Evaluation « objective » du photoréalisme d'images issues d'un simulateur de conduite ferroviaire

Julien MUZEAU Denis FAURE-VINCENT Patricia LADRET Alice CAPLIER
Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

Résumé – Les contenus numériques à base de modélisation 3D sont omniprésents de nos jours, que ce soit à travers les jeux vidéos, les films d'animation ou les applications industrielles. De plus, les récents développements matériels et logiciels donnent lieu à une approximation toujours plus précise du monde réel. Une problématique soulevée est celle du photoréalisme des images extraites de tels simulateurs 3D, en particulier de ceux de la conduite ferroviaire. Dans cet article, nous présentons une méthode à base de réseau de neurones convolutif capable d'associer un score de photoréalisme à une image donnée en entrée. Une deuxième contribution porte sur la validation perceptive de cette métrique. Pour ce faire, nous avons développé une expérience subjective permettant à des utilisateurs d'évaluer, parmi différentes paires d'images, laquelle est la plus photoréaliste. Nous concluons ainsi à une corrélation d'environ 90% entre le modèle proposé et la perception visuelle humaine.

Abstract – Digital content based on 3D modelling is ubiquitous nowadays, whether through video games, animated films or industrial applications. Moreover, recent hardware and software developments lead to an ever more accurate approximation of the real world. One issue raised is that of the photorealism of the images extracted from such 3D simulators, in particular those of train driving. In this article, we present a convolutional neural network-based method able to associate a photorealism score to any image given as input. A second contribution concerns the perceptual validation of this metric. To this end, we developed a subjective experiment allowing users to evaluate, among several pairs of images, which is the most photorealistic. We conclude to a 90% correlation between the proposed model and the human visual perception.

1 Introduction

La modélisation 3D intervient abondamment aujourd'hui grâce aux dernières avancées, et ce à différents niveaux : jeux vidéos, images de synthèse, reconstruction d'objets, *etc*. Certaines applications requièrent une réplique aussi précise que possible de la physique du monde réel afin d'aboutir à une expérience visuelle confortable. Nous nous intéressons plus précisément dans ce travail à la simulation de conduite ferroviaire.

La figure 1a montre deux images extraites d'un tel simulateur 3D, proposé par une entreprise grenobloise, sur des lignes de chemin de fer et sous des conditions météorologiques (nuageuse et pluvieuse) différentes : nous les qualifions de « synthétiques » dans la suite de cet article. Par opposition, des images réelles sont affichées en figure 1b. Elles sont issues de vidéos enregistrées via une caméra fixée à l'avant de trains en service. Nous considérons ces images comme photoréalistes malgré les différents défauts qui peuvent survenir, en lien avec la lentille de la caméra par exemple.

En comparant ces deux types d'images, une question qui vient est celle de la qualité de reconstruction du monde réel, et plus particulièrement du photoréalisme des images 2D extraites de cette représentation. La première contribution de cet article, détaillée en section 2, est la proposition d'une métrique permettant d'associer un score de photoréalisme à une image. Dans un deuxième temps, une expérience d'évaluation subjective est développée dans le but de valider l'indicateur proposé, c'est-à-dire de vérifier s'il est en adéquation avec la perception visuelle humaine (*cf.* section 3). Enfin, des conclusions et perspectives sont données en section 4.

(a) Images extraites du simulateur. (b) Images réelles.

FIGURE 1 – Exemples d'images synthétiques et réelles de conduite ferroviaire.

2 Métrique de photoréalisme

Suite à un bref état de l'art, nous présentons dans cette section un indicateur permettant d'évaluer le degré de photoréalisme d'une image.

