{w_i(CV_i + SV_i)} \tag{II-19}$$

De manière analogue, le rapport de contraste concernant la composante temporelle est donné par

$$\frac{(E_3 - L_3)}{w_3(E_3 + L_3)} \tag{II-20}$$

Les valeurs de w_i^{-1} pour i = 0,1,2,3 ont été obtenues par étalonnage.

Les images de la réponse de contraste sont calculées en tant que versions écrêtées des valeurs absolues des quantités définies par les deux relations précédentes. Ces quantités sont calculées comme suit

$$H_i = \max\left(0, \left|\frac{(SH_i - CH_i)}{w_i(CH_i + SH_i)}\right| - \varepsilon\right), \; V_i = \max\left(0, \left|\frac{(SV_i - CV_i)}{w_i(CV_i + SV_i)}\right| - \varepsilon\right) \tag{II-21}$$

i = 0,1,2, et

$$T_3 = \max\left(0, \left|\frac{(E_3 - L_3}{w_3(E_3 + L_3)}\right| - \varepsilon\right), \; avec \; \varepsilon = 0{,}75. \tag{II-22}$$

Le signe algébrique de chaque valeur de rapport de contraste du pixel, avant l'opération mise en valeur absolue (étapes E, F de la Figure II.13) doit être conservé pour être utilisé ultérieurement à l'étape H.

II.5.2.5 Masquage de contraste de luminance

Le masquage de contraste est une fonction non linéaire appliquée à chacune des réponses de contraste calculées au § II.5.2.4. Elle modélise l'incidence de la structure spatio-temporelle de la séquence d'images de référence sur la discrimination de distorsion dans la séquence d'images d'essai.

Considérons par exemple une image d'essai et une image de référence dont la différence est constituée par une onde sinusoïdale spatiale de faible amplitude. On sait que cette différence est plus visible lorsque les deux images ont en commun une onde sinusoïdale de contraste moyen dont la fréquence spatiale est identique, que si les deux images contiennent une trame. Toutefois, si le contraste de l'onde sinusoïdale commune est trop élevé la visibilité de la différence d'image diminue. On peut signaler en outre que les ondes sinusoïdales dont les fréquences spatiales sont différentes peuvent avoir un effet sur la visibilité de la différence de contraste. Il est possible de modéliser ce comportement au moyen d'une non-linéarité de type sigmoïde aux faibles énergies de contraste et par une fonction de puissance croissante pour les énergies de contraste élevées. En outre, les règles suivantes peuvent être généralement observées en ce qui concerne la vision humaine. Chaque canal se masque lui-même les fréquences spatiales élevées masquant les fréquences faibles (mais non l'inverse). Tandis que la scintillation temporelle masque la sensibilité de contraste spatial (et également l'inverse).

Compte tenu de ces propriétés de la vision, le présent modèle utilise la forme ci-dessous de non-linéarité (appliquée pixel par pixel):

$$T(y, D_i) = \frac{d_y Z_i^{\beta}}{az_i + c} \tag{II-23}$$

$$\text{avec } z_i = \frac{y}{[1 + m_f (D_i - y)]} \text{ pour i = 0,1,2, et } z_3 = \frac{y}{(1 + D_3 - m_t y)}.$$

Dans laquelle y désigne le contraste à masquer: spatial, H_i ou V_i (équation II-21) ou temporel (T_3) (équation II-22)). La valeur D_i (pixel par pixel) à une image qui dépend du niveau pyramidal i auquel y appartient. Les valeurs B, a, c, m_f, et m_t ont été obtenues par étalonnage. d_y désigne le signe algébrique du contraste y, enregistré préalablement à la mise en valeur absolue.

Le calcul de D_i exige une construction pyramidale (filtrage suivi d'un sous-échantillonnage) et une reconstitution pyramidale (suréchantillonnage suivi d'un filtrage), tel qu'il ressort de la Figure II.13 et des relations indiquées ci-dessous. Dans ces relations, fds1() désigne un filtrage de 3×3 suivi d'un sous-échantillonnage à un niveau pyramidal et usf1 () désigne un suréchantillonnage d'un niveau pyramidal suivi d'un filtrage de 3×3 (voir fin du § II.5.2.2). En premier lieu, le tableau E_0 est calculé comme suit

$$E_0 = H_0 + V_0 \tag{II-24}$$

Puis, pour i = 1, 2 les tableaux font l'objet d'un calcul récursif

$$E_i = H_i + V_i + fds1(E_{i-1}), \text{ pour i = 1,2} \tag{II-25}$$

$$E_3 = fds1(E_2) \tag{II-26}$$

Les tableaux E_i sont ensuite combinés à l'image de contraste temporel T_3 et aux images T_i de façon à obtenir les tableaux D_i de dénominateur de contraste:

$$D_3 = m_t T_3 + m_{ft}\, fds1(E_2), \tag{II-27}$$

$$T_2 = usf1(D_3), \quad T_i = usf1(T_{i+1}), \text{ pour i = 1,0, et}$$

$$D_i = E_i + T_i, \text{ pour i = 0,1,2} \tag{II-28}$$

Dans ce cas, le paramètre m_{ft} module l'intensité de masquage du canal de luminance temporelle (scintillement) par l'ensemble des canaux spatiaux de luminance; tandis que le paramètre m_t module l'intensité de masquage de chaque canal de luminance spatiale par le canal de luminance temporelle (scintillement).

Le traitement décrit ci-dessus montre que les fréquences spatiales plus élevées ont pour effet de masquer les fréquences plus faibles (puisque les tableaux D_i sont moins affectés par les niveaux pyramidaux d'indices inférieur ou égal à i), tandis que le canal temporel masque l'ensemble des canaux spatiaux et inversement. Cela confirme globalement les observations psychophysiques. Comme on pourra le constater, les grandeurs D_i, i = 0,1,2 masquent également les contrastes de chrominance (mais non l'inverse).

II.5.2.6 Construction de la carte JND de luminance

Le processus de construction décrit ci-dessous s'applique à toutes les images de contraste masqué obtenue au terme de l'étape H ci-dessus (voir Figure II.13).

- Les images des pyramides H et V (c'est-à-dire images H_0, V_0, H_1, V_1, H_2, et V_2);
- L'image T_3 (de résolution au niveau 3);

- Les images correspondantes obtenues à partir des séquences de référence (désignées par l'exposant ^{ref} à la Figure II.13).

Les quatre premières étapes du processus suivant s'appliquent séparément aux images précédentes. Dans la présentation des étapes, X désigne toute image obtenue à partir de la séquence d'essai et X^{ref} l'image correspondante obtenue à partir de la séquence de référence. Une fois cette notation précisée, les étapes sont les suivantes:

- séparer l'image X en deux alternances redressées, l'une pour les contrastes positifs et l'autre pour les contrastes négatifs. Dans l'image correspondant au contraste positif (notée X_+), les signes du contraste X (enregistrés séparément lors de l'étape E) servent à attribuer des zéros à tous les pixels de X_+ dont les contrastes sont négatifs. On procède à l'inverse dans le cas de l'image X_- correspondant au contraste négatif;
- pour chaque image I_+ et X_- procédez à une sommation locale, conformément aux observations psychophysiques, en convolutionnant l'image avec le noyau 0,25 (1,2,1) horizontalement et verticalement;
- sous-échantillonner ces images par un facteur 2 dans chaque direction, afin d'éliminer la redondance créée par la sommation effectuée à l'étape précédente. En supposant que l'image de référence correspondante X^{ref} a fait l'objet d'un traitement identique à celui de l'image X, calculer pixel par pixel les images de différences absolues $| X_+ - X_+^{ref}|$ et $| X_- - X_-^{ref}|$. Les images ainsi obtenues constituent les cartes JND.

