Lensless opto-electronic neural network architecture for processing multi-color-channel signals

Wanxin Shi
Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University
Beijing, China
shiwx18@mails.tsinghua.edu.cn

Zheng Huang
Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University
Beijing, China
z-huang20@mails.tsinghua.edu.cn

Yuyang Han
Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University
Beijing, China
yy-han22@mails.tsinghua.edu.cn

Sigang Yang
Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University
Beijing, China
ysg@tsinghua.edu.cn

Hongwei Chen*
Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University
Beijing, China
chenhw@tsinghua.edu.cn

***Abstract*—The lensless opto-electronic neural network based on a multi-channel mask enables completing the convolution operations of multi-color-channel signals on the light field, significantly improving the task accuracy, eliminating the sensor components, and reducing energy consumption.**

Keywords—lensless, multi-channel, opto-electronic neural network, passive optical mask

I. Introduction

Conventional neural networks (CNN) have become instrumental in a variety of tasks ranging from image processing and computer vision to natural language processing [1]. However, it remains difficult to deploy a CNN in edge devices due to the tight energy and computational resources. Optical neural networks (ONNs) can increase computing speed and overcome the challenge of the electrical neural network [2], but most of them require a coherent laser and can not work in the natural light environment. On the other hand, many opto-electrical neural networks [3] consist of groups of lenses, which makes them difficult to use in edge devices. Our previous work has proposed a lensless opto-electrical network (LOEN) [4] architecture, which can complete feature extraction on the natural light field. Based on the LOEN, a new end-to-end scheme is proposed in this paper, which can effectively process multi-color-channel signals. When applied to a computer vision system, our method can only use a simple monochromatic sensor rather than an RGB sensor to process the RGB color image, eliminating some hardware components and saving computational consumption.

II. Principle and Method

As demonstrated in Fig.1, the object can be seen as a set of point light sources; the light from the object transmits through the mask onto the sensor, and when the size of the mask and the distance between the mask and the object are adjusted, the signal captured by the sensor (Y) is the convolution result of the object (I_{RGB})and the mask (PSF_{RGB}) [4]. The convolution kernel on the mask is the point spread function (PSF) of the optical system. However, the previous LOEN system can only process gray signals. In other words, it cannot perform different convolution feature extraction operations on different color channel inputs, even with an RGB sensor. However, many images in the real world are in color, and the color channels carry important information, which play an essential role in image classification and recognition tasks.

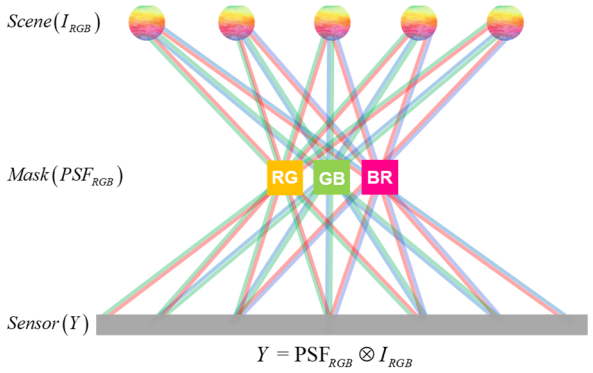

Fig. 1. Schematic diagram of the optical convolution. ⊗ represents convolution operation.

The proposed architecture in this paper can deal with multi-color-channel signals only by using a monochrome sensor. As shown in Fig.2, RGB images have three color channels, and the PSFs of different channels use filter films with different transmission spectra. The PSFs of three channels can be spatially overlapped, while the corresponding transmission spectra are superimposed. It can be proved that the output of the system is naturally the superposition of the corresponding three channels' convolution outputs, which is equivalent to the RGB three-channel input convolution operation in the electrical neural network.

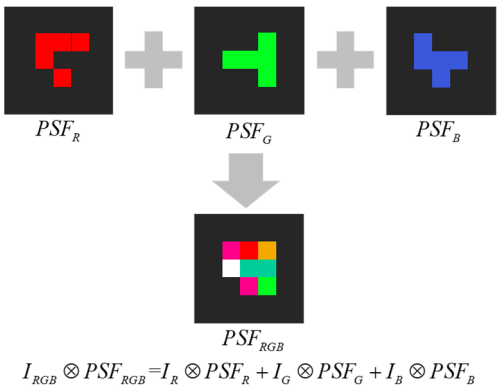

Fig. 2. Schematic diagram of the multi-color-channel optical mask.

