Analyse de métriques de qualité pour des images en couleurs

N. Maignan[1] F. Pierre[1] F. Sur[1]

[1] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

nicolas.maignan@univ-lorraine.fr

Résumé

L'évaluation par des métriques adaptées est cruciale dans de nombreux problèmes de l'analyse des images. Si l'utilisation de métriques basées sur une image de référence comme PSNR ou SSIM est classique, ceci n'est pas possible dans certaines applications où aucune image de référence n'est disponible. Il faut alors s'appuyer sur des métriques d'évaluation dites « sans références », comme BRISQUE ou NIQE, censées être bien corrélées à la perception humaine. Ces métriques furent initialement développées pour des images en niveaux de gris. Nous proposons une extension de BRISQUE et NIQE aux images en couleur, et nous évaluons leur intérêt face à différentes dégradations pouvant toucher l'information chromatique d'une image.

Mots Clef

Métriques de qualité, perception, couleur.

Abstract

Assessing results with metrics is crucial in many problems in image analysis. While the use of metrics based on a reference image such as PSNR or SSIM is conventional, this is not possible in some applications where no reference image is available. In these cases, we have to rely on so-called reference-free evaluation metrics, such as BRISQUE or NIQE, which are supposed to correlate well with human perception. These metrics were initially developed for grayscale images. We propose an extension of BRISQUE and NIQE to color images, and assess their usefulness when considering various degradations that can affect the chromatic information of an image.

Keywords

Quality metric, perception, color.

1 Introduction

Les méthodes modernes de traitement d'image ou de vision par ordinateur (édition ou génération d'images) rendent crucial de disposer de mesures fiables pour évaluer la qualité perceptuelle des images. Par exemple, un filtre peut dégrader la qualité d'une image, une reconstruction 3D nécessite d'utiliser des images de bonne qualité, la compression et les pertes lors de la transmission peuvent considérablement altérer la perception, etc.

Tandis que des métriques telles que PSNR ou SSIM sont utilisées de longue date, celles-ci présentent un certain nombre de défauts tels que le manque de corrélation à la perception humaine ou l'incapacité à détecter des défauts localisés. En outre, PSNR comme SSIM nécessitent une image de référence, ce qui est impossible dans certaines applications telles que la génération d'image ou la colorisation, où plusieurs résultats plausibles sont possibles.

Notre étude porte spécifiquement sur l'adaptation de métriques « sans références » utilisées pour quantifier la qualité d'images en niveau de gris, NIQE [8] et BRISQUE [7], à des images couleurs. L'application visée est celle de la colorisation d'images et de vidéos, qui consiste à ajouter de la couleur à des images en niveaux de gris. Les images issues de photographies ou de films historiques en noir et blanc n'ont pas de référence en couleur, rendant inutilisables les mesures basées sur une « image de référence » comme PSNR ou SSIM. De plus, la colorisation présente des défauts spécifiques, différents de ceux rencontrés dans la transmission d'images (application préférentielle de NIQE, BRISQUE, SSIM) : on observe des couleurs qui débordent d'un objet à un autre, un problème rencontré dans des techniques comme Deep Video Color Propagation [6], Deep Exemplar-Based Video Colorization [15] et Learned Variational Video Color Propagation [2]. On peut obtenir des couleurs ternes ou inappropriées, comme celles générées par Colorful Image Colorization [16] et DeOldify [9].

Pour évaluer les colorisations, diverses métriques sont utilisées dans l'état de l'art afin de mesurer la qualité des résultats, avec l'objectif d'obtenir de « bons » scores pour des images jugées comme convenablement colorisées. Cependant, l'absence de normes communes complique les comparaisons entre les différents modèles de colorisation.

Nos contributions principales sont :

- l'adaptation de métriques existantes de mesure de la qualité perceptuelle sans image de référence (BRISQUE [7] et NIQE [8]) à des images couleurs ;
- l'évaluation de la corrélation entre les scores de ces métriques et la perception humaine de la qualité d'images couleurs, en utilisant le jeu de données LIVE IQA [10] ;
- l'illustration de la cohérence de ces métriques dans des exemples de dégradation d'images en couleur.

2 Métriques étudiées

Nous utilisons les notations suivantes : I représente l'image dont on cherche à quantifier la qualité, $\hat{I}$ désigne l'image couleur de référence si besoin, H et W désignent respectivement la hauteur et la largeur des images.

2.1 PSNR

Le rapport du pic du signal sur bruit (PSNR) [3] est une mesure basée sur une image de référence qui compare pixel par pixel les images traitées aux images de référence. Pour cette mesure, un score plus élevé signifie que l'image I est proche de la référence $\hat{I}$.
Le PSNR d'une image couleur s'exprime comme suit :

$$\mathrm{PSNR}(I,\hat{I}) = 10\log_{10}\left(\frac{r^2}{\mathrm{EQM}(I,\hat{I})}\right) \tag{1}$$

avec $r = 255$ la plage des valeurs d'intensité dans un codage de 8 bits, et l'erreur quadratique moyenne (EQM) :

$$\mathrm{EQM}_{\mathrm{RVB}}(I,\hat{I}) = \frac{1}{3HW}\sum_{c\in\{\mathrm{R,V,B}\}}\sum_{i=1}^{H}\sum_{j=1}^{W}\Big(I_c(i,j)-\hat{I}_c(i,j)\Big)^2. \tag{2}$$

