Apport des Descripteurs Visuels à la Détection d'Images Générées par IA

Abderrezzaq SENDJASNI Mohamed-Chaker LARABI
CNRS, Univ. Poitiers, XLIM, UMR 7252, France

Résumé – Cette étude examine le défi croissant de distinguer les images naturelles des images générées par l'intelligence artificielle (IA). En explorant trois descripteurs visuels distincts, MSCN (statistiques des scènes naturelles), CLIP (embeddings sémantiques) et M-LBP (caractérisation de textures), nous évaluons leur efficacité individuelle et combinée dans la classification d'images. Les résultats montrent que l'intégration de ces descripteur améliore les performances. Cette recherche souligne l'importance de combiner des caractéristiques complémentaires pour améliorer la distinction entre images naturelles et synthétiques, fournissant une base pour le développement futur d'outils de détection des médias synthétiques.

Abstract – This study addresses the increasingly complex challenge of distinguishing between natural and AI-generated (GenAI) images produced by diverse generative models. By leveraging the complementary strengths of three distinct feature spaces, including MSCN (spatial and statistical features), CLIP (high-level semantic embeddings), and M-LBP (texture-based features), we evaluate their individual and combined efficacy in image classification. Our findings reveal that while each feature space contributes unique insights, their integration achieves superior classification performance. This work highlights the critical role of combining complementary feature spaces to enhance the robustness and accuracy of distinguishing natural from synthetic images. Our results advance the current understanding of GenAI content analysis and provide a foundation for future research in developing more effective tools for detecting GenAI media in an era of rapidly evolving generative technologies.

1 Introduction

La prolifération des modèles d'intelligence artificielle générative (IA générative) a révolutionné la création de contenu, notamment la synthèse d'images, la génération de vidéos et l'édition de médias. Des modèles avancés comme StyleGAN [7], DALL·E [12], et Stable Diffusion [14] produisent des contenus visuels extrêmement réalistes, remettant en question les notions traditionnelles d'authenticité. Cependant, évaluer la naturalité et la qualité perceptuelle des médias générés par l'IA reste un défi majeur. La naturalité se réfère ici à l'alignement des médias synthétiques avec les attentes perceptuelles humaines et leur ressemblance avec le contenu réel.

Les avancées récentes en synthèse d'images ont soulevé des préoccupations quant à la distinction entre images naturelles et générées par l'IA. Les premières méthodes de détection [9, 8] s'appuyaient sur les statistiques des scènes naturelles (NSS), utilisées dans des modèles comme BRISQUE [10]. Ces méthodes analysent les régularités statistiques des images naturelles pour prédire la qualité perceptuelle, mais peinent face au contenu complexe et convaincant généré par les modèles avancés d'IA générative. Les méthodes d'intégration sémantique, telles que CLIP [11], alignent les informations visuelles et textuelles pour une compréhension contextuelle des images, démontrant une forte performance dans la capture des différences sémantiques entre images naturelles et synthétiques. Parallèlement, l'analyse des artefacts de fréquence [1, 13] révèle des anomalies subtiles dans la texture et la cohérence des images synthétiques. Les descripteurs orientés texture [9, 5, 6], dérivés de motifs locaux ou d'analyses de fréquence, capturent les incohérences micro-structurales et les motifs répétitifs caractéristiques des médias générés par l'IA. Ces méthodes sont efficaces pour détecter des artefacts tels que l'éclairage non naturel ou les textures répétées.

Cette étude évalue la robustesse des descripteurs visuels NSS, sémantiques et orientés texture. Les caractéristiques NSS, modélisées par le descripteur MSCN, capturent les régularités statistiques des scènes naturelles. Les caractéristiques sémantiques, modélisées par CLIP, fournissent des informations contextuelles, tandis que les caractéristiques orientées texture, dérivées de l'analyse M-LBP, détectent les anomalies de surface. En comparant ces descripteurs, cette étude vise à identifier l'approche la plus efficace pour capturer les divergences perceptuelles entre images naturelles et générées par l'IA. Les résultats contribueront au développement de métriques d'évaluation plus robustes pour les modèles génératifs.

