Réseau de neurones convolutif pour l'extraction d'attributs de texture à partir d'images multispectrales

Anis Amziane, Olivier Losson, Benjamin Mathon, et Ludovic Macaire
Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

Résumé

Les caméras multispectrales de type "snapshot" équipées d'une matrice de filtres optiques multispectraux (MSFA) acquièrent instantanément plusieurs bandes spectrales et fournissent une image brute dans laquelle un seul canal est disponible pour chaque pixel. Les caractéristiques de texture sont classiquement extraites d'images entièrement définies qui sont estimées par dématriçage. Cette procédure peut toutefois générer des artefacts spatio-spectraux. En outre, les coûts de calculs de l'extraction d'attributs de texture ainsi que la dimension de ces derniers augmentent avec le nombre de bandes spectrales échantillonnées par les filtres de la caméra. Dans cet article, nous proposons une approche originale basée sur un réseau neuronal convolutif appelé MSFA-Net pour capturer des interactions spatio-spectrales dans les images brutes à coûts de calcul réduits. Les expériences de classification d'images multispectrales et de segmentation d'images acquises en conditions extérieures montrent que l'approche proposée surpasse plusieurs descripteurs de l'état de l'art.

Mots clefs

Imagerie multispectrale, texture, matrice de filtres multispectraux (MSFA), classification, segmentation.

1 Introduction

Les caméras multispectrales intègrent plusieurs filtres optiques, ce qui permet d'observer les surfaces des matériaux dans plusieurs bandes spectrales. Selon le type de filtres qui échantillonnent la lumière incidente (radiance), les images multispectrales peuvent contenir de l'information spectrale associée au domaine du visible (VIS), du proche infrarouge (NIR) et/ou de l'infrarouge. Les dispositifs « multi-shot » [1] produisent une image multispectrale en empilant plusieurs *frames* acquises successivement. À l'inverse, les dispositifs « snapshot » fournissent une image multispectrale à partir d'une seule acquisition [2]. Les caméras snapshot multicapteurs utilisent des prismes dichroïques pour scinder le faisceau de lumière entrant sur plusieurs capteurs selon des plages de longueurs d'onde. Ils sont donc coûteux et ne peuvent échantillonner que quelques bandes spectrales. Les dispositifs snapshot monocapteur intègrent une matrice de filtres optiques multispectraux (MSFA) recouvrant le capteur, comme le filtre de Bayer (CFA) largement utilisé en imagerie couleur, afin d'échantillonner spatialement et spectralement la radiance incidente en fonction de l'emplacement des photo-capteurs. Chaque filtre du MSFA est sensible à une bande spectrale étroite spécifique, de sorte que chaque pixel de l'image *brute* ainsi acquise représente une seule bande. Les autres bandes manquantes sont estimées par dématriçage pour reconstruire l'image multispectrale pleinement définie [3]. Certaines applications (comme l'identification d'adventices en plein champ) nécessitent des signatures spectrales indépendantes de l'éclairage. Pour ce faire, les images de réflectance sont classiquement estimées à partir des images de radiance dématricées, et des attributs de texture en sont extraits. Comme le dématriçage génère des artefacts et augmente les coûts de calcul, certains auteurs proposent de traiter directement les images brutes pour l'estimation de la réflectance [4] ou l'extraction d'attributs [5]. Dans [6], des attributs de texture basés sur les motifs locaux binaires (LBPs) sont directement extraits des images brutes. Dans le même esprit, nous exploitons ici les avantages de l'apprentissage profond et proposons un réseau neuronal convolutif (CNN) qui agit comme un extracteur d'attributs de texture à partir d'images brutes [7].

2 Extraction d'attributs de textures bruts par CNN

2.1 Image brute acquise via un MSFA

Pour classer les images de texture fournies par une caméra mono-capteur échantillonnant B^2 bandes via un MSFA, le dématriçage estime généralement des images pleinement définies (sur B^2 canaux) à partir d'images brutes. De ces images pleinement définies sont extraits des attributs de texture [8]. Cette approche peut se révéler gourmande en temps de calcul et en mémoire, surtout avec des images à haute définition spectrale. De plus, les interactions spatio-spectrales peuvent ne pas être prises en considération de manière efficace, ce qui affaiblit le pouvoir discriminant des attributs extraits.