2.1 Etat de l'art

Nous nous appuyons dans cet article sur les nombreux travaux récents visant à distinguer les images générées par ordinateur (*Computer-Generated* en anglais ou CG) des PhotoGraphies (PG), domaine crucial notamment pour la détection d'une quelconque falsification au sein de contenus numériques abondants de nos jours. Les approches modernes se basent majoritairement sur la puissance de l'apprentissage profond. En effet, il a

été montré dans [8] que, contrairement aux méthodes traditionnelles, les caractéristiques profondes extraites de réseaux neuronaux fournissent une représentation d'une image fortement corrélée à celle donnée par la perception visuelle humaine. Par exemple, [2] et [7] utilisent des réseaux de neurones convolutifs classiques auxquels est ajouté une dernière couche de classification binaire dans le but de dinstiguer les images CG et PG. Dans [6], l'information globale d'une image est combinée à celle locale, c'est-à-dire au niveau de ses différents patchs constitutifs, afin d'améliorer les performances de l'état de l'art. [5] recommande l'utilisation de filtres passe-hauts dans le but de se concentrer sur le critère discriminant qu'est le bruit ambiant plutôt que sur le contenu des basses fréquences. Quant à [9], les auteurs s'attardent sur la corrélation entre les canaux de couleur ainsi que celle des pixels voisins. Enfin, [4] propose d'associer les traitements d'une même image sous différents espaces de couleurs (RGB, LCH, HSV) grâce à l'apprentissage ensembliste.

2.2 Modèle proposé

Les approches décrites précédemment, bien que performantes, ne s'attachent qu'à la distinction binaire entre les images générées par ordinateur et les photographies, sans réellement s'attarder sur le degré d'appartenance à telle ou telle classe. De plus, l'aspect photoréaliste ou non d'une image n'est pas abordé.

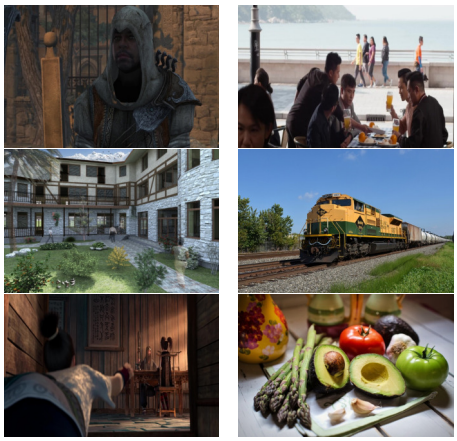
Nous nous inspirons donc de ces méthodes en émettant l'hypothèse que les images PG sont photoréalistes, contrairement aux images CG, qui elles sont supposées synthétiques. Le modèle proposé est un réseau de neurones convolutif suivant l'architecture ResNet50 [3] pour des raisons de performances. Ce dernier est pré-entraîné sur la base de données ImageNet[1]. La dernière couche de classification est supprimée au profit d'un seul neurone de régression logistique. La première partie du réseau est figée et est donc considérée comme un simple extracteur de caractéristiques profondes, seule la deuxième est affectée par l'entraînement. Nous sommes donc en présence d'une classification à deux catégories : image photoréaliste ($Y = \text{PR}$) ou non-photoréaliste ($Y = \text{NPR}$). Suite à l'apprentissage, en phase de test, l'indice de photoréalisme $\mathcal{S}$ d'une image X est déterminé selon la probabilité *a posteriori* suivante :

$$\mathcal{S}(X) = \mathbb{P}(Y = \text{PR}|X). \quad (1)$$

Une métrique proche de 1 correspond à une image photoréaliste, alors qu'une valeur tendant vers 0 est associée à une image synthétique ou générée par ordinateur. L'équation 1 permet de retrouver la classe binaire prédite pour l'image X par seuillage : un score supérieur (resp. inférieur) à 0,5 correspond à une image photoréaliste (resp. non-photoréaliste). Cette valeur de seuil est retrouvée empiriquement (*cf.* figure 3).

2.3 Données d'apprentissage

En ce qui concerne les données d'entraînement, nous utilisons la base de données *Large-Scale CG images Benchmark* (LSCGB) [1]. Cette dernière est constituée d'environ 140000 images réparties équitablement en deux catégories : synthétiques (CG) et photographiques (PG). Des exemples pris dans ce jeu de données sont affichés en figure 2. Les images CG sont extraites de jeux vidéos, films d'animation, modélisations 3D ou même de modèles génératifs. Quant aux photographies, elles proviennent de films, d'autres bases d'images publiques ou sont le fruit de différentes recherches Internet.



(a) Images CG. (b) Images PG.

FIGURE 2 – Exemples d'images issues de la base de données LSCGB.