2 Méthode Proposée

L'étude proposée vise à relever le défi de la distinction entre images naturelles et images générées par l'IA (GenAI) en évaluant l'efficacité des descripteurs MSCN, CLIP embeddings et M-LBP. Chaque type offre une perspective unique pour analyser le contenu et la qualité des images, fournissant des informations diverses mais complémentaires. Cette section décrit la méthodologie adoptée pour exploiter et évaluer les performance de ces descripteurs visuels dans la détection de contenu GenAI.

La Fig. 1 illustre le schéma proposé, qui comprend trois étapes clés : le codage des caractéristiques, la normalisation et la détection. Ce cadre traite à la fois les images naturelles et GenAI pour générer des représentations de caractéristiques optimisées pour la détection. Plus précisément, chaque image subit un codage des caractéristiques pour produire un vecteur de caractéristiques, tel que défini par la fonction générale de

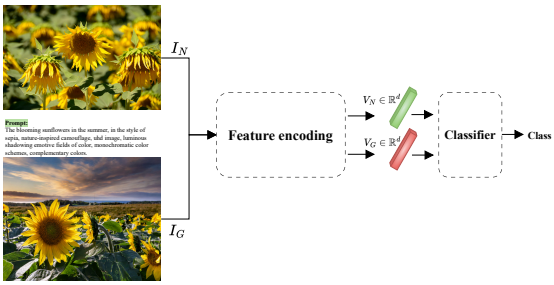

FIGURE 1 : Schéma de classification d'images.

génération de caractéristiques :

$$\Psi_n^{\text{Descripteur}} : I \in \mathbb{R}^{x,y,3} \rightarrow \mathcal{V} \in \mathbb{R}^d, \tag{1}$$

où I représente l'image d'entrée, d désigne la dimension des caractéristiques générées, et $\mathcal{V}$ est le vecteur de caractéristiques résultant. Le terme $Descripteur$ spécifie le descripteur de caractéristiques utilisé, permettant la flexibilité d'explorer divers descripteurs adaptés à la tâche en cours.

2.1 Génération des descripteurs visuels

MSCN : Le descripteur MSCN (Mean Subtracted Contrast Normalized) est conçu pour extraire des caractéristiques statistiques de second ordre à partir des statistiques d'image en utilisant la matrice de co-occurrence des niveaux de gris (GLCM). L'image d'entrée $I \in \mathbb{R}^{x,y,3}$ est d'abord convertie en niveaux de gris, réduisant sa dimension à $I_{\text{gray}} \in \mathbb{R}^{x,y}$. La GLCM est ensuite construite pour analyser les relations spatiales entre les intensités des pixels à des distances $d \in \{1, 2, 3\}$ et des angles $\theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$, permettant de capturer des statistiques de scène sensibles à l'échelle et à la direction. Six propriétés statistiques sont dérivées de chaque paire distance-angle : contraste, dissimilarité, homogénéité, moment angulaire secondaire (ASM), énergie et corrélation. Ces caractéristiques décrivent divers aspects statistiques tels que les variations d'intensité, l'uniformité et la dépendance des pixels. Les caractéristiques extraites sont ensuite aplaties et concaténées en un vecteur unique, $\mathbf{f}_{\text{MSCN}} \in \mathbb{R}^d$, où $d = 72$ représente le nombre total de caractéristiques.

La Fig. 2 présente les cartes de contraste d'une image naturelle et de deux versions générées. Dans l'image naturelle, la carte de contraste GLCM capture des transitions douces et graduelles autour des fleurs, reflétant les variations naturelles d'ombrage. En revanche, les images GenAI présentent des contrastes relativement nets, soulignant la délimitation artificielle des fleurs et contribuant à leur apparence synthétique.

Embeddings de CLIP : Les embeddings basés sur CLIP génèrent des représentations de caractéristiques de haute dimension des images, en se concentrant sur leur contenu sémantique et leurs relations contextuelles. Pour calculer ces plongements, l'image d'entrée $I \in \mathbb{R}^{x,y,3}$ est traitée par l'architecture Vision Transformer (ViT) du modèle CLIP. Le ViT divise l'image en patchs, applique des mécanismes d'auto-attention pour encoder chaque patch, et agrège les informations pour produire un vecteur de caractéristiques global. Le vecteur de caractéristiques résultant, $\mathbf{f}_{\text{CLIP}} \in \mathbb{R}^d$, où $d = 512$ représente la dimension des embeddings, capture les propriétés sémantiques globales, telles que les objets, les scènes et les interactions contextuelles. Ce processus peut être exprimé comme $\Psi_{\text{CLIP}} : I \rightarrow \mathbf{f}_{\text{CLIP}}$.