Pour prendre en compte la corrélation spatio-spectrale, certaines études traitent directement les images brutes [5, 6]. Lorsqu'un descripteur analyse efficacement une image brute, il peut atteindre des performances de classification similaires, voire supérieures, à celles obtenues à partir d'une image pleinement définie, car le dématriçage génère des artefacts susceptibles d'altérer la représentation de la texture. Dans [6], les attributs de texture sont directement

calculés à partir d'images brutes, ce qui évite l'étape de dématriçage et fournit des attributs discriminants. Plus précisément, la méthode analyse une image brute en fonction du motif de base du MSFA et de sa disposition pour construire un descripteur de texture basé sur l'opérateur LBP. En s'inspirant de ces travaux, nous proposons ici une nouvelle architecture CNN adaptée aux images brutes. Les MSFAs utilisés dans ce travail sont définis par la répétition d'un motif de base $B \times B$ qui échantillonne B^2 bandes différentes. Il n'existe pas de consensus concernant la taille du motif de base, et la recherche d'un compromis entre les échantillonnages spatiaux et spectraux reste un problème ouvert difficile [9] qui dépasse le cadre du présent article. Par conséquent, nous suivons les dispositions MSFA de deux caméras snapshot fabriquées par IMEC [10] et opérant dans les domaines VIS ($B = 4$) et NIR ($B = 5$) (voir Fig. 1).

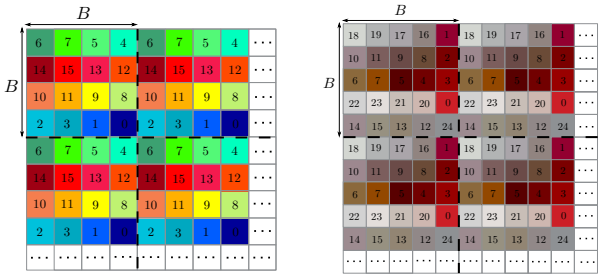

FIGURE 1 – *MSFAs utilisés : IMEC VIS* 4×4 *(*$\lambda^b \in \{469\,nm,\ \ldots,\ 633\,nm\}$*,* $b \in [\![0, 15]\!]$*) (gauche) et NIR* 5×5 *(*$\lambda^b \in \{678\,nm,\ \ldots,\ 960\,nm\}$*,* $b \in [\![0, 24]\!]$*) (droite).*

2.2 Architecture CNN proposée

L'architecture CNN proposée, nommée MSFA-Net, extrait directement des attributs de texture à partir de patchs carrés de l'image brute de taille $X \times X$ pixels, où $X = m \cdot B$ est un multiple de la largeur du motif de base du MSFA. MSFA-Net est composée de trois blocs convolutifs, suivis d'une couche de sous-échantillonnage (« pooling ») qui moyenne les cartes d'attributs et de deux couches entièrement connectées. La première couche de convolution est la plus importante, car elle guide l'extraction d'attributs selon le motif de base du MSFA. Elle utilise 128 noyaux convolutifs $\{H_n\}_{n=0}^{127}$ de taille $B \times B$ et de profondeur 1, avec un pas de B pixels dans les deux dimensions spatiales et sans remplissage par zero (« zero padding »). Un pas de B pixels garantit que chaque coefficient du noyau est toujours associé à la même bande du MSFA pour toutes les convolutions. Cette première couche apprend des interactions spatiales et spectrales entre les valeurs des canaux dans chaque patch brut qui correspond au motif élémentaire du MSFA. La convolution entre un patch brut P^{raw} et un noyau H_n, $n \in [\![0, 127]\!]$, est définie en chaque pixel $(x, y) \in [\![0, m-1]\!]^2$ par :

$$O_n(x,y) = \sum_{i=0}^{B-1} \sum_{j=0}^{B-1} H_n(i,j) \cdot P^{\text{raw}}(B \cdot x + i, B \cdot y + j). \quad (1)$$

Les 128 cartes d'attributs résultantes $\{O_n\}_{n=0}^{127}$, de taille $m \times m$, sont introduites dans le deuxième bloc convolutif qui utilise 256 noyaux de taille 3×3 avec un pas et un remplissage par zero d'un pixel, de sorte que les tailles des cartes d'attributs d'entrée et de sortie soient identiques. Le dernier bloc convolutif utilise 384 noyaux de taille 3×3 avec un pas d'un pixel et sans remplissage par zero. Pour être invariant aux translations spatiales et robuste au bruit, les cartes d'attributs de la dernière couche convolutive sont injectées dans une couche de sous-échantillonnage. cette dernière moyenne les cartes d'attributs canal par canal. Afin d'introduire une non-linéarité et de réduire la dimension des attributs, le vecteur d'attributs de dimension 384 est injecté dans une couche entièrement connectée qui fournit le vecteur de texture final de taille 128.

