Depth from Defocus en configuration stéréoscopique - Apports et application à l'imagerie 3D infra-rouge

P. TROUVÉ-PELOUX, B. BUAT, F. CHAMPAGNAT, G. LE BESNERAIS, C. COUDRAIN et G. DRUART

ONERA, Université Paris-Saclay, FR-91123 Palaiseau, France
pauline.trouve@onera.fr

Résumé – Nous présentons une nouvelle méthode d'estimation de profondeur, qui repose sur l'utilisation de deux caméras en configuration stéréoscopique avec chacune une mise au point différente. La profondeur est estimée via un critère dérivé d'un maximum de vraisemblance qui intègre conjointement des informations de disparité et les flous de défocalisation de chaque caméra. Nous étudions les apports de cette approche, notamment sur des scènes à motifs répétitifs, par rapport à la stéréoscopie classique sur des images issues de caméras opérant dans le visible. Enfin nous montrons des exemples de résultats sur des images IR thermique acquises en extérieur.

Abstract – We present a new method for depth estimation, based on a stereoscopic camera with various camera focus setting. Depth is estimated using a criterion derived from a maximum likelihood estimator, which jointly analyses the data likelihood with respect to the disparity and the defocus blur of each camera. Benefit of this approach is studied, in particular for scene having repetitive patterns with respect to classical stereoscopy, then we present experimental results on outdoor scenes from a real infra-red stereoscopic system.

1 Introduction

Parmi les différentes approches de mesure 3D passives, les caméras stéréoscopiques sont aujourd'hui couramment utilisées, pour une grande variété de domaines d'application tels que la robotique autonome, la défense ou l'inspection industrielle. Si la stéréoscopie est aujourd'hui bien maîtrisée d'un point de vue algorithmique, elle repose néanmoins sur la possibilité d'associer des pixels entre les deux images [4]. La qualité de la mesure 3D peut donc être dégradée lorsque la scène présente une répétition d'un même motif, ce qui se produit notamment pour des objets mono-directionnels tels que des grillages, câbles ou tuyaux. Cette limitation est d'autant plus grande en imagerie infrarouge (IR) thermique, où les objets sont en général moins texturés que dans le visible.

Par ailleurs, les approches de *Depth from Defocus* (DFD) reposent sur une estimation locale du flou de défocalisation pour en déduire la profondeur [5]. Ces approches ne souffrent donc pas d'une répétition d'un motif dans l'image. La difficulté algorithmique provient en revanche du fait que la scène est inconnue. Cette difficulté a mené notamment au développement d'approches basées sur l'estimation du flou via des critères de vraisemblance marginalisée par rapport à la scène, dans un contexte bayésien [10]. Enfin, la précision du DFD est dépend de son ouverture [7] dont la valeur reste limitée par des contraintes de faisabilité.

Dans cet article nous proposons de fusionner les approches de stéréoscopie et de DFD en utilisant deux caméras en configuration stéréoscopique, mais présentant un réglage de la mise au point différent. Nous proposons une approche algorithmique originale qui permet d'estimer la profondeur conjointement en fonction d'indices liés à la parallaxe et au flou de défocalisation de chaque caméra. Nous étudions ensuite l'intérêt de cette approche, que nous appelons SDFD pour *stereoscopic depth from defocus* pour des scènes à motifs répétitifs, par rapport à la stéréoscopie classique. Enfin, une validation expérimentale est ensuite menée en infrarouge thermique.

