UNION INTERNATIONALE DES TÉLÉCOMMUNICATIONS

UIT-T J.144

(03/2001)

SECTEUR DE LA NORMALISATION
DES TÉLÉCOMMUNICATIONS
DE L'UIT

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

Mesure de la qualité de service

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

Recommandation UIT-T J.144

(Antérieurement Recommandation du CCITT)

RECOMMANDATIONS UIT-T DE LA SÉRIE J

RÉSEAUX CÂBLÉS ET TRANSMISSION DES SIGNAUX RADIOPHONIQUES, TÉLÉVISUELS ET AUTRES SIGNAUX MULTIMÉDIAS

Recommandations générales	J.1–J.9
Spécifications générales des transmissions radiophoniques analogiques	J.10–J.19
Caractéristiques de fonctionnement des circuits radiophoniques analogiques	J.20–J.29
Equipements et lignes utilisés pour les circuits radiophoniques analogiques	J.30–J.39
Codeurs numériques pour les signaux radiophoniques analogiques	J.40–J.49
Transmission numérique de signaux radiophoniques	J.50–J.59
Circuits de transmission télévisuelle analogique	J.60–J.69
Transmission télévisuelle analogique sur lignes métalliques et interconnexion avec les faisceaux hertziens	J.70–J.79
Transmission numérique des signaux de télévision	J.80–J.89
Services numériques auxiliaires propres aux transmissions télévisuelles	J.90–J.99
Prescriptions et méthodes opérationnelles de transmission télévisuelle	J.100–J.109
Services interactifs pour la distribution de télévision numérique	J.110–J.129
Transport des signaux MPEG-2 sur les réseaux par paquets	J.130–J.139
Mesure de la qualité de service	**J.140–J.149**
Distribution de la télévision numérique sur les réseaux locaux d'abonnés	J.150–J.159
IPCablecom	J.160–J.179
Divers	J.180–J.199
Application à la télévision numérique interactive	J.200–J.209

Pour plus de détails, voir la Liste des Recommandations de l'UIT-T.

Recommandation UIT-T J.144

Techniques de mesure 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 dispose du signal vidéo de référence complet.

Source

La Recommandation J.144 de l'UIT-T, élaborée par la Commission d'études 9 (2001-2004) de l'UIT-T, a été approuvée le 9 mars 2001 selon la procédure définie dans la Résolution 1 de l'AMNT.

AVANT-PROPOS

L'UIT (Union internationale des télécommunications) est une institution spécialisée des Nations Unies dans le domaine des télécommunications. L'UIT-T (Secteur de la normalisation des télécommunications) 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.

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 n'avait pas 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 responsables de la mise en œuvre de consulter la base de données des brevets du TSB.

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. 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.

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 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.

2 Références

2.1 Références normatives

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.

– UIT-R BT.500-9 (1998), *Méthodologie d'évaluation subjective de la qualité des images de télévision*.

2.2 Références informatives

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

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

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

– Commission d'études 9 de l'UIT-T, Contribution COM 9-80 (2000), *Rapport final du Groupe d'experts sur la qualité vidéo concernant la validation de modèles d'évaluation objective de la qualité vidéo*.

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.

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).

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.

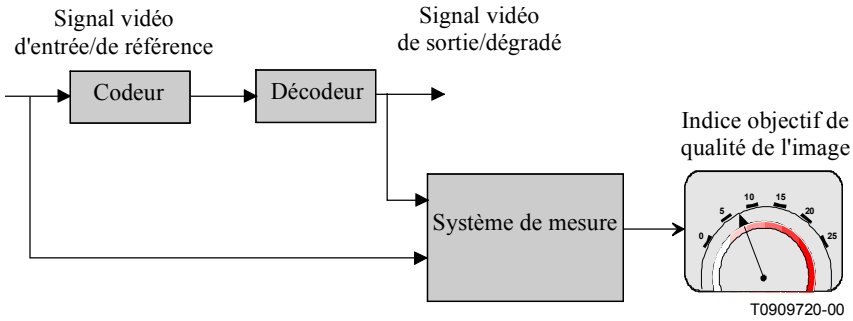


Figure 1/J.144 – Application de la méthode de mesure de la qualité perçue avec référence complète pour tester un codec en laboratoire

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.

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.

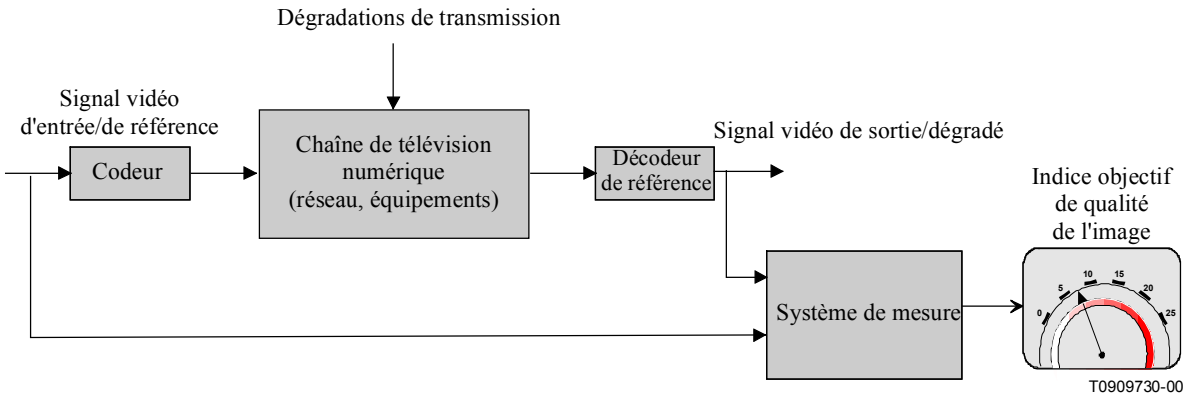


Figure 2/J.144 – Application de la méthode de mesure de la qualité perçue avec référence complète pour tester une chaîne de transmission

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.

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 aux Commissions d'études 12 et 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 fondés sur la méthode aux deux extrémités.

Le VQEG a publié un projet de rapport final détaillé sur la première phase de ses travaux en mars 2000.

Il est conseillé aux lecteurs d'étudier ce rapport pour avoir une vue d'ensemble des travaux du VQEG jusqu'à cette date. En résumé, le rapport contient les résultats de tests réalisés sur dix modèles soumis au VQEG par dix proposants différents, des notes objectives ayant été calculées au moyen de ces modèles et comparées à l'évaluation subjective, et ce pour une large diversité de systèmes vidéo et de séquences source. Les tests ont permis de comparer la qualité des modèles des proposants par rapport à des tests d'évaluation subjective des mêmes images ainsi que par rapport à l'algorithme "de référence" PSNR valeur de crête du rapport signal sur bruit (PSNR, *peak signal-to-noise ratio*). 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).

Plus de 26 000 notes d'opinion subjective ont été produites sur la base de vingt séquences source différentes traitées par seize systèmes vidéo différents et ont été évaluées dans huit laboratoires indépendants dans le monde entier.

Les tests subjectifs organisés ont été classés en quatre catégories: haute qualité 50 Hz, faible qualité 50 Hz, haute qualité 60 Hz et faible qualité 60 Hz. Dans ce contexte, haute qualité renvoie à la qualité de production et faible qualité renvoie à la qualité de distribution. Les catégories haute qualité comprenaient des signaux vidéo à des débits binaires compris entre 3 Mbit/s et 50 Mbit/s. Les catégories faible qualité comprenaient des signaux vidéo à des débits binaires compris entre 768 kbit/s et 4,5 Mbit/s.

Les procédures de la Rec. UIT-R BT.500-9 relatives à la méthode à double stimulus utilisant une échelle de qualité continue (DSCQS, *double stimulus continuous quality scale*) ont 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 des données objectives avec les résultats subjectifs.

Outre l'analyse fondée sur la totalité des données, des sous-ensembles fondés sur les quatre catégories de tests subjectifs et l'ensemble des données à l'exclusion de celles obtenues avec certains systèmes de traitement vidéo ont été analysés pour déterminer la sensibilité des résultats à divers paramètres dépendant de l'application.

Les résultats obtenus au moyen des deux algorithmes qui n'avaient pas été entièrement testés ou qui avaient présenté des problèmes de mise en œuvre ont été rejetés. Les résultats du test du VQEG, fondés sur l'analyse obtenue pour les quatre quadrants de test subjectif individuel, montrent pour l'essentiel ce qui suit:

- aucun des systèmes de mesure objective essayés ne peut remplacer un test subjectif;
- aucun modèle de mesure objective n'est supérieur aux autres dans l'ensemble des conditions de référence;
- aucun modèle de mesure objective n'est statistiquement supérieur au rapport PSNR dans l'ensemble des conditions de référence;
- sur la base de ces constatations, on ne peut, à l'heure actuelle, recommander à l'UIT une méthode unique.

Les travaux effectués par le VQEG ont tout de même permis de bien mieux cerner le problème de l'évaluation de la qualité vidéo perçue ainsi que les besoins des utilisateurs, ce qui conduira probablement à l'élaboration de modèles de perception améliorés, qui seront implémentés dans des équipements proposés sur le marché.

Il est prévu que le sous-comité G-2.1.6 de l'IEEE mène des études en vue de fournir un ensemble de scènes de test dégradées de façon contrôlée. A chaque scène sera associée une échelle de perception correspondante, étalonnée par pas successifs de différences de dégradation tout juste perceptibles. Ces scènes sont censées représenter un bon ensemble de données de référence pour tester les systèmes à venir.

7 Conclusions

Comme aucune méthode de mesure ne peut être recommandée à l'heure actuelle, on formule, dans le présent paragraphe, un avis général sur les modèles d'évaluation de la qualité vidéo utilisant la méthode avec référence complète. Les modèles actuels évalués par le VQEG sont présentés en détail dans les appendices. Il est prévu, compte tenu des travaux futurs du VQEG et d'autres groupes, d'adopter un ou plusieurs de ces modèles sous forme de normes. Les futurs travaux du VQEG porteront sans doute aussi sur l'examen d'autres conditions de test, des distances de visionnement moins grandes et d'autres types de gammes de distorsions, par exemple, qui permettront peut-être de

faire une meilleure distinction au sein des modèles objectifs et entre chaque modèle et rapport signal sur bruit.

Avis général

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.

Lorsqu'ils rendent compte des résultats des mesures de la qualité vidéo perceptuelle, les exploitants devraient toujours mentionner la marque, le modèle et les réglages de l'équipement utilisé et les images tests correspondantes. Cela permettrait notamment aux exploitants de comparer les résultats de ces tests avec ceux de tests effectués par d'autres exploitants.

Cette précaution est nécessaire car il est possible que les équipements de mesure de la qualité perçue fondés sur la méthode avec référence complète fournissent un degré de corrélation avec les tests d'évaluation subjective qui dépend, entre autres, de l'ensemble d'images de test sélectionnées, du degré de compression appliqué au flux binaire vidéo testé et du nombre de choix d'implémentation que le fabricant peut avoir faits dans sa conception.

Lorsque les exploitants hésitent entre plusieurs modèles d'équipement de mesure de la qualité perçue fondés sur la méthode avec référence complète présents sur le marché ou avant qu'ils ne décident de choisir un nouveau modèle, il leur est conseillé d'effectuer un ensemble de tests avec le nouvel équipement, pour vérifier la corrélation de ses indications avec celles obtenues au moyen de tests d'évaluation subjective sur un ensemble approprié d'images ou de séquences de test.

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

Enfin, les informations données dans l'Appendice I aideront les exploitants à choisir le modèle de mesure de la qualité perçue qui répond le mieux à leurs besoins. L'Appendice I est fondé sur le rapport final du VQEG figurant dans la Contribution COM 9-80 de la Commission d'études 9 de l'UIT-T, juin 2000.

L'Appendice I sera mis à jour régulièrement afin de tenir compte des travaux en cours, notamment au sein du VQEG, ainsi que de l'expérience pratique que les participants aux travaux de l'UIT-T pourront acquérir dans l'utilisation d'équipements de mesure de la qualité perçue.

Au fur et à mesure que les méthodes indiquées dans l'Appendice I (ou d'autres méthodes qui pourront être proposées ultérieurement) évolueront, qu'elles seront de plus en plus connues et qu'elles obtiendront de plus en plus de reconnaissance, elles pourront être adoptées comme des sections normatives de la présente Recommandation. 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.

APPENDICE I

Modèles de mesure de la qualité vidéo perçue avec référence complète

I.1 Description des modèles

Les Appendices I à IX décrivent les 8 modèles que le VQEG a validés et qui sont présentés dans le rapport final du VQEG publié en mars 2000. Le texte qui suit contient un bref exposé de ces modèles ainsi qu'une description de l'algorithme PSNR.

I.1.1 Algorithme PSNR

L'algorithme PSNR est défini par les formules suivantes:

$$PSNR = 10 \log_{10}\left(\frac{255^2}{MSE}\right)$$

$$MSE = \frac{1}{(P2-P1+1)(M2-M1+1)(N2-N1+1)} \sum_{p=P1}^{p=P2} \sum_{m=M1}^{m=M2} \sum_{n=N1}^{n=N2} (d(p,m,n)-o(p,m,n))^2$$

où *d(p,m,n)* et *o(p,m,n)* représentent respectivement la valeur de dégradation et la valeur d'origine du pixel situé dans l'image p, à la ligne m et à la colonne n.

NOTE – L'algorithme PSNR nécessite qu'un très haut degré de normalisation soit utilisé avec confiance. La normalisation nécessite un alignement spatial et un alignement temporel ainsi que des corrections pour le gain et le décalage. La méthode de normalisation fait l'objet d'une autre Recommandation (à l'étude).

I.1.2 CPqD

Le modèle du CPqD présenté pour les tests du VQEG a temporairement été appelé CPqD-IES évaluation d'image fondée sur la segmentation (IES, *image evaluation based on segmentation*) version 2.0. La première version de ce système d'évaluation objective de la qualité, CPqD-IES v.1.0, était un système conçu pour faire des prédictions de qualité sur un ensemble de scènes prédéfinies.

Dans le système CPqD-IES v.1.0, l'évaluation de la qualité vidéo est implémentée au moyen de paramètres objectifs fondés sur la segmentation d'image. Les scènes naturelles sont segmentées en régions planes, régions de bord et régions de texture et un ensemble de paramètres objectifs est assigné à chacun de ces contextes. On définit un modèle de perception qui prédit des indices subjectifs en calculant la relation entre des mesures objectives et les résultats de tests d'évaluation subjective, dans le cas d'un ensemble de scènes naturelles traitées par des systèmes de traitement vidéo. Dans ce modèle, la relation entre chaque paramètre objectif et le niveau de dégradation subjectif est approchée par une courbe logistique, d'où il résulte un niveau de dégradation estimé pour chaque paramètre. On obtient le résultat final en procédant à une combinaison des niveaux de dégradation estimés, sur la base de leurs fiabilités statistiques.

Un classificateur de scènes a été ajouté au CPqD-IES v.2.0 afin d'obtenir un système d'évaluation indépendant de la scène. Il utilise des informations spatiales (fondées sur l'analyse DCT) et des informations temporelles (fondées sur les modifications de segmentation) de la séquence d'entrée pour obtenir des paramètres de modèle à partir d'une base de données contenant douze scènes (525/60 Hz) (Appendice II).

I.1.3 Tektronix/Sarnoff

Le modèle soumis par Tektronix/Sarnoff est fondé sur un modèle de discrimination visuelle qui simule les réponses de mécanismes visuels spatio-temporels humains et les amplitudes perçues des différences dans les sorties de mécanisme entre la séquence source et la séquence traitée. A partir de ces différences, on calcule un modèle global relatif à la capacité de discrimination des deux séquences. Le modèle de mesure a été conçu avec, comme contrainte, un fonctionnement à grande

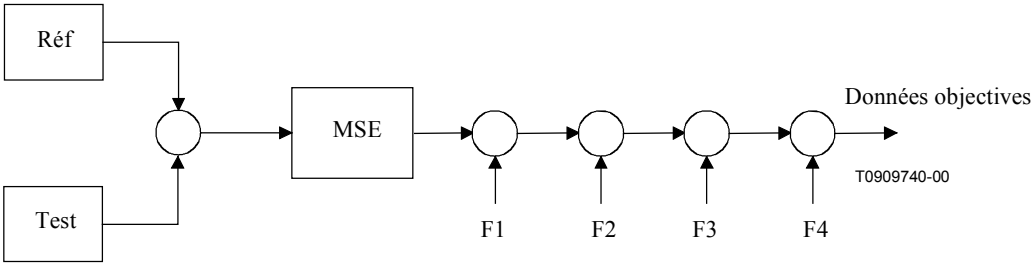
vitesse dans le matériel de traitement d'image standard et représente donc une solution relativement immédiate, avec des calculs simples (Appendice III).

I.1.4 NHK/Mitsubishi Electric Corp.

Le modèle correspond à une émulation des caractéristiques visuelles humaines au moyen de filtres 3D (spatio-temporels), appliqués aux différences entre le signal source et le signal traité. On fait varier les caractéristiques des filtres en fonction du niveau de luminance. La note de qualité en sortie est calculée comme étant une somme pondérée de mesures obtenues avec les filtres. La version matérielle maintenant disponible permet de mesurer la qualité d'image en temps réel et sera utilisée dans divers environnements de diffusion (contrôle en temps réel de signaux diffusés par exemple) (Appendice IV).

I.1.5 KDD

Voir Figure I.1.



F1 filtrage spatial fondé sur les pixels
F2 filtrage fondé sur les blocs (effet du masquage de bruit)
F3 filtrage fondé sur les images (dispersion du point de visualisation)
F4 filtrage fondé sur les séquences (vecteur mouvement + segmentation d'objet, etc.)

Figure I.1/J.144 – Description du modèle

L'erreur quadratique moyenne (MSE, *mean square error*) est calculée en soustrayant le signal de test (test) du signal de référence (réf) et elle est pondérée par des filtres visuels humains F1, F2, F3 et F4.

Le modèle soumis correspond à F1+F2+F4 (Version 2.0, août 1998) (Appendice V).

I.1.6 EPFL

Le modèle de mesure de la distorsion perçue (PDM, *perceptual distortion metric*) soumis par l'EPFL est fondé sur un modèle spatio-temporel du système visuel humain. Il comprend quatre phases, qui s'appliquent aussi bien à la séquence de référence qu'à la séquence traitée. Dans la première phase, l'entrée est convertie en espace de couleurs opposées. Dans la deuxième, on procède à une décomposition spatio-temporelle en canaux visuels distincts de fréquence temporelle, fréquence spatiale et orientation différentes. Dans la troisième, on modélise les effets du masquage de motif en simulant des mécanismes excitateurs et inhibiteurs conformément à un modèle de contrôle du gain de contraste. Dans la quatrième et dernière phase du modèle, qui sert de phase de rassemblement et de détection, on calcule une mesure de distorsion à partir de la différence entre la sortie du capteur de la séquence de référence et celle du capteur de la séquence traitée (Appendice VI).

I.1.7 NASA

Le modèle proposé par la NASA est appelé DVQ qualité vidéo numérique (DVQ, *digital video quality*) et sa version est la version 1.08b. Dans ce modèle de mesure, on tente d'incorporer de nombreux aspects de la sensibilité visuelle humaine dans un algorithme de traitement d'image simple. La simplicité est un but important, étant donné qu'on aimerait que le modèle fonctionne en

temps réel et ne nécessite que des ressources de calcul modestes. L'un des éléments les plus complexes et les plus gourmands en temps des autres modèles proposés correspond aux opérations de filtrage spatial employées pour réaliser les multiples filtres spatiaux passe-bande qui sont caractéristiques de la vision humaine. Nous accélérons cette étape en utilisant la transformée discrète en cosinus (DCT, *discrete cosine transform*) pour la décomposition en canaux spatiaux. Cela constitue un avantage certain parce que des logiciels et des matériels efficaces sont disponibles pour cette transformée et parce que, dans de nombreuses applications, la transformée peut déjà avoir été faite dans le cadre du processus de compression.

L'entrée du modèle est un couple de séquences d'images en couleur: une séquence de référence et une séquence de test. La première phase consiste à appliquer diverses opérations (échantillonnage, recadrage, transformations de la couleur), qui servent à restreindre le traitement à une région particulière et à exprimer les séquences dans un espace de couleurs perçues. Dans cette phase, on procède aussi au désentrelacement et à la dé-gamma-correction du signal vidéo d'entrée. Les séquences sont alors soumises à une subdivision en blocs et à une transformée discrète en cosinus et les résultats sont ensuite transformés en contraste local. Les phases suivantes correspondent au filtrage temporel et spatial et à une opération de masquage de contraste. Enfin, on procède à une sommation des différences masquées sur les dimensions spatiale, temporelle et chromatique pour calculer une mesure de la qualité (Appendice VII).

I.1.8 KPN/Swisscom CT

La mesure de la qualité vidéo perçue (PVQM, *perceptual video quality measure*) mise au point par KPN/Swisscom CT utilise la même approche pour mesurer la qualité vidéo que celle utilisée par la mesure de la qualité vocale perçue (PSQM, *perceptual speech quality measure* [1], Rec. UIT-T P.861 [2]) pour mesurer la qualité vocale. La méthode est censée tenir compte des distorsions spatiales, des distorsions temporelles et des distorsions localisées spatio-temporellement, par exemple celles que l'on rencontre dans les conditions d'erreur. Elle utilise les séquences vidéo au format d'entrée Rec. UIT-R BT.601-5 [3] (entrée et sortie) et les rééchantillonne au format 4:4:4, Y, Cb, Cr. Un alignement de luminance spatio-temporel est inclus dans l'algorithme. Comme les changements généraux de brillance et de contraste n'ont qu'une incidence limitée sur la qualité perçue subjectivement, la mesure PVQM utilise une adaptation spéciale brillance/contraste de la séquence vidéo distordue. La procédure d'alignement spatio-temporel est effectuée dans le cadre d'une sorte de procédure de mise en correspondance de bloc. La partie spatiale de l'analyse de la luminance est fondée sur la détection de bord du signal Y, tandis que la partie temporelle est fondée sur l'analyse d'images différente du signal Y. Chacun sait que le système visuel humain (HVS, *human visual system*) est beaucoup plus sensible à l'acuité de la composante de luminance qu'à celle des composantes de chrominance. En outre, le système HVS a une fonction de sensibilité au contraste qui décroît aux hautes fréquences spatiales. Ces fondements du système HVS sont reflétés dans la première étape de l'algorithme de mesure PVQM qui fournit une approximation au premier ordre des fonctions de sensibilité au contraste des signaux de luminance et de chrominance. Dans la deuxième étape, l'irrégularité de la luminance Y est calculée sous la forme d'une représentation de signal contenant les aspects les plus importants de l'image. Pour obtenir cette irrégularité, on calcule d'abord le gradient local du signal de luminance (au moyen d'un filtrage spatial de type Sobel) dans chaque image puis on fait la moyenne de cette irrégularité dans l'espace et dans le temps. Dans la troisième étape, l'erreur de chrominance est calculée comme une moyenne pondérée de l'erreur de couleur des composantes Cb et Cr avec une dominance de la composante Cr. Dans la dernière étape, les trois différents indicateurs sont traduits sous la forme d'un indicateur de qualité unique, au moyen d'une simple régression linéaire multiple, qui donne une bonne corrélation de la qualité vidéo d'ensemble de la séquence telle qu'elle est perçue subjectivement (Appendice VIII).

I.1.9 NTIA

Ce modèle de mesure de qualité vidéo utilise des caractéristiques de largeur de bande réduites qui sont extraites des régions spatio-temporelles (S-T) des scènes vidéo d'entrée et de sortie traitées. Ces

caractéristiques correspondent aux détails spatiaux, au mouvement et à la couleur qui sont présents dans la séquence vidéo. Les caractéristiques spatiales correspondent à l'activité des contours d'image ou aux gradients spatiaux. Les systèmes vidéo numériques peuvent ajouter des contours (par exemple bruit de contour, subdivision en blocs) ou réduire les contours (par exemple flou). Les caractéristiques temporelles correspondent à l'activité des différences temporelles ou aux gradients temporels entre des images successives. Les systèmes vidéo numériques peuvent ajouter un mouvement (par exemple blocs d'erreur) ou réduire le mouvement (par exemple répétitions d'image). Les caractéristiques de chrominance correspondent à l'activité de l'information de couleur. Les systèmes vidéo numériques peuvent ajouter une information de couleur (par exemple couleur croisée) ou réduire l'information de couleur (par exemple sous-échantillonage de la couleur). Pour calculer les paramètres de gain et de perte, on compare deux flux parallèles d'échantillons de caractéristiques, l'un provenant de l'entrée et l'autre de la sortie. Les paramètres de gain et de perte sont examinés séparément pour chaque couple de flux de caractéristiques car ils mesurent fondamentalement des aspects différents de la perception de la qualité. Les fonctions de comparaison de caractéristiques utilisées pour calculer le gain et la perte tentent de simuler la perceptibilité de dégradations en modélisant des seuils de perceptibilité, un masquage visuel et un rassemblement des erreurs. On utilise une combinaison linéaire des paramètres pour estimer l'indice de qualité subjectif (Appendice IX).

I.2 Références

[1] BEERENDS (J.G.), STEMERDINK (J.A.): A perceptual speech quality measure based on a psychoacoustic sound representation, *J. Audio Eng. Soc*. 42, 115-123, 1994.

[2] UIT-T P.861 (1998), *Mesure objective de la qualité des codecs vocaux fonctionnant en bande téléphonique* (300-3400 Hz).

[3] 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)*.

APPENDICE II

CPqD

Evaluation de la qualité vidéo en utilisant des paramètres objectifs fondés sur la segmentation de l'image

Résumé

Le présent appendice présente une méthode d'évaluation de la qualité vidéo utilisant des paramètres objectifs fondés sur la segmentation de l'image. Les scènes naturelles sont segmentées en régions planes, régions de bord et régions de texture et un ensemble de paramètres objectifs sont assignés à chacun de ces contextes. On définit un modèle de perception qui prédit des indices subjectifs en calculant la relation entre des mesures objectives et les résultats de tests d'évaluation subjective, dans le cas d'un ensemble de scènes naturelles et de codecs vidéo MPEG-2. Dans ce modèle, la relation entre chaque paramètre objectif et le niveau de dégradation subjectif est approchée par une courbe logistique, d'où il résulte un niveau de dégradation estimé pour chaque paramètre. On obtient le résultat final en procédant à une combinaison linéaire des niveaux de dégradation estimés, dans laquelle le poids de chaque niveau de dégradation est proportionnel à sa fiabilité statistique. Les résultats présentés dans le présent appendice montrent que l'utilisation de mesures objectives fondées sur des régions permet d'obtenir des prédictions plus précises que celles qui sont fondées sur des paramètres généraux.

II.1 Introduction

L'évaluation de la qualité vidéo est devenue une question cruciale, compte tenu de l'utilisation croissante de systèmes de compression vidéo numérique et des services vidéo qui en découlent, tels que la distribution primaire et secondaire de télévision numérique, la vidéo à la demande, le visiophone, la visioconférence, etc. En raison de la souplesse des normes sur le codage vidéo, les différents codecs n'offrent pas la même qualité d'image. Par conséquent, les méthodes d'évaluation de la qualité vidéo représentent des outils importants pour comparer la qualité vidéo de différents codecs et pour quantifier leur performance dans un grand nombre d'applications.

La difficulté rencontrée dans l'élaboration de techniques permettant d'évaluer la qualité de systèmes de compression vidéo provient en partie du fait que les algorithmes de compression introduisent des dégradations vidéo qui sont fortement dépendantes des niveaux de détail et du mouvement dans les scènes. De plus, la perception visuelle des dégradations vidéo dépend aussi des détails et du mouvement dans les scènes. Les méthodes d'évaluation classiques, qui sont fondées sur des signaux de test statiques, ne sont donc pas adaptées pour quantifier la performance des systèmes de compression vidéo.

Dans le présent appendice, on présente une méthode d'évaluation de la qualité vidéo à utiliser lorsque la vidéo est traitée par des systèmes de transmission unidirectionnels qui utilisent des interfaces numériques et, idéalement, des moyens de transport numériques. On a appliqué la méthode pour évaluer des systèmes de compression vidéo conformes à la norme MPEG [1] et [2], mais on pourrait aussi l'utiliser pour évaluer d'autres types de systèmes, tels que les codecs vidéo fondés sur d'autres techniques d'analyse (c'est-à-dire vaguelettes et filtres de prédiction) et les codeurs/décodeurs de signaux composites.