L'espace colorimétrique couramment utilisé est l'espace RVB [5, 4, 11, 14, 1], mais des articles récents présentent une évaluation sur l'espace CIELAB [6, 2], qui calcule la moyenne de l'EQM sur les canaux de chrominance a* et b* uniquement :

$$\mathrm{EQM}_{\mathrm{a^*b^*}}(I,\hat{I}) = \frac{1}{2HW}\sum_{c\in\{\mathrm{a^*,b^*}\}}\sum_{i=1}^{H}\sum_{j=1}^{W}\Big(I_c(i,j)-\hat{I}_c(i,j)\Big)^2. \tag{3}$$

Lorsque la luminance reste inchangée, l'utilisation de mesures comme le PSNR dans l'espace RVB peut donc atténuer la détection des erreurs de colorisation. En effet, les trois canaux RVB incorporent l'information de luminance, qui est identique entre la prédiction et la référence, ce qui diminue proportionnellement l'impact des inexactitudes chromatiques. Cela motive l'évaluation du PSNR dans les canaux de chrominance seulement, car ils sont indépendants de la luminance. PSNR étant calculé par moyenne sur toute l'image, des défauts localisés à quelques pixels de l'image I l'affectent peu.

2.2 SSIM

L'indice de similarité structurelle (SSIM) [12] est une mesure qui étudie la structure des images, basée sur une image de référence. Dans SSIM, ce ne sont pas les valeurs de chaque pixel de I qui sont comparées au pixel correspondant de $\hat{I}$ (comme dans PSNR), mais les moyennes et les écarts types de leurs voisinages de taille 11 par 11 pixels, qui sont comparés au voisinage correspondant dans l'image de référence. La mesure est basée, pour chaque canal de l'image I, sur trois grandeurs supposées indépendantes : la luminance l (notons qu'il ne s'agit pas de la luminance L* de l'espace CIELAB), le contraste c et la structure s. Ces quantités sont calculées pour chaque pixel (i,j) :

$$\mathrm{l}(I,\hat{I},i,j) = \frac{2\,\mu_I(i,j)\,\mu_{\hat{I}}(i,j)+C_1}{\mu_I^2(i,j)+\mu_{\hat{I}}^2(i,j)+C_1} \tag{4}$$

$$\mathrm{c}(I,\hat{I},i,j) = \frac{2\,\sigma_I(i,j)\,\sigma_{\hat{I}}(i,j)+C_2}{\sigma_I^2(i,j)+\sigma_{\hat{I}}^2(i,j)+C_2} \tag{5}$$

$$\mathrm{s}(I,\hat{I},i,j) = \frac{\sigma_I(i,j)\,\sigma_{\hat{I}}(i,j)+C_3}{\sigma_I(i,j)\,\sigma_{\hat{I}}(i,j)+C_3} \tag{6}$$

avec, pour deux images x et y, $C_1 = 2.55^2$, $C_2 = 7.65^2$, $C_3 = C_2/2$, et une évaluation sur un voisinage de (i,j) de taille 11 par 11 pixels (soit $N = 121$ pixels) :

$$\mu_x(i,j) = \frac{1}{N}\sum_{l=-5}^{5}\sum_{m=-5}^{5} w(l,m)\,x(i+l,j+m) \tag{7}$$

$$\sigma_x(i,j) = \left(\frac{1}{N-1}\sum_{l=-5}^{5}\sum_{m=-5}^{5} w(l,m) \times \big(x(i+l,j+m)-\mu_x(i+l,j+m)\big)^2\right)^{\frac{1}{2}} \tag{8}$$

$$\begin{aligned}\sigma_{xy}(i,j) = \frac{1}{N-1}\sum_{l=-5}^{5}\sum_{m=-5}^{5} & w(l,m) \\ & \times \big(x(i+l,j+m)-\mu_x(i+l,j+m)\big) \\ & \times \big(y(i+l,j+m)-\mu_y(i+l,j+m)\big).\end{aligned} \tag{9}$$

Ces différentes quantités sont pondérées par w, un noyau gaussien circulaire symétrique échantillonné avec un écart type de $1,5$. La mesure finale sur l'ensemble de l'image est calculée comme la moyenne sur tous les pixels du produit de l, c et s :

$$\mathrm{SSIM}(I,\hat{I}) = \frac{1}{3HW}\sum_{c\in\{\mathrm{R,V,B}\}}\sum_{i=1}^{H}\sum_{j=1}^{W} \mathrm{l}\Big(I_c,\hat{I}_c,i,j\Big)\,\mathrm{c}\Big(I_c,\hat{I}_c,i,j\Big)\,\mathrm{s}\Big(I_c,\hat{I}_c,i,j\Big). \tag{10}$$

Une première amélioration par rapport au PSNR est la prise en compte du voisinage de chaque pixel, ce qui permet de mieux refléter les variations locales. En effet, les écarts types permettent, en particulier, à la structure s de pénaliser plus sévèrement les textures qui disparaissent ou sont supprimées entre I et $\hat{I}$. Cependant, le calcul de la moyenne sur les canaux de l'image empêche le SSIM de prendre en compte la corrélation entre ces canaux. Finalement, le SSIM offre une faible amélioration par rapport au PSNR et devrait être utilisé dans les mêmes cas.

