Hand segmentation using a chromatic 3D camera

P. Trouvé, F. Champagnat, M. Sanfourche, G. Le Besnerais

HAND SHAPE SEGMENTATION WITH A CHROMATIC 3D CAMERA

P. Trouvé, F. Champagnat, M. Sanfourche and G. Le Besnerais

ONERA-The French Aerospace Lab
Corresponding author : pauline.trouve@onera.fr

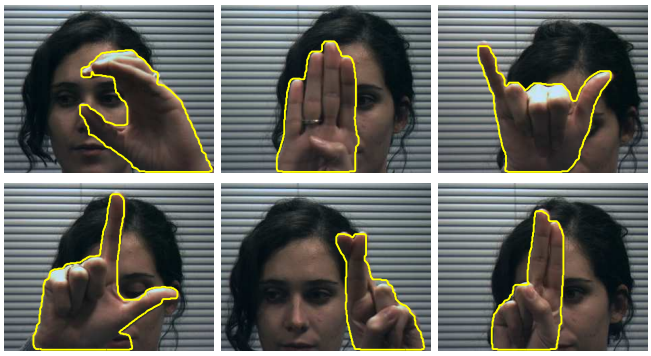


Fig. 1. Hand segmentation results obtained with the proposed 3D chromatic camera.

ABSTRACT

In this paper we present a new approach for hand shape segmentation using a passive monocular 3D camera. The camera depth estimation is based on the *Depth from Defocus* principle improved with the use of enhanced chromatic aberration. As this camera is passive and monocular, it can be used for indoor and outdoor application on a compact device, in contrast with active IR depth sensors that can be disturbed with the sun's infrared illumination. We show on various experimental examples that hand can be segmented using a depth cue from the proposed 3D camera.

***Index Terms*—** Hand segmentation, partial hand pose estimation, Depth from defocus, chromatic aberration.

1. INTRODUCTION

Man-machine interface through vision has known an increasing development in the last years, mostly due to miniaturization of the cameras that are now present on everyday objects such as laptops, smartphones and touchpads. In particular, hand pose estimation (HPE) is a challenging task, because of the 20 degrees of freedom of the hand, the requirement of a processing speed compatible with man/machine interactions, the possible uncontrolled environment and rapid hand motions [1, 2]. A simplified problem is the case of *Partial hand pose estimation* [2], where only the positions and orientations of the fingertips and the palm are looked for. In this field of application, an important issue is the hand localization and shape estimation. In this paper we propose to handle this problem using a passive 3D monocular camera. This camera 3D ability comes from a *Depth from Defocus* (DFD) approach enhanced by the use of chromatic aberration that creates spectrally varying defocus blur in the acquired image. We present results of hand segmentation obtained with the proposed chromatic 3D camera and a dedicated segmentation algorithm.

1.1. Related works

State of the art approaches for hand localization through vision are usually either based on color cues, background subtraction, motion cue and tracking or classification object detection (see [1, 2] and references herein). However those approaches are prone to illumination changes, use assumption of a single motion in the scene, or require to build a large samples training set. On the other hand a very efficient solution for hand segmentation in the context of uncontrolled background and illumination is now to use an active 3D camera, also referred to as RGB-D camera. Impressive hand segmentation results can be obtained using active sensors such as the Kinect [3, 4], however this stereoscopic active device relies on IR illumination of the scene that can be disturbed by the IR illumination of the sun, which is a drawback for outdoor applications. On the other hand, passive stereoscopic devices have also been proposed for HGR in [5, 6]. However passive and active stereoscopic devices both require the use of at least two cameras, which leads to bulky devices.

In this paper, we propose to conduct hand segmentation using a 3D camera based on DFD, i.e. a local estimation of the defocus blur [7]. Compared to traditional stereoscopic ranging system, a DFD camera requires only a single lens and thus leads to a more compact and simple experimental setting. Besides, in contrast to active ranging systems, it can be used in outdoor as well as in indoor situations. Several 3D cameras using DFD have been recently developed [8, 9, 10], however to the best of our knowledge, no practical application of these cameras for HPE have yet been thoroughly investigated.

