Régression locale de la profondeur grâce au flou de défocalisation et à un réseau de neurones entraîné par *soft-assignment*

R. LEROY[1], P. TROUVÉ-PELOUX[1], B. LE SAUX[2], B. BUAT[1], F. CHAMPAGNAT[1]

[1]DTIS, ONERA - Université Paris-Saclay, F-91123, Palaiseau, France

[2]Φ-lab, ESA-ESRIN, Frascati, Italie

remy.leroy@onera.fr

Résumé – Nous présentons une nouvelle approche de régression de la profondeur exploitant le flou de défocalisation dans des patchs d'image. La plupart des méthodes dites de *Depth from Defocus* (DFD) reposent sur une classification locale de flou parmi un ensemble de flous potentiels liés à une profondeur, ce qui induit des erreurs dues à sa variation continue en pratique. Nous proposons d'utiliser un encodage de la profondeur réelle en un vecteur de probabilité d'appartenance à plusieurs labels et d'entraîner un réseau de neurones de régression à partir de ce vecteur. Notre méthode surpasse à la fois le codage dur (*i.e.* classification) et les approches de régressions directes, sur des images simulées provenant de jeux de données de textures structurées, et sur un jeu de données réelles provenant d'une expérience de DFD actif.

Abstract – We present a novel patch-based approach for depth regression from defocus blur. Most state of the art methods for *Depth from Defocus* (DFD) use a patch classification approach among a set of potential defocus blurs related to a depth, which induces errors due to the continuous variation of the depth. Here, we propose to use a soft assignment encoding of the true depth into a membership probability vector and train a simple regression model so that its output vector fits this probability. Our method outperforms both hard encoding (*i.e.* classification) and direct regression approaches, on simulated images from structured texture datasets, and on a real dataset from an active DFD experiment.

1 Introduction

L'estimation monoculaire de la profondeur a été largement étudiée ces dernières années, en particulier lorsque des informations 3D sont requises dans un environnement contraint. Plusieurs méthodes sont basées sur l'apprentissage d'un modèle de réseau de neurones profond à partir d'une seule image, par régression d'une carte de profondeur ou de nuages de points 3D grâce aux informations contextuelles d'images complètes [1, 2]. L'image d'entrée étant redimensionnée en fonction de l'architecture du réseau de neurones, cela affecte le flou du capteur dans l'image, en particulier le flou de défocalisation qui fournit des informations sur la profondeur.

D'autres travaux exploitent des indices de défocalisation à l'aide de l'apprentissage profond, soit pour la prédiction de cartes de profondeur [3, 4] soit pour la prédiction de carte de défocalisation relative comme une étape intermédiaire pour le défloutage d'image [5]. Ces méthodes appliquent le même filtrage à tous les pixels de l'image, et seront perturbées par une variation spatiale de la PSF due aux aberrations optiques. En outre, ces méthodes utilisent en entrée des morceaux d'images, réduisant l'information de contexte au bénéfice de l'information du flou. Une méthode de prédiction locale de la profondeur semble donc plus adaptée, notamment pour les capteurs à faible coût dont les aberrations optiques ne sont pas corrigées. Des méthodes locales de *Depth from Defocus* (DFD) ont été proposées dans la littérature, à commencer par les travaux pionniers de A. Pentland [6]. Dans le cas mono-image, l'approche commune consiste à sélectionner un flou parmi un ensemble fini de flous potentiels, en utilisant un critère de sélection dérivé des approches de maximum de vraisemblance [7, 8, 9] ou bien directement un réseau de neurones de classification [10]. Or en pratique les données réelles présentent une variation continue de la profondeur. De plus, ces approches de DFD ne considèrent pas la relation existante entre les classes dans la prédiction de la profondeur [1]. À notre connaissance, seules quelques méthodes traitent la régression du flou localement à partir d'un patch préfiltré, en utilisant soit des arbres de régression [11], un GRNN (*general regression neural network*) [12] ou un réseau de neurones convolutif [4].