A la base de données LSCGB sont ajoutées environ 5000 images (soit 3,3% de l'ensemble de la base) de conduite de trains synthétiques et réelles, telles que celles présentées en figure 1, afin d'inclure de l'information liée à l'application pendant l'entraînement.

Le jeu de données complet est divisé en trois sous-ensembles : 70% pour la phase d'entraînement, 10% en validation et 20% en test.

2.4 Performances

L'évaluation du modèle proposé précédemment se fait via la précision (*accuracy* en anglais), c'est-à-dire la proportion de bonnes décisions par rapport au nombre total de prédictions réalisées. Cet indicateur est en effet suffisant du fait que les deux classes $Y = \text{PR}$ et $Y = \text{NPR}$ sont équilibrées. Nous obtenons une précision de 90% sur l'ensemble d'entraînement, 90,3% en validation et 79,6% en phase de test. Cependant, si l'on ne tient uniquement compte que des images de conduite de trains, une précision de 94,9% est atteinte.

Nous donnons en figure 3 la distribution des scores de photoréalisme associés aux images ferroviaires, synthétiques et réelles, obtenus par la méthode proposée. On peut constater que la quasi-totalité des indicateurs pour les images CG sont inférieurs à 0,5, avec un pic proche de 0. Toutefois, en ce qui concerne les images réelles, malgré un mode autour de 1 et une majorité de scores supérieurs à 0,5, l'histogramme est plus éparse entre 0 et 1. On peut notamment expliquer cette différence par la grande variabilité dans les images naturelles par rapport aux images générées par ordinateur selon le même modèle graphique.

Quatre illustrations sont données en figure 4. En figure 4a est affichée une image CG obtenant un score de photoréalisme d'environ 0,3. Même si certains éléments, tels que la linéarité des caténaires, des rails ou des glissières de sécurité, rendent le contenu global plutôt synthétique, le ciel nuageux ou la végétation apportent une touche de photoréalisme. En revanche,

[1] https://www.image-net.org/

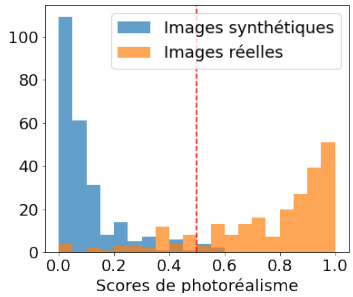
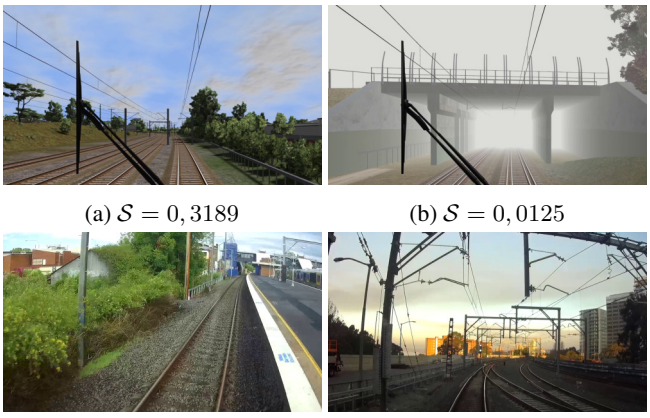


FIGURE 3 – Histogrammes des scores de photoréalisme obtenus sur les images de test de conduite ferroviaire. La séparation à 0,5 entre les deux classes est représentée par une ligne verticale rouge.

l'image de la figure 4b est évaluée comme étant fortement synthétique, notamment à cause de l'effet de brouillard. Pour ce qui est des images réelles, la métrique de photoréalisme proposée dans cet article attribue un score proche de 1 pour l'image 4c. Toutefois, on peut noter un indicateur inférieur à 0,5 en ce qui concerne l'image réelle de la figure 4d : ceci s'explique principalement par le faible contraste d'une part, par la saturation et la teinte jaune à l'horizon d'autre part.



(a) $\mathcal{S} = 0,3189$ (b) $\mathcal{S} = 0,0125$

(c) $\mathcal{S} = 0,9172$ (d) $\mathcal{S} = 0,3860$

FIGURE 4 – Exemples de la métrique de photoréalisme $\mathcal{S}$ appliquée à des images de conduite ferroviaire. Les figures (a) et (b) sont issues d'un simulateur 3D, alors que les figures (c) et (d) sont réelles.

