Benchmark de métriques de qualité sur bases de données d'images compressées

M. Nauge M.-C. Larabi C. Fernandez

Institut XLim, dépt. SIC, Université de Poitiers

Bât. SP2MI, Téléport 2, Bvd Marie et Pierre Curie, BP 30179
86962 Futuroscope Chasseneuil Cedex France

{nauge, larabi, fernandez}@sic.univ-poitiers.fr

Résumé

Il est utile de pouvoir quantifier les dégradations perçues afin de juger l'intérêt de nouveaux algorithmes de compression, tatouage ou des techniques de transmission. Nous proposons d'évaluer les performances d'un lot de métriques de qualité d'images couleurs en visant la corrélation avec le jugement humain. Le but escompté est de faciliter le choix d'une métrique parmi les nombreuses disponibles, en fournissant des scores de performance standards et exhaustifs. Pour cette étude quatre bases de données d'images sont utilisées et analysées afin d'élargir au maximum le jeu d'images tests et de mesurer la flexibilité des métriques.

Mots clefs

Métrique de qualité, système visuel humain, qualité d'expérience, évaluation objective, tests subjectifs.

1 Introduction

On note l'émergence des problématiques liées à l'augmentation de la qualité de service et de la qualité d'expérience entre l'homme et les machines. La preuve est l'explosion des recherches dans le domaine des interfaces plus intuitives, plus naturelles, plus rapides, qui nous comprennent facilement, partout, tout le temps. Il semble indispensable de disposer d'outils et de méthodologies capables de mesurer la qualité de ces nouveaux systèmes.

Pour mesurer la qualité des images numériques, il existe deux méthodes. La première consiste à effectuer des tests subjectifs, où un jeu d'images est présenté à un panel d'évaluateurs chargé de donner leurs avis sur la qualité perçue, généralement sur une échelle graduée de valeurs allant de Très mauvais à Très bonne. Pour garantir une certaine fiabilité des résultats, des recommandations strictes sont à respecter. L'ITU-R BT.500 [1] définit et formalise les contraintes pour effectuer les campagnes de tests subjectifs. Elles concernent le calibrage du système de visualisation, la distance et le positionnement de chaque observateur, leur nombre ainsi que les outils pour le dépouillement des résultats. Ces expérimentations sont délicates, coûteuses en temps et en argent. Il existe une seconde méthode pour effectuer les mesures de qualité : il s'agit des tests objectifs par l'utilisation de métriques de qualité. Il en existe 3 types : avec référence, avec référence réduite et sans référence. Les métriques avec référence utilisent l'intégralité de l'image sans dégradation pour effectuer les comparaisons. Les métriques sans référence utilisent seulement l'image dégradée pour juger la qualité. C'est une tâche facile pour l'homme mais très complexe pour une machine. Les métriques avec référence réduite extraient un minimum d'attributs de l'image sans distorsion, puis la comparaison s'effectue sur l'image dégradée avec ce minimum d'informations.

Dans cette étude, notre démarche vise à récolter un maximum de métriques de qualité et de les évaluer de manière rigoureuse sur un grand nombre d'images. Nous souhaitons faciliter le choix de métriques de qualité pour toute personne désireuse de tester l'apport de son nouvel algorithme en termes de réduction de dégradation perceptible. L'idée est également de faciliter les démarches de comparaison de nouvelles métriques de qualité en les comparant de manière équitable sur un grand nombre de bases de données d'images. Nous proposons dans une première partie de faire un tour d'horizon des métriques actuellement disponibles et utilisées pour l'étude. Leur comparaison est effectuée dans une troisième partie. Mais avant cela, une analyse des différentes bases de données d'images utilisées est détaillée dans une deuxième partie. La partie quatre est dédiée aux expérimentations suivie de ses conclusions et perspectives.

2 Description des métriques utilisées

Il existe une grande variété de métriques de qualité d'images numériques. Les travaux de Pederson [2] font état de plus de 111 métriques existantes. Depuis cette étude, nous pouvons très facilement ajouter à ce nombre plusieurs dizaines de métriques et ceci est dû à l'effervescence que connaît cette problématique. On peut les classer en 3 grandes catégories aux frontières plus ou moins floues : mathématiques, mesures pondérées par quelques propriétés du Système Visuel Humain (SVH) et la modélisation complète du SVH . Les premières métriques telles

MSE, SNR, RMSE, PSNR sont purement mathématiques, basées sur une comparaison pixel à pixel dans l'espace RVB. La métrique Peak Signal to Noise Ratio (PSNR) est la plus connue et la plus utilisée, grâce à son implémentation simple et son exécution rapide. Cependant le PSNR montre une faible corrélation avec le jugement humain sur certaines images (figure 1). De nombreux travaux visent à développer des métriques plus proches de la perception humaine.

