L'importance de la mesure de la qualité en matière d'imagerie médicale

Par Jean-Marie MOUREAUX
Professeur à l'Université de Lorraine et chercheur au Centre de recherche en automatique de Nancy (CRAN)

Ces dernières années, le développement des technologies numériques a connu un essor considérable dans le domaine de la santé. Dans ce contexte, les images et les vidéos médicales représentent des données essentielles dans les parcours de soins. Celles-ci sont aujourd'hui produites massivement et nécessitent des traitements pour pouvoir être transmises, archivées ou servir pour d'autres usages. Ces traitements algorithmiques pouvant causer des dégradations plus ou moins importantes, la mesure de la qualité représente un enjeu scientifique et économique majeur du fait de la sensibilité de ces données et de leurs applications en santé. Dans cet article, nous décrirons les concepts de *qualité « subjective »* et de *qualité « objective »* et nous décrirons certains des outils associés à ces concepts.

Introduction

Le développement des techniques d'imagerie numérique dans le monde médical a permis des progrès considérables, tant en matière de diagnostic que dans la prise en charge thérapeutique des patients. L'image est aujourd'hui partout dans l'hôpital, depuis le service de radiologie jusqu'au bloc opératoire, où elle guide le geste du chirurgien. La contrepartie de ces progrès liés à l'imagerie réside dans une quantité considérable de données qui contribue largement à alimenter les *Big data* à l'échelle planétaire.

En 2014, on estimait ainsi que les images médicales occupaient 30 % de la capacité mondiale d'archivage des images [1]. Cette quantité de données a donc un coût et représente un enjeu majeur de la médecine d'aujourd'hui et de demain, notamment en matière d'archivage [2]. Cependant, la multiplication des supports physiques et l'augmentation des capacités de stockage qu'elle entraîne, ainsi que le développement du « *cloud* » ne sont que des réponses partielles au problème ainsi posé. Mais il existe aussi des alternatives moins coûteuses parmi lesquelles figure la compression de données. Celle-ci est naturellement utilisée dans le milieu médical, dans sa version dite sans perte afin de préserver l'intégrité des données. Cependant, de multiples études scientifiques ont montré que même en matière d'imagerie radiologique, il existait une certaine tolérance aux pertes. Ainsi, des techniques de compression avec pertes peuvent être utilisées de façon maîtrisée, car elles offrent des taux de compression bien supérieurs à ceux que permettent les techniques sans perte. C'est sur la base de ces études que la Société canadienne de radiologie (CAR – *Canadian Association of Radiology*) et l'*American College of Radiology* ont émis tous deux des recommandations en matière d'usage de la compression avec perte pour l'archivage des données radiologiques [1]. La tolérance aux pertes (que nous avons évoquée plus haut) traduit le fait que même l'œil d'un expert avisé (tel qu'un radiologue) ne saurait tout déceler dans une image, et ce, du fait des limites du système visuel humain. Ces données sont exploitées en recourant aux techniques de la compression afin de minimiser les dégradations perceptibles. La principale difficulté réside dans la détermination du seuil de compression au-delà duquel la qualité perceptuelle des images n'est plus compatible avec une utilisation médicale. Dans cet article, nous allons nous attacher à définir les mesures qui permettent d'évaluer la qualité d'une image ou d'une vidéo dans un contexte médical. Après avoir rappelé la notion de qualité en matière d'images et de vidéos, nous en définirons la mesure subjective et la mesure objective, puis nous conclurons sur des considérations générales.

(1) https://blogs.msdn.microsoft.com/healthblog/2014/08/28/medical-image-archiving-in-the-cloud-consider-the-4-ss/
(2) En France, la durée légale d'archivage des images médicales est de 20 ans.

La notion de qualité en matière d'images et de vidéos

D'après l'encyclopédie en ligne Wikipedia, « *la qualité perceptuelle d'image est une mesure de la perception de la dégradation des images (souvent par comparaison à une image non dégradée dite de référence). Les systèmes de traitement des signaux introduisent souvent des artefacts (ou distorsions) dans le signal. Aussi, la mesure de la qualité est devenue importante…* »

La perception de la dégradation s'effectue ainsi souvent par comparaison avec une image ou une vidéo de référence (l'originale). On distingue ainsi la mesure de fidélité d'une image traitée, qui permet de calculer la distance entre l'image traitée et l'originale (la référence), de la mesure de sa qualité qui consiste à apprécier une image en tant que telle, en fonction de sa conscience, de la perception que l'on en a et des émotions qu'elle procure. Ce processus est souvent étudié au travers d'expériences psychophysiques. Notons que, par abus de langage, on emploie souvent le terme de « qualité » d'une image même lorsque l'on parle en réalité de sa fidélité. En médecine, la référence fait foi. C'est sans doute la raison pour laquelle la Société canadienne de radiologie (la CAR) préconise d'utiliser l'image compressée comme image de référence, plutôt que l'image native ! Le radiologue établit donc son diagnostic à partir d'une image compressée (suivant en cela les recommandations de la CAR).

