Calibration de la métrique MS-SSIM pour les distorsions de compression à l'aide d'une échelle perceptive des différences.

Christophe CHARRIER[1], Kenneth KNOBLAUCH[2], Laurence T. MALONEY[3], Alan C. BOVIK[4]

[1]Université de Caen Basse-Normandie, GREYC UMR CNRS 6072, Équipe Image, Campus II, Caen, France

[2] INSERM, U846, Stem Cell and Brain Research Institute, Bron, France ; Université Lyon 1, Lyon, France

[3] University of New-York, Department of Psychology, Center for Neural Science, NY, USA

[4]University of Texas at Austin, LIVE lab, Austin, TX, USA

christophe.charrier@unicaen.fr,ken.knoblauch@inserm.fr
ltm1@nyu.edu,bovik@ece.utexas.edu

Résumé – La problématique abordée dans cet article concerne le développement et l'optimisation d'une métrique de qualité pour les images compressées. Nous utilisons une méthode d'optimisation basée sur l'utilisation conjointe d'une échelle perceptive des différences et d'un algorithme génétique pour trouver les paramètres de la fonction de qualité développée. Les résultats obtenus montrent que les valeurs optimisées des paramètres induisent une augmentation significative de la corrélation avec les valeurs objectives obtenues par les observateurs humains.

Abstract – In this paper, we describe a recently developed method for assessing perceived image quality, Maximum Likelihood Difference Scaling (MLDS), and use it to assess the performance of MS-SSIM on compression distored images. MLDS allows us to quantify suprathreshold perceptual differences between pairs of images and to examine how perceived image quality, estimated through MLDS, changes as the compression rate is increased. We show how the data collected by MLDS allows us to recalibrate MS-SSIM to improve its performance.

1 Introduction

L'ubiquité de l'information visuelle numérique, sous la forme d'images compressées dans presque tous les secteurs économiques nécessite fiabilité et efficacité des méthodes pour évaluer la qualité de ce support visuel de l'information. Bien qu'un grand nombre de méthodes d'évaluation de la qualité des images avec référence aient été développées et ont montrées de très bons résultats (en terme de forte corrélation avec l'évaluation subjective de la qualité), les travaux de recherche liés à ce domaine continuent à prospérer. Une des techniques les plus couramment utilisées pour mener une évaluation subjective est de demander aux observateurs humains de noter la qualité de l'image présentée, généralement sur une dynamique d'échelle de 0 à 100 ; échelle normalisée par l'ITU (*International Telecommunications Union*) [1]. Obtenues à partir d'un large panel d'observateurs, la moyenne des notes est calculée pour chaque image. Cette moyenne est également appelée MOS (*Mean Opinion Score*). L'efficacité d'une métrique de qualité est habituellement résumée par la mesure d'une corrélation entre les scores algorithmiques et les valeurs MOS. Les mesures classiques de corrélation incluent le coefficient linéaire de Pearson, l'erreur quadratique moyenne et le coefficient de Spearman. Parmi tous les algorithmes de mesure de qualité des images avec références, l'indice MS-SSIM (*Multi-Scale Structural SIMilarity*) permet d'atteindre d'importantes valeurs de corrélation avec les notes MOS. Malgré cela, MS-SSIM utilise de nombreux paramètres qui n'ont pas été optimisés et qui restent quelque peu *ad hoc*. Dans cet article, nous présentons une méthode d'optimisation des valeurs des paramètres mis en jeu dans le cadre applicatif de la compression. La méthode d'optimisation est basée sur l'utilisation des échelles des différences (*MLDS–Maximum Likelihood Difference Scaling*).