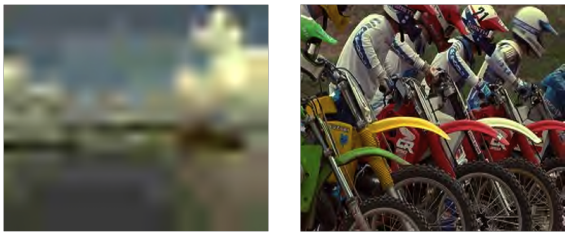

Figure 1 – *deux images avec un PSNR de 24dB*

On y retrouve les métriques VDP [3], Pdiff [4], NQM [5] qui tentent de modéliser finement le Système Visuel Humain (SVH) et le système de visualisation. Mais ces méthodes sont complexes et couteûses en temps d'exécution. La troisième catégorie de métriques tente de trouver un équilibre en effectuant des mesures pondérées par le SVH. On peut citer PSNR-HVS , PSNR-HVSM [6] qui effectuent des mesures proches de PSNR après avoir effectuées une pondération par une fonction de sensibilité aux contrastes (CSF) et un effet de masquage simplifié. Le but est de mimer les capacités changeantes de perception de stimuli par l'homme, en fonction de la fréquence, l'orientation, la couleur et la présence d'autres stimuli sur une image.
Il existe PSNR-E76, PSNR-E94 et PSNR-E00 [7] qui effectuent la comparaison pixel à pixel dans un domaine plus perceptuel. DeltaE2000 est une mesure de distance couleur à partir d'un espace CIELAB, spécialement conçue pour s'adapter aux différences de perception du SVH. Typiquement, les différences de couleurs dans les tons bleus sont moins perceptibles que les différences dans les tons verts. Les métriques S-Cielab [8], SHAME I et SHAME II [9] utilisent également une mesure de distance couleur dans l'espace CIELAB, après avoir effectuées un prétraitement de filtrage spatial, afin de considérer la distance d'observation comme un facteur agissant sur la détection d'artefact.
Il existe d'autres métriques telles qu'UQI [10], SSIM [11], MSSIM [12] et ses nombreuses dérivées qui ne prennent en considération que les changements structurels dans l'image, sans se soucier des conditions de visualisation ou des caractéristiques poussées du SVH. L'idée sous-jacente est que l'homme est naturellement très sensible aux changements de structures dans une image.
D'autres métriques exploitent des statistiques de scènes naturelles, par exemple IFC [13] ainsi que VIF et VIFP [14] qui utilisent conjointement l'analyse statistique de scène et pondération par CSF. Il existe VSNR [15] et WSNR [16] qui exploitent quelques propriétés du SVH. L'originalité réside dans l'utilisation d'ondelette pour effectuer les mesures. Etant donnée la variété des métriques, il semble intéressant de les comparer. Afin d'effectuer des comparaisons de métriques et de juger leurs corrélations avec le jugement humain, il est indispensable de disposer de bases de données d'images dont on connait les scores subjectifs d'humains. Ces scores sont les MOS (Mean Opinion Score).

3 Bases d'images utilisées

Il y a peu de bases de données d'images permettant de vérifier les performances des métriques. Ces bases sont LIVE [17], Toyama [18], IVC [19] et TID2008 [20]. Le tableau 1 présente les caractéristiques de chaque base utilisée pour l'étude.

Tableau 1 – *Caractéristiques des bases de données*

Car.	LIVE 1	Toyama	TID2008	IVC
Distors.	JPG, J2K	JPG, J2K	JPG, J2K	JPG, J2K
Codeur Jpg	Matlab imwrite	cjpeg	Non précisé	Non précisé
Codeur jp2k	Kakadu v2.2	Jasper v1.7	kakadu	Non précisé
Méthode	Single stimulus	Single stimuli	Double stimuli	Double stimuli
Image	Couleurs RGB avec 24 bits/pixels			
Résolution	entre 768x512 et 634x438	768x512	812x384	512x512
Nb img.	29	14	25	10
écran	CRT 21-inch (1024x768) (non calibrés)	CRT 17-inch (1024x768)	LCD et TFT 17 et 19 inches 1152x864	Non précisé.
Distance déobservation	2-2.5H (hauteur écran)	4H (hauteur image)	Trés varié	6H (hauteur écran)
Eclairage ambiant	Bureau	Faible	Trés varié	Normalisé
Nb. observateurs	J2K [20-25] JPG 20	16	654	15
Type observateur	Etudiants Univ.Texas	Non expert , étudiants	Non précisé	Non précisé
Ecart type et IC	Calculable	Fournis	Incalculable	Incalculable