La Figure II.1 montre la configuration du processus de calcul des paramètres objectifs utilisé pour l'évaluation de la qualité vidéo. Le format de fichier des signaux vidéo numériques d'entrée et de sortie est YCbCr4:2:2, comme défini dans la Rec. UIT-R BT.601-5 [3].

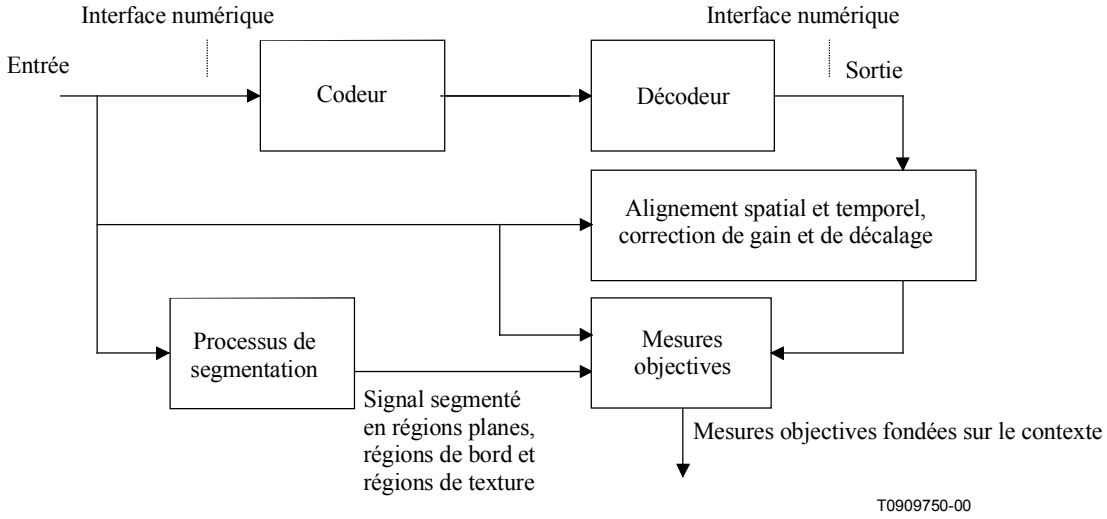


Figure II.1/J.144 – Calcul des paramètres objectifs

Sur la Figure II.1, chaque paramètre objectif est calculé séparément dans les contextes suivants des scènes: régions planes, régions de bord et régions de texture. Il s'agit de l'un des aspects les plus importants de cette méthode. Une distorsion liée à la subdivision en blocs, par exemple, peut être mesurée par un détecteur de bord appliqué aux régions planes de la scène vidéo, dans lesquelles la perception visuelle de cette distorsion est plus nette. On réduit la complexité de calcul de la méthode en utilisant des estimateurs à faible complexité et en restreignant le calcul aux contextes correspondants des scènes. Ces contextes sont définis par un algorithme de segmentation d'image qui

est appliqué aux scènes naturelles d'origine (c'est-à-dire au signal test d'entrée). Ce type d'algorithme nécessite normalement des calculs très complexes, il n'est toutefois exécuté qu'une seule fois. Il est à noter que l'alignement spatial et temporel entre les signaux vidéo d'entrée et de sortie et la correction de gain et de décalage sont également nécessaires pour pouvoir calculer les paramètres objectifs correctement. On trouvera des informations sur l'alignement et l'étalonnage dans la référence [4].

Pour calculer les paramètres objectifs, on procède à une comparaison directe entre les scènes d'origine et les scènes dégradées. Tous les estimateurs sont appliqués aux trames et non aux images vidéo afin de garantir la fiabilité statistique des mesures dans les scènes comportant beaucoup de mouvements.

On définit un modèle de perception qui prédit des indices subjectifs en calculant la relation entre des mesures objectives et les résultats de tests d'évaluation subjective, dans le cas d'un ensemble de scènes naturelles et de codecs vidéo MPEG-2. Ces modèles de perception dépendants des scènes sont définis en deux étapes comme suit:

1) la relation entre chaque paramètre et le niveau de dégradation subjectif est approchée par une courbe logistique, d'où il résulte un niveau de dégradation estimé pour chaque paramètre;
2) on obtient le résultat final en procédant à une combinaison linéaire des niveaux de dégradation estimés, dans laquelle le poids de chaque niveau de dégradation est proportionnel à sa fiabilité statistique.

Les détails de la méthode dont on vient de donner un aperçu sont présentés dans les paragraphes qui suivent. Le paragraphe II.2 contient une brève description des conditions adoptées pour les tests d'évaluation subjective. Les méthodes permettant de déterminer les paramètres objectifs et de segmenter les scènes naturelles sont décrites au paragraphe II.3. Dans le paragraphe II.4, on introduit les modèles de perception pour l'évaluation subjective de la qualité et on rapporte les résultats qui ont été obtenus dans cette étude. Le paragraphe II.4 indique également les avantages liés à l'utilisation de paramètres objectifs fondés sur la segmentation d'image pour l'évaluation subjective de la qualité et on présente la dépendance des modèles de perception vis-à-vis de la catégorie d'évaluateurs et vis-à-vis de la distance de visualisation par rapport au moniteur (4H ou 6H). Les conclusions de cette contribution sont présentés au paragraphe II.5.

II.2 Tests d'évaluation subjective

Le laboratoire de traitement d'image du centre CPqD/TELEBRÁS (Centre brésilien de recherche et développement sur les télécommunications) possède une salle spéciale pour les essais d'évaluation subjective, conformément à la Rec. UIT-R BT.500-7 [5]. On a utilisé cette salle pour évaluer la performance de certains codecs vidéo MPEG-2 matériels et logiciels sur un sous-ensemble des scènes naturelles proposées dans la Rec. UIT-R BT.802-1 [6]. Les codecs MPEG-2 matériels ont été fournis par TV Globo (une société de télévision brésilienne). Les scènes ont également été traitées par le logiciel de codage MPEG-2 disponible au centre CPqD/TELEBRÁS.

On décrit brièvement ci-après les conditions utilisées pour l'évaluation subjective des codecs vidéo MPEG-2 précités.

II.2.1 Sessions d'évaluation subjective

Le Tableau II.1 présente un résumé des conditions utilisées pour les tests d'évaluation subjective.

Tableau II.1/J.144 – Conditions relatives aux tests d'évaluation subjective

Conditions pour l'évaluation	Conformément au § 2.1 de la Rec. UIT-R BT.500-7 [5]
Source des signaux	Magnétoscope D1
Moniteur	Moniteur de studio de 20" avec interface numérique
Distances de visualisation	4H et 6H
Méthode d'évaluation	Méthode à double stimulus utilisant une échelle de dégradation (DSIS, *double-stimulus impairment scale*) avec neuf points entre 1 et 5 [5].
Séquences de test	5 scènes de télévision numérique à définition classique (voir II.2.2)
Durée de présentation	10 secondes (signal d'origine) + 3 secondes (signal de gris) + 10 secondes (signal à évaluer) + 5 secondes pour le vote, comme proposé sur la Figure 3.a de la Rec. UIT-R BT.500-7 [5]
Evaluateurs	14 spécialistes et 34 profanes
Evaluateurs par session	5
Sessions par évaluateur	2
Présentations par session	48
Eléments évalués	Voir II.2.3
Présentation des résultats	Moyenne et écart-type du niveau de dégradation par rapport au signal de référence (scène d'origine) Elimination de notes et d'évaluateurs comme proposé dans la Rec. UIT-R BT.500-7 [5]

II.2.2 Scènes naturelles

Dans les sessions d'évaluation subjective, on a utilisé un ensemble de cinq scènes naturelles (voir Tableau II.2), qui sont définies comme étant des séquences de test pour la télévision classique dans la Rec. UIT-R BT.802-1 [6]:

Tableau II.2/J.144 – Scènes naturelles utilisées pour l'évaluation subjective

Nom de la séquence	Numéro de la scène dans la Rec. UIT-R 802-1
Jardin fleuri	15
Mobile et calendrier	30
Tennis de table	29
Diva avec bruit	17
Port de Kiel 4	26

II.2.3 Systèmes testés

Au total, 26 systèmes ont été inclus dans les sessions d'évaluation subjective. Il sont présentés dans le Tableau II.3.

Tableau II.3/J.144 – Systèmes testés

Groupe	Type	Caractéristiques	Systèmes testés
1	Codec MP@ML MPEG-2 matériel, pour des applications à débit constant (CBR, *constant bit rate*)	Débits: 5, 10 et 15 Mbits/s N = 12 et M = 2	6
2	Codec MP@ML MPEG-2 logiciel, pour des applications CBR	Débits: 2,5; 5; 7,5; 10; 12 et 15 Mbits/s N = 12 et M = 1 et 2	12
3	Codec 422P@MPL MPEG-2 matériel, pour des applications CBR	Débit: 18 Mbits/s N = 2 et M = 2	1
4	Codec MP@ML MPEG-2 logiciel, pour des applications à débit variable (VBR, *variable bit rate*) utilisant uniquement le codage intra-image	Echelle de quantificateur fixe [2] en 4, 8, 16, 32 et 62	5
5	Conversion de signaux composites	NTSC et PAL-M	2

II.3 Mesures objectives fondées sur le contexte

Le présent paragraphe décrit les données vidéo utilisées pour l'évaluation objective (c'est-à-dire les données utilisées pour le calcul des paramètres objectifs – voir II.3.1), propose trois méthodes de segmentation d'image qui peuvent être utilisées pour subdiviser les données vidéo en régions planes, régions de bord et régions de texture (voir II.3.2) et présente les paramètres objectifs qui ont été adoptés dans cette étude (voir II.3.3).

II.3.1 Données vidéo utilisées pour l'évaluation objective

Les données vidéo utilisées pour l'évaluation objective sont constituées d'une séquence vidéo de 17 secondes, composée de dix clips de scènes naturelles et de deux clips de signaux de test artificiels.

Cinq clips de scènes naturelles, de 2 secondes chacun, ont été choisis parmi les scènes naturelles présentées au II.2.2. On a utilisé des clips de 2 secondes et non des clips de 10 secondes, comme dans les tests subjectifs décrits au paragraphe II.2, afin de réduire la complexité du calcul du processus d'évaluation objective. Le choix des clips de 2 secondes était fondé sur les critères suivants:

- le clip de 2 secondes d'une scène donnée représente un segment critique des 10 secondes de données par rapport au "caractère critique" moyen de la scène. Ce caractère critique a été défini comme étant le nombre de bits par image résultant du processus de codage d'un codec MP@ML MPEG-2 (N = 12 et M = 2) avec débit variable et échelle de quantificateur égale à 16;
- ce clip de 2 secondes représente aussi un segment critique du point de vue subjectif, lorsque la scène est traitée par un codec MP@ML MPEG-2 (N = 12 et M = 2) à 5 Mbits/s.

Les cinq autres clips, de 1 seconde chacun, sont constitués de scènes comportant peu ou pas de mouvement. Ces scènes ont été utilisées dans le processus d'évaluation objective, en intercalation avec les précédents clips de 2 secondes, afin de tester le comportement adaptatif des codecs vidéo MPEG-2 (c'est-à-dire le comportement concernant le débit et le contrôle de qualité, la performance en régime et après transition de scène). Elles sont également spécifiées dans la Rec. UIT-R BT.802-1 [6]. Bien que cela n'entre pas dans le cadre de cette contribution, il est important de signaler que la détermination de la variation de performance (dispersion du rapport signal sur bruit) après chaque transition de scène et en régime (différence de performance sur les images I, P et B) a été utilisée pour caractériser le comportement dynamique des codecs vidéo MPEG-2 matériels.

Les signaux de test artificiels sont (1 seconde chacun):

- bruit à bande étroite [4] – Signal vidéo statique et trichromatique défini par du bruit avec une résolution d'environ 1/25 de la limite de Nyquist et avec un histogramme approximativement uniforme pour chacune des composantes Y, Cb et Cr;
- zone plate circulaire [4] – Signal vidéo statique et trichromatique défini par une séquence sinusoïdale pour les composantes Y, Cb et Cr, avec des fréquences horizontale et verticale constantes respectivement le long de la même colonne et le long de la même ligne d'une trame vidéo donnée et des fréquences croissantes à partir du centre de l'image vers l'extérieur.

On a utilisé ces signaux artificiels pour déterminer les paramètres suivants:

- déplacement des signaux vidéo actifs;
- zone vidéo active;
- gain et décalage;
- réponse fréquentielle 2D;
- déplacement entre la chrominance et la luminance (un déplacement vertical entre ces composantes a été observé très souvent dans les systèmes MP@ML MPEG-2 matériels, en raison des conversions YCbCr4:2:0 ⇔ YCbCr4:2:2, d'où la création d'une auréole de chromaticité parasite sur les bords du signal de sortie).

Les données de test successives de 17 secondes sont décrites dans le Tableau II.4:

Tableau II.4/J.144 – Données de test pour l'évaluation objective

Code horaire (mm:ss:ff)	Scène	Nom abrégé	Caractéristique temporelle	Durée (secondes)
00:00:00	Bruit à bande étroite	Bruit	Statique	1
00:01:00	Jardin fleuri	Jardin	Dynamique	2
00:03:00	Arbre	Arbre	Statique	1
00:04:00	Mobile et calendrier	Mobile	Dynamique	2
00:06:00	Clown	Clown	Statique	1
00:07:00	Tennis de table	Tennis	Dynamique	2
00:09:00	Pelotes de laine	Pelotes	Dynamique	1
00:10:00	Diva avec bruit	Diva	Dynamique	2
00:12:00	Garçon avec jouets	Garçon	Statique	1
00:13:00	Port de Kiel 4	Kiel	Dynamique	2
00:15:00	Jeune couple	Couple	Statique	1
00:16:00	Zone plate circulaire	Zone plate	Statique	1

II.3.2 Segmentation spatiale

Trois algorithmes ont été élaborés pour la segmentation d'image [7]. Le premier est un algorithme de segmentation d'image fondée sur la détection de bord au moyen d'un filtrage récurrent (voir II.3.2.1), le deuxième est un algorithme de segmentation d'image floue fondée sur des fonctions spatiales (voir II.3.2.2) et le troisième est un algorithme de segmentation d'image fondée sur l'algorithme du bassin versant (voir II.3.2.3). Les résultats de l'évaluation objective faite au moyen de ces algorithmes sont discutés au II.4.3. Dans ces algorithmes de segmentation, la stratégie consiste à classer la composante de luminance de chaque trame vidéo dans l'un des trois contextes

mutuellement exclusifs suivants: régions planes, régions de bord et régions de texture. Ces algorithmes sont brièvement décrits ci-après:

II.3.2.1 Algorithme I: segmentation d'image fondée sur la détection de bord au moyen d'un filtrage récurrent

Au départ, cet algorithme classe chaque pixel comme appartenant ou n'appartenant pas aux régions planes de l'image, sur la base de la variance de brillance calculée au voisinage du pixel. L'image binaire résultante est alors lissée par un filtre médian [7]. L'algorithme applique aussi à l'image d'origine un détecteur de bord fondé sur un filtrage récurrent. Les bords à la frontière des régions planes sont classés comme appartenant aux régions de bord. Les régions de texture sont les autres régions de l'image.

A titre d'exemple, la Figure II.2 montre une partie de la scène Mobile et calendrier. Le résultat de la segmentation par l'algorithme I de cette partie est représenté sur la Figure II.3. Il est à noter que les régions planes sont représentées par des pixels blancs, les régions de bord par des pixels gris et les régions de texture par des pixels noirs.

Figure II.2/J.144 – Partie de Mobile et calendrier

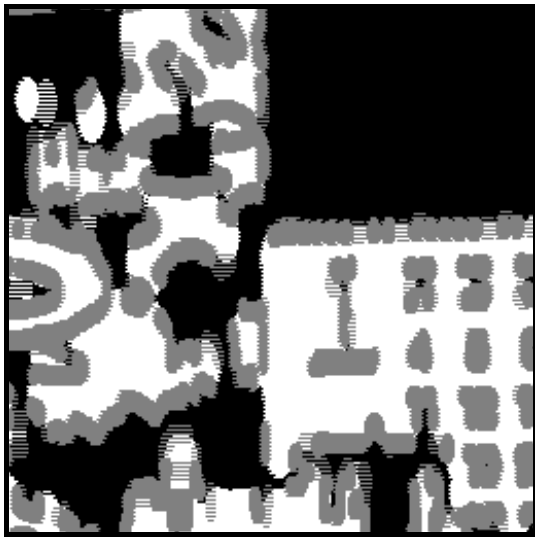


Figure II.3/J.144 – Résultat de la segmentation

II.3.2.2 Algorithme II: segmentation d'image floue fondée sur des fonctions spatiales

Cet algorithme comprend deux étapes. Dans la première étape, l'algorithme assigne une fonction de membre, définie dans l'intervalle [0, 1], à chacun des trois contextes de la classification. Dans la fonction de membre des régions planes, la valeur de membre d'un pixel est définie comme étant inversement proportionnelle à la variance de brillance calculée au voisinage du pixel. Le gradient morphologique [8] appliqué à cette fonction définit la fonction de membre des régions de bord. Le complément de l'union floue [9] de ces deux fonctions de membre définit la fonction de membre des régions de texture. Dans la deuxième étape, chaque pixel est classé comme appartenant au contexte avec la valeur la plus élevée de membre parmi ses trois valeurs de membre déterminées dans l'étape précédente.

II.3.2.3 Algorithme III: algorithme de segmentation d'image fondée sur l'algorithme du bassin versant

Dans cet algorithme, on commence par simplifier la composante de luminance en augmentant ses régions homogènes par l'application d'un filtre de lissage préservant les bords [10]. Ensuite, on applique un algorithme du bassin versant au gradient morphologique de l'image simplifiée. Cet

algorithme détecte les régions homogènes, appelées bassins versants, avec un contraste relatif minimal spécifié. Les régions planes sont les bassins versants dont la surface est supérieure à un seuil. Les régions de texture sont données par l'érosion du complément des régions planes. Les régions de bord sont les autres régions de ce processus.

II.3.3 Paramètres objectifs

Les paramètres objectifs sont obtenus pour chaque contexte (plan, bord et texture) et à partir des échantillons de luminance et de chrominance des signaux d'entrée (Y_{ref}, Cb_{ref} et Cr_{ref}) et de sortie (Y_{dec}, Cb_{dec} et Cr_{dec}), après alignement spatial et temporel et correction de gain et de décalage, comme indiqué sur la Figure II.1. Les mesures et le processus sous-jacent permettant de les calculer sont décrits ci-dessous:

- erreur quadratique moyenne (MSE, *mean square error*);
- différence de Sobel positive (PSD, *positive Sobel difference*);
- différence de Sobel négative (NSD, *negative Sobel difference*);
- différence de Sobel absolue (ASD, *absolute Sobel difference*).

Soit X(i,j) le j^e pixel de la ligne du signal d'entrée, Z(i,j) le j^e pixel de la i^e ligne du signal de sortie et les éléments $X_m(i,j)$ et $Z_m(i,j)$ le pixel du signal d'entrée et le pixel du signal de sortie après un filtrage médian.

Le calcul de l'erreur MSE dans un contexte R (plan, bord ou texture) est défini par la valeur médiane de:

$$SE(i,j) = [\ X(i,j) - Z(i,j)\ ]^2, \text{ où } (i,j) \in R.$$

Le calcul de la différence PSD dans un contexte R (plan, bord ou texture) est défini par la valeur médiane de:

$$PS(i,j) = \max [\ sobel(X_m(i,j)) - sobel(Z_m(i,j))\ ,\ 0\ ], \text{ où } (i,j) \in R.$$

Le calcul de la différence NSD dans un contexte R (plan, bord ou texture) est défini par la valeur moyenne de:

$$NS(i,j) = -\max [\ sobel(Z_m(i,j)) - sobel(X_m(i,j))\ ,\ 0\ ], \text{ où } (i,j) \in R.$$

Le calcul de la différence ASD dans un contexte R (plan, bord ou texture) est défini par la valeur médiane de:

$$AS(i,j) = |\ sobel(X_m(i,j)) - sobel(Z_m(i,j))\ |, \text{ où } (i,j) \in R.$$

Autrement dit, ASD = PSD + NSD.

Les paramètres objectifs utilisés pour l'évaluation subjective de la qualité correspondent à la valeur moyenne des mesures précitées calculées sur un ensemble de 2N trames de la portion finale (c'est-à-dire en régime) de chacun des cinq clips indiqués dans le Tableau II.4 et appartenant aux scènes soumises à l'évaluation subjective. La valeur N est un multiple de l'intervalle entre images codées intra-image (type I), c'est-à-dire que c'est un multiple de la longueur d'un groupe d'images [1, 2]. Afin que cette condition soit satisfaite pour tous les systèmes définis dans le Tableau II.3, on choisit N = 12.

II.4 Evaluation subjective de la qualité

Le présent paragraphe décrit la façon dont les modèles d'évaluation subjective de la qualité sont définis pour chaque scène. Le paragraphe II.4.1 décrit un modèle de perception pour évaluer le niveau de dégradation subjectif sur la base d'un seul paramètre. On combine linéairement les résultats de cette approximation pour chaque paramètre objectif pour évaluer le niveau de

dégradation subjectif final. Le modèle de prédiction linéaire est présenté au paragraphe II.4.2. Le paragraphe II.4.3 contient une présentation et une discussion des résultats de cette étude.

II.4.1 Evaluation subjective de la qualité fondée sur un seul paramètre: approximation logistique

Pour chaque scène, la relation entre chaque paramètre objectif D et le résultat subjectif U est défini au départ comme suit.

Un niveau de dégradation normalisé compris entre 0% et 100% est défini dans la référence [5] comme étant:

$$d = (U_{\max} - U)/(U_{\max} - U_{\min}) \times 100\%$$

La relation entre d et chaque paramètre objectif D est approchée par la fonction logistique suivante [5]:

$$\underline{d} = \frac{1}{1 + \left(\frac{D_M}{D}\right)^G} \times 100\%$$

où les valeurs D_M et G sont calculées de manière à minimiser l'erreur quadratique moyenne:

$$e = E\left[\{d - \underline{d}\}^2\right]$$

pour chaque scène et pour chaque paramètre objectif séparément. La fiabilité statistique de $\underline{d}$ est définie par 1/e.

II.4.2 Evaluation subjective de la qualité: prédiction linéaire en 3 étapes

L'évaluation du niveau de dégradation normalisé d par un ensemble de niveaux de dégradation estimés $\underline{d}$ (un par paramètre comme défini au II.4.1) est mise en œuvre en trois étapes de prédiction linéaire comme décrit ci-dessous.

Etape 1

Considérons d'abord les ensembles suivants de niveaux de dégradation estimés qui sont choisis pour la composante de luminance:

- $\underline{d}^{MSE}$
- $\underline{d}^{PSD}$ et $\underline{d}^{NSD}$
- $\underline{d}^{ASD}$
- $\underline{d}^{MSE}$, $\underline{d}^{PSD}$ et $\underline{d}^{NSD}$
- $\underline{d}^{MSE}$ et $\underline{d}^{ASD}$

Pour une scène donnée et un contexte donné de cette scène (plan, bord ou texture), le meilleur ensemble est celui qui est associé à la plus faible erreur de prédiction. En utilisant ce critère pour choisir un ensemble de niveaux de dégradation estimés pour chaque contexte, on procède, dans cette étape, à la combinaison linéaire des niveaux de dégradation de chaque ensemble sélectionné, ce qui donne en résultat trois valeurs d'estimation (une par contexte) désignées par: $\underline{d}_{YP}$, $\underline{d}_{YE}$ et $\underline{d}_{YT}$.

De même, les ensembles considérés de niveaux de dégradation estimés pour les composantes de chrominance de la scène sont les suivants:

- $\underline{d}^{MSE(Cb)}$ et $\underline{d}^{MSE(Cr)}$
- $\underline{d}^{ASD(Cb)}$ et $\underline{d}^{ASD(Cr)}$

et les trois valeurs d'estimation résultantes (une par contexte) sont désignées par: $\underline{d}_{CP}$, $\underline{d}_{CE}$ e $\underline{d}_{CT}$.

Etape 2

Les valeurs d'estimation $\underline{d}_P$, $\underline{d}_E$ et $\underline{d}_T$ résultent d'une prédiction linéaire fondée respectivement sur les vecteurs ($\underline{d}_{YP}$, $\underline{d}_{CP}$), ($\underline{d}_{YE}$, $\underline{d}_{CE}$) et ($\underline{d}_{YT}$, $\underline{d}_{CT}$).

Etape 3

Les valeurs d'estimation $\underline{d}_P$, $\underline{d}_E$ et $\underline{d}_T$ sont combinées par prédiction linéaire pour produire le niveau de dégradation estimé $\underline{d}$.

Dans toutes les étapes ci-dessus, les prédicteurs satisfont aux restrictions suivantes.

Soit ($\underline{d}_1, \underline{d}_2, ..., \underline{d}_P$) le vecteur d'entrée du prédicteur linéaire. La sortie $\underline{d}_o$ est donnée par:

$$\underline{d}_o = \sum a_i \underline{d}_i$$

où les poids $\{a_i\}$ sont calculés de manière à minimiser l'erreur quadratique moyenne:

$$E\left[\{d - \underline{d}_o\}^2\right], \text{ avec}$$

$$\sum a_i = 1 \text{ et}$$

$$a_i / a_k = e_k / e_i$$

où la fiabilité statistique de $\underline{d}_i$ est $1/e_i$, comme défini au II.4.1.

On a observé qu'une prédiction de ce type était plus robuste que celle obtenue par des prédicteurs optimaux, car elle dépend moins de la base de données de conditionnement. Elle permet d'obtenir de meilleurs résultats lorsqu'elle est appliquée à des bases de données de test, comme le montre l'exemple du II 4.3.

II.4.3 Evaluation subjective de la qualité: présentation et discussion des résultats

Le présent paragraphe est subdivisé en trois principaux sujets. Au II.4.3.1, on décrit les résultats et les modèles de perception obtenus par l'évaluation subjective de la qualité fondée sur l'algorithme I (l'algorithme de segmentation d'image précédemment décrit au II.3.2.1). Au § II.4.3.1 on présente également la dépendance entre les modèles de perception et la catégorie d'évaluateurs (spécialistes et profanes) ainsi qu'entre les modèles de perception et la distance de visualisation par rapport au moniteur (4H et 6H). La variation de la précision de l'estimation en fonction des algorithmes de segmentation d'image est discutée au II.4.3.2. Au II.4.3.3, on indique les avantages de la méthode d'évaluation subjective proposée par rapport aux autres méthodes qui sont fondées sur des mesures générales ou sur une prédiction optimale.

II.4.3.1 Résultats: modèles de perception et performance

Le Tableau II.5 présente les résultats de la méthode d'évaluation subjective fondée sur l'algorithme I (voir II.3.2.1) pour la segmentation des scènes suivantes: jardin, mobile, tennis, diva et Kiel. Dans le Tableau II.5:

- les poids de la prédiction linéaire décrite à l'étape 2 du II.4.2 sont équivalents aux poids subjectifs relatifs des dégradations de la luminance (Y) et de la chrominance (C) dans les régions planes, les régions de bord et les régions de texture. La valeur moyenne globale calculée sur toutes les scènes est donnée à la dernière ligne de ce tableau.
- Les poids de la prédiction linéaire décrite à l'étape 3 du II.4.2 sont équivalents aux poids subjectifs relatifs de la dégradation dans les régions planes (P), les régions de bord (E) et les régions de texture (T). La valeur moyenne globale calculée sur toutes les scènes est donnée à la dernière ligne du tableau.
- L'erreur quadratique moyenne (MSE) et l'erreur absolue moyenne (MAE, *mean absolute error*) entre le niveau de dégradation normalisé d et le niveau de dégradation estimé $\underline{d}$, compte tenu d'une échelle de normalisation comprise entre 0% et 100%, sont indiquées dans les deux dernières colonnes de ce tableau. L'erreur entre le niveau moyen normalisé de dégradation et le niveau moyen estimé de dégradation, calculée sur toutes les scènes, est donné à la dernière ligne de ces colonnes.

Les résultats présentés dans le Tableau II.5 correspondent aux modèles de perception obtenus à partir des notes subjectives des 34 évaluateurs profanes du Tableau II.1 et des 26 systèmes testés du Tableau II.3.