2 Etat de l'art

Plusieurs travaux de la littérature se sont intéressés à la fusion entre DFD et stéréoscopie. Dans [1] le flou relatif entre les deux images acquises par un banc stéréoscopique avec une mise au point différente est décrit par un polynôme de degré 2, en se basant sur un modèle géométrique du flou de défocalisation. La profondeur est obtenue par une optimisation alternée entre disparité et flou relatif. A chaque itération, le modèle polynomial est imposé et ses coefficients sont ré-estimés. Plus proche de nos travaux, dans la référence [9] la profondeur est estimée à partir d'une erreur de reconstruction pixel à pixel: à chaque hypothèse de profondeur est associée une estimation de la scène par déconvolution dans l'espace de Fourier basée sur un modèle de bruit blanc. Cette scène est alors utilisée pour re-prédire les images des deux caméras ce qui permet d'identifier localement l'hypothèse qui minimise l'écart entre images acquises et images re-prédites. La validation expérimentale reste cependant limitée à la translation d'une même caméra pour produire les deux acquisitions. Enfin notons que des méthodes de fusion de stéréoscopie et de DFD à partir de réseaux de neurones ont été proposés récemment [2], mais nécessitent des bases de données de grandes tailles, qui sont simulées.

A la différence des références citées précédemment, grâce à un formalisme bayésien, nous proposons une définition directe de la vraisemblance des données vis à vis de la profondeur qui prend en compte conjointement la disparité et le flou de défocalisation, à partir d'un modèle de scène naturelle. Notre critère est calculé sur des fenêtres, avec une adaptation locale du rapport signal à bruit.De plus, nous investiguons l'apport de l'approche SDFD par rapport à la stéréoscopie classique, notamment sur des scènes à motifs répétitifs, dans le visible. Enfin, à notre connaissance il n'existe pas de travaux expérimentaux de DFD en imagerie IR, ce que nous présentons ici.

3 Depth from Defocus stéréoscopique

L'approche de SDFD que nous proposons est dérivée de l'approche de DFD multi-images présentée dans [10]. Nous généralisons tout d'abord son principe dans le cas de $k \geq 1$ images et décrivons ensuite son extension dans le cas d'une configuration stéréoscopique.

3.1 Depth from Defocus multi-image

Pour k images d'une scène acquise par une même caméra avec différents réglages, le modèle de formation d'images dans une région supposée à profondeur constante p s'écrit :

$$\boldsymbol{Y} = H\boldsymbol{X} + \boldsymbol{B}, H = \left[\begin{array}{cccc} H_1(p)^t & H_2(p)^t & ... & H_k^t(p) \end{array} \right]^t, \quad (1)$$

avec $\boldsymbol{Y} = [\boldsymbol{y}_1^t \ \boldsymbol{y}_2^t \ ... \ \boldsymbol{y}_k^t]^t$ où les vecteurs $\boldsymbol{y}_s$ sont formés par la concaténation des N pixels des fenêtres locales issues des k images. $\boldsymbol{X}$ est un vecteur de dimension M qui concatène les valeurs d'intensité de la scène imagée par les caméras. $\boldsymbol{B}$ correspond au bruit d'acquisition modélisé par un processus aléatoire gaussien centré de variance σ_b^2. Chaque matrice $H_i(p)$ est une matrice de convolution de taille $N \times M$ relative à la fonction d'étalement de point (FEP) de l'image i, qui dépend de la profondeur de la scène p et du réglage de la caméra. Nous cherchons à estimer la profondeur p alors que la scène $\boldsymbol{X}$ est inconnue. Dans [10] est définie une vraisemblance marginale où la scène a été intégrée grâce à un a priori de la forme:

$$p(\boldsymbol{X};\sigma_x^2) \quad \propto \quad exp\left(-\frac{\|D\boldsymbol{X}\|^2}{2\sigma_x^2}\right). \quad (2)$$

D correspond à la concaténation verticale des matrices de convolution associées aux vecteurs $[-1 \ \ 1]$ et $[-1 \ \ 1]^t$. La vraisemblance marginale s'écrit alors:

$$p(\boldsymbol{Y}|H(p),\sigma_b^2,\alpha) \propto |\sigma_b^{-2}P(\alpha,p)|_+^{\frac{1}{2}} e^{-\frac{\boldsymbol{Y}^t P(\alpha,p)\boldsymbol{Y}}{2\sigma_b^2}}, \quad (3)$$

