Extraction du contour de la main à l'aide d'une caméra 3D chromatique

Pauline TROUVÉ-PELOUX, Frédéric CHAMPAGNAT, Guy LE BESNERAIS, Martial SANFOURCHE

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

Résumé – Nous proposons une nouvelle approche d'extraction du contour de la main à partir d'une caméra 3D monovoie passive. La capacité 3D de cette caméra repose sur l'estimation de la profondeur grâce au flou de défocalisation, amélioré par la présence d'aberration chromatique dans l'optique. Le contour de la main est extrait à l'aide d'une segmentation de la carte de profondeur obtenue par la minimisation d'une énergie définie à l'aide d'un modèle de champ de Markov aléatoire sur l'image. Nous montrons des résultats expérimentaux obtenus avec l'approche proposée.

Abstract – We propose a new approach for hand shape segmentation using a passive monocular 3D camera. The camera depth estimation is based on the estimation of the defocus blur improved by the use of an optic with enhanced chromatic aberration. The hand shape is extracted from a segmentation of the depth map formulated in terms of an energy minimisation using a Markov random field model on the image. We show experimental examples obtained with the proposed approach.

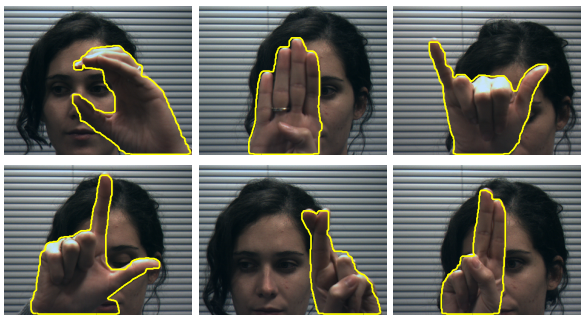


FIGURE 1 – Résultats d'estimation du contour de la main obtenus à l'aide d'une caméra chromatique.

1 Introduction

L'interaction homme machine par la vision se développe de plus en plus notamment grâce à l'intégration de caméras miniaturisées dans les appareils grand public. Or cette interaction nécessite d'estimer la posture de la main, opération délicate car elle recquiert l'estimation de près de 20 paramètres, avec un temps de calcul compatible avec les applications envisagées et dans un environnement non contrôlé [13, 4]. Un problème plus simple est celui de l'estimation partielle de la posture de la main [4], dans laquelle seulement la position et l'orientation des doigts de la main sont estimés. Dans ce domaine, l'enjeu principal est donc de localiser et de segmenter le contour de la main. Nous proposons ici de traiter ce problème à l'aide d'une caméra 3D monovoie passive. Sa capacité 3D repose sur le principe d'estimation de profondeur grâce au flou de défocalisation (DFD pour *Depth from Defocus*) [14], ce qui lui permet d'être utilisée à l'extérieur comme à l'intérieur et d'être plus compact qu'un système stéréoscopique. L'estimation de profondeur par DFD est améliorée à l'aide d'un optique chromatique, qui permet d'avoir un flou de défocalisation différent suivant les canaux rouge, vert et bleu de l'image couleur. Nous considérons ici le cas d'une personne qui présente sa main devant elle, en face de la caméra et présentons un ensemble de traitements des images issues de la caméra chromatique permettant d'extraire le contour de la main, à l'aide d'une segmentation de la carte de profondeur via la minimisation d'une énergie définie par un modèle de champ de Markov. Nous présentons des résultats expérimentaux obtenus par l'approche proposé et discutons des perspectives de ce travail.

2 Etat de l'art

Les méthodes de segmentation de la main à partir d'images 2D utilisent le plus souvent des informations sur la couleur, la soustraction du fond, le mouvement, le pistage ou de l'apprentissage [13, 4]. Cependant, ces approches peuvent être perturbées par des changements d'illumination de la scène, par la présence de plusieurs objects mobiles, ou par la nécessité d'avoir une grande base d'apprentissage. Une information 3D sur la scène peut permettre de surmonter ces difficultés. Notamment de très bons résultats de segmentation de la main ont été obtenus à l'aide de caméras actives telles que la Kinect ou les caméras Time of Flight[6, 17, 7], cependant ces systèmes reposent sur la projection d'un signal infrarouge, signal qui peut être perturbé par un autre appareil, ou par le rayonnement du soleil. Des approches de stéréoscopie passives ont également

été proposées [11, 10], mais cette approche augmente l'encombrement et nécessite une synchronisation précise des deux caméras.