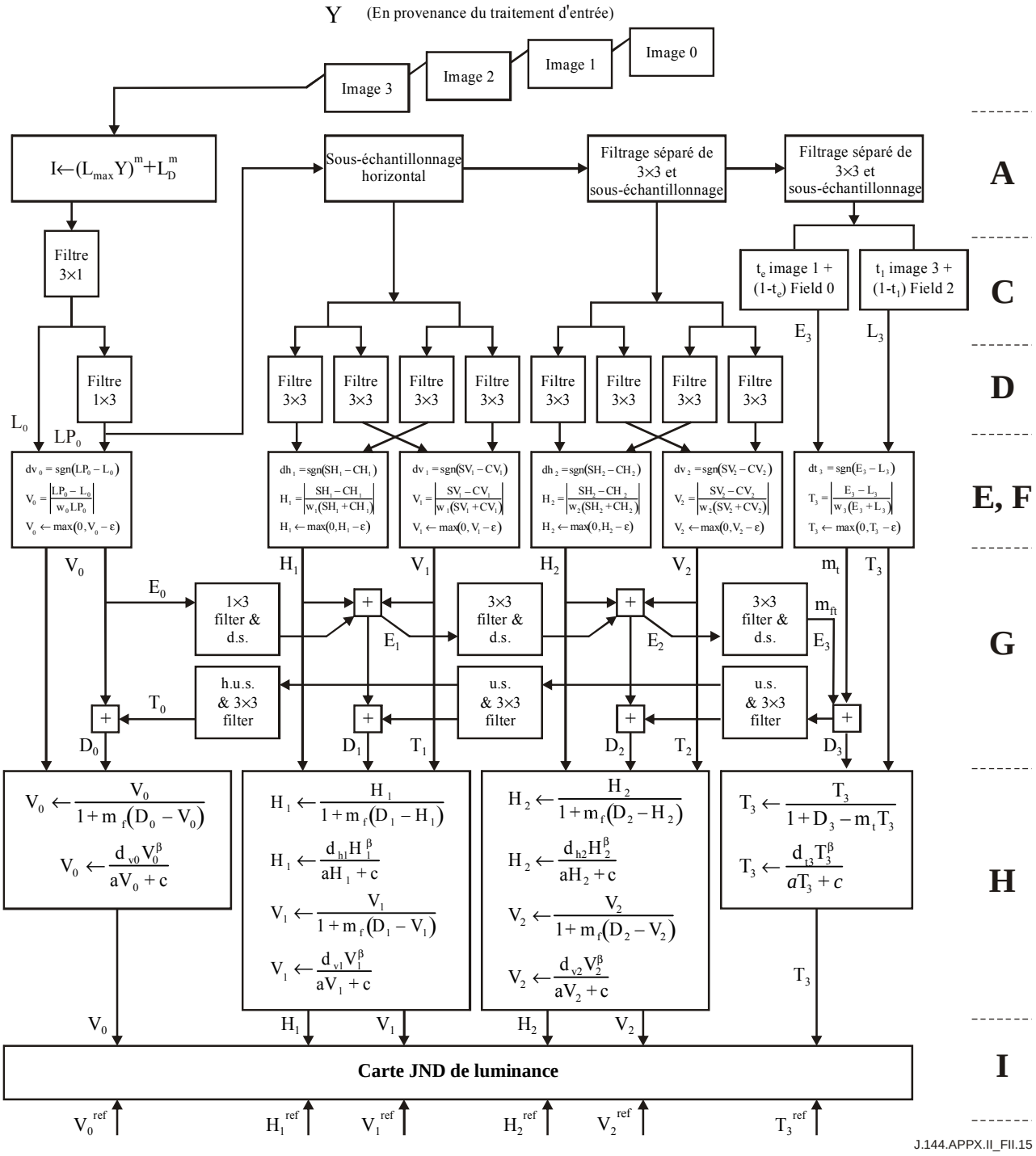
Une fois ces opérations terminées pour toutes les paires X, X^{ref}, procéder pour toutes les images au suréchantillonnage, au filtrage et à la sommation jusqu'au niveau requis de façon à calculer les mesures récapitulatives. A cet effet, procéder comme suit:

- initialiser une image de sommation courante contenant la somme de la puissance $Q^{ème}$ des images de niveau 3 calculées à partir de T_3, T_3^{ref}: $|T_{3+} - T_{3+}^{ref}|^Q$ et $|T_{3-} - T_{3-}^{ref}|^Q$. Ici Q a la valeur 2;
- suréchantillonner/filtrer l'image de sommation courante de façon à ce qu'elle contienne une image de niveau 2;
- actualiser l'image de sommation courante en lui ajoutant la puissance Qième des images de niveau 2 calculée à partir H_2, H_2^{ref} et V_2^{ref};
- suréchantillonner/filtrer l'image de sommation courante de façon à ce qu'elle contienne une image de niveau 1;
- actualiser l'image de sommation courante en lui ajoutant la puissance $Q^{ème}$ des images de niveau 1 calculées à partir de H_1, H_1^{ref}, V_1, V_1^{ref};
- suréchantillonner/filtrer l'image de sommation courante de façon à ce qu'elle contienne une image de niveau 0;
- actualiser l'image de sommation courante en lui ajoutant la puissance $Q^{ème}$ des images de niveau 0 calculées à partir de H_0, H_0^{ref}, V_0 et V_0^{ref}. Utiliser directement cette image pour le calcul des mesures récapitulatives (voir Figure II.9 et § II.5.4).

Il convient de noter qu'au terme de ce processus, l'image résultante présente une résolution moitié par rapport à l'originel. Dans le même ordre d'idée, chaque indice de niveau pyramidal indiqué dans le présent paragraphe correspond au niveau pyramidal à partir duquel il a été calculé à l'origine, qui correspond à une résolution double de celle associée à ce niveau près filtrage/sous-échantillonnage.

II.5.2.7 Traitement mi-hauteur de luminance

Si les images mi-hauteur doivent être transmises directement sans remplissage par des zéros pour atteindre la hauteur véritable de l'image, alors il faut modifier le traitement de luminance ci-dessus de façon à tenir compte du fait que la résolution verticale inhérente atteint seulement la moitié de la résolution horizontale inhérente. La Figure II.15 récapitule la luminance obtenue par l'algorithme mi-hauteur.



3×3 filter & d.s.: Filtres 3×3 distincts et sous-échantillonnage
u.s. & 3×3 filter: Suréchantillonnage filtre de 3×3

Figure II.15 – Description détaillée du traitement de luminance (mi-hauteur)

La comparaison de ce schéma et du schéma correspondant relatifs au traitement hauteur totale (Figure II.13) fait apparaître les principales modifications suivantes:

1) le canal horizontal de résolution la plus élevée H_0 est supprimé;

2) après l'étape A, l'image de résolution la plus élevée a fait l'objet d'un filtrage passe-bas vertical (c'est-à-dire le long des colonnes) au moyen de trois filtres "Kell", avec 1/8, 3/4 et

1/8 pour coefficients de pondération. Cette opération correspond au filtrage conjoint du filtre de désentrelacement supposé, ainsi qu'au filtrage effectué par les composantes verticales des filtres 3×3 à l'étape D de l'algorithme hauteur totale. L'image filtrée verticalement ainsi obtenue L_0 fait ensuite l'objet d'un filtrage horizontal, au moyen d'un filtre de 1×3 (noyau 0,25[1,2,1]). L'image obtenue (LP_0) est une version horizontale de L_0 après filtrage passe-bas.

3) $L_{0 \text{ et } LP0}$ sont combinées aux étapes E,F afin de fournir une réponse de passe-bande ($LP_0 - L_0$) divisée par une réponse orientée passe-bas (LP_0), analogue aux réponses S–C/(S+C) des autres canaux orientés.

4) L'image LP_0 (image mi-hauteur de 720×240 pixels) est sous-échantillonnée horizontalement à l'étape A pour obtenir une image hauteur totale de demi-résolution (360×240). Le traitement de cette image, et pour chacun des trois niveaux pyramidaux restants, se poursuit comme pour les options hauteur totale.

5) A l'étape G, le sous-échantillonnage et le suréchantillonnage entre les images mi-hauteur à partir du niveau zéro et les images hauteur totale de niveau 1 s'effectue par filtrage 1×3/sous-échantillonnage horizontal (noté filtre 1×3 et sous-échantillonnage) et par suréchantillonnage horizontal (h.u.s.)/*(horizontal up-sampling)*/filtrage 1×3 respectivement. Le sous-échantillonnage horizontal désigne une décimation par un coefficient 2 dans le sens horizontal, c'est-à-dire une opération consistant à supprimer une colonne sur deux de l'image. Le suréchantillonnage horizontal désigne l'introduction d'une colonne de zéros entre deux colonnes successives de l'image existante. Le noyau de filtrage suite au suréchantillonnage est défini par les coefficients 0,5 [1,2,1) pour la raison indiquée à la fin du § II.5.2.2.

En outre, dans la construction de la carte JND, le filtre 3×3 et le sous-échantillonnage à partir de V_0 est remplacé par un filtrage 1×3 et un sous-échantillonnage horizontal.

II.5.3 Traitement de la chrominance

Les § II.5.3.1 à II.5.3.6 décrivent le traitement hauteur totale. Le § II.5.3.7 présente les modifications nécessaires pour réaliser le traitement mi-hauteur.