$$P(\alpha,p) = I - H(p)(H(p)^t H(p) + \alpha D^t D)^{-1} H(p)^t.$$

$\alpha = \sigma_b^2/\sigma_x^2$ est un terme de régularisation permettant de tenir compte de la variation locale du rapport signal à bruit et I la matrice identité. Enfin, on peut montrer que maximiser la vraisemblance marginale revient à minimiser le terme :

$$GL_k(p,\alpha) = \boldsymbol{Y}^t P(\alpha,p)\boldsymbol{Y}|P(\alpha,p)|_+^{-1/(kN-1)}, \quad (4)$$

où $|A|_+$ correspond au produit des valeurs propres non nulles de la matrice A. A partir d'une famille de profondeurs potentielles, la profondeur est alors obtenue en minimisant GL_k par rapport à α puis en minimisant $GL_k(p,\widehat{\alpha})$ sur les valeurs discrètes de p.

3.2 Cas d'une configuration stéréoscopique

Dans ce cas $k = 2$ avec un décalage entre les caméras. L'enjeu est de trouver les fenêtres des images gauche et droite ($\boldsymbol{y}_G$ et $\boldsymbol{y}_D$) qui sont associées à la même partie de la scène. Or, cette association dépend de la profondeur. Nous proposons donc d'ajouter une contrainte supplémentaire au critère GL_k de (6) liée à la disparité entre les deux images.

Soit $\boldsymbol{y}_{G,(i_0,j_0)}$ la fenêtre de l'image de gauche centrée sur (i_0,j_0) dont on veut estimer la profondeur. En considérant que les images sont rectifiées, la fenêtre $\boldsymbol{y}_{D,(i,j)}$ correspondante dans l'image de droite est également sur la ligne (i_0). Le numéro de colonne j est lié à une hypothèse de profondeur par la relation classique en stéréoscopie :

$$disp(p) = \frac{Bf}{p}, \ j(p) = j_0 - disp(p). \quad (5)$$

Où f est la focale des caméras en pixel et B l'écart entre les caméras. Notons qu'en pratique $j(p)$ est arrondi à l'entier le plus proche, afin de ne pas introduire d'interpolation entre les fenêtres $\boldsymbol{y}_G$ et $\boldsymbol{y}_D$, ce qui qui pourrait modifier l'information de flou. Ainsi, on peut écrire :

$$GL_{SDFD}(p,\alpha) = \boldsymbol{Y}(p)^t P(\alpha,p)\boldsymbol{Y}(p)|P(\alpha,p)|_+^{-1/(2N-1)} \quad (6)$$

avec $\boldsymbol{Y}(p) = [\boldsymbol{y}_G^t \boldsymbol{y}_D(p)^t]^t$. Ainsi, pour une hypothèse de profondeur, deux contraintes liées à la profondeur sont évaluées simultanément à l'aide de (6) : d'une part la cohérence de l'appariement entre les fenêtres $\boldsymbol{y}_G$ et $\boldsymbol{y}_D$ et d'autre part la cohérence entre les données et le flou de défocalisation.

A titre d'exemple, nous considérons un banc de deux caméras chacune munie d'un objectif de focale 16 mm, d'ouverture N=2.6 avec des pixels de 4.5 μm. La caméra de droite a une mise au point à l'infini, la caméra de gauche une mise au point à 1.5 m. A partir d'une scène texturée considérée comme plane, illustrée à la figure 1, nous simulons les images issues par ce banc stéréo, lorsque celle-ci est placé à 2.5 m. Nous considérons alors deux cas : $B = 0$, sans effet de parallaxe entre les images et le cas où $B = 60$ mm, avec parallaxe. Nous évaluons respectivement le critère GL_2 et GL_{SDFD} pour ces deux configurations pour différentes hypothèses de profondeur pour une fenêtre de l'image de taille 21×21. Les courbes obtenues sont présentées à la figure 1. Les deux courbes atteignent un minimum proche de la valeur de profondeur attendue 2.5 m. La courbe GL_{SDFD} apparaît nettement plus piquée que la courbe GL_2 aux alentours du minimum ce qui indique que la contrainte de disparité ajoute un gain en précision sur l'estimation de la profondeur. Ceci n'est pas surprenant car la configuration stéréoscopique permet de bénéficier de deux indices sur la profondeur: la parallaxe et le flou. En revanche, on peut noter qu'il existe plus de minima locaux dans la courbe de SDFD.