Tout d'abord nous remarquons une différence d'exhaustivité des informations disponibles pour certaines bases. On peut critiquer l'absence d'informations relatives à l'écart type des MOS qui rend impossible certains calculs nécessaires à l'évaluation des performances de métriques.
Bien que LIVE1 ne les fournissent pas, les notes de chaque observateur sont fournies, ce qui permet de calculer les informations manquantes. Nous pouvons remarquer que les recommandations de l'ITU-R relatives aux protocoles d'expérimentation ne sont pas toujours respectées. Par exemple les distances d'observations ne sont pas assurées. Les conditions d'éclairage ne sont pas contrôlées. Les dispositifs d'affichage ne sont ni identiques ni réglés pour chaque expérimentation. Dans de telles conditions allons-nous réellement tester les performances des métriques ou les conditions de l'évaluation subjective ?
En ce qui concerne le respect des recommandations, certaines études se veulent rassurantes et montrent que les différences ne sont pas notables. La base TID2008 qui a eu recours a une très large expérimentation (3 laboratoires de

pays différents, ainsi que des dispositifs d'affichages TFT et CRT mélangés, associés à des distances d'observation et des conditions d'éclairage variées) affirme que les résultats obtenus entre chaque laboratoire sont corrélés à 97%. Une autre étude a tenté de vérifier l'impact de la différence de culture (Japon/France) ainsi que l'incidence du type d'affichage (CRT/LCD). Les résultats numériques démontrent une corrélation à plus de 95%. Donc la combinaison de toutes ces contre-indications semble être négligeable.

Abordons maintenant le choix des images et de la magnitude des distorsions. Pour tester les performances des métriques de qualité, il est important d'avoir des images variées, représentatives de la diversité des images échangées à travers le monde. Nous pouvons noter que les 3 bases LIVE, Toyama et TID2008 utilisent les mêmes images sources (12 images communes). Ce panel d'images est tout de même intéressant car il contient des images d'objets manufacturés, de visages, d'animaux, de paysages naturels, différentes prises de vue avec des premiers et arrières plans plus ou moins distincts. Mais si ce choix d'images initial n'est pas correct, les trois bases subissent le même discrédit. Bien que les images sources soient les mêmes, la magnitude des distorsions et les codeurs sont différents. Par exemple la base TID2008 a des distorsions qui rendent le contenu des images indiscernable tandis que toutes les images de Toyama restent très correctes. Tandis que LIVE propose des distorsions réparties de manière plus homogène en terme de magnitude. On se rend compte que les échelles, qui ont été proposées aux utilisateurs, n'ont pas le même sens pour une base ou pour une autre. Quand les valeurs donnent un état « Bad » pour une image de la base Toyama, il s'agit finalement d'un état « Good » de la base TID2008. Une métrique performante avec la base Toyama montre une grande justesse de mesure pour les images très peu dégradées, nécessitant une analyse très précises des dégradations. Mais si cette métrique donne de mauvais résultats avec la base Toyama qui crée d'importantes dégradations, cela sous entend qu'elle n'est pas très robuste aux importantes distorsions. L'idée est d'exploiter la complémentarité des bases de données pour juger les métriques.

Si certains sont sceptiques sur la diversité des images de ces bases de données et qu'ils espèrent trouver d'autres images avec la base IVC, il faudra être très prudent. Bien que cette base affirme respecter de manière rigoureuse le protocole d'évaluation (environnement normalisé), le choix des images sources peut laisser perplexe. Les images semblent éloignées des images actuelles. La dynamique des couleurs et la résolution des images sources sont très basses, bien en dessous des capacités d'acquisition des capteurs grand public. Il est très difficile de disposer de bases de données liant respect des protocoles, qualité de contenu, et exhaustivité des résultats (détailler tous les paramètres de l'évaluation, fournir les scores subjectifs qui ont permis le calcul du MOS etc.). Cependant il semble nécessaire et suffisant d'utiliser ces bases comme référence actuelle pour effectuer des tests de performance de métrique.