II.5.3.1 Décomposition pyramidale de la chrominance

Outre les niveaux pyramidaux 0, 1,2 (calculés pour Y au titre du traitement de la luminance)

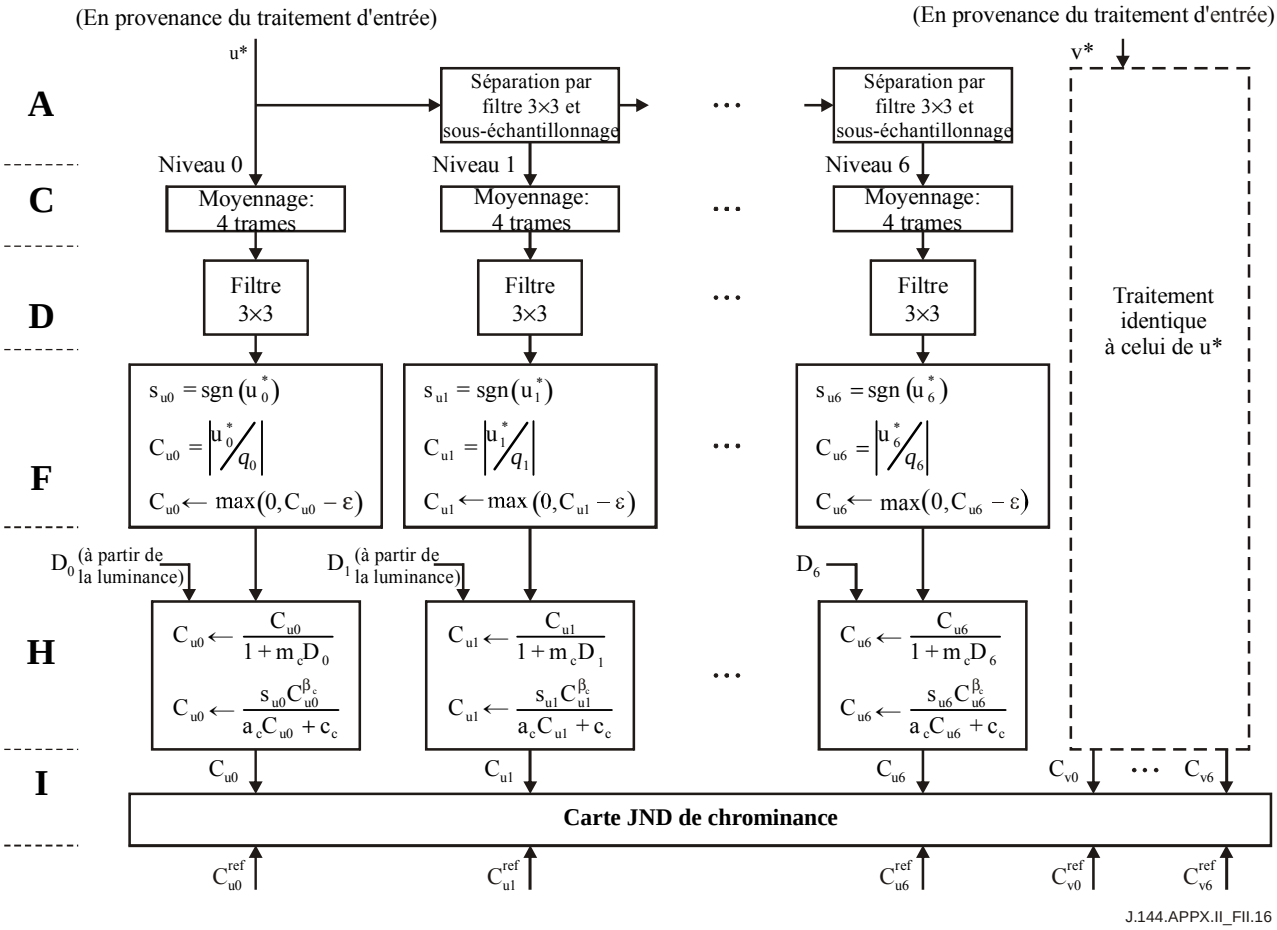
Calculons les niveaux pyramidaux 3, 4, 5, 6 pour u* et v*. Soit:

$$u_0 = u^*,\ v_0 = v^*,\ u_i = fds1(u_{i-1}),\ v_i = fds1(v_{i-1}),\ i = 1,...,6, \qquad \text{(II-29)}$$

Où fds(1) désigne l'opération de filtrage et de sous-échantillonnage décrite au § II.5.2.2. Voir Figure II.16.

La résolution spatiale du canal de chrominance de résolution la plus élevée est égale à celle du canal de luminance de niveau 0, puisque la résolution est fonction de l'espacement des pixels et non de l'espacement des récepteurs. L'espacement entre récepteur est de 0,007 degrés d'angle visuel, et l'espacement entre pixels de 0,03 degrés – calculé sur un écran comportant 480 pixels en hauteur, visualisés à une distance de 4 fois sa hauteur. De plus, la résolution du canal chromatique bleu-jaune, est limitée par le fat que le système visuel est de type tritanopique (cécité au bleu) pour des angles d'observation sous-tendus inférieurs à environ 2' (ou 0.033 degré). La résolution des pixels de 0,03 degré d'angle visuel est si proche de la valeur maximale que la solution de sécurité consiste à choisir les mêmes résolutions en pixel pour les canaux de luminance et de chrominance.

La pyramide de chrominance s'étend jusqu'au niveau 6 et non 2. Cela confirme la différence perceptible observée entre grandes trames de couleur spatialement uniforme.



NOTE – D_3, ... , D_6 sont calculés successivement par filtrage et sous-échantillonnage successifs de D_2 (à partir de la luminance)

Figure II.16 – Description détaillée du traitement de chrominance[23]

II.5.3.2 Traitement temporel de la luminance

Pour chaque niveau de résolution i, procéder à une mise en moyenne de 4 trames des images u_i et également des images v_i, avec des coefficients de pondération identiques (0,25, 0,25, 0,25, 0,25), soit, en posant:

$$u_i \leftarrow \frac{1}{4}\sum_{j=0}^{3} u_i^j \qquad v_i \leftarrow \frac{1}{4}\sum_{j=0}^{3} v_i^j \tag{II-30}$$

Avec j indice de trame.

Cette étape correspond au filtrage temporel passe-bas inhérent des canaux de couleur et remplace le traitement des images anticipées-tardives du canal de luminance temporel.

[23] Les marques d'étape sont ainsi présentées afin de maintenir la continuité avec le marquage des étapes adopté dans le cas du traitement de la luminance.

II.5.3.3 Filtrage spatial de la chrominance

Appliquer un filtre spatial Laplacien non orienté à chacune des images u_i et v_i. Le filtre utilisé dans chaque cas est le noyau 3×3 suivant:

$$1/4\begin{bmatrix} 1 & 2 & 1 \\ 2 & -12 & 2 \\ 1 & 2 & 1 \end{bmatrix} \qquad \text{(II-31)}$$

Choisir de façon à obtenir un poids total nul et produire une réponse d'intensité maximum 1 à tout contour rectiligne entre deux zones uniformes présentant entre elles une différence de valeur unitaire (la réponse maximale est obtenue par un contour vertical ou horizontal). Cela a pour effet de transformer les images u_i et v_i en cartes de différences de chrominance évaluées en unités d'espace de couleur uniforme (JND).

II.5.3.4 Calcul de contraste de chrominance

Les images u_i et v_i obtenues à l'étape D en tant que pyramides de contraste et de chrominance, à interpréter d'une manière analogue par comparaison aux contrastes de Michelson calculés à l'étape E du modèle de luminance. A l'instar des contrastes de luminance, les contrastes de chrominance sont calculés via les comparaisons intra-image affectées par les pyramides laplaciennes. De même que la différence laplacienne divisée par une moyenne spatiale représente le contraste de Michelson, qui selon la loi de Weber suppose une valeur constante au d'une unité JND (seuil de détection), la pyramide laplacienne appliquée à u_i et v_i comporte une interprétation d'une unité JND. Comme cela était le cas dans le modèle de luminance, il faut modifier cette interprétation au cours de l'étalonnage. La modification traduit l'interaction des différentes parties du modèle et le fait que les stimuli provoquant la réponse d'une unité JND ne sont pas simples une fois exprimés dans les termes du modèle.

Ensuite, niveau par niveau, le calcul consiste à diviser les images pyramidales de contraste par 7 constantes q_i (i= 0…,6) dont les valeurs sont déterminées par étalonnage. Ces constantes sont analogues aux quantités w_i (i = 0,1,2,3) dans le modèle de luminance.

On calcule ensuite les valeurs absolues écrêtées de tous les contrastes u_i et v* [avec clip(x) = max(0, x – ε)], avec ε = 0,75. Les signes algébriques sont conservés jusqu'à l'étape H, puis sont réaffectés. On évite ainsi le risque d'enregistrer des différences JND nulles entre des images distinctes, en raison de l'ambiguïté de la perte de signe liée à l'opération de mise en valeur absolue. Les résultats sont constitués par deux pyramides C_u et C_v de contraste de chrominance.

II.5.3.5 Masquage du contraste de chrominance

Les niveaux pyramidaux dénominateurs D_m (m = 0,1, 2) sont adoptés directement à partir de l'étape G du modèle de luminance, sans autre modification. En ce qui concerne les niveaux 3,…6, procéder à un filtrage et un sous-échantillonnage séquentiels de D_2 en utilisant la même méthode que dans le traitement de luminance, mais sans adjonction de nouveaux termes. Ces valeurs D_m sont utilisées à l'étape H dans l'optique de la théorie des perturbations. Puisque dans la plupart des cas les effets de luminance sont censés être prédominants par rapport aux effets de chrominance, le modèle de chrominance peut être considéré comme une perturbation de premier ordre du modèle de luminance. Les effets de la luminance (les niveaux D_m) sont donc modélisés en tant que masquage de la chrominance et non l'inverse.