2.3 BRISQUE et NIQE

Les métriques d'évaluation BRISQUE (pour *Blind Referenceless Image Spatial QUality Evaluator*) [7] et NIQE

(pour *Natural Index Quality Evaluator*) [8] sont des mesures ne nécessitant pas d'image de référence pour évaluer la qualité des images, en utilisant des caractéristiques extraites sur un jeu d'images. BRISQUE utilise des images issues de LIVE IQA [10] ayant subi différentes dégradations et évaluées par des humains (selon le score DMOS, voir Section 3). NIQE n'utilise que des images supposées de bonne qualité, la base Pristine [8]. Plus les valeurs de ces mesures sont faibles, meilleure est la qualité perceptuelle des images correspondantes. Les deux mesures sont basées sur l'extraction d'un ensemble de caractéristiques de l'image testée I, ces caractéristiques permettant dans un second temps d'établir un score de qualité perceptuelle censé être corrélé à l'évaluation par des humains.
Nous décrivons à présent les métriques BRISQUE et NIQE telles que présentées par leurs auteurs, dans le cadre de l'évaluation d'images en niveau de gris.

Métriques originelles. Dans les deux cas, il faut d'abord estimer des caractéristiques sur les images. Cela nécessite une normalisation préalable de l'intensité lumineuse de l'image I pour obtenir $\tilde{I}$. Pour tout pixel (i,j) :

$$\tilde{\mathrm{I}}(i,j) = \frac{I(i,j) - \mu_I(i,j)}{\sigma_I(i,j) + 1} \tag{11}$$

$$\mu_I(i,j) = \sum_{l=-3}^{3}\sum_{m=-3}^{3} w(l,m)\, I(i+l, j+m) \tag{12}$$

$$\sigma_I(i,j) = \left(\sum_{l=-3}^{3}\sum_{m=-3}^{3} w(l,m) \times \left(I(i+l,j+m) - \mu_I(i+l,j+m)\right)^2\right)^{\frac{1}{2}} \tag{13}$$

avec w, un noyau gaussien circulaire symétrique de taille 7 par 7, échantillonné avec un écart type de 3.
Un premier ensemble de caractéristiques est estimé par ajustement d'une distribution gaussienne généralisée (abrégé dans la suite GGD, pour *Generalized Gaussian Distribution*) aux valeurs de $\tilde{I}$:

$$\mathrm{GGD}\left(\tilde{I}, \alpha, \sigma^2\right) = \frac{\alpha}{2\beta\,\Gamma\left(1/\alpha\right)} \exp\left(-\left(\frac{\left|\tilde{I}\right|}{\beta}\right)^{\alpha}\right) \quad \text{où} \quad \beta = \sigma\sqrt{\frac{\Gamma\left(1/\alpha\right)}{\Gamma\left(3/\alpha\right)}} \tag{14}$$

avec Γ la fonction gamma. Les valeurs que l'on considère comme caractéristiques d'intérêt sont d'une part le paramètre de forme α et d'autre part la variance σ^2.
Ensuite, quatre autocorrélations spatiales sont calculées en translatant $\tilde{I}$ dans quatre directions - horizontale, verticale et le long des deux diagonales - et en multipliant $\tilde{I}$ par ces translations.

$$\mathrm{H}(i,j) = \tilde{\mathrm{I}}(i,j)\,\tilde{\mathrm{I}}(i,j+1) \tag{15}$$

$$\mathrm{V}(i,j) = \tilde{\mathrm{I}}(i,j)\,\tilde{\mathrm{I}}(i+1,j) \tag{16}$$

$$\mathrm{D1}(i,j) = \tilde{\mathrm{I}}(i,j)\,\tilde{\mathrm{I}}(i+1,j+1) \tag{17}$$

$$\mathrm{D2}(i,j) = \tilde{\mathrm{I}}(i,j)\,\tilde{\mathrm{I}}(i+1,j-1) \tag{18}$$

Un second ensemble de caractéristiques est estimé à partir d'une distribution gaussienne généralisée asymétrique (abrégé dans la suite AGGD, pour *Asymetric Generalized Gaussian Distribution*) ajustée à chaque corrélation, ce qui fournit quatre paramètres pour chacun des produits définis précédemment. Pour BRISQUE, les valeurs d'intérêt retenues sont le paramètre de forme ν, la moyenne η, la variance gauche σ_l et la variance droite σ_r, tandis que pour NIQE, il s'agit de ν, η, β_l et β_r. Cette étape permet d'obtenir un total de seize caractéristiques à partir du modèle AGGD donné par :

$$\mathrm{AGGD}\left(x,\nu,\eta,\sigma_l^2,\sigma_r^2\right) = \begin{cases} \frac{\nu}{(\beta_l+\beta_r)\,\Gamma(1/\nu)} \\ \quad \times \exp\left(-\left(-x/\beta_l\right)^{\nu}\right) \\ \text{où} \quad \beta_l = \sigma_l\sqrt{\frac{\Gamma(1/\nu)}{\Gamma(3/\nu)}}, \\ \text{si} \quad x < 0 \\ \\ \frac{\nu}{(\beta_l+\beta_r)\,\Gamma(1/\nu)} \\ \quad \times \exp\left(-\left(x/\beta_r\right)^{\nu}\right) \\ \text{où} \quad \beta_r = \sigma_r\sqrt{\frac{\Gamma(1/\nu)}{\Gamma(3/\nu)}}, \\ \text{si} \quad x \geq 0 \end{cases} \tag{19}$$

$$\eta = (\beta_r - \beta_l)\frac{\Gamma\left(2/\nu\right)}{\Gamma\left(1/\nu\right)}. \tag{20}$$