Dans la suite de cet article, nous étudierons l'évaluation de la qualité d'une image ou d'une vidéo ayant subi un traitement (par exemple, une compression) par rapport à une référence.

L'évaluation subjective de la qualité

L'évaluation subjective de la qualité des images (ou des vidéos) dans le domaine médical [2, 3] dépend de la question à laquelle on souhaite répondre. Celle-ci peut être de deux natures différentes. Il peut s'agir :

- de détecter une pathologie ou une caractéristique importante dans une image ayant subi un traitement. Par exemple, peut-on encore détecter sur une image scanner compressée l'ensemble des nodules pulmonaires que l'on avait identifiés sur l'image native ?
- d'estimer une qualité globale compatible avec l'usage de l'image ou de la vidéo. Par exemple, cette vidéo endoscopique compressée permet-elle encore de guider correctement le chirurgien dans son geste ?

Dans les deux cas, les tests sont effectués par un panel représentatif des usagers (ici, les professionnels de santé). Celui-ci est généralement constitué au minimum de 3 experts (des « séniors » reconnus dans la spécialité médicale concernée) ou de 15 praticiens de la spécialité (incluant juniors et séniors). Le test doit inclure au minimum 30 images et se dérouler dans les mêmes conditions pour chaque panéliste, selon un protocole rigoureux à construire autour du type de pathologie, du type d'image/de vidéo, de la (des) question(s) posée(s), du nombre de patients, de leur diversité, etc.

Le cas de la détection

Dans le cas de la détection, la méthodologie le plus fréquemment utilisée est celle de l'analyse ROC (*Receiver Operating Characteristics*). Cette méthodologie s'appuie sur des indices pour analyser la fidélité diagnostique, en particulier :

- la sensibilité : $Se = VP / (VP + FN)$,
- la spécificité : $Sp = VN / (VN + FP)$,

où *VP, VN, FP et FN* désignent respectivement les Vrais Positifs (lésions réelles correctement détectées), les Vrais Négatifs (absence de lésion correctement détectée), les Faux Positifs (lésions détectées à tort) et les Faux Négatifs (lésions réelles non détectées).

La sensibilité désigne le pourcentage de vraies lésions détectées par rapport au nombre total de lésions réelles. La spécificité, quant à elle, désigne au contraire le pourcentage de « fausses » lésions identifiées par rapport au nombre total des fausses lésions. La méthodologie ROC nécessite au préalable la définition d'un standard doré (« *gold standard* ») qui établit une vérité « vraie » (la référence) à laquelle vont se confronter les panélistes. Une fois les tests effectués, on trace la courbe ROC qui représente la sensibilité *Se* en fonction du complément de la spécificité (1-*Sp*). Cette courbe est construite à partir du nuage de points fourni par les observateurs (les panélistes) [2]. Chaque point de la courbe fait alors apparaître un seuil de décision (un compromis entre sensibilité et spécificité). La courbe peut ensuite être interprétée de manière qualitative, en fonction de son allure, et permettre de répondre à la question initiale posée. Ainsi, par exemple, dans le cas de l'évaluation de l'impact d'un (ou de plusieurs) algorithme(s) de compression sur la détection des lésions, la courbe ROC permettra de conclure qu'un algorithme est moins impactant qu'un autre, ou encore qu'un algorithme n'impacte pas le compromis sensibilité/spécificité ou, au contraire, qu'il l'impacte fortement.