Nous proposons de segmenter la main à l'aide d'une caméra 3D monovoie passive reposant sur le principe de DFD. Si plusieurs caméras 3D par DFD ont été développées dans la littérature, avec notamment des optiques non conventionnelles permettant d'améliorer les performances d'estimation de profondeur [8, 1, 16], à notre connaissance aucune n'a été utilisée pour une application d'extraction du contour de la main.

3 Caméra 3D chromatique

L'estimation de profondeur par *Depth from Defocus* [14] repose sur le lien entre le flou de défocalisation et la profondeur. En effet, comme illustré à la Figure 2, si un point source est placé à une distance p de la caméra, en dehors du plan de mise au point, son image, appelée fonction d'étalement du point (FEP), a une taille donnée géométriquement par la relation :

$$\epsilon = D d_{det} \left| \frac{1}{f} - \frac{1}{p} - \frac{1}{d_{det}} \right|, \qquad (1)$$

où f et la distance focale, D le diamètre de la lentille, d et d_{det} respectivement la distance entre la lentille et le détecteur.

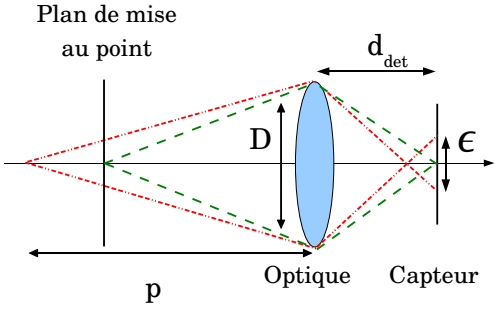
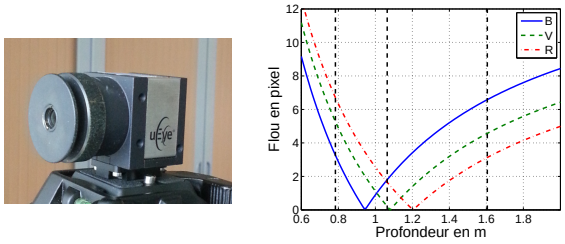


FIGURE 2 – Principe du DFD.

Nous proposons d'estimer la profondeur avec une optique dont le chromatisme axial n'a pas été corrigé. En effet, le chromatisme permet d'avoir une variation de la distance focale avec la longueur d'onde et donc des flous de défocalisation différents suivant les canaux R,V ou B de l'image couleur, comme illustré à la figure 3. Ainsi, comme discuté dans [16], contrairement à une caméra classique, il n'existe pas d'ambiguité sur la profondeur (devant ou derrière le plan de mise au point), ni de zone aveugle dans la profondeur de champ de la caméra, car il existe un triplet unique de flou par profondeur. En pratique, nous utilisons une optique chromatique ouverte à F/4 avec une focale de 25 mm dont le chromatisme axial vaut 200μm, associée à un capteur uEye 1240 dont les pixels font 5.3 μm avec une résolution de 1280×1080. Le champ de la caméra est de l'ordre de 20^o. La Figure 3 (a) montre une photographie de la caméra et (b) la variation théorique du flou de défocalisation pour les trois canaux R,V et B.



(a) (b)

FIGURE 3 – (a) Caméra. (b) Variation théorique du flou de défocalisation avec la profondeur pour les canaux R,V et B. Les lignes verticales correspondent aux profondeurs utilisées en 5.

4 Extraction du contour de la main

Notre approche consiste à obtenir une carte segmentée de la profondeur de la scène puis à la seuiller afin d'extraire un masque binaire contenant le contour de la main. La segmentation est réalisée en minimisant une énergie définie à l'aide d'un modèle de champ de Markov aléatoire par :

$$E(p) = \sum_r V(p_r) + \lambda \sum_{r,s \in N_r} f(p_r, p_s), \qquad (2)$$

où N_r est le voisinage au premier ordre du pixel r, p_r est la valeur de la profondeur estimée au pixel r. V constitue le terme d'attache aux données qui dépend de la profondeur. La fonction f est utilisée pour régulariser la carte de profondeur, en introduisant des contraintes sur les valeurs de profondeur entre pixels voisins. Les deux fonctions f et V sont décrites dans les paragraphes suivants.