Le masquage des pyramides de contraste de chrominance, à tous les niveaux pyramidaux m = 0,…,6 utilise les niveaux pyramidaux dénominateurs D_m du canal de luminance et les formes de fonction employées pour le transducteur de luminance.

$$C_{um} \leftarrow \frac{s_{um} z_{um}^{\beta c}}{a_c C_{um} + c_c} \qquad \text{(II-32)}$$

$$\text{avec} \quad z_{um} = \frac{C_{um}}{(1 + m_c D_i)}$$

Avec D_i version filtrée et sous-échantillonnée de D_2, avec i > 2. De manière analogue,

$$C_{vm} \leftarrow \frac{s_{vm} z_{vm}^{\beta c}}{a_c C_{vm} + c_c} \qquad \text{(II-33)}$$

$$\text{avec} \quad z_{vm} = \frac{C_{vm}}{(1 + m_c D_i)}$$

Il convient de noter que le signe algébrique éliminé à l'étape F a été réaffecté au moyen des facteurs s_{um} et s_{vm}. Cette opération permet d'obtenir des pyramides de contraste masqué pour u_i et v_i. Les valeurs a_c, c_c, β_c, m_c et m_f sont obtenues par étalonnage.

II.5.3.6 Construction de la carte JND de chrominance

La carte JND de chrominance se construit exactement de la même façon que la carte JND de luminance (§ II.5.2.6). Dans ce cas, la procédure s'applique à toutes les images de chrominance à contraste masquée produite par l'étape H ci-dessus (voir Figure II.16).

- images C_{u0}, C_{v0}… C_{u6}, C_{v6}
- images correspondantes obtenues à partir des séquences de référence (désignées par l'exposant ^{ref} Figure II.11).

Les trois premières étapes du processus suivant s'appliquent séparément aux images ci-dessus. Pour les présenter nous notons X toute image obtenue à partir de séquences d'essai et X^{ref} l'image correspondante obtenue à partir de la séquence de référence. Cette notation permet de décrire comme suit les étapes successives:

- séparer l'image X en deux alternances redressées, l'une pour les contrastes positifs et l'autre pour les contrastes négatifs. Dans l'image correspondant au contraste positif (notée X_+), les signes du contraste X (enregistrés séparément lors de l'étape E) servent à attribuer des zéros à tous les pixels de X_+ dont les contrastes sont négatifs. On procède à l'inverse dans le cas de l'image $X_–$ correspondant au contraste négatif;
- pour chaque image X_+ et $X_–$ procédez à une sommation locale, conformément aux observations psychophysiques, en convolutionnant l'image avec le noyau 0,5(1,2,1) horizontalement et verticalement. Ensuite sous-échantillonner les images obtenues par un facteur 2 dans chaque sens, pour éliminer la redondance créée par la sommation;
- en supposant que l'image de référence correspondante X^{ref} a fait l'objet d'un traitement identique à celui de l'image X, calculer pixel par pixel les images de différences absolues $|X_+ – X_+^{ref}|$ et $|X_– – X_–^{ref}|$. Les images ainsi obtenues constituent les cartes JND.

Une fois ces opérations terminées pour toutes les paires X, X^{ref}, procéder pour toutes les images au suréchantillonnage, au filtrage et à la sommation jusqu'au niveau requis, de façon à calculer les mesures récapitulatives. A cet effet, procéder comme suit:

- initialiser une image de sommation courante contenant la somme de la puissance $Q^{ème}$ des images de niveau à partir de C_{u6}, C_{u6}^{ref}, C_{v6}, et C_{v6}^{ref}. Ici, Q = 2.

Effectuer ensuite les deux étapes suivantes pour le niveau pyramidal m, en commençant au niveau 5 et en terminant à 0, par étapes d'un niveau:

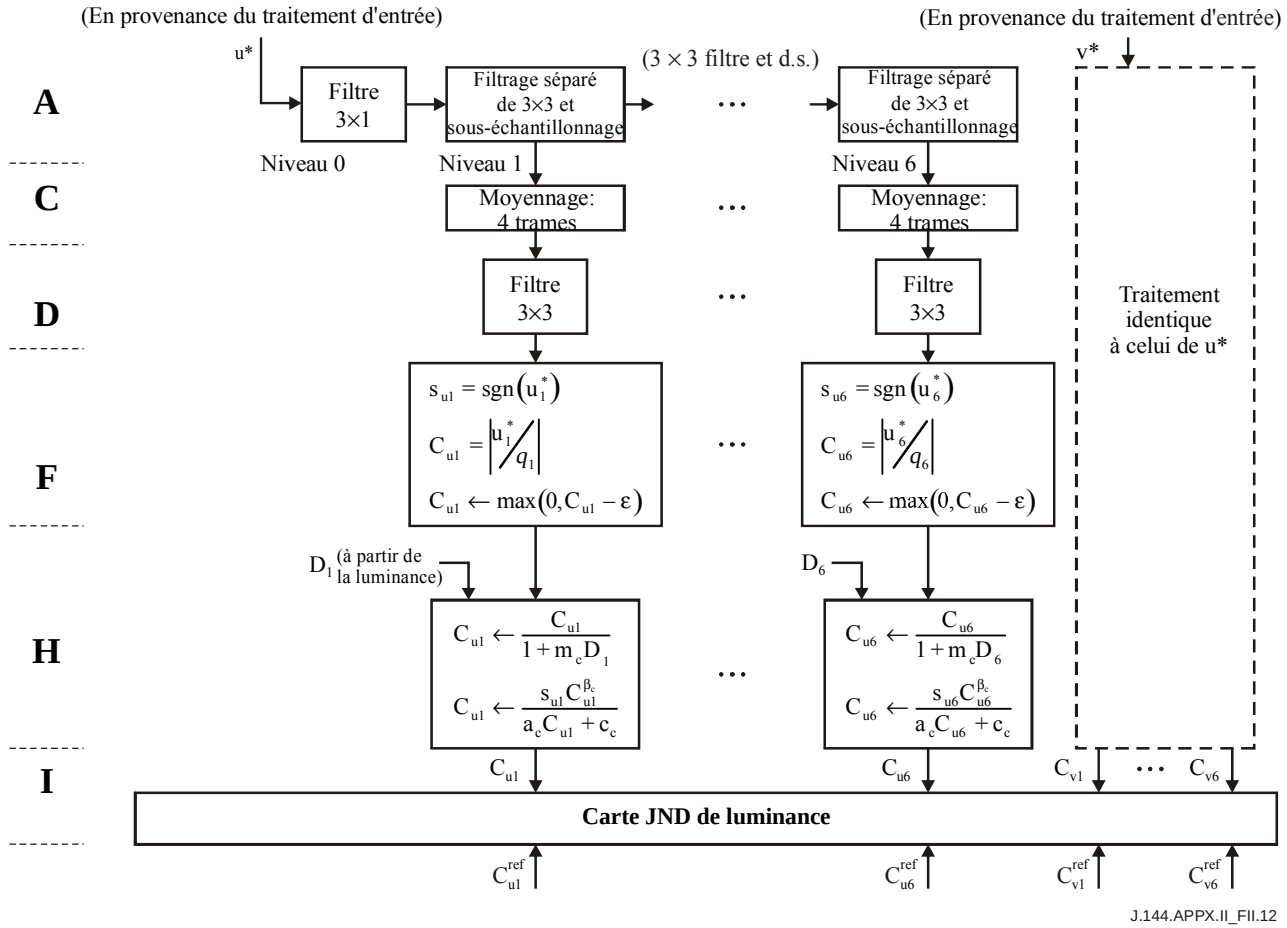
- suréchantillonner/filtrer l'image de sommation courante jusqu'à obtention d'une image de niveau m;

- actualiser l'image de sommation courante en lui ajoutant les puissance $Q^{ème}$ des images de niveau m obtenues à partir de $C_{um;}$ C_{um}^{ref}, C_{vm} et C_{vm}^{ref}.

A l'instar du traitement de la luminance, l'image résultante obtenue aux termes de ces opérations présente une résolution moitié de l'original. Dans le même ordre d'idée, chaque indice de niveau pyramidal indiqué dans le présent paragraphe correspond au niveau pyramidal à partir duquel il a été calculé à l'origine, qui correspond à une résolution double de celle associée à ce niveau après filtrage/sous-échantillonnage. L'image de niveau 0 fait ensuite directement l'objet du traitement récapitulatif (voir Figure II.9 et § II.5.4).

II.5.3.7 Traitement de la chrominance mi-hauteur

Les images de mi-hauteur doivent être transmises directement sans remplissage par des zéros pour atteindre la hauteur véritable de l'image, alors il faut modifier le traitement de luminance ci-dessus de façon à tenir compte du fait que la résolution verticale inhérente atteint seulement la moitié de la résolution horizontale inhérente. La Figure II.17 récapitule la chrominance obtenue par l'algorithme de mi-hauteur.



