Single-Pixel Salient Object Detection via Discrete Cosine Spectrum Acquisition and Deep Learning

Yonghao Li, Jianhong Shi, Lei Sun, Xiaoyan Wu, and Guihua Zeng

***Abstract*— Single-pixel imaging (SPI) can reduce the cost and have the potential of being competent for some challenging tasks. However, for a SPI system, acquiring detailed information from a complex scene for complex vision task is a measurements-consuming process, by which low efficiency resulted is one of the important obstructions of SPI for practical application. Reasonable allocation of resources of measurements and calculations is one of the effective solutions to this problem. As a preprocessing procedure with a role of guidance, salient object detection can help the vision system focus more attention on the area with more important information to improve the efficiency. Therefore, in this letter, we explore the implement of salient object detection based on SPI system and present a scheme via discrete cosine spectrum (DCS) acquisition and deep learning model.**

***Index Terms*— Single-pixel imaging, salient object detection.**

I. Introduction

As A computational imaging system, SPI can break through the limit of hardware by powerful data post-processing [1]. Meanwhile, compared with a pixelated detector, the fabrication of a single-pixel detector with fast timing response, high sensitivity to low light or effectivity to invisible light is cheaper and easier. Thus, SPI has potential of economically completing the tasks which conventional imaging system can do and applications on some challenging scenarios such as remote sensing [2], weak light imaging [3] and scatter imaging [4]. But for a SPI system, acquiring detailed information from a complex scene to meet the needs of some complex vision tasks always consumes a great many measurements and results low efficiency, which is one of the important causes to block the practical process of SPI. It is worth noting that there is a certain amount of redundant information in nature scene and the detection of them is usually useless for many vision tasks. Therefore, reasonably allocating the resource of detection and calculation is an effective way to increase efficiency of a SPI system for practical vision application.

Salient object detection, commonly serving as preprocessing procedure in many complex vision tasks including object recognition [5], visual tracking [6], robot navigation [7], etc., is in imitation of the human's talent that the regions of attractive objects in complex nature scene can be rapidly captured and more information from these areas is processed in full detail leaving the inessential from other areas relatively unprocessed, which greatly ensures the human vision system efficiently and normally working. Such an ability can be introduced into a SPI system as a preprocessing step to locate the areas with more important information in the scene, which can direct the SPI system to focus more detection and calculation on these areas to improve the efficiency of subsequent vision tasks. Meanwhile combined with the SPI's advantages of compressive sensing at data acquisition step and stronger adaptability to challenging applications, salient object detection based on a SPI system has potential to be a low-measurements-consumption preprocessing procedure with potential applications on some challenging scenarios. Therefore, the research on salient object detection based on a SPI system is valuable.

In this letter, we explore the implement of salient object detection based on a SPI system and present a scheme via more efficient acquisition of DCS and deep learning model. In proposed scheme, the DCS of scene is acquired. And for taking the SPI's advantage of compressive sensing at data acquisition step, a CNN-based model which can detect the regions of salient objects in the scene reconstructed from the undersampled DCS is trained. Simulation and experiment are conducted to verify feasibility of the proposed scheme.

II. Theories and Methods

The two set of discrete cosine basis patterns [8] for illumination can be expressed as

$$P_{\pm}(x, y; u, v) = \pm a \cdot c(u)c(v)C(x, u, M)C(y, v, N) + b \quad (1)$$

where (x, y) is the spatial coordinates; (u, v) is the spatial frequency; a is the contrast; $c(\tau) = 1/\sqrt{2}$ when $\tau = 0$; $c(\tau) = 1$ when $\tau \neq 0$; $C(\delta, \omega, \kappa) = cos\,[(\delta + 0.5)\omega\pi/\kappa]$; M and N indicate that size of image $R(x, y)$ to be reconstructed is $M \times N$; b is a constant to ensure no negative number in $P_{\pm}(x, y; u, v)$. To make the digital micromirror device (DMD) with high ceiling of modulation speed available, we use upsampling and Floyd-Steinberg error diffusion dithering [9] to binarize the gray-scale patterns $P_{\pm}(x, y; u, v)$, which can be denoted as