Tableau II.5/J.144 – Modèles de perception et résultats: évaluateurs profanes

	Etape 2: plan		**Etape 2: bord**		**Etape 2: texture**		**Etape 3**			**Erreur**	
Scène	**Y(%)**	**C(%)**	**Y(%)**	**C(%)**	**Y(%)**	**C(%)**	**P(%)**	**E(%)**	**T(%)**	**MSE**	**MAE**
Jardin	61	39	70	30	37	63	13	37	51	18,1	3,0
Mobile	74	26	75	25	63	37	83	7	9	24,2	3,6
Tennis	67	33	65	35	70	30	45	13	42	25,3	3,5
Diva	49	51	92	8	42	58	27	59	14	5,4	1,5
Kiel	62	38	66	34	40	60	32	39	29	22,7	3,6
Global	**63**	**37**	**73**	**27**	**50**	**50**	**40**	**31**	**29**	**6,2**	**1,8**

Les Tableaux II.6 et II.7 montrent la dépendance entre les modèles de perception et les résultats pour:

- des évaluateurs profanes et des évaluateurs spécialistes;
- deux distances de visualisation (4H et 6H) par rapport au moniteur (chaque fois avec 50% du nombre total d'évaluateurs).

Tableau II.6/J.144 – Modèles de perception et résultats: évaluateurs profanes et évaluateurs spécialistes

	Evaluateurs profanes						Evaluateurs spécialistes					
	Région			Composante		Erreur	Région			Composante		Erreur
Scène	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**
Jardin	13	37	51	52	48	18,1	12	53	34	51	49	23,5
Mobile	83	7	9	73	27	24,2	72	13	15	72	28	73,4
Tennis	45	13	42	68	32	25,3	47	12	41	70	30	48,1
Diva	27	59	14	73	27	5,4	22	42	36	55	45	21,2
Kiel	32	39	29	57	43	22,7	43	36	21	47	53	44,1
Global	**40**	**31**	**29**	**62**	**38**	**6,2**	**39**	**31**	**30**	**58**	**42**	**12,1**

Tableau II.7/J.144 – Modèles de perception et résultats: distances de visualisation de 6H et de 4H

	Distance de visualisation de 6H						Distance de visualisation de 4H					
	Région			Composante		Erreur	Région			Composante		Erreur
Scène	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**
Jardin	15	40	45	51	49	20,0	9	35	55	47	53	16,8
Mobile	83	8	9	77	23	24,7	71	13	16	62	38	59,4
Tennis	47	15	38	60	40	40,9	42	10	48	74	26	21,4
Diva	41	36	22	64	36	15,5	21	46	33	54	46	12,5
Kiel	34	40	26	54	46	18,3	31	46	23	57	43	26,0
Global	**44**	**28**	**28**	**61**	**39**	**7,9**	**35**	**30**	**35**	**59**	**41**	**9,6**

Les résultats présentés dans les Tableaux II.5, II.6 et II.7 sont commentés ci-dessous:

- l'évaluation subjective de la qualité en utilisant des paramètres objectifs fondés sur la segmentation d'image et calculés pour les 26 systèmes décrits au II.2.3, donne une erreur absolue moyenne (MAE) inférieure à 4% pour chacune des scènes et une erreur MAE globale de 1,8%, dans le cas d'évaluateurs profanes;
- si on compare les modèles de perception sur la base de l'opinion des évaluateurs spécialistes et de celle des évaluateurs profanes, le poids des dégradations de la chrominance est légèrement supérieur dans le cas des évaluateurs spécialistes;
- si on compare les modèles de perception sur la base des distances de visualisation de 4H et de 6H, le poids des dégradations dans les régions de bord et dans les régions de texture est nettement supérieur dans le cas de la distance de visualisation de 4H, comme on pouvait s'y attendre.

II.4.3.2 Variation de la précision de l'évaluation en fonction de l'algorithme de segmentation d'image

Les résultats de l'évaluation subjective de la qualité, fondés sur les algorithmes II et III (brièvement décrits au II.3.2) et obtenus à partir des notes des 34 évaluateurs profanes, sont montrés dans le Tableau II.8. Si on compare les résultats de ce tableau avec les résultats présentés précédemment

côté gauche du Tableau II.6 (pour l'algorithme I), on remarque que la précision de l'évaluation présente de faibles variations pour une scène donnée en fonction de l'algorithme de segmentation d'image. En revanche, il n'y a pas de variation significative dans la précision d'évaluation globale si on considère les trois algorithmes de segmentation d'image. Cela implique que des algorithmes de segmentation d'image encore plus simples peuvent donner des résultats satisfaisants.

Tableau II.8/J.144 – Modèles de perception et résultats: algorithmes II et III

	Algorithme II						**Algorithme III**					
	Région			**Composante**		**Erreur**	**Région**			**Composante**		**Erreur**
Scène	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**
Jardin	9	32	59	53	47	18,0	10	44	46	53	47	15,8
Mobile	65	26	9	59	41	20,7	82	11	6	60	40	18,7
Tennis	54	27	19	70	30	28,5	68	21	11	72	28	31,3
Diva	25	50	24	75	25	7,1	50	31	19	60	40	7,4
Kiel	23	31	46	64	36	25,9	28	33	38	59	41	22,4
Global	**35**	**33**	**31**	**66**	**34**	**7,4**	**48**	**28**	**24**	**63**	**37**	**6,5**

II.4.3.3 Avantages de la méthode d'évaluation subjective de la qualité adoptée

L'exemple illustré dans le Tableau II.9 porte sur deux propriétés très importantes de toute méthode d'évaluation subjective de la qualité fondée sur des paramètres objectifs: précision et robustesse [11] à [14]. Dans cet exemple, on compare la méthode d'évaluation subjective de la qualité adoptée, qui utilise des mesures objectives fondées sur le contexte et la méthode de prédiction linéaire décrite au II.4.2, avec les méthodes suivantes:

- une méthode qui utilise les mêmes mesures objectives fondées sur le contexte mais qui emploie la prédiction optimale;
- une méthode qui emploie la prédiction linéaire décrite au II.4.2 mais qui utilise des mesures globales.

Dans l'exemple, on a utilisé les systèmes du Groupe 2 et le système NTSC du Groupe 5 pour la base de données de conditionnement et les systèmes du Groupe 1 et le système PAL-M du Groupe 5 pour la base de données de test (voir Tableau II.3). Les paramètres objectifs utilisés dans cet exemple correspondent à ceux fondés sur l'erreur MSE et décrits au II.3.3. Les résultats ont été obtenus à partir des notes des évaluateurs profanes. Les valeurs d'entrée du tableau sont des erreurs quadratiques moyennes de prédiction. La dernière ligne du tableau indique la valeur moyenne de ce paramètre calculée sur l'ensemble des scènes.

Tableau II.9/J.144 – Comparaison: robustesse et précision

	Méthode adoptée		Prédicteur optimal		Mesures globales	
Scène	**Condition-nement**	**Test**	**Condition-nement**	**Condition-nement**	**Test**	**Condition-nement**
Jardin	3,9	87,6	2,8	71,8	3,9	62,3
Mobile	30,1	48,6	10,5	82,1	179,1	162,5
Tennis	10,8	91,3	7,7	335,0	108,9	221,2
Diva	1,4	8,9	0,8	17,7	1,8	34,3
Kiel	22,4	9,3	20,5	13,4	30,6	27,7
Moyenne	**13,7**	**49,1**	**8,5**	**104,0**	**64,9**	**101,6**

L'avantage lié au calcul des paramètres objectifs sur la base du contexte devient manifeste lorsque la procédure décrite aux II.4.1 et II.4.2 est également appliquée aux mesures globales. Il est à noter que l'utilisation de mesures fondées sur le contexte peut permettre d'améliorer nettement les résultats d'évaluation dans toutes les scènes (à l'exception de celle du Jardin fleuri). Elle indique éventuellement que le processus de segmentation d'image pour la scène du Jardin fleuri a besoin d'être affiné.

L'exemple montre aussi que le processus de prédiction décrit au II.4.2 est plus robuste (c'est-à-dire qu'il est moins dépendant de la base de données de conditionnement) que le prédicteur optimal, ce qui permet d'améliorer les résultats de prédiction relatifs à la base de données de test.

II.5 Conclusions

Dans le présent appendice, on présente une méthode d'évaluation subjective de la qualité utilisant des paramètres objectifs fondés sur la segmentation d'image. Les paramètres objectifs sont calculés dans les régions planes, les régions de bord et les régions de texture résultant du processus de segmentation d'image.

Les résultats présentés dans le présent appendice montrent que l'utilisation de paramètres objectifs fondés sur le contexte au lieu de paramètres globaux conduit à des prédictions plus précises. Cet aspect est renforcé par l'utilisation du modèle de perception fondé sur la méthode de prédiction linéaire décrite au II.4.2. Cette méthode a donné des résultats de prédiction plus robustes que la méthode de prédiction optimale.

On peut encore améliorer les résultats si:

- l'information temporelle est incluse dans le processus de segmentation d'image (par exemple les régions de bord pourraient être scindées en régions de bord comportant peu de mouvement et en régions de bord comportant beaucoup de mouvement);
- les régions planes, les régions de bord et les régions de texture de la chrominance sont également prises en compte dans le processus de segmentation d'image, car les algorithmes I, II et III ont été utilisés pour segmenter la composante de luminance uniquement.

Nous proposons donc d'inclure la méthode de prévision linéaire présentée dans le présent appendice ainsi que les mesures objectives fondées sur le contexte dans de nouvelles Recommandations de l'UIT portant sur l'évaluation objective de la qualité vidéo.

II.6 Références

[1] 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*.

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

[3] 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)*.

[4] ANSI T1.801.03 (1996), *Digital transport of one-way video signals – Parameters for objective performance assessment*.

[5] UIT-R BT.500-7 (1995), *Méthodologie d'évaluation subjective de la qualité des images de télévision*.

[6] 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*.

[7] GONZALEZ, WINTZ (P.), Digital Image Processing, *Addison Wesley*, 1987.

[8] DOUGHERTY: An Introduction to Morphological Image Processing, *SPIE Optical Engineering Press*, Bellingham, WA, Vol. TT9, 1992.

[9] KAUFMANN (A.): Introduction to The Theory of Fuzzy Subsets, *Academic Press*, New York, NY, Vol. 1, 1975.

[10] BARRERA (J.), BANON (J.F.), LOTUFO (R.A.): Mathematical Morphology Toolbox for the Khoros System, *Conference on Image Algebra and Morphological Image Processing, V International Symposium on Optics, Imaging, and Instrumentation, SPIE's Annual Meeting*, San Diego, USA, 24-29 juillet 1994.

[11] Contribution COM 12-66 de l'UIT-T, *Selections from the draft American National Standard – Digital transport of one-way signals – Parameters for objective performance assessment*, Etats-Unis d'Amérique, janvier 1996.

[12] Contribution tardive D021 de la Commission d'études 12 de l'UIT-T, *Objective and subjective measures of MPEG video quality: summary of experimental results*, Etats-Unis d'Amérique, avril 1997.

[13] Contribution tardive D101 de la Commission d'études 12 de l'UIT-T, *A Two-Stage Objective Model for Video Quality Evaluation*, Bellcore, mai 1996.

[14] ANSI T1A1 Contribution Number T1A1.5/96-121, *Objective and subjective measures of MPEG video quality*, GTE Labs., NTIA/ITS, Octobre 1996.

APPENDICE III

Tektronix/Sarnoff

Introduction

Les nouveaux services de la télévision numérique nécessitent que la qualité de service soit contrôlée par des instruments de mesure très différents de ceux mis en œuvre pour les services analogiques. Une corrélation étroite entre les mesures objectives de qualité de l'image et l'évaluation subjective de la qualité constitue une exigence clé.

Le présent appendice décrit un instrument de mesure, fondé sur le modèle de la vision humaine, que l'on peut utiliser au sein d'un système de télévision numérique. Son implémentation pratique fournit des résultats qui montrent une corrélation élevée avec les évaluations subjectives faites conformément à la Rec. UIT-R BT.500-7.

Sont traitées les questions particulières suivantes:

- mesure de l'indice objectif de qualité de l'image (PQR, *picture quality rating*) au sein des réseaux numériques de transmission vidéo;
- exigences relatives au prétraitement vidéo antérieur à l'analyse;
- détails de l'algorithme utilisé pour l'analyse;
- résultats des tests montrant la corrélation entre les mesures objectives et les évaluations subjectives de qualité de l'image.

Une solution permettant de mesurer la qualité vidéo perçue au sein des systèmes vidéo numériques est décrite. Elle est implémentée dans une application commerciale.

III.1 Indice objectif de qualité de l'image (PQR) dans les environnements opérationnels

Il est bien connu que l'on peut accroître la précision des mesures objectives de qualité de l'image si l'on connaît la vidéo de référence. Sur le diagramme générique de la Figure III.1, le signal vidéo parvient au système de transmission (référence vidéo) qui le transmet vers la sortie où il est l'objet de contrôles (vidéo traitée). L'analyse des différences entre les vidéos de référence et traitée avec un modèle de la vision humaine fournit une mesure précise de l'indice objectif de qualité de l'image (PQR). (Le paragraphe III.4 détaille l'algorithme de calcul de l'indice objectif de qualité de l'image (PQR), qui s'appuie sur le modèle de la vision humaine élaboré conjointement par Sarnoff et Tektronix.)

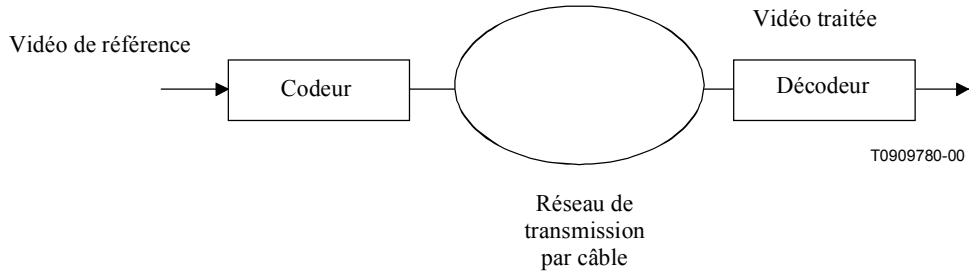


Figure III.1/J.144 – Diagramme générique d'un système de transmission de la vidéo compressée

La normalisation de la vidéo traitée constitue un prérequis essentiel à l'analyse suivant le modèle de la vision humaine. Le codage et le décodage peuvent engendrer des déplacements d'image horizontaux et verticaux et des recadrages, ainsi que des modifications de gain et de niveau de luminance et de chrominance. Ils doivent faire l'objet d'une normalisation avant la mise en œuvre du modèle de la vision humaine. (Les détails de ce processus de normalisation figurent dans le paragraphe III.2.

Certains systèmes de transmission peuvent nécessiter l'extension du diagramme générique pour prendre en compte les codecs concaténés et/ou l'utilisation du codage et du décodage PAL. Les principes restent toutefois identiques et l'on peut continuer d'appliquer le processus fondé sur l'indice objectif de qualité de l'image (PQR).

On peut envisager en laboratoire d'utiliser des séquences de test vidéo en lieu et place de données vidéo réelles. Ces séquences constituent une source vidéo pouvant être répétée et facilitent les mesures communes à différents laboratoires. La prise en compte d'une large gamme de données de programmes normalisés dans des séquences vidéo tests assurent une validité optimale des données de programmes réels.

Dans un environnement opérationnel, on peut remplacer les séquences de test vidéo par des données de programmes réels. La vidéo de référence et la vidéo traitée par un décodeur placé, comme l'indique la Figure III.2, au niveau de la source de transmission, permettent de mesurer l'indice objectif de qualité de l'image (PQR) pour le système opérationnel.

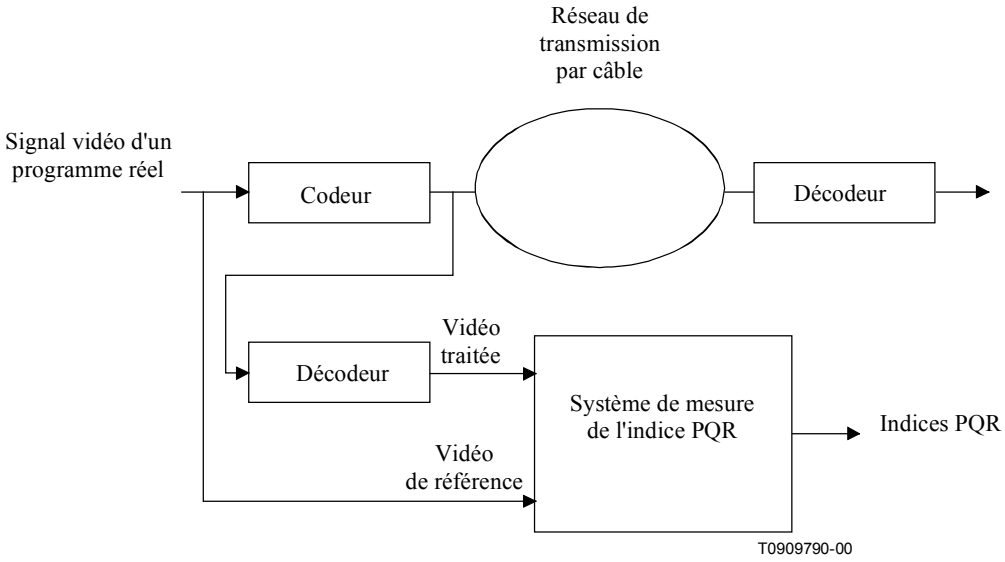


Figure III.2/J.144 – Indice objectif de qualité de l'image (PQR) lorsque la référence est disponible

Il est possible que la difficulté d'accès à la vidéo de référence restreigne l'utilisation de données réelles continues. La surveillance de l'alimentation d'un satellite au niveau d'une tête de réseau par câble en constitue un exemple clair, pour lequel on peut choisir pour vidéo de référence une séquence vidéo commune telle qu'un logo de station. Celui-ci pourrait être fourni au câble-opérateur et stocké localement pour servir de référence de comparaison par rapport à la vidéo traitée. Voir la Figure III.3.

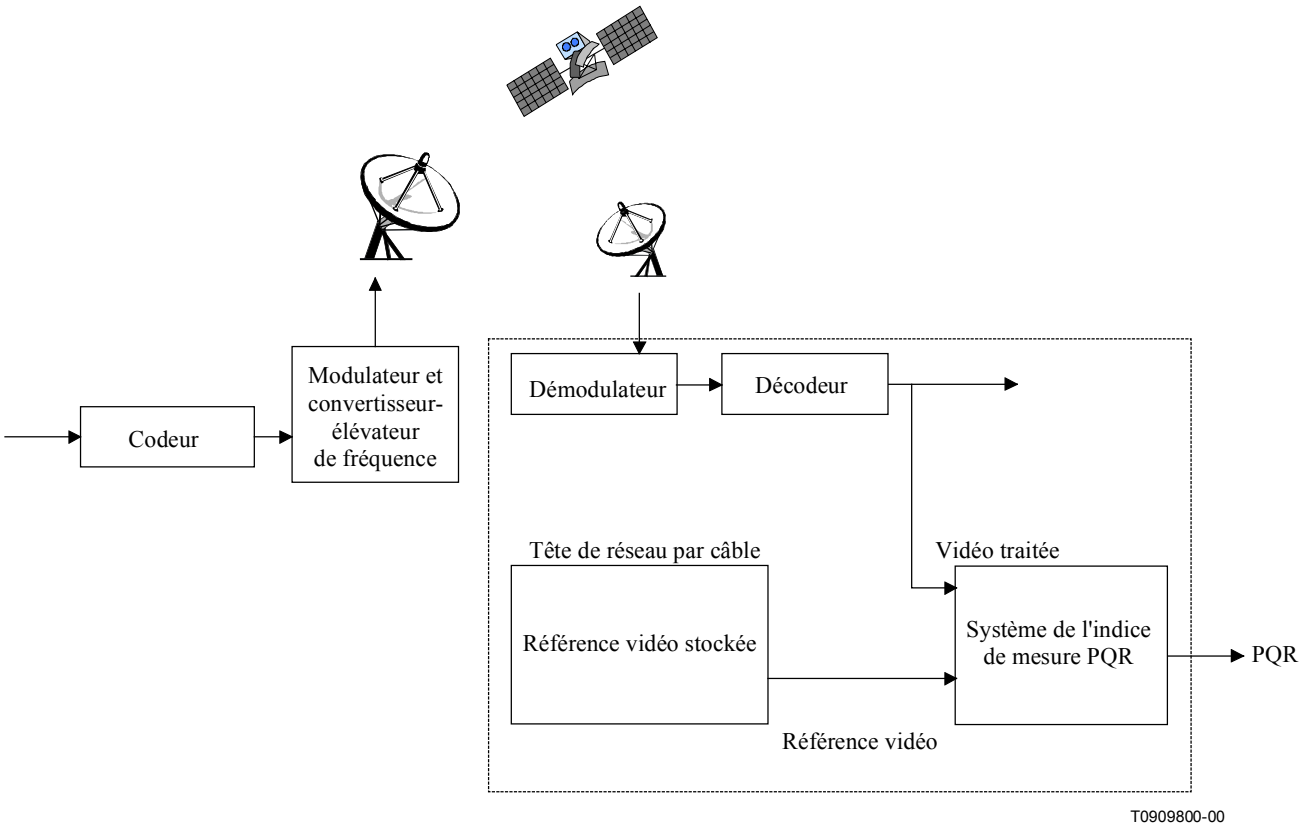


Figure III.3/ J.144 – Indice objectif de qualité de l'image (PQR) pour une tête du réseau éloignée, avec stockage de la référence

III.2 Prétraitement vidéo – Normalisation

Appliquer la méthode de l'indice objectif de qualité de l'image (PQR) à un système vidéo quelconque nécessite la normalisation de la vidéo traitée. 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 (HVS, *human visual system*). Méthode de mesure objective de qualité de l'image la plus sensible et la plus précise, 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 subdivision de la mesure en deux parties (normalisation et calcul de l'indice PQR) est nécessaire pour obtenir les résultats les plus significatifs.

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 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. Les changements non stationnaires sur les images

dus au contenu vidéo et au système de compression sont mesurés dans le cadre du calcul de l'indice PQR.

Normaliser avant d'évaluer la qualité de l'image est un concept qui se révèle également nécessaire dans les normes de mesure subjective, afin de tenir compte de l'exploitation typique d'un système. Il faut inclure l'énoncé suivant à la Rec. UIT-T P.911 ("Méthodes d'évaluation subjective de la qualité audiovisuelle pour applications multimédias") et à la Rec. UIT-T P.910 ("Méthodes subjectives d'évaluation de la qualité vidéographique pour les applications multimédias"): "Les paramètres d'exploitation (tels que le niveau de signal) associés aux séquences de test devront correspondre à ceux des signaux d'alignement utilisés pour la vérification des conditions d'observation (et d'écoute). Il faudra faire état de tout réglage en exploitation effectué pour que les séquences source ou traitées satisfassent à la présente exigence".

La Figure III.4 présente le fonctionnement du système de mesure de l'indice PQR du point de vue de la normalisation. 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 III.1.

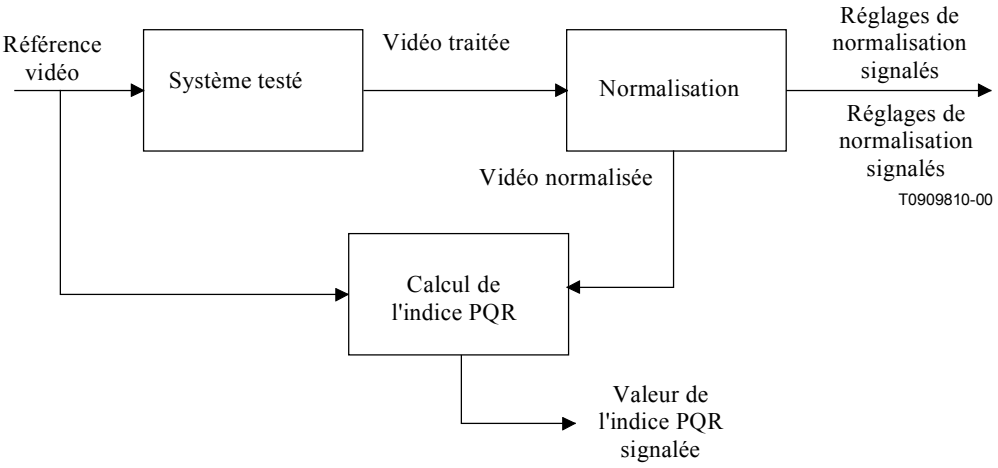


Figure III.4/J.144 – Exploitation du système de mesure de PQR

Tableau III.1/J.144 – Paramètres et tolérance de normalisation

Paramètre	Tolérance de normalisation
Gain de luminance	< 0,2 dB
Gain (de différence) de couleur	< 0,2 dB
Niveau continu de luminance	< 0,5 % du signal max
Niveau continu (de différence) de couleur	< 0,5% du signal max
Décalage temporel canal à canal	< 2 ns
Déplacement horizontal de pixel	< 0,1 pixel
Déplacement vertical de la ligne	< 0,1 ligne

III.3 Aperçu du système

Les indices objectifs de qualité de l'image (PQR) figurent parmi les principaux résultats du système de mesure de l'indice PQR décrit plus haut. L'objet du présent paragraphe est de décrire le modèle de la vision humaine utilisé par ce système.

Le modèle de la vision humaine de Sarnoff/Tektronix est une méthode permettant 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*). Lubin (1993, 1995) décrit une version du modèle qui ne s'applique qu'aux images statiques et achromatiques.

Le modèle de la vision humaine se révèle utile dans un contexte général (voir Figure III.5). 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 III.5 intitulée "système sous test" correspond schématiquement à l'une quelconque de ces possibilités. Normalement, l'évaluation de la qualité subjective de l'image de test, par comparaison avec la séquence de référence, nécessiterait un observateur et un véritable dispositif de visualisation. Cette évaluation se trouverait facilitée si on remplaçait le dispositif de visualisation et l'observateur par le modèle de la vision humaine. On substituerait ainsi à une comparaison subjective, une comparaison des séquences de test et de référence permettant de générer une séquence de cartes JND.

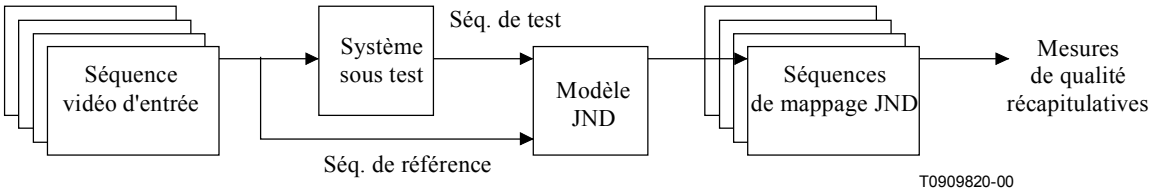


Figure III.5/J.144 – Modèle de la vision humaine pour l'évaluation du système

La Figure III.6 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 III.6), 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.

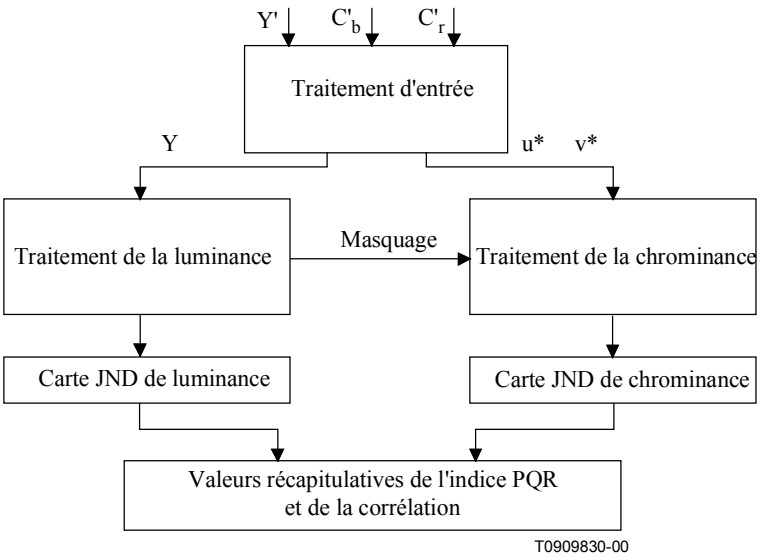


Figure III.6/J.144 – Ordinogramme du modèle de vision humaine Sarnoff/Tektronix

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. 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.