1.2. Paper organization

Section 2 is dedicated to the derivation of a depth cue for hand segmentation based on DFD and a chromatic camera. In sec-

tion 3 are presented the algorithms for hand segmentation and fingertips localization from this depth cue and the results obtained on experimental images. We discuss the perspectives of the proposed approach for partial HPE before to conclude in section 4.

2. 3D CHROMATIC CAMERA

2.1. Principle

Depth from defocus (DFD) is a passive depth estimation method based on the relation between defocus blur and depth [7]. Indeed, as illustrated in Fig.2, if a point source is placed out of the in-focus plane of an imaging system, its image, which corresponds to the point spread function (PSF), has a size given by the geometrical relation:

$$\epsilon = Ds \left| \frac{1}{f} - \frac{1}{d} - \frac{1}{s} \right|, \quad (1)$$

where f is the focal length, D the lens diameter, d and s respectively the distance of the point source and the sensor with respect to the lens. Knowing f and s, the depth d can be inferred from a local estimation of the PSF size, or in other words, of the local blur amount.

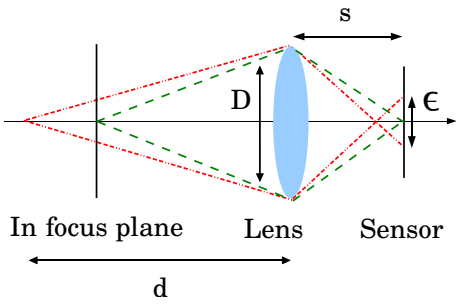
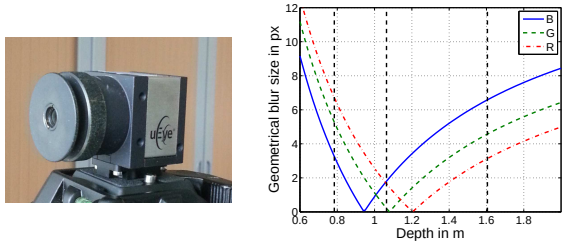


Fig. 2. Principle of Depth from Defocus

In this paper we propose to conduct hand shape segmentation using a DFD camera having enhanced chromatic aberration. Chromatic aberration implies a variation of the in-focus plane with respect to wavelength, thus as illustrated in figure 3, a variation of the defocus blur with respect to the color channel. In contrast with conventional camera, a chromatic camera has no depth ambiguity before or after the in-focus plane, and no dead zone [10], which improves both the accuracy in depth estimation and the range where depth can be estimated.

We use the same F/4 25 mm chromatic lens, with axial chromatic aberration of 100μm, as proposed in [10] but with a different setting. Here we use a uEye 1240 camera with pixel size of 5.3 μm and a resolution of 1280 $\times$ 1080. The camera field of view is about 20°. Fig. 3 (a) shows a picture of the camera and Fig. 3 (b) the camera theoretical RGB geometric blur size variation with respect to depth.



(a) (b)

Fig. 3. (a) Picture of the 3D camera. (b) Blur size variation of the proposed chromatic camera. Black vertical lines correspond to the depths that are used to segment the hand.

2.2. Depth cue from a chromatic 3D camera

In [10] we have presented a DFD algorithm dedicated to the processing of images produced by a chromatic camera. This algorithm is based on the derivation of a *generalized likelihood* calculated for each image patch and each depth. Here we propose to use this likelihood has a depth cue for hand segmentation.