4.1 Dérivation d'une vraisemblance généralisée

4.1.1 Modèle de formation d'image

La relation entre la scène et l'image acquise est traditionnellement modélisée par une convolution entre une scène nette et une FEP. Comme le fou de défocalisation varie spatialement, cette relation n'est valable que localement, sur des fenêtres de l'image pour lesquelles la profondeur est supposée constante. Dans le cas d'une caméra chromatique associée à un capteur couleur, chaque canal RVB possède un flou différent, donc une FEP différente. En utilisant le formalisme matriciel, les données peuvent être modélisées par :

$$\boldsymbol{Y}_C = H_C(p)\boldsymbol{X}_C + \boldsymbol{B}, \qquad (3)$$

où $\boldsymbol{Y}_C = [\boldsymbol{y}_R^T\, \boldsymbol{y}_V^T\, \boldsymbol{y}_B^T]^T$ et $\boldsymbol{X}_C = [\boldsymbol{x}_R^T\, \boldsymbol{x}_V^T\, \boldsymbol{x}_B^T]^T$ représentent respectivement la concaténation des pixels des scènes et des images RVB, $\boldsymbol{B}$ le bruit qui affecte les trois canaux, $H_C(p)$ est une matrice diagonale par blocs contenant les matrices de convolution associées aux FEP des trois canaux. En supposant un bruit blanc gaussien sur les images on peut écrire :

$$P(\boldsymbol{Y}_C | \boldsymbol{X}_C, \sigma_b^2) \propto \exp\left(-\frac{\|\boldsymbol{Y}_C - H_C \boldsymbol{X}_C\|^2}{2\sigma_b^2} \right). \qquad (4)$$

4.1.2 Modèle de scène

Chaque image RVB est considérée comme issue de trois scènes : $\boldsymbol{x}_R, \boldsymbol{x}_V$ et $\boldsymbol{x}_B$. Ces scènes étant partiellement corrélées nous utilisons plutôt une décomposition en luminance et chrominances telle que $\boldsymbol{X}^{LC} = [\boldsymbol{x}_l^T\ \boldsymbol{x}_{c_1}^T\ \boldsymbol{x}_{c_2}^T]^T$ défini par

$$\boldsymbol{X}_C = T \otimes \mathbf{I}_M \boldsymbol{X}^{LC} \text{ avec } T = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \end{bmatrix}, \quad (5)$$

et $\otimes$ correspond au produit de Kronecker et $\mathbf{I}_M$ est la matrice identité de taille $M \times M$. Les trois composantes luminance et chrominances peuvent alors être considérées comme indépendantes. Le modèle d'observation s'écrit alors :

$$\boldsymbol{Y}_C = H_{LC}(p)\boldsymbol{X}^{LC} + \boldsymbol{B} = H_C(p)T \otimes \mathbf{I}_M + \boldsymbol{B}. \quad (6)$$

Nous proposons d'utiliser un modèle de scène gaussien [9, 16], sur les gradients de la luminance et des chrominances qui s'expriment alors de la manière suivante :

$$P(\boldsymbol{X}^{LC}, \sigma_x^2, \mu) \propto \exp\left(-\frac{\|D_C(\mu)\boldsymbol{X}^{LC}\|^2}{2\sigma_x^2}\right), \quad (7)$$

où D_C est une matrice diagonale par blocs contenant respectivement $\mu D, D$ et D et où D est la concaténation verticale des matrices de convolutions relatives aux opérateurs de dérivations horizontaux et verticaux du premier ordre. Le paramètre μ permet de modéliser le rapport entre la variance des gradients de la luminance par rapport à ceux des chrominances, il est fixé à 0.04 [16, 2].

4.1.3 Vraisemblance généralisée

En utilisant les modèles de scène et d'image précédent, une vraisemblance marginale associée aux données peut être exprimée analytiquement [5, 16]. La maximisation de cette vraisemblance suivant le paramètre de bruit permet d'exprimer analytiquement une *vraisemblance marginalisée*, et maximiser cette vraisemblance revient à minimiser le terme :

$$GL_C(p, \alpha) = \frac{\boldsymbol{Y}_C^T \Psi(\alpha, p) \boldsymbol{Y}_C}{|\Psi(\alpha, p)|_+^{1/(3N-3)}} \quad (8)$$
$$\Psi(\alpha, p) = I_N - H_{LC}(H_{LC}^T H_{LC} + \alpha D_C^T D_C)^{-1} H_{LC}^T,$$

où $|A|_+$ est le produit des valeurs propres non nulles de A, $3N$ est la taille du vecteur $\boldsymbol{Y}_C$ et $\alpha = \sigma_b^2/\sigma_x^2$ peut être interprété comme l'inverse d'un rapport signal à bruit. Notons que le critère GL_C peut être calculé à n'importe quelle profondeur pour laquelle les FEP des trois canaux sont connues.

