Mesures de qualité d'images avec référence : Limitation

ALADINE CHETOUANI

Laboratoire PRISME
12 rue de Blois, 45067 Orléans, France

aladine.chetouani@univ-orleans.fr

Résumé – À travers cet article, nous proposons d'étudier l'universalité des mesures de qualité d'images avec référence. L'objectif est ici de montrer l'importance de considérer la dégradation dans le processus d'estimation de la qualité. Différents tests expérimentaux ont été menés afin d'analyser leurs performances. Huit mesures de qualité ont été utilisées et comparées en termes de corrélation avec les jugements subjectifs. Les résultats obtenus montrent que le rendement d'une mesure donnée diffère totalement d'une dégradation à une autre. Nous concluons enfin par la pertinence de certains travaux récents qui proposent des solutions alternatives pour résoudre cette limitation et ainsi optimiser le processus d'estimation de la qualité de l'image.

Abstract - In this work, we propose to study the universality of Full-Reference Image Quality metrics (FR-IQMs) and show the no-relevance to use this kind of metrics without considering the degradation type contained in the image. Different experimental tests have been done in order to analyze its performance. Eight common FR-IQMs have been used and compared in terms of correlation with the subjective judgments. Obtained results show that the performance of a given FR-IQM differs totally from a degradation type to another. We finally conclude by the pertinence of some recent works that propose alternative solutions to solve this limitation and then optimize the image quality estimation process.

1 Introduction

Ces dernières années, un nombre important de mesures de qualité d'images avec référence (FR-IQM : Full Reference Image Quality Metric) a été proposé dans la littérature. Dans [1], plus d'une centaine de mesures ont été répertoriées. Certaines d'entres elles sont basées sur des notions purement mathématique, tandis que d'autres sont basées sur une analyse structurelle [2] ou bien encore sur des modèles perceptuels [3].

Les performances de ce type de mesures sont évaluées en calculant des coefficients de corrélation entre les scores objectifs, obtenus par la mesure, et les scores subjectifs, obtenus à partir de tests psycho-visuels. Ces corrélations sont généralement affichées pour chaque dégradation [4]. L'universalité de ces méthodes est ainsi implicitement supposée.

Dans cet article, nous proposons d'étudier l'universalité de ce type de mesures et d'essayer de montrer l'importance de considérer la dégradation dans le processus d'estimation de la qualité des images. En effet, pour une mesure FR-IQM donnée, les performances obtenues diffèrent grandement d'une dégradation à une autre. Huit mesures fréquemment utilisées ont été ici sélectionnées avec comme critère d évaluation le coefficient de corrélation de Pearson.

Ce papier est organisé comme suit: Dans la section 2, nous présentons certaines bases de données d'images subjectives disponibles dans la littérature ainsi que les mesures de qualité sélectionnées. L'analyse des performances est discutée dans la section 3, suivie de la conclusion à la section 4.

2 Base de données d'images subjectives et mesures de qualité avec référence

2.1 Base de données

Afin d'évaluer les performances d'une mesure FR-IQM donnée, nous devons disposer d'une base de données la plus complète, couvrant le plus de distorsions possibles. Cette évaluation s'effectue en comparant les indices de qualité objectifs obtenus (issus de la mesure FR-IQM) et les notes subjectives correspondantes (MOS: Mean Opinion Scores). Certaines de ces bases sont ici présentées.

2.1.1 LIVE

La base de données d'images, nommée LIVE, est composée de cinq types de dégradations différents (bruit, flou, JPEG, JPEG2000 et « fast fading ») [10]. Les notes subjectives (DMOS : Difference MOS) sont fournies pour chaque image dégradée. Sa valeur varie entre 0 et 100, où 0 et 100 désignent respectivement le meilleur et le mauvais scores. Le nombre d'images dégradées, obtenues à partir de 29 images de référence, est différent pour chaque type de dégradations considéré. Au total, 982 images dégradées sont disponibles.

2.1.2 IVC

Pour la base de données IVC [11], des évaluations subjectives ont été faites avec une distance d'observation de 6 fois la hauteur de l'écran en utilisant le protocole DSIS (Double-Stimulus Impairment Scale) avec une échelle de 5 catégories et 15 observateurs. Une note subjective égale à 0 indique une mauvaise qualité et une note égale à 5 indique une bonne qualité. Les

auteurs proposent aussi différents types de dégradations, à savoir : JPEG2000, JPEG, LAR codage, Blur.