2.2.1. Image formation model

The relation between the scene and the recorded image is usually modeled as a convolution with the PSF. In the general case, defocus blur varies spatially in the image and this model is only valid on image patches, where the PSF is supposed to be constant. In the case of a lens having spectrally varying defocus blur combined with a color sensor, each RGB channel has a different PSF. Using the matrix formalism on image and scene patches, this case can be modeled as:

$$\boldsymbol{Y}_C = H_C(d)\boldsymbol{X}_C, \quad (2)$$

where $\boldsymbol{Y}_C = [\boldsymbol{y}_R^T\, \boldsymbol{y}_G^T\, \boldsymbol{y}_B^T]^T$ and $\boldsymbol{X}_C = [\boldsymbol{x}_R^T\, \boldsymbol{x}_G^T\, \boldsymbol{x}_B^T]^T$ represent the concatenation of the pixels of three RGB scene and image patches respectively, $\boldsymbol{N}$ stands for the noise which affects the three channels and $H_C(d)$ is a block diagonal matrix containing each RGB channel convolution matrix [10]. Assuming a zero-mean white Gaussian noise process with variance σ_n^2 the data likelihood writes:

$$p(\boldsymbol{Y}_C|\boldsymbol{X}_C, \sigma_n^2) \propto \exp\left(-\frac{\|\boldsymbol{Y}_C - H_C\boldsymbol{X}_C\|^2}{2\sigma_n^2}\right). \quad (3)$$

2.2.2. Scene model

With a chromatic camera the RGB images originate from three distinct scenes: $\boldsymbol{x}_R, \boldsymbol{x}_G$ and $\boldsymbol{x}_B$, which are partially correlated. We propose to decompose the images into the luminance and the red-green and blue-yellow chrominance

decomposition $\boldsymbol{X}^{LC} = [\boldsymbol{x}_l^T\ \boldsymbol{x}_{c_1}^T\ \boldsymbol{x}_{c_2}^T]^T$ defined as:

$$\boldsymbol{X}_C = T \otimes \mathbf{I}_M \boldsymbol{X}^{LC} \text{ where } T = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \end{bmatrix}, \quad (4)$$

and $\otimes$ stands for the Kronecker product and $\mathbf{I}_M$ is the $M \times M$ identity matrix. Following [11], the three components of the luminance/chrominance decomposition are assumed independent. The observation model then writes:

$$\boldsymbol{Y}_C = H_{LC}(d)\boldsymbol{X}^{LC} + \boldsymbol{N} = H_C(d)T \otimes \mathbf{I}_M + \boldsymbol{N}. \quad (5)$$

We propose to use the Gaussian prior [12, 10], on each luminance and chrominance components, which leads to:

$$p(\boldsymbol{X}^{LC}, \sigma_x^2, \mu) \propto \exp\left(-\frac{\|D_C(\mu)\boldsymbol{X}^{LC}\|^2}{2\sigma_x^2}\right), \quad (6)$$

where D_C is a block diagonal matrix whose blocks are respectively $\mu D, D$ and D given that D is the vertical concatenation of the convolution matrices relative to the horizontal and vertical first order derivation operator [10]. Parameter μ models the ratio between the gradient variances of the luminance and the chrominance components and is fixed at 0.04 [10, 11].

2.2.3. *Generalized likelihood derivation*

Using the previous scene and image models, a marginal likelihood function can be analytically derived. Following the derivation of [10] maximization of the marginal likelihood over the noise parameter gives a *generalized likelihood*. In [10] we show that maximization of this likelihood is equivalent to minimize the criterion:

$$GL_C(d,\alpha) = \frac{\boldsymbol{Y}_C^T P(\alpha,d)\boldsymbol{Y}_C}{|P(\alpha,d)|_+^{1/(3N-3)}} \quad (7)$$
$$P(\alpha,d) = I_N - H_{LC}(H_{LC}^T H_{LC} + \alpha D_C^T D_C)^{-1} H_{LC}^T,$$

where $|A|_+$ stands for the product of the non zeros eigenvalues of matrix A, $3N$ is the length of $\boldsymbol{Y}_C$ and $\alpha = \sigma_n^2/\sigma_x^2$ can be interpreted as the inverse of a signal to noise ratio. Note that the criterion GL_C can be calculated on any depth.