Notons que le modèle présenté ici repose sur deux hypothèses fondamentales:

a) chaque pixel est carré et sous-tend un angle d'observation de 0,03 degrés. Ce chiffre correspond à une hauteur d'écran de 480 pixels et une distance d'observation de quatre hauteurs d'écran (ce qui constitue la distance d'observation minimale, selon la Rec. UIT-R BT.500). Au-delà de cette distance, le modèle surestime la sensibilité humaine aux détails spatiaux. En l'absence de contraintes sévères relatives à la distance d'observation, on choisit le modèle le plus "sensible" possible respectant les prescriptions de la Rec. UIT-R BT.500;

b) le modèle s'applique à des luminances d'écran de 0,01 à 100 ft-L (valeurs pour lesquelles on a étalonné la sensibilité globale), la plus grande précision correspondant toutefois à environ 20 ft-L (valeur pour laquelle on a étalonné toutes les fréquences spatio-temporelles). On suppose qu'un changement de luminance induit des modifications proportionnelles de sensibilité pour toutes les fréquences spatio-temporelles. Cette hypothèse s'avère moins importante près de 20 ft-L, valeur pour laquelle des étalonnages plus nombreux ont été effectués.

Les traitements mentionnés dans certaines cases de la Figure III.6 sont décrits de manière plus détaillée ci-après, en particulier sur les Figures III.7, III.8 et III.9.

III.4 Aperçu de l'algorithme

III.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 III.7 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'.

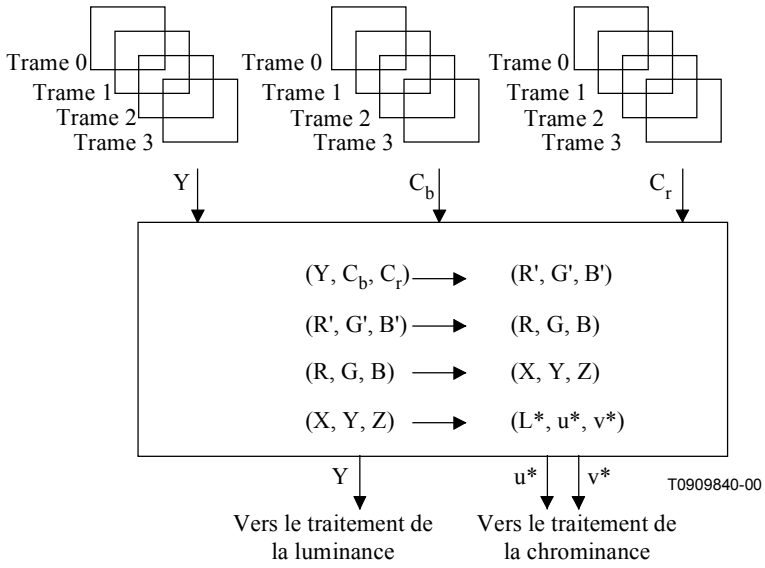


Figure III.7/J.144 – Traitement d'entrée

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.

On peut ensuite choisir l'une des deux options de traitement: mi-hauteur ou hauteur totale. Dans le cas de balayages avec entrelacement, les images mi-hauteur[1] 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.

[1] Les lignes d'une image mi-hauteur correspondent à une trame, c'est-à-dire aux lignes paires ou impaires d'une image.

Pour garantir (pour chaque pixel) une uniformité approximative de perception de l'espace de couleur pour ce qui est des différences de couleur isoluminantes, on mappe chaque pixel dans l'espace CIELUV, un espace de couleur uniforme répondant à une norme internationale (voir Wyszecki et Stiles, 1982). Les composantes de chrominance u* et v* de cet espace parviennent ensuite aux phases de traitement de la chrominance du modèle[2].

III.4.2 Traitement de la luminance

Comme l'indique la Figure III.8, chaque valeur de luminance fait d'abord l'objet d'une compression non linéaire. Chaque trame de luminance est ensuite filtrée et sous-échantillonée 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.

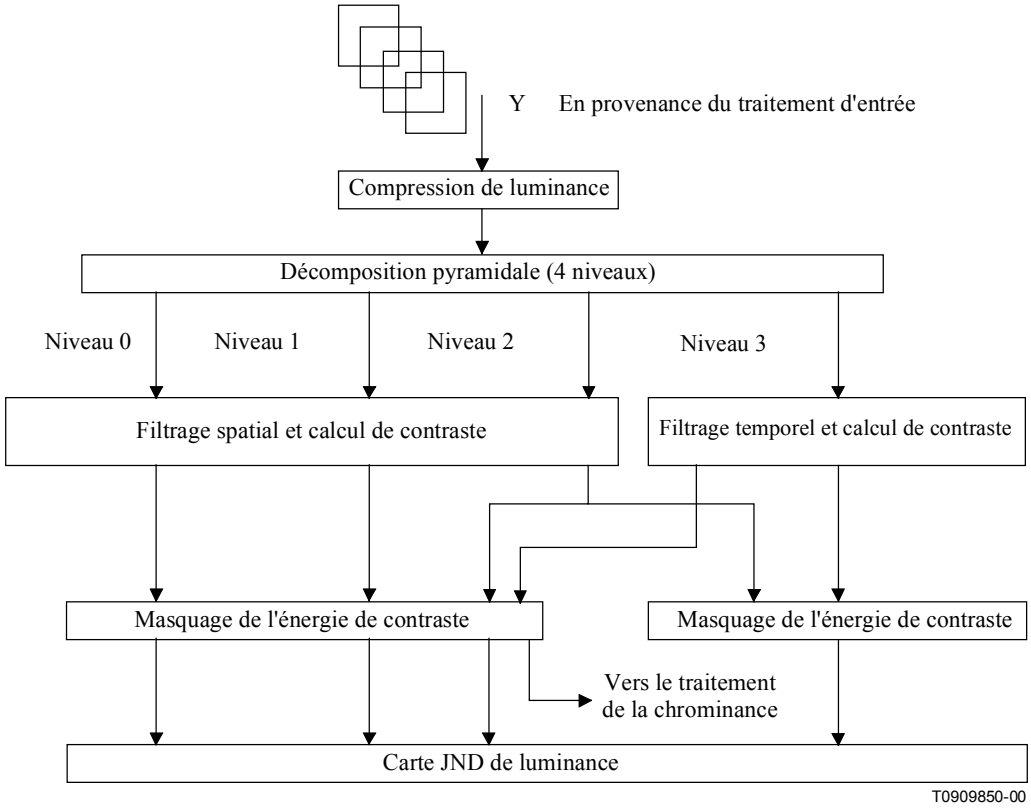


Figure III.8/J.144 – Aperçu du traitement de la luminance

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

[2] 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.

phases ultérieures du modèle[3]. (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 (moyennage et sous-échantillonage, 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;
- suréchantillonage 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.

III.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 III.9 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 (Mullen, 1985), 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 de 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).

---

[3] Associer un contraste constant à 1 JND constitue une implémentation connue sous le nom de loi de Weber pour la vision.

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 la fin du III.4.2).

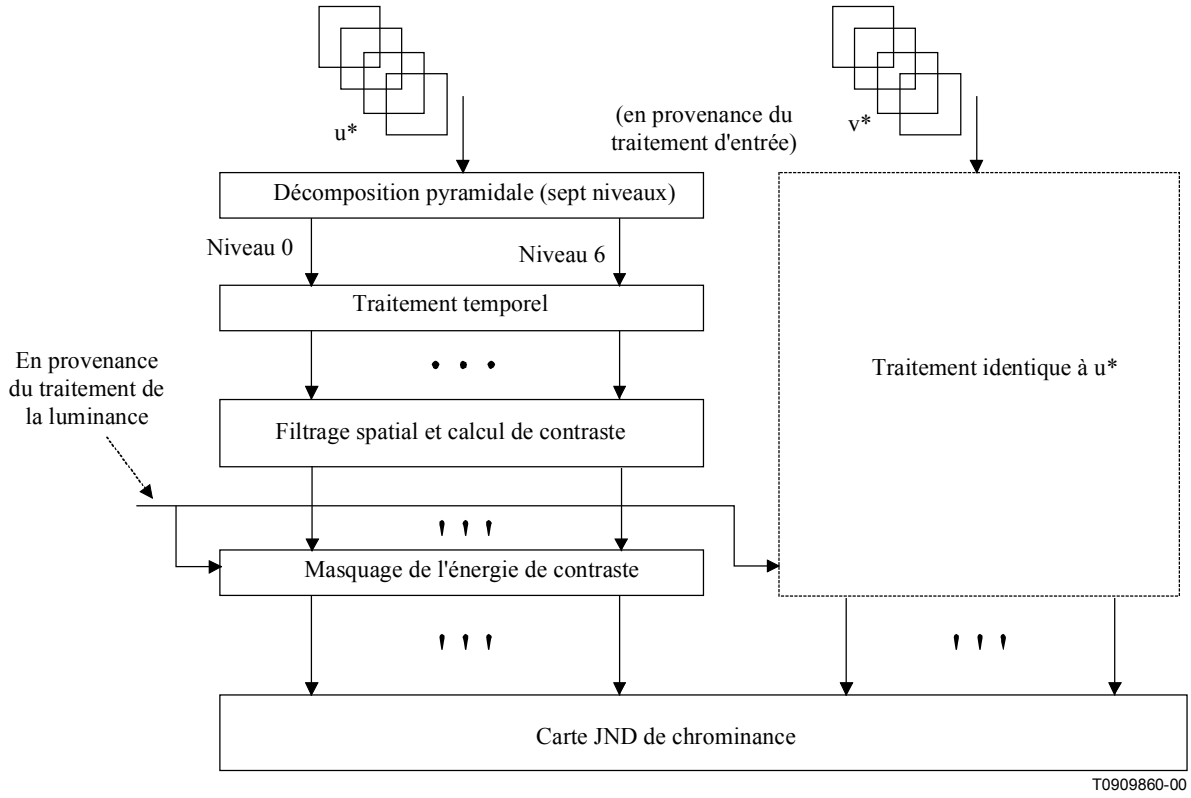


Figure III.9/J.144 – Aperçu général du traitement de la chrominance

III.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 PQR (indice de qualité de l'image, *picture quality rating*), calculé à l'aide de la norme Q de Minkowski. Pour ce faire, on élève à la puissance Q^e chaque valeur de pixel d'une carte JND. L'indice PQR se calcule alors comme la racine Q^e de la somme normalisée de toutes ces valeurs de pixel à la puissance Q^e.

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 de qualité de l'image relatif à la séquence entière, en utilisant à nouveau une norme Q de Minkowski.

III.5 Corrélation avec les résultats subjectifs

III.5.1 Aperçu général

L'IRT (*Institut für Rundfunktechnik GmbH* de Munich, en Allemagne) et Tektronix ont récemment achevé la phase initiale d'une enquête relative à la performance d'une méthode de calcul de l'indice objectif de qualité de l'image (PQR, *picture quality rating*), fondée sur le modèle de la vision humaine élaboré conjointement par Sarnoff et Tektronix. Le présent paragraphe propose un bref résumé des résultats d'un test en aveugle qui compare la méthode de mesure de qualité de l'image (PQR) aux notes moyennes d'opinion (MOS, *mean opinion score*) subjectives de téléspectateurs. L'IRT a généré pour l'expérience 60 scènes vidéo à partir de cinq séquences vidéo différentes que deux codeurs MPEG-2 ont compressées à des débits de 2; 3; 4,5; 7 et 10 Mbit/s. Les notes moyennes d'opinion ont été déterminées par l'IRT et les évaluations objectives de l'indice PQR ont été déterminées par Tektronix. La procédure de notation subjective reposait sur un échantillon de 25 personnes et était en stricte conformité avec les procédures de la Rec. UIT-R BT.500-7 (méthode DSCQS). Tektronix a calculé les indices objectifs PQR à partir du modèle de la vision humaine Sarnoff/Tektronix fondé sur les principes des différences tout juste perceptibles. Aucun paramètre du modèle n'a été modifié pour s'adapter aux données de l'IRT. Pour éviter tout risque de biais lors de l'expérience, Tektronix et l'IRT n'ont échangé les évaluations subjectives et objectives qu'une fois que chaque groupe avait terminé son évaluation. Bien que les paramètres du modèle (qui sont fondés sur la science de la vision humaine) n'aient fait l'objet d'aucun réglage particulier, on a observé une forte corrélation (taux de 0,88) entre les résultats issus des tests subjectifs et des tests objectifs (la qualité de diffusion présentant une corrélation typique de 0,91). Ces résultats, présentés sur la Figure III.12, se révèlent prometteurs quant à l'utilisation future de méthodes objectives pour la caractérisation et le contrôle de la qualité de l'image vidéo.

III.5.2 Matériel de test vidéo et traitement

L'IRT a fourni à Tektronix les scènes de test vidéo suivant le format SMPTE 125M 422-625/50 Hz (c'est-à-dire le format PAL pour bande D1). La durée de chaque scène est de 9 secondes. Dans la suite (HRC, *hypothetical reference circuit*) désigne le "circuit fictif de référence", conformément à la norme ANSI T1A1.5. Tektronix a ajouté un code à barres en haut de chaque image vidéo, avant le passage de la vidéo par les HRC. Ce code permet de déterminer le défaut d'alignement horizontal et vertical des pixels, le numéro de l'image ainsi que d'autres facteurs. On a caché la bande d'alignement durant les tests subjectifs, mais les résultats d'un test conduit avec un petit groupe de contrôle qui pouvait voir la bande ont montré que celle-ci n'avait qu'un effet minime sur l'évaluation des téléspectateurs. Après ajout des bandes d'alignement, l'IRT a transmis les séquences dans les HRC. On a utilisé deux codeurs vidéo (IRT[4] et Thomson) fonctionnant à des débits de 2,0; 3,0; 4,5; 7,0 et 10,0 Mbit/s. On a inclus aux tests des scènes à 2,0 Mbit/s (bien qu'il soit peu probable que des systèmes de diffusion commerciaux fonctionnent à moins de 3,0 Mbit/s) afin de pouvoir étudier la performance en delà des limites usuelles. Un dernier ensemble de HRC a permis de faire suivre une phase de conversion PAL par deux codeurs identiques fonctionnant à 3 Mbit/s. Il est probable que la conversion PAL en particulier engendre quelques défauts d'alignement sous-pixel. Les séquences d'origine et le traitement qui leur a été appliqué pour créer les scènes de test sont résumés ci-après.

---

[4] Le codeur "IRT" a été élaboré par l'IRT et plusieurs partenaires européens dans le cadre des projets Eureka 625 VADI, Race HD-SAT et Race DISTIMA.

Séquences d'origine	HRC	Débits binaires (Mbit/s)
1. Barcelone	1 Codeur IRT	2,0
2. Mobile et calendrier	2 "	3,0
3. NDR	3 "	4,5
4. Football (*soccer*)	4 "	7,0
5. Jardin fleuri	5 "	10,0
	6 Codeur Thomson	2,0
	7 "	3,0
	8 "	4,5
	9 "	7,0
	10 "	10,0
	11 PAL + MPEG (Thomson)	3,0
	12 PAL + MPEG (IRT)	3,0
	13 Référence – Pas de compression	

Ensemble de **60 scènes de test** = (**5** séquences) × [(**2** codeurs) × (**5** débits binaires) + **2** PAL]

Barcelone: Formation d'un défilé haut en couleur et plein de fantaisie sur un vaste terrain de sport (voir Figure III.13). La caméra effectue lentement un zoom arrière, le mouvement est lent. Les tribunes au second plan sont riches en fins détails. La séquence est colorée, lente, riche en fins détails.

Mobile et calendrier: séquence familière d'animation souvent utilisée par la communauté de compression vidéo. Cette animation colorée comprend des animaux de dessins animés, un petit train en mouvement, un ballon qui roule et un calendrier présentant des détails textuels. La séquence est colorée, lente et riche en fins détails.

NDR: chroniqueur radiophonique debout devant un mur de pierres agrégées. Le mur, peu coloré, se compose de détails très fins. La caméra effectue lentement un zoom arrière. Les détails du mur de pierres constituent la principale difficulté de compression. Le mouvement est très lent. La séquence est lente, riche en fins détails.

Football (désigné aux Etats-Unis d'Amérique sous le terme de "soccer"): la caméra filme le jeu avec un angle large, assez loin de l'action, avec des mouvements de vitesse moyenne. On observe une certaine défocalisation sur la scène d'origine lors de la première seconde de la vidéo. La séquence est rapide, riche en fins détails.

Jardin fleuri: cette séquence est largement utilisée au sein de la communauté de recherche sur la compression vidéo. La caméra, installée dans un véhicule ouvert, se déplace à vitesse modérée et filme un jardin de fleurs plein de couleurs. A l'arrière-plan se trouvent un moulin à vent en action et quelques personnes. Le jardin et de grosses branches d'arbre dénudées fournissent de fins détails. Le mouvement apparent est modéré. La séquence est colorée, lente, riche en fins détails.

La Figure III.13 présente une image typique de chacune des séquences décrites ci-dessus.

III.5.3 Evaluation subjective

On a utilisé pour les tests la méthode à double stimulus utilisant une échelle de qualité continue (DSCQS, *double stimulus continuous quality scale*), décrite dans la Rec. UIT-R BT.500-7.

La structure de présentation comprenait les phases suivantes illustrées sur la Figure III.10.

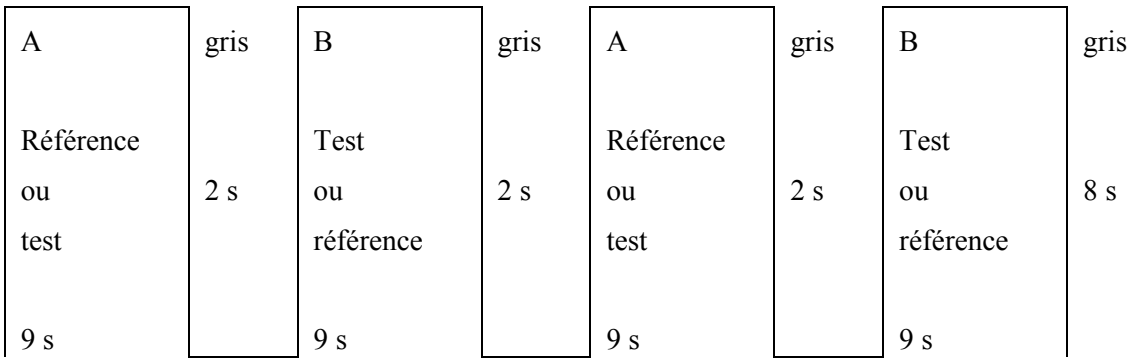


Figure III.10/J.144 – Ordre de présentation pour la méthode DSCQS

A constituait la référence et B le HRC, ou vice et versa, en fonction du test considéré. L'ordre était inconnu des évaluateurs. La durée totale d'un test était de 50 secondes.

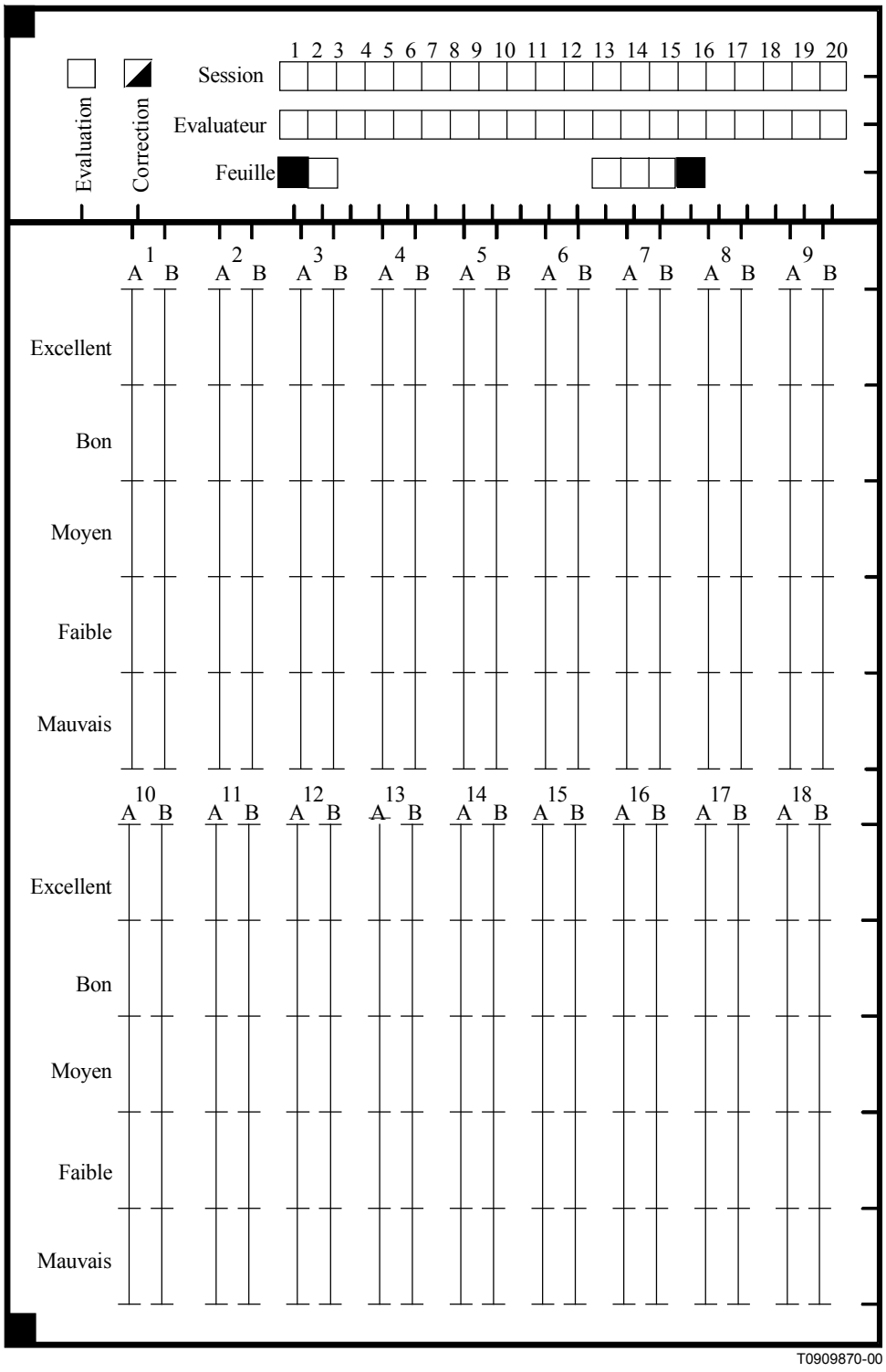


Figure III.11/J.144 – Feuille de test utilisée pour l'évaluation des séquences de test

Pour l'évaluation des séquences de test, on a utilisé une feuille de test semblable à celle présentée sur la Figure III.11.

Les évaluateurs ont indiqué les qualités de A et B sur une échelle linéaire. Les termes de notation de la qualité, qui se trouvent dans la marge gauche de la feuille, signifient: excellent, bon, moyen, faible, mauvais. Les résultats ont été évalués électroniquement et la distance entre l'extrémité basse de l'échelle et l'indicateur de qualité estimé par l'évaluateur a été calculé en millimètres dans chaque cas. La différence entre les résultats relatifs à la référence et ceux relatifs au HRC constituait le résultat important.

Outre le test réel, on a présenté des exemples et des séquences de conditionnement. Quatre exemples ont été présentés au début de la première session, illustrant la méthode de test et couvrant la gamme de qualité prévue. On a demandé aux téléspectateurs de ne pas évaluer les séquences, qui n'étaient que des exemples. Les exemples sont énumérés dans le Tableau III.2 ci-après.

Tableau III.2/J.144 – Exemples de séquences

Numéro	Séquence de test	Codeur	Débit binaire (Mbit/s)
1	Zoom sur une rue	IRT	3
2	Barcelone 2	Thomson	4
3	Zoom sur une rue	IRT	10
4	Barcelone 2	Thomson	2

"Zoom sur une rue" est une production bien connue de la BBC qui montre une scène de rue à Edimbourg. Barcelone 2 est une scène provenant de la même production que "Barcelone", mais pour laquelle les participants sont filmés en gros plan.

Il fallait que les séquences de conditionnement soient estimées par les personnes ne sachant pas que les résultats n'étaient pas évalués. Ces séquences sont énumérées dans le Tableau III.3 ci-après.

Tableau III.3/J.144 – Séquences de conditionnement

Numéro	Séquence de test	Codeur	Débit binaire (Mbit/s)
1	Renata	Thomson	2
2	Tennis de table	IRT	10
3	Renata	Thomson	4
4	Tennis de table	IRT	2
5	Renata	Thomson	10
6	Tennis de table	IRT	4

"Renata" et "Tennis de table" sont des séquences de test bien connues.

Les sessions de test étaient structurées de la manière suivante:

Session 1: exemples (4) – Séquences de conditionnement (6) – Tests réels (31)

Session 2: séquences de conditionnement (6) – Tests réels (34)

La durée totale s'élevait à 34 minutes et 10 secondes pour la session 1, et 33 minutes et 20 secondes pour la session 2. Vingt-cing évaluateurs ont pris part aux séries de test, dont 15 "extérieurs" (femmes au foyer, étudiants, etc…) et 10 personnes appartenant au personnel de l'IRT (à l'exclusion des spécialistes). La distance d'observation était de 6 H (H: hauteur d'image). Toutes les autres conditions étaient conformes à la Rec. UIT-R BT.500-7. On a utilisé des moniteurs Sony.

Les bandes de code à barres situées en haut de chaque image étaient cachées par une feuille noire fixée à l'écran. Un test conduit au sein d'un petit groupe de cinq évaluateurs (du personnel de l'IRT, à l'exclusion des spécialistes) durant lequel la bande n'était pas cachée a montré que cette condition n'avait pas d'incidence significative sur les résultats.

Les résultats essentiels des tests subjectifs étaient les valeurs moyennes (note moyenne d'opinion subjective, MOS) et les intervalles de confiance à 95% des différences entre les résultats relatifs à la référence et ceux relatifs au HRC. Comme l'échelle s'étale sur 100 millimètres, 100 constitue le pire des résultats, et 0 le meilleur. Un résultat de 20 correspond à la différence entre "excellent" et "bon", ou entre "bon" et "moyen", etc.

III.5.4 Evaluation objective de la qualité de l'image

Après que l'IRT a traité les séquences vidéo au moyen des HRC pour générer, comme décrit plus haut, les données de test, Tektronix a effectué les évaluations objectives de qualité PQR. On peut décrire brièvement le processus de la façon suivante:

- la vidéo est stockée sur des fichiers informatiques à partir d'une bande D1, en vue d'un traitement numérique;
- des algorithmes d'alignement temporel et spatial sont utilisés pour déterminer les erreurs d'alignement;
- la vidéo est ensuite réalignée spatialement et temporellement. Pour cet ensemble de données, le réalignement spatial n'a été effectué qu'au pixel entier le plus proche, ce qui fait qu'aucun filtre d'interpolation n'a été nécessaire. Le réalignement temporel est réalisé par décalage d'image et ne modifie les données en aucune façon;
- la vidéo a ensuite été traitée par la méthode objective d'évaluation de la qualité de l'image (PQR) élaborée par Sarnoff/Tektronix. On a effectué cette analyse par le biais d'une version logicielle du modèle de qualité implantée sur une station de travail SUN Sparc. La méthode a généré un historique de la qualité image par image pour toute la durée de la vidéo, de manière à pouvoir analyser la qualité en continu. Pour la comparaison avec les évaluations subjectives, on a condensé les historiques de chaque scène en un indice de qualité de l'image (PQR) global, qui constituait une mesure de la qualité globale pour toute la durée de la scène.

III.5.5 Comparaison des évaluations subjective et objective

La Figure III.12 montre les notes moyennes d'opinion subjectives déterminées par l'IRT et les indices PQR objectifs estimés par Tektronix. Les barres d'erreurs verticales correspondent aux intervalles de confiance à 95% couvrant l'ensemble des estimations subjectives des téléspectateurs. La relation entre les évaluations subjective et objective est bien établie, constante, avec un taux de corrélation important s'élevant à 0,88. On peut voir dans l'aplatissement de la courbe vers la droite une compression de l'évaluation de la qualité de l'image par le téléspectateur, à mesure que la qualité se dégrade vers des valeurs très basses. Cet effet est bien connu dans le domaine des tests subjectifs; il est cohérent avec les effets de compression observés dans d'autres domaines de la perception humaine, tels que le volume sonore et la luminosité. Le groupe des trois points situés dans le coin supérieur droit correspond à des scènes pour lesquelles soit le codeur a échoué de manière catastrophique sur certaines régions de la scène, soit la qualité était très faible. Si l'on exclut ces points, le coefficient de corrélation s'élève alors à 0,91. Etant donné que les estimations objectives de qualité n'ont nécessité aucun réglage ni optimisation des paramètres pour s'adapter aux données de test, les résultats sont tout à fait encourageants et laissent à penser que des méthodes objectives contribueront à réduire le temps, les coûts et les éventuels biais associés à la caractérisation et au contrôle vidéo.