3 Évaluation expérimentale

3.1 Description

Nous évaluons notre approche de classification et de segmentation d'images multispectrales à l'aide de deux bases : HyTexiLa [11], formée de 112 images pour la classification de textures, et une base de 96 images dont nous disposons, dédiée à la reconnaissance de cultures et adventices (« mauvaises herbes »).

Nous comparons notre descripteur avec ceux de l'état de l'art, à savoir trois descripteurs basés sur un apprentissage profond et quatre autres basés sur l'opérateur LBP. Nous adaptons d'abord à nos images le modèle SegNet-Basic (version simplifiée de SegNet [12]) en retenant uniquement l'encodeur, complété d'une couche de vectorisation d'attributs et deux couches entièrement connectées pour obtenir un vecteur de 512 attributs. Nous testons également le modèle S-CNN [13], composé de trois couches convolutives et de deux couches entièrement connectées, dont la première fournit un vecteur de 1024 attributs. Enfin, nous considérons l'extraction d'attributs par apprentissage résiduel profond [14] grâce à l'architecture à 18 couches (ResNet18), qui fournit un vecteur de 512 attributs. Pour les descripteurs basés sur l'opérateur LBP, nous considérons le LBP marginal (comme descripteur de base) [6], les motifs angulaires locaux (LAP) [15], ainsi que les descripteurs LBP-LCC [16] et M-LBP [6].

3.2 Résultats et discussions

Le tableau 1 montre les résultats de classification obtenus par chaque attribut avec, comme classifieur, le plus proche voisin (1-ppv) couplé avec la distance euclidienne. Parmi les descripteurs non basés sur un apprentissage, M-LBP est plus performant que les autres descripteurs basés sur l'opérateur LBP, car il prend en compte la corrélation

spatio-spectrale au sein de l'image brute et évite l'étape de dématriçage, qui peut affecter la représentation des textures. Globalement, les performances de tous les descripteurs augmentent avec la taille des patchs, en particulier celles du LBP marginal et du LAP, qui sont sensibles au nombre de pixels du patch. Les encodeurs SegNet-Basic, S-CNN et ResNet18 sont peu affectés par la taille des patchs et ne fournissent pas de meilleurs résultats que M-LBP avec le MSFA IMEC 5×5 dans la plupart des cas. L'approche que nous proposons est classée première cinq fois parmi les cas testés, suivie de ResNet18. Cela confirme que les attributs fournis par MSFA-Net sont discriminants malgré leur petite taille. Dans l'ensemble, MSFA-Net fournit des performances meilleures que les autres approches ou comparables à des coûts de calcul nettement inférieurs (voir Fig. 2). Le tableau 1 montre également que les performances de MSFA-Net sont moins affectées par la taille des patchs que celles des descripteurs calculés sans apprentissage, ce qui le rend intéressant pour effectuer des tâches de segmentation.

La Fig. 3 montre les résultats de la segmentation obtenue par MSFA-Net et SegNet-Basic sur deux images tests pour le problème de détection et d'identification des cultures de betteraves et de leurs adventices. Elle montre des performances comparables en matière de détection des adventices entre SegNet-Basic et MSFA-Net. Elle montre également que pour l'identification des betteraves et des adventices, les attributs extraits par MSFA-Net permettent au classifieur de mieux distinguer les betteraves et les feuilles de chénopode.

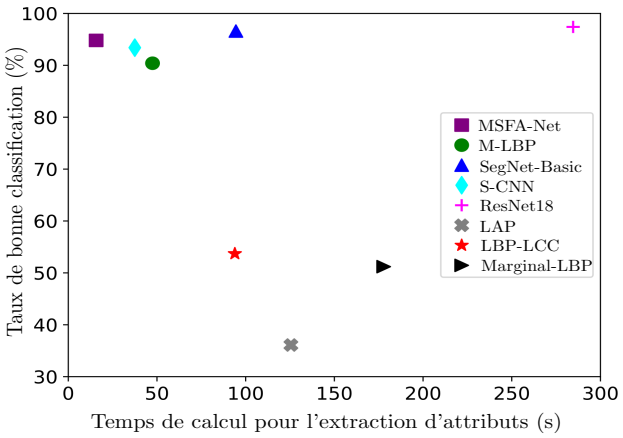

FIGURE 2 – *Taux de bonne classification par 1-ppv vs. temps d'extraction d'attributs pour* $\approx 35,9 \cdot 10^3$ *patchs apprentissage de* 65×65 *pixels (simulés avec le MSFA IMEC* 5×5*) de la base HyTexiLa. Les temps de dématriçage et d'apprentissage des réseaux ne sont pas pris en compte.*