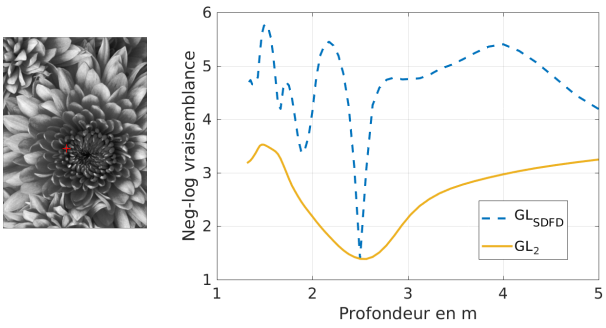

FIG. 1: *Gauche : image de gauche simulée. Droite : Courbes de GL_{SDFD} (B=60 mm) et GL_2 (B=0) (échelle logarithmique) sur la fenêtre de donnée centrée sur la croix rouge.*

4 Cas de scènes à motifs répétitifs

Comme discuté en introduction, les systèmes stéréoscopiques souffrent de la présence d'objets à motifs répétitifs le long des lignes épipolaires. Dans cette section, nous comparons les performances d'un algorithme de stéréoscopie classique, par rapport à l'approche SDFD sur une scène présentant ce type de motif sur des données réelles, dans le visible.

4.1 Images réelles

4.1.1 Configuration expérimentale et étalonnage

Nous effectuons des tests sur des données réelles à partir d'un banc de caméras dans le visible munies d'objectifs identiques à ceux utilisés en section 3.2. L'étalonnage des paramètres intrinsèques et extrinsèques des caméras est réalisé avec des outils ROS [1], à partir de mires circulaires, car elles sont moins sensibles au flou de défocalisation que les mires de type damier. Les couples $\{disp(p), H(p)\}$ sont obtenus en utilisant la mire d'estimation de FEP proposée par [3] avec la disparité mesurée sur la mire à partir de SGBM [4]. La profondeur est obtenue à l'aide de la relation (5) et des paramètres issus de l'étalonnage géométrique.

4.1.2 Résultats expérimentaux

La figure 2 présente les résultats d'estimation de profondeur obtenus avec les algorithmes SDFD, SGBM et le critère GL_1 sur les mêmes données, pour deux types de scènes différentes : une grille placée devant un objet texturé et une tige placée devant une scène plus naturelle. Pour SDFD et GL_1, la profondeur est estimée sur des fenêtres de taille 21x21, avec 50% de recouvrement et un filtrage médian de taille 3x3 est appliqué en post-traitement. Pour GL_1, la profondeur est estimée à partir de l'image de droite uniquement qui grâce à une mise au point à l'infini, ne souffre pas de l'ambiguité sur le flou de défocalisation comme discuté dans [10]. Ceci revient à n'utiliser