2.1.3 TOYAMA

La base TOYAMA est quant à elle composée de deux types de dégradations (images compressées JPEG et JPEG2000) avec 98 images par distorsion [12], issues de 14 images de référence. Les scores subjectifs sont divisés en cinq notes (1: mauvais, 2: faible, 3: moyen, 4: bon et 5: excellent) en accord avec les recommandations de l'UIT.

2.1.4 TID 2008

La base d'images de Tampere (TID 2008) est constituée de 17 types de dégradations (les artefacts de compression tels que JPEG et JPEG2000, flou, le bruit, …) avec 100 images dégradées par distorsion, obtenues à partir de 25 images de référence [13]. Les notes subjectives sont aussi disponibles et varient entre 0 et 9 avec 9 la meilleure qualité.

2.2 Mesures de qualité avec référence

Les mesures de qualité FR-IQMs sélectionnées sont brièvement présentées dans cette section [15].

Universal image Quality Index (UQI)

La mesure UQI [9] est basée sur une analyse locale de l'image. Après une décomposition de l'image en blocs de taille NxN, plusieurs paramètres statistiques locaux sont extraits. Ces paramètres sont ensuite utilisés pour estimer la qualité des images.

Structural SIMilarity et sa version multi-échelle (SSIM)

La mesure SSIM est une version améliorée de la mesure UQI et est aussi basée sur l'extraction d'informations structurelles locales [2]. La mesure finale est composée de trois facteurs: un facteur de luminance (L), de contraste (C) et de structure (S). Dans [17], une version multi-échelle a été proposée où les mêmes facteurs (L, C et S) sont extraits. Le facteur L est dérivé du dernier niveau de décomposition, tandis que C et S sont calculés à chaque niveau de décomposition. L'indice global de la qualité de l'image est finalement obtenu par la multiplication du facteur L et la somme des facteurs C et S.

Weighted Signal to Noise Ratio (WSNR)

Contrairement aux mesures précédentes, la mesure WSNR est calculée dans le domaine fréquentiel [7]. Elle correspond au rapport signal sur bruit pondéré et est exprimée comme étant le rapport entre le spectre de Fourier et la fonction de sensibilité au contraste (CSF : Contrast Sensitivity Contrast) qui est une représentation de la sensibilité de notre système visuel aux changements de contraste.

PSNRHVS

Dans [3], une mesure basée sur le PSNR, appelée PSNRHVS, a été proposée. Les auteurs proposent d'améliorer les performances du PSNR (critiqué par plusieurs auteurs [16]) en intégrant certaines caractéristiques du SVH. Pour ce faire, les auteurs ont intégré dans le calcul du PSNR un modèle de la fonction CSF dans le domaine DCT.

Information Fidelity Criterion (IFC)

Basée sur la théorie de l'information, la mesure IFC est calculée en utilisant un modèle de source (C) et de distorsion (D) [6]. Ces modèles sont estimés dans le domaine des ondelettes, uniquement sur certaines sous-bandes.

Visual Information Fidelity (VIF)

Une extension de la mesure IFC, appelée VIF, a aussi été proposée dans [5]. L'amélioration réside dans l'intégration de certaines propriétés du SVH. Une autre version de la mesure VIF, nommée VIFP, a également été proposée. Cette dernière est calculée dans le domaine spatial.

Visual Signal to Noise Ratio (VSNR)

Cette méthode est basée sur deux étapes principales [8]. La première étape consiste à calculer un seuil de contraste dans le domaine des ondelettes. Ce seuil vise à déterminer la visibilité de la distorsion. Si la valeur obtenue est inférieure au seuil de détection, l'image est censée être parfaite (VSNR = Inf). Sinon, l'indice de qualité VSNR est calculé.

3 Analyse des performances et discussions

Comme mentionné précédemment, l'objectif de cette étude est de montrer la non-universalité des mesures de qualité d'images avec référence et de mettre en avant la pertinence d'utiliser un système intégrant le type de dégradations.

Tab 1 : Base d'images TID 2008: dégradations considérées.