$$P_{B\pm}(x, y; u, v) = F\{U_{\eta}\{P_{\pm}(x, y; u, v)\}\} \quad (2)$$

where $P_{B\pm}(x, y; u, v)$ are binary discrete cosine basis patterns; a and b are set to 0.5; $U_{\eta}\{\bullet\}$ denotes η-folds upsampling

Manuscript received May 25, 2020; revised August 8, 2020; accepted September 17, 2020. Date of publication September 24, 2020; date of current version October 6, 2020. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61631014, Grant 61905140, and Grant 61471239 and in part by the Startup Fund for Youngman Research at SJTU under Grant 18 × 100040014. *(Corresponding author: Jianhong Shi.)*

The authors are with the State Key Laboratory of Advanced Optical Communication Systems and Networks, Center of Quantum Sensing and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: purewater@sjtu.edu.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LPT.2020.3026472

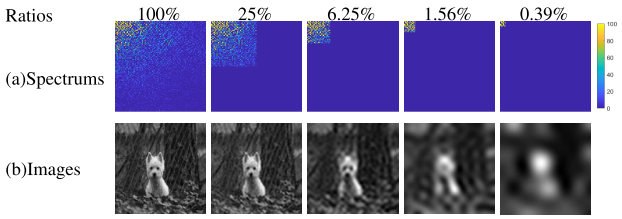

Fig. 1. Training images with different utilization ratios of DCS.

by 'bicubic' image interpolation algorithm; $F\{\bullet\}$ denotes Floyd-Steinberg error diffusion dithering. After the projected patterns $P_{B\pm}(x, y; u, v)$ illuminating the scene $T(x, y)$, the corresponding responses $D_{\pm}(u, v)$ of detector to reflected light can be shown as follow:

$$D_{\pm}(u,v) = k \cdot \sum_{x=0}^{\eta M-1} \sum_{y=0}^{\eta N-1} P_{B\pm}(x,y;u,v) \cdot T(x,y) + e \quad (3)$$

where k is the scale factor depending on the size and location of detector and e is the response of detector to background illumination. The DCS $RD(u, v)$ of the reconstructed image $R(x, y)$ can be obtained by canceling out the effect of constant b introduced in Eqs. 1, 2 and the response to background illumination e, which can be shown as

$$RD(u,v) = \{D_{+}(u,v) - \frac{D_{+}(U,V) + D_{-}(U,V)}{2}\}/h \quad (4)$$

where (U, V) is a selected and fixed frequency; h is a constant. According to this method, sampling for a full DCS consumes only $M \times N + 1$ measurements which are nearly 50% fewer than [8] and 33% fewer than [9].

After the DCS $RD(u, v)$ acquired, the image $R(x, y)$ of the scene is reconstructed by inverse DCT to be the input of the CNN-based model for extracting regions of the salient objects. The CNN-based model in our scheme adopts the structure of PoolNet [10] with ResNet-50 [11] as backbone but no joint training with edge detection. For training the network, firstly the 10553 images in training set of DUTS [12] are converted to gray and resized to 128×128. Then, DCT is applied to the preprocessed images to obtain their corresponding DCS. Next, according to 4 squares with different sizes (respectively 64×64, 32×32, 16×16, 8×8), we segment each spectrum to generate other 4 spectrums, as shown in row (a) of Fig. 1. Then, the corresponding images are acquired by inverse DCT, as shown in row (b) of Fig. 1. So, 5 sets of images (in total 52765) with different utilization ratios (respectively 100%, 25%, 6.25%, 1.56%, 0.39%) of spectrum are generated as training set. In the training phase, the 52765 training images with 128×128 pixels are mixed together to train the network. And as the settings in [10], the loss function is standard binary cross entropy loss. The network is trained for 24 epochs by Adam [13] optimizer with weight decay of 5e-4. The learning rate is initially 5e-5 and changed to 5e-6 after 15 epochs. Parameters of ResNet-50 pretrained on the ImageNet dataset [14] is used to initial the backbone and rest parameters of the network are randomly initialized. The program of the network ran on a graphics processing unit (NVIDIA TITAN Xp) with Python version 3.7 using PyTorch 1.1.0.