NOTE – D_3, ... , D_6 sont calculés par filtrage et sous-échantillonnage successifs de D_2 (à partir de la luminance)

Figure II.17 – Description détaillée du traitement de la chrominance (mi-hauteur)

La comparaison de ce schéma et du schéma correspondant relatifs au traitement hauteur totale (Figure II.11) fait apparaître les principales modifications suivantes:

1) les canaux, u_0* et v_0* de chrominance de résolution la plus élevée sont supprimés. Puisque la sensibilité de la chrominance est faible aux fréquences spatiales élevées, cette étape ne comporte aucune perte d'information significative;

2) à l'étape A, afin d'obtenir l'image de chrominance u_1* et v_1* de niveau de résolution immédiatement inférieur au niveau maximal, un noyau de filtrage passe-bas "Kell" avec vecteur de pondération (1/8, 3/4, 1/8) est appliqué verticalement (c.-à-d. aux colonnes). Cette opération correspond à l'exécution conjointe du filtrage de désentrelacement supposé et du filtrage effectué par les composantes verticales des filtres 3×3 à l'étape D de l'algorithme hauteur totale. Les images filtrées verticalement ainsi obtenues font ensuite l'objet d'un filtrage horizontal à l'aide d'un noyau 1×3 avec coefficients de pondération 0,25(1,2,1). Ce filtrage des images u* et v* confère aux images mi-hauteur une isotropie de résolution. La résolution est celle du niveau pyramidal 1 des images hauteur totale;

3) étant donné que le flux de norme Q est entièrement accumulé au niveau pyramidal 1 dans le modèle de chrominance, la carte JND de chrominance des mesures récapitulatives présente des dimensions moitié (dans le sens horizontal comme dans le sens vertical) par rapport à la carte de luminance entièrement accumulée. Avant de combiner les cartes de chrominance et de luminance afin d'obtenir la carte JND totale, il faut dans un premier temps amener la carte de chrominance à la résolution de la carte de luminance. A cet effet, un suréchantillonnage suivi d'un filtrage 3×3 permet d'obtenir la carte JND de chrominance pour les mesures récapitulatives.

II.5.4 Valeurs récapitulatives de sortie

Tel qu'indiqué dans les paragraphes précédents, les cartes JND de luminance et de chrominance passées à l'étape des valeurs récapitulatives de sortie sont des images JND et sont représentées à une résolution moitié de l'image d'origine. La redondance inhérente à l'exécution de la sommation à chaque étape du contraste masqué est ainsi mise à profit.

Ensuite, les cartes JND de luminance et de chrominance JND_L et JND_c sont combinées en une carte JND de trame totale, JND_T. La loi de combinaison est une norme Q de Minkowski (Q = 2) par analogie avec la combinaison des canaux permettant d'obtenir des cartes JND_L et JND_C:

$$JND_T(i,j) = [JND_L(i,j)^Q + JND_C(i,j)^Q]^{1/Q} \quad \text{(II-34)}$$

Ensuite chacune des trois cartes JND (luminance, chrominance et luminance-chrominance combinées) est réduite à une seule valeur récapitulative appelée valeur d'indice de qualité d'image (PQR, *picture quality rating*). Tel qu'indiqué ci-après la norme Q de Minkowski calcule les valeurs numériques récapitulatives.

Chacune des images JND résolution moitié (3 pour chaque trame – luminance, chrominance et trame totale) est ramenée à une mesure unique de qualité de fonctionnement appelée objectif de qualité de l'image (PQR) au moyen des formules ci-dessous:

$$PQR_{luma} = \left[\left(\frac{1}{N_p}\right)\sum_{i,j} JND_L(i,j)^Q\right]^{1/Q}$$

$$PQR_{chroma} = \left[\left(\frac{1}{N_p}\right)\sum_{i,j} JND_C(i,j)^Q\right]^{1/Q} \quad \text{(II-35)}$$

$$PQR_{total} = \left[\left(\frac{1}{N_p}\right)\sum_{i,j} JND_T(i,j)^Q\right]^{1/Q}$$

dans lesquelles la sommation est effectuée sur tous les pixels de la carte JND, avec Q = 4, et N_p nombre de pixels de la carte. Ainsi, trois mesures récapitulatives correspondant effectivement à JND_L, JND_C et JND_T sont calculées pour chaque trame k d'une séquence vidéo.

A partir de N valeur de l'indice PQR sur une trame unique, correspondant à une séquence vidéo,[24] une seule mesure de la performance PQR_N est calculée par la formule suivante de norme Q de Minkowski:

$$PQR_N = \left[\left(\frac{1}{N} \right) \sum_k PQR_{field}(k)^Q \right]^{1/Q} \tag{II-36}$$

NOTE – Les données d'évaluation subjective sont bruitées et peu fiables dans le cas des séquences vidéo courtes (moins d'une demi-seconde, soit 15 trames). Les estimations PQR sont médiocrement corrélées avec les évaluations subjectives dans le cas des séquences de courte durée.

II.5.5 Traitement des bordures d'image

Pour limiter les recadrages et donc éviter les artefacts de bordure, la méthode PQR remplace la bordure de l'écran par un cadre gris de dimension infinie, sans toutefois augmenter la taille réelle de l'image de plus de six pixels d'un même côté. Le recours à ce cadre virtuel supprime la nécessité de recadrer la carte JND afin d'éviter les artéfacts de bordure. Le cadre gris infini modélise les conditions de visualisation et peut donc être considéré comme étant non artéfactuel. Dans ce sens, la totalité de la carte JND est exempte d'artéfacts.

Le présent paragraphe décrit l'algorithme de traitement des bordures. Dans l'exposé ci-après une image, complétée par six pixels sur les quatre côtés est dite image *bourrée* tandis qu'une *image non bourrée* ou son emplacement à l'intérieur d'une *image bourrée* est appelée *image proprement dite.*

II.5.5.1 Couleur du cadre

Puisque les transformations de l'image sont effectuées localement, le cadre pratiquement infini peut être mis en place de façon judicieuse. Suffisamment loin de l'image proprement dite, la présence d'un cadre infini se traduit par l'existence d'un ensemble de valeurs identiques constantes à un stade de modélisation donnée quelconque. L'effet des transformations de l'image (par exemple, le filtrage) dans cette région constante peut être calculé a priori. Ainsi, une bordure étroite (six pixels dans les réalisations actuelles) peut assurer une transition adéquate depuis l'image proprement dite vers le cadre infini.

Au niveau du traitement d'entrée, les valeurs Y' = 90, U' = V' = 0 sont attribuées au cadre. (La valeur Y' = 90 correspond à la moitié de la valeur de fond de la Rec. UIT-R BT.500-11 égale à 15% de la luminance maximale de l'affichage). Toutefois, le cadre est inutile tant que le traitement d'entrée n'est pas terminé, puisque les interactions spatiales au-delà des bordures de l'image ne se produisent pas auparavant. Dans le canal de luminance aucune bordure (et donc aucune valeur du cadre) n'est associée aux images préalablement à la phase de compression de luminance. Dans le canal de chrominance, les bordures sont ajoutées une fois le traitement d'entrée terminé.

Dans le canal de luminance, la première valeur de cadre, après compression de la luminance est donné par:

$$first\_luma\_bezel = \left[L_{\max} \left(\frac{90}{255} \right)^{\gamma} \right]^m + L_d^m \tag{II-37}$$

Dans les canaux u* et v*, les premières valeurs de cadre sont nulles dans les deux cas.

---

[24] Dans ce contexte le mot *trame* désigne une valeur de PQR dans l'une quelconque des trois séquences, luminance, chrominance ou total.

Ces valeurs sont transmises tout au long des étapes suivantes de la modélisation de trois façons:

1) les fonctions pixel par pixel utilisent les anciennes valeurs du cadre afin d'obtenir les nouvelles. Par exemple, la valeur du cadre fournie par la fonction de puissance (équation II-23) est la suivante:

$$bezel\_out = (bezel\_in)^R \quad \text{(II-38)}$$

2) les filtres spatiaux 3×3 dont la somme du contenu des lignes et des colonnes est égale à P, mettent la valeur de sortie du cadre à P fois la valeur d'entrée;

3) les numérateurs de fonction de contraste et les filtres temporels sur quatre trames (qui introduisent des sommes de tableaux de valeur nulle) mettent la valeur de sortie du cadre à la valeur 0.

A partir du stade du traitement du contraste, le cadre est mis à la valeur zéro dans les canaux de luminance et de chrominance – conséquence logique d'un fonctionnement avec un noyau linéaire à somme nulle sur un tableau de valeurs spatiales constantes.

Les trois catégories ci-dessus introduisent une partie seulement des complexités nécessaires pour comprendre et appliquer l'algorithme de bordure. Le paragraphe ci-après introduit le niveau de détail suivant.

II.5.5.2 Intégration de l'image et du cadre

A partir des étapes pyramidales du modèle, les bordures doivent être fournies. La première opération sur les bordures pour une image d'entrée N sur M consiste à bourrer l'image de 6 pixels (des 4 côtés) avec la valeur approchée du cadre (first_luma_bezel pour l'image de luminance comprimée et 0 pour les images u* et v*). Les dimensions de l'image bourrée sont $(N + 12) \times (M + 12)$. Dans le cas du $k^{ème}$ niveau pyramidal (avec k allant de 0 à 7)[25], les dimensions de l'image bourrée sont $([N/2^k] + 12) \times ([M/2^k] + 12)$, avec "[*x*]" désignant la partie entière de *x*.

