Self-Supervised Focus Measure Fusing for Depth Estimation from Computer-Generated Holograms

Nabil Madali, Antonin Gilles, Patrick Gioia, Luce Morin

Self-Supervised Focus Measure Fusing for Depth Estimation from Computer-Generated Holograms

Nabil Madali[1,3*] Antonin Gilles[1] Patrick Gioia[1,2] Luce Morin[1,3]

[1] IRT b<>com Cesson-Sévigné France
[2] Orange Labs Cesson-Sévigné France
[3] INSA Rennes Rennes France

Abstract

Depth from focus is a simple and effective methodology for retrieving the scene geometry from a hologram when used with the appropriate focus measure and patch size. However, fixing those parameters for every sample may not be the right choice, as different scenes can be composed with various types of textures. In this work, we propose a self-supervised learning methodology for fusing the depth maps produced using different focus measures with variable patch sizes applied to the holographic reconstruction volume. Experimental results show that fusing depth information produces more accurate and smoother depth maps, which can be directly used for alternative tasks such as motion estimation.

Keywords : 3D Imaging, Holography, Depth Estimation, Self-supervised learning, Depth from focus.

1 Introduction

Designing an efficient holographic video codec is still an open problem in the current state of the art, due to the lack of a fast and accurate motion estimation algorithm. Indeed, traditional video codecs based on block-matching are not appropriate for holographic data because there is no clear spatial correlation between consecutive frames [1]. To accurately estimate motion from one frame to another, the scene geometry must first be recovered to calculate the 3D motion vectors.

In [2], the authors demonstrated that the *Depth-From-Focus* (DFF) [3–5] method is a simple and effective technique for retrieving the scene geometry when implemented within a well-defined framework. This technique involves building a reconstruction volume by performing a series of numerical reconstructions at sampled distances. Then, the level of focus for each pixel in the reconstruction volume is evaluated using a *Focus Measure* (FM) on a patch centered around the pixel, and the distance at which the pixel reaches its optimal focus level is selected as its depth value. Experimental results have shown that the focus measure and patch size significantly impact the precision of the estimated scene geometry. Different focus measures may perform differently depending on whether the input region is uniform or has high texture variation. Furthermore, the authors point out that, due to the unique nature of holographic numerical reconstruction, the focus curve can be subject to polarity change when transitioning from uniform to textured regions, depending on the reconstruction interval and patch size. To address this issue, an automatic switch between the global minimum and maximum of the focus curve is implemented. This allows for the handling of polarity change and ultimately improves the final performance.

Inspired by the latest breakthroughs in self-supervised image denoising [6–10], we propose in this paper an enhanced version of the automatic switch process which determines the appropriate depth value using both local and global information. The method involves extracting multiple depth maps from a reconstruction volume using various focus measures and patch sizes, then training a neural network in a self-supervised manner to identify the depth map with the lowest average deviation. The experimental results show that incorporating multiple focus measures and different patch sizes significantly improves the accuracy and smoothness of the predicted depth map. The remaining of the article is organized as follows: Section 2 details the different components of the proposed method. Then, in Section 3 a series of experiments are conducted to validate the proposed approach. Finally, in Section 4, we discuss the advantages and limitations of the presented method.

*This work has been achieved within the Research and Technology Institute b<>com, dedicated to digital technologies. It has been funded by the French government through the National Research Agency (ANR) Investment referenced ANR-A0-AIRT-07. Authors can be reached at {nabil.madali, antonin.gilles, patrick.gioia, luce.morin}@b-com.com.

2 Methodology

2.1 Overview

The proposed approach is composed of three steps: First, the input hologram H of size $L \times L$ is used to compute a reconstruction volume. This step involves several hyperparameters that can affect the final results, so to ensure fair and accurate evaluation, the same methodology as proposed in [2] is consistently used. Next, the DFF method is applied to the reconstruction volume using various focus measures and patch sizes, each of which produces a depth map referred to as measurement. All these measurements can be considered as different distorted versions of the ground truth depth map. Finally, a neural network is used to estimate a depth map that is as close as possible to the ground truth by minimizing the average deviation from the measurements.