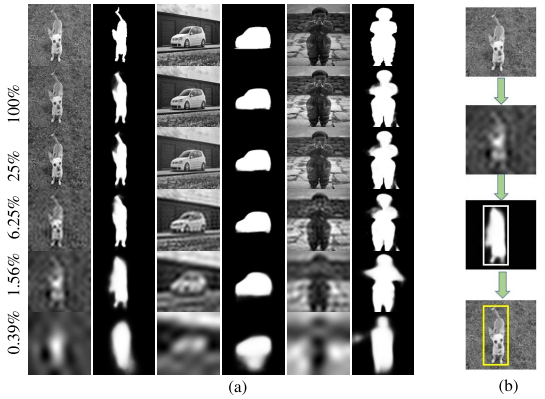

Fig. 2. (a) The results of several example object scenes. (The top row is object scenes and corresponding ground truths; the rest of rows are the reconstructed images at 5 different sampling ratios and the corresponding saliency maps); (b) Example application on visual tracking with sampling ratio 1.56%.

III. Experimental Results and Discussion

A. Computational Simulations

In simulation, the theoretical feasibility of DCS-acquisition method in our scheme is verified and the performance of the trained CNN-model to input images reconstructed at different sampling ratios of DCS is observed. We set $M = N = 128$, $\eta = 2$ and $U = V = 8$. The 1000 images in ECSSD [15] are converted to gray and resized to 256×256 as the test object scenes to be illuminated. According to the segmentation strategy shown in Fig. 1, the images of the test object scenes are reconstructed at 5 different sampling ratios of DCS. Then according to different sampling ratios, the 5000 test reconstructed images with 128×128 pixels are divided into 5 batches and respectively input into the trained CNN-based model. The reconstructed images and saliency maps of several test examples are shown in Fig. 2 (a). And because F-measure score and mean absolute error (MAE) are widely-used metrics for evaluating the performance of salient object detection models, we present the mean values of max F_β and MAE on the overall test object scenes at different sampling ratios in TABLE I. The higher F-measure score and the lower MAE mean better accuracy. F-measure is denoted as F_β:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \quad (5)$$

where $Precision = |B \cap G|/|B|$; $Recall = |B \cap G|/|G|$; B is a mask generated by binarizing the prediction saliency map S with a threshold; G is ground truth; $|\bullet|$ denotes the accumulation of non-zero entries; β^2 is empirically set to 0.3. MAE is calculated by

$$MAE = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} |S(i,j) - G(i,j)| \quad (6)$$

As the results shown in Fig. 2 (a) and TABLE I, accuracy of salient object detection is decreased with the decrease of sampling ratio. When the sampling ratio is reduced to 25% which is sub-Nyquist sampling, the decrease of accuracy on the whole test object scenes is small. It shows the possible of saving measurements-consuming on the premise of ensuring

TABLE I
MAX $F_\beta s$ AND MAEs ON TEST OBJECT SCENES

	100%	25%	6.25%	1.56%	0.39%
MaxF_β	0.8581	0.8528	0.7820	0.6909	0.6548
MAE	0.0798	0.0834	0.0995	0.1292	0.1701

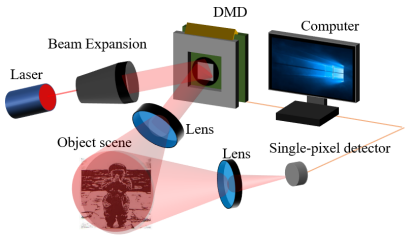

Fig. 3. Experimental setup. DMD: Digital micromirror device.