2 L'échelle des différences MLDS

La méthode MLDS présente l'indéniable avantage de pouvoir être utilisée pour estimer les effets de la compression sur la perception de la qualité des images, peu importe le schéma de compression utilisé. L'objectif de cette méthode est de trouver une représentation numérique pour un ensemble d'objets à partir de l'ordonnancement des distances qui les séparent les uns des autres [2]. L'utilisation de telles techniques de construction d'une échelle des différences repose sur l'hypothèse qu'il existe des valeurs numériques $(n_0, n_1, \ldots, n_N)$, telles que l'observateur réponde que $(a_i : a_j)$ est plus grand que $(a_k : a_l)$, si et seulement si $\|n_i - n_j\| > \|n_k - n_l\|$. Ces valeurs définissent l'échelle des différences. La construction d'une telle échelle doit répondre à trois critères qui sont [3] :

un critère d'ordonnancement. Ce critère est en fait un critère de transitivité $(a_i \succ_1 a_j)\ \&\ (a_j \succ_1 a_k) \Rightarrow (a_i \succ_1 a_k)$

où le symbole $\succ_1$ signifie « plus grand que », au sens de la perception de l'observateur. Ce premier critère permet un ordonnancement des stimuli le long d'un axe suivant une propriété commune des stimuli.

un critère interne. Ce critère est aussi connu comme la condition des six points, qui se traduit par l'équation suivante :

$$\left.\begin{array}{l}(a_i : a_j) \succ_2 (a_l : a_m) \\ \text{et} \\ (a_j : a_k) \succ_2 (a_m : a_n)\end{array}\right\} \Rightarrow (a_i : a_k) \succ_2 (a_l : a_n)$$

où $\succ_2$ signifie « intervalle plus grand que », au sens de la perception de l'observateur. Ce critère peut être considéré comme une double transitivité.

un critère d'axiomes techniques. Ils permettent d'établir une relation entre les stimuli (a_i) et les valeurs numériques (n_i). Ainsi, nous avons $a_i \succ_1 a_j \Leftrightarrow n_i > n_j$ et $(a_i : a_j) \succ_2 (a_k : a_l) \Leftrightarrow \|n_i - n_j\| > \|n_k - n_l\|$

Maloney et Yang [4] ont proposé une méthode d'estimation des valeurs d'échelle par maximisation de la vraisemblance. Néanmoins, étant donné que la règle de décision implique une simple combinaison des réponses internes, elles peuvent également être estimée selon un modèle linéaire généralisé (*GLM-Generalized Linear Model*) [5, 6].

Cette technique a été appliquée pour évaluer la qualité d'un ensemble de 15 images originales compressées par le standard JPEG2000 sur 9 niveaux 0.1000, 0.3057, 0.5627, 0.7684, 0.9741, 1.1798, 1.3854, 1.5912 bpp. Nous avons utilisé l'implantation fournie par le projet JasPer [7]. Trente observateurs humains ont participé aux tests psychophysiques. Au final, nous avons obtenu des échelles de différence pour chaque observateur humain et pour chaque image.

3 L'indice MS-SSIM et sa corrélation avec les valeurs MLDS.

L'indice multi-échelle de qualité MS-SSIM [8] implique un facteur de distorsion dédié à la luminance (l) auquel sont adjoints un facteur de distorsion de contraste (c) et un critère de mesure de distorsion de structure (s) calculés entre deux images f et g, tel que :

$$\text{MS-SSIM}(f,g) = \underbrace{\left[\frac{2\mu_f\mu_g + C_1}{\mu_f^2 + \mu_g^2 + C_1}\right]^{\alpha_M}}_{l(f,g)} \prod_{i=1}^{M} \underbrace{\left[\frac{2\sigma_f\sigma_g + C_2}{\sigma_f^2 + \sigma_g^2 + C_2}\right]^{\beta_i}}_{c(f,g)} \underbrace{\left[\frac{2\sigma_{f,g} + C_3}{\sigma_f^2\sigma_g^2 + C_3}\right]^{\gamma_i}}_{s(f,g)} \quad (1)$$