Les bases d'images sont importantes et il semble qu'une vague de prise en compte de l'aspect subjectif de la qualité des traitements déferle sur le monde scientifique. De plus en plus de laboratoires envisagent de disposer d'une salle permettant de réaliser des tests subjectifs. Nous conseillons de veiller à respecter les standards existants afin de minimiser les inquiétudes des futurs utilisateurs. Il est important de veiller à l'exhaustivité des résultats obtenus afin d'assurer une transparence des résultats et permettre plus de flexibilité des analyses. Puisqu'il est intéressant d'utiliser plusieurs bases d'images, il faut également veiller à faciliter leurs utilisations. Pour réaliser cette étude utilisant seulement 4 bases, il a fallu consacrer beaucoup de temps pour uniformiser les informations du fait de la variété des formats de fichier contenant les résultats et des différences de hiérarchie des dossiers. Dans le cas de cette étude, nous avons normalisé toutes les bases de données. Cela passe par une spécification de la hiérarchie des dossiers contenant les images, une convention de nomination de fichier, le stockage des résultats dans des formats de fichier standard et non propriétaires. Le respect des standards garantit flexibilité et interopérabilité des différents composants acteurs de tous projets.

4 Expérimentation

4.1 Procédure

Afin d'évaluer équitablement les performances des métriques de qualité, nous suivons le plan de test du VQEG [21]. La méthode consiste à disposer d'un maximum d'images dont on connait les MOS subjectifs d'un panel d'observateurs (les MOS sont obtenus en respect des recomandations de [1]). Nous utilisons les bases d'images LIVE, Toyama, IVC et TID2008 pour les dégradations jpg et jp2k. La compression JPEG introduit des effets de bloc, tandis que JPEG2000 a tendance à ajouter du flou à l'image. Ces deux types de dégradation permettent de tester la robustesse des métriques. Nous exécutons 27 métriques de qualité avec référence sur chaque image de chaque base afin d'obtenir leurs prédictions (MOSp). Les prédictions MOSp sont classées par base et par type de distorsion. Les analyses permettant de tester les métriques portent sur trois facteurs : un facteur de corrélation grâce au calcul de la corrélation de Pearson, un facteur de précision avec un calcul de racine de l'erreur quadratique moyenne (RMSE), et un facteur de cohérence par un calcul de taux de rejet (OR). L'analyse des résultats de corrélation de Pearson pour chaque métrique permet d'étudier l'existence de relation entre les valeurs MOS subjectives et les MOSp des métriques. Le plan de test du VQEG préconise d'appliquer une régression non linéaire sur chaque série de MOSp avant d'effectuer les mesures de performance. En appliquant ce prétraitement la majorité des métriques affichent des scores de corrélation proches de 99%, ce qui rend leurs comparaisons délicates. La section « 4.7 Costs and Benefits of the logistic transformation » d'une version antérieure du plan de test du VQEG [22] permet d'expliquer ces fortes

corrélations. Il est expliqué que la regression non linéaire peut introduire un gain important de corrélation quand les métriques ont une corrélation inférieure à 80% sans regression. Or beaucoup de métriques affichent des résultats sous ce palier. Les figures 2 et 4 font donc apparaître des scores de corrélation sans regression afin de ne pas introduire un biais trop significatif. Plus les valeurs sont proches de 1 ou -1 plus les résultats sont corrélés.

Pour les calculs d'OR et de RMSE, il faut mesurer les « Perror » représentant la différence entre la valeur subjective MOS et la valeur prédite MOSp. Cependant la plage de valeurs MOSp est très variable d'une métrique à l'autre. Typiquement la plage de prédictions MOSp de PSNR est de [17-45] tandis qu'elle est de [0.34-0.99] pour SSIM pour la base LIVE_jp2k. Une normalisation est appliquée sur chaque série de MOSp afin d'effectuer des comparaisons sur une échelle unique de [0.0-1.0]. L'étude du RMSE (figure 5) permet de mesurer avec précision les écarts de prédiction qui sont meilleures quand les valeurs sont proches de zéro. L'objectif de l'OR est de vérifier que les mesures sont contenues dans un intervalle acceptable. Les valeurs proches de zéro indiquent qu'il y a peu de prédictions trop éloignées des MOS subjectifs. Les écarts autorisés pour les métriques sont définis pour chaque image, et sont relatifs à la variabilité des scores des évaluateurs. Si pour une même image, les évaluateurs donnent des scores variés, l'écart autorisé pour les métriques sera important. L'objectif est de pénaliser les métriques qui donnent de mauvaises prédictions là où les observateurs sont unanimes sur la qualité d'une image. Les OR sont restreints aux bases LIVE et Toyama car elles sont les seules à fournir suffisament d'informations. L'utilisation combinée des facteurs, corrélation, RMSE et OR permet de juger les métriques avec équité et permet de faciliter la prise de décision dans le choix de métriques pour un type d'utilisation.