Enfin, l'ensemble du processus est répété sur une version de l'image, $I_{1/2}$, qui est redimensionnée par interpolation bicubique à la moitié de sa résolution d'origine, ce qui permet d'extraire des caractéristiques supplémentaires à une résolution inférieure, ce qui fournit un total de trente-six caractéristiques.
BRISQUE procède à l'estimation des paramètres sur l'ensemble de l'image, ce qui a pour inconvénient de perdre l'information locale. Pour pallier ce défaut, NIQE estime les caractéristiques sur des imagettes de taille 96 par 96 pixels, chaque image étant décrite par l'ensemble des descripteurs extraits de toutes les imagettes.
Une fois que toutes les caractéristiques du jeu de données considéré ont été extraites (sur les images entières de LIVE IQA [10] pour BRISQUE et sur les blocs des images de Pristine [8] pour NIQE), la métrique elle-même peut être calculée, de la manière suivante.
BRISQUE entraîne une machine à vecteurs supports de régression (ϵ-SVR) pour prédire un score de qualité (le score DMOS, indiqué dans LIVE IQA) à partir des caractéristiques extraites. Le score BRISQUE d'une nouvelle image I est alors le score inféré par la SVR entraînée.

NIQE, pour sa part, estime un modèle gaussien multivarié (abrégé dans la suite MVG, pour *multivariate Gaussian*) à partir des caractéristiques des imagettes de Pristine. Le score NIQE est finalement obtenu comme la distance entre le MVG ajusté sur les caractéristiques issues de Pristine et le MVG ajusté sur les caractéristiques de I :

$$\mathrm{NIQE}(I) = \frac{1}{N_{\text{car}}} \sqrt{(\mu - \hat{\mu})^\top \left(\frac{\Sigma + \hat{\Sigma}}{2}\right)^{-1} (\mu - \hat{\mu})} \tag{21}$$

avec $N_{\text{car}} = 36$ le nombre de caractéristiques, la moyenne $\hat{\mu}$ et la covariance $\hat{\Sigma}$ empiriques des vecteurs de caractéristiques des imagettes extraites de l'image évaluée et la moyenne μ et la covariance Σ empiriques des vecteurs des caractéristiques des blocs du jeu de données de référence. Notons que seules les imagettes de référence ayant un indice de *sharpness* au-dessus d'un certain seuil sont considérée (voir Section II.B de [8]).

Nous avons remarqué que le calcul des équations (12) et (13) entraîne sur certaines images des différences assez faibles lors du calcul des convolutions (de l'ordre de 10^{-14}) entre l'implémentation Matlab originale et notre implémentation Python. Au fur et à mesure des calculs de caractéristiques et d'estimation des modèles, cela donne des différences notables sur les scores finaux (de l'ordre de 10^{-1}), expliquant une différence parfois significative entre l'implémentation originale de BRISQUE [1] et NIQE [2] et notre ré-implémentation.

BRISQUE et NIQE traitent les images en niveaux de gris et ne sont pas conçues pour mesurer l'impact d'artefacts spécifiques à des problèmes de colorisation par exemple (l'usage qui en est fait dans [13] semble donc impropre).

Extension aux images couleurs. Nous étendons à présent BRISQUE et NIQE aux images couleurs, en calculant des corrélations supplémentaires sur les canaux chromatiques. Cette approche permettra d'évaluer plus précisément les artefacts de colorisation et d'obtenir des scores qui reflètent mieux la qualité perceptuelle des images en couleur.

Pour BRISQUE et NIQE, trois nouvelles versions sont proposées. Dans la version que l'on a nommée "Features RVB", les caractéristiques ne sont plus calculées sur l'intensité lumineuse uniquement, mais sur chacun des trois canaux RVB normalisés :

$$\tilde{\mathrm{R}}(i,j) = \frac{I_R(i,j) - \mu_{I_R}(i,j)}{\sigma_{I_R}(i,j) + 1} \tag{22}$$

$$\tilde{\mathrm{V}}(i,j) = \frac{I_V(i,j) - \mu_{I_V}(i,j)}{\sigma_{I_V}(i,j) + 1} \tag{23}$$

$$\tilde{\mathrm{B}}(i,j) = \frac{I_B(i,j) - \mu_{I_B}(i,j)}{\sigma_{I_B}(i,j) + 1}. \tag{24}$$

1. http://live.ece.utexas.edu/research/quality/BRISQUE_release.zip
2. http://live.ece.utexas.edu/research/quality/niqe_release.zip

Cela donne un total de trois fois 36 caractéristiques, soit 108 caractéristiques.

Dans la version nommée "Correl RVB", des corrélations chromatiques sont ajoutées aux corrélations calculées sur l'intensité lumineuse normalisée. Ces corrélations sont calculées entre les canaux RVB normalisés, ce qui donne 4 caractéristiques supplémentaires extraites pour chaque AGGD correspondant à RV, RB, ou VB :

$$\mathrm{RV}(i,j) = \tilde{\mathrm{R}}(i,j)\,\tilde{\mathrm{V}}(i,j) \tag{25}$$

$$\mathrm{RB}(i,j) = \tilde{\mathrm{R}}(i,j)\,\tilde{\mathrm{B}}(i,j) \tag{26}$$

$$\mathrm{VB}(i,j) = \tilde{\mathrm{V}}(i,j)\,\tilde{\mathrm{B}}(i,j). \tag{27}$$

Ceci donne un total de 36 caractéristiques plus 24 caractéristiques, soit 60 caractéristiques pour "Correl RVB".

