Réseaux de neurones profonds pour estimer la profondeur grâce au flou de défocalisation

Thierry DUMAS, Pauline TROUVÉ-PELOUX, Bertrand LE SAUX

ONERA - The French Aerospace Lab
F-91761 Palaiseau, France
pauline.trouve@onera.fr, bertrand.le_saux@onera.fr

Résumé – Nous proposons un nouveau concept de caméra 3D monovoie passive reposant sur l'association d'une optique non conventionnelle optimisée pour renforcer l'estimation de la profondeur à l'aide du flou de défocalisation (DFD pour *Depth from defocus*) avec un algorithme d'estimation de ce flou reposant sur un apprentissage par réseau de neurones. En particulier, pour éviter l'ambiguïté sur la profondeur et les zones aveugles du DFD classique, l'optique possède du chromatisme, ce qui induit un flou de défocalisation variable suivant les canaux rouge (R), vert (G) et bleu (B) de l'image couleur. L'approche par apprentissage par réseau de neurones profond apprend directement la fonction entre le flou et la profondeur, ce qui permet d'éviter une calibration du flou de la caméra pour chaque profondeur. Nous validons notre approche avec des exemples de cartes de profondeur obtenues avec notre algorithme sur des scènes extérieures.

Abstract – We propose a new monocular and passive 3D camera based on the combination of an unconventional lens dedicated to Depth from Defocus (DFD) and a new DFD algorithm based on a deep network learning. DFD is a depth estimation technique that uses the relation between defocus blur and depth. To avoid depth ambiguity and dead zone of DFD, the proposed camera has a uncorrected longitudinal chromatic aberration and thus captures in a single snapshot a RGB image with different defocus blurs. The deep neural network directly learns a function between defocus blur and depth, which avoids a time consuming camera defocus blur calibration. We assess our approach by showing experimental depth maps obtained with the proposed method on outdoor scenes.

1 Introduction

Les imageurs capables d'estimer la profondeur (distance à la caméra), communément appelés caméras 3D ou RGB-D (*Red Green Blue - Depth*), ont de multiples applications, depuis la vision robotique, les interfaces pour les jeux vidéos et les smartphones jusqu'à la vidéo 3D. Dans tous ces cas, les caméras 3D doivent fournir des informations 3D précises en respectant des contraintes d'encombrement, de consommation d'énergie et de temps de calcul. Nous présentons ici une nouvelle caméra 3D passive et monoculaire qui estime la profondeur à partir du flou de défocalisation (*Depth-From-Defocus - DFD*). Les points-clés de cette approche sont une optique non-conventionnelle avec différents flous de défocalisation pour chaque canal couleur et une estimation de la profondeur utilisant des réseaux de neurones profonds.

Plus précisément, la caméra proposée a une lentille avec une aberration chromatique longitudinale non-corrigée. Grâce à un capteur couleur, elle acquiert en une seule prise une image couleur avec différents plans de focalisation selon les canaux, et donc différents flous de défocalisation. Nous montrerons dans la suite que cela permet d'éviter l'ambiguïté de profondeur et la zone aveugle qui apparaissent en DFD avec une optique conventionnelle. De plus, pour avoir une mesure de profondeur exacte et non uniquement relative, les systèmes avec DFD requièrent une étape de calibration coûteuse en temps d'opérateur et délicate dans sa réalisation pour associer une valeur de flou à une profondeur sur l'axe et hors axe. Notre approche vise à contourner ce problème en construisant cette fonction d'association par apprentissage sur un grand volume de données : à cet effet un large ensemble de couples d'images défocalisées et des profondeurs associées a été acquis. Ensuite, un réseau Bayésien profond (*Deep Belief Network - DBN*) avec contrainte de parcimonie est utilisé pour encoder la relation entre le flou et la profondeur. Nous montrons des résultats expérimentaux qui valident notre approche par rapport aux techniques de DFD de l'état de l'art.