2.2 Reconstruction volume acquisition

Given an input hologram H, a reconstruction volume is computed by performing a set of numerical reconstructions using the Angular Spectrum Method [11] at uniformly sampled reconstruction distances inside a manually defined depth interval $[z_{\min}, z_{\max}]$, where $z_{\min}$ and $z_{\max}$ are the minimal and maximal reachable depths values respectively. Each numerical reconstruction is defined as

$$\mathcal{P}_{z_i}\{H\} = \mathcal{F}^{-1}\left\{\mathcal{F}\{H\}(f_x, f_y)e^{j2\pi z_i\sqrt{\lambda^{-2}-f_x^2-f_y^2}}\right\}, \tag{1}$$

where $\mathcal{F}$ is the Fourier transform, f_x and f_y are the spatial frequencies along the X and Y axis, λ is the acquisition wavelength, and z_i is the sampled reconstruction depth, given by

$$z_i = \frac{z_{\max} - z_{\min}}{N} i + z_{\min}, \tag{2}$$

where N is the number of sampled distances.

2.3 Depth From Focus (DFF) method

The focus level of each pixel in the reconstruction volume is evaluated locally using a patch $R_{m,n,i}$ of size $s \times s$ centered around each pixel (m, n), defined as

$$R_{m,n,i}(u, v) = \left|\mathcal{P}_{z_i}\{H\}(m + u - s/2, n + v - s/2)\right|. \tag{3}$$

Then, the reconstruction distances at which the focus curve reaches its maximum $d^{max}_{m,n}$ and minimum $d^{min}_{m,n}$ values are extracted, such that

$$d^{max}_{m,n}(FM, s) = \underset{i\in[1,N]}{\arg\max}\left\{FM(R_{m,n,i})\right\}, \tag{4}$$

$$d^{min}_{m,n}(FM, s) = \underset{i\in[1,N]}{\arg\min}\left\{FM(R_{m,n,i})\right\}. \tag{5}$$

According to [2], if the optimal focus measure is employed, and the reconstruction interval is close to the boundary of the scene, the sharpness measurement for highly textured patches will achieve its highest value at the depth of focus. Conversely, for uniform patches, the sharpness measurement will reach its lowest value at the depth of focus. However, if the optimal conditions are not met, the focus curve is likely to have multiple extrema, with the optimal focus depth located at the second or third extremum value. The objective of this study is to enhance the fusion of the textured and smooth regions obtained from the estimated depth maps d^{max} and d^{min}, respectively. This will be achieved by considering not only the information provided by each pixel focus curve but also the local neighborhood information at various patch resolutions.

Assuming a uniform distribution of both regions throughout the depth map, each depth map can be perceived as a degraded version of the ground truth depth map. The degradation can be visually assessed using the in-focus map linked with each depth map, where the well-estimated areas appear sharp and the remaining areas are contaminated by speckle noise. All possible predictions produced using different pairs of focus measurements and patch size can be united into a set given by:

$$\mathcal{D} = \bigcup_{FM\in\Phi}\bigcup_{s\in\mathcal{S}}\left\{d^{min}(FM, s), d^{max}(FM, s)\right\} \tag{6}$$

where Φ and $\mathcal{S}$ are all the possible focus measures and patch sizes.

2.4 Self-supervised learning

Given the set $\mathcal{D}$, the task at hand is to determine the depth map d that has the minimum average deviation, according to the loss function $\mathcal{L}$, by minimizing

$$\arg\min_{d} \mathbb{E}_{y\in\mathcal{D}}(\mathcal{L}(d, y)). \tag{7}$$

A methodology similar to the one proposed in [7] is used, which reformulates image restoration as an optimization problem. Given a degraded image $\dot{x}$ and a clean image x, the conventional image denoising problem can be written as

$$\hat{x} = \arg\min E(\dot{x}; x) + R(x), \tag{8}$$

where $E(\dot{x}; x)$ is a fidelity term and $R(x)$ is a regularization term. The optimal value of x is iteratively searched in the image space starting from a random initial point until an optimal value is reached according to predefined criteria. An alternative approach outlined in [7] involves constructing a parameterized function G with parameters θ and optimizing them in the parameter space using gradient descent until convergence.