Les embeddings CLIP sont particulièrement efficaces pour des tâches telles que la distinction entre images naturelles et GenAI, les recherches de similarité sémantique et l'alignement croisé image-texte. En exploitant un espace latent partagé pour les informations visuelles et textuelles, CLIP offre des représentations sémantiques robustes, le rendant bien adapté à la détection d'incohérences subtiles.

Descripteur M-LBP : Le descripteur Multi-scale Local Binary Pattern (M-LBP) extrait des informations texturales et structurelles en analysant les motifs d'intensité locaux à plusieurs échelles. L'image d'entrée $I \in \mathbb{R}^{x,y,3}$ est d'abord convertie en niveaux de gris $I_{\text{gray}} \in \mathbb{R}^{x,y}$, garantissant que l'analyse de texture repose uniquement sur les valeurs d'intensité. L'opérateur Local Binary Pattern (LBP) est ensuite appliqué à chaque pixel en comparant son intensité à celle de ses voisins dans un rayon défini. La Fig. 2 illustre la différence de complexité texturale entre les images naturelles et celles générées par GenAI. La carte LBP de l'image naturelle met en évidence des textures douces et naturellement variées dans différentes régions. En revanche, les images générées tendent à présenter des motifs plus répétitifs et une variation texturale moins organique, indiquant une régularité synthétique.

Dans le M-LBP, l'analyse est étendue à plusieurs échelles en faisant varier le rayon $r \in \{1, 2, 3\}$, permettant au descripteur de capturer les détails de texture à différents niveaux de granularité. Pour chaque rayon, l'opérateur LBP génère des motifs binaires en seuillant les intensités des pixels voisins par rapport au pixel central. Ces motifs binaires sont convertis en valeurs décimales et compilés en histogrammes, qui résument les distributions de texture locales. Les caractéristiques de toutes les échelles sont concaténées en un vecteur de caractéristiques unique $\mathbf{f}_{\text{M-LBP}} \in \mathbb{R}^d$, où $d = 36$ représente la dimension totale des caractéristiques à travers toutes les échelles. Le processus peut être représenté par $\Psi_{\text{M-LBP}} : I \rightarrow \mathbf{f}_{\text{M-LBP}}$, soulignant sa capacité à encoder des informations texturales multi-échelles pour une analyse robuste des textures.

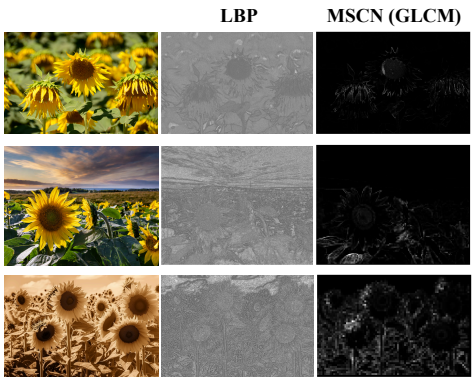

FIGURE 2 : Distribution des textures LBP et du contraste GLCM pour une image naturelle (1ère), image générée par Firefly (2ème) et image générée par Midjourney (3ème).

2.2 Classification

Une fois les vecteurs de caractéristiques générés pour chaque image à l'aide des trois descripteurs visuels (MSCN, CLIP et

M-LBP), l'étape suivante consiste en une normalisation pour garantir une échelle comparable entre ces descripteurs. Nous utilisons la normalisation par *z-score* [2] pour assurer que chaque caractéristique ait une moyenne de 0 et un écart-type de 1.