où μ_f et μ_g représentent l'intensité moyenne des images f et g ; σ_f et σ_g sont les écarts-types calculés sur les images f et g ; $\sigma_{f,g} = \frac{1}{N-1}\sum_{i=1}^{N}(f_i - \mu_f)(g_i - \mu_g)$. C_1, C_2 et C_3 sont des constantes de stabilisation. Dans cet article, $M = 5$ et à l'échelle $i = 1$, l'image est de résolution originale. Les valeurs originales des exposants sont $\beta_1 = \gamma_1 = 0.0448$, $\beta_2 = \gamma_2 = 0.2856$, $\beta_3 = \gamma_3 = 0.3001$, $\beta_4 = \gamma_4 = 0.2363$, et $\alpha_5 = \beta_5 = \gamma_5 = 0.1333$ [8].

Afin de comparer les valeurs MLDS avec les scores obtenus par application de l'indice MS-SSIM, nous avons calculé les valeurs MS-SSIM entre deux images compressées consécutives par ordre croissant du taux de compression. Pour une même image originale et ses versions compressées, les valeurs cumulatives ont été calculées. Dans [9], les auteurs ont montré que malgré un fort taux de corrélation avec les observateurs humains, l'indice de qualité MS-SSIM échoue à prédire de façon pertinente les changements perceptuels entre les images à mesure que le taux de compression augmente. Plus précisément, le troisième facteur $s(f,g)$ est celui qui présente la moins bonne corrélation avec les valeurs MLDS. Plus spécifiquement, ce manque de corrélation est flagrant au début de chacune des échelles de différences. Dès lors, afin de pallier ce manque de corrélation, et ainsi augmenter le taux de corrélation avec les valeurs MLDS, il convient d'optimiser les valeurs des 15 coefficients α_i, α_i et γ_i, $\forall i \in [1, \cdots, 5]$.

4 Processus d'optimisation

Dans ce cas, l'expression de l'indice MS-SSIM peut être revue selon une formule paramétrique à 15 inconnues, telle que :

$$\text{MS-SSIM}(I, J, \alpha_i, \beta_i, \gamma_i; i = 1, \cdots, M) = \prod_{i=1}^{M} \left[l_i(I,J)^{\alpha_i} c_i(I,J)^{\beta_i} s_i(I,J)^{\gamma_i}\right] \quad (2)$$

sous les contraintes $\sum_{i=1}^{M} \alpha_i + \beta_i + \gamma_i = 1$ et $\forall i \in [1, \cdots, M]$, $0 \leq \alpha_i \leq 1, 0 \leq \beta_i \leq 1, 0 \leq \gamma_i \leq 1$.

La recherche des valeurs optimales des exposants peut alors être formulée comme la minimisation de la fonction d'erreur :

$$E(\alpha_i, \beta_i, \gamma_i; i = 1, \cdots, M) = \min\left(\sum_{j=1}^{K} (\text{MLDS}_j(I,J) - \text{fSSIM}_j(I, J, \alpha_i, \beta_i, \gamma_i))^2\right) \quad (3)$$

avec K le nombre des images testées pour lesquelles les valeurs MLDS sont disponibles, et $\text{fSSIM}_j(.)$ les valeurs MS-SSIM calculées entre deux images de taux de compression consécutifs et pour lesquelles une régression logistique a été réalisée. Étant donné que la fonction d'erreur est non-convexe et peut contenir plusieurs minima locaux, le choix d'un stratégie de recherche adaptée est crucial. Parmi toutes les techniques d'optimisation, les algorithmes génétiques offrent de notre point de vue toutes les garanties nécessaires à la recherche d'un minimum global [10]. Dans cet article, un chromosome est représenté par un vecteur à 15 dimension $(\alpha_1, \cdots, \alpha_M, \beta_1, \cdots, \beta_M, \gamma_1, \cdots, \gamma_M)$, et l'équation (3) représente la fonction de fitness.