Nous proposons une nouvelle méthode de régression de la profondeur à partir d'un patch d'image en utilisant les indices de défocalisation d'un capteur monoculaire. Notre méthode ne nécessite qu'un ensemble d'apprentissage composé de correspondances patch/valeur, sans modèle analytique de flou ou de scène, ni préfiltrage de patch. Nous proposons un modèle de régression simple, entraîné à l'aide d'un codage par affectation souple de la profondeur réelle dans un vecteur de probabilité d'appartenance. Notre méthode surpasse les approches par *hard-assignment* (*i.e.* classification) ainsi que la régression directe, sur des jeux de données de textures structurées. Enfin, nous testons notre méthode sur un jeu de données réelles de motifs binaires aléatoires pour l'inspection de surfaces [8].

2 Travaux Connexes

La plupart des méthodes locales de DFD monoculaire utilisent une classification parmi un ensemble fini de flous potentiels par un critère de sélection dérivé d'une approche par maximum de vraisemblance dans un cadre bayésien [7, 9, 8]. Des modèles statistiques simples des gradients de la scène sont généralement utilisés pour obtenir une expression analytique de la vraisemblance d'un flou donné. Des méthodes d'apprentissage profond ont également été développées pour la classification locale des paramètres du flou pour l'estimation monoculaire de la profondeur [10]. À notre avis, les méthodes de DFD par classification ne sont pas adaptées car la profondeur varie continûment dans les images réelles, ce qui introduit une erreur d'estimation systématique due au pas de quantification. La réduction de ce pas implique une augmentation du coût de calcul pour les méthodes par maximum de vraisemblance, et un entraînement des méthodes par apprentissage compliqué par le nombre réduit d'exemples par classe dans le cas d'un faible volume de données d'entraînement. Enfin, les approches de classification ne tiennent pas compte de la relation de voisinage existante entre différentes classes de profondeur [1]. Il existe des méthodes de régression locale des paramètres de flou à partir d'un patch pré-filtré, via un GRNN [12] ou bien un arbre de régression [11]. À notre connaissance, seul Kashiwagi *et al*. utilise un réseau de neurones convolutif profond pour la régression locale de la profondeur. Il est alimenté par les gradients des patchs, ainsi que par une branche exploitant la localisation du patch dans l'image pour construire des cartes d'attention afin de gérer la dépendance aux aberrations optiques. L'entraînement est effectué en minimisant directement une erreur L_1 sur la profondeur scalaire [4], approche qui est sujette à un problème de régression vers la moyenne comme le démontre la section 4.1.

Nous présentons ici une nouvelle méthode locale de DFD par entraînement d'un réseau de régression simple en utilisant un codage par *soft-assignment* de la valeur de profondeur réelle en un vecteur d'appartenance à des classes de profondeurs prédéfinies. Nous utilisons un terme d'attache aux données sur le vecteur d'appartenance correspondant (Section 3). Notre approche nécessite uniquement un ensemble d'apprentissage de paires patch/valeur, sans modèle analytique de flou ou de scène, ni filtrage préalable des patchs. Nous validons notre méthode sur des données synthétiques (Section 4.1) et réelles provenant d'une expérience de DFD active [8] (Section 4.2). Nous comparons les résultats d'estimation utilisant le *soft-assignment* proposé avec les méthodes de régression directe ou par encodage.

3 Méthode

A partir des bons résultats de classification de flou [10], nous proposons de transformer un réseau de classification de profondeur, sujet à des erreurs issues de la quantification, en un réseau de régression. La figure 1 illustre la méthode et l'architecture proposées. Un réseau de neurones entièrement convolutif (CNN), prenant un patch d'image, renvoie un vecteur de logit $\mathbf{y}$ correspondant, puis, un vecteur d'appartenance $\tilde{\mathbf{p}} = \{\tilde{p}_i\}_{i=1}^N$ est obtenu par un opérateur softmax, N étant le nombre de classes. Une couche linéaire, appelée *échelle de régression*, paramétrée par $\mathbf{z} = \{z_i\}_{i=1}^N$, est appliquée à $\tilde{\mathbf{p}}$ pour donner notre valeur estimée $\tilde{z}$. Nous décrivons ici plusieurs approches de régression, dont celle exploitant le codage par *soft-assignment* que nous proposons.