	Dégradation		Dégradation
1	Additive Gaussian noise	10	JPEG compression
2	Additive noise in color components	11	JPEG 2000 compression
3	Spatially correlated noise	12	JPEG transmission errors
4	Masked noise	13	JPEG2000 transmission errors
5	High frequency noise	14	Non eccentricity pattern noise
6	Impulse noise	15	Local block-wise distortions
7	Quantization noise	16	Mean shift (intensity shift)
8	Gaussian blur	17	Contrast change
9	Image denoising		

Dans [14], les auteurs mentionnent qu'une mesure donnée doit être évaluée sur plusieurs bases d'images. Cependant, dans le cadre du présent travail, nous souhaitons aller plus loin et analyser les performances d'une mesure donnée pour différents types de dégradations. Ainsi, une seule base d'images peut être utilisée mais elle doit être la plus large possible. Par conséquent, la base TID 2008 a ici été choisie (voir tableau 1).

Différents critères peuvent être utilisés pour évaluer les performances d'une mesure de qualité d'images [4]. La corrélation de Pearson (PCC) a été ici utilisée et est

calculée entre les notes objectives et les notes subjectives. Nous utilisons également le gain de corrélation pour mesurer l'amélioration obtenue en termes de pourcentage. Ce gain correspond au rapport de corrélations des mesures comparées.

Le tableau 2 présente pour chaque dégradation, le classement (en termes de corrélation) des mesures de qualité d'images sélectionnées. On peut aisément constater que le classement obtenu diffère totalement d'une dégradation à une autre. En effet, la métrique PSNRHVS obtient le meilleur classement pour neuf types de dégradations (dégradation : 1-3, 5, 7, 9-11 et 13) et de mauvais résultats pour les autres (dégradation : 12).

Pour mieux comprendre l'importance de considérer le type de dégradations dans le processus d'estimation de la qualité de l'image, nous présentons les valeurs de corrélations obtenues pour trois mesures fréquemment utilisées: PSNRHVS, VIF et SSIM (voir le tableau 3). Plusieurs observations peuvent être faites:

- Les valeurs des PCCs diffèrent fortement, même entre deux positions de classement adjacentes. En effet, pour la dégradation 1, les corrélations des mesures PSNRHVS et VIF sont de 0.93 et 0.87 (soit un gain d'environ 7%).
- Pour certaines dégradations, la corrélation obtenue reste faible, même pour celle classée en première position. C'est le cas notamment pour la dégradation 16 (0.74, voir tableau 4).
- Le gain de corrélation varie fortement selon la mesure utilisée. A titre d'exemple, en utilisant le PSNRHVS pour la dégradation 1, les gains obtenus sont respectivement de 7% (VIF) et 15% (SSIM), tandis que pour la dégradation 16, les gains sont de -20% (VIF) et -24 % (SSIM), respectivement.
- Comme nous pouvons le constater en analysant le classement obtenu (tableau 2), les performances de la mesure IFC sont mauvaises pour la majorité des dégradations. Cependant, elle est classée en première position (0.83) pour la dégradation 14 avec des gains de 28% (PSNRHVS), 11% (VIF) et 32% (SSIM). Ainsi, une mesure de qualité d'images donnée peut obtenir de très mauvais résultats globalement, mais son utilisation pour un type de dégradation particulier reste pertinente.

Ainsi, nous pouvons conclure qu'une mesure donnée ne doit pas être utilisée pour évaluer la qualité d'une image quel que soit son type de dégradation.

Afin de pallier ce problème, certains auteurs ont proposés d'intégrer la notion de dégradation. A titre d'exemple, nous pouvons citer la méthode présentée dans [18]. Les auteurs proposent d'insérer une étape de classification qui vise à diviser les scores objectifs en cinq classes. L'étape de classification est basée sur des modèles statistiques de la scène et un SVM est utilisé comme classifieur. A noter ici que la dégradation n'est pas directement considérée mais le niveau de dégradation est décomposé.

Dans [19], les auteurs ajoutent une étape de classification afin de détecter le type de dégradations contenu dans l'image. A l'issue de cette classification, la métrique la plus performante est utilisée. L'étape de classification est ici basée sur une modélisation des dégradations par un réseau de neurones. Cette méthode a été testée sur plusieurs bases d'images. Les gains de corrélation obtenus ont été comparés à la métrique SSIM et varient de 0 à 32% pour la base TID 2008 et de 0% à 14% pour la base LIVE.