More formally

$$\hat{x} = \arg\min E(G_\theta(z); \dot{x}) + R(G_\theta(z)), \tag{9}$$

where G_θ is a convolution network and z is random noise. The authors claim that due to the surjective nature of the function G, the two problems are equivalent. Experimental results indicate that using a suitable network architecture G initialized randomly and optimized with gradient descent, the network can converge to a naturally-looking local optimum or, at least, pass near one before reaching the optimal solution $\dot{x} = G_\theta(z)$. To prevent the network from overfitting to the input, the authors employed an early stopping technique by manually setting a limit on the number of iterations.

In the present work, we used the assembled encoder-decoder proposed in [7] for the inpainting task [12]. It has a depth of 5 layers, 128 output channels at each stage and does not use skip connections. We used the ℓ_2 and Sobel gradient for the fidelity and regularization terms, respectively. Overall, the training process can be formalized as

$$\hat{x} = \arg\min \frac{1}{\mid L \mid^2} [\sum_{i,j} \|\dot{x} - \hat{x}\|_2 + (\nabla_x \|\dot{x} - \hat{x}\|_2 + \nabla_y \|\dot{x} - \hat{x}\|_2)], \tag{10}$$

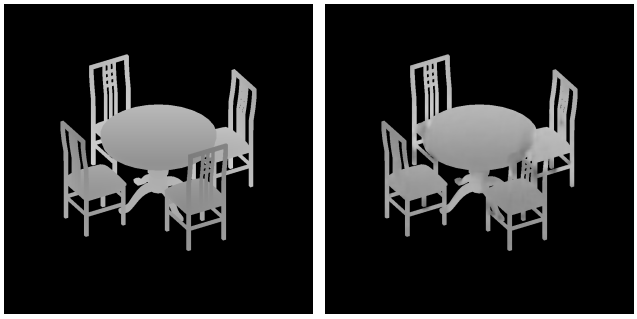
where $\dot{x}$ is a random sample from $\mathcal{D}$, $\hat{x}$ is the predicted value, and ∇_x and ∇_y are the image gradients along the x and y directions respectively.

3 Experiments

To evaluate the performance of the proposed methodology, a collection of 500 holograms, divided into five categories (*Piano*, *Table*, *Wood*, *Dices*, *Cars*), were obtained using a layer-based method [13], with a resolution of 1024×1024, a pixel pitch of $6\mu m$, and 100 different acquisition angles for each scene.

Each hologram is then reconstructed at 256 different distances within the range of $[z_{min}, z_{max}]$, with z_{min} and z_{max} set to 0.0049cm and 1.23cm, respectively. Four of the best-performing focus operators from [2], namely the Variance of Laplacian [14] (LAPV), the Ratio (WAVR) and Variance (WAVV) of wavelet coefficients [15], and the Normalized Graylevel variance [16] (GLVN), are used with variable patch-sizes in the set $\{5, 9, 13, 17, 21, 33\}$.

For each hologram, the network is trained using the depth maps $\mathcal{D}$ for 20000 iterations with an exponential learning rate decay starting from 0.1 with a factor of 0.8 every 100 iterations.



(a) (b)

Figure 1: An illustration of the estimates produced on the *Table* scene, where occluded areas in the ground truth depth map in (a) are misestimated on the predicted depth map in (b).

3.1 Results

Table 1 gives the obtained ℓ_1 norm for both the auto-switch method [2] and the proposed method.

	Piano	Table	Woods	Cars	Dices
LAPV	22.4	37.17	7.65	9.47	21.4
WAVV	24.46	39.4	7.12	9.53	19.01
WAVR	10.24	22.64	4.89	6.1	9.99
GLVN	15.63	26.3	4.2	7.24	5.54
Ours	4.41	8.54	2.53	3.82	3.02

Table 1: The $\ell 1$ norm evaluates the number of reconstruction planes between the predicted and the ground truth depth. Green for the first, blue for the second, and red for the third best-performing method.