Exposant	α_1	α_2	α_3	α_4	α_5
Value	0.1920	0.2169	0.2026	0.2136	0.1749
CI	[0.0989,0.2415]	[0.1877,0.2791]	[0.1692,0.2384]	[0.1765,0.2868]	[0.0814,0.2304]
Exposant	β_1	β_2	β_3	β_4	β_5
Value	0.9612	0.0097	0.0097	0.0097	0.0097
CI	[0.8288,0.9681]	[-0.0145,0.0933]	[0.0084,0.0112]	[0.0084,0.0112]	[-0.0133,0.1012]
Exposant	γ_1	γ_2	γ_3	γ_4	γ_5
Value	0.0082	0.1586	0.8167	0.0083	0.0082
CI	[0.0073,0.0086]	[0.1241,0.2530]	[0.7250,0.8501]	[0.0073,0.0086]	[0.0073,0.0086]

TAB. 1 – Valeur des 15 paramètres minimisant l'équation d'erreur 3 avec les intervalles de confiances à 95% associés.

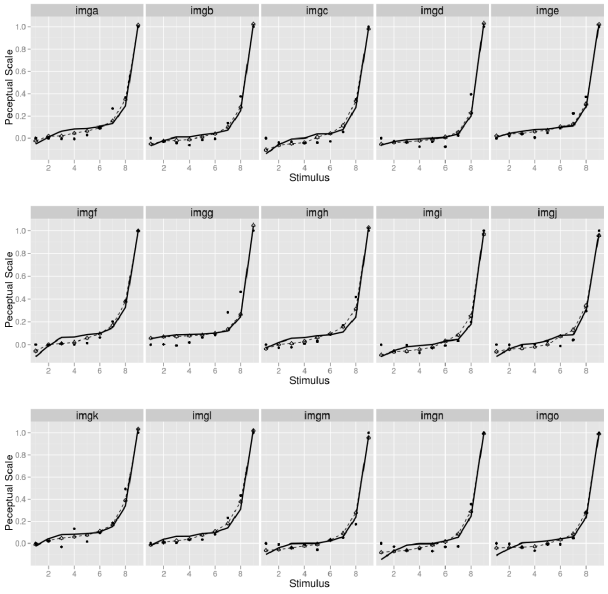

FIG. 1 – Résultats obtenus pour les images testées. Les points noirs représentent les valeurs MLDS, la courbe continue noire les valeurs MS-SSIM calculées avec les exposants originaux, et la courbe noire pointillée fait référence aux valeurs MS-SSIM calculées avec les nouveaux exposants du tableau 1.

5 Résultats

Le tableau 1 présente les résultats obtenus pour chacun des 15 exposants ainsi que l'intervalle de confiance à 95% associé, après minimisation de (3). Ces valeurs ont été obtenues selon un processus de bootstrap à 999 réplications. Afin de quantifier l'impact de ces nouvelles valeurs sur la prédiction des scores de qualité obtenus, les corrélations de Pearson, de Kendal et de Spearman ont été calculées et comparées avec les valeurs de corrélation obtenues lors de l'utilisation des valeurs initiales des coefficients α_i, α_i et γ_i. Deux bases d'images, contenant divers types de dégradation, ont été utilisées : la base LIVE v2 [11] et la base TID2008 [12].

La figure 1 présente le résultat de cette comparaison où les points noirs représentent les valeurs MLDS, la courbe continue noire les valeurs MS-SSIM calculées avec les exposants originaux, et la courbe noire pointillée fait référence aux valeurs MS-SSIM calculées avec les nouveaux exposants du tableau 1. Pour chacune des images testées, on observe un meilleur niveau

	JP2K		JPEG		Bruit blanc	
	Original	New	Original	New	Original	New
CC	0.783	0.810	0.730	0.742	0.9153	0.9142
KROCC	0.884	0.884	0.849	0.852	0.8887	0.8878
SROCC	0.980	0.991	0.962	0.981	0.9825	0.9813
	Flou gaussien		FastFading		All	
	Original	New	Original	New	Original	New
CC	0.8864	0.8623	0.725	0.788	0.7980	0.8142
KROCC	0.8591	0.8413	0.859	0.876	0.8021	0.8543
SROCC	0.9725	0.9627	0.965	0.974	0.9464	0.9762