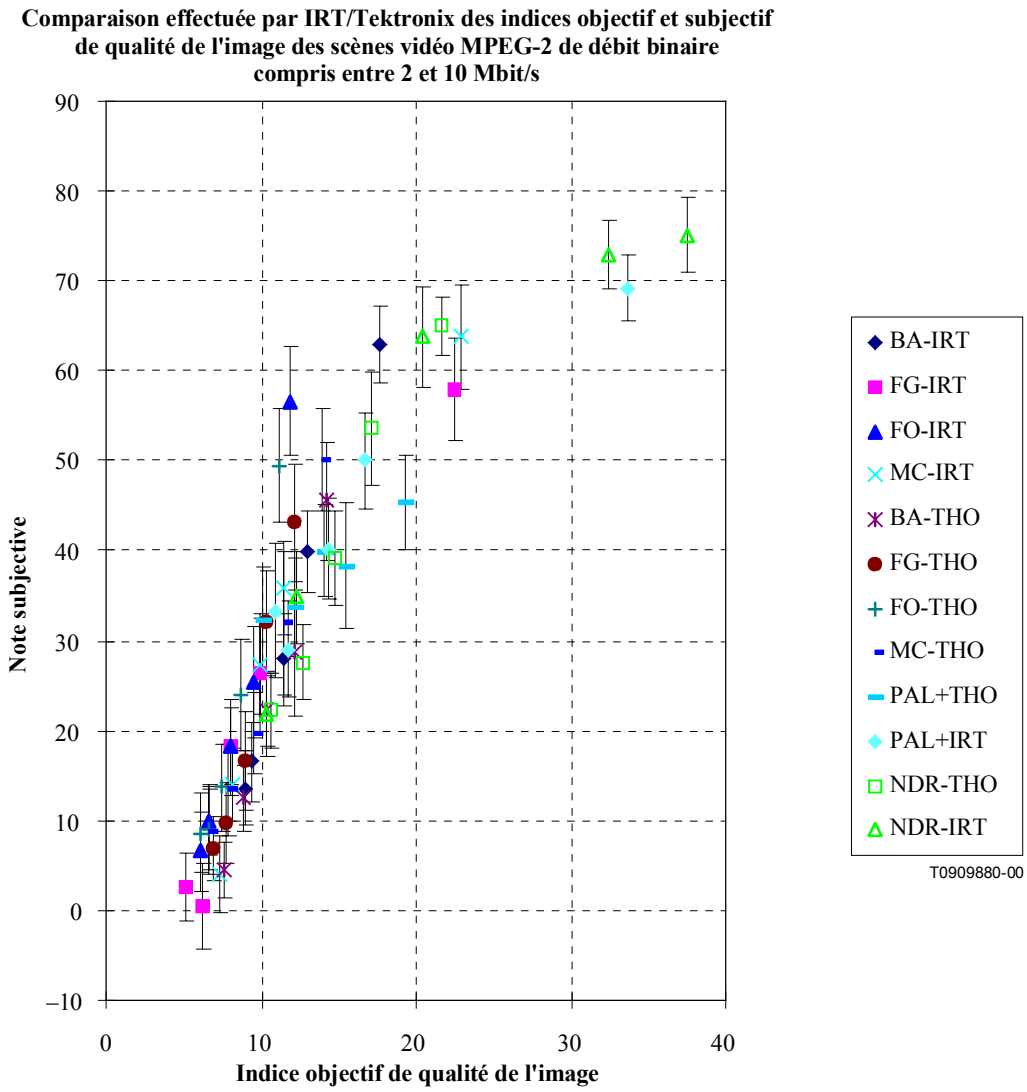
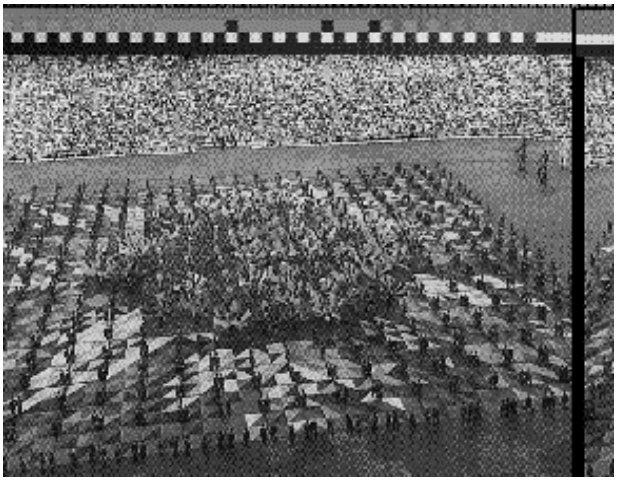
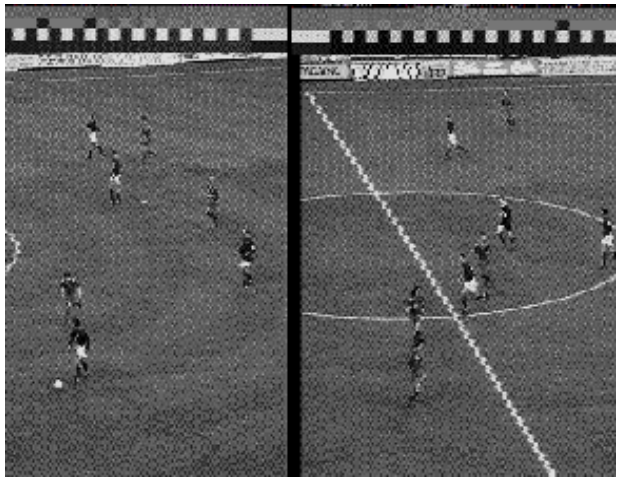


Figure III.12/J.144 – Comparaison des notes moyennes d'opinion (MOS) subjectives déterminées par l'IRT et des indices objectifs de qualité de l'image (PQR) estimés par Tektronix, pour 60 scènes de test MPEG-2 et PAL de débit binaire compris entre 2 et 10 Mbit/s

Les intervalles de confiance à 95% relatifs aux notes subjectives sont indiqués par des barres verticales. La corrélation entre les évaluations objective et subjective est de 0,88 pour l'ensemble des données, et la compression par le téléspectateur de l'évaluation de la qualité est visible dans le coin supérieur droit de la figure, pour les scènes présentant les qualités les plus faibles. La corrélation est de 0,91 si l'on exclut les scènes de plus faible qualité situées dans le coin supérieur droit.

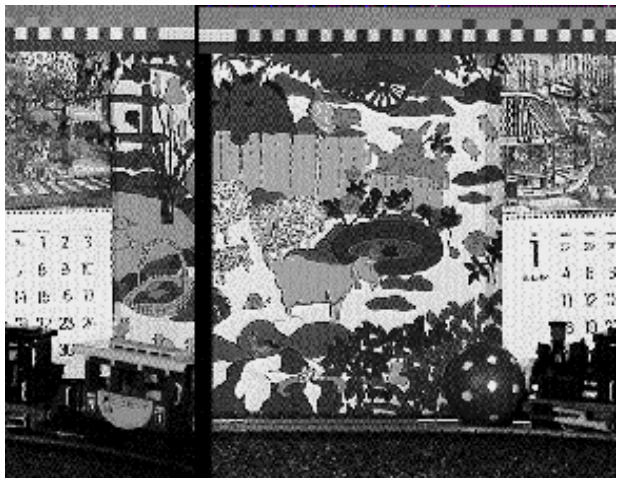


Barcelone



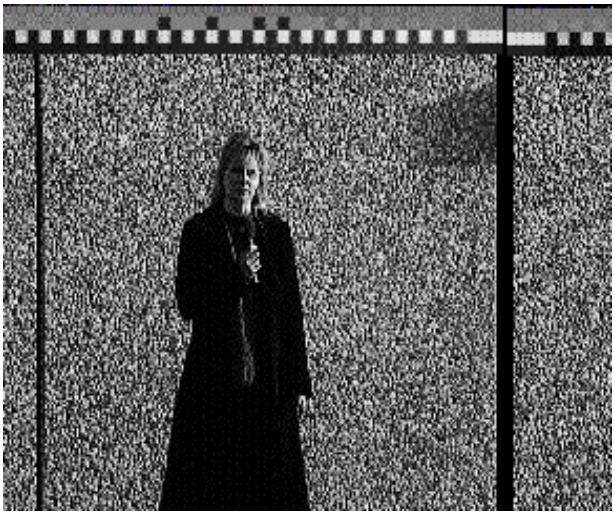
Football

Jardin fleuri



Mobile et calendrier

T0909890-00



NDR

Figure III.13/J.144 – Images typiques des séquences de test vidéo

III.6 Références

– LUBIN (J.): The use of psychophysical data and models in the analysis of display system performance, in A.B. Watson (ed.), *Digital Images and Human Vision*, *MIT Press*, pp. 163-178, 1993.

– LUBIN (J.): A visual system discrimination model for imaging system design and evaluation, in E. Peli (ed.), *Visual Models for Target Detection and Recognition*, *World Scientific Publishers*, 1995.

– MULLEN (K.T.): The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings, *J. Physiol.* 359, 381-400, 1985.

– WYSZECKI (G.), STILES (W.S.): Color Science, 2nd ed., *Wiley*, 1982.

APPENDICE IV

NHK/Mitsubishi Electric Corp.

Résumé

Un système d'évaluation de la qualité d'image numérique compressée a été mis au point, dans lequel la dégradation d'image est calculée en temps réel compte tenu de la perception visuelle humaine. Dans ce système, on considère la sensibilité au bruit en fonction des fréquences spatio-temporelles, compte tenu de la brillance de l'image. Cette approche a permis d'améliorer la précision de l'évaluation de la qualité de l'image concernant de nombreux types de dégradation.

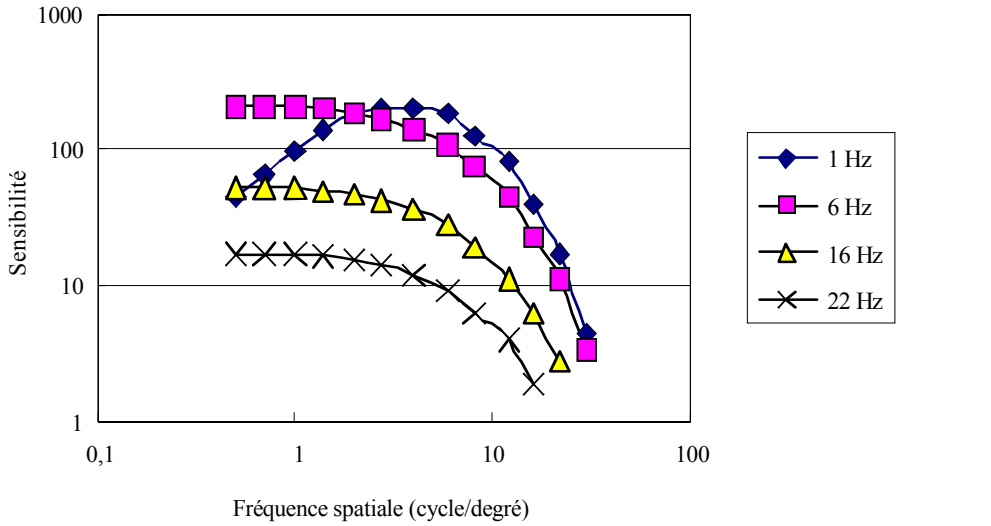
IV.1 Méthode d'évaluation objective de la détérioration de la qualité

Le modèle correspond à une émulation des caractéristiques visuelles humaines au moyen de filtres 3D (spatio-temporels), appliqués aux différences entre le signal source et le signal traité. En ce qui concerne l'implémentation des filtres, on n'emploie pas de méthode d'analyse fréquentielle fondée sur les blocs comme la transformation DCT afin d'éviter les effets mutuels potentiels entre systèmes de codage et d'évaluation. On fait varier les caractéristiques des filtres en fonction du niveau de luminance. La note de qualité en sortie est calculée comme étant une somme pondérée de mesures obtenues avec les filtres. Le système doit permettre d'évaluer la qualité d'image en termes de finesse et de reproductibilité, en reflétant exactement les fonctions visuelles. Dans les paragraphes qui suivent, on décrit ces caractéristiques visuelles humaines puis on explique l'implémentation matérielle.

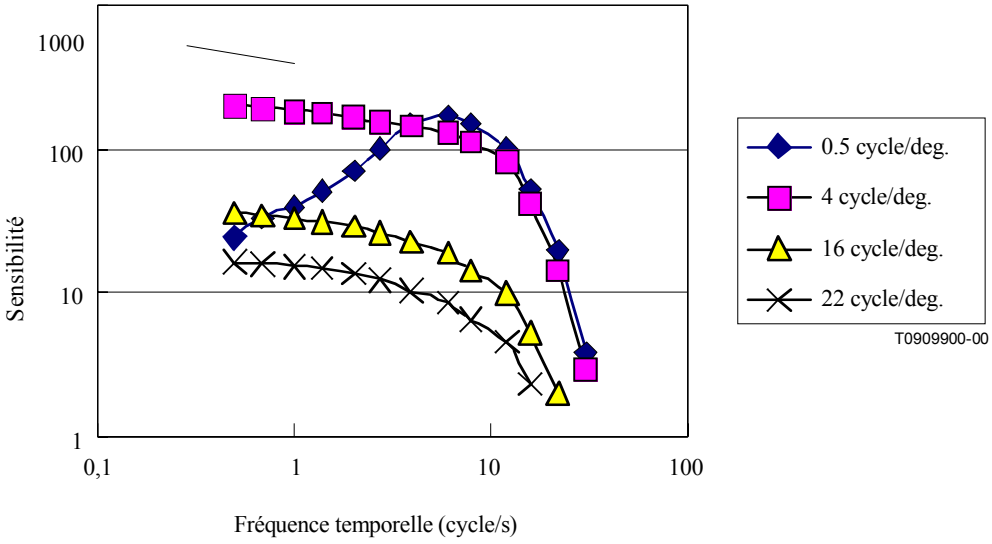
IV.2 Caractéristiques visuelles humaines

IV.2.1 Réponse visibilité/fréquence spatiale

La réponse visibilité humaine/fréquence spatio-temporelle représentée sur la Figure IV.1 a notamment été mesurée par J.G. Robson [1]. La réponse visibilité/fréquence spatiale présente une caractéristique de perpendicularité par rapport à la réponse visibilité/fréquence temporelle, d'où une symétrie de rotation avec pour centre les axes optiques.



a) Réponse en fonction de la fréquence spatiale



b) Réponse en fonction de la fréquence temporelle

Figure IV.1/J.144 – Réponse visibilité/fréquence spatio-temporelle

IV.2.2 Réponse visibilité/fréquence pour diverses valeurs de brillance de l'image

En ce qui concerne la dépendance de la réponse visibilité/fréquence vis-à-vis de la brillance, des mesures, faites notamment par Kelly [2], montrent que non seulement la réponse en fonction de la fréquence spatiale mais aussi la réponse en fonction de la fréquence temporelle dépendent de la brillance de l'image. La Figure IV.2 illustre la dépendance de la réponse visibilité/fréquence spatiale vis-à-vis de la brillance de l'image dans le cas d'une image pratiquement fixe avec une fréquence temporelle de moins de 4 Hz pour la sensibilité visuelle. "td" est une unité associée à la luminance d'une image de fond d'œil.

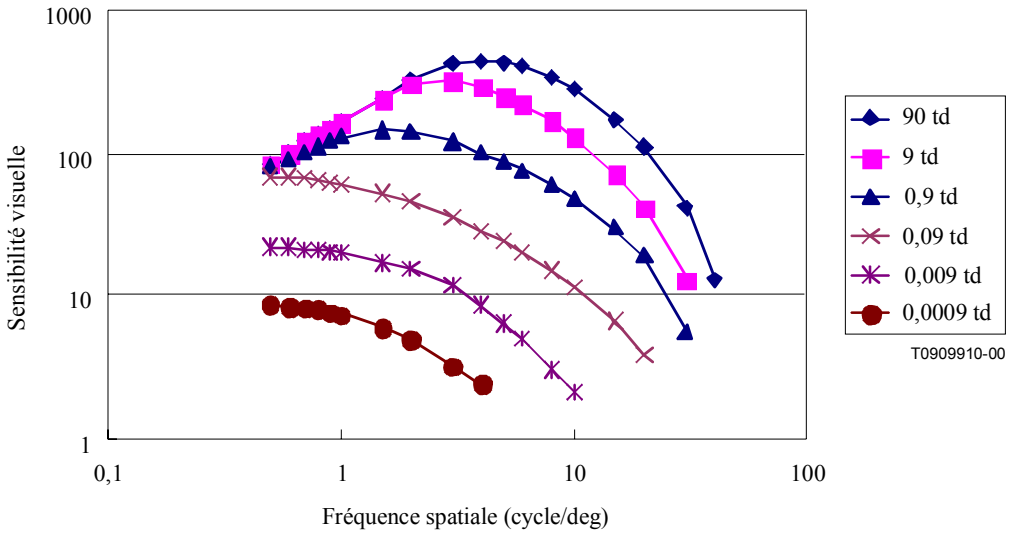


Figure IV.2/J.144 – Dépendance de la réponse visibilité/fréquence spatiale vis-à-vis de la brillance

La Figure IV.3 illustre la dépendance de la réponse visibilité/fréquence temporelle vis-à-vis de la brillance dans le cas d'une image uniforme. L'œil humain est généralement sensible à un papillotement d'environ 10 Hz lorsque l'image est très brillante. Lorsque l'image est très peu brillante, le papillotement est largement invisible.

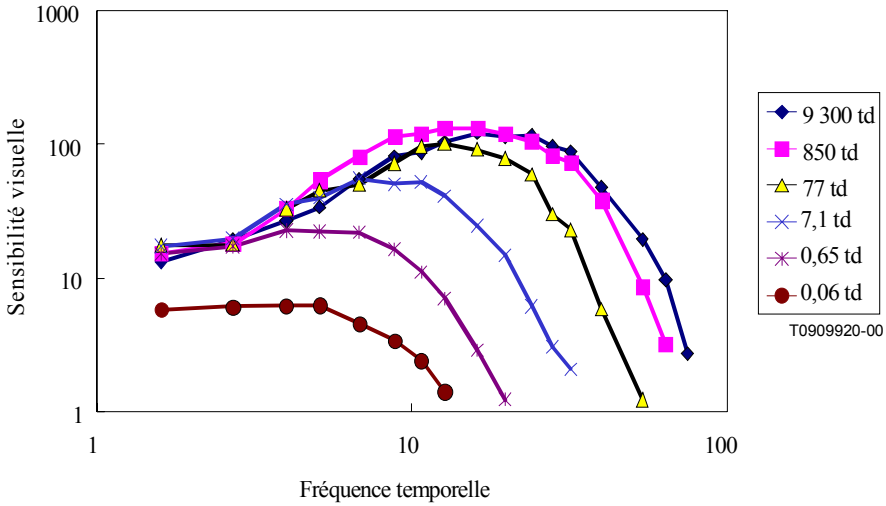


Figure IV.3/J.144 – Dépendance de la réponse visibilité/fréquence temporelle vis-à-vis de la brillance

IV.2.3 Sensibilité visuelle pour diverses valeurs de brillance

La Figure IV.4 montre les limites de perception d'un bruit aléatoire sur un écran de télévision [3] pour différents niveaux de brillance. On constate que la sensibilité visuelle dépend de la brillance.

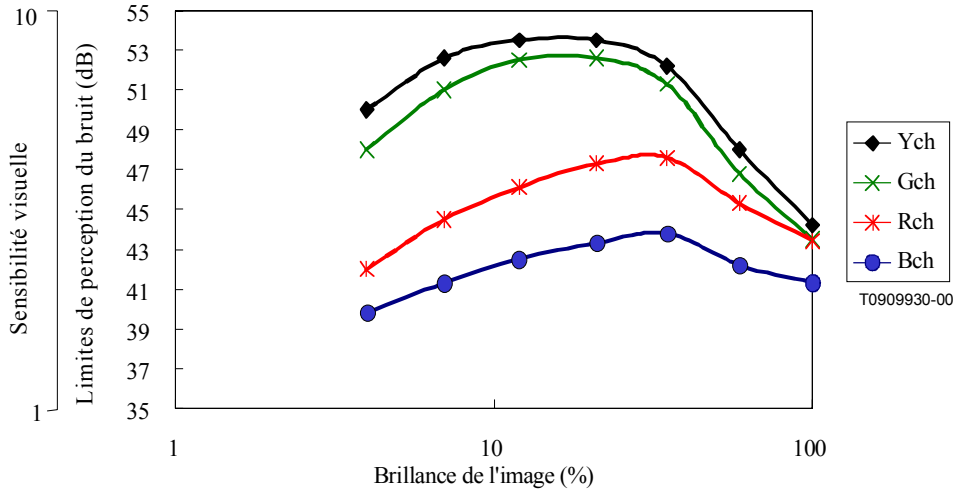


Figure IV.4/J.144 – Limites de perception d'un bruit aléatoire sur un écran de télévision

IV.3 Réalisation de fonctions visuelles dans un filtre numérique

IV.3.1 Structure du système d'évaluation

La Figure IV.5 montre la structure du système d'évaluation. Des signaux de différence sont d'abord produits à partir d'images de la séquence d'origine et d'images de la séquence de test puis injectés dans le filtre numérique 3D commandé par la brillance avec la même dépendance de la réponse visibilité/fréquence 3D vis-à-vis de la brillance. On compare alors les signaux de différence filtrés avec le perception visuelle de chaque pixel. On obtient en résultat une expression numérique de la distorsion au-delà des limites de perception de l'œil humain.

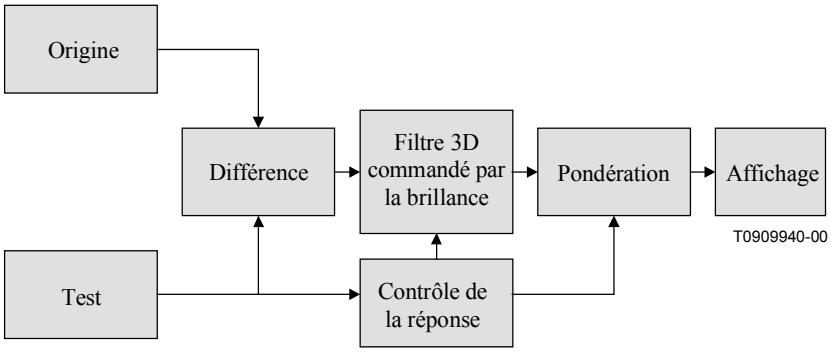


Figure IV.5/J.144 – Structure d'un système d'évaluation expérimental

IV.3.2 Filtre numérique 3D commandé par la brillance

La Figure IV.6 montre la composition de filtres numériques 3D, dont la réponse fréquentielle et la sensibilité varient en fonction de la brillance. En combinant les filtres spatiaux et les filtres temporels en fonction de la brillance de l'image, on obtient une émulation de la réponse visibilité humaine/fréquence.

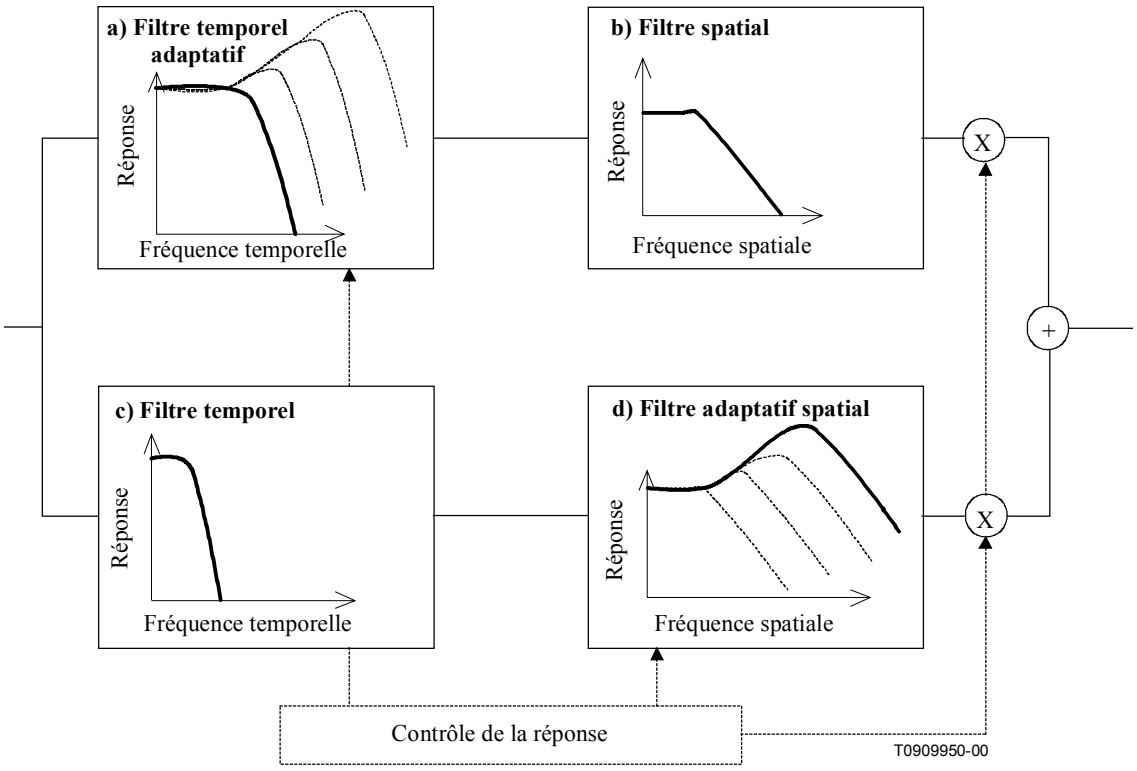


Figure IV.6/J.144 – Composition de filtres numériques 3D, dont la réponse fréquentielle et la sensibilité varient en fonction de la brillance.

IV.3.3 Filtre spatial adaptatif dépendant de la brillance de l'image

La Figure IV.7 montre le filtre spatial adaptatif d) de la Figure IV.6, qui est obtenu par commutation de filtres spatiaux ayant des caractéristiques différentes.

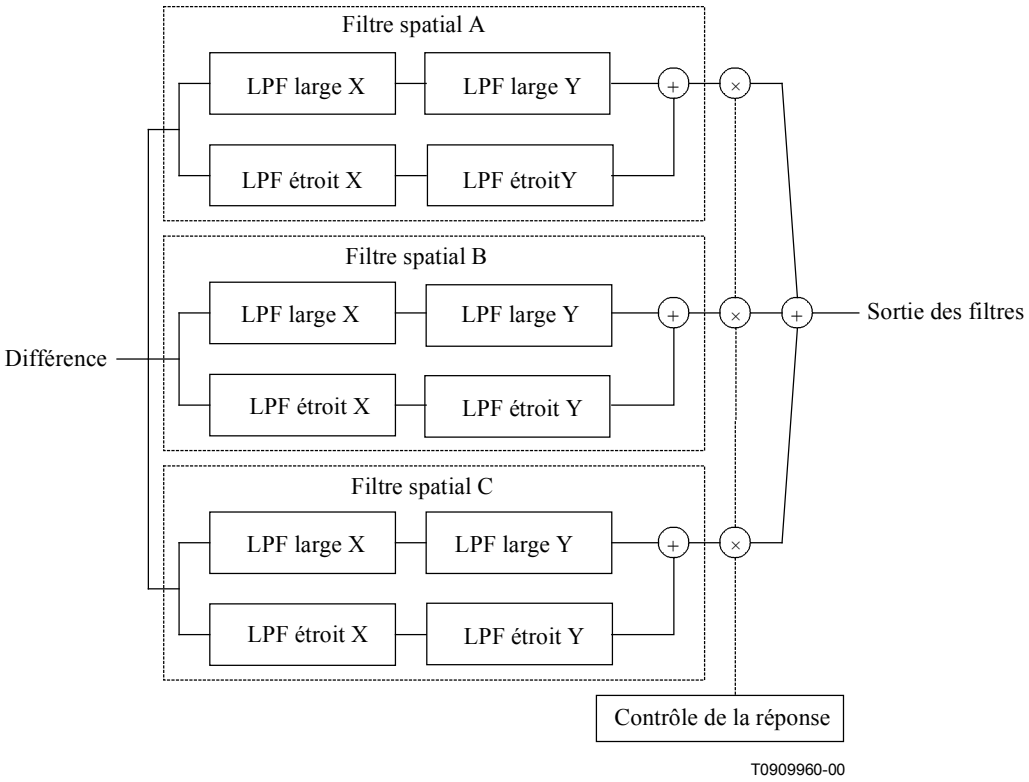


Figure IV.7/J.144 – Filtre spatial adaptatif obtenu par commutation de filtres spatiaux

IV.3.4 Réponse en fonction de la fréquence spatiale ayant la forme d'un volcan

Les fonctions visuelles possèdent les caractéristiques d'un filtre spatial passe-bande dans les directions horizontale et verticale. En représentant ces caractéristiques par un filtre numérique 3D, on obtient le graphe ayant la forme d'un volcan représenté sur la Figure IV.8. Ce graphe représente la réponse de l'œil humain, indiquant que la détérioration est manifeste sur les bords de l'image.