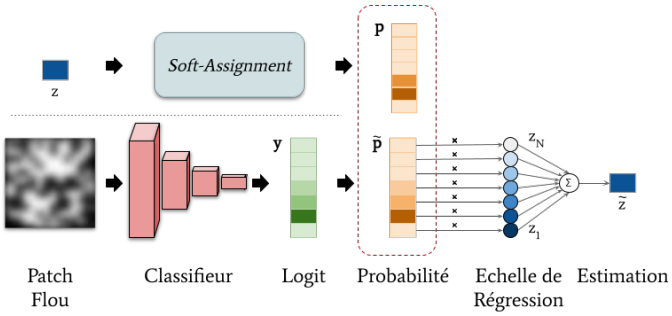


FIGURE 1 – Illustration de la méthode proposée. Un classifieur entièrement convolutif estime un vecteur logit à partir d'un patch flou. La valeur de profondeur régressée est obtenue par combinaison linéaire du vecteur d'appartenance $\tilde{\mathbf{p}}$ et d'une échelle de régression. La profondeur réelle est encodée dans un vecteur d'appartenance cible à l'aide d'un *soft-assignment*.

3.1 Régression dans l'espace de sortie

L'approche classique de régression consiste à apprendre les paramètres du modèle, l'échelle de régression et un biais en minimisant une perte L_2 directement sur les vraies valeurs de profondeur exprimées par $\mathcal{L}_{out} = (\tilde{\mathbf{p}}^T\mathbf{z} + b - z)^2 + \lambda_r\|\mathbf{y}\|_1$. avec $\mathbf{z}$ et b étant des paramètres appris. Dans ces conditions $\mathcal{L}_{out}$ est invariante par permutation simultanée de $\tilde{\mathbf{p}}$ et $\mathbf{z}$, ce qui multiplie les minima locaux et complique l'apprentissage.

3.2 Régression dans un espace latent

Pour cette approche, l'espace de profondeur est divisé en une échelle de valeurs discrètes de profondeur $\mathbf{z}$, comme en classification, à laquelle nous accolons un vecteur de probabilité $\mathbf{p}$. Nous réalisons la régression sur ce vecteur de probabilité $\mathbf{p}$ au moyen d'un codage d'une valeur vraie z en un vecteur de probabilité $\mathbf{p}$. Nous présentons deux formes d'encodages capable de construire ce vecteur.

Hard-Assignment Dans cet encodage, une profondeur z est attribuée à une unique classe j, où $j = \arg\min_i |z - z_i|$. Cela se traduit par un vecteur de probabilité $\mathbf{p}$ ayant $p_j = 1$ et $p_{i\neq j} = 0$, comme le montre la figure 2. Le décodage pour obtenir la valeur estimée $\tilde{z}$ est effectué grâce à l'opérateur $\arg\max$: $\tilde{z} = z_{\arg\max_i(\tilde{p}_i)}$. Ce codage est celui utilisé usuellement par les méthodes d'estimation par classification. Le décodage par

arg max rend les estimations sensibles aux erreurs de classification. Une façon d'atténuer cette sensibilité est de considérer une opération de soft arg max sur le vecteur de probabilité, qui est définie comme $\tilde{z} = \sum_{i=1}^{N} \tilde{p}_i \cdot z_i = \tilde{\mathbf{p}}^T \mathbf{z}$.

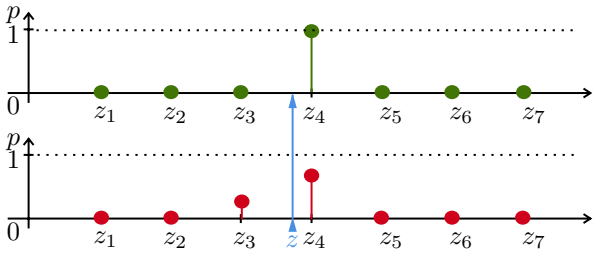


FIGURE 2 – Représentation d'un encodage sur 7 classes d'une valeur z par *hard-assignment* (haut) et *soft-assignment* (bas).