3 Evaluation subjective du photoréalisme

L'objectif de cette section est la validation du score de photoréalisme proposé précédemment. Est-il corrélé à la perception visuelle humaine ? Pour ce faire, nous avons développé un outil en ligne demandant à des utilisateurs de choisir l'image la plus photoréaliste parmi deux proposées, et ce sur un grand nombre d'images à différents niveaux de réalisme.

3.1 Protocole

La base de données pour cette expérience d'évaluation subjective est constituée de $P = 400$ paires d'images au total. Parmi celles-ci, l'utilisateur doit en annoter $N = 200$, correspondant à une durée d'expérimentation d'environ 10 minutes. Avec cette configuration, un minimum de $K = 75$ annotateurs est nécessaire afin d'assurer une précision significative dans les résultats finaux et dans l'interprétation qui en est faite. Les différentes paires sont définies au préalable et restent inchangées pour l'ensemble des utilisateurs. En revanche, l'ordre d'affichage est aléatoire pour chaque annotateur, de même que l'ordre gauche/droite.

En ce qui concerne les images impliquées dans cette application, quatre types sont utilisés de manière équilibrée :

1. des images synthétiques extraites d'une première version du simulateur 3D de conduite de trains (figure 1a).
2. des images issues d'une deuxième version du simulateur 3D mettant en oeuvre un moteur graphique plus récent.
3. des images réelles telles que celles de la figure 1b.
4. des images générées par CycleGAN [10] dans le but d'améliorer le photoréalisme des images simulées.

Chaque paire est composée de deux types d'images différents tirés aléatoirement. De cette manière, une plus grande variabilité est permise et un possible phénomène d'adaptation du côté de l'utilisateur est évité.

Nous sommes donc en présence d'une expérimentation dite *subjective*, par implication de la perception humaine, ainsi que *relative* du fait de la comparaison entre images. Une capture d'écran de cette dernière est donnée en figure 5.

FIGURE 5 – Capture d'écran de l'expérience subjective.

3.2 Résultats

Pour chaque paire, la méthode proposée en section précédente permet de conclure quant à l'image qui obtient le score de photoréalisme le plus élevé. On peut procéder de la même manière grâce à l'expérience d'évaluation subjective : l'image qui recueille le plus de suffrages de la part des annotateurs est considérée comme la plus photoréaliste. Il est ainsi possible de comparer les deux approches par la détermination d'une matrice de confusion, donnée en table 1.

Les résultats colonne par colonne se rapportent à la métrique présentée précédemment (M), les lignes à l'expérimentation subjective (ES). Les signes « < » et « > » se réfèrent à la comparaison du photoréalisme. Enfin, nous dénommons « img **g**auche » et « img **d**roite » les deux images d'une même paire. Par exemple, le chiffre supérieur droit du tableau représente le nombre de paires, ici 20, pour lesquelles l'image de gauche est évaluée algorithmiquement plus photoréaliste que celle de droite, alors que les annotateurs humains jugent de l'inverse

en moyenne. La somme des différentes valeurs est égale à 400, soit le nombre total P de paires de la base de données.

TABLE 1 – Matrice de confusion de l'évaluation du photoréalisme par la métrique (M) et l'expérience subjective (ES).

ES \ M	img g < img d	img g > img d
img g < img d	191	20
img g > img d	28	161

Le pourcentage des éléments diagonaux s'élève à $(191 + 161)/400 = 0,88$. Autrement dit, la métrique de photoréalisme est corrélée à 88% à la perception visuelle humaine sur les images présentées lors de l'expérimentation.

La figure 6 montre un exemple de désaccord entre les deux approches. En effet, l'image réelle (6a) est considérée à l'unanimité des annotateurs humains comme plus photoréaliste que l'image synthétique (6b). En revanche, la métrique associe un score de réalisme égal à 0,08 pour la première, 0,15 pour l'image issue du simulateur 3D. Ceci peut notamment s'expliquer par les défauts sur l'image réelle, tels que la saturation et les stries verticales qui en découlent.