TABLE II
MAEs ON TEST OBJECT SCENES

Training	Test	100%	25%	6.25%	1.56%
128×128	64×64	0.1738	0.1814	0.1973	0.2138
128×128	128×128(upsampling)	0.0892	0.0998	0.1291	0.1705
64×64	64×64	0.1322	0.1342	0.1490	0.1782

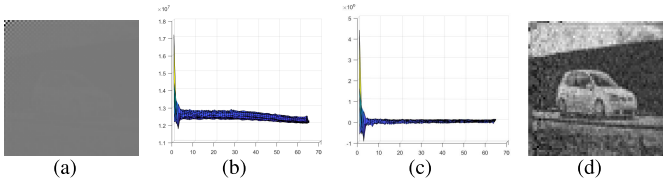

Fig. 4. (a) reconstructed image without denoising; (b) measured $D_+(u, v)$; (c) denoised $RD(u, v)$; (d) reconstructed image with denoising.

good accuracy. And when the sampling ratio is reduced to 6.25%, the system still works not badly. When the reconstructed image become quite fuzzy with the sampling ratio further declining to 1.56%, the rough shapes can also be output on saliency maps, as shown in Fig. 2 (a). Even when the reconstructed images are barely recognizable at 0.39% sampling ratio, the sketchy location can be obtained. From the results, we can see when the sampling ratio is too small, the reconstructed image is greatly distorted and the accuracy of salient object detection is obviously reduced. But for some applications, the required accuracy of salient object detection is low. For example, in the visual tracking [6], only location information of object is needed but the contour information is unimportant. So, there is no need to use more measurements to reconstruct less distorted image used in [6] for locating the salient object, as shown in Fig. 2 (b). It shows the potential to save unnecessary measurements-consuming, according to the requirements of different applications.

B. Laboratory Experiments

In laboratory experiments, we verify the practical feasibility of the scheme by detecting two kinds of object scenes (one is 2-dimensional (2D) picture printed on photo papers and another is a complex object scene with 3D objects as salient objects and 2D picture as background.). The experimental setup is schematically shown in Fig. 3. The laser with wavelength of 633 nm is expanded by a beam expansion. A DMD is used to modulate the laser beam. The size of all object scenes is $3.39cm \times 3.39cm$. And a charge-coupled device (CCD), which works as a single-pixel detector, is used to collect the intensity of light reflected from the object scene.

We set $M = N = 64$, $\eta = 4$ and $U = V = 8$. There are 4097 binary discrete cosine basis patterns (4906 $P_{B+}(x, y; u, v)$ and 1 $P_{B-}(x, y; 8, 8)$) with size 256×256 loaded into DMD for light modulation. By processing the measured light's intensities ($D_+(u, v)$ and $D_-(8, 8)$), the image with 64×64 pixels is reconstructed at 4 different sampling ratios (respectively 100%, 25%, 6.25%, 1.56%) of DCS which is segmented by 4 squares with different sizes (respectively 64×64, 32×32, 16×16, 8×8) in a similar way in Fig. 1. Before inputting the reconstructed images into the trained network for salient object detection, we choose to resize them to 128×128 by 'bicubic' image interpolation algorithm. The reason for upsampling is as follows. We have done simulation experiments similar with the one in Sec. III A. In simulation, we set $M = N = 64$ and $\eta = 4$. And the images of the test object scenes are reconstructed at 4 different sampling ratios of DCS. Other settings are the same as ones in Sec. III A. The size of test reconstructed image is 64×64. We compare the performances with and without upsampling. Moreover, we also consider the result of that the test reconstructed image without upsamling is input into the network trained by images with size 64×64 generated in similar way in Sec. II. From the MAEs shown in TABLE II, it shows better performance that reconstructed images are resized to 128×128 before input into the network trained by the method in Sec. II.