4.2 Résultats et discussion

Les résultats des figures 2 et 4 montrent que la métrique PSNR_HVSM a la meilleure corrélation avec le jugement humain avec des scores de 94% sur TID2008_jpeg et 96% sur TID2008_jp2k. La magnitude de distorsion de la base TID_2008 est très importante, il est donc interressant d'utiliser cette métrique dans des contextes où les images ont une plage importante de dégradation. Pour les contextes où les dégradations sont minimes, mais que l'on souhaite tout de même détecter d'infimes dégradations, il faut s'intérresser aux résultats obtenus sur la base Toyama. Pour les dégradations JPEG dans ce contexte c'est la métrique VIF qui surpasse les autres avec un score de 0.98 avec près de 5% de corrélation de plus que la seconde métrique IFC. C'est également VIF qui l'emporte avec un score de 0.949 sur Toyama_jp2k. Il semble que cette métrique se révèle la plus précise sur les 2 types de distorsions, elle est donc également assez robuste. Pour des contextes d'utilisation où une grande robustesse est attendue, il est interressant d'observer le comportement moyen des métriques sur toutes les bases de données. C'est VIF avec sont 0.849 qui est en tête suivi de près par PSNR_HVSM sur les dégradations du type JPEG. Pour les dégradations jp2k le trio de tête est composé de VIF (0.915), VIFP (0.905) et PSNR_HVSM (0.907). Dans ce cas il semble difficile de départager le 0.905 de VIFP du 0.907 de PSNR_HVSM. Bien que ces métriques soient fortement corrélées, elles n'ont pas la même précision de prédiction. On peut utiliser le RMSE pour départager les métriques ayant des scores de corrélation trop proches.

Les résultats de RMSE (figure 5) montrent que c'est la métrique VIFP qui est la plus stable sur l'intégralité des bases d'images. Les RMSE montrent également que peu de métriques ont une erreur moyenne inférieure à 20%. Le classique PSNR reste tout même assez précis sur l'ensemble des bases d'images. On remarque également que les erreurs de prédictions sont les plus importantes sur la base Toyama et IVC. Cette difficulté de précision peut s'expliquer sur la base Toyama connaissant sa faible magnitude de distorsions. Il est donc difficile d'être précis sur ces faibles dégradations. Les mesures pixel à pixel des différentes versions de PSNR sont donc efficaces malgré leur faible complexité.

Pour affiner le jugement, il est indispensable d'étudier l'OutlierRatio (figure 3) afin de vérifier que les mesures restent dans un intervalle de confiance acceptable. Cette mesure d'OR accentue de manière significative les écarts entre les différentes bases et les différentes métriques. On remarque encore une fois que la base Toyama est vraiment très exigente sur la précision des prédictions. Même la très reconnue SSIM affiche un taux de rejet proche des 70%. Alors que ce taux est en dessous des 20% sur la base LIVE. On peut donc s'interroger sur de tels écarts. La principale raison de ces forts taux de rejet vient du fait que les écarts autorisés fournis sur la base Toyama sont très restrictifs.

Finalement la modélisation complète du SVH n'est pas la solution car la très complexe HDR_VDP ne s'est jamais démarquée. Mais l'approche structurelle uniquement de la famille des SSIM n'est pas suffisante non plus. Il faut également noter que le PSNR bien que très critiquable et pouvant facilement être mis en échec se révèle en moyenne relativement performant malgré sa faible complexité. Il semble qu'ajouter peu de paramètres du SVH sur des mesures très simples augmente la qualité des prédictions. Typiquement PSNR-HVSM n'introduit qu'un filtrage CSF et un effet de masquage simplifié afin de garantir un temps de calcul limité et une robustesse aux différents types de bases de données et de distorsions.