Travaux connexes : estimation de la profondeur Dans le domaine des capteurs, la stéréovision basée sur la géométrie multi-vue [4, 11] est un moyen largement étudié qui permet d'obtenir une carte de profondeur. Son principe a récemment été étendu aux caméras actives (avec projecteur infra-rouge) telles que la Kinect. Ces systèmes sont précis, mais plus encombrant qu'une seule caméra. Les caméras plénoptiques, avec une matrice de micro-lentilles placée devant le détecteur, sont une autre approche novatrice, mais à la résolution spatiale encore limitée. Les approches par DFD constituent une autre famille de caméras 3D passives et monoculaires. La principale difficulté rencontrée par ces systèmes est d'estimer avec précision la valeur du flou quand la scène est inconnue. L'estimation de profondeur peut être basée sur plusieurs images [8], ou sur

une seule image acquise avec des optiques non-conventionnelles (comme les pupilles codées [7] ou les ouvertures chromatiques [1]) pour renforcer la discrimination du flou avec la profondeur.

Une approche récente d'apprentissage avec une caméra classique consiste à estimer la profondeur à partir des statistiques de caractéristiques (locales et contextuelles, champs de Markov etc) de l'image [10]. Plus récemment, [2] utilise des larges ensembles d'images RGB-D pour entraîner des réseaux de neurones convolutionnels profonds. Ces approches sont pour l'instant appliquées à des images focalisées et se basent donc principalement sur des statistiques de la scène.

Notre contribution principale réside donc dans la combinaison d'un apprentissage statistique de la profondeur appliqué aux images produites par une caméra chromatique.

2 *Deep Belief Networks* pour estimer le flou de défocalisation

2.1 DFD avec aberration chromatique

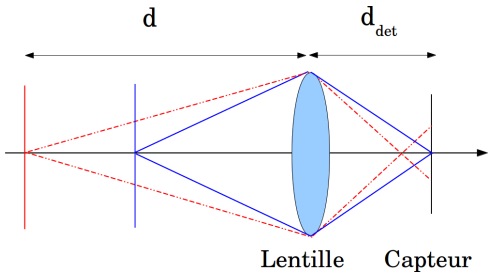

FIGURE 1 – Illustration du principe de la DFD.

La DFD est une approche passive d'estimation de la profondeur : les objets de la scène apparaîtront plus ou moins flous en fonction de leur distance au plan focal (cf. Fig. 1). Cependant une optique conventionnelle présente deux inconvénients : une ambiguïté sur la position par rapport à ce plan (en avant ou en arrière), et une incapacité à mesurer un flou dans la région de profondeur de champ. L'aberration chromatique permet de séparer les plans de mise au point des canaux R, G et B ce qui induit un unique triplet de flou RGB pour une distance donnée [12]. L'ambiguïté a donc disparue ainsi que les zones aveugles liées à la profondeur de champ. Ceci permet entre autre d'acquérir de larges ensembles facilement. Nous utilisons la caméra chromatique proposée dans [12], optimisée pour la plage 1 à 5 m : distance focale de 25 mm, ouverture de 3, une différence de focale entre les canaux rouge et bleu de 100 μm ce qui place les plans focaux RGB à respectivement 3.4, 2.7 et 2.2 m, et une résolution spatiale de 3.45 μm. Les images RGB produites ont une taille de 1226×1028.

2.2 DBN pour l'estimation de la profondeur

Notre but est de construire une fonction qui prédit la profondeur d en fonction de l'information locale d'une fenêtre $\mathbf{v}$ de pixels de l'image RGB : $d = f(\mathbf{v})$. En utilisant un espace