Soft-Assignment Le *soft-assignment* distribue des probabilités d'appartenance sur plusieurs classes. Cette approche permet d'injecter une relation de voisinage entres les valeurs de l'échelle de régression , comme dans les travaux de Proença *et al.* [13] où le soft-assignment est utilisé pour encoder des poses de satellites. Le vecteur de probabilité $\mathbf{p}$, correspondant à la décomposition de z dans l'espace des z_i, est obtenu par la règle classique [14] $p_i = K(z_i, z) / \sum_j K(z_j, z)$ où K est un noyau préalablement spécifié. Nous optons pour un noyau B-spline d'ordre 1, notamment pour sa caractéristique de reconstruction sans perte : $K(z_i, z) = [\delta - |z_i - z|]_+$, où δ est la distance entre deux points consécutifs de l'échelle. Enfin, la valeur estimée $\tilde{z}$ est reconstruite à partir de $\mathbf{p}$ grâce à l'opération de soft arg max : $\tilde{z} = \tilde{\mathbf{p}}^T \mathbf{z}$. La figure 2 représente le vecteur de probabilité $\mathbf{p}$ correspondant à une valeur vraie z en utilisant une échelle de régression de 7 valeurs $\{z_i\}_{i=1}^{7}$. L'apprentissage minimise un coût d'entropie croisée sur $\tilde{\mathbf{p}}$ et $\mathbf{p}$: $\mathcal{L}_{CE} = -\sum_i p_i \log \tilde{p}_i$.

La table 1 résume le type d'encodage, la fonction de perte, ainsi que l'opération servant à l'estimation pour chacune des approches ci-dessus.

TABLE 1 – Résumé des fonctions de coût et des règles d'estimation pour chaque méthode de régression considérée.

Méthode	Code	Fct. coût	Estimation
Argmax	Hard	$\mathcal{L}_{CE}(\tilde{\mathbf{p}}, \mathbf{p})$	$\tilde{z} = z_{\text{argmax}(\tilde{\mathbf{p}})}$
Soft-Argmax	Hard		$\tilde{z} = \tilde{\mathbf{p}}^T \mathbf{z}$
Soft-Assign.	Soft		$\tilde{z} = \tilde{\mathbf{p}}^T \mathbf{z}$
Rég. Scalaire	-	$(\tilde{z} - z)^2 + \lambda_r \|\mathbf{y}\|_1$	$\tilde{z} = \tilde{\mathbf{p}}^T \mathbf{z} + b$

3.3 Architecture du réseau

Nous considérons un CNN simple, similaire au *inner-net* proposé dans [10]. Il est conçu pour prédire une valeur scalaire pour un patch 32×32 en niveaux de gris donné en entrée, voir figure 1. Il se compose de 7 couches de convolution avec des filtres de dimensions respective : 9×9×64, 5×5×64, 5×5×64, 5×5×64, 5×5×64 et 1×1×N, 1×1×1, et avec un *stride* de 2.

FIGURE 3 – (a) Exemple d'un patch net MBA, et pour des flou de paramètre σ : (b) 0.4, (c) 1.0, (d) 2.0, et (e) 3.0 pixels.