The obtained results show that for the DFF method with auto-switch, the GLVN and WAVR focus measures generally perform the best across all scenes, with the highest ℓ_1 norm for the *Table* scene due to multiple scene objects occlusions. The performance of the focus measures can be attributed to three factors. First, the size of the reconstruction interval and the number of sampled reconstruction distances: when the reconstruction interval and depth sampling rate increase, the focus curves become noisier, which complicates the optimal focus plane estimation. Second, estimating the depth map using a single patch size may not be adaptable as different regions may require larger or smaller patches to have a proper focus curve with single extremum values. Third, the performance also depends on the nature of the used operators. For example, gradient-based operators will have higher performance in textured regions and poor performance in smooth regions.

Fusing the depth maps leads to improved performance across all scenes. The difference in performance between the auto-switch approach and the proposed

approach varies depending on the initial depth estimation. For instance, the improvement in *Dices* and *Woods* class, where initial estimations are close to the ground truth, is relatively small compared to other scenes where the performance is twice as good as the auto-switch. Despite the network improved handling of the occlusions in the *Table* scene, the performance remains the weakest. Indeed, while the estimated depth map presents clear boundaries between occluded objects, our proposed method fails to accurately predict the depth in occluded regions, as depicted in Figure 1.

The final results of the network are determined by the chosen regularization term and the used output space $\mathcal{D}$. When no regularization is applied, the network is prone to converge to an over-smoothed solution. To remedy this, the Sobel loss is employed to increase the output depth map sharpness. The total variation method was also considered, however, it hinders the convergence speed and gives poor results.

3.2 Discussion

In addition, the images in the set $\mathcal{D}$ should have a consistent degradation compared to the ground truth depth map and each region of the depth map should be accurately estimated in at least half of the images utilized. If a region is poorly estimated in several images, the network tends to average those estimates instead of propagating the well-estimated values from the other regions. In order to achieve accurate depth map estimation for each region, it may be necessary to reduce the number of images used to a minimum number where each region is accurately estimated in at least half of them. Another potential solution is to directly train the network using the depth maps produced using the auto-switch method. For example, in Figure 2, the network was trained only with the depth map generated by the WAVR operator with a patch size of 33×33. This resulted in a smoother depth map, but it lacked spatial coherence. In addition, the network struggles to properly address discontinuous areas. By supervising the network with more depth maps, the results are significantly improved and the predicted depth map is close to the ground depth.

Although the network yields meaningful results, it has limitations. Firstly, the choice of neural network architecture influences the final results. The selected network should have a long path from input to output while avoiding too many parameters that cause fast overfitting. Secondly, finding the optimal number of iterations can be challenging, since stopping too soon leads to poor depth maps, while stopping too late degrades the results as shown in Figure 3. Lastly, the optimization step takes considerable time, around 10-15 minutes in our experiments, which may hinder online use.

4 Conclusion

In this work, we present a learning-based approach to fuse information obtained by applying the DFF method on a holographic reconstruction volume using various focus measures with varying patch sizes.

The experimental results showed that combining different focus measures can significantly improve the obtained results when selecting the appropriate hyperparameters, such as network architecture, number of iterations, optimization, and a regularization term.

References

[1] David Blinder, Ayyoub Ahar, Stijn Bettens, Tobias Birnbaum, Athanasia Symeonidou, Heidi Ottevaere, Colas Schretter, and Peter Schelkens, "Signal processing challenges for digital holographic video display systems," *Signal Processing: Image Communication*, vol. 70, pp. 114–130, 2019.

[2] Nabil Madali, Antonin Gilles, Patrick Gioia, and Luce Morin, "Automatic depth map retrieval from digital holograms using a depth-from-focus approach," *Appl. Opt.*, vol. 62, no. 10, pp. D77–D89, Apr 2023.

[3] Lihong Ma, Hui Wang, Yong Li, and Hongzhen Jin, "Numerical reconstruction of digital holograms for three-dimensional shape measurement," *Journal of Optics*, vol. 6, pp. 396–400, 2004.