Cas de l'estimation globale de qualité

Pour l'évaluation de la qualité subjective des images, l'ITU (*International Telecommunications Union*) impose des normes très strictes aux conditions de l'observation. Ainsi, l'environnement dans lequel les panélistes effectuent les tests est normalisé (conditions d'éclairement, distance et angle d'observation…) et est strictement identique pour chacun d'eux. La norme ITU-BT.500-13 « Méthodologie d'évaluation subjective de la qualité des images de télévision » [4] formalise les conditions de l'observation et définit les méthodes générales d'essai, ainsi que les échelles de notation. Le choix entre les différentes méthodes et échelles dépend de l'application visée. Ainsi, les méthodes à stimulus unique permettent d'observer la réponse du panel à une donnée sans la comparer à une référence, contrairement aux méthodes à double stimulus. En effet, dans ces dernières, la donnée de référence est en permanence mise en regard de la donnée altérée (sans, bien entendu, que le panéliste sache laquelle est altérée et laquelle ne l'est pas).

Que ce soit en simple ou en double stimulus, l'observateur doit évaluer la qualité de l'image sur une échelle qui

comporte généralement cinq repères sémantiques (voir la Figure 1 ci-dessous). Cette échelle peut limiter l'évaluation au choix d'un adjectif parmi cinq ou être utilisée de façon continue, permettant alors à l'observateur de mettre une note (en cochant un point) parmi une infinité de possibilités entre 1 et 5. Dans le cas des méthodes à double stimulus, l'observateur devra attribuer une paire de notes (une pour chaque stimulus) sur deux échelles identiques, selon les principes énoncés plus haut.

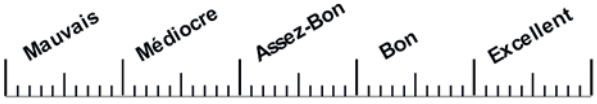

Figure 1 : Échelle de notation continue pour les tests subjectifs de qualité.

Comme le recommande la norme ITU-BT.500-13, un test de cohérence est effectué pour déceler d'éventuels observateurs incohérents. Ce test a pour but de normaliser la capacité d'un observateur à répondre de façon cohérente par rapport à l'ensemble du panel. S'il s'avère qu'un observateur répond systématiquement de façon éloignée par rapport au panel, il est alors rejeté. Une fois ces étapes franchies, une base de données reprenant les notations des observateurs jugés cohérents est constituée. Elle permet de définir un score MOS (*Mean Opinion Score*) représentant, pour chaque image ou vidéo, la moyenne des notes données par les observateurs. Ce score d'opinion moyen est l'unité de perception subjective de qualité obtenue pour un panel d'observateurs ayant réalisé un test strictement identique. Le MOS est considéré comme le score de perception de la qualité le plus fiable, il est donné par la formule :

$$\bar{u}_{jk} = \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} u_{ijk}$$

où N_{obs} représente le nombre d'observateurs et u_{ijk} la note que l'observateur i a attribuée à l'image ou à la vidéo médicale k. L'indice j représente la variable étudiée (par exemple, le taux de compression).

Selon la question initiale posée, les points du MOS sont placés sur un graphe en fonction de la variable étudiée dans la question. Pour reprendre l'exemple de la compression, on pourra ainsi tracer le MOS en fonction du taux de compression pour un algorithme donné et en déduire un seuil ou une zone d'acceptabilité de la compression avec perte pour un usage médical. Si l'on souhaite mettre en concurrence plusieurs algorithmes de compression, on tracera autant de courbes MOS fonctions du taux de compression que d'algorithmes et l'on en déduira quel est l'algorithme le plus compatible avec la pratique médicale pour un taux de compression donné (ou pour une gamme de taux de compression).

La mesure objective de la qualité

En raison de son coût élevé, la mise en œuvre de tests subjectifs n'est pas toujours réalisable. C'est l'une des raisons pour lesquelles les chercheurs se sont penchés depuis de nombreuses années sur le développement d'outils mathématiques permettant de mesurer la qualité. Ces outils sont généralement appelés « critères objectifs » ou « métriques objectives ». La qualité image/vidéo est un domaine qui occupe ainsi de nombreux chercheurs et l'on dénombre aujourd'hui dans la littérature plus d'une centaine de ces métriques [5] ! Celles-ci ont chacune leur spécificité et ont souvent été construites pour répondre à une question précise ou pour étudier un phénomène particulier (par exemple, la mesure du célèbre « effet de bloc » induit par le codeur JPEG). Certaines tendent à se rapprocher du système visuel humain (SVH), avec toute la difficulté que l'on peut imaginer du fait de la complexité de celui-ci.