Après normalisation, l'étape suivante est la classification. Pour classer les images comme naturelles ou GenAI, nous utilisons XGBoost [3], un algorithme de gradient boosting qui s'est avéré très efficace pour les tâches de classification. De plus, XGBoost est bien adapté aux caractéristiques de haute dimension, notamment dans les cas où les relations entre les caractéristiques et les étiquettes cibles sont complexes et non linéaires. Cette tâche est représentée par :

$$y = \hat{f}(\mathbf{f}_{\text{Descripteur}}) \tag{2}$$

où y représente l'étiquette de classe prédite (naturelle ou GenAI), $\mathbf{x}$ est le vecteur de caractéristiques normalisé, et $\hat{f}$ est la fonction apprise représentant le modèle entraîné.

3 Expérience et Discussion

Cette section présente la configuration expérimentale, les résultats et les discussions de l'étude. Les expériences évaluent la capacité de divers types de caractéristiques à distinguer les images naturelles des images GenAI, en utilisant des métriques telles que la précision, le coefficient de corrélation de Matthews (MCC) et l'AUC-ROC. Une validation croisée k-fold est utilisée pour garantir des résultats robustes.

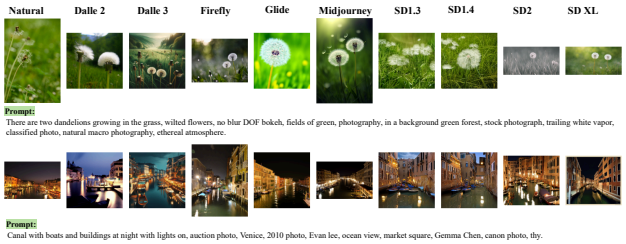

FIGURE 3 : Échantillons de de la base de données, incluant des images naturelles et leurs contreparties générées par l'IA.

Base de Données : Nous utilisons la base de données de Bammey *et al.* [1], conçu pour évaluer les méthodes de détection sur des images générées par des modèles d'IA générative modernes, tels que Stable Diffusion et DALL·E. Les images synthétiques sont générées à partir de prompts textuels inspirés de l'ensemble de données RAISE-1k [4], couvrant diverses catégories. Le classifieur est entraîné avec une division 80/20 pour l'entraînement et les tests.

La Fig. 3 montre des échantillons de la base de données utilisée, illustrant le réalisme des images générées par les modèles d'IA. Les images naturelles servent de référence, soulignant les progrès rapides de l'IA générative et le défi croissant de la classification.

3.1 Comparaison de performances

La comparaison des performances, telle que rapportée dans le Tableau 1, met en évidence la manière dont chaque descripteur capture des aspects distincts en fonction du modèle génératif.

MSCN : L'utilisation du descripteur MSCN montre son efficacité à capturer les statistiques globales et les caractéristiques basées sur le contraste pour distinguer les images naturelles des images GenAI. MSCN offre des performances robustes à travers divers modèles génératifs, avec des variations notables. Les performances sur la série Stable Diffusion se distinguent, avec des scores de précision de 96,74% et 96,39%, des valeurs MCC de 0,94 et 0,93, et des scores ROC-AUC de 0,99 pour SD 1.3 et 1.4 respectivement. Ces résultats indiquent la capacité de MSCN à capturer des variations subtiles de contraste et de luminance, menant à une haute performance de classification.

CLIP : L'utilisation des embeddinsg de CLIP montrent une large gamme de résultats, avec des différences notables dans la capacité à distinguer les modèles d'IA générative. CLIP excelle avec DALL-E 3, atteignant une précision de 99,70%, un MCC de 0,99, et un ROC-AUC de 0,99, démontrant sa capacité à capturer des motifs sémantiques et visuels distincts. Avec Firefly et Glide, CLIP montre également de fortes performances, avec des précisions de 92,35% et 96,40% respectivement. Cependant, CLIP rencontre des défis avec des modèles comme DALL-E 2, où le contenu généré et plus réaliste.

M-LBP : L'analyse de M-LBP révèle des résultats perspicaces, avec des modèles comme DALL-E 2 atteignant une précision de 95,35%, un MCC de 0,91, et un ROC-AUC de 0,99. Cependant, avec DALL-E 3, M-LBP montre des scores plus faibles, indiquant des difficultés avec des techniques de génération plus nuancées. Avec Firefly et Glide, des performances modérées sont obtenues, tandis qu'avec les modèles Stable Diffusion, M-LBP montre des améliorations progressives, avec des précisions allant de 84,64% à 92,74%.