The colorful images are composed of three channels: R, G, and B. Hence, the multi-channel mask consists of eight different filter films, such as filters that pass through a single channel of RGB, filters that pass through two channels in RGB, and fully transparent or completely opaque in RGB three channels. The completely opaque part is a chromium-plated film, and the fully transparent part should be a glass substrate without coating. The schematic diagrams of the transmission spectra of the other six filter films are shown in Fig.3. It should satisfy that the filter can pass through the corresponding light and the transmittance is relatively uniform. In contrast, the transmittance of the other bands should be as small as possible.

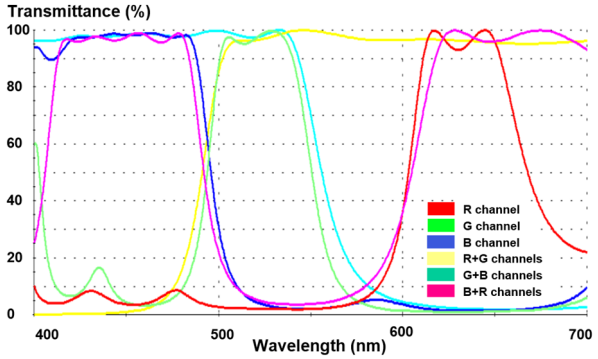

Fig. 3. Transmission spectrum schematic diagram of the mask's filter films in the system. The filter films have eight types; the other two settings are fully transparent in the visible light band (uncoated) and completely opaque in the visible light band (coated chromium film).

III. Simulation and Results

Fig.4 shows the end-to-end framework. It consists of four components: the RGB images in the natural scenes or displayed by a screen, the mask with filter films that implements the first convolution layer of the network, a monochrome sensor, and a digital processor that finishes the rest layers of the network. In order to achieve the highest network accuracy, the optical convolution kernels are joint-optimized with the digital layer. The architecture removes the first convolution layer from the electrical to the optical domain, making the system lensless and significantly reducing the size of the system. Meanwhile, the image signal processing (ISP) of the digital processor is also removed from the pipeline, which will reduce nearly half of the energy consumption of the sensor.

In this paper, we chose image classification as the vision task, where CIFAR-10 and Fruits 360 were used as the datasets, respectively. The architecture was evaluated on the Resnet network for CIFAR-10 classification, and the network used in the fruits classification task consisted of two convolution layers and three fully connected layers. The multi-channel (RGB) optical layer replaced the first convolution layer of the two networks.

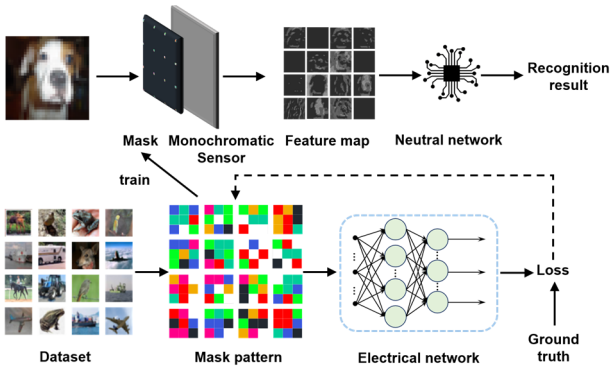

Fig. 4. Flow chart of the lensless optoelectronic neural network joint-optimazation architecture based on the multi-channel mask.

While dealing with multi-color signals, we can use multiple kernels under the condition of spatial reusability to increase classification accuracy. The results of the multi-channel convolution system are compared with that of the single-channel convolution system. The input of the single-channel system is grayscale images, while the input of the multi-channel system is RGB three-channel images. As shown in Fig.5 (a) and (b), the classification accuracy of the multi-convolution kernel system based on the multi-channel mask is significantly improved compared with that based on the single-channel mask. As for the CIFAR-10 task, the accuracies of the multi-channel mask system can achieve up to 93.17%, which have increased by 2.7475% on average compared to the same structure of single-channel mask systems. While for the fruits classification task, the accuracies based on multi-channel mask systems can achieve 97.13% and have an average 6.991% increase compared to single-channel mask systems.