Enfin, on rajoute aux caractéristiques RVB de "Features RVB" les caractéristiques de corrélation entre canaux de "Corrélation RVB" dans la version "All RVB". Les caractéristiques sont extraites sur chacun des canaux normalisés, tant en GGD qu'en AGGD, et les corrélations intercanaux donnent des caractéristiques par AGGD. Ceci donne au total 132 caractéristiques.

3 Vérification de cohérence

Pour vérifier la cohérence des métriques couleur proposées, nous reproduisons le protocole expérimental de [10].

Le jeu de données LIVE IQA [10] a été constitué en appliquant diverses modifications (typiques de problèmes de transmission) à 29 images de base. Des compressions JPEG (JPEG dans les tableaux) et JPEG2000 (JP2K) ont été appliquées, du bruit blanc (WN) a été ajouté, et un flou gaussien (Blur) ainsi qu'un évanouissement de Rayleigh (FF) ont été appliqués. Ces transformations ont abouti à un ensemble de 779 images. Chacune de ces images a été évaluée par environ 20 personnes, attribuant une note comprise entre 0 et 100.

Après la collecte des notes, toutes les évaluations ont été réalignées selon une méthode décrite dans la section C de [10]. Un score de différence moyenne d'opinion (DMOS) a été obtenu, reflétant la qualité perceptuelle moyenne de chaque image.

Pour chaque distorsion, la base de données a été divisée en 1000 ensembles d'entraînement et de test distincts. Cette approche de validation croisée permet de garantir que les résultats obtenus sont robustes et généralisables. L'ensemble d'entraînement a été utilisé pour entraîner les différents SVR nécessaires pour BRISQUE et ses variantes. Concernant les paramètres de l'entraînement de la ϵ-SVR, nous utilisons ceux connus des auteurs : un noyau à base radiale de paramètre $\gamma = 0,05$ et un coefficient de régularisation $C = 1024$. Les valeurs de ϵ ont été estimées par validation croisée : $\epsilon = 2,78$ pour "Original", $\epsilon = 2,78$ pour "Features RVB", $\epsilon = 3,44$ pour "Correl RVB" et $\epsilon = 1,39$ pour "All RVB". Quant à l'ensemble de test, il a été utilisé pour obtenir les scores avec les différentes

TABLE 1 – Médianes des coefficients de corrélation de Spearman sur 1000 divisions d'ensembles d'entraînement et de test sur la base de données LIVE IQA.

Metrics	JP2K	JPEG	WN	Blur	FF	All
PSNR - RVB	0.8811	0.8931	0.9793	0.7773	0.8793	0.8718
PSNR - a*b*	0.7613	0.8926	0.9781	0.7746	0.6374	0.6453
SSIM	0.9316	0.9423	0.9542	0.8924	0.9305	0.8831
BRISQUE	0.9395	0.9765	0.9867	0.9690	0.9138	0.9644
BRISQUE - Original	0.8706	0.9584	0.9823	0.8983	0.8229	0.9056
BRISQUE - Correl	0.9222	0.9745	0.9847	0.9643	0.8966	0.9610
BRISQUE - Features	0.9225	0.9728	0.9852	0.9621	0.8798	0.9527
BRISQUE - All	0.9129	0.9711	0.9828	0.9510	0.8889	0.9501
NIQE	0.9117	0.9347	0.9640	0.9246	0.8557	0.9071
NIQE - Original	0.9011	0.9286	0.9591	0.9167	0.8502	0.9074
NIQE - Correl	0.9044	0.9303	0.9690	0.8798	0.8623	0.8036
NIQE - Features	0.8961	0.9175	0.9645	0.8970	0.8547	0.8767
NIQE - All	0.8894	0.9206	0.9700	0.8658	0.8446	0.7785

TABLE 2 – Médianes des coefficients de corrélation de Pearson sur 1000 divisions d'ensembles d'entraînement et de test sur la base de données LIVE IQA.

Metrics	JP2K	JPEG	WN	Blur	FF	All
PSNR - RVB	0.8701	0.8922	0.9785	0.7810	0.8765	0.8439
PSNR - a*b*	0.7744	0.9014	0.9687	0.7833	0.6677	0.6487
SSIM	0.9149	0.9350	0.9482	0.8598	0.9092	0.7933
BRISQUE	0.9479	0.9862	0.9928	0.9703	0.9368	0.9664
BRISQUE - Original	0.8696	0.9681	0.9910	0.9036	0.8564	0.9115
BRISQUE - Correl	0.9318	0.9846	0.9930	0.9702	0.9246	0.9629
BRISQUE - Features	0.9270	0.9834	0.9925	0.9672	0.9072	0.9536
BRISQUE - All	0.9179	0.9810	0.9895	0.9585	0.9109	0.9513
NIQE	0.9203	0.8760	0.8589	0.9266	0.8723	0.6552
NIQE - Original	0.9018	0.8522	0.8522	0.9164	0.8608	0.7126
NIQE - Correl	0.9062	0.8306	0.6242	0.7841	0.8865	0.3339
NIQE - Features	0.8965	0.8269	0.6352	0.9068	0.8679	0.3498
NIQE - All	0.8907	0.8102	0.6880	0.6538	0.8656	0.3724

métriques. Pour chaque division, les coefficients de corrélation de Spearman et de Pearson ont été calculés entre les scores de qualité prédits par les différentes métriques (PSNR sur RVB, PSNR sur a*b*, SSIM, les différentes versions de BRISQUE et de NIQE) et les scores subjectifs fournis dans la base de données LIVE IQA. La valeur médiane de ces corrélations est rapportée pour chaque distorsion de LIVE IQA et chaque métrique, respectivement dans la Table 1 pour les médianes de corrélations de Spearman et dans la Table 2 pour les médianes de corrélations de Pearson. Ces tables donnent pour chaque colonne la division du jeu de donnée correspondant à une modification donnée et la colonne "All" regroupe l'ensemble de LIVE IQA. La ligne "PSNR - RVB" utilise l'Équation (2) alors que la ligne "PSNR - a*b*" utilise l'Équation (3). Les lignes "BRISQUE" et "NIQE" utilisent les estimateurs SVR et MVG fournis par les auteurs tandis que les lignes "Original" correspondent à nos entrainements et estimations de ces descripteurs, et les lignes "Correl", "Features" et "All" correspondent aux versions adaptées aux images en couleur définies dans la Section 2.