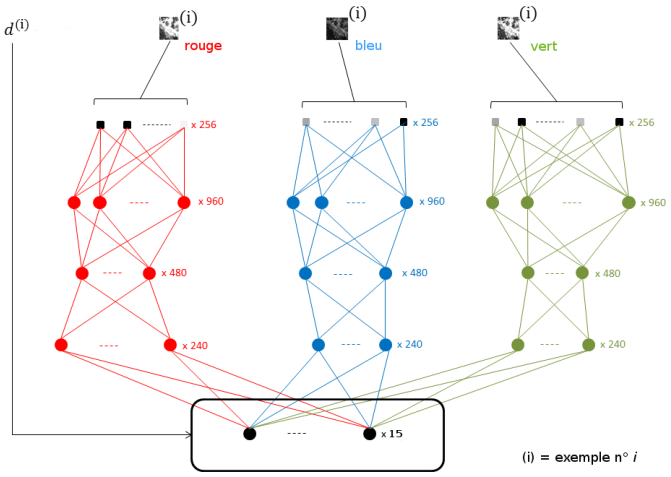

FIGURE 2 – Architecture globale du DBN à 3 canaux avec contrainte de parcimonie pour encoder l'information locale.

de profondeurs quantifiées, cela revient à un problème de classification multi-classes, que nous résolvons par un réseau de neurones adapté, composé de trois couches d'auto-encodage et d'un classifieur *softmax*. La création de l'ensemble d'apprentissage est expliquée dans la partie 3.

Le flou local peut être caractérisé par les corrélations entre pixels voisins. [6] montre que ces corrélations peuvent être apprises à partir de données non-étiquetées grâce à un encodage parcimonieux et des Machines de Boltzmann Restreintes (RBMs) empilées. Une RBM est un graphe à 2 couches bipartite et non-orienté qui consiste en un ensemble de variables observées $\mathbf{v}$, de variables cachées $\mathbf{h}$ et des connexions symétriques entre ces deux couches représentées par une matrice de poids $\mathbf{W}$.

Nous utilisons un réseau avec 3 RBMs empilées pour chacun des canaux R, G et B (cf. Fig. 2). Pour la première RBM, nous modélisons les observations constituées d'une fenêtre R,G ou B par des unités linéaires avec bruit Gaussien, et la fonction d'énergie s'écrit :

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{N_v}\sum_{j=1}^{N_h}\frac{v_i}{\sigma_i}W_{ij}h_j + \sum_{i=1}^{N_v}\frac{(v_i-a_i)^2}{2\sigma_i^2} - \sum_{j=1}^{N_h}h_j, b_j \quad (1)$$

où σ_i est l'écart-type du bruit de l'unité i. Chaque fenêtre est pré-normalisée dans $[0,1]$ avec moyenne à 0.5. Pour les couches suivantes, les variables observées sont les sorties de la RBM précédente, donc binaires. Alors la fonction d'énergie s'écrit :

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{N_v}\sum_{j=1}^{N_h}v_iW_{ij}h_j - \sum_{i=1}^{N_v}v_ia_i - \sum_{j=1}^{N_h}h_jb_j. \quad (2)$$

Chaque RBM est entraînée sur un ensemble d'apprentissage $[\mathbf{v}^{(1)},...,\mathbf{v}^{(m)}]$ de manière à optimiser la vraisemblance $\log P(\mathbf{v})$ de cet ensemble. Nous contraignons de plus les variables cachées à être éparses en pénalisant l'écart de leur probabilité d'activation q à une probabilité cible p (0.02 pour la première RBM, 0.04 pour la deuxième et 0.08 pour la dernière), ce qui

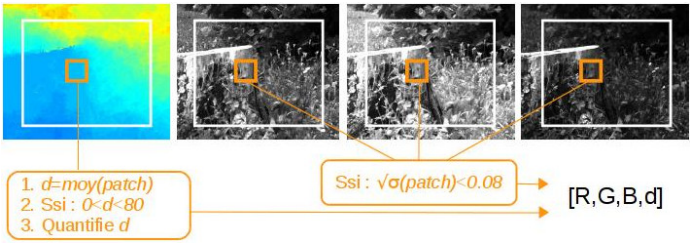

FIGURE 3 – Schéma de sélection de fenêtres (*patch* $[R, G, B, d]$ dans un triplet d'images RGB chromatique et la carte de profondeur obtenue par stéréo.