1. http://www.ros.org/

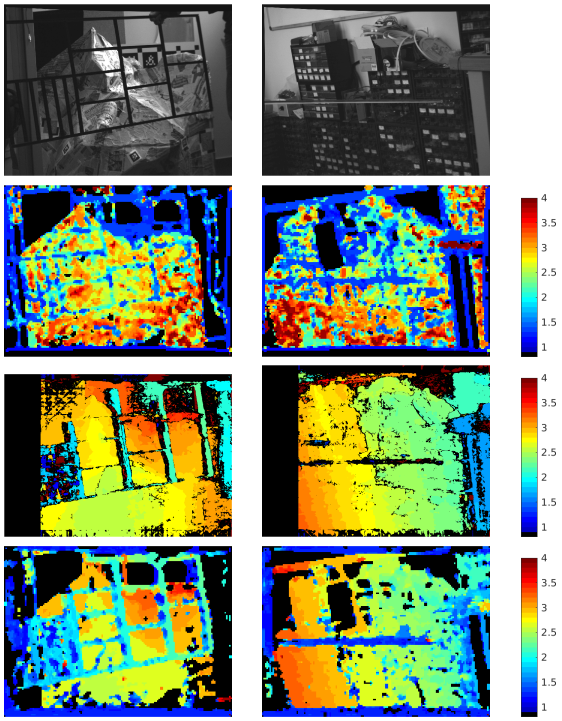

FIG. 2: *Images de la caméra de gauche (ligne1) et cartes de profondeur en m en utilisant le critère GL_1 sur l'image de droite (ligne 2), SGBM (ligne 3) et SDFD (ligne 4).*

que l'information de flou. Le label noir correspond aux fenêtres pour lequel le paramètre α est supérieur à un seuil fixé empiriquement. L'algorithme de stéréoscopie classique fonctionne bien sur les régions texturées de manière isotrope, bien que les mises au point des caméras soient différentes. Par contre il ne permet pas une bonne estimation de la position de la grille ou de la tige. Le critère GL_1 permet de distinguer en partie ces objets mais du bruit et de nombreuses zones aberrantes. En revanche, les cartes SDFD permettent de bien distinguer les différents objets de la scène avec peu de bruit dans la carte de profondeur. Notons cependant que la résolution en profondeur du résultat SDFD est limitée par le jeu de profondeurs calibrées.

Ces résultats montrent que l'utilisation conjointe du flou de défocalisation et de la disparité permet de mieux distinguer les objets à motifs répétitifs, tels que des grilles, tiges ou câbles qu'avec les algorithmes classiques de stéréoscopie.

5 Application à l'imagerie 3D IR

5.1 Configuration expérimentale et étalonnage

Deux caméras micro-bolomètre avec un pas pixel de $17\mu m$ et munies d'objectifs 25 mm ouvert à f/1.2 sont alignées sur un banc. Les mises au point des caméras sont fixées autour de 2.5 m pour la caméra de gauche et à environ 5 m pour la caméra de droite. Au contraire de l'imagerie visible, en IR thermique il n'est pas facile d'imprimer des mires d'étalonnage texturées telles que celle proposée dans [3]. Nous proposons donc d'utili-

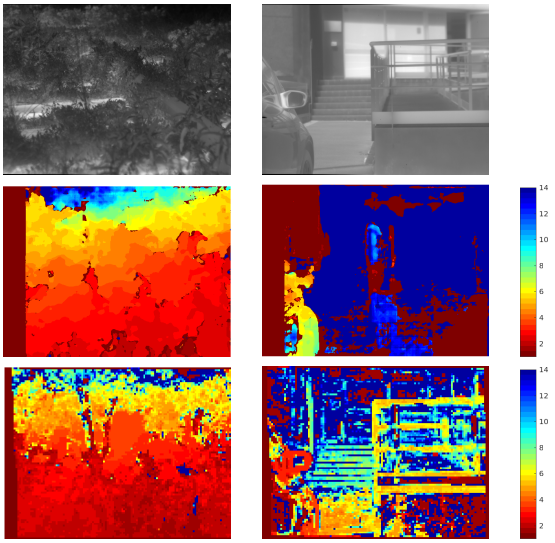

FIG. 3: *Ligne 1: images réelles en IR thermique, cartes de profondeur en m par SGBM (ligne 2) et SDFD (ligne 3).*

ser comme mire une plaque trouée placée devant un corps noir. Cette mire sert simultanément pour l'étalonnage géométrique et l'étalonnage du flou de défocalisation: les centres des trous sont estimés sur une séquence d'images pour en déduire une estimation des paramètres intrinsèques et extrinsèques des deux caméras. Le flou de défocalisation est calibré à l'aide d'une approche de type lame de couteau [6] que nous généralisons au cas de cercles. Ceci revient à modéliser les FEP par des gaussiennes paramétrées par leurs écarts-types.