Les images à tous les niveaux pyramidaux sont enregistrées mutuellement dans le coin supérieur gauche de l'image proprement dite. Les indices de l'image proprement dite vont de 0 = y = hauteur, 0 = x = largeur. Le coin supérieur gauche de l'image proprement dite a toujours les indices (0,0). Les indices des pixels du cadre prennent en hauteur et en largeur des valeurs inférieures à zéro. Le pixel supérieur gauche du cadre est repéré par (–6, –6). Dans le sens de la largeur, à partir du bord gauche d'une image de largeur w (image +cadre w+12), les pixels du cadre sont indicés par *x* = (–6,–5,...,–1), tandis que l'image réelle est indexée par les valeurs (0,1,…,w–1), les indices du cadre droit étant (w,w+1,….,w+5).

Si l'on considère une image bourrée, quatre situations peuvent se présenter selon le stade suivant de traitement. Dans la description présentée ci-dessous la récapitulation du traitement spatial porte sur des lignes d'images individuelles (sachant que des opérations analogues interviennent dans le sens vertical).

a) *Opérations pixel par pixel.* Lorsque l'opération suivante doit se faire pixel par pixel (par exemple, avec une non-linéarité), l'image bourrée est simplement transmise par l'opération, les dimensions de l'image de sortie étant les mêmes que celles de l'image d'entrée. Il en va ainsi lorsque l'opération s'effectue entre les pixels correspondants de différentes trames ou de différentes bandes de couleurs.

b) *Filtres spatiaux 3×3.* Considérons (dans une dimension) une image d'entrée non bourrée de dimension N_k. Ensuite, l'image d'entrée bourrée est de dimension N_k+12, l'image de sortie bourrée étant également de dimension N_k+12. La valeur du cadre de sortie est calculée dans un premier temps (par exemple, au moyen de l'équation 37 ci-dessus et inscrite dans au moins les pixels du cadre qui ne sont pas remplis d'une autre façon par

[25] Seule la construction de la carte de chrominance exige le niveau 7.

l'opération suivante. Ensuite, en commençant un pixel à partir du bord gauche de l'image d'entrée bourrée, le noyau 3×3 commence à traiter l'image d'entrée et à écrire en surimpression les valeurs du cadre de l'image de sortie, en s'arrêtant à un pixel du bord droit (ou du bord inférieur) de l'image (lorsque la valeur du cadre d'origine est conservée). En raison de la valeur préinscrite du cadre il est inutile que le noyau utilise des points situés à l'extérieur de l'image d'origine (bourrée) pour calculer ces valeurs.

c) *Pour le filtrage et le sous-échantillonnage de l'algorithme REDUCE.* Considérons une image bourrée d'entrée de dimension N_k+12, et un tableau de sortie auquel la dimension $[N_k/2]$+12 a été attribuée. La valeur du cadre (calculée par exemple au moyen de l'équation 37 est inscrite au moins dans les pixels du cadre qui ne sont pas remplis d'une autre façon par les opérations ultérieures de filtrage et de sous-échantillonnage. Ensuite, l'image d'entrée est filtrée selon les indications de b ci-dessus, le filtre étant néanmoins appliqué aux pixels –4, –2, 0, 2, 4 jusqu'à épuisement des pixels de l'image d'entrée, et les valeurs de sortie sont inscrites dans les pixels consécutifs –2, –1, 0, 1, 2 jusqu'à ce qu'il n'y ait plus de place dans l'image de sortie. Il est à noter que l'emplacement du pixel 0 dans la nouvelle image se situe à sept pixels de l'extrémité gauche de la nouvelle image. L'application du filtre au dernier pixel dirige le pixel d'entrée $N_k + 3$ vers le pixel de sortie $[N_k/2] + 2$ si N_k est impair et dirige le pixel d'entrée $N_k + 4$ vers les pixels de sortie $[N_k/2 + 2]$ si N_k est pair (ici, le pixel d'entrée du filtre est défini comme le pixel correspondant au centre du noyau de trois pixels).

Ci-dessous figurent quatre exemples simplifiés de traitement des bordures selon l'algorithme REDUCE. Dans chaque cas, les pixels sont étiquetés de façon consécutive à l'intérieur de crochets et en gras, en ce qui concerne *l'image proprement dite* et en ce qui concerne le cadre préécrit.

EXEMPLE 1: $N_k = 3$ (dimension d'entrée impaire, dimension de sortie impaire.)

Entrée: –6 –5 –4 –3 –2 –1 **[0 1 2]** 3 4 5 6 7 8

Sortie: –6 –5 –4 –3 –2 –1 **[0]** 1 2 3 4 5 6

EXEMPLE 2: $N_k = 4$. (dimension d'entrée paire, dimension de sortie paire.)

Entrée: –6 –5 –4 –3 –2 –1 **[0 1 2 3]** 4 5 6 7 8 9

Sortie: –6 –5 –4 –3 –2 –1 **[0 1]** 2 3 4 5 6 7

EXEMPLE 3: $N_k = 5$. (Dimension d'entrée impaire, dimension de sortie paire.)

Entrée: –6 –5 –4 –3 –2 –1 **[0 1 2 3 4]** 5 6 7 8 9 10

Sortie: –6 –5 –4 –3 –2 –1 **[0 1]** 2 3 4 5 6 7

EXEMPLE 4: $N_k = 6$. (dimension d'entrée paire, dimension de sortie impaire.)

Entrée: –6 –5 –4 –3 –2 –1 **[0 1 2 3 4 5]** 6 7 8 9 10 11

Sortie: –6 –5 –4 –3 –2 –1 **[0 1 2]** 3 4 5 6 7 8

d) *Echantillonnage et filtrage selon l'algorithme EXPAND*. Supposons une image d'entrée *bourrée* au niveau k+1, de dimension $N_{k+1} + 12$, un tableau de sortie au niveau k de dimension $N_k + 12$ lui étant attribué et initialisé à 0 (on notera que N_k+1 a été prédéfini comme égal à ($[N_k/2]$). Ensuite, les pixels d'entrée –2,–1,0,1,… sont affectés aux pixels de sortie –4,–2,0,2,4,… Ensuite l'opération de filtrage décrite dans b ci-dessus est réalisée sur l'image obtenue. Enfin, la valeur du cadre au niveau k est calculée par exemple au moyen de l'équation 37, et affectée aux trois pixels extérieurs des quatre côtés de l'image de sortie. Il est à noter que l'emplacement du pixel 0 sur la nouvelle image se situe à sept pixels de l'extrémité gauche de la nouvelle image. L'application du filtre au dernier pixel dirige le pixel d'entrée $[N_k/2] + 2$ vers les pixel de sortie $N_k + 3$, si N_k est impair et le pixel d'entrée $[N_k/2] + 2$ vers le pixel de sortie $N_k + 4$ si N_k est pair. (Là encore, le pixel d'entrée du filtre est défini comme le pixel correspondant au centre du noyau de trois pixels).

Ci-dessous figurent quatre exemples simplifiés de traitement de bordure au moyen de l'algorithme EXPAND. Dans chaque cas, les pixels sont étiquetés consécutivement entre crochets et en gras pour l'*image proprement dite*, et soulignés pour les pixels du cadre inscrits a posteriori.

EXEMPLE 1: $N_k = 3$. (Dimension d'entrée impaire, dimension de sortie impaire.)

Entrée: –6 –5 –4 –3 –2 –1 **[0]** 1 2 3 4 5 6

Sortie: –6 –5 –4 –3 –2 –1 **[0 1 2]** 3 4 5 6 7 8

EXEMPLE 2: $N_k = 4$. (dimension d'entrée paire, dimension de sortie paire.)

Sortie: –6 –5 –4 –3 –2 –1 **[0 1]** 2 3 4 5 6 7

Entrée: –6 –5 –4 –3 –2 –1 **[0 1 2 3]** 4 5 6 7 8 9

EXEMPLE 3: $N_k = 5$. (dimension d'entrée paire, dimension de sortie impaire.)

Sortie: –6 –5 –4 –3 –2 –1 **[0 1]** 2 3 4 5 6 7

Entrée: –6 –5 –4 –3 –2 –1 **[0 1 2 3 4]** 5 6 7 8 9 10

EXEMPLE 4: $N_k = 6$. (Dimension d'entrée impaire, dimension de sortie paire.)

Sortie: –6 –5 –4 –3 –2 –1 **[0 1 2]** 3 4 5 6 7 8

Entrée: –6 –5 –4 –3 –2 –1 **[0 1 2 3 4 5]** 6 7 8 9 10 11

Ces exemples montrent que l'inscription en superposition des pixels du cadre n'a pas d'incidence lorsque l'algorithme EXPAND est répété à des niveaux successifs.

Appendice II.A

Bibliographie

– Recommandation UIT-R BT.500-10 (2002), *Méthodologie d'évaluation subjective de la qualité des images de télévision*.

– USA Standards Committee T1, technical report T1.TR.73-2001: Video Normalization Methods Applicable to Objective Video Quality Metrics Utilizing a Full Reference technique.

– Burt PJ, and Adelson EH (1983), *The laplacian pyramid as a compact image code.* IEEE Trans Comm 31-532-540. <http://www.citeseer.ifi.unizh.ch/burt83laplacian.html>.

Appendice II.B

Facteurs d'essai, techniques de codage et applications

Voir rapport final VQEG (Document de référence UIT-T II.A) pour indications complémentaires concernant les données fournies dans ces tableaux. Toutes les données concernent le système à 525 lignes.