revient à rajouter un terme de régularisation basé sur l'entropie croisée de p et q :

$$C(p, q) = -p \log(q) - (1 - p) \log(1 - q). \quad (3)$$

La fonction objective pour l'entraînement non-supervisé d'une RBM s'écrit alors (avec δ le coût de parcimonie, typiquement fixé à $\frac{1}{p}$) :

$$min_{\mathbf{W},\mathbf{a},\mathbf{b}} \quad -\sum_{i=1}^{m} \log \left(\sum_{\mathbf{h}} P(\mathbf{v}^{(i)}, \mathbf{h}^{(i)} \right) + \delta C(p, q), \quad (4)$$

Eq. 4 est optimisée de manière standard par *divergence contrastive* [5]. Enfin un classifieur multi-classe basé sur une fonction non-linéaire softmax combine les sorties des 3 encodeurs et les classes de profondeurs associés aux fenêtres initiaux. Les poids de la couche softmax sont d'abord entraînés de manière directe à la suite des RBMs. Dans un deuxième temps, un apprentissage fin de tous les poids du réseau est réalisé par rétro-propagation.

3 Génération d'un ensemble d'apprentissage

Base d'images RGB-D Des données avec vérité-terrain sont nécessaires pour apprendre la relation entre les images RGB de la caméra chromatique et la profondeur. À cette fin, nous avons combiné la caméra avec flou chromatique présentée en section 2.1 avec deux caméras conventionnelles pour produire les cartes de profondeur par stéréo. Les données ont été acquises en extérieur en évitant les flous de bougé. Dans cette étude, la disparité entre les images stéréo est estimée par flot optique [9]. La carte de profondeur stéréo est ensuite projetée dans la géométrie de la caméra chromatique et interpolée de manière à obtenir un triplet d'images RGB et la carte de profondeur de même résolution (cf. Fig. 3). Une portion de l'ensemble d'images RGB-D est mise à part comme ensemble de test de la méthode (et n'entre pas dans l'apprentissage).

Ensemble d'apprentissage de fenêtre $\{\mathbf{RGB}, d\}$ Pour éviter les déformations du flou les plus critiques au bord de l'image dues aux aberrations optiques, les fenêtres 16×16 sont sélectionnés dans une zone centrée de l'image de taille 960×800 (cf. Fig. 3). Ces fenêtres sont retenues si elles sont suffisamment texturés pour contenir une information de flou (écart-type de l'intensité > 0.08). Pour chaque fenêtre, on retient sa distance moyenne qui est quantifiée pour former plusieurs classes d. Pour tenir compte des spécifications optiques de la caméra, le pas est de 50 cm entre 1 et 5 m, 1 m entre 5 et 10 m, et une classe pour les fenêtres au-delà de 10 m, soit 15 classes au total. Ces conditions sont synthétisées dans la Figure 3 et permettent d'obtenir un ensemble d'environ 830000 couples $\{\text{RGB}, d\}$.

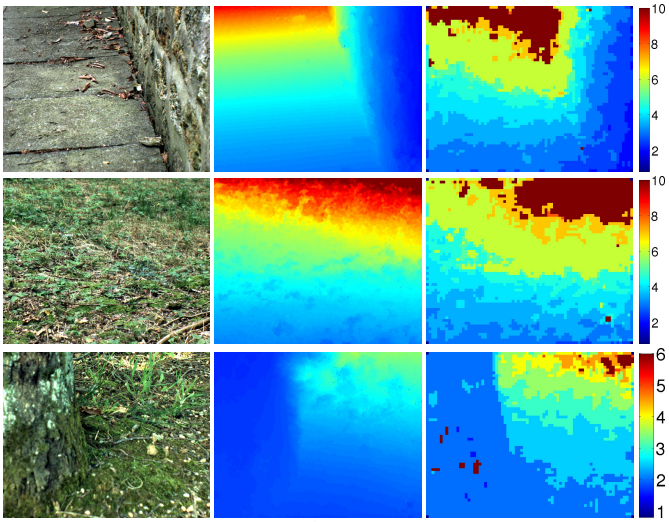

FIGURE 4 – Exemples de reconstructions de cartes de profondeur par DBN parcimonieux. (a) Image couleur, (b) Carte de profondeur stéréo (vérité-terrain), (c) Notre approche.