Les mesures "PSNR - RVB" et "SSIM" obtiennent de très fortes corrélations avec les DMOS, de l'ordre de 0.8 à 0.9 pour la plupart des altérations. Ce sont des mesures qui sont extrêmement sensibles au bruit gaussien comme le montre les corrélations de Spearman supérieures à 0.95 (0.9793 pour "PSNR - RVB" et 0.9542 pour "SSIM") et les corrélations de Pearson supérieures à 0.94 (0.9785 pour "PSNR - RVB" et 0.9482 pour "SSIM").

La mesure "PSNR - a*b*" obtient des corrélations plus faibles que sa contrepartie RVB pour la compression JPEG 2000 et pour l'évanouissement de Rayleigh, ce qui montre que l'inclusion de la luminance reste primordiale pour l'évaluation de certaines altérations dues à la transmission d'images.

Concernant BRISQUE, notre entraînement avec les caractéristiques originales montre une corrélation plus faible que celle des auteurs. Comme cette version ("BRISQUE" dans le tableau) a été entrainée sur l'intégralité de LIVE IQA (et donc sur les bases test), le biais introduit pourrait expliquer la différence obtenue. Une autre explication pourrait être les erreurs numériques mentionnées dans la Section 2. Il faut cependant noter que l'ajout de la corrélation entre canaux, qui est visible en comparant les lignes "Original" avec "Correl", et "Features" avec "All", augmente significativement la corrélation entre BRISQUE et les DMOS, avec des résultats comparables à la version des auteurs pour "Correl". Ceci montre que l'information importante pour capturer les défauts ajoutés parait être contenue dans les corrélations entre canaux.

Concernant NIQE, il n'y a pas de biais à utiliser le modèle "NIQE" fourni par les auteurs, car il est entrainé sur une base d'image différente de la base utilisée dans cette expérience. On remarque dans les différences de corrélations entre "NIQE" et notre version réestimé "NIQE - Original" que les erreurs numériques entre MATLAB et Python explicitées précédemment ont une influence certaine mais négligeable sur les corrélations. On note que même s'il y a des corrélations similaires entre les lignes "Original" avec "Correl", et "Features" avec "All", une différence est visible pour le défaut de flou gaussien "Blur" et l'ensemble du jeu de donnée "All". Notons qu'une partie des différences entre les variantes de NIQE et de BRISQUE réside dans l'utilisation d'une base d'images de référence supposée de qualité pour NIQE, et d'images variées évaluées par des humains pour BRISQUE. Néanmoins, en s'attardant sur les corrélations linéaires de Pearson obtenues par NIQE sur l'ensemble des données, on observe une très faible corrélation entre le score des différentes versions de NIQE et le DMOS, une corrélation autour de 0.7 pour "NIQE" et "NIQE - Original" et autour de 0.3 pour les autres versions "Correl", "Features" et "All". Ceci est dû à certaines images de la classe WN, ce qui est bien visible sur le nuage de points de la Figure 1. La corrélation de Spearman n'étant pas sensible à cette non-linéarité, elle reste élevée pour "All".

4 Sensibilité des métriques perceptuelles à des perturbations chromatiques

Afin d'étudier plus précisément la sensibilité des différentes métriques à la couleur, on applique sur des images un ensemble de défauts qui sont typiques du domaine de la colorisation, un domaine où la luminance des images reste inchangée, mais où les défauts perceptuels apparaissent

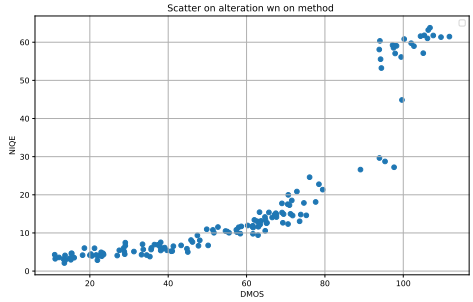

FIGURE 1 – Nuage de points entre les DMOS et les résultats de NIQE ("NIQE" sur les Tables 1 et 2) sur la partie bruit blanc du jeu de donnée LIVE IQA ("WN" sur les Tables 1 et 2).

uniquement dans les canaux de chrominances a* et b*. On applique à sept images couleurs les altérations suivantes :

- une rotation de la teinte de l'image (Figure 2) en la passant dans l'espace colorimétrique HSV, où la teinte est exprimée dans le canal circulaire H avec des valeurs de 0° à 360° ;
- une désaturation de l'image (Figure 5) en la passant dans l'espace HSV et en appliquant sur le canal S un facteur de désaturation allant de 0 (aucune désaturation) à 1 (image en niveaux de gris) ;
- l'application d'un bruit blanc gaussien sur les chrominances de l'image en CIELAB (Figure 8), paramétré par un écart type allant de 0 à 2 (pour des chrominances entre 0 et 1).