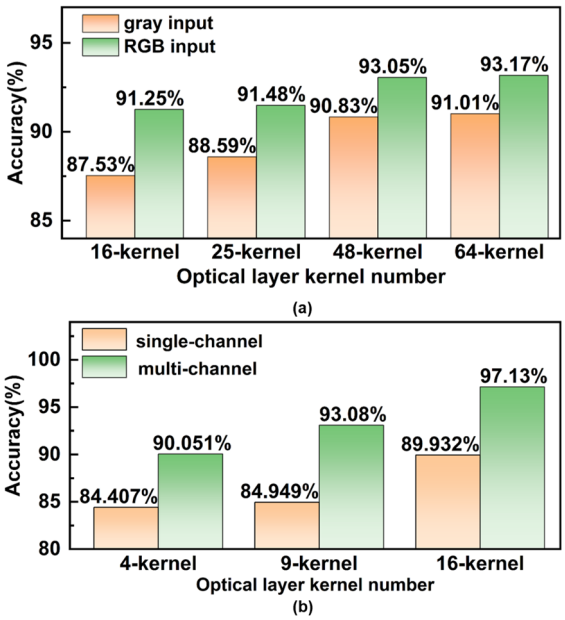

Fig. 5. (a) Results of the CIFAR-10 classification task. The 'kernel number' means the convolution kernel amount of the optical convolution layer in the multi-channel and single-channel mask systems. The 'single-channel' legend represents a single color channel mask system, while the 'multi-channel' is for a multiple color channels mask system. (b) Results of the Fruits 360 classification task.

Part of the confusion matrices of the classification results of CIFAR10 and fruits-360 are shown in Fig.6 and Fig.7,

respectively. The average precision and average recall of the corresponding tasks are shown in Tab.1 and Tab.2. For the CIFAR10 task, the average precision and average recall rate of multi-channel convolution are higher than the corresponding value of single-channel convolution system, the increase of average precision is up to 3.255%, and that of average recall rate is up to 3.72% (corresponding to the case of 16 convolution kernels systems). Compared with the single-channel system, the classification accuracy of the multi-channel system for cats, dogs and birds has improved. For the Fruits 360 task, the average precision and average recall of the multi-channel system with different numbers of convolution kernels is about 2%-6% higher than that of the single-channel system. In the 9 convolution kernels systems, the average precision increased by 5.58%, and the average recall rate increased by 6.61%. Compared with the single-channel system, the multi-channel system has improved the classification accuracy of different categories and different maturity levels apples.

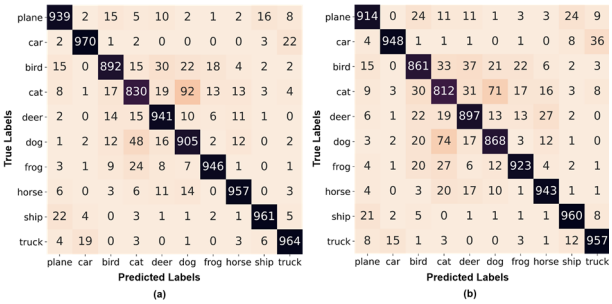

Fig. 6. Confusion matrices of the 64 kernel optical convolution systems for CIFAR10 task. (a) The multi-channel mask system. (b) The single-channel system.

TABLE I. AVERAGE PRECISION AND AVERAGE RECALL FOR CIFAR10 IMAGE CLASSIFICATION TASK.

optical layer kernel number	*average precision*	*average recall*
16 multi-channel kernels	91.29%	91.25%
16 single-chanel kernels	88.04%	87.53%
25 multi-channel kernels	91.46%	91.48%
25 single-chanel kernels	89.34%	88.59%
48 multi-channel kernels	93.07%	93.05%
48 single-chanel kernels	90.84%	90.83%
64 multi-channel kernels	93.16%	93.17%
64 single-chanel kernels	91.02%	91.01%

TABLE II. AVERAGE PRECISION AND AVERAGE RECALL FOR FRUITS 360 IMAGE CLASSIFICATION TASK.