TAB. 2 – Coefficients de corrélation calculés sur la base d'images LIVE pour les valeurs originales des exposants utilisées par l'indice de qualité MS-SSIM et les nouvelles valeurs.

de correspondance entre les valeurs MLDS et les nouvelles valeurs MS-SSIM. Ceci valide l'hypothèse sur l'utilisation de valeurs différentes pour chacun des quinze exposants pour optimiser l'indice MS-SSIM. Si l'on considère les valeurs des exposants pour le facteur de distorsion de structure, on observe que le troisième niveau de décomposition est celui de prime importance puisque la valeur de l'exposant associé est la plus grande au regard des quatre autres valeurs. Si l'on considère la situation de la mesure de distorsion de luminance, le résultat montre que les cinq niveaux de décomposition contribuent de manière quasiment identique dans l'estimation de la dégradation de luminance.

Pour affirmer qu'une différence de corrélation est statistiquement significative entre les deux indices MS-SSIM, un test d'hypothèse basé sur la variance des erreurs résiduelles entre les valeurs DMOS et les scores prédits par chacun des deux indices MS-SSIM est réalisé. Ce test est basé sur le test de Fisher permettant de déterminer si les variances de deux populations sont égales, ce qui est réalisé en comparant le ratio des deux variances calculées. Dans notre cas, l'hypothèse nulle correspond au cas où les erreurs résiduelles de l'indice MS-SSIM original sont statistiquement non distinguables (à un niveau de confiance de 95%) des erreurs résiduelles de l'indice MS-SSIM amélioré. Le tableau 2 présente les résultats obtenus sur la base LIVE. À la lecture des résultats, il convient d'affirmer que la performance de l'algorithme d'évaluation de la qualité MS-SSIM utilisant les nouvelles valeurs des exposants est meilleure que l'indice MS-SSIM original. Cependant, ceci n'est pas exact pour les images dégradée par un bruit blanc ou un flou gaussien. En considérant la totalité de la base LIVE, la valeur

	Dégradations											
	Bruit gaussien		Bruit couleur		Bruit corrélé		bruit masqué		bruit haute freq.		bruit impulsionnel	
	Original	New	Original	New	Original	New	Original	New	Original	New	Original	New
CC	0.7994	0.7700	0.8151	0.7913	0.8278	0.8340	0.8341	0.8224	0.8861	0.8333	0.6672	0.6399
KROCC	0.6139	0.5767	0.6013	0.5677	0.6148	0.6241	0.6117	0.5977	0.6419	0.5887	0.4846	0.4575
SROCC	0.8099	0.7767	0.8055	0.7748	0.8215	0.8265	0.8099	0.7923	0.8706	0.8211	0.6899	0.6547
	Bruit quantification		flou gaussien		Débruitage		JPEG		JPEG2000		transmission JPEG	
CC	0.8524	0.8355	0.9384	0.9292	0.9638	0.9485	0.9629	0.9796	0.9727	0.9823	0.8784	0.8983
KROCC	0.6569	0.6514	0.8169	0.7793	0.8316	0.8013	0.7489	0.7664	0.8559	0.8876	0.6637	0.6891
SROCC	0.8488	0.8361	0.9563	0.9355	0.9587	0.9458	0.9328	0.9571	0.9697	0.9812	0.8663	0.8852
	transmission JPEG2000		bruit non eccentré		Distorsions de bloc		Décalage d'intensité		Chgt. contraste		All	
CC	0.8414	0.8437	0.7417	0.7388	0.7290	0.8666	0.7322	0.7259	0.7721	0.5468	0.8332	0.8532
KROCC	0.6766	0.6957	0.5254	0.5335	0.5038	0.6309	0.5345	0.5427	0.4748	0.4068	0.6577	0.6699
SROCC	0.8609	0.8849	0.7375	0.7434	0.7109	0.8353	0.7239	0.7402	0.6349	0.5430	0.8543	0.8601