Les graphiques montrés représentent en ordonnée la valeur médiane sur les sept images des différents scores de métriques selon le paramètre gouvernant la transformation étudiée en abscisse.

Dans la première expérience, la rotation des teintes entraîne des couleurs qui ne correspondent pas aux couleurs possibles pour les différents objets. Par exemple, le miel dans la Figure 2a devient bleu dans la Figure 2b, ce qui ne correspond pas à une couleur naturelle. On peut voir que ceci impacte peu les métriques des auteurs, en bleu dans les Figures 3 et 4. Nos métriques entraînées, en revanche, réagissent aux changement (rappelons qu'une augmentation de métrique signe ici une dégradation perceptuelle). Pour BRISQUE (Figure 3), la version "Features" présente une variabilité plus importante que la version "Correl", ces effets s'additionnant dans la version "All". Pour NIQE (Figure 4), l'ajout de la corrélation entre les canaux RVB de "Correl" permet de détecter la variation de teinte, contrairement aux versions basées sur la luminance, à "Features" et "All" qui montrent des variations moindres.

Dans la seconde expérience, une désaturation de 50% n'altère que peu la qualité de l'image (Figure 5a). Par contre, lorsque la désaturation devient supérieure à 90 %, les couleurs disparaissent (Figure 5b). Cette altération est détectée par BRISQUE dans la Figure 6 par l'ajout de la couleur

(a) Image originale (0°) (b) Rotation de 171°

FIGURE 2 – Rotation de la teinte H dans l'espace HSV.

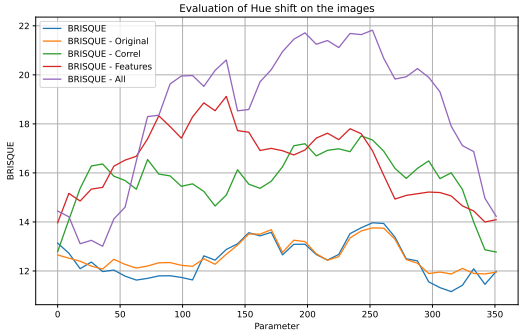

FIGURE 3 – Évaluation des différents BRISQUE sur une rotation de la teinte.

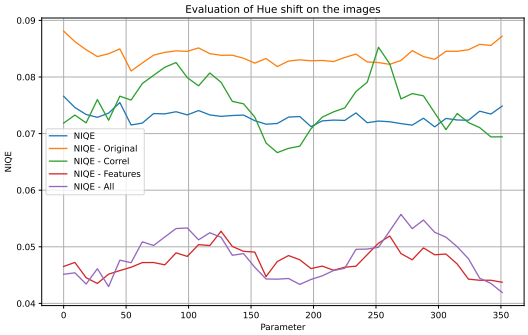

FIGURE 4 – Évaluation des différents NIQE sur une rotation de la teinte.

dans les nouvelles versions. C'est également le cas pour NIQE dans la Figure 7, mais uniquement pour une valeur très élevée de désaturation, et seulement lorsque la corrélation entre les canaux RVB est ajoutée, comme dans les versions "Correl" et "All". On peut noter une décroissance des différentes métriques lorsque le paramètre de désaturation augmente, ce qui signifie que les images légèrement désaturées sont jugées perceptuellement plus satisfaisantes. Il s'agit sans doute là d'un signe que les images initiales présentaient une saturation exagérément élevée, ce qui est souvent le cas de base d'images disponibles sur Internet.

(a) Désaturation de 50 % (b) Désaturation de 95 %

FIGURE 5 – Désaturation dans l'espace HSV.

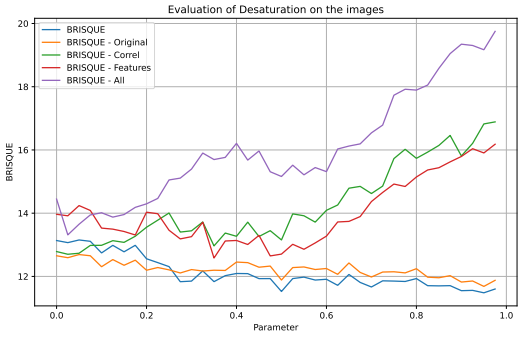

FIGURE 6 – Évaluation des différents BRISQUE sur une désaturation.

Dans la troisième expérience, le bruitage dégrade la qualité de l'image assez rapidement avec l'augmentation de l'écart type (Figure 8). Autant BRISQUE dans la Figure 9 que NIQE dans la Figure 10 sont sensibles au bruitage des chrominances. Toutes les variantes présentant une augmentation du score, y compris celles ne prenant en comptant que la luminance. L'explication est que la luminance sur laquelle les descripteurs de BRISQUE et NIQE sont calculés n'est pas le canal L*, et cette luminance (image I dans la Section 2.3) est impactée par le bruit ajouté sur les canaux a* et b*. On note cependant que, tant pour BRISQUE que pour NIQE, la version "Correl" est celle qui présente

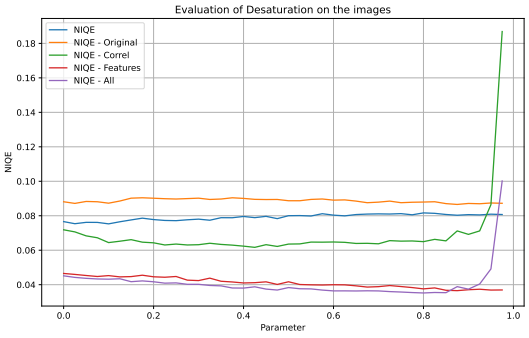

FIGURE 7 – Évaluation des différents NIQE sur une désaturation.

la plus forte pénalisation de la baisse de qualité de l'image.