5 Conclusion

Pour valider l'apport de nouveaux algorithmes de traitement d'images, il est nécessaire d'utiliser plusieurs métriques de qualité sur plusieurs bases de données d'images. Nous conseillons les métriques PSNR, PNSR_HVSM [23], VIFP et MSSIM [24] pour garantir une certaine variété de résultats tout en maintenant une bonne corrélation avec le

jugement humain. PSNR est utile pour sa popularité dans le monde scientifique, et son exécution rapide. PSNR_HVSM pour son intégration de composantes du SVH tout comme VIFP qui introduit en plus quelques informations statistiques de scènes naturelles. Et MSSIM pour son approche pertinente basée sur la perte de structure. Il faut pondérer ces conseils, car aucune métrique n'associe à la fois précision et robustesse. De plus ces analyses reposent uniquement sur l'étude des distorsions introduites par compression. De nouvelles distorsions ou de nouvelles métriques donneraient lieu à de nouveaux résultats. Nous souhaitons enrichir et maintenir à jour cette étude avec plus de métriques et de bases de données. Bien que ce travail soit long et fastidieux, il permet entre autre de connaître à tout instant les performances de chaque métrique par une étude standard et indépendante. Elle permet également de faciliter l'utilisation et le choix des métriques en fonction des applications visées. Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-VERS-002 dans le cadre du projet Caiman.

Références

[1] Methodology for the subjective assessment of the quality of television pictures. Rapport technique RECOMMENDATION ITU-R BT.500-11, ITU-R.

[2] Marius Pedersen et Jon Yngve Hardeberg. Survey of full-reference image quality metrics. Dans *GCIS'2009 Global Congress on Intelligent Systems*, Gjovik, Norway, Juin 2009.

[3] R. Mantiuk, K. Myszkowski, et H.-P. Seidel. Visible difference predicator for high dynamic range images. Dans *IEEE International Conference on Systems, Man and Cybernetics*, pages 2763–2769, Octobre 2004.

[4] H. Yee. A perceptual metric for production testing. *Journal of Graphics Tool*, pages 33–40, 2004.

[5] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, et A. C. Bovik. Image quality assessment based on a degradation model. Dans *IEEE transactions on image processing*, pages 636–650, 2000.

[6] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, et M. Carli.. Two new full-reference quality metrics based on hvs. Dans *Second International Workshop on Video Processing and Quality Metrics*, Scottsdale USA, 2006.

[7] G. Sharma, W. Wu, et E. N. Dalal. The ciede2000 color-difference formule : Implementation notes, supplementary test data, and mathematical observations. *Color Research and Application*, Février 2004.

[8] X. Zhang et B.A. Wandell. A spatial extension of cielab for digital color image reproduction. Dans *Soc. Inform. Display 96 Digest*, pages 731–734, San Diego, 1996.

[9] M. Pedersen et J.Y. Hardeberg. Shame : A new spatial hue angle metric for perceptual image difference. Dans *Color Research and Application*, Naples, FL, USA, Mai 2009.

[10] Z. Wang et A. C. Bovik. A universal image quality index. Dans *IEEE Signal Processing Letters*, pages 81–84, 2002.

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, et E. P. Simoncelli. Image quality assessment : From error visibility to structural similarity. Dans *IEEE Transactios on Image Processing, vol. 13, no. 4*, pages 600–612, Avril 2004.

[12] Z. Wang, E. P. Simoncelli, et A. C. Bovik. Multiscale structural similarity for image quality assessment. Dans *37th IEEE Asilomar Conference on Signals, Systems and Computers*, Novembre 2003.

[13] H. R. Sheikh. *Image Quality Assessment Using Natural Scene Statistics*. PhD thesis, University of Texas at Austin, 2004.

[14] H. R. Sheikh et A. C. Bovik. Image information and visual quality. Dans *IEEE Transactions on Image Processing*, pages 430–444, 2006.

[15] D. M. Chandler et S. S. Hemami. Vsnr : A wavelet-based visual signal-to-noise ratio for natural images. Dans *IEEE Transactions on Image Processing*, Septembre 2007.

[16] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, et A. C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4) :636–650, Avril 2000.

[17] H.R. Sheikh, Z. Wang, L. Cormack, et A.C. Bovik. Live image quality assessment database release 1. http ://live.ece.utexas.edu/research/quality.

[18] Y. Horita, Y. Kawayoke, et Z. M. Parvez Sazzad. Image quality evaluation database. http ://mict.eng.utoyama.ac.jp/mict/index2.html.

[19] P. Le Callet et F. Autrusseau. Subjective quality assessment irccyn/ivc database, 2005. http ://www.irccyn.ec-nantes.fr/ivcdb/.

[20] N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola, M. Carli, et F. Battisti. Color image database for evaluation of image quality metrics, 2008. http ://www.ponomarenko.info/tid2008.htm.