4 Expériences

4.1 Analyse qualitative

La Figure 4 montre quelques exemples de cartes de profondeur obtenues par notre DBN parcimonieux à partir d'images de l'ensemble de test. Ces résultats montrent qu'en dépit d'erreurs résiduelles à grande distance, les estimations de notre algorithme sont qualitativement proches de celles de la stéréo, dans la plage de fonctionnement de la caméra $[1\ m\ :\ 5\ m]$: les graduations de profondeurs sont similaires et l'on distingue bien les discontinuités de profondeurs liées à des objets (un coin de mur en Figure 4 (a) et un arbre en Figure 4 (c)).

Fig. 6(b-d) compare des estimations de cartes de profondeur obtenues avec différentes approches : stéréo, notre méthode (DBN parcimonieux à 3 canaux) et une méthode de DFD non supervisée [12], pour laquelle le flou de défocalisation a été calibré sur l'axe. Les cartes obtenues par [12] sont correctes au centre de l'image mais déformées hors-axe. Corriger ce défaut nécessiterait une calibration coûteuse en temps d'opérateur du flou sur toute l'image, alors que les cartes estimées par le DBN sont directement valides sur toute l'image. De plus, les temps de calculs sont de 2 s pour [12] contre 0.9 s pour le DBN sur un processeur Intel Xeon CPU E5-1603@ 2.80 GHz, à nombre de classes et tailles d'image et de fenêtre égaux.

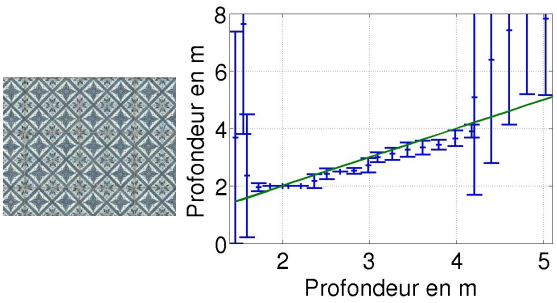

FIGURE 5 – Scène et courbe des erreurs moyennes et écarts-types en fonction de la profondeur.

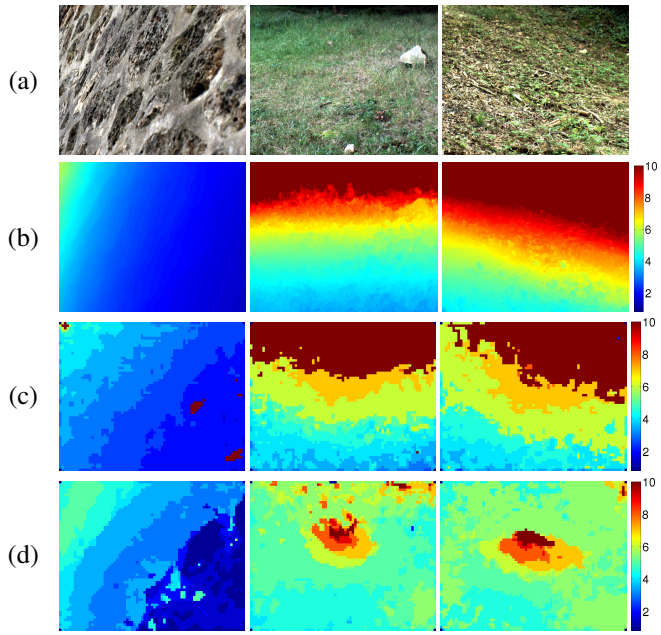

FIGURE 6 – (a) Images couleurs. (b) à (d), cartes de profondeur obtenues par (de haut en bas) : stéréo, notre approche, méthode non supervisée [12].