In our laboratory experiment, the measured $D_+(u, v)$ is introduced into some low-frequency noises which can be inferred from that the mean plane of surface in Fig. 4 (b) has a slow descending with the u and v increasing. Such noises result some pixels in the area nearing the top row and leftmost column of reconstructed image with high value, which makes the trained network can detect nothing because the contrast of input image is decreased as shown in Fig. 4 (a). To alleviate the influence of the noises, we utilize the symmetry of DCT and inverse DCT to design an simple denoising algorithm. First, we apply DCT to the measured $D_+(u, v)$ for getting $DD(u', v')$, which makes the noises more concentrated on the left and top boundaries of $DD(u', v')$. Then we use a method based on histogram statistics to generate a mask, of which the flowchart is shown in Fig. 5. Next, we use $MASK$ to get rid of the high values in $DD(u', v')$, which can be expressed as

$$DD'(u', v') = DD(u', v') \times MASK(u', v') \tag{7}$$

Then according to the direction represented by orange arrows in Fig. 5, we use the mean value of adjacent pixel values to sequentially make up the hollows in $DD'(u', v')$ and get the $DD''(u', v')$. The denoised $RD(u, v)$ with flat mean plane as is shown in Fig. 4 (c) can be obtained by applying inverse DCT to $DD''(u', v')$.

In our laboratory experiments, because the influence of noises is small at sampling ratios 1.56% and 6.25%, $RD(u, v)$ is obtained by Eq. 4 but no denoising. At sampling ratios

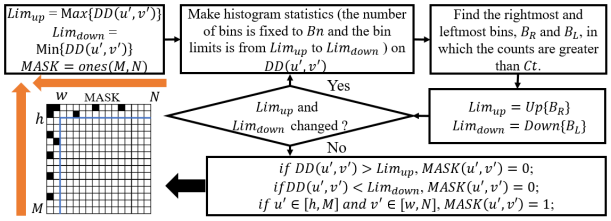

Fig. 5. Flowchart of mask generation. ($Max\{\bullet\}$ and $Min\{\bullet\}$ respectively denote returns of the max and min value of the matrix; $ones(M, N)$ denotes the generation of a $M \times N$ matrix with all value 1; Bn and Ct are constants preseted according to the data characteristic; $Up\{\bullet\}$ and $Down\{\bullet\}$ respectively denote returns of the upper and lower bounds of the bins; h and w is used to limit the scope in $DD(u', v')$ to be denoised).

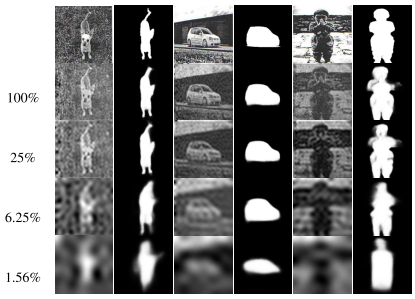

Fig. 6. The results of 2D object scenes. (The top row is object scenes and corresponding ground truths; the rest of rows are the reconstructed images at 4 different sampling ratios and the corresponding saliency maps).

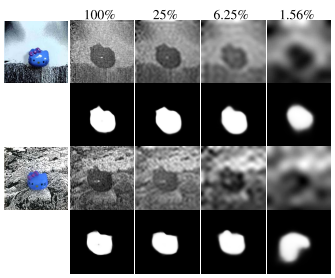

Fig. 7. The results of 3D object scenes. (The leftmost column is object scenes; the rest of columns are the reconstructed images at 4 different sampling ratios and the corresponding saliency maps).