[21] Final report from the video quality experts group on the validation of objective models of multimedia quality assessement. Rapport technique PHASE I 2008, VQEG.

[22] Final report from the video quality experts group on the validation of objective models of video quality assessement. Rapport technique PHASE II 2003, VQEG.

[23] Ni. Ponomarenko. Psnrhvsm code, 2006. http ://ponomarenko.info.

[24] M. Gaubatz. Metrix mux, 2007. http ://foulard.ece.cornell.edu/gaubatz/metrix_mux.

Bdd	UQI	Shamel	IFC	MSE	SCielab	VIFP	PSNR	MSSIM	NQMRFZ3	Pdiff	PSNR_HVS	PSNR_E76	ShameII	NQM	DSSIM	VSNR	SVD	HdrVdp	PSNR_E94	VIF	PSNR_HVSM	RMSE	SSIM	PSNRLum	WSNR	SNR	PSNR_E00
Toyama	0,77	-0,463	**0,858**	-0,229	-0,24	0,753	0,356	0,794	0,389	0,709	0,621	0,445	-0,417	0,652	-0,657	**0,809**	-0,128	0,334	0,392	**0,898**	0,735	-0,259	0,621	0,434	0,437	0,296	0,384
LIVE1	**0,847**	-0,797	**0,828**	**-0,899**	-0,368	**0,935**	**0,886**	**0,884**	0,752	0,704	**0,911**	**0,885**	-0,781	**0,853**	**-0,844**	**0,923**	-0,216	0,784	**0,89**	**0,935**	**0,918**	**-0,881**	**0,909**	**0,865**	0,731	**0,888**	**0,892**
Tid2008	0,793	-0,761	0,78	**-0,916**	-0,148	**0,918**	**0,889**	**0,929**	0,73	0,643	**0,943**	**0,895**	**-0,811**	**0,869**	**-0,859**	**0,9**	-0,328	**0,831**	**0,888**	**0,932**	**0,944**	**-0,897**	**0,901**	**0,818**	0,66	**0,849**	**0,926**
IVC	**0,819**	-0,583	0,921	-0,617	-0,105	0,79	0,591	**0,821**	0,534	0,181	0,662	0,544	-0,389	0,429	-0,5	0,654	-0,257	0,63	0,565	**0,922**	0,695	-0,638	0,7	0,341	0,265	0,617	0,546
Toutes	**0,807**	-0,651	**0,847**	-0,665	-0,215	**0,849**	0,681	**0,857**	0,601	0,559	0,784	0,692	-0,599	0,701	-0,715	**0,821**	-0,232	0,645	0,684	**0,922**	**0,823**	-0,669	0,783	0,615	0,523	0,663	0,687