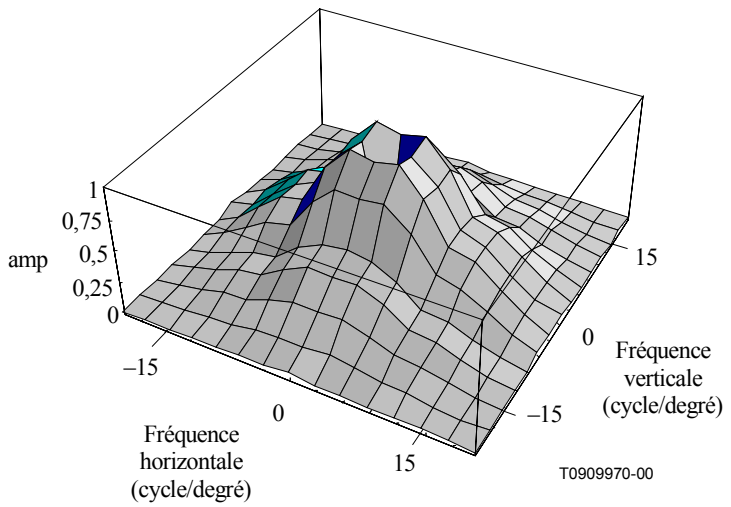


Figure IV.8/J.144 – Filtre spatial dont la réponse a la forme d'un volcan

IV.4 Exemple d'évaluation au moyen du système d'évaluation de la qualité d'image

La Figure IV.9 illustre la relation entre des notes d'évaluation subjective obtenues par 20 spécialistes en vidéo conformément à la Rec. UIT-R BT.500 et des notes d'évaluation objective que nous avons obtenues au moyen de notre système d'évaluation. Pour notre évaluation, nous avons utilisé des images en composantes et des images composites pour la séquence de test et des images en composantes pour la séquence d'origine.

En ce qui concerne non seulement la distorsion liée à la compression mais aussi la détérioration de la qualité, y compris la conversion composite/composantes et les limites de largeur de bande, on constate que la qualité d'image (PQ, *picture quality*) obtenue par le système d'évaluation objective concorde bien avec celle obtenue par la méthode à double stimulus utilisant une échelle de qualité continue (DSCQS) d'évaluation subjective.

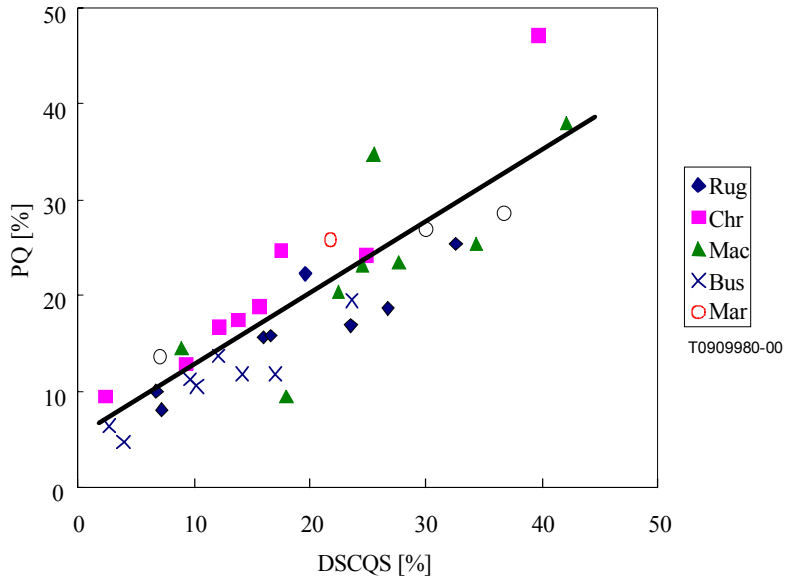


Figure IV.9/J.144 – Relation entre des notes subjectives et des résultats d'évaluation objective

A titre de référence, la Figure IV.10 montre une relation entre des erreurs quadratiques moyennes des images traitées et des notes subjectives. Contrairement à la Figure IV.9, ce graphe présente une corrélation relativement faible.

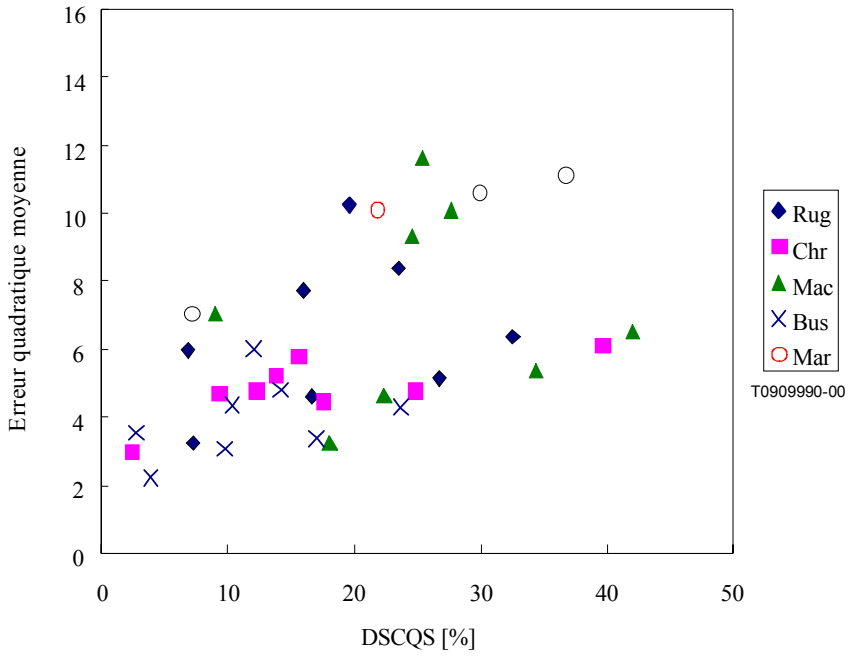


Figure IV.10/J.144 – Relation entre des erreurs quadratiques moyennes et des notes subjectives

IV.5 Système d'évaluation de la qualité d'image en temps réel

Sur la Figure IV.11, on peut voir l'apparence extérieure du système d'évaluation. Les caractéristiques du système sont les suivantes:

1) mesure en temps réel;

2) ajustement automatique du temps de traitement du CODEC et du déplacement de phase pour la synchronisation;

3) facilité de mesure car des images d'origine sont incorporées dans le système.

Le système a été amélioré en termes de précision par la représentation fidèle des fonctions de visibilité humaine. Généralement, l'œil humain peut voir des détails lorsque l'image est brillante, mais ne peut les voir qu'indistinctement lorsque l'image est sombre. Il est manifeste que la réponse en fonction de la fréquence spatiale varie suivant la brillance. Par ailleurs, nos yeux peuvent distinguer nettement un papillotement sur l'écran lorsque l'image est brillante. Lorsque l'image devient sombre, la réponse en fonction de la fréquence temporelle varie suivant la persistance de la vision. Dans le système, nous avons représenté au mieux de notre capacité les fonctions de visibilité humaine qui varient beaucoup en fonction du niveau de brillance.

Le système d'évaluation a permis d'obtenir des résultats de mesure bien représentés et fortement corrélés avec les notes obtenues par évaluation subjective, quel que soit le type de signal vidéo.

Figure IV.11/J.144 – Apparence extérieure du système d'évaluation

IV.6 Références

[1] ROBSON (J.G.): Spatial and Temporal Contrast-Sensitivity Functions of the Visual System, *J. Opt. Soc. Am.*, pp. 1141-1142, août 1966.

[2] KELLY (D.H.): Visual Responses to Time-Dependent Stimuli. I. Amplitude Sensitivity Measurements, *J. of the Opt. Soc. of Am.*, Vol. 51, No. 4, pp. 422-429, avril 1961.

[3] NISHIDA (Y.), KOIKE (J.), OHTAKE (H.), ABE (M.), YOSHIKAWA (S.): Design Concept for a Low-Noise CCD Image Sensor Based on Subjective Evaluation, *IEEE Trans. ED.*, Vol. 36, No. 2, 1989.

APPENDICE V

KDD

Système d'évaluation objective de la qualité vidéo et détermination de la performance

V.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) [1], 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.

- Evaluation 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, nous proposons 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, comme indiqué au [2]. Dans le présent appendice, nous rapportons des résultats de vérification de ce système.

V.2 Système d'évaluation objective de la qualité vidéo

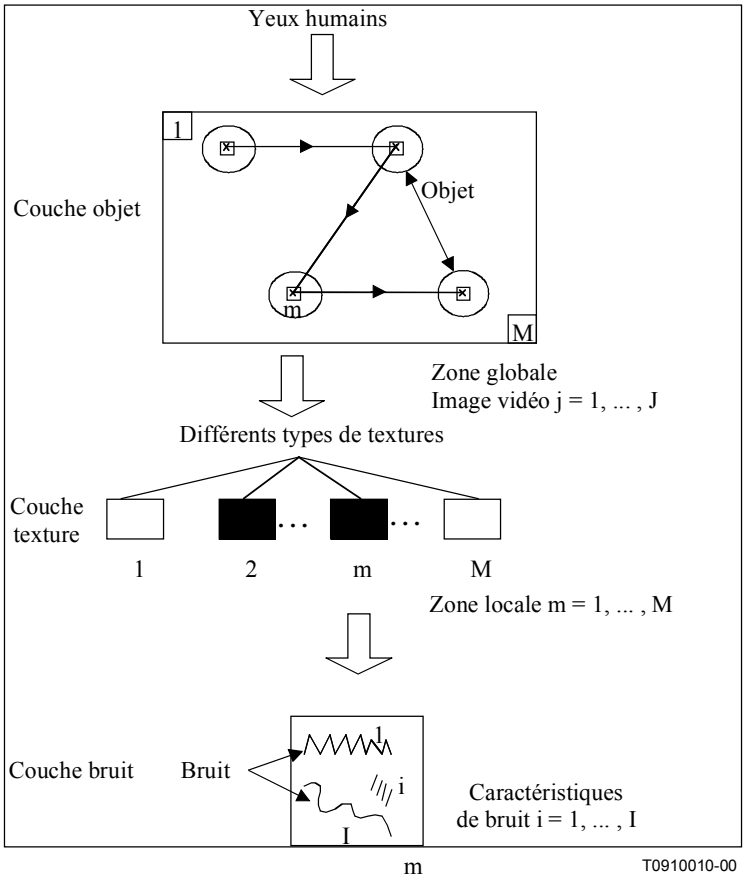


Figure V.1/J.144 – Modèle à trois couches pour le signal vidéo

La Figure V.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 V.2).

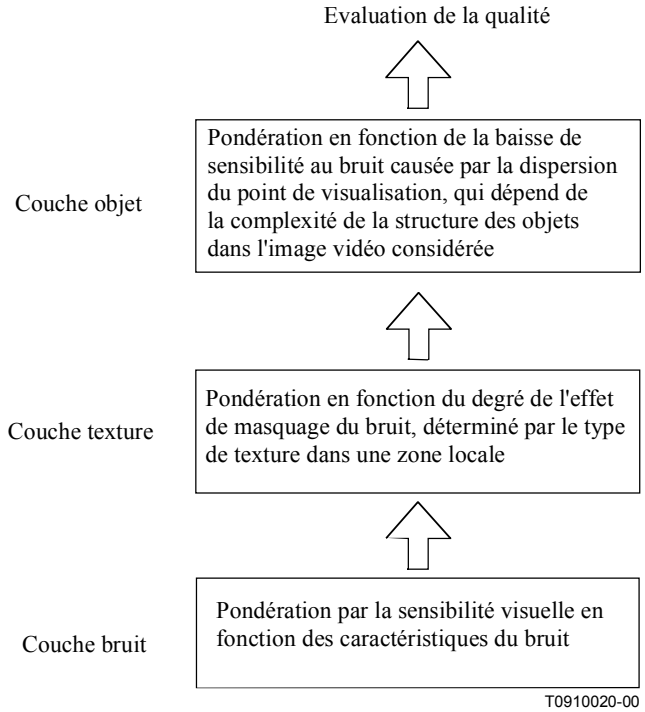


Figure V.2/J.144 – Système de pondération du bruit de bas en haut à trois couches

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)

t_c: fonction de pondération de bruit dans une texture C

G(j): paramètre caractérisant la complexité de la structure des objets dans l'image j

q_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 h_i correspondant aux caractéristiques de bruit dans la zone locale (j,m), on calcule l'erreur $WMSE_{NL}$ comme suit:

$$WMSE_{NL}(j,m) = \frac{1}{I}\sum_{i=1}^{I} h_i \cdot P(j,m,i) \tag{V-1}$$

Ensuite, dans la couche texture, en sommant l'erreur $WMSE_{NL}(j,m)$ sur la totalité de l'image (m = 1, …, M) pondérée par t_c correspondant à la texture C(j,m) dans la zone locale (j,m), on calcule l'erreur $WMSE_{TL}(j)$ comme suit:

$$WMSE_{NL}(j) = \frac{1}{M}\sum_{m=1}^{M} t_c(j,m) \cdot WMSE_{NL}(j,m) \tag{V-2}$$

Enfin, dans la couche objet, en prenant une valeur moyenne de l'erreur $WMSE_{TL}$ sur les images j = 1, …, J pondérée par G(j) correspondant au degré de dispersion du point de visualisation, on calcule l'erreur $WMSE_{OL}$ comme suit:

$$WMSE_{OL} = \frac{1}{J}\sum_{j=1}^{J} q_G(j) \cdot WMSE_{TL}(j) \tag{V-3}$$

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.500-7) comme suit:

$$WSMR(dB) = 10\ \log_{10}\frac{255^2}{WMSE} \tag{V-4}$$

$$D(\%) = f(WSNR) \tag{V-5}$$

V.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 V.3 montre la configuration du système et le Tableau V.1 décrit les principaux paramètres. Comme l'indique le Tableau V.1, les signaux composites (NTSC) et les signaux en composantes avec échantillonnage total sont pris en charge.

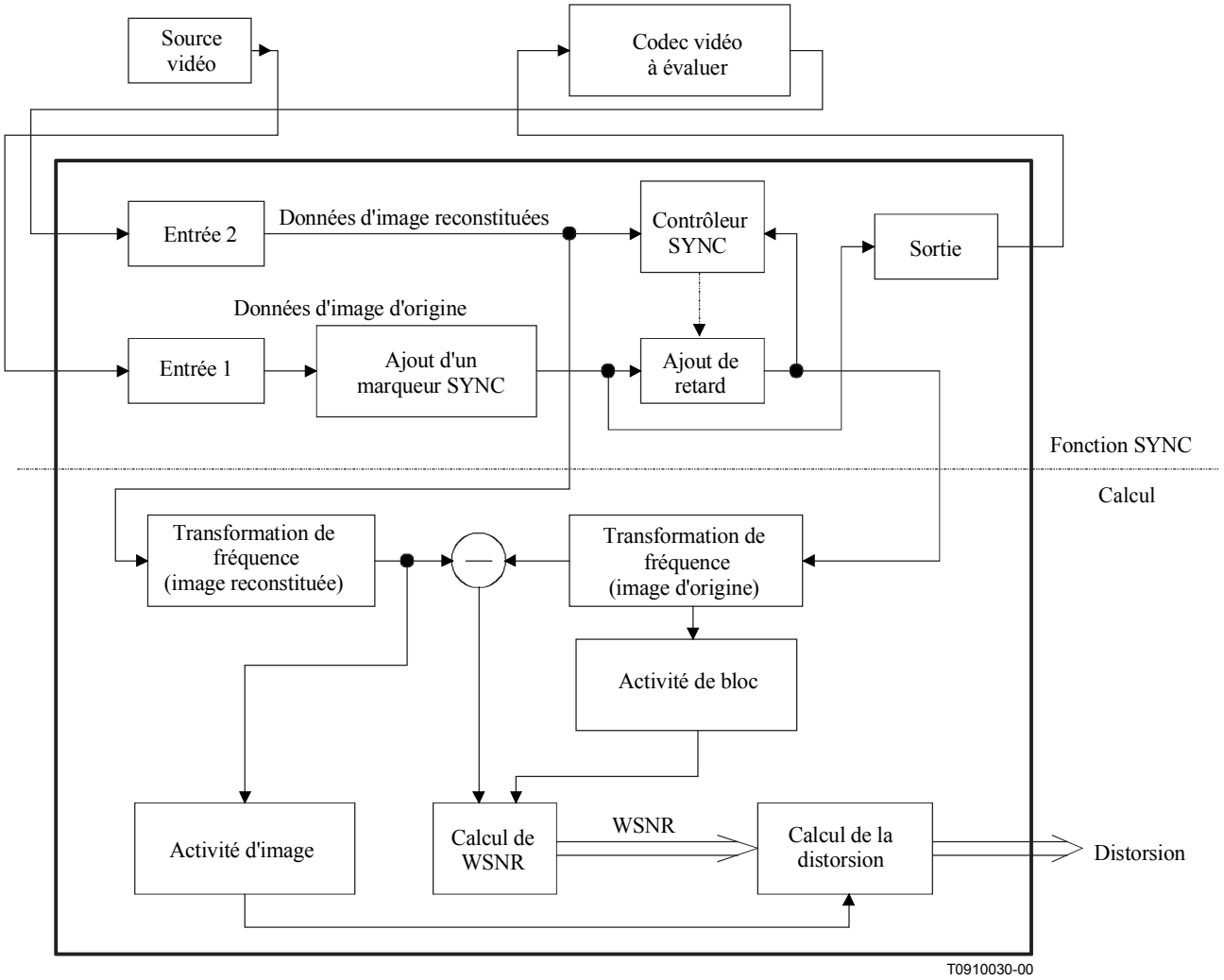


Figure V.3/J.144 – Configuration du système

V.3.1 Module 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. 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.).

V.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 V.1/J.144 – 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

V.4 Résultats de vérification

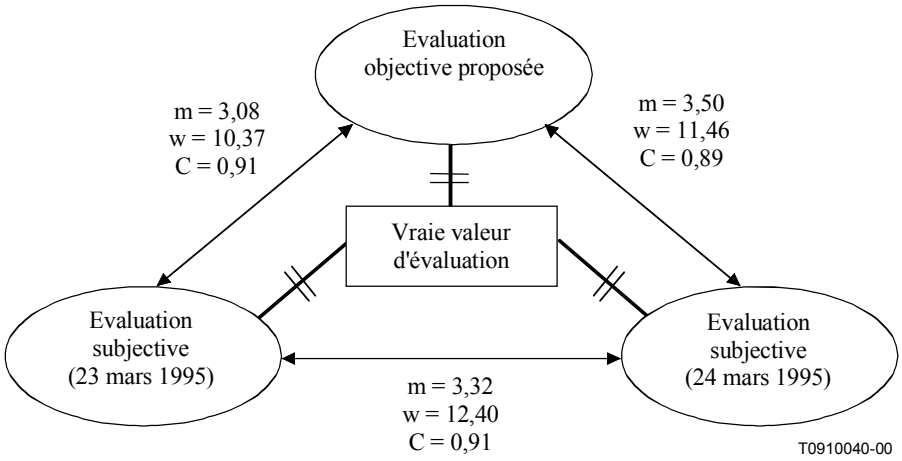
Nous avons comparé les résultats obtenus par le système d'évaluation proposé avec les résultats de tests d'évaluation subjective qui ont déjà été associés à des notes conformément à la Rec. UIT-R BT.500-7. Les cibles d'évaluation sont MPEG-2 SP@ML avec 5 Mbit/s, 7 Mbit/s et 10 Mbit/s appliqués aux signaux de test de télévision en composantes 4:2:2 de la Rec. UIT-R BT.601. Elles comprennent 17 données comprenant Mobile, Jardin fleuri, Leaders, etc. On a donc au total 17 données × 3 débits binaires = 51 échantillons (Tableau V.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 V.4.

Tableau V.2/J.144 – 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	Un générique
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



m erreur quadratique moyenne
w erreur quadratique dans le cas le plus défavorable
C corrélation

Figure V.4/J.144 – Comparaisons avec les tests d'évaluation subjective

La Figure V.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 V.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.500-7.

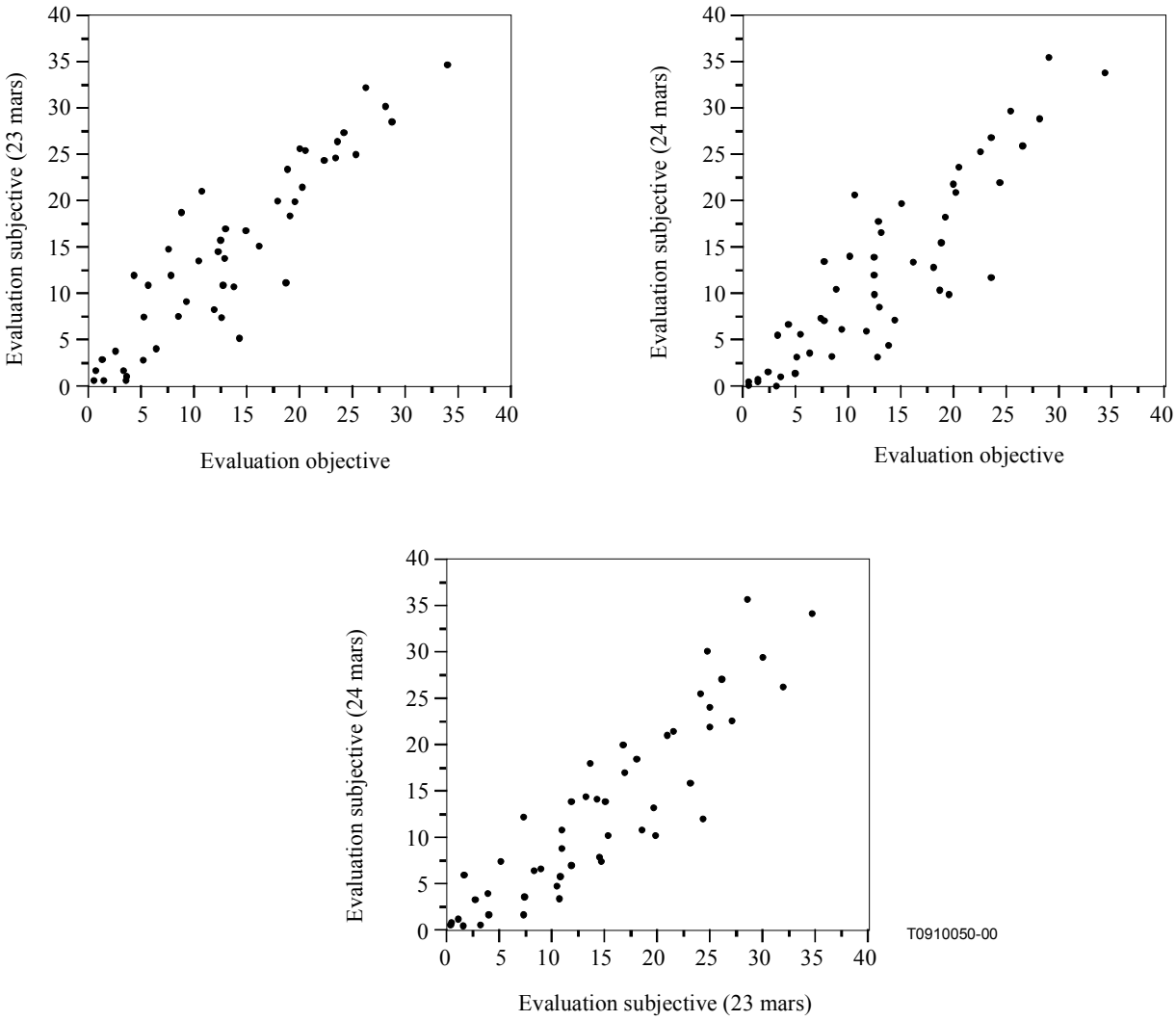


Figure V.5/J.144 – Comparaisons entre l'évaluation objective et les deux évaluations subjectives

V.5 Références

[1] *2nd version of Table defining video quality classes*, Expert meeting on subjective and objective video quality assessment, Turin, 14-16 octobre 1997.

[2] *Progress report on development of digital compressed picture quality assessment system in Japan*, SG 9 Document D15 Genève, 21-25 avril 1997.

APPENDICE VI

EPFL

Le modèle de mesure de la distorsion perçue (PDM, *perceptual distortion metric*) soumis par l'EPFL est fondé sur un modèle spatio-temporel du système visuel humain. Il comprend quatre phases, qui s'appliquent aussi bien à la séquence de référence qu'à la séquence traitée. Dans la première phase, l'entrée est convertie en espace de couleurs opposées. Dans la deuxième, on procède à une décomposition spatio-temporelle en canaux visuels distincts de fréquence temporelle, fréquence spatiale et orientation différentes. Dans la troisième, on modélise les effets du masquage de motif en simulant des mécanismes excitateurs et inhibiteurs conformément à un modèle de contrôle du gain de contraste. Dans la quatrième et dernière phase du modèle, qui sert de phase de rassemblement et de détection, on calcule une mesure de distorsion à partir de la différence entre la sortie du capteur de la séquence de référence et celle du capteur de la séquence traitée.

APPENDICE VII

NASA

VII.1 Introduction

L'infrastructure naissante de la vidéo numérique exige une composante déterminante: un moyen fiable permettant la mesure automatique de la qualité visuelle. Une telle possibilité se révèle cruciale pour pouvoir évaluer les codecs, contrôler les émissions de radiodiffusion et garantir une efficacité maximale quant à la compression des sources et l'utilisation des largeurs de bande de communication. Le présent appendice présente une nouvelle méthode de mesure de la qualité vidéo, appelée qualité vidéo numérique (DVQ, *digital video quality*), qui peut être utilisée pour la mesure automatique de la qualité visuelle.

VII.2 La méthode DVQ

Toutes les méthodes de mesure de la qualité vidéo sont intrinsèquement fondées sur des modèles de la vision humaine. Dans cette méthode de mesure DVQ, on tente d'incorporer de nombreux aspects de la sensibilité visuelle humaine dans un algorithme de traitement d'image simple. La simplicité est un but important, étant donné qu'on aimerait que le modèle fonctionne en temps réel et ne nécessite que des ressources de calcul modestes. L'un des éléments les plus complexes et les plus gourmands en temps des autres modèles proposés correspond aux opérations de filtrage spatial employées pour réaliser les multiples filtres spatiaux passe-bande qui sont caractéristiques de la vision humaine. Nous accélérons cette étape en utilisant la transformée discrète en cosinus (DCT) pour la décomposition en canaux spatiaux. Cela constitue un avantage certain parce que des logiciels et des matériels efficaces sont disponibles pour cette transformation et parce que, dans de nombreuses applications, la transformation peut déjà avoir été faite dans le cadre du processus de compression.

La Figure VII.1 propose un aperçu général des étapes de traitement de la méthode de mesure DVQ. Ces étapes sont décrites de manière plus détaillée dans d'autres documents [1] à [3] et nous n'en donnons ici qu'un bref aperçu. L'entrée du modèle est un couple de séquences d'images en couleur: une séquence de référence (R) et une séquence de test (T). La première étape consiste à appliquer diverses opérations (échantillonnage, recadrage, transformations de la couleur), qui servent à restreindre le traitement à une région particulière et à exprimer les séquences dans un espace de couleurs perçues. Dans cette étape, on procède aussi au désentrelacement et à la dé-gamma-correction du signal vidéo d'entrée. Les séquences sont alors soumises à une subdivision en blocs (BLK, *blocking*) et à une transformée discrète en cosinus (DCT) et les résultats sont ensuite transformés en contraste local (LC, *local contrast*). Le contraste local est le rapport entre l'amplitude

DCT et l'amplitude DC pour le bloc correspondant. L'étape suivante consiste en une opération de filtrage temporel (TF, *temporal filtering*), qui implémente la composante temporelle de la fonction de sensibilité au contraste, grâce à l'utilisation d'un filtre du second ordre discret, récursif et adapté. On convertit ensuite les résultats en différences tout juste perceptibles en divisant chaque coefficient DCT par le seuil visuel associé, ce qui correspondant à l'implémentation de la composante spatiale de la fonction de sensibilité au contraste (CSF, *contrast sensitivity function*). On procède lors de l'étape suivante à la différence entre les deux séquences. Cette séquence différence fait ensuite l'objet d'une opération de masquage de contraste (CM, *contrast masking*), qui dépend aussi de la séquence de référence. On peut enfin procéder à une sommation des différences masquées de plusieurs manières différentes, ce qui permet d'illustrer l'erreur perçue selon diverses dimensions (POOL). L'erreur sommée peut être convertie en qualité visuelle (VQ, *visual quality*).