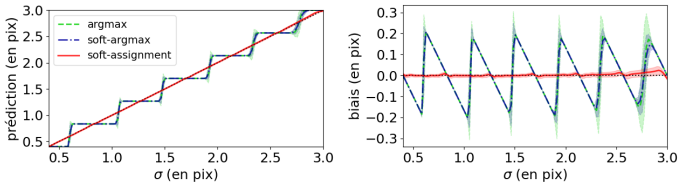
4 Expériences

4.1 Estimation du flou sur données simulées

Nous testons d'abord notre méthode sur un problème d'estimation de l'écart-type σ d'un flou gaussien sur des patchs simulés. Chaque exemple d'entraînement consiste en un patch d'image nette floutée et bruitée, avec le σ correspondant en pixel. Afin de comparer notre approche aux méthodes de l'état de l'art, nous considérons un ensemble d'images de motifs binaires aléatoires (MBA) utilisé dans [8] pour le DFD chromatique actif. Ce jeu de données comprend 10000 motifs différents, 7500 sont utilisés pour la formation et 2500 pour le test. Les modèles sont entraînés à estimer 70 tailles de flou uniformément espacées entre 0.4 et 3.0 pixels. La figure 3 montre des exemples de patch à différents niveaux de flou. La table 2 montre les valeurs de RMSE et MAE, absolues et relatives, pour le meilleur modèle de chacune des différentes approches de régression entraînées sur MBA. Comme discutée en Sec. 3.1, l'approche de régression scalaire est mal conditionnée et tend vers un modèle qui souffre de la régression vers la moyenne [10].

TABLE 2 – RMSE et MAE, absolues et relatives pour différentes méthodes de régression entraînées sur MBA.

Méthode	RMSE (en pix)	MAE (en pix)	RMSE relative (en %)	MAE relative (en %)
Rég. Scalaire	3.07	2.57	42.4	270.4
Ordinal [1]	0.35	0.19	6.9	24.9
Argmax	0.13	0.11	1.1	8.3
Soft-Argmax	0.12	0.10	1.1	7.9
Soft-Assign.	**0.01**	**0.01**	**0.01**	**0.6**



(a) (b)

FIGURE 4 – Valeur moyenne de flou σ estimée (a), et biais (b) avec l'écart-type de prédiction, pour les méthodes par *hard-assignment* and *soft-assignment* entraînées sur MBA.

En comparaison, exploiter l'ordonnancement des profondeurs comme dans [1] conduit à des performances nettement meilleures. Pour les approches de *hard-assignment*, les erreurs de l'arg max sont légèrement diminuées par l'utilisation du soft arg max.

Enfin, notre méthode surpasse nettement les autres. La figure 4 montre pour chaque valeur de flou σ, la moyenne, l'écart-type et le biais du flou estimé. On observe une quantification de l'estimation pour les deux approches par *hard-assignment* (courbes quasiment superposées). D'ailleurs la dispersion des estimations est localisée à la transition entre classes. En revanche notre approche est proche de l'identité.

4.2 Estimation de profondeur sur données réelles

Nous utilisons une base de données réelles de DFD active constituée d'images d'un MBA projeté sur une surface plane qui balaye une plage de distance allant de 300 mm à 350 mm avec un pas de 0,2 mm [8]. La caméra est munie d'un objectif de focale f=25 mm, ouvert à $f/4$, et caractérisé par une aberration chromatique de 200 μm pour renforcer le flou de défocalisation [8]. Nous traitons les images brutes au format Bayer sans dématriçage. L'échelle de régression choisie a 15 profondeurs espacées régulièrement entre 300 et 350 mm. La table 3 montre les performances de notre méthode, comparée aux méthodes par *hard-assignment* et à la méthode proposée par Buat *et al.* [8], pour 51 classes de profondeur régulièrement espacées. Ces résultats confirment les meilleures performances de notre approche par rapport au *hard-assignment*. La figure 5 superpose le biais par profondeur, ainsi que l'écart-type d'estimation pour notre méthode et celle proposée par Buat *et al.* Cette dernière présente un biais d'estimation faible qui augmente avec la profondeur avec une dispersion élevée des estimations causée par la discrétisation des estimations. Notre méthode présente un biais et un écart-type nettement inférieurs.

TABLE 3 – RMSE et MAE, absolues et relatives pour les méthodes de régression et la méthode de Buat *et al.* [8] entraînées sur données expérimentales.