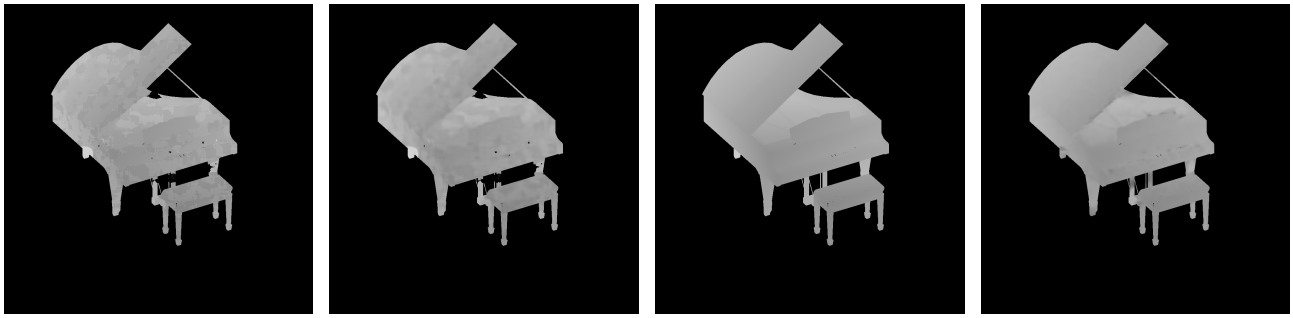
[4] Conor P. McElhinney, Jonathan Maycock, Bryan M. Hennelly, Thomas J. Naughton, John B. McDonald, and Bahram Javidi, "Extraction and reconstruction of shape information from a digital hologram of three-dimensional objects," 2006.

[5] Baturay Ozgurun and Mujdat Cetin, "Depth extraction from a single compressive hologram," *arXiv preprint arXiv:2102.13371*, 2021.

[6] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila, "Noise2noise: Learning image restoration without clean data," *CoRR*, vol. abs/1803.04189, 2018.

[7] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky, "Deep image prior," *CoRR*, vol. abs/1711.10925, 2017.

[8] Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo, "Self-supervised super-resolution for multi-exposure push-frame satellites," 2022.



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

Figure 2: In (b) the obtained depth map when the network is supervised using only one depth map computed using the WAVR operator in (a), (c) the depth map when trained using the proposed methodology, and (d) the ground truth depth map.

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

Figure 3: The intermediate results generated by the network after 1,000, 4,000, and 20,000 iterations in (a), (b), and (c), respectively, are compared to the ground truth depth map in (d). During the early iteration stages, the network converges towards a relevant solution but lacks the intricate details present in the ground truth depth map. Nevertheless, as the iteration count increases, the network gradually integrates the missing details into its predicted depth map.

[9] Saeed Izadi, Darren Sutton, and Ghassan Hamarneh, "Image denoising in the deep learning era," 2022.

[10] Thibaud Ehret, Axel Davy, Gabriele Facciolo, Jean-Michel Morel, and Pablo Arias, "Model-blind video denoising via frame-to-frame training," *CoRR*, vol. abs/1811.12766, 2018.

[11] Joseph W Goodman, "Introduction to fourier optics," *Introduction to Fourier optics, 3rd ed., by JW Goodman. Englewood, CO: Roberts & Co. Publishers, 2005*, vol. 1, 2005.

[12] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," *CoRR*, vol. abs/1603.06937, 2016.

[13] Antonin Gilles, Patrick Gioia, Rémi Cozot, and Luce Morin, "Hybrid approach for fast occlusion processing in computer-generated hologram calculation," *Appl. Opt.*, vol. 55, no. 20, pp. 5459–5470, Jul 2016.

[14] J.L. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia, "Diatom autofocusing in brightfield microscopy: a comparative study," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, 2000, vol. 3, pp. 314–317 vol.3.

[15] Ge Yang and B.J. Nelson, "Wavelet-based autofocusing and unsupervised segmentation of microscopic images," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453)*, 2003, vol. 3, pp. 2143–2148 vol.3.

[16] A Santos, Carlos Ortiz-de Solorzano, Juan Jose Vaquero, J Peña, Norberto Malpica, and Francisco Del Pozo Guerrero, "Evaluation of autofocus functions in molecular cytogenetic analysis," *Journal of microscopy*, vol. 188, pp. 264–72, 01 1998.