En résumé, MSCN, CLIP et M-LBP montrent des forces et des faiblesses distinctes. MSCN excelle dans la capture des variations globales de contraste, CLIP dans les motifs sémantiques, et M-LBP dans les textures locales. Intégrer ces descripteurs visuels pourrait améliorer la classification, notamment pour des modèles génératifs divers et subtils.

3.2 Evaluation de la généralisation

Pour évaluer la précision et la capacité de généralisation de chaque type de descripteur dans la classification des images naturelles par rapport aux images GenAI, nous avons utilisé une stratégie de création d'ensembles de données où les images GenAI sont sélectionnées aléatoirement parmi l'ensemble des modèles de génération disponibles, afin d'assurer une représentation variée. Les performances sont présentées dans la Fig. 4 ou les barres représente la moyenne sur 10 itérations. Les résultats montrent que les embeddings de CLIP surpasse les autres, avec une précision de 89,5%, un MCC de 0,790 et un ROC-AUC de 0,962, soulignant l'efficacité des caractéristiques sémantiques élevées pour capturer les aspects contextuels des images.

Avec les caractéristiques MSCN, solides performances sont également obtenus, avec une précision de 88,0%, mais restent inférieures à CLIP, indiquant que les caractéristiques NSS sont robustes pour identifier les écarts statistiques mais peuvent manquer de capacité à distinguer les contenus génératifs très réalistes. Le descripteurs M-LBP a montré des performances inférieures, avec une précision de 83,4%, suggérant que les caractéristiques orientées texture manquent d'informations contextuelles globales nécessaires pour différencier les images

TABLE 1 : Performance du classifieur entraîné avec les différents descripteurs viuelles, en termes de Précision, MCC et ROC-AUC, respectivement. Les meilleures performances par type de caractéristique sont sont mises en **rouge**.

Model	MSCN			CLIP			M-LBP		
	Accuracy	MCC	ROC-AUC	Accuracy	MCC	ROC-AUC	Accuracy	MCC	ROC-AUC
DALL-E 2	0.933	0.866	0.982	0.831	0.663	0.919	**0.953**	**0.907**	**0.991**
DALL-E 3	0.958	0.917	0.993	**0.997**	**0.994**	**0.997**	0.763	0.527	0.840
Firefly	0.937	0.875	0.986	0.923	0.848	0.977	0.819	0.638	0.904
GLIDE	0.906	0.813	0.967	0.964	0.928	0.995	0.889	0.779	0.958
Midjourney-V5	0.920	0.843	0.967	0.915	0.831	0.974	0.909	0.818	0.967
SD 1.3	**0.967**	**0.935**	**0.992**	0.923	0.847	0.981	0.846	0.693	0.920
SD 1.4	0.964	0.928	0.992	0.920	0.841	0.980	0.866	0.733	0.925
SD 2.0	0.935	0.870	0.982	0.933	0.867	0.986	0.864	0.728	0.927
SD XL	0.926	0.853	0.980	0.939	0.877	0.985	0.927	0.855	0.976

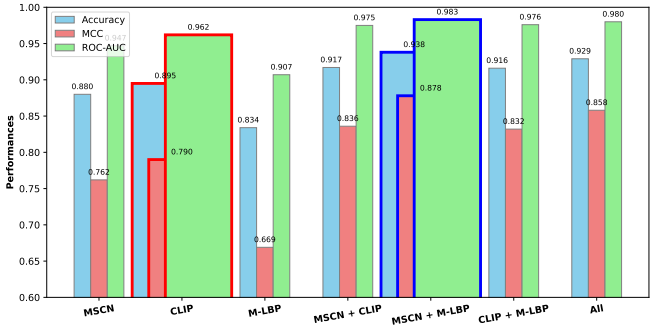

FIGURE 4 : Comparaison des performances des descripteurs visuels individuels et combinés. Les résultats sont présentés sous forme de scores moyens et les meilleures performances sont mises en évidence par des bords colorés.

naturelles de celles générées par des modèles avancés. La combinaison MSCN + M-LBP a atteint la meilleure performance globale, avec une précision de 93,8%, démontrant que l'intégration de caractéristiques spatiales et texture maximise les performances en capturant les motifs d'image. Ces résultats soulignent la nature complémentaire de plusieurs descripteurs où leur combinaison améliore la classification des images naturelles par rapport aux images générées par l'IA.