Tab 2 : Classement des mesures de qualité sélectionnées pour chaque degradation de la base TID 2008.

Dégradation	Classement							
	1	2	3	4	5	6	7	8
1	PSNRHVS	VIF	WSNR	SSIM	VIFP	VSNR	IFC	UQI
2	PSNRHVS	VIF	VIFP	SSIM	WSNR	VSNR	IFC	UQI
3	PSNRHVS	VIF	WSNR	VIFP	SSIM	VSNR	IFC	UQI
4	VIF	VIFP	SSIM	PSNRHVS	UQI	VSNR	IFC	WSNR
5	PSNRHVS	VIF	WSNR	VSNR	VIFP	SSIM	IFC	UQI
6	WSNR	PSNRHVS	VIF	VIFP	SSIM	VSNR	IFC	UQI
7	PSNRHVS	WSNR	VSNR	SSIM	VIF	VIFP	UQI	IFC
8	VIF	VIFP	WSNR	VSNR	PSNRHVS	SSIM	IFC	UQI
9	PSNRHVS	VSNR	WSNR	VIFP	VIF	SSIM	UQI	IFC
10	PSNRHVS	VIF	WSNR	VIFP	VSNR	SSIM	IFC	UQI
11	PSNRHVS	WSNR	VIFP	VSNR	VIF	UQI	SSIM	IFC
12	VIF	VIFP	UQI	SSIM	IFC	PSNRHVS	VSNR	WSNR
13	PSNRHVS	WSNR	SSIM	VIF	VIFP	VSNR	IFC	UQI
14	IFC (0.83)	VIFP	VIF	UQI	WSNR	PSNRHVS	SSIM	VSNR
15	SSIM	VIFP	UQI	VIF	IFC	PSNRHVS	VSNR	WSNR
16	WSNR (0.74)	SSIM	PSNRHVS	VIFP	VIF	IFC	UQI	VSNR
17	VIF	VIFP	SSIM	WSNR	UQI	PSNRHVS	VSNR	IFC

Tab 3 : Corrélations de Pearson obtenues et classement pour les mesures PSNRHVS, VIF et SSIM

Dégradation	Corrélation (Classement)		
	PSNRHVS	VIF	SSIM
1	**0.93 (1)**	0.87 (2)	0.81 (4)
2	**0.91 (1)**	0.90 (2)	0.83 (4)
3	**0.95 (1)**	0.87 (2)	0.82 (6)
4	0.81 (4)	**0.89 (1)**	0.83 (3)
5	**0.97 (1)**	0.94 (2)	0.87 (6)
6	**0.86** (2)	0.82 (3)	0.74 (5)
7	**0.89 (1)**	0.77 (5)	0.80 (4)
8	0.91 (5)	**0.94 (1)**	0.90 (6)
9	**0.96 (1)**	0.92 (5)	0.91 (6)
10	**0.97 (1)**	0.95 (2)	0.90 (6)
11	**0.95 (1)**	0.93 (6)	0.87 (8)
12	0.79 (6)	**0.88 (1)**	0.82 (4)
13	**0.92 (1)**	0.82 (5)	0.83 (4)
14	0.65 (7)	**0.75 (4)**	0.63 (8)
15	0.68 (6)	0.84 (4)	**0.89 (1)**
16	0.70 (3)	0.54 (6)	**0.73 (2)**
17	0.58 (6)	**0.89 (1)**	0.66 (3)

Ainsi, ces solutions alternatives permettent d'optimiser l'utilisation de toutes les mesures de qualité d'images avec référence et in fine d'améliorer le processus d'estimation globale de la qualité.

Tab 4 : Gains de corrélation.