4.2 Fonction de régularisation

Notre objectif est d'obtenir une carte de profondeur segmentée cohérente avec les contours effectifs des objets observés par la caméra. Nous proposons donc d'utiliser comme fonction de régularisation la fonction :

$$f(p_r, p_s, \sigma) = \exp\left(-\frac{||y_g(r) - y_g(s)||^2}{2\sigma^2}\right)(1 - \delta(p_r, p_s)), \quad (9)$$

avec y_g l'image couleur convertie en niveau de gris. Si les pixels r et s ont la même profondeur, cette fonction vaut 0. Sinon, la fonction de régularisation est une gaussienne qui est maximale lorsque les niveaux de gris sont identiques et minimale lorsque l'écart entre les niveaux de gris est important. Autrement dit cette fonction autorise les ruptures des niveaux de profondeurs uniquement lorsqu'elles coïncident avec une rupture dans les niveaux de gris de l'image.

4.3 Segmentation de la carte de profondeur

En insérant les fonctions définies dans les équations (8) et (9), l'énergie (2) s'écrit alors :

$$E(p) = \sum_r GL_C(p_r, \widehat{\alpha}) + \lambda \sum_{r,s \in N_r} \exp\left(-\frac{||y_g(r) - y_g(s)||^2}{2\sigma^2}\right)(1 - \delta(p_r, p_s)), \quad (10)$$

avec $\widehat{\alpha}$ obtenue par une minimisation 1D de (8) [16]. Nous proposons de minimiser ce critère à l'aide d'un algorithme de type *graphcut*, ou courbure de graphe [15].

4.4 Extraction du contour de la main

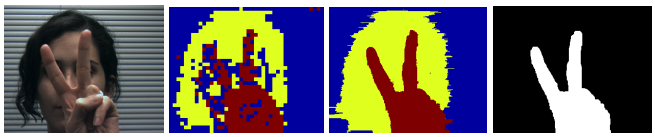
Puisque nous considérons que la main est l'objet le plus proche vu par la caméra, à partir de la carte de profondeur segmentée obtenue par minimisation de (10), un seuil sur les valeurs de profondeur permet directement d'extraire un masque binaire contenant uniquement la forme de la main segmentée et d'en extraire le contour.

5 Résultats expérimentaux

La caméra est réglée afin d'avoir les plans de mise au point des canaux R,V et B positionnés respectivement à 0.9, 1.05 et 1.2 m. Le critère (10) peut être minimisé pour un ensemble quelconque de profondeurs potentielles, cependant la scène que nous considérons est composée principalement de trois niveaux de profondeurs : la main, le visage et le fond. Pour simplifier la minimisation du critère (10), nous segmentons donc la carte de profondeur en seulement trois niveaux de profondeur : 0.7 m, 1 m et 1.6 m, représenté en noir dans la figure 3(b), correspondant approximativement aux positions de ces trois éléments de la scène par rapport à la caméra. Pour calculer le terme de vraisemblance généralisée, les FEP des trois canaux R,V et B sont calibrées à ces profondeurs l'aide de la méthode de [3].

Les Figures 4 (a) à (c) montrent respectivement un exemple d'image acquise avec la caméra et deux cartes de profondeur segmentées. Dans (b) $\sigma = 4.5 \times 10^{-4}$ et $\lambda = 0$. Ceci correspond à minimiser uniquement le terme de vraisemblance généralisée. La carte de profondeur obtenue permet de distinguer les trois plans de profondeurs, mais le résultat présente des valeurs aberrantes et elle a un aspect pixelisé, du fait du non-recouvrement des fenêtres pour lesquelles la vraisemblance généralisée est calculée. La Figure (c), obtenue cette fois avec

$\sigma = 4.5 \times 10^{-4}$ et $\lambda = 5$ montre une carte de profondeur cohérente avec la scène, et fait bien apparaître le contour de la main. La Figure 4 (d) présente le résultat de l'extraction du premier niveau de profondeur et la figure 1 la superposition du contour de la main et de l'image couleur pour différents exemples, avec les mêmes paramètres algorithmiques que pour la figure 4 (d). Pour ces différents exemples, le contour de la main est correctement estimé. Le temps de calcul total entre l'acquisition de l'image et l'extraction du contour de la main est de 0.8s avec un processeur Corei7 3930k @ 3.26 GHz et une implémentation des algorithmes en `Matlab`.



(a) (b) (c) (d)

FIGURE 4 – (a) Image acquise. Résultats de la minimisation de (10) avec : $\sigma = 4.5 \times 10^{-4}$ et (b)λ = 0, (c) λ=5 (jaune : 1m, bleu : 1.6 m, rouge : 0.7m). (d) Contour extrait.