TAB. 3 – Coefficients de corrélation calculés sur la base d'images TID2008 pour les valeurs originales des exposants utilisées par l'indice de qualité MS-SSIM et les nouvelles valeurs.

de corrélation SROCC est statistiquement significativement plus élevée lorsque les nouvelles valeurs des exposants est utilisée. Ceci est dû à l'utilisation de la dégradation JPEG2000 pour construire l'échelle des différences. Le tableau 3 présente les résultats obtenus lors de l'utilisation de la base TID2008. Si l'on considère les artéfacts liés à la compression (dégradation 10 à 13), on constate que les différences sont statistiquement significatives au profit de l'indice MS-SSIM calculé avec les nouvelles valeurs d'exposants. En outre la dégradation 15 correspond à des distorsions locales de blocs d'intensité différentes, peut également être considérée comme le résultat d'un processus de compression. Pour cette dégradation, le gain de corrélation est statistiquement significatif. En revanche, le gain n'est pas statistiquement significatif lorsque la base TID est évaluée dans sa totalité.

6 Conclusion

Dans ces travaux, nous avons défini une méthodologie permettant de mesurer l'évolution des corrélations entre les scores prédits par une solution algorithmique et les notes de qualité attribuées par un être humain. Nous avons appliqué une échelle des différences calculée par maximum de vraisemblance MLDS sur un large panel d'images afin d'évaluer l'impact du schéma de compression JPEG2000 sur la précision de la prédiction de l'indice MS-SSIM vis-à-vis des jugements humains. En optimisant les valeurs de pondération des trois facteurs de distorsion, la corrélation avec le jugement humain peut être améliorée.

Références

[1] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," tech. rep., International Telecommunication Union, Geneva, Switzerland, 2002.

[2] F. W. Young, "Nonmetric multidimensional scaling : Recovery of metric information," *Psychometrika*, vol. 35, pp. 455–473, 1970.

[3] J. N. Yang and L. T. Maloney, "Difference scaling in color space near the neutral point," in *Investigative Ophtalmology and Visual Science*, vol. 39 of *ARVO annual meeting*, (Fort Lauderdale, Florida), p. 160, May 1998. abstracts.

[4] L. T. Maloney and J. N. Yang, "Maximum likelihood difference scaling," *Journal of Vision*, no. 3, pp. 573–585, 2003.

[5] C. Charrier, L. T. Maloney, H. Cherifi, and K. Knoblauch, "Maximum likelihood difference scaling of image quality in compression-degraded images," *Journal of the Optical Society of America*, vol. 24, no. 11, pp. 3418–3426, 2007.

[6] K. Knoblauch and L. T. Maloney, "MLDS : Maximum likelihood difference scaling in R," *Journal of Statistical Software*, vol. 25, pp. 1–26, 1 2008.

[7] U. University of Victoria, "The JasPer projet home page," *http ://www.ece.uvic.ca/ mdadams/jasper/*.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment : From error measurement to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, 2004.

[9] C. Charrier, K. Knoblauch, A. K. Moorthy, A. C. Bovik, and L. T. Maloney, "Comparison of image quality assessment algorithms on compressed images," in *SPIE,Image Quality and System Performance VII*, (San-Jose, California), Jan. 2010.

[10] J. H. Holland, *Adaptation in natural and artificial systems*. Cambridge, MA, USA : MIT Press, 1992.

[11] Laboratory for Image & Video Engineering, University of Texas (Austin), "LIVE Image Quality Assessment Database," *http ://live.ece.utexas.edu/research/Quality*, 2002.

[12] N. Ponomarenko, M. Carli, V. Lukin, K. E. ans J. Astola, and F. Battisti, "Color image database for evaluation of image quality metrics," in *International Workshop on Multimedia Signal Processing*, (Australia), pp. 403–408, Oct. 2008.