Dégradation	Gains de correlation	
	Gain (First, Second)	Gain (First, Last)
1	7 (0.93, 0.87)	78 (0.93, 0.52)
2	1 (0.91, 0.90)	93 (0.90, 0.47)
3	10 (0.95, 0.87)	75 (0.95, 0.54)
4	4 (0.89, 0.86)	49 (0.89, 0.60)
5	2 (0.97, 0.94)	40 (0.97, 0.69)
6	3 (0.89, 0.86)	85 (0.89, 0.48)
7	4 (0.89, 0.85)	666 (0.89, 0.12)
8	1 (0.94, 0.93)	8 (0.94, 0.87)
9	2 (0.96, 0.94)	25 (0.96, 0.77)
10	2 (0.96, 0.95)	22 (0.96, 0.79)
11	0 (0.94, 0.94)	11 (0.94, 0.85)
12	2 (0.88, 0.86)	19 (0.88, 0.73)
13	11 (0.92, 0.83)	36 (0.92, 0.68)
14	10 (0.83, 0.76)	46 (0.83, 0.57)
15	5 (0.89, 0.85)	242 (0.89, 0.23)
16	2 (0.74, 0.73)	183 (0.74, 0.26)
17	4 (0.89, 0.86)	200 (0.89, 0.3)

4 Conclusion

Cette étude vise à montrer la pertinence de considérer le type de dégradations dans le processus d'estimation de la qualité d'images lorsque les mesures avec référence sont utilisées. Les gains de corrélation obtenus son considérable. Une utilisation optimale de ce type de mesures passe donc par l'ajout d'une étape intermédiaire, où la dégradation joue un rôle important.

5 Références

[1] M. Pedersen and J. Y. Hardeberg, " Survey of full-reference image quality metrics," Høgskolen i Gjøviks rapportserie ISSN: 1890-520X, Number 5, 2009.

[2] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, Vol. 13, issue 4, pp. 600-612, 2004.

[3] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, M. Carli, "New full-reference quality metrics based on HVS," International Workshop on Video Processing and Quality Metrics, 2006.

[4] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," Mar. 2000, ftp://vqeg.its.bldrdoc.gov/Documents/

[5] H.R. Sheikh.and A.C. Bovik, "Image information and visual quality," IEEE Transactions on Image Processing, Vol.15, no.2, pp. 430-444, 2006.

[6] H.R. Sheikh, A.C. Bovik and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," IEEE Transactions on Image Processing, Vol.14, no.12, pp. 2117-2128, 2005.

[7] T. Mitsa and K. Varkur, "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms", IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 301-304, 1993.

[8] D. Chandler, S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," IEEE Transactions on Image Processing, Vol. 16, pp. 2284-2298, 2007.

[9] Z. Wang and A. Bovik, "A universal image quality index," IEEE Signal Processing Letters, Vol. 9, pp. 81-84, 2002.

[10] H.R. Sheikh, Z. Wang, L. Cormack and A.C. Bovik, LIVE Image Quality Assessment Database, http://live.ece.utexas.eduesearch/quality

[11] P. Le Callet and F. Autrusseau, Subjective quality assessment IRCCyN/IVC database http://www.irccyn.ecnantes. fr/ivcdb/

[12] Z. M. Parvez Sazzad, Y. kawayoke, and Y. Horita, "Image Quality Evaluation Database," http://mict.eng.u-toyama.ac.jp/database_toyama/

[13] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola and F. Battisti, "Color Image Database for Evaluation of Image Quality Metrics," International Workshop on Multimedia Signal Processing, Australia, pp. 403-408, 2008.

[14] S. Tourancheau, F. Autrusseau, P. Z. M. Sazzad, Y. Horita, "Impact of the subjective dataset on the performance of image quality metrics," IEEE International Conference on Image Processing, pp. 365-368, 2008.

[15] M. Gaubatz, "Metrix MUX Visual Quality Assessment Package: MSE, PSNR, SSIM, MSSIM, VSNR, VIF, VIFP, UQI, IFC, NQM, WSNR, SNR," http://foulard.ece.cornell.edu/gaubatz/metrix_mux/

[16] Z. Wang and A. C. Bovik "Mean squared error: love it or leave it? - A new look at signal fidelity measures," IEEE Signal Processing Magazine, Vol. 26, no. 1, pp. 98-117, 2009.

[17] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," Invited Paper, IEEE Asilomar Conference on Signals, Systems and Computers, 2003.

[18] C. Charrier, O.Lezoray, and G. Lebrun, "A machine learning regression scheme to design a FR-image quality assessment algorithm," *European Conference on Colour in Graphics, Imaging,* 2012.

[19] A. Chetouani, A. Beghdadi and M. Deriche, "A hybrid system for distortion classification and image quality evaluation," Signal Processing: Image Communications, Vol. 27, pp. 948-960, 2012.