4.2 Analyse quantitative

Nous avons évalué la profondeur estimée par notre caméra et le DBN sur plusieurs scènes texturées, placées à différentes distances de la caméra connues à l'aide d'un télémètre. Fig. 5 montre un exemple de résultat. Entre 2 et 4 m le biais d'estimation varie entre 1 et 30 cm, et l'écart-type moyen est de l'ordre de 20 cm, ce qui est acceptable compte tenu du pas de quantification de 50 cm des classes du DBN. En-dessous de 2 m l'estimation est faussée : le rayon de l'étalement du flou est trop proche de la taille des fenêtres. Au-dessus de 4 m, le biais reste faible mais la variation des estimations est beaucoup plus forte, pour deux raisons : la distance entre deux classes consécutives augmente (1 m) et la variation de flou diminue. Notons que le choix de l'algorithme d'estimation de la disparité entre les images stéréo peut avoir une influence sur les performances de l'algorithme proposé. Une comparaison des performances de notre approche, en utilisant d'autres algorithmes de la littérature tels que [3] pour le calcul de disparité constitue une perspective de ce travail.

5 Conclusion

Nous avons présenté une nouvelle méthode pour l'estimation de profondeur utilisant une optique chromatique en combinaison avec un réseau de neurone profond. Ce *Deep Belief Network* parcimonieux a été construit en utilisant un large ensemble d'images chromatiques associées à une vérité terrain obtenue avec un banc stéréo. Les tests de robustesse montrent que notre algorithme estime des cartes de profondeur avec une précision de l'ordre de 20 cm dans une plage de 2 à 4 m, ce qui permet d'envisager des applications de types détection et évitement d'obstacles. Les avantages de l'approche résident dans l'évitement d'une procédure de calibration coûteuse en temps d'opérateur (remplacée par un apprentissage sur un jeu de données facile à acquérir) et des temps d'estimation divisés par 2. Ces travaux montrent que la combinaison d'une optique non conventionnelle avec un apprentissage sur de grands jeux de données ouvre la voie à de nouveaux capteurs 3D rapides, compacts et passifs.

Références

[1] A. CHAKRABARTI et T. ZICKLER : Depth and deblurring from a spectrally varying depth of field. *In ECCV*, 2012.

[2] D. EIGEN, C. PUHRSCH et R. FERGUS : Depth map prediction from a single image using a multi-scale deep network. *In NIPS*, 2014.

[3] A. GEIGER, M. ROSER et R. URTASUN : Efficient large-scale stereo matching. *In Asian Conference on Computer Vision (ACCV)*, 2010.

[4] R. I. HARTLEY et A. ZISSERMAN : *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN : 0521540518, second édition, 2004.

[5] G. E. HINTON : Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8), 2002.

[6] H. LEE, E. CHAITANYA et A. Y. NG : Sparse deep belief net model for visual area v2. *NIPS*, 2007.

[7] A. LEVIN, R. FERGUS, F. DURAND et W. T. FREEMAN : Image and depth from a conventional camera with a coded aperture. *ACM Trans. on Graphics*, 26, 2007.

[8] A. P. PENTLAND : A new sense for depth of field. *IEEE Trans. on PAMI*, 9, 1987.

[9] A. PLYER, G. LE BESNERAIS et F. CHAMPAGNAT : Massive parallel lucas kanade optical flow for real time video processing applications. *J. of Real-Time Image Processing*, 2014.

[10] A. SAXENA, H. S. CHUNG et A.Y. NG : Learning depth from single monocular images. *Adv. in Neural Information Processing Systems*, 2005.

[11] D. SCHARSTEIN et R. SZELISKI : A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. of Comp. Vision*, 47, 2002.

[12] P. TROUVÉ, F. CHAMPAGNAT, J. SABATER, T. AVIGNON, G. LE BESNERAIS et J. IDIER : Passive depth estimation using chromatic aberration and a depth from defocus approach. *Applied Optics*, 52(29), 2013.