(a) Écart type de 0.5 (b) Écart type de 1.5

FIGURE 8 – Bruitage blanc des chrominances dans l'espace CIELAB.

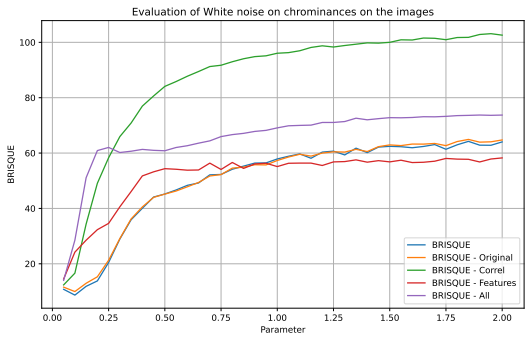

FIGURE 9 – Évaluation des différents BRISQUE sur un bruitage blanc des chrominances.

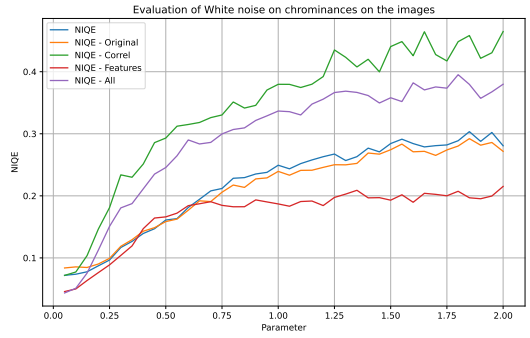

FIGURE 10 – Évaluation des différents NIQE sur un bruitage blanc des chrominances.

5 Conclusion

Nous avons introduit des extensions aux images couleur des métriques d'évaluation « sans référence » BRISQUE et NIQE (Section 2). Nous avons vérifié la cohérence de ces nouvelles métriques face à des distorsions classiquement utilisées pour vérifier la pertinence des métriques (Section 3) : si les résultats sont semblables aux métriques de la littérature, l'utilisation de la couleur permet d'obtenir des corrélations au score de qualité perceptuelle un peu plus élevées pour BRISQUE, mais l'amélioration n'est pas notable avec NIQE. La raison est que cette expérience n'est pas dédiée à des défauts de nature chromatique mais à des dégradations de transmission du signal. Notre second ensemble d'expériences introduit une évaluation perceptuelle face à des défauts typiques du traitement des images couleurs (Section 4). Il montre que BRISQUE et NIQE permettent souvent de bien quantifier la qualité perceptuelle. Par ailleurs, et sans surprise, l'évaluation dépend de la base d'images qui a été choisie comme référence d'images naturelles de qualité, ce qui se traduit par les différences de variations de BRISQUE et NIQE, les deux métriques utilisant des bases différentes.

Références

[1] S. Chen, X. Li, X. Zhang, M. Wang, Y. Zhang, J. Han, and Y. Zhang. Exemplar-based video colorization with long-term spatiotemporal dependency. *Knowledge-Based Systems*, 284 :111240, 2024.

[2] M. Hofinger, E. Kobler, A. Effland, and T. Pock. Learned Variational Video Color Propagation. In *Proc. ECCV*, pages 512–530, 2022.

[3] Q. Huynh-Thu and M. Ghanbari. The accuracy of PSNR in predicting video quality for different video scenes and frame rates. *Telecommunication Systems*, 49(1) :35–48, 2012.

[4] S. Iizuka and E. Simo-Serra. DeepRemaster : Temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics*, 38(6) :176 :1–176 :13, 2019.

[5] C. Lei and Q. Chen. Fully Automatic Video Colorization With Self-Regularization and Diversity. In *Proc. CVPR*, pages 3748–3756, 2019.

[6] S. Meyer, V. Cornillère, A. Djelouah, C. Schroers, and M. Gross. Deep Video Color Propagation. In *Proc. BMVC*, 2018.

[7] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing*, 21(12) :4695–4708, 2012.

[8] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3) :209–212, 2013.

[9] A. Salmona, L. Bouza, and J. Delon. DeOldify : A Review and Implementation of an Automatic Colorization Method. *Image Processing On Line*, 12 :347–368, 2022.

[10] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Transactions on Image Processing*, 15(11) :3440–3451, 2006.

[11] Z. Wan, B. Zhang, D. Chen, and J. Liao. Bringing Old Films Back to Life. In *Proc. CVPR*, pages 17673–17682, 2022.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment : From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4) :600–612, 2004.

[13] R. Ward, D. Bigioi, S. Basak, J. G. Breslin, and P. Corcoran. LatentColorization : Latent Diffusion-Based Speaker Video Colorization. *IEEE Access*, 12 :81105–81121, 2024.

[14] Y. Yang, J. Pan, Z. Peng, X. Du, Z. Tao, and J. Tang. Bistnet : Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8) :5612–5624, 2024.

[15] B. Zhang, M. He, J. Liao, P. V. Sander, L. Yuan, A. Bermak, and D. Chen. Deep Exemplar-Based Video Colorization. In *Proc. CVPR*, pages 8044–8053, 2019.

[16] R. Zhang, P. Isola, and A. A. Efros. Colorful Image Colorization. In *Proc. ECCV*, pages 649–666, 2016.