De manière générale, outre l'intérêt lié à l'économie des tests subjectifs qu'elles permettent, les métriques objectives sont utilisées pour l'optimisation des algorithmes de traitement d'images/vidéos afin de minimiser les dégradations engendrées par leur traitement. Comme nous l'avons évoqué précédemment, il existe un très grand nombre de métriques liées à la qualité, ce qui conduit à différentes façons de les classer ou de les regrouper. Une façon usuelle consiste à distinguer :

- *les métriques avec référence,* qui utilisent l'intégralité de l'image (de la vidéo) originale pour effectuer les comparaisons nécessaires ;
- *les métriques sans référence,* qui sont basées uniquement sur l'image (sur la vidéo) dégradée (mais leur calcul par une machine est très difficile à réaliser) ;
- *les métriques avec référence réduite* basées sur l'image (la vidéo) dégradée, à laquelle on ajoute un minimum d'attributs de l'image (vidéo) originale.

Ces métriques sont d'autant plus pertinentes qu'elles sont corrélées au jugement humain. Ainsi, l'on représente souvent le MOS en fonction d'un critère objectif pour vérifier le degré de corrélation entre les deux outils.

À ce jour, le critère le plus utilisé dans le domaine de l'image et de la vidéo est sans contexte le rapport signal/bruit-pic (*Peak Signal to Noise Ratio* en anglais). Exprimé en décibel (dB), il correspond à la formule suivante :

$$PSNR = 10 log_{10} \frac{(2^b-1)^2}{\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}(x(m,n) - \hat{x}(m,n))^2} dB$$

où b est le nombre de bits/pixel de codage des images. Pour les images à niveaux de gris grand public, b vaut généralement 8, alors qu'il est souvent égal à 12 pour les images radiologiques au format DICOM (*Digital Imaging and COmmunications in Medicine*). Le dénominateur représente l'erreur quadratique moyenne entre l'image originale et l'image traitée : M et N sont respectivement le nombre de lignes et de colonnes de l'image, x(m,n) la valeur du pixel de coordonnées m et n dans l'image d'origine, et $\hat{x}(m,n)$, la valeur du même pixel dans l'image après traitement. Notons que la formule ci-dessus peut être étendue au cas des images couleur de plusieurs façons différentes. L'une d'entre elles consiste à calculer le PSNR dans chacun des trois plans colorimétriques, puis de faire la moyenne des trois valeurs obtenues.

Le principal avantage du PSNR est la simplicité de son calcul. L'inconvénient en est qu'il représente très mal le jugement humain et qu'il ne permet pas de localiser les dégradations dans l'image, ni leur impact sur la perception visuelle, comme l'illustrent les figures 2 et 3 qui montrent qu'à PSNR égal une image demeure exploitable (voir la Figure 3 ci-contre), alors que l'autre ne l'est pas (voir la Figure 2 ci-dessous). Le PSNR est donc aujourd'hui concurrencé par un grand nombre de métriques. Nous ne les présenterons pas ici (le lecteur intéressé pourra consulter les références suivantes [2, 3, 5, 6] – voir la Bibliographie en fin d'article). L'imagerie médicale peut s'appuyer sur l'ensemble de ces métriques, en particulier sur celles qui sont basées sur le SVH. On peut également ajouter des critères basés sur l'étude de paramètres cliniques (par exemple, le diamètre d'une tumeur ou la surface d'un organe) afin d'évaluer l'impact d'un traitement (la compression, par exemple) sur ces paramètres [2].

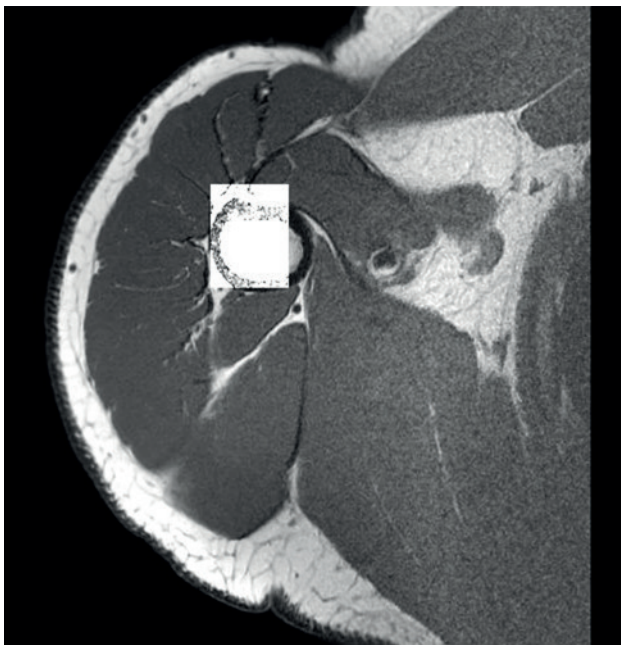

Figure 2 : IRM bruitée d'une épaule (facteur multiplicatif de 10 appliqué sur une zone rectangulaire), PSNR = 46,25 dB.