Tableau II.B.1 – Facteurs d'essai, techniques de codage et applications pour lesquels la méthode PQR a vérifié la précision spécifiée au § II.1.3.4

Débit binaire	Résolution	Méthode	Observations
2 Mbit/s	¾ résolution	mp@ml	Uniquement réduction de résolution horizontale
2 Mbit/s	¾ résolution	sp@ml	
4.5 Mbit/s		mp@ml	Avec erreurs
3 Mbit/s		mp@ml	Avec erreurs
4.5 Mbit/s		mp@ml	
3 Mbit/s		mp@ml	
4.5 Mbit/s		mp@ml	NTSC et/ou PAL composite
6 Mbit/s		mp@ml	
8 Mbit/s		mp@ml	NTSC et/ou PAL composite
8 & 4,5 Mbit/s		mp@ml	Deux codecs concaténés
19 Mbit/s – NTSC- 19 Mbit/s – NTSC- 12 Mbit/s		422p@ml	NTSC 3 générations
50-50-… -50 Mbit/s		422p@ml	7e génération avec décalage/I trame
19-19-12 Mbit/s		422p@ml	3e génération
s/o		s/o	Betacam multigénération avec compensation des évanouissements (4 ou 5, composite/composant)

Tableau II.B.2 – Facteurs d'essai, techniques de codage et applications pour lesquels la méthode PQR n'a pas vérifié la précision spécifiée

Débit binaire	Résolution	Méthode	Observations
1,5 Mbit/s	CIF	H.263	Plein écran
768 kbit/s	CIF	H.263	Plein écran
Autre			La méthode PQR spécifiée à l'Appendice I n'est pas adaptée aux applications de vidéoconférence qui répètent des trames ou ne vérifie pas les spécifications de latence et de délai des classes vidéo. En outre, la méthode PQR est uniquement applicable aux systèmes de transmission de télédiffusion, comportant de très faibles taux d'erreur par exemple, les systèmes pris en compte dans les essais VQEG.

Tableau II.B.3 – Séquences d'essai utilisées afin de déterminer les facteurs d'essai, les techniques de codage et les applications pour lesquels la méthode PQR a vérifié la précision spécifiée au § II.1.3.4

Séquence	Caractéristiques
Baloon-pops	Film, couleurs saturées, mouvement
NewYork 2	Effet de masquage, mouvement)
Mobile&Calendar	Disponible dans les deux formats, couleur, mouvement
Betes_pas_betes	Couleur, synthétique, mouvement, plan de coupe
Le_point	Couleur, transparence, mouvement dans toutes les directions
Autumn_leaves	Couleur, paysages, travelling optique, mouvement de masses d'eau
Football	Couleur, mouvement
Sailboat	Pratiquement fixe
Susie	Couleur de la peau
Tempete	Couleur, mouvement

Appendice II.C

Classification des erreurs

La classification des erreurs est une façon d'évaluer d'efficacité de la mesure de la qualité vidéo perçue (VQM, *video quality metric*). Le présent appendice examine la signification des erreurs de classification, sous forme de diagramme de l'indice subjectif z en fonction de Δ-VQM défini dans le corps de la Recommandation. Pour l'exposé ci-dessous sera utilisé l'intervalle commun [0,1] pour les indices subjectifs et objectifs. Dans cet intervalle commun [0,1], "0" correspond à l'absence de dégradation et "1" correspond à la dégradation maximale.

Pour tout essai subjectif il est possible de fixer un seuil Δz qui indique, d'un point de vue statistique, quand deux points de données (A, B) sont équivalents et quand il est possible de les distinguer[26]. Une fois ce seuil défini, les résultats des essais subjectifs permettent de classer chaque paire de points de données (A, B) dans l'une des trois catégories suivantes:

$\Delta z_{AB} < -\Delta z$ → A est meilleur que B → Bs

$-\Delta z \leq \Delta z_{AB} \leq \Delta z$ → A est identique à B → Es

$\Delta z < \Delta z_{AB}$ → A est plus mauvais que B → Ws

Les abréviations utilisées pour chacune des trois catégories (Bs, Es et Ws) signifient respectivement, qualité subjective supérieure, qualité subjective équivalente, et qualité subjective inférieure).

Considérons à présent un seuil analogue pour les valeurs VQM Δ_0

VQM(A)– VQM (B) < – Δo → A est meilleur que B → Bo

–Δo ≤ VQM (A) – VQM (B) ≤ Δo → A est identique à B → Eo

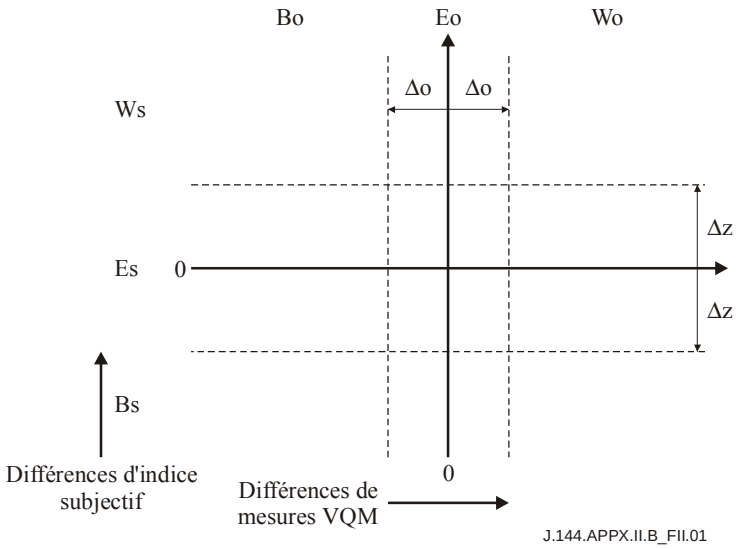
Δo < VQM (A) – VQM (B) → A est plus mauvais que B → Wo

Les abréviations correspondant aux trois catégories (Bo, Eo et Wo) signifient respectivement qualité objective supérieure, qualité objective équivalente et qualité objective inférieure.

Puisque chaque paire de points de données peut être classées dans trois catégories au terme des essais subjectifs et dans trois autres catégories pour l'indice VQM, on distingue neuf résultats possibles. Dans l'espace à deux dimensions des différences d'indice subjectif en fonction des différences de mesures VQM, les lignes en pointillés définissent ces neuf zones de résultats.

---

[26] Les points de données A et B représentent en fait des séries d'observations de deux combinaisons SRC/HRC. Tel qu'indiqué dans le corps de la Recommandation, le terme Δz_{AB} est égal à la différence entre les moyennes de A et B $(\hat{S}_{A\bullet} - \hat{S}_{B\bullet})$, divisée par l'écart-type calculé $\sqrt{(V_A / N_A + V_B / N_B)}$, avec V_A variance des notes relatives à la situation A, et N_A nombre d'observations liées à la situation A, etc.

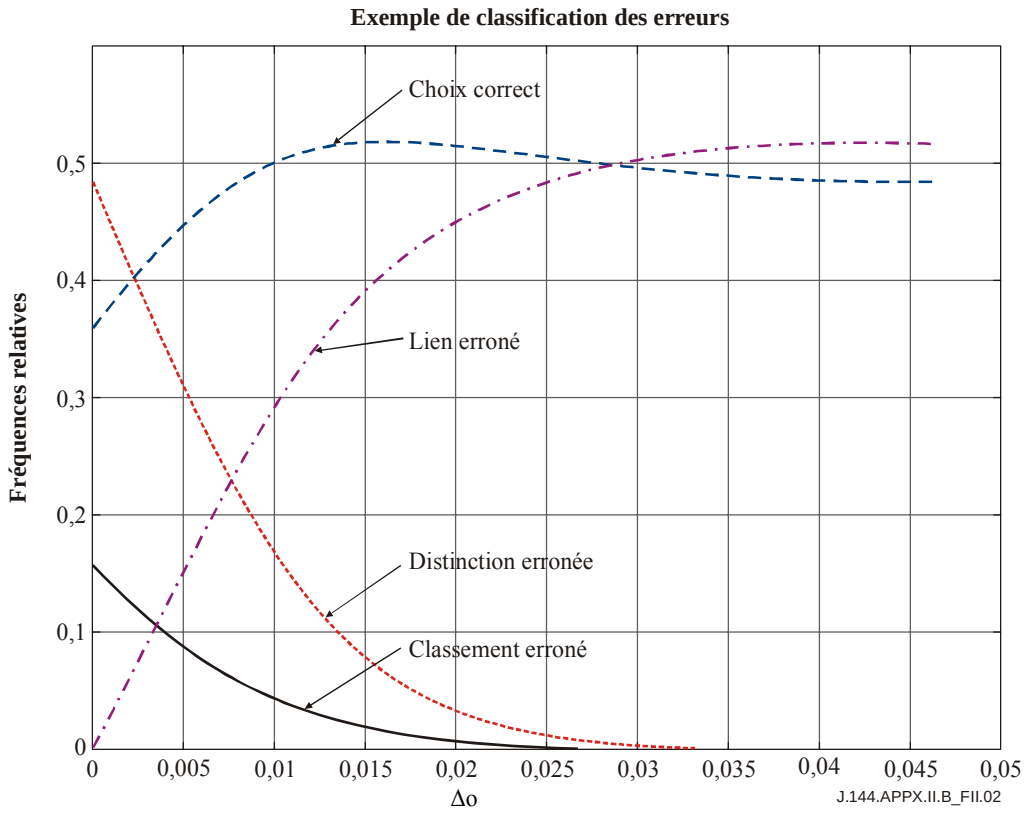


Dans le tableau ci-dessous chacun des neuf résultats sont désignés du point de vue de la réponse à la question "comment se situe la classification VQM à trois niveaux par rapport à la classification à trois niveaux des essais subjectifs?"