5.2 Résultats expérimentaux

La figure 3 présente deux exemples d'images acquises avec les caméras IR : une image de végétation et une image dans un contexte urbain. Nous présentons ensuite les cartes de profondeurs obtenues en utilisant SGBM ou SDFD sur ces images. Dans le cas de scènes IR, qui sont généralement moins texturées que les images visibles, l'approche SDFD donne des résultats globalement plus satisfaisants que la stéréoscopie. Ces résultats confirment également que l'approche SDFD permet d'estimer la profondeur des objets fins et/ou à motifs répétitifs tels que les tiges de la végétation dans l'image de gauche, ou bien les rambardes dans l'image de droite.

6 Conclusion et perspectives

Nous avons présenté une nouvelle méthode d'estimation de profondeur, combinant les approches de *Depth from Defocus* et de stéréoscopie. Cette approche permet d'estimer la profondeur des objects à motifs répétitifs qui constituent une difficulté en stéréoscopie, ce qui nous semble prometteur en imagerie IR thermique, où les objets sont moins texturés que dans le visible. Cependant l'approche de SDFD proposée repose sur un critère de vraisemblance calculé pour un ensemble discret d'hypothèses de profondeur, ce qui réduit la précision d'estimation par rapport aux traitements de stéréoscopie qui sont subpixelliques. SDFD pourrait donc être utilisé en initialisation des algorithmes de stéréoscopie classiques ou seulement pour les objets donnant des résultats aberrants avec ces méthodes.

Dans la suite de ce travail, nous nous intéresserons également à l'utilisation d'optiques non conventionnelles en configuration stéréoscopique, telles qu'une pupille codée (en IR et en visible) ou une optique chromatique [8, 10] (dans le visible) afin d'améliorer les précisions d'estimation de profondeur. Enfin, la question du réglage optimal de la mise au point des deux caméras pour maximiser les performances du système peut se poser. Une des perspectives de ce travail est donc de modéliser la performance d'un système de SDFD en fonction de ses paramètres optiques et de traitement ce qui permettrait de mener l'optimisation conjointe optique/traitement du système pour une application donnée.

7 Remerciements

Ce travail a été réalisé dans le cadre du partenariat de recherche entre SNCF Réseau, Altametris et l'ONERA.

Références

[1] C. Chen, H. Zhou, and T. Ahonen. Blur-aware disparity estimation from defocus stereo images. In *Proc. of the IEEE International Conference on Computer Vision*, 2015.

[2] Z. Chen, X. Guo, S. Li, X. Cao, and J. Yu. A learning-based framework for hybrid depth-from-defocus and stereo matching. *arXiv preprint arXiv:1708.00583*, 2017.

[3] M. Delbracio, P. Musé, A. Almansa, and J-M Morel. The non-parametric sub-pixel local point spread function estimation is a well posed problem. *Int. Journal of Computer Vision*, 2012.

[4] H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[5] A. P. Pentland. A new sense for depth of field. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9, 1987.

[6] S. E Reichenbach, S. K Park, and R. Narayanswamy. Characterizing digital image acquisition devices. *Optical Engineering*, 30(2), 1991.

[7] Y. Y. Schechner and N. Kiryati. Depth from defocus vs. stereo: how different really are they? In *Proc. of International Conference on Pattern Recognition*, 1998.

[8] A. Sellent and P. Favaro. Which side of the focal plane are you on? In *IEEE International Conference on Computational Photography*, 2014.

[9] Y. Takeda, S. Hiura, and K. Sato. Fusing depth from defocus and stereo with coded apertures. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[10] P. Trouvé, F. Champagnat, G. Le Besnerais, J. Sabater, T. Avignon, and J. Idier. Passive depth estimation using chromatic aberration and a depth from defocus approach. *Applied optics*, 52(29), 2013.