6 Conclusion et perspectives

Nous avons présenté les premiers résultats d'extraction du contour de la main à partir d'une caméra 3D monovoie passive utilisant le principe de DFD et du chromatisme. L'approche proposée peut être utilisée à l'intérieur et à l'extérieur, avec un système compact et un mode d'acquisition simple. Cependant, nous avons pu observer parfois des erreurs de segmentation de la carte de profondeur comme illustré à la figure 6. Une perspective de ce travail est donc de faire une étude statistique approfondie des résultats de notre méthode sur une large base d'images afin de mieux caractériser les erreurs de segmentation et optimiser les paramètres de notre approche, ainsi qu'une comparaison approfondie avec l'état de l'art [6, 17, 7].

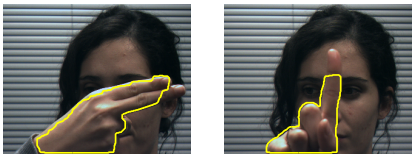


FIGURE 5 – Exemples typiques d'erreur de segmentation.

Une fois le contour de la main estimé, plusieurs applications sont possibles comme l'identification d'un geste fait avec la main, ou bien la localisation de la pointe des doigts, ce qui permettrait par exemple de faire de l'epellation digitale[10]. La Figure montre des exemples de localisation de la position des doigts, en utilisant une recherche de maxima de la courbure du contour de la main [12].

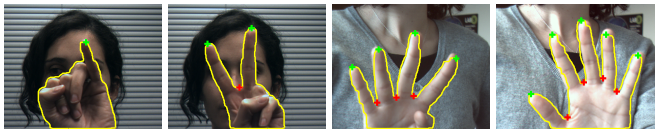


FIGURE 6 – Exemples de résultats de localisation de l'extrémité des doigts (croix vertes) et de la jonction entre deux doigts successifs (croix rouges), à partir du contour de la main.

Références

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

[2] L. CONDAT : A generic variational approach for demosaicking from an arbitrary color filter array. *In IEEE ICIP*, 2009.

[3] M. DELBRACIO, P. MUSÉ, A. ALMANSA et J.M. MOREL : The non-parametric sub-pixel local point spread function estimation is a well posed problem. *IJCV*, pages 1–20, 2012.

[4] A. EROL, G. BEBIS, M. NICOLESCU, R. D BOYLE et X. TWOMBLY : Vision-based hand pose estimation : A review. *Comp. Vision and Image Understanding*, 108(1), 2007.

[5] J. IDIER : *Approche bayésienne pour les problémes inverses*. Traité IC2, Série traitement du signal et de l'image, Hermés, Paris, nov. 2001.

[6] C. KESKIN, F. KIRAÇ, Y. KARA et L. AKARUN : Hand pose estimation and hand shape classification using multi-layered randomized decision forests. *IEEE ECCV*, 2012.

[7] H. LAHAMY et D. LICHTI : Real-time hand gesture recognition using range cameras. *In Proceedings of the Canadian Geomatics Conference, Calgary, Canada*, 2010.

[8] 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(3), 2007.

[9] A. LEVIN, Y. WEISS, F. DURAND et W.T. FREEMAN : Understanding and evaluating blind deconvolution algorithms. *In IEEE CVPR*, 2009.

[10] K. LIU et N. KEHTARNAVAZ : Real-time robust vision-based hand gesture recognition using stereo images. *J. of Real-Time Image Proc.*, 2013.

[11] S. MAHOTRA, C. PATLOLLA et N. KEHTARNAVAZ : Real-time computation of disparity for hand-pair gesture recognition using a stereo webcam. *J. of Real-Time Image Proc.*, 7(4), 2012.

[12] S. MALIK : Real-time hand tracking and finger tracking for interaction csc2503f project report. *Department of Computer Science, University of Toronto, Tech. Rep*, 2003.

[13] V.I. PAVLOVIC, R. SHARMA et T.S HUANG : Visual interpretation of hand gestures for human-computer interaction : A review. *IEEE Trans. on PAMI* , 19(7), 1997.

[14] A. PENTLAND : A new sense for depth of field. *IEEE Trans. on PAMI* , 4, 1987.

[15] S. ROY et I. J. COX : A maximum-flow formulation of the n-camera stereo correspondence problem. *IEEE ICCV*, 1998.

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

[17] C. XU et L. CHENG : Efficient hand pose estimation from a single depth image. *In IEEE ICCV*, 2013.