4 Conclusion et perspectives

Dans cet article, nous avons proposé une approche originale pour l'extraction d'attributs de texture à partir d'images brutes grâce à une architecture CNN appelée

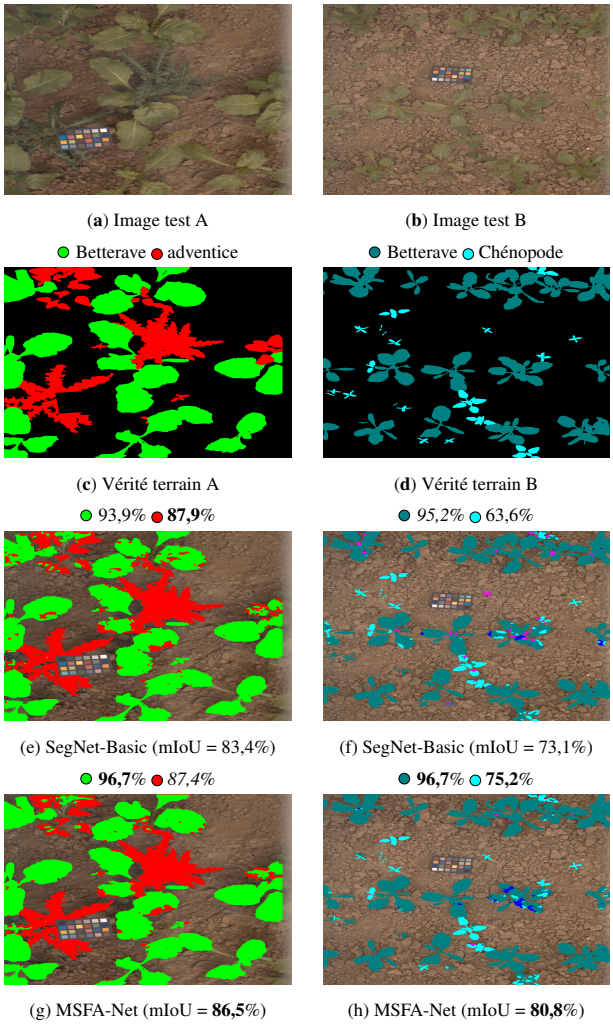

FIGURE 3 – *Résultats de segmentation (Intersection-sur-union moyenne (mIoU) et taux de bonne classification par classe) obtenus par les attributs extraits par SegNet-Basic et MSFA-Net. (a, b) : rendus RGB de deux images multispectrales tests ; (c, d) : vérités terrain ; (e, k) : résultats de détection betterave/adventice ; (f, h) : résultats d'identification betterave/adventice. Les valeurs en gras indiquent les meilleurs résultats. Les couleurs magenta et bleu dans (f, h) correspondent aux pixels chénopode classés comme datura ou chardon, respectivement.*

MSFA-Net. Cette approche évite l'étape de dématriçage qui peut être gourmande en temps de calculs et peut altérer la représentation des textures. Elle nécessite l'apprentissage de beaucoup moins d'hyper-paramètres que les autres architectures CNN testées. Des expériences sur la classification d'images et la segmentation des cultures/adventices montrent que MSFA-Net est globalement plus performante que les autres approches testées, avec des coûts de calcul bien moindres. Les travaux futurs se focaliseront sur la conception d'architectures CNN plus robustes aux perturbations extérieures, liées par exemple à la variation de l'éclairage et aux ombres portées.

TABLEAU 1 – *Taux de bonne classification (%) obtenu par 1-ppv et les attributs extraits à partir d'images brutes ou dématricées, sur la base HyTexiLa. Le meilleur résultat de chaque colonne est affiché en gras, le second meilleur en italique. Le symbole* * *fait référence au MSFA IMEC* 4×4*,* † *à IMEC* 5×5*. L'architecture de ResNet18 utilisée est disponible sur* `https://paperswithcode.com/model/resnet`.