3. HAND SEGMENTATION FROM DFD

3.1. Segmented depth map

We propose to conduct hand segmentation using a Markov random field model. The data fidelity term is the likelihood defined in (7) and a segmented depth map is obtained by minimization of the energy :

$$E(d) = \sum_p GL_C(d_p, \widehat{\alpha}) + \lambda \sum_{p,q \in N_p} \exp\left(-\frac{||y_g(p) - y_g(q)||^2}{2\sigma^2}\right)(1 - \delta(d_p, d_q)), \quad (8)$$

where N_p is a first order neighborhood of the pixel p, d_p is the estimated depth value at pixel p and y_g is the result of color image conversion in grayscale and $\widehat{\alpha}$ is obtained for each depth with a 1D continuous minimization of (7) [10]. This energy favors depth jumps located on image edges. We minimize this criterion using a graphcut algorithm [13].

Figure 4 (a) to (c) are respectively example of image acquired with the chromatic camera and segmented depth maps. We choose examples of hand in front of a person head on purpose, to illustrate the effectiveness of our approach on scenes that can be complex to process in particular with only skin color cues. The likelihood is calculated for three depths: 0.7 m, 1 m and 1.6 m. Figure 4 (b) is obtained with $\sigma = 4.5 \times 10^{-4}$ and $\lambda = 0$, thus corresponds to a minimization of $GL_C(d, \widehat{\alpha})$ with no regularization. The result is noisy and presents wrong depth labels. In contrast, Figure 4 (c), obtained with the same value of σ and $\lambda = 5$, shows a smooth segmented depth map consistent with the scene.

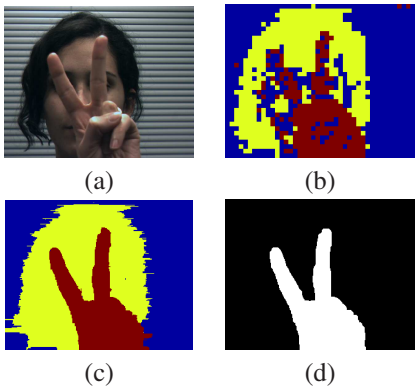


Fig. 4. (a) Acquired image. Results of the minimization of (8) with (b) λ = 0 (c) λ=5 (yellow label is for 1m, blue label for 1.6 m and red label for 0.7m. (d) Extracted hand shape.

3.2. Hand shape extraction

Given the segmented depth map obtained with the minimization of equation (8), hand shape is extracted using a threshold in depth values, fixed at 0.9 m. Then the objects in the binary image are labeled and we extract the hand as the object with the largest area. Finally, hand contours are determined using a boundary trace function, after erosion and dilatation of the contour with a disk of 7 pixels diameter. Figure 4 (d) shows the extracted hand shape, using the segmented depth map of Fig 4 (c).

3.3. Fingertips localization

An interesting application of hand shape estimation is fingertips localization, for instance to conduct finger spelling [6]. Using the extracted hand shape, we use the approach proposed in [14] where peaks and valleys are detected from the

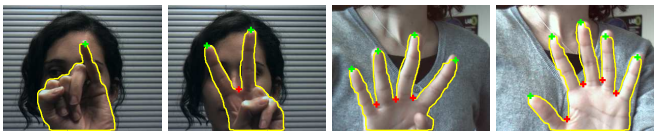


Fig. 5. Fingertips detection results. Green crosses are for fingertips while red crosses are for junctions between fingers.

hand contour using the k-curvature, which is the angle between the two vectors [C(j), C(j - k)] and [C(j), C(j + k)] at each pixel j in a hand contour C, with k a constant. Small k-curvatures are typical of a peak or a valley, while large k-curvatures correspond to a smooth region. A threshold on the k-curvature thus returns the peaks and valleys of the hand shape, i.e fingertips or junction between fingers. Moreover, the sign of the cross products between the two vectors indicates if the selected point is actually a fingertip or not. Figure 5 shows fingers detection results obtained with a threshold of 30^o and k=50. The overall time processing speed is of 0.8s with a processor Corei7 3930k @ 3.26 GHz, with a Matlab implementation of the proposed algorithms.