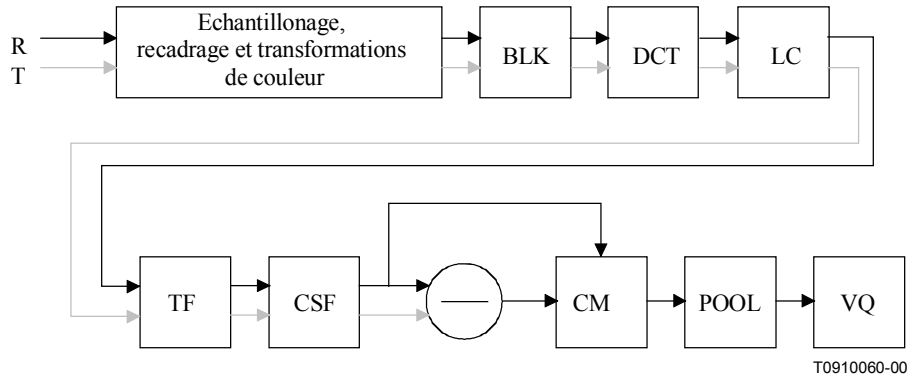


Figure VII.1/J.144 – Aperçu général des étapes de traitement de la méthode de mesure DVQ

On a estimé les paramètres de la méthode à partir de données psychophysiques provenant à la fois de la littérature existante et de mesures de visibilité de l'erreur de quantification dynamique DCT.

VII.2.1 Entrée

L'entrée du modèle est un couple de séquences d'images en couleur. Ses dimensions sont {s,f,c,y,x}, avec: s = séquence (2), f = images, c = couleur (3), y = lignes et x = colonnes. La première des deux séquences est la référence, la seconde est le test. Le test sera généralement différent de la référence en présence d'artéfacts de compression. L'espace de couleurs en entrée doit être défini de manière suffisamment détaillée pour pouvoir être transformé en coordonnées CIE, en spécifiant par exemple les coordonnées gamma et de chromatisme pour chaque couleur primaire. On utilise dans le présent appendice deux exemples courants: un espace linéaire RGB (gamma = 1) et un espace YCbCr avec gamma = 2,2.

VII.2.2 Transformations de couleur

La première étape du processus consiste à convertir les deux séquences d'images vers l'espace de couleurs YOZ, que nous avons déjà utilisé pour modéliser les erreurs perçues relatives à la compression d'images fixes. Les trois composantes de cet espace sont: Y (luminance CIE en candelas/m^2), O (un canal de couleur opposée donné par $O = \{X = 0{,}47;\ Y = -0{,}37;\ Z = -0{,}1\}$ et un canal bleu donné par la coordonnée Z CIE. La transformation vers l'espace YOZ comprend généralement:

1) une transformation gamma suivie;

2) d'une transformation linéaire de couleurs.

Ces opérations ne modifient pas les dimensions de l'entrée.

VII.2.3 DCT par blocs

Une DCT par blocs est à présent appliquée à chaque image dans chaque canal de couleur. {s, f, c, by, bx, v, u} constituent les dimensions du résultat. by et bx désignent les nombres de blocs selon les directions verticale et horizontale, et maintenant v = u = 8.

VII.2.4 Contraste local

Les coefficients DCT sont convertis en unités de contraste local de la manière suivante. Les coefficients DC sont d'abord extraits de tous les blocs. Ils sont ensuite filtrés temporellement par un filtre IIR passe-bas du premier ordre de gain 1 et de constante de temps τ_1. Puis on divise bloc par bloc les coefficients DCT par les coefficients DC filtrés: les blocs Y et Z sont divisés par les coefficients DC Y et Z, et O est divisé par le coefficient DC Y. On ajoute dans chaque cas une très petite constante au dénominateur pour empêcher la division par zéro. Enfin, on module les quotients grâce aux amplitudes relatives de leurs coefficients correspondants à une fonction unité de base de contraste. Ces opérations convertissent chaque coefficient DCT en un nombre compris entre –1 et 1, qui exprime l'amplitude de la fonction de base correspondante comme une fraction de la luminance moyenne dans le bloc considéré.

Les coefficients DC sont eux-mêmes convertis de manière similaire: on soustrait la moyenne continue pour l'ensemble de l'image et on divise le résultat par cette moyenne.

VII.2.5 Filtrage temporel

Les deux séquences font l'objet d'un filtrage temporel. On utilise un filtre IIR temporel du second ordre, semblable à celui décrit plus haut pour l'adaptation du bruit dynamique DCT. L'utilisation d'un filtre IIR minimise le nombre de données à stocker en mémoire. Pour une simplicité plus grande encore, on peut utiliser un filtre du premier ordre.

VII.2.6 Conversion JND

Les coefficients DCT, maintenant exprimés sous forme de contraste local, sont à présent convertis en différences tout juste perceptibles (JND, *just-noticeable differences*) par division par leur seuil spatial respectif. Ces seuils sont d'abord multipliés par un facteur de sommation spatiales, dont le rôle et l'évaluation sont décrits plus bas. Les seuils s'appliquant aux deux canaux de couleurs résultent des seuils de luminance[3] ou reposent sur d'autres seuils chromatiques. Après conversion aux différences JND, les coefficients des deux séquences font l'objet d'une soustraction qui génère une *séquence différence*.

VII.2.7 Masquage de contraste

Pour le masquage de contraste, on commence par élaborer une *séquence de masquage*, dont le début correspond à la séquence de référence, après conversion JND. Cette séquence est rectifiée, puis filtrée temporellement par un filtre IIR discret passe-bas du premier ordre présentant un gain g_1 et une constante de temps τ_2. On élève ensuite ces valeurs à la puissance *m* (toute valeur inférieure à 1 étant remplacée par 1) et on utilise le résultat pour diviser la séquence différence. Ce processus simule le résultat classique de masquage de contraste selon lequel les contrastes inférieurs au seuil n'ont pas d'effet de masquage, alors qu'au-dessus du seuil, les effets s'accroissent comme la puissance mième du masque de contraste exprimé en différences JND.

VII.2.8 Sommation de Minkowski

{f, c, by, bx, v, u} constituent les dimensions du résultat de cette étape. Rappelons que *f* désigne les images, *c* les canaux de couleur, *by* et *bx* les nombres de blocs selon les directions verticale et horizontale, *v* et *u (v = u)* les fréquences verticale et horizontale. On peut ensuite combiner les erreurs élémentaires suivant diverses dimensions (ou toutes les dimensions) pour générer des

mesures récapitulatives de l'erreur visuelle. On effectue cette sommation en utilisant la méthode de mesure de Minkowski:

$$J_x = M\left(j_{f,c,by,bx,y,x},\beta\right)=\left(\sum_x \left|j_{f,c,by,bx,y,x}\right|^{\beta}\right)^{\frac{1}{\beta}} \qquad \text{(VII-1)}$$

Nous avons indiqué dans cette équation une sommation suivant l'ensemble des six dimensions, mais tout sous-ensemble de ces dimensions peut également être envisagé. Un avantage de la formule de Minkowski est qu'elle permet le calcul en cascade. Par exemple, on peut dans un premier temps sommer uniquement suivant la dimension de couleur *(c)*, puis l'on peut par la suite procéder à d'autres sommations, par exemple suivant les dimensions de blocs (*by* et *bx)*.

VII.3 Evaluation

Nous avons évalué la performance de la méthode de mesure de la qualité vidéo DVQ en comparant ses prédictions aux avis sur les détériorations émis par 25 observateurs ayant visionné cinq séquences de référence traitées par 12 HRC. Le modèle de mesure DVQ présente des performances nettement meilleures que des modèles fondés uniquement sur un débit binaire ou sur une erreur quadratique moyenne (rms, *root mean square*). La qualité des prédictions laisse à penser que la méthode peut être utilisée dans la pratique. Nous avons soumis il y a peu notre algorithme au projet d'évaluation du groupe d'experts sur la qualité vidéo (VQEG, *video quality experts group*). Cet algorithme a donné des résultats tout à fait satisfaisants pour une large gamme de sous-ensembles HRC, en particulier dans la plage haute qualité, dans laquelle on a observé une corrélation de classement de 0,72. Deux des conditions testées (cycles enregistrement/lecture professionnels 1/2 pouce multi-génération et erreurs de transmission) sortent du cadre de notre modèle. Après élimination de ces HRC, la corrélation de classement de Spearman était de 0,82.

VII.4 Références

[1] WATSON (A.B.): Toward a perceptual video quality metric in Human Vision, *Visual Processing, and Digital Display VIII*, San Jose, CA: SPIE, Bellingham, WA, 1998.

[2] WATSON (A.B.), *et al*.: Design and performance of a digital video quality metric in Human Vision, *Visual Processing, and Digital Display IX.* San Jose, CA: SPIE, Bellingham, WA, 1999.

[3] WATSON (A.B.), HU (J.), McGOWAN (J.F.), III: *DVQ:* A digital video quality metric based on human vision, *Journal of Electronic Imaging*, 2000. in press.

APPENDICE VIII

KPN/Swisscom CT

VIII.1 Introduction

Dans le cadre de mesure de la qualité vidéo perçue (PVQM*, perceptual video quality mesure*), les signaux physiques d'entrée et de sortie de l'instrument testé (par exemple un codec, ou une chaîne de transmission) sont traduits sous forme de représentations psychophysiques (voir Figure VIII.1) qui se rapprochent autant que possible des représentations internes des signaux audio/vidéo (représentations internes au cerveau humain). On juge de la qualité de l'instrument en se fondant sur les différences de représentation interne. Dans le cadre de la PVQM, la représentation interne, de laquelle on déduit la qualité, est telle que les distorsions spatiales et temporelles sont toutes deux prises en compte par la méthode de mesure.

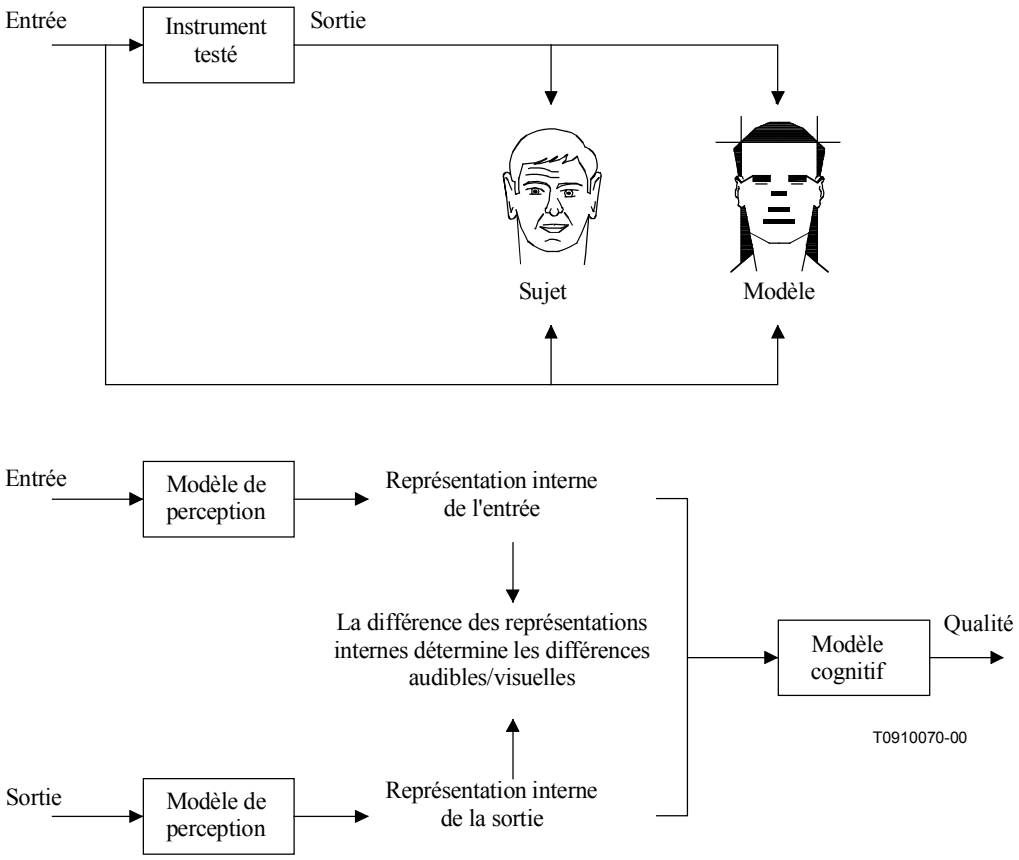


Figure VIII.1/J.144

Aperçu général des principes de base utilisés pour l'élaboration de la PVQM. Un modèle informatique du sujet, constitué d'un modèle de perception et d'un modèle cognitif, sert à comparer la sortie de l'instrument testé (par exemple un codec vidéo) à l'entrée, en utilisant un signal vidéo quelconque.

L'algorithme comprend un alignement spatio-temporel de luminance qui permet d'utiliser la PVQM dans la pratique. Il est bien connu que les changements généraux de brillance et de contraste n'ont qu'une incidence limitée sur la qualité perçue de façon subjective, surtout si on les compare aux incidences de distorsions comme le tuilage. Dans le cadre de la PVQM, on quantifie cet effet en utilisant une adaptation spéciale brillance/contraste de la séquence vidéo distordue. Il est de plus évident que l'on ne peut calculer une mesure significative de la distorsion que si l'on sait quelles parties du signal d'entrée et du signal de sortie doivent être comparées. C'est pourquoi la PVQM utilise une sorte de procédure d'alignement spatio-temporel avec mise en correspondance des blocs avant que les véritables mesures ne soient effectuées.

La partie spatiale de l'analyse de la luminance est fondée sur la détection de bord du signal Y, tandis que la partie temporelle est fondée sur l'analyse d'images différence du signal Y. Chacun sait que le système visuel humain (HVS) est beaucoup plus sensible à l'acuité de la composante de luminance qu'à celle des composantes de chrominance. En outre, le système HVS a une fonction de sensibilité au contraste qui décroît aux hautes fréquences spatiales. Ces fondements du système HVS sont reflétés dans la première étape de l'algorithme de mesure PVQM qui fournit une approximation au premier ordre des fonctions de sensibilité au contraste des signaux de luminance et de chrominance.

Dans la deuxième étape, l'irrégularité de la luminance Y est calculée sous la forme d'une représentation de signal contenant les aspects les plus importants de l'image. Pour obtenir cette irrégularité, on calcule d'abord le gradient local du signal de luminance dans chaque image. L'erreur relative d'irrégularité entre la vidéo d'entrée et la vidéo de sortie est sommée dans l'espace et le temps en utilisant les mesures p de Lebesgue.

Dans la troisième étape, l'erreur de chrominance est calculée comme une moyenne pondérée de l'erreur de couleur des composantes Cb et Cr (normalisées suivant la saturation locale) avec une dominance de la composante Cr.

Dans la dernière étape, les trois différents indicateurs sont traduits sous la forme d'un indicateur de qualité unique, au moyen d'une simple régression linéaire multiple, qui donne une bonne corrélation de la qualité vidéo d'ensemble de la séquence telle qu'elle est perçue subjectivement. Le centre de recherche KPN a validé la méthode à partir d'une grande variété de bases de données contenant à la fois des distorsions générées par des codecs (MPEG, UIT-T H.263, etc.) et des distorsions créées artificiellement. La corrélation entre les valeurs objectives PVQM et les notes moyennes d'opinion subjectives, établie sur l'ensemble des bases de données significatives, est supérieure à 0,9.

VIII.2 Références

[1] BEERENDS (J.G.), HEKSTRA (A.P.): Objective measurement of video quality, *Commission d'études 12 de l'UIT-T, Document COM 12-7,* février 1997.

APPENDICE IX

NTIA

Introduction

Le présent appendice présente en détail l'algorithme utilisé pour une méthode de mesure de la qualité vidéo (VQM) dont les résultats sont bien corrélés avec les jugements de qualité subjectifs relatifs à des scènes vidéo. La présente version de la méthode VQM comporte plusieurs améliorations du modèle qui a été soumis au groupe d'experts sur la qualité vidéo (VQEG). Ces améliorations ont été apportées avant que les données subjectives du VQEG ne soient disponibles [1]. Outre le fait qu'elle fournisse des estimations fondées sur la perception et indépendantes de la technologie, la méthode VQM présente une complexité calculatoire faible et peut être utilisée pour les applications de surveillance de la qualité s'effectuant en service, en temps réel et en continu. Les résultats présentés comparent les mesures VQM aux notes moyennes d'opinion issues de neuf tests subjectifs différents avec double stimulus couvrant une grande diversité de scènes, de systèmes vidéo et de techniques de codage. Sept de ces groupes de données contiennent essentiellement des scènes vidéo issues d'applications de radiodiffusion de qualité contribution et de qualité distribution (> 1,5 Mbit/s), alors que les deux autres groupes de données comportent surtout des scènes vidéo issues d'applications multimédia (< 1,5 Mbit/s).

IX.1 Description de l'algorithme VQM

L'algorithme VQM s'appuie sur une combinaison linéaire de quatre paramètres qui ont été optimisés pour la distance d'observation normalisée de six hauteurs d'image. Trois paramètres sont extraits des gradients spatiaux de la composante de luminance Y issue des flux vidéo d'entrée et de sortie de la Rec. UIT-R BT.601 [2], alors que le quatrième paramètre provient du vecteur formé par les composantes de chrominance (C_B, C_R).

On suppose que les flux vidéo échantillonnés d'entrée et de sortie ont été étalonnés avant que les processus décrits ici ne s'exécutent. L'étalonnage comprend la compensation du gain du système et du décalage de niveau, ainsi que l'alignement spatial et temporel des images.

IX.2 Paramètres de gradient spatial

La Figure IX.1 donne un aperçu général de l'algorithme utilisé pour extraire les paramètres de gradient spatial. Les composantes Y des flux vidéo d'entrée et de sortie sont traitées par des filtres de souligné des contours suivant les directions horizontale et verticale. Ces flux vidéo traités sont ensuite subdivisés en régions spatio-temporelles (S-T) à partir desquelles on extrait des caractéristiques (ou des statistiques récapitulatives) qui quantifient l'activité spatiale en fonction de l'angle d'orientation. Ces caractéristiques sont ensuite écrêtées à l'extrémité inférieure pour simuler les seuils de perceptibilité. Puis les distorsions de qualité vidéo consécutives aux gains et pertes associés aux valeurs des caractéristiques sont calculées pour chaque région S-T; on compare pour cela leurs valeurs d'entrée et de sortie en utilisant des relations fonctionnelles qui simulent le masquage visuel des dégradations. Ces distorsions sont ensuite rassemblées spatialement (regroupement spatial) et temporellement (regroupement temporel) afin de générer des paramètres de qualité pour un clip vidéo de 5 à 10 secondes de durée nominale.

Les filtres de souligné des contours, la taille de la région S-T et les seuils de perceptibilité présentés ici ont été optimisés en se fondant sur la corrélation avec les distorsions perçues à une distance de 6 hauteurs d'image.

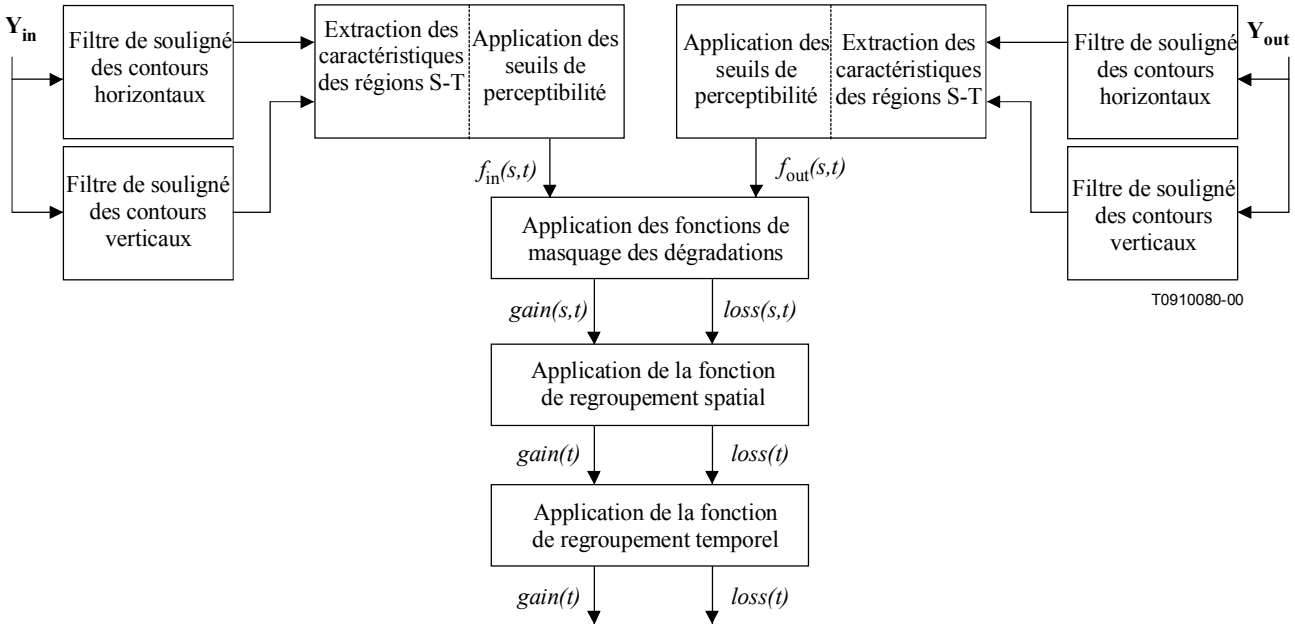


Figure IX.1/J.144 – Aperçu de l'algorithme utilisé pour l'extraction des paramètres de gradient spatial

IX.3 Filtres de souligné des contours

Les *images* vidéo d'entrée et de sortie sont traitées en premier lieu par des filtres de souligné des contours horizontaux et verticaux qui soulignent les contours tout en réduisant le bruit. Les deux filtres présentés sur la Figure IX.2 sont appliqués séparément. Le premier (filtre de gauche) a pour rôle de souligner les différences horizontales entre pixels tout en procédant à un lissage vertical, tandis que le second (filtre de droite) souligne les différences verticales entre pixels tout en effectuant un lissage horizontal.

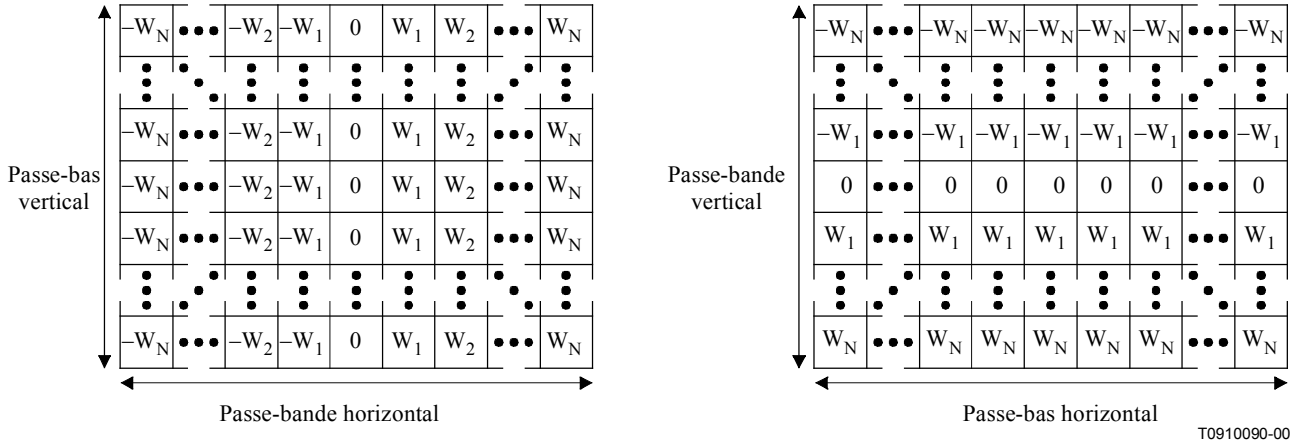


Figure IX.2/J.144 – Filtres de souligné des contours

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

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

x représente le déplacement en pixels compté à partir du 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. L'expérience a montré que le filtrage optimal en passe-bande horizontal pour une distance d'observation 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:

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

Il faut remarquer que les filtres de la Figure IX.2 présentent une réponse passe-bas plate. Cette réponse a généré la meilleure estimation de qualité et présente de plus l'avantage d'offrir une grande efficacité calculatoire (dans le cas du filtre de gauche de la Figure IX.2 par exemple, il suffit de sommer les pixels d'une colonne et de multiplier le résultat par le coefficient de pondération).

IX.4 Taille de région S-T

Les flux vidéo d'entrée et de sortie aux contours horizontaux et verticaux soulignés sont tous subdivisés en régions S-T localisées. La Figure IX.3 donne la taille de région S-T (8 pixels horizontaux × 8 lignes verticales × 6 images vidéo) qui permet d'obtenir la corrélation maximale avec les évaluations subjectives. Il faut noter toutefois qu'on a observé que la corrélation se détériorait *lentement* à mesure que l'on s'éloignait de l'optimum. L'utilisation de largeurs horizontales ou verticales allant jusqu'à 32 pixels ou lignes et de largeurs temporelles allant jusqu'à 30 images conduit à des résultats satisfaisants, ce qui donne au concepteur de système de mesure objective une très grande souplesse pour adapter les techniques présentées ici aux différentes tailles de région S-T.

On extrait les caractéristiques de chaque région S-T par le calcul de statistiques récapitulatives sur cette région. Une description détaillée des caractéristiques extraites est donnée dans le paragraphe IX.5.

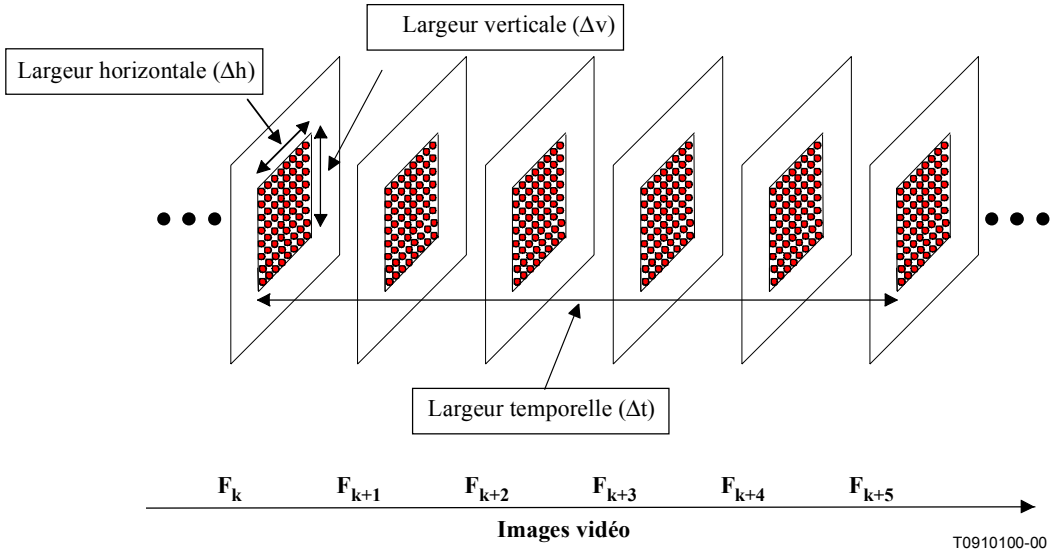


Figure IX.3/J.144 – Taille optimale de région spatio-temporelle (S-T) pour l'extraction des caractéristiques

IX.5 Description des caractéristiques

Le présent paragraphe décrit l'extraction de deux caractéristiques d'activité spatiale issues des régions S-T des flux vidéo d'entrée et de sortie aux contours soulignés qui proviennent du paragraphe IX.4. Le filtre présent sur la Figure IX.2 (à gauche) souligne les gradients spatiaux suivant la direction horizontale (H) alors que le transposé de ce filtre renforce 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 IX.4: la réponse du filtre H y constitue 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:

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

et:

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

La première caractéristique f_1 se calcule simplement comme l'écart type (*stdev, standard deviation*) sur la région S-T des échantillons *R(i, j, t)*. Elle est ensuite écrêtée suivant le seuil de perceptibilité P (ce qui signifie que f_1 est fixé à P si le calcul de stdev donne un résultat inférieur à P). On obtient ainsi:

$$f_1 = \left\{ stdev\left[R(i,j,t)\right] \right\}\Big|_P : \; i,j,t \in \left\{ \text{Région S-T} \right\}$$

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 niveau de seuil P recommandé pour cette caractéristique est de 12.