4 Conclusion

Cette étude met en lumière le rôle crucial de l'intégration des descripteurs visuels dans la classification des images naturelles par rapport aux images générées par l'IA. En examinant les performances individuelles et combinées des caractéristiques MSCN, CLIP et M-LBP, nous démontrons que, bien que chaque descripteur de caractéristiques apporte des informations précieuses, la combinaison des trois offre les résultats de classification les plus précis et robustes. Plus précisément, l'approche MSCN + M-LBP surpasse les caractéristiques individuelles, illustrant la capacité des descripteurs visuels hybrides à relever les défis complexes de la distinction entre images naturelles et celles générées par des modèles génératifs avancés. Ces résultats soulignent la nécessité de techniques d'extraction de caractéristiques multidimensionnelles et complémentaires, ouvrant une voie prometteuse pour les recherches futures sur l'analyse des contenus GenAI et la vérification de l'authenticité des images. De plus, les conclusions de cette étude pourraient guider le développement d'outils de détection plus efficaces pour faire face à la nature de plus en plus sophistiquée des médias générés par l'IA.

Références

[1] Q. BAMMEY : Synthbuster : Towards detection of diffusion model generated images. *IEEE OJSP*, 5:1–9, 2024.

[2] CM. BISHOP et NM. NASRABADI : *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[3] T. CHEN et C. GUESTRIN : XGBoost : A scalable tree boosting system. *In ACM SIGKDD*, page 785–794, 2016.

[4] D. DANG-NGUYEN, C. PASQUINI, V. CONOTTER et G. BOATO : Raise : A raw images dataset for digital image forensics. *In ACM MSC*, pages 219–224, 2015.

[5] R. DURALL, M. KEUPER et J. KEUPER : Watch your up-convolution : CNN Based generative deep neural networks are failing to reproduce spectral distributions. *In IEEE/CVF CVPR*, June 2020.

[6] J. FRANK, T. EISENHOFER, L. SCHÖNHERR, A. FISCHER, D. KOLOSSA et T. HOLZ : Leveraging frequency analysis for deep fake image recognition. *In ICML*, volume 119, pages 3247–3258. PMLR, 13–18 Jul 2020.

[7] T. KARRAS, S. LAINE et T. AILA : A style-based generator architecture for generative adversarial networks. *IEEE TPAMI*, 43(12):4217–4228, 2021.

[8] F. MARRA, D. GRAGNANIELLO, D. COZZOLINO et L. VERDOLIVA : Detection of gan-generated fake images over social networks. *In IEEE MIPR*, pages 384–389, Miami, FL, USA, 2018.

[9] F. MATERN, C. RIESS et M. STAMMINGER : Exploiting visual artifacts to expose deepfakes and face manipulations. *In IEEE WACVW*, pages 83–92, Waikoloa, HI, USA, 2019.

[10] A. MITTAL, Ak. MOORTHY et AC. BOVIK : No-reference image quality assessment in the spatial domain. *IEEE TIP*, 21(12):4695–4708, 2012.

[11] A. RADFORD et et AL. : Learning transferable visual models from natural language supervision. *In ICML*, volume 139, pages 8748–8763, 2021.

[12] A. RAMESH, M. PAVLOV, G. GOH, S. GRAY, C. VOSS, A. RADFORD, M. CHEN et I. SUTSKEVER : Zero-shot text-to-image generation. *In ICML*, volume 139, pages 8821–8831, 2021.

[13] J. RICKER, S. DAMM, T. HOLZ et A. FISCHER : Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv :2210.14571*, 2022.

[14] R. ROMBACH, A. BLATTMANN, D. LORENZ, P. ESSER et B. OMMER : High-resolution image synthesis with latent diffusion models. *In IEEE/CVF CVPR*, pages 10684–10695, June 2022.