Conclusion

L'image et la vidéo sont aujourd'hui omniprésentes dans le traitement des patients, depuis le diagnostic jusqu'aux soins prodigués, incluant éventuellement une intervention chirurgicale. Les enjeux en termes d'archivage, de transmission et de traitement de données sont très importants. Les méthodes associées à ces activités engendrant des modifications de la qualité de ces images et de ces vidéos, l'étude de la qualité est essentielle pour garantir une fiabilité des données sur lesquelles le praticien puisse s'appuyer en toute confiance. Comme nous l'avons vu, cette qualité peut être évaluée à l'aide de tests subjectifs normalisés ou grâce à des métriques objectives. Dans tous les cas, dans le domaine de la santé, la qualité des images et des vidéos ne saurait avoir le sens vague que ce terme a généralement dans le grand public. Il s'agit d'une

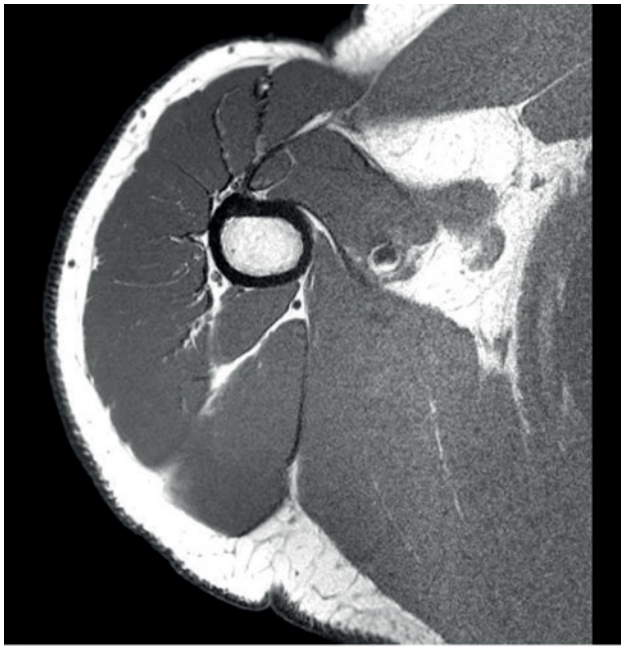

Figure 3 : IRM bruitée d'une épaule (facteur multiplicatif de 1,205 appliqué sur toute l'image), PSNR = 46,3 dB.

problématique médicale majeure : c'est en effet cette qualité des données qui permet d'apporter toutes les garanties en matière de justesse du diagnostic ou d'apport de soins adaptés.

Bibliographie

[1] Canadian Association of Radiologists, *CAR standards for irreversible compression in digital diagnostic within radiology*, http://www.car.ca/, juin 2011.

[2] NAIT ALI (A.) & CAVARO-MENARD (C.), « Compression des images et des signaux médicaux », *Information et sciences du vivant*, Éditions Hermès Lavoisier, 2007, ISBN 978-2-7462-1493-4.

[3] CHAABOUNI (A.), GAUDEAU (Y.), LAMBERT (J.), MOUREAUX (J.-M.) & GALLET (P.), *Subjective and Objective Quality Assessment for H264 Compressed Medical Video Sequences*, 4th IEEE International Conference on Image Processing, Theory, Tools and Applications, IPTA 2014, Paris, 14-17 octobre 2014.

[4] ITU-R. Recommendation 500-13, *Methodology for the subjective assessment of the quality of television pictures*, ITU-R Rec-BT.500, 2012.

[5] PEDERSEN (M.) & HARDEBERG (J. Y.), *Full-Reference Image Quality Metrics: Classification and Evaluation*, Foundations and Trends® in Computer Graphics and Vision, vol. 7, n°1, 2012, pp. 1-80. Doi :10.1561/0600000037.

[6] ALBANESI (M. G.) & AMADEO (R.), "A New Categorization of Image Quality Metrics Based On a Model of Human Quality Perception", *World Academy of Science, Engineering and Technology International Journal of Computer, Information, Systems and Control Engineering*, vol. 8, n°6, 2014.