Méthode (# classes)	RMSE (en mm)	MAE (en mm)	RMSE relative (en %)	MAE relative (en %)
Argmax (15)	1.1	9.2 e-1	3.4 e-2	2.8 e-1
Soft-Argmax (15)	8.5 e-1	7.1 e-1	2.6 e-2	2.2 e-1
Soft-Assign. (15)	**6.7 e-2**	**5.0 e-2**	**2.0 e-3**	**1.6 e-2**
Buat *et al.* [8] (51)	9.7 e-1	5.7 e-1	2.9 e-2	1.7 e-1

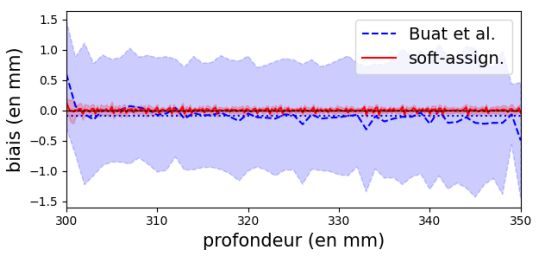


FIGURE 5 – Biais avec écart-type d'estimation par profondeur et biais moyen (pointillé) pour la méthode de Buat *et al.* [8] et de soft-assignment, entrainées sur les données expérimentales.

5 Conclusion

Nous avons proposé une nouvelle approche de régression locale de la profondeur reposant sur un apprentissage guidé par un encodage de la profondeur en un vecteur de probabilité d'appartenance à plusieurs labels. Notre méthode ne nécessite aucun *a priori* image et peut être appliquée à n'importe quelle correspondance image/valeur. De plus, notre approche est plus performante que des approches de classification ou de régression directe. Dans la suite, nous analyserons l'influence de divers paramètres tels que le nombre de classes, la robustesse à différents niveaux de bruit et la taille du patch. Enfin, nous explorerons le cas d'un flou variant spatialement pour gérer les aberrations optiques de champ.

Références

[1] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *CVPR*, 2018.

[2] R. Leroy, P. Trouvé-Peloux, F. Champagnat, B. Le Saux, and M. Carvalho, "Pix2Point : Learning outdoor 3D using sparse point clouds and optimal transport," in *MVA*, 2021.

[3] M. Carvalho, B. Le Saux, P. Trouvé-Peloux, A. Almansa, and F. Champagnat, "Deep Depth from Defocus : How can defocus blur improve 3D estimation using dense neural networks ?," in *ECCV Worskhops*, 2018.

[4] M. Kashiwagi, N. Mishima, T. Kozakaya, and S. Hiura, "Deep depth from aberration map," in *ICCV*, 2019.

[5] H. Ma, S. Liu, Q. Liao, J. Zhang, and J. Xue, "Defocus image deblurring network with defocus map estimation as auxiliary task," *IEEE Trans. on Image Processing*, 2022.

[6] A. P. Pentland, "A new sense for depth of field," *IEEE TPAMI*, 1987.

[7] P. Trouvé, F. Champagnat, G. Le Besnerais, and J. Idier, "Single image local blur identification," in *ICIP*, 2011.

[8] B. Buat, P. Trouvé-Peloux, F. Champagnat, and G. Le Besnerais, "Learning scene and blur model for active chromatic depth from defocus," *Applied Optics*, vol. 60, no. 31, 2021.

[9] X. Zhu, S. Cohen, S. Schiller, and P. Milanfar, "Estimating spatially varying defocus blur from a single image," *IEEE Trans. on Image Processing*, 2013.

[10] H. Haim, S. Elmalem, R. Giryes, A. M. Bronstein, and E. Marom, "Depth estimation from a single image using deep learned phase coded mask," *IEEE Trans. on Computational Imaging*, 2018.

[11] L. D'Andrès, J. Salvador, A. Kochale, and S. Süsstrunk, "Non-parametric blur map regression for depth of field extension," *IEEE Trans. on Image Processing*, 2016.

[12] R. Yan and L. Shao, "Blind image blur estimation via deep learning," *IEEE Trans. on Image Processing*, 2016.

[13] P. F Proença and Y. Gao, "Deep learning for spacecraft pose estimation from photorealistic rendering," in *ICRA*. IEEE, 2020.

[14] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *ICCV*. IEEE, 2011.