optical layer kernel number	*average precision*	*average recall*
4 multi-channel kernels	91.17%	91.61%
4 single-chanel kernels	88.92%	86.68%
9 multi-channel kernels	95.31%	94.22%
9 single-chanel kernels	89.73%	97.61%
16 multi-channel kernels	97.76%	93.05%
16 single-chanel kernels	93.02%	91.64%

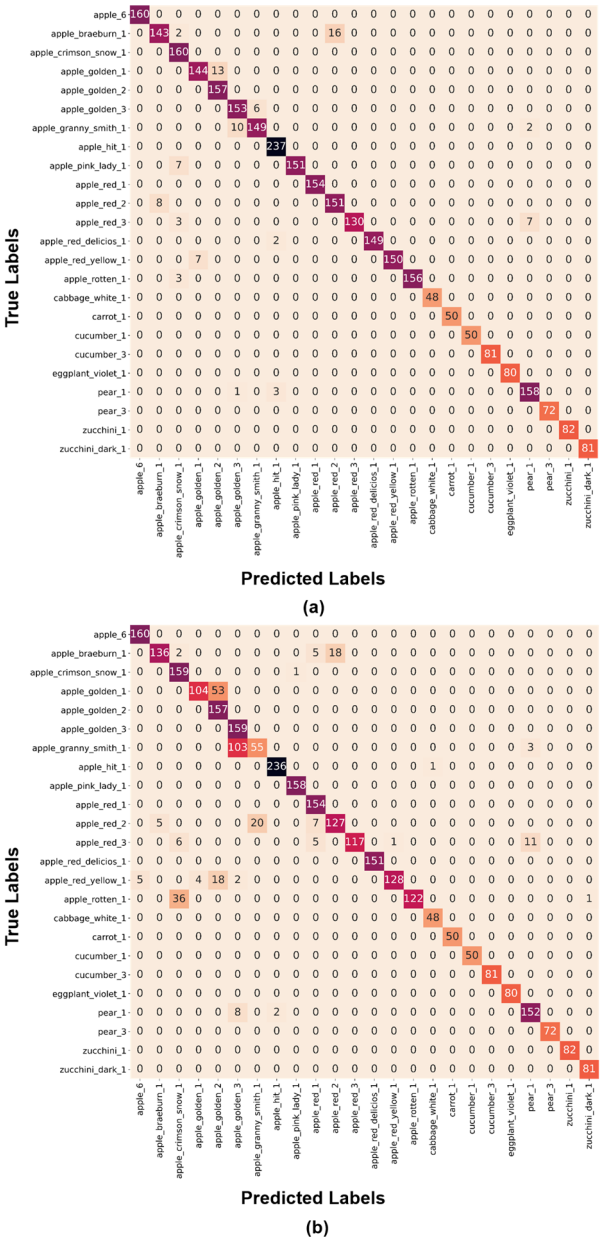

Fig. 7. Confusion matrices of the 16 kernel optical convolution systems for Fruits 360 task. (a) The multi-channel mask system. (b) The single-channel system.

IV. DISCUSSION

In this paper, a lensless opto-electrical neural network based on multi-color-channel mask is demonstrated. Through the mask with filter films, the processing and computing of multi-color signals can be completed on the light field while the size and weight of the system are greatly reduced. The system can significantly increase task accuracy compared with the single-channel mask system. Meanwhile, the architecture simplifies the sensor design, which can only use a simple monochromatic sensor to perform different convolution operations on different color channels signal rather than an RGB sensor, and reduce the energy and computation cost of the network and image signal processing (ISP) in the whole task pipeline.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (NSFC) (62135009), and by a grant from the Institute for Guo Qiang Tsinghua University.

REFERENCES

[1] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[2] Lin, X., Rivenson, Y., Yardimci, N. T., Veli, M., Luo, Y., Jarrahi, M., & Ozcan, A. (2018). All-optical machine learning using diffractive deep neural networks. Science, 361(6406), 1004-1008.

[3] Chang, J., Sitzmann, V., Dun, X., Heidrich, W., & Wetzstein, G. (2018). Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. Scientific reports, 8(1), 12324.

[4] Shi, W., Huang, Z., Huang, H., Hu, C., Chen, M., Yang, S., & Chen, H. (2022). LOEN: Lensless opto-electronic neural network empowered machine vision. Light: Science & Applications, 11(1), 121.