3.4. Discussion

In the proposed approach the quality of hand shape depends on the regularization stage. Indeed a underregularization leads to a noisy depth map, while overregularization leads to oversmooth depth map, where fingers can be merged with the background level (see for instance 6). Thus other segmentation approaches, based for instance on bilateral filters are to be studied to make the hand segmentation more robust. Note that depth information with the proposed camera could also be merged with the usual hand localization cues such as color cue or motion cue to improve hand localization performance.

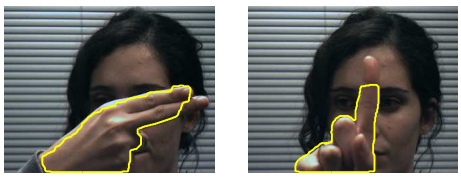


Fig. 6. Examples of overregularization.

4. CONCLUSION

In this paper we have presented the first results of hand shape segmentation using a chromatic 3D camera using DFD. In contrast to state of the art methods, our approach is suitable for indoor and outdoor applications on a compact device. A perspective of this work is to develop a learning stage for partial HPE with an experimental evaluation of the pose recognition rate. Moreover with the proposed 3D camera and the genericity of the likelihood model, we could also consider to build a depth cue from DFD with more than 3 depths and thus conduct a full degrees of freedom HPE.

Ackowledgments

The authors wish to thank L. Jacubowiez and J. Sabater of the IOGS for the design of the chromatic lens, B. Le Saux, A. Plyer and J. Abou for fruitfull discussions.

5. REFERENCES

[1] V. I. Pavlovic et al., "Visual interpretation of hand gestures for human-computer interaction: A review," *IEEE Trans. on Pattern Anal. Mach. Intell.* , vol. 19, no. 7, 1997.

[2] A. Erol et al., "Vision-based hand pose estimation: A review," *Comp. Vision and Image Understanding*, vol. 108, no. 1, 2007.

[3] C. Keskin et al., "Hand pose estimation and hand shape classification using multi-layered randomized decision forests," *IEEE Europ. Conf. Comp. Vision*, 2012.

[4] C. Xu and L. Cheng, "Efficient hand pose estimation from a single depth image," in *IEEE Int. Conf. Comp. Vision*, 2013.

[5] S. et al. Mahotra, "Real-time computation of disparity for hand-pair gesture recognition using a stereo webcam," *J. of Real-Time Image Proc.*, vol. 7, no. 4, 2012.

[6] K. Liu and N. Kehtarnavaz, "Real-time robust vision-based hand gesture recognition using stereo images," *J. of Real-Time Image Proc.*, 2013.

[7] A. Pentland, "A new sense for depth of field," *IEEE Trans. on Pattern Anal. Mach. Intell.* , vol. 4, 1987.

[8] A. Levin et al., "Image and depth from a conventional camera with a coded aperture," *ACM Trans. on Graphics (TOG)*, vol. 26, no. 3, 2007.

[9] A Chakrabarti and T. Zickler, "Depth and deblurring from a spectrally varying depth of field," in *IEEE Europ. Conf. Comp. Vision*, 2012, vol. Part V, pp. 648–661.

[10] P. Trouvé and et al., "Passive depth estimation using chromatic aberration and a depth from defocus approach," *Appl. Opt.*, vol. 52, no. 29, 2013.

[11] L. Condat, "A generic variational approach for demosaicking from an arbitrary color filter array," in *IEEE Int. Conf. on Image Processing*, 2009.

[12] A. Levin and et al., "Understanding and evaluating blind deconvolution algorithms," in *IEEE Conf. Comp. Vision Pattern Recogn.*, 2009.

[13] S. Roy and Ingemar J. Cox, "A maximum-flow formulation of the n-camera stereo correspondence problem," *IEEE Int. Conf. Comp. Vision*, Jan 1998.

[14] S. Malik, "Real-time hand tracking and finger tracking for interaction CSC2503F project report," *Department of Computer Science, University of Toronto, Tech. Rep*, 2003.