(a) Image réelle, $\mathcal{S} = 0,08$. (b) Image synthétique, $\mathcal{S} = 0,15$.

FIGURE 6 – Exemple d'erreur d'évaluation du photoréalisme par la métrique par rapport aux annotateurs humains.

Enfin, notons que l'ordonnancement des différents types d'images est conservé. En effet, la métrique de même que l'expérience subjective permettent de conclure que les images réelles sont les plus photoréalistes, puis les images créées par un moteur 3D récent, celles générées par CycleGAN, enfin les images synthétiques.

4 Conclusion et perspectives

Nous avons proposé dans cet article une méthode à base de réseau neuronal permettant d'évaluer automatiquement le degré de photoréalisme d'une image issue d'un simulateur 3D de conduite de trains, avec une précision de classification binaire CG/PG à hauteur de 95%. Dans le but de valider cette métrique, nous avons également développé une expérience d'évaluation subjective pendant laquelle les utilisateurs ont à comparer différentes paires d'images et à juger laquelle des deux est la plus photoréaliste. Cette expérimentation permet de conclure à une corrélation entre notre méthode et la perception visuelle humaine de 88%.

En termes de perspectives, une première consiste à approfondir la recherche d'un modèle donnant lieu à de meilleures performances. Ce dernier permettrait de pouvoir généraliser l'idée de métrique photoréalisme à d'autres domaines d'application. Deuxièmement, une idée serait de mettre en place une expérience subjective dite *absolue*, contrairement à celle présentée en section 3. De cette manière, du fait de la cohérence entre les résultats donnés par la métrique photoréalisme et ceux des annotateurs humains, la validation de la méthode proposée serait facilitée.

Références

[1] Weiming BAI, Zhipeng ZHANG, Bing LI, Pei WANG, Yangxi LI, Congxuan ZHANG et Weiming HU : Robust texture-aware computer-generated image forensic : Benchmark and algorithm. *Transactions on Image Processing*, 30:8439–8453, octobre 2021.

[2] Edmar R. S. de REZENDE, Guilherme C. S. RUPPORT, Antonio THEOPHILO et Tiago CARVALHO : Exposing computer generated images by using deep convolutional neural networks. *Signal Processing : Image Communication*, 66:113–126, août 2018.

[3] Kaiming HE, Xiangyu ZHANG, Shaoqing REN et Jian SUN : Deep residual learning for image recognition. *In CVPR*, pages 770–778, juin 2016.

[4] P. Gangan MANJARY, K. ANOOP et Lajish V. L. : Distinguishing natural and computer-generated images using multi-colorspace fused efficientnet. *arXiv*, pages 1–13, octobre 2021.

[5] Kunj Bihari MEENA et Vipin TYAGI : Distinguishing computer-generated images from photographic images using two-stream convolutional neural network. *Applied Soft Computing*, 100(4):1–10, mars 2021.

[6] Weize QUAN, Kai WANG, Dong-Ming YAN et Xiaopeng ZHANG : Distinguishing between natural and computer-generated images using convolutional neural networks. *Transactions on Information Forensics and Security*, 13(11):2772–2787, mai 2018.

[7] Nicolas RAHMOUNI, Vincent NOZICK, Junichi YAMAGISHI et Isao ECHIZEN : Distinguishing computer graphics from natural images using convolution neural networks. *In Workshop on Information Forensics and Security*, pages 1–6, décembre 2017.

[8] Richard ZHANG, Phillip ISOLA, Alexei A. EFROS, Eli SHECHTMAN et Oliver WANG : The unreasonable effectiveness of deep features as a perceptual metric. *In CVPR*, pages 586–595, juin 2018.

[9] Rui-Song ZHANG, Wei-Ze QUAN, Lu-Bin FAN, Li-Ming HU et Dong-Ming YAN : Distinguishing computer-generated images from natural images using channel and pixel correlation. *Journal of Computer Science and Technology*, 35(3):592–602, mai 2020.

[10] Jun-Yan ZHU, Taesung PARK, Philip ISOLA et Alexei A. EFROS : Unpaired image-to-image translation using cycle-consistent adversarial networks. *In ICCV*, pages 2223–2232, octobre 2017.