Patch d'entrée	Attribut	taille	IMEC $4 \times 4^*$			IMEC $5 \times 5^\dagger$		
			200×200	124×124	64×64	200×200	125×125	65×65
MSFA	MSFA-Net	$128^{*,\dagger}$	**99,5**	**98,3**	**98,7**	**99,0**	**98,4**	95,1
	M-LBP [6]	$4096^*/6400^\dagger$	*97,2*	*96,9*	*94,6*	96,9	95,4	90,4
Dématricé	SegNet-Basic [12]	$512^{*,\dagger}$	86,2	83,5	86,4	97,1	96,9	*96,6*
	S-CNN [13]	$1024^{*,\dagger}$	82,5	83,6	81,1	94,4	97,4	93,4
	ResNet18	$512^{*,\dagger}$	95,5	88,1	81,7	*98,5*	*97,8*	**97,4**
	LAP [15]	$256^{*,\dagger}$	80,0	69,1	41,3	68,4	65,6	36,1
	LBP-LCC [16]	$512^{*,\dagger}$	87,0	83,8	69,6	70,9	71,4	53,7
	Marginal LBP [6]	$4096^*/6400^\dagger$	81,5	76,4	45,9	77,2	71,6	51,2

Références

[1] Julien Pichette, Wouter Charle, et Andy Lambrechts. Fast and compact internal scanning CMOS-based hyperspectral camera: the Snapscan. Dans *Procs. SPIE*: *Photonic Instrumentation Engineering IV*, volume 10110, pages 1–10, San Francisco, USA, 2017.

[2] Nils Genser, Jürgen Seiler, et André Kaup. Camera array for multi-spectral imaging. *IEEE Transactions on Image Processing*, 29:9234–9249, 2020.

[3] Vishwas Rathi et Puneet Goyal. Generic multispectral demosaicking based on directional interpolation. *IEEE Access*, 10:64715–64728, 2022.

[4] Vlado Kitanovski, Jean-Baptiste Thomas, et Jon Yngve Hardeberg. Reflectance estimation from snapshot multispectral images captured under unknown illumination. Dans *Procs. 29th Color and Imaging Conference*, pages 264–269, Online, 2021.

[5] Wei Zhou, Shengyu Gao, Ling Zhang, et Xin Lou. Histogram of oriented gradients feature extraction from raw Bayer pattern images. *IEEE Transactions on Circuits and Systems II*: *Express Briefs*, 67(5):946–950, 2020.

[6] Sofiane Mihoubi, Olivier Losson, Benjamin Mathon, et Ludovic Macaire. Spatio-spectral binary patterns based on multispectral filter arrays for texture classification. *Journal of the Optical Society of America A*, 35(9):1532–1542, 2018.

[7] Anis Amziane, Olivier Losson, Benjamin Mathon, et Ludovic Macaire. MSFA-Net: a convolutional neural network based on multispectral filter arrays for texture feature extraction. *Pattern Recognition Letters*, 168:93–99, 2023.

[8] Alice Porebski, Mohamed Alimoussa, et Nicolas Vandenbroucke. Comparison of color imaging vs. hyperspectral imaging for texture classification. *Pattern Recognition Letters*, 161:115–121, 2022.

[9] Travis W. Sawyer, Michaela Taylor-Williams, Ran Tao, Ruqiao Xia, Calum Williams, et Sarah E. Bohndiek. Opti-MSFA: a toolbox for generalized design and optimization of multispectral filter arrays. *Optics Express*, 30(5):7591–7611, 2022.

[10] Bert Geelen, Nicolaas Tack, et Andy Lambrechts. A compact snapshot multispectral imager with a monolithically integrated per-pixel filter mosaic. Dans *Procs. SPIE*: *Advanced Fabrication Technologies for Micro/Nano Optics and Photonics VII*, volume 8974, pages 1–8, San Francisco, USA, 2014.

[11] Haris Ahmad Khan, Sofiane Mihoubi, Benjamin Mathon, Jean-Baptiste Thomas, et Jon Yngve Hardeberg. HyTexiLa: high resolution visible and near infrared hyperspectral texture images. *Sensors*, 18(7):2045, 2018.

[12] Vijay Badrinarayanan, Alex Kendall, et Roberto Cipolla. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[13] Vivek Sharma, Ali Diba, Tinne Tuytelaars, et Luc Van Gool. Hyperspectral CNN for image classification & band selection, with application to face recognition. *Technical report KUL/ESAT/PSI/1604, KU Leuven, ESAT, Leuven, Belgique*, 2016.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et Jian Sun. Deep residual learning for image recognition. Dans *Procs. IEEE CVPR*, pages 770–778, Las Vegas, USA, 2016.

[15] Claudio Cusano, Paolo Napoletano, et Raimondo Schettini. Local angular patterns for color texture classification. Dans *Procs. 18th ICIAP Workshops*, volume 9281 de *LNCS*, pages 111–118, Gênes, Italie, 2015.

[16] Claudio Cusano, Paolo Napoletano, et Raimondo Schettini. Combining local binary patterns and local color contrast for texture classification under varying illumination. *Journal of the Optical Society of America A*, 31(7):1453–1461, 2014.