25% and 100%, $RD(u, v)$ is directly obtained by applying the proposed denoising algorithm($Bn = 256$, $Ct = 8$ and $h = w = 3$) to measured $D_+(u, v)$. The reconstructed images and saliency maps of 2D object scenes are shown in Fig. 6. And the ones of complex object scene are shown in Fig. 7. The results show the feasibility and validity of our scheme to the practical 2D and 3D object scenes. As the results and the denoising process show, in the practical process of data measurements, the introduced noises may influence the quality of the reconstructed image. When the noises severely destroy the global information of the reconstructed scene like in Fig. 4 (a), the trained network may lose efficacy to the acquired scenes. But because of the excellent antinoise characteristic of DCT, some main noises could be alleviated by designing simple denoising algorithm. And from Fig. 6 and Fig. 7, we can see the trained network still works for the reconstructed images with some noises which don't greatly influence the global information.

IV. Conclusion

In this letter, we explore the implement of salient object detection based on SPI system and present a scheme via DCS acquisition and deep learning model. Fewer measurements for DCS are needed. And a CNN-based model is used to detect salient objects in the scene reconstructed from the undersampled data. The proposed scheme shows the potential of efficiency improvement, brought by salient object detection, of a SPI system for complex vision tasks. And experimental results and analyses also shows the flexibility of a single-pixel salient object detection system for adapting different application's need. When well-defined boundaries are needed, the more data can be measured in some applications such as image segmentation [16] and photo synthesis [17]. When only the sketchy location and region are enough, the more measurements can be reduced in some applications such as visual tracking [6] and object locating. It is believed that the proposed scheme provides a promising strategy for single-pixel salient object detection, and the single-pixel salient object detection has potential to promote the process of SPI applied on complex and practical tasks.

References

[1] M. P. Edgar, G. M. Gibson, and M. J. Padgett, "Principles and prospects for single-pixel imaging," *Nature Photon.*, vol. 13, pp. 13–20, Dec. 2018.
[2] X. Liu, J. Shi, L. Sun, Y. Li, J. Fan, and G. Zeng, "Photon-limited single-pixel imaging," *Opt. Express*, vol. 28, no. 6, pp. 8132–8144, 2020.
[3] Y. Yang, J. Shi, F. Cao, J. Peng, and G. Zeng, "Computational imaging based on time-correlated single-photon-counting technique at low light level," *Appl. Opt.*, vol. 54, no. 31, pp. 9277–9283, 2015.
[4] Q. Fu, Y. Bai, X. Huang, S. Nan, P. Xie, and X. Fu, "Positive influence of the scattering medium on reflective ghost imaging," *Photon. Res.*, vol. 7, no. 12, pp. 1468–1472, 2019.
[5] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR*, Jun. 2004, p. 2.
[6] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, "Adaptive object tracking by learning background context," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 23–30.
[7] C. Craye, D. Filliat, and J.-F. Goudou, "Environment exploration for object-based visual saliency learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 2303–2309.
[8] B.-L. Liu, Z.-H. Yang, X. Liu, and L.-A. Wu, "Coloured computational imaging with single-pixel detectors based on a 2D discrete cosine transform," *J. Modern Opt.*, vol. 64, no. 3, pp. 259–264, Feb. 2017.
[9] Z. Zhang, X. Wang, G. Zheng, and J. Zhong, "Fast Fourier single-pixel imaging via binary illumination," *Sci. Rep.*, vol. 7, no. 1, pp. 1–9, Dec. 2017.
[10] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3917–3926.
[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[12] L. Wang *et al.*, "Learning to detect salient objects with image-level supervision," in *Proc. CVPR*, Jul. 2017, pp. 136–145.
[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[14] A. Krizhevsky *et al.*, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[15] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. CVPR*, Jun. 2013, pp. 1155–1162.
[16] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *Proc. CVPR*, Sep. 2009, pp. 817–824.
[17] C. Goldberg, T. Chen, F.-L. Zhang, A. Shamir, and S.-M. Hu, "Data-driven object manipulation in images," *Comput. Graph. Forum*, vol. 31, no. 2, pp. 265–274, 2012.