Figure 2 – *Scores de corrélations sur images dégradées par JPEG*

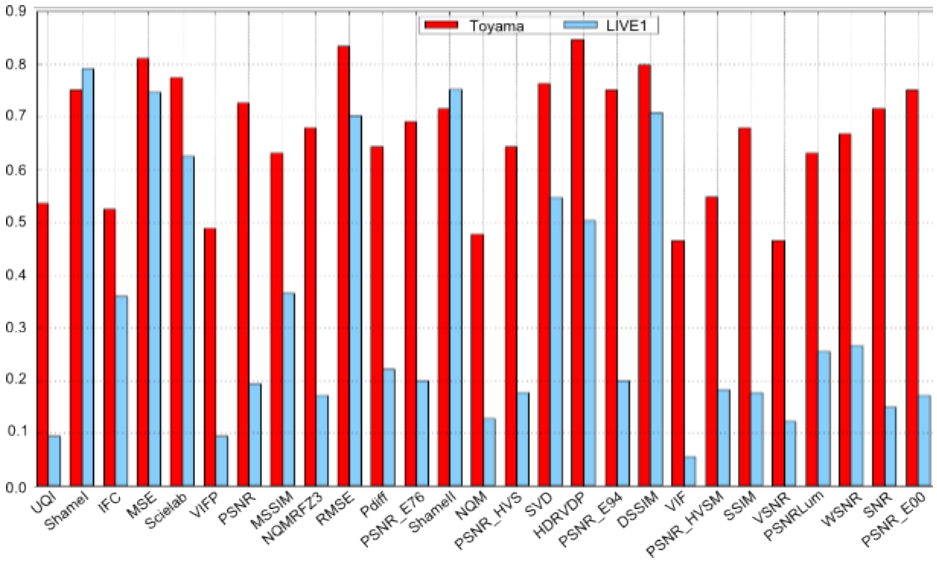

Figure 3 – *Scores des OR sur images dégradées par JPEG*

Bdd	UQI	Shamel	IFC	MSE	SCielab	VIFP	PSNR	MSSIM	NQMRFZ3	Pdiff	PSNR_HVS	PSNR_E76	ShameII	NQM	DSSIM	VSNR	SVD	HdrVdp	PSNR_E94	VIF	PSNR_HVSM	RMSE	SSIM	PSNRLum	WSNR	SNR	PSNR_E00
Toyama	0,631	-0,701	0,796	-0,788	0,058	**0,927**	**0,843**	**0,825**	0,614	0,767	**0,866**	0,778	-0,658	**0,84**	-0,782	**0,912**	-0,179	0,552	0,797	**0,949**	**0,893**	**-0,809**	**0,847**	**0,878**	0,673	0,787	0,767
LIVE1	**0,822**	-0,522	0,816	-0,793	-0,301	**0,907**	**0,843**	**0,8**	0,78	0,672	**0,864**	0,787	-0,575	**0,871**	-0,726	**0,903**	-0,199	0,737	**0,82**	**0,903**	**0,883**	-0,794	**0,857**	**0,866**	0,79	0,743	0,795
Tid2008	**0,913**	-0,762	**0,82**	**-0,842**	0,172	**0,941**	**0,866**	**0,936**	**0,884**	0,64	**0,959**	**0,821**	-0,751	**0,934**	**-0,8**	**0,93**	-0,252	**0,925**	**0,847**	**0,916**	**0,96**	**-0,835**	**0,864**	**0,866**	**0,837**	**0,831**	0,89
IVC	**0,8**	-0,649	**0,882**	-0,709	-0,052	**0,848**	0,77	**0,77**	0,755	0,363	**0,875**	0,649	-0,528	0,673	-0,548	**0,803**	-0,091	0,69	0,708	**0,893**	**0,893**	-0,731	0,76	0,661	0,633	0,734	0,678
Toutes	0,791	-0,659	**0,826**	-0,783	-0,031	**0,905**	**0,83**	**0,833**	0,758	0,611	**0,891**	0,758	-0,628	**0,829**	-0,714	**0,887**	-0,18	0,726	0,793	**0,915**	**0,907**	-0,792	**0,832**	**0,818**	0,733	0,774	0,783

Figure 4 – *Scores de corrélations sur images dégradées par jp2k*

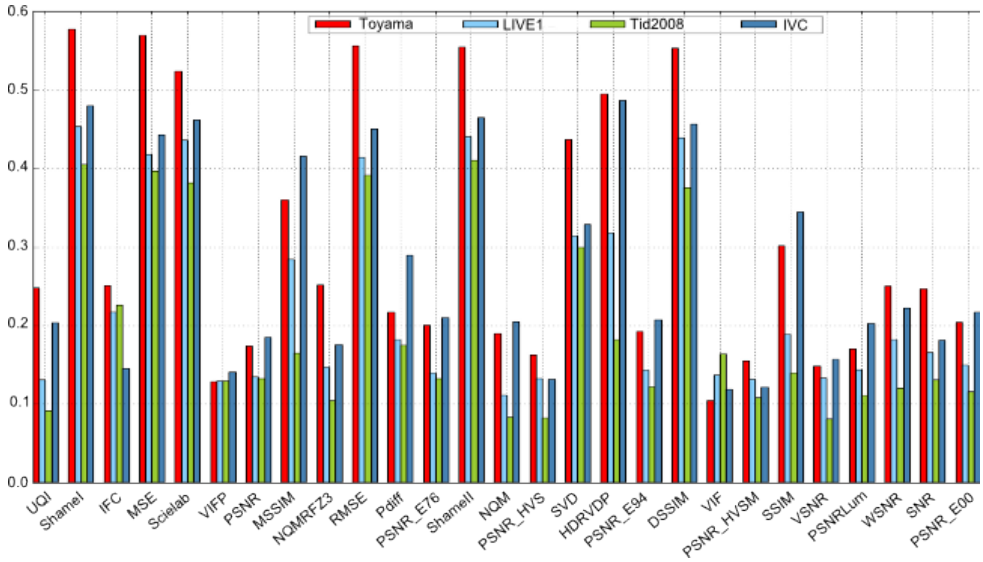

Figure 5 – *Scores de RMSE sur images dégradées par jp2k*