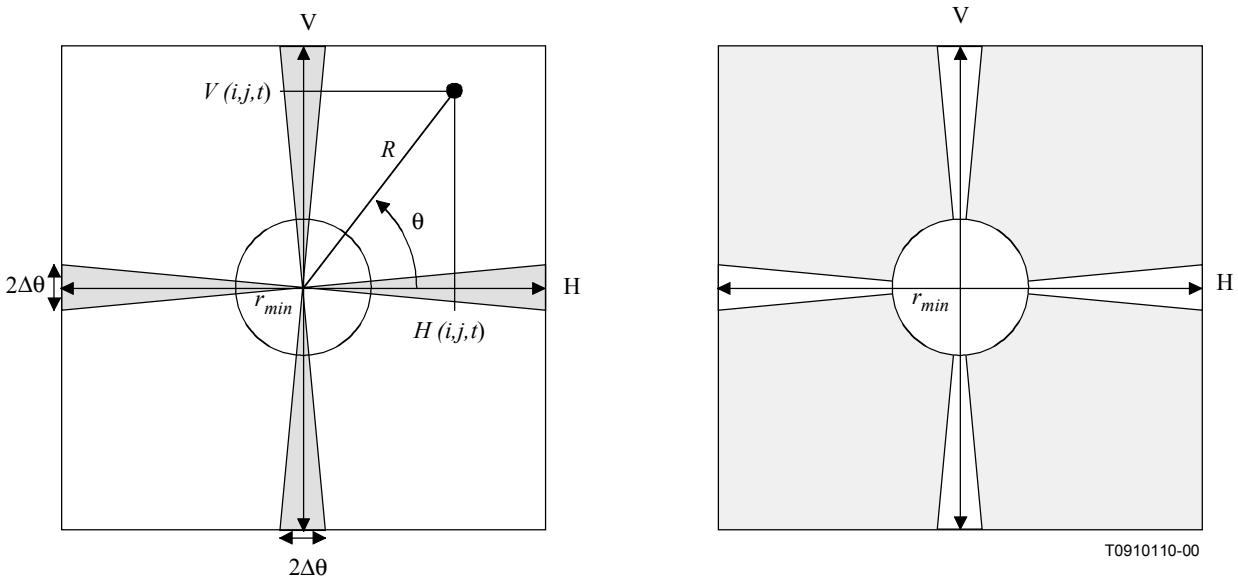


Figure IX.4/J.144 – Subdivision de l'activité spatiale horizontale (H) et verticale (V) en distributions *HV* (gauche) et $\overline{HV}$ (droite)

La seconde caractéristique f_2 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 IX.4. L'image avec les gradients horizontaux et verticaux, notée *HV*, contient les pixels *R(i, j, t)* possédant des contours horizontaux ou verticaux (les pixels dont les contours sont des diagonales sont mis à zéro). L'image avec les gradients en diagonal, notée $\overline{HV}$, contient les pixels *R(i, j, t)* présentant des contours en diagonale (les pixels de contours horizontaux ou verticaux sont annulés). Les amplitudes de gradient *R(i, j, t)* inférieures à $r_{\min}$ sont mis 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:

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

avec:

$$i,j,t\in\{\text{Région S - T}\}$$

Pour les calculs de HV et $\overline{HV}$ ci-dessus, on recommande la valeur 20 pour $r_{\min}$ et 0,05236 radians pour $\Delta\theta$. La caractéristique f_2 pour une région S-T est ensuite obtenue comme le rapport entre la valeur moyenne de HV et la valeur moyenne de $\overline{HV}$, ces valeurs moyennes étant écrêtées à leur seuil de perceptibilité P. On obtient ainsi:

$$f_2 = \{mean[HV(i,j,t)]\}\big|_P \Big/ \{mean[\overline{HV}(i,j,t)]\}\big|_P$$

Le seuil de perceptibilité P recommandé pour les moyennes de HV et $\overline{HV}$ est de 3. La caractéristique f_2 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 en diagonale, la valeur de f_2 en sortie sera moins élevée qu'en entrée. D'autre part, si des contours horizontaux ou verticaux aberrants sont introduits (par exemple sous forme de distorsions liée à une subdivision en blocs ou à un quadrillage), la valeur de f_2 sera alors plus élevée en sortie qu'en entrée. La caractéristique f_2 fournit aussi un moyen simple pour tenir compte des variations de sensibilité du système visuel humain en fonction de l'orientation angulaire.

Dans la suite de l'étude, on notera $f_{in}(s, t)$ le flux caractéristique d'entrée et $f_{out}(s, t)$ le flux correspondant en sortie. Les indices s et t désignent respectivement les positions spatiale et temporelle de la région S-T comprise dans les flux vidéo étalonnés d'entrée et de sortie.

IX.6 Fonctions de masquage des détériorations

On calcule ensuite la détérioration perçue dans chaque région S-T à l'aide d'une fonction modélisant le masquage visuel des détériorations. Les gains et les pertes doivent être traités séparément, puisqu'ils ont des effets fondamentalement différents sur la perception de la qualité (par exemple, les pertes d'activité spatiale dues au flou et les gains d'activité spatiale dus au bruit ou à la subdivision en blocs). Parmi les nombreuses fonctions de comparaison que nous avons évaluées, deux ont logiquement généré la meilleure corrélation avec les évaluations subjectives. Ces fonctions de comparaison modélisent la perceptibilité des dégradations spatiales ou temporelles. Pour une région S-T donnée, les distorsions de gain et de perte sont calculées de la manière suivante:

$$gain(s,t) = pp\left\{\log_{10}\left[\frac{f_{out}(s,t)}{f_{in}(s,t)}\right]\right\}$$

et:

$$loss(s,t) = np\left\{\frac{f_{out}(s,t) - f_{in}(s,t)}{f_{in}(s,t)}\right\}$$

pp désigne la partie positive de l'exploitant (les valeurs négatives sont ainsi remplacées par zéro) alors que np en désigne la partie négative (les valeurs positives sont mises à zéro). Ces fonctions de masquage visuel ont pour effet que la perception de dégradation est inversement proportionnelle à la quantité d'activité spatiale ou temporelle localisée sur la scène d'entrée. En d'autres termes, les dégradations spatiales deviennent moins visibles à mesure que l'activité spatiale de la scène d'entrée s'accroît (masquage spatial), tandis que les dégradations temporelles perdent en visibilité à mesure que l'activité temporelle de la scène d'entrée augmente (masquage temporel). Alors que les fonctions de comparaison logarithmique et de comparaison de ratio présentent un comportement à peu près similaire, la fonction logarithmique tend à être légèrement plus avantageuse sur le plan des gains tandis que la fonction de ratio semble présenter des avantages plus nombreux lorsqu'on s'intéresse aux pertes.

IX.7 Fonction de regroupement spatial

Les détériorations des régions S-T de même indice temporel *t* sont ensuite regroupées à l'aide d'une fonction de regroupement spatial. Des études poussées ont montré que les fonctions de regroupement spatial optimales nécessitent en général une certaine forme de traitement du cas le plus défavorable. Les dégradations localisées tendent en effet à attirer l'attention du téléspectateur, ce qui fait que la partie la plus détériorée de l'image devient le facteur prédominant dans la prise de décision de la qualité subjective. La fonction de regroupement spatial est calculée pour chaque indice temporel *t* comme la moyenne (que l'on note *worst*_5%) des 5% de distorsions mesurées les plus mauvaises sur l'indice spatial *s*. Cela revient à classer les distorsions de gain pour chaque indice temporel et à moyenner les distorsions au-dessus du seuil de 95%. De la même façon, les distorsions de perte sont classées pour chaque indice temporel t, mais on utilise la moyenne des distorsions inférieures au seuil de 5% (puisque les pertes sont négatives). Appliquer la fonction $worst\_5\%_{space}$ génère un historique des échantillons de gain et de perte [c'est-à-dire *gain(t)* et *loss(t)*], qui doivent ensuite être regroupés temporellement.

IX.8 Fonctions de regroupement temporel

En fin de compte, on rassemble les résultats de la fonction de regroupement spatial en utilisant une fonction de regroupement temporel qui génère un paramètre objectif pour le clip vidéo considéré, d'une durée nominale de 5 à 10 secondes. Les téléspectateurs semblent utiliser plusieurs fonctions de regroupement temporel lorsqu'ils évaluent subjectivement les clips vidéo d'une durée de 9 à 10 secondes. Une des fonctions de regroupement temporel reflète le niveau de qualité moyen du clip tandis que l'autre révèle la plus mauvaise qualité en transitoire du clip (en l'occurrence, les erreurs de transmission numérique génèrent en principe 1 à 2 secondes de perturbation de la vidéo de sortie).

La moyenne temporelle (notée $mean_{time}$) paraît révélatrice de la qualité moyenne observée durant cette période. En ce qui concerne la plus mauvaise qualité en transitoire, le niveau pendant 10% du temps pour les paramètres de perte (noté $10\%_{time}$) et le niveau pendant 90% du temps pour les paramètres de gain (noté $90\%_{time}$) semblent avoir l'impact subjectif le plus important (en d'autres termes, les échantillons de l'historique du paramètre de perte sont classés et le niveau à 10% est utilisé; les échantillons de l'historique du paramètre de gain sont classés et le niveau à 90% est utilisé). Il est nécessaire de procéder à des études ultérieures pour optimiser ces fonctions de regroupement temporel.

IX.9 Trois paramètres de gradient spatial

Les trois paramètres de gradient spatial utilisés pour calculer la qualité VQM sont donnés par:

f_1*_loss* (utilisation de la fonction de regroupement temporel $10\%_{time}$);

f_2*_loss* (utilisation de la fonction de regroupement temporel $mean_{time}$);

f_2*_gain* (utilisation de la fonction de regroupement temporel $mean_{time}$).

Les caractéristiques f_1 et f_2 sont décrites dans le paragraphe IX.5, les fonctions de gain et de perte sont données dans le paragraphe IX.6, la fonction de regroupement spatial est donné au paragraphe IX.7 et les fonctions de regroupement temporel sont données dans le paragraphe IX.8.

IX.10 Paramètre de chrominance

Le présent paragraphe présente un paramètre de distorsion de chrominance unique intervenant dans le calcul de la qualité VQM. Soit $C_B(i, j, t)$ et $C_R(i, j, t)$ les valeurs de C_B et C_R (définies dans la Rec. UIT-R BT.601) pour un pixel donné d'une image repéré par sa ligne *i*, sa colonne *j* et le temps *t*. Les composantes du vecteur à deux dimensions caractéristique de la chrominance, $\underline{f}_C$, sont calculées

respectivement de manière simple comme étant la moyenne (*mean*) sur la région S-T des échantillons $C_B(i,j,t)$ et $C_R(i,j,t)$, avec un poids de perception plus grand attribué à la composante C_R:

$$\underline{f}_C(s,t) = \left(mean\left[C_B(i,j,t)\right], W_R * mean\left[C_R(i,j,t)\right]\right) : \; i,j,t \in \{\text{Région S-T}\}, \text{ et } W_R = 1{,}5 .$$

La taille recommandée pour la région S-T est de 8 pixels horizontaux × 8 lignes verticales × 1 image vidéo (il s'agit en fait de 4 pixels C_B and C_R horizontaux, puisque ces signaux sont sous-échantillonés d'un facteur deux dans la Rec. UIT-R BT.601). La distorsion de chrominance pour chaque région S-T est notée d_C (s, t), où s et t sont les indices repérant les positions spatiale et temporelle de la région S-T au sein des flux vidéo étalonnés entrant et sortant. Elle se calcule comme la distance euclidienne entre les vecteurs des caractéristiques de chrominance d'entrée et de sortie $\underline{f}_{Cin}$ et $\underline{f}_{Cout}$ et figure en pointillé sur la Figure IX.5. Elle s'exprime par:

$$d_C(s,t) = \left\| \underline{f}_{Cout}(s,t) - \underline{f}_{Cin}(s,t) \right\|$$

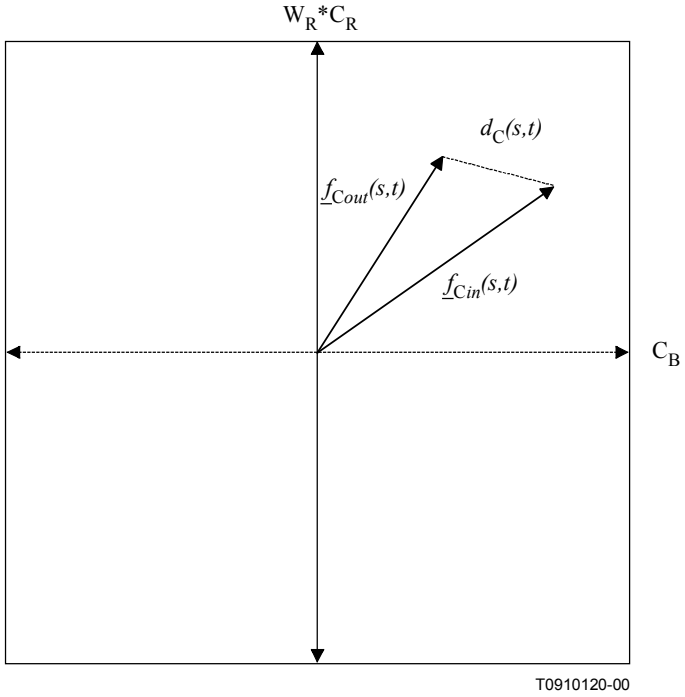


Figure IX.5/J.144 – Calcul de la distorsion de chrominance $d_C(s, t)$ pour une région S-T

L'écart type spatial (noté $stdev_{space}$) constitue la fonction de regroupement spatial optimale pour d_C (s, t) et s'apparente à la fonction $worst\_5\%_{space}$ définie plus haut. La fonction de regroupement temporel optimale correspond au niveau pendant 10% du temps (noté $10\%_{time}$), qui représente le niveau de distorsion que l'on trouve presque toujours. La distorsion de chrominance présente après les regroupements spatial et temporel est écrêtée suivant le seuil de perceptibilité P = 0,8, puis on lui soustrait P, ce qui constitue la mesure d_C. En résumé, le paramètre de distorsion de chrominance d_C est donné par:

$$d_C = \left\{ 10\%_{\text{time}} \left[stdev_{\text{space}}\left(d_C(s,t)\right) \right] \right\}\Big|_P - P$$

IX.11 Calcul de la qualité VQM

La qualité VQM est calculée de la manière suivante:

$$VQM = -0{,}3609*f_1\_loss + 0{,}5031*(f_2\_loss)^2 + 0{,}1390*f_2\_gain + 0{,}0295*d_C$$

L'élévation au carré du paramètre f_2_loss est nécessaire pour en linéariser la réponse. Le paramètre f_1_loss doit être précédé d'un coefficient multiplicatif négatif, car il est toujours inférieur ou égal à zéro (ce qui est aussi toujours le cas du paramètre f_2_loss, mais c'est le carré de ce paramètre que l'on utilise dans le calcul de la qualité VQM). Les paramètres f_2_gain et d_C sont toujours supérieurs ou égaux à zéro. La qualité VQM calculée de cette manière présentera toujours des valeurs positives ou nulles et une valeur maximale nominale égale à un. La mesure de la qualité VQM peut parfois dépasser un pour des scènes vidéo extrêmement distordues.

IX.12 Description des groupes de données subjectives

Les neuf expériences subjectives ont eu lieu de 1992 à 1999. Tous les groupes de données ont été traités conformément à la version la plus récente de la Rec. UIT-R BT.500-9 [3] disponible au moment de l'expérience. Tous les groupes de données ont utilisé des scènes ayant une durée de 9 à 10 secondes ainsi qu'une visualisation à double stimulus (les téléspectateurs ont vu à la fois les séquences d'origine et les séquences détériorées). A des fins de concision, seul un résumé de chaque expérience subjective est proposé ici. Le lecteur désirant des descriptions plus complètes est invité à se reporter aux références jointes.

Groupe de données numéro un [4, 5]

Un groupe de 48 téléspectateurs a évalué un total de 132 clips vidéo générés par le couplage aléatoire ou déterministe entre 36 scènes de test et 27 systèmes vidéo. Les 36 scènes de test comprenaient des quantités très diverses d'informations spatiales et temporelles. Les 27 systèmes vidéo comportaient des systèmes numériques de compression vidéo fonctionnant à des débits binaires allant de 56 kbit/s à 45 Mbit/s avec des taux d'erreurs contrôlés, des cycles de codage/décodage NTSC, des cycles enregistrement/lecture VHS et S-VHS et une diffusion en ondes métriques. Les téléspectateurs ont d'abord visionné la version d'origine, puis la version dégradée avant d'être invités à évaluer la différence de qualité perçue à l'aide de l'échelle de détérioration à 5 notes (imperceptible, perceptible mais non gênant, légèrement gênant, gênant, très gênant).

Groupe de données numéro deux [6, 7]

Les groupes de visionnage comprenaient un total de 30 téléspectateurs détachés par trois laboratoires différents. Ces téléspectateurs ont évalué 600 clips vidéo générés par le couplage entre 25 scènes de test et 24 systèmes vidéo. Les 25 scènes de test comportaient des scènes issues de 5 catégories:

1) une personne, dont on voit surtout la tête et les épaules;
2) une personne avec des graphiques et/ou des détails supplémentaires;
3) plus d'une personne;
4) des graphiques avec des pointillés;
5) un mouvement rapide d'objet et/ou de caméra.

Les 24 systèmes vidéo comprenaient des systèmes de téléconférence vidéo propriétaires ou normalisés fonctionnant à des débits binaires compris entre 56 kbit/s à 1,5 Mbit/s, avec des taux d'erreur contrôlés, un codec de 45 Mbit/s et un cycle d'enregistrement/lecture VHS. La procédure de test subjectif était la même que pour le groupe de données numéro un.

Groupe de données numéro trois [8]

Un groupe de 32 téléspectateurs a évalué la différence de qualité entre des scènes d'entrée avec des quantités contrôlées de bruit ajouté et la sortie correspondante compressée suivant la

norme MPEG-2. Le groupe de données était constitué d'un total de 105 clips vidéo générés par le couplage entre sept scènes de test (présentant trois niveaux de bruit différents) et cinq systèmes vidéo MPEG-2. Les sept scènes de test ont été choisies de manière à couvrir une gamme de détails spatiaux, de mouvement, de brillance et de contraste. Les cinq systèmes vidéo MPEG-2 ont fonctionné à des débits binaires compris entre 1,8 Mbit/s et 13,9 Mbit/s. Les téléspectateurs ont visionné l'entrée et la sortie traitée selon un ordre A/B aléatoire et ont été invités à évaluer la qualité de B en utilisant A comme référence. L'expérience s'est appuyée sur une échelle de comparaison à sept notes (B beaucoup plus mauvais que A, B plus mauvais que A, B légèrement plus mauvais que A, B identique à A, B légèrement meilleur que A, B meilleur que A, B bien meilleur que A).

Groupe de données numéro quatre [9]

Un groupe de 32 téléspectateurs a évalué un total de 112 clips vidéo générés par le couplage entre des sous-groupes comptant chacun huit scènes (le test comprenait un nombre total de 16 scènes) et 14 systèmes vidéo différents. Les 16 scènes de test couvraient une large gamme de détails spatiaux, de mouvement, de brillance et de contraste et comptaient des données scéniques provenant de scènes de test relatives au cinéma, au sport, à la nature ou à des thèmes classiques issus de la Rec. UIT-R BT.601. Les 14 systèmes vidéo étaient constitués de systèmes MPEG-2 fonctionnant à des débits binaires compris entre 2 et 36 Mbit/s, avec des taux d'erreur contrôlés, un MPEG-2 multi-génération, des cycles enregistrement/lecture professionnels 1/2 pouce multi-génération, le VHS et des systèmes de téléconférence vidéo fonctionnant à des débits binaires allant de 768 kbit/s à 1,5 Mbit/s. La procédure de test subjectif était la même que pour le groupe de données numéro trois.

Groupe de données numéro cinq [9]

Un groupe de 32 téléspectateurs a évalué un total de 42 clips vidéo générés par le couplage entre des sous-groupes comptant chacun six scènes (le test comprenait un nombre total de 12 scènes) et sept systèmes MPEG-2 différents. Les 12 scènes de test comprenaient des données relatives au sport et des scènes de test issues des thèmes classiques de la Rec. UIT-R BT.601. Les neuf systèmes MPEG-2 fonctionnaient à des débits binaires compris entre 2 Mbit/s et 8 Mbit/s. La procédure de test subjectif était la même que pour le groupe de données numéro trois.

Groupes de données numéro six à neuf [10]

Quatre groupes de données (haute qualité 525 lignes, basse qualité 525 lignes, haute qualité 625 lignes, basse qualité 625 lignes) composés chacun de 90 clips vidéo ont été générés par couplage entre dix scènes et neuf systèmes vidéo. Pour chaque groupe de données, un nombre total de 60 à 80 téléspectateurs provenant de quatre laboratoires différents (il y avait donc 15 à 20 téléspectateurs par laboratoire) ont évalué la qualité subjective en utilisant une échelle de qualité continue à double stimulus (DSCQS, *double stimulus continuous quality scale*). Les vingt scènes de test différentes (dix pour les 525 lignes, dix pour les 625 lignes) comprenaient des données de sport, des scènes de test classiques issues de la Rec. UIT-R BT.601, des graphiques animés et des photographies. Les systèmes vidéo comportaient des systèmes MPEG-2 fonctionnant à des débits binaires allant de 2 à 50 Mbit/s, des systèmes vidéo de téléconférence fonctionnant à 768 kbit/s et 1,5 Mbit/s, quelques systèmes avec des erreurs de transmission numérique, un MPEG-2 multi-génération, des cycles enregistrement/lecture professionnels 1/2 pouce multi-génération, pour lesquels on a utilisé des formats de signal en composantes et/ou composite.

IX.13 Résultats

Le Tableau IX.1 donne le coefficient de corrélation linéaire de Pearson entre les mesures VQM et chaque groupe de données subjectives. Le coefficient moyen est de 0,90.

La Figure IX.6 présente la courbe de dispersion des jugements de qualité subjectifs en fonction des mesures VQM pour les neuf groupes de données subjectives. Les notes moyennes d'opinion subjectives des neuf groupes de données y ont été tracées de façon à être comprises entre 0 et 1. On y

observe un coefficient de corrélation linéaire de Pearson entre les notes subjectives et les mesures VQM de 0,94 (cette valeur est supérieure à la moyenne des valeurs du Tableau IX.1 parce que l'éventail de qualité des données combinées est plus important que pour l'un quelconque des groupes de données). La majorité des points éloignés de la courbe de dispersion est issue de systèmes présentant une certaine forme de bruit non stationnaire en sortie (par exemple, émission en ondes métriques, cycles enregistrement/lecture professionnels 1/2 pouce multi-génération, cycles codage/décodage composites, erreurs de transmission numérique produisant des blocs d'erreurs en transitoire). Des améliorations futures à la méthode VQM sont en cours de développement. Elles feront appel à des paramètres fondés sur la perception pour mesurer ces effets de bruit non stationnaire. Les paramètres de qualité résultant des informations de gradient temporel (c'est-à-dire l'activité temporelle) constituent un domaine de recherche prometteur.

Tableau IX.1/J.144 – Coefficient de corrélation linéaire de Pearson pour la méthode VQM

Groupe de données	Coefficient de corrélation linéaire de Pearson
Un	0,92
Deux	0,90
Trois	0,94
Quatre	0,88
Cinq	0,91
Six à neuf combinés	0,86

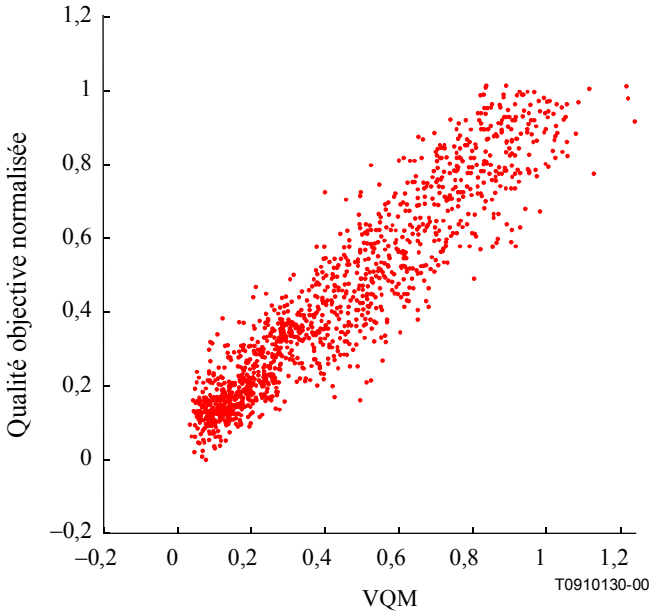


Figure IX.6/J.144 – Courbe de dispersion de la qualité subjective en fonction des mesures VQM pour les neuf groupes de données

IX.14 Références

[1] WOLF (Stephen), PINSON (Margaret H.): Spatial-temporal distortion metrics for in-service quality monitoring on any digital video system, *SPIE International Symposium on Voice, Video, and Data Communications*, Boston, MA, 11-22 septembre 1999.

[2] 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*).

[3] UIT-R BT.500-9 (1998), *Méthodologie d'évaluation subjective de la qualité des images de télévision*.

[4] VORAN (Stephen), WOLF (Stephen): The Development and evaluation of an objective video quality assessment system that emulates human viewing panels, *International Broadcasting Convention (IBC)*, juillet 1992.

[5] WEBSTER (Arthur A.), JONES (Coleen T.), PINSON (Margaret H.), VORAN (Stephen D.), WOLF (Stephen): An objective video quality assessment system based on human perception, *Human Vision, Visual Processing, and Digital Display IV, Proceedings of the SPIE*, Vol. 1913, février 1993.

[6] ANSI Accredited Standards Working Group T1A1 contribution number T1A1.5/94-118R1, "*Subjective test plan (tenth and final draft)*", Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington, DC, 3 octobre 1993.

[7] ANSI T1.801.01 (1995), *Digital Transport of Video Teleconferencing/Video Telephony Signals – Video Test Scenes for Subjective and Objective Performance Assessment*.

[8] FENIMORE (Charles), *et al.*: Perceptual effects of noise in digital video compression, *SMPTE Journal*, Vol. 109, pp. 178-186, mars 2000.

[9] WOLF (S.), PINSON (M.): In-service performance metrics for MPEG-2 video systems, *Made to Measure 98 – Measurement Techniques of the Digital Age Technical Seminar, technical conference jointly sponsored by the International Academy of Broadcasting (IAB), ITU, and the Technical University of Braunschweig (TUB)*, Montreux, Suisse, 12-13 novembre 1998.

[10] Final report from the video quality experts group (VQEG) on the validation of objective models of video quality assessment, *VQEG meeting number 4*, Ottawa, Canada, mars 2000.

SÉRIES DES RECOMMANDATIONS UIT-T

Série A	Organisation du travail de l'UIT-T
Série B	Moyens d'expression: définitions, symboles, classification
Série C	Statistiques générales des télécommunications
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	RGT et maintenance des réseaux: systèmes de transmission, circuits téléphoniques, télégraphie, télécopie et circuits loués internationaux
Série N	Maintenance: circuits internationaux de transmission radiophonique et télévisuelle
Série O	Spécifications des appareils de mesure
Série P	Qualité de transmission téléphonique, installations téléphoniques et réseaux locaux
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 et communication entre systèmes ouverts
Série Y	Infrastructure mondiale de l'information et protocole Internet
Série Z	Langages et aspects généraux logiciels des systèmes de télécommunication

Imprimé en Suisse
Genève, 2002