	Bs	**Es**	**Ws**
Wo	Classement erroné	Différence erronée	Choix correct
Eo	Lien erroné	Choix correct	Lien erroné
Bo	Choix correct	Différence erronée	Classement erroné

On notera que pour trois des conclusions, la classification VQM concorde avec celle des essais subjectifs. Ces trois résultats sont désignés par la mention "Choix correct". Les six autres résultats correspondent à trois types différents d'erreurs susceptibles de résulter de l'utilisation d'une mesure VQM. Le lien erroné est sans doute l'erreur la moins grave. Il intervient lorsque deux points de données sont différents d'après les essais subjectifs, mais sont identiques d'après la mesure VQM. Une distinction erronée est généralement plus préjudiciable. Celle-ci intervient lorsque d'après les essais subjectifs deux points de données sont identiques, mais différents d'après la mesure VQM. Le classement erroné est en général l'erreur la plus grave. Dans ce cas, A est meilleur que B d'après les essais subjectifs, mais B est meilleur que A d'après la mesure VQM.

Un essai subjectif et une mesure VQM quelconques permettent de constituer toutes les paires distinctes de points de données possibles et de compter ensuite le nombre de paires dans chacune des quatre catégories distinctes de résultats: choix correct, lien erroné, distinction erronée et classement erroné. On peut ensuite normaliser le nombre d'occurrences en le divisant par le nombre total de paires distinctes et établir les fréquences relatives correspondant aux quatre catégories de résultats. En règle générale, ces résultats seront fonction conjointement de Δs et Δo. Le diagramme ci-dessous donne des exemples de résultats correspondant à une mesure VQM fictive. Δz a été choisi de façon à obtenir un seuil de confiance estimé à 95% dans la classification des mesures subjectives, Δo étant le paramètre libre représenté en abscisse.

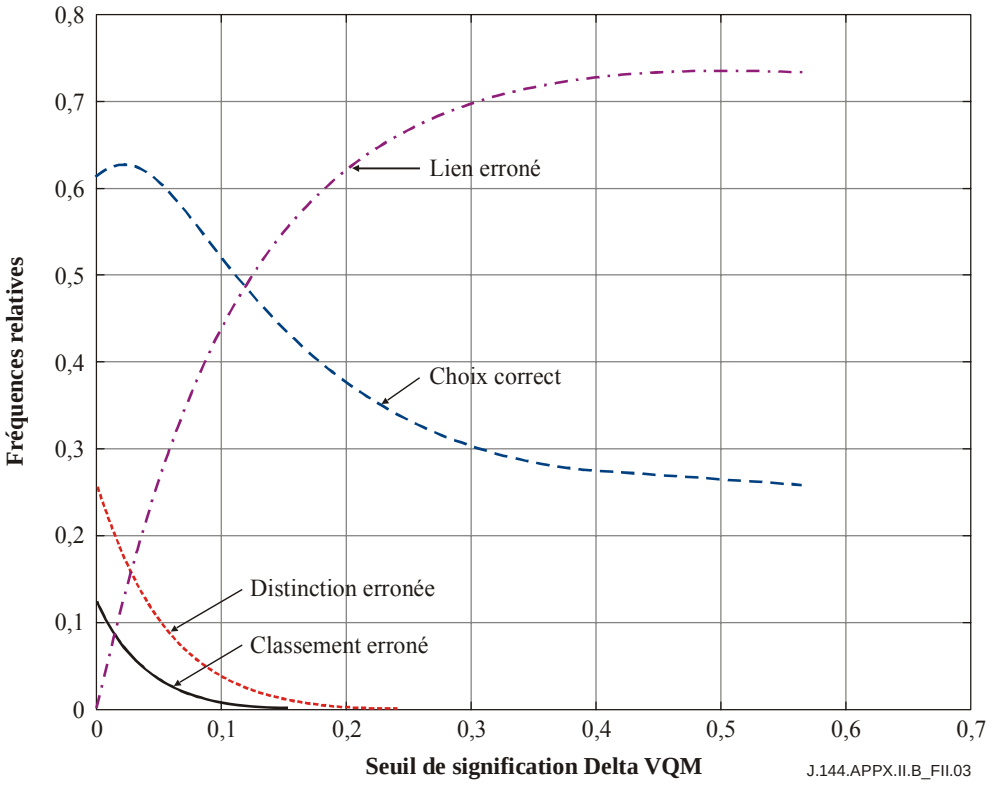


On notera que pour des valeurs croissantes de Δo, la mesure VQM indique un nombre croissant de paires de points de données équivalents. Il en résulte une diminution des occurrences des distinctions et des classements erronés, mais une augmentation des liens erronés. Lorsque le Δo atteint 0,5, la fréquence des liens erronés tend vers 0,52. A ce stade, la mesure VQM déclare l'équivalence de toutes les paires et ce faisant donne une indication fausse dans 52% des cas et exacte dans 48%. Ce résultat est compatible avec le fait selon lequel dans cet essai 48% des paires de points de données ont été déclarés équivalents d'après les mesures subjectives. Un diagramme de ce type pourrait servir à choisir une valeur appropriée de Δo. Ainsi, Δo pourrait être choisi de façon à maximiser la probabilité de choix correct ou de minimiser certaines sommes pondérées de fréquences relatives d'erreurs.

Dans le programme qui a permis d'établir la figure ci-dessus (élément du programme MATLAB mentionné à l'Annexe B) le seuil utilisé pour l'essai subjectif est noté subj_th. Le seuil utilisé pour ΔVQM est noté vqm_th et constitue une valeur paramétrique. Le programme trace la fréquence d'occurrence des trois types d'erreurs différents et la fréquence de l'absence d'erreurs en fonction de vqm_th. Une valeur optimale de vqm_th pourrait être définie comme celle qui maximise la fréquence d'occurrence de l'absence d'erreur ou celle qui minimise une somme des erreurs pondérée par les coûts. Il est à noter en règle générale que les erreurs dites de "lien erroné" sont vraisemblablement les moins préjudiciables, tandis que les distinctions erronées le sont davantage et que les classements erronés sont les plus préjudiciables.

NOTE – Les neuf conclusions et le tableau trois dans l'espace (ΔVQM indice Z subjectif) constitue l'illustration la plus naturelle de cette analyse. Cette représentation suppose pour ΔVQM des valeurs bipolaires. Or, le programme a déjà mis ΔVQM en valeur absolue (et remplacé Z par –Z) pour tous les points comportant des valeurs négatives de ΔVQM. Bien que les relations mathématiques restent inchangées, on obtient alors une description plus naturelle de la situation avec six conclusions et un tableau de 2 sur 3. Deux des conclusions correctes (A meilleur que B et A pire que B) ont été regroupées. Il subsiste alors deux conclusions intitulées "lien erroné" mais seulement une conclusion "distinction erronée" et une conclusion "classement erroné".

La classification suivante des erreurs s'applique à la méthode PQR spécifiée dans le présent appendice.



SÉRIES DES RECOMMANDATIONS UIT-T

Série A	Organisation du travail de l'UIT-T
Série D	Principes généraux de tarification
Série E	Exploitation générale du réseau, service téléphonique, exploitation des services et facteurs humains
Série F	Services de télécommunication non téléphoniques
Série G	Systèmes et supports de transmission, systèmes et réseaux numériques
Série H	Systèmes audiovisuels et multimédias
Série I	Réseau numérique à intégration de services
Série J	**Réseaux câblés et transmission des signaux radiophoniques, télévisuels et autres signaux multimédias**
Série K	Protection contre les perturbations
Série L	Construction, installation et protection des câbles et autres éléments des installations extérieures
Série M	Gestion des télécommunications y compris le RGT et maintenance des réseaux
Série N	Maintenance: circuits internationaux de transmission radiophonique et télévisuelle
Série O	Spécifications des appareils de mesure
Série P	Terminaux et méthodes d'évaluation subjectives et objectives
Série Q	Commutation et signalisation
Série R	Transmission télégraphique
Série S	Equipements terminaux de télégraphie
Série T	Terminaux des services télématiques
Série U	Commutation télégraphique
Série V	Communications de données sur le réseau téléphonique
Série X	Réseaux de données, communication entre systèmes ouverts et sécurité
Série Y	Infrastructure mondiale de l'information, protocole Internet et réseaux de prochaine génération
Série Z	Langages et aspects généraux logiciels des systèmes de télécommunication

Imprimé en Suisse
Genève, 2009