2017 17th International Conference on Control, Automation and Systems (ICCAS 2017)
Oct. 18–21, 2017, Ramada Plaza, Jeju, Korea

Understanding Visual Information with Compound Eye Camera

Hwiyeon Yoo, Geonho Cha, and Songhwai Oh

Department of Electrical and Computer Engineering and ASRI, Seoul National University,
Seoul, Korea ({hwiyeon.yoo, geonho.cha, songhwai.oh}@cpslab.snu.ac.kr)

Abstract: The compound eye, which is the eye structure of insects, has many advantages such as a large field of view, low aberrations, and an infinite depth of field. In this paper, we introduce a new hardware platform that mimics the eye structure of insects and propose some applications using the proposed platform. Specifically, we propose objectness estimation scheme and semantic segmentation scheme, which are important problems in computer vision, based on the compound images. In the objectness estimation task, we achieved the accuracy of 77.14% on a combined data set of PASCAL VOC 2012 and COCO-Stuff 10K data sets, and in the case of semantic segmentation, we achieved the mean IU of 0.432 on the simulated COCO-Stuff 10K data set.

Keywords: Compound Eye Camera, Compound Images, Objectness Estimation, Semantic Segmentation

1. INTRODUCTION

Many devices have been designed using inspirations of nature. Especially, we could design the refined camera structure inspired by the eye structure of insects. It has many advantages such as a large field of view (FOV), low aberrations, short image processing time, and an infinite depth of filed [1]. In this paper, we introduce a new hardware platform that mimics a compound eye of insects and propose some higher-level vision applications based on it. Specifically, we propose a objectness estimation scheme and a semantic segmentation scheme based on the compound images which are obtained from the proposed platform.

2. COMPOUND EYE CAMERA

We have designed a compound camera prototype which consists of six single-lens reflex cameras. The camera modules are on the hemisphere-shaped metal frame. Each camera module can capture 1280×960 size images at 24.6Hz. The proposed camera system is shown in Figure 1 (a). Time stamps of all cameras are synchronized with respect to the master camera which is at the center. These synchronized cameras emulate densely distributed single eyes on the hemisphere surface based on multi-view geometry as shown in Figure 1 (c). Here, each emulated single eye captures a rectangular low-resolution image, $e.g.$, 10×10 pixels. Emulated single eyes with the same polar angle are evenly distributed along the surface line of their latitude. Also, they have constant angular stride of the latitudes of single eyes.

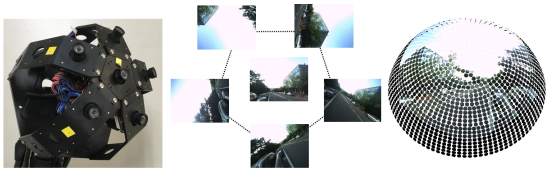

Fig. 1 (a) The proposed compound eye camera platform, (b) An example of images taken from the proposed compound eye camera platform, (c) An example of emulated densely distributed single eye images.

3. PROPOSED SCHEMES

3.1 Objectness Estimation

An overview of the proposed objectness estimation scheme is described in Figure 2. The emulated densely distributed single eye images which compose a compound image are locally flat and we call it as a compound eye data. We define region proposals on the hemisphere of a compound image which are candidates of object existence. For region proposals, we choose the k-nearest neighbor (k-NN) method with $k = 9$ among the single eyes since conventional bounding box method cannot handle the irregular distribution of single eyes on the hemisphere surface. A compound eye data and the proposed regions become inputs of the proposed objectness estimation network. The network consists of two steps. First, each of the single eye images is encoded to a 256-dimensional feature by an identical convolutional neural network (CNN). Then, in region convolutional network, information of each of the proposed regions are hierarchically merged by convolutions. The objective function of the objectness estimation network is designed as follows:

$$\text{Loss} = \|G - D\|_1 + (0.5 - G)^T D, \quad (1)$$

where G is the ground-truth and D is the estimated objectness score which is the output of the region convolutional network.

3.2 Semantic Segmentation

We introduce the proposed semantic segmentation scheme in this section. To leverage the conventional CNN scheme, we transform the compound image into a tensor representation by vectorizing each single eye image. This tensor represented compound image is fed into the proposed semantic segmentation network. The basic structure of the network is inspired by [2], and it consists of four convolutional layers. A leaky ReLU activation layer is followed after each convolutional layer. Note that in

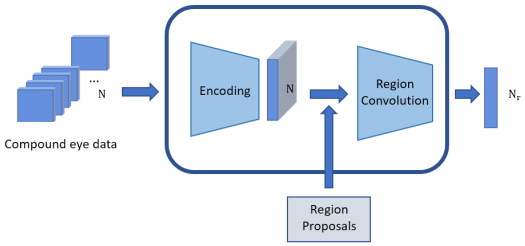

Fig. 2 An overview of the objectness estimation network. A compound eye data and the region proposals are fed to the network as inputs. It consist of two networks: an encoding network and a region convolutional network. The encoding network learns a feature embedding of each single eye. The region convolutional network learns to predict objectness based on neighboring single eyes.

the last layer, no activation layer is applied, but a pixel-wise softmax layer is applied. All convolution filters have the same size of 3×3, and the number of channels are 100, 50, 20, 5, respectively. Here, the number of channel of the last convolutional layer is the same as the number of semantic classes. The output of the network is pixel-wise confidence distribution which shows how much the single image is classified to different classes.

4. EXPERIMENTAL RESULTS

4.1 Objectness Estimation

We trained and evaluated the proposed network with the combined data set of the PASCAL VOC 2012 [4] and the COCO-Stuff 10K [3] data sets. Table 1 shows the results with various single eye sizes ($S \times S$ pixels). Here, the compound eye consists of 441 single eyes. It shows that it is important to find an appropriate overlap ratio between single eyes by varying their sizes. The configuration with a small single eye cannot cover the entire area of the original scene. On the other hand, too large single eyes are not effective since objects do not have smooth boundaries on the compound eye image. In our experiments, the configuration of $S = 10$ has the highest accuracy of 77.14% by covering the entire scene smoothly without missing patches. In this configuration, the angular stride of the latitudes of single eyes is 3 degrees, and 53% of a single eye region overlaps with neighboring single eyes on average.

For a baseline experiment, we divided 210×210 pixels size image into patches that have the same number and size of single eyes, $i.e.$, 21×21 patches with 10×10 pixels. As shown in Table 1, the proposed network achieved better accuracy by using overlapping single eyes.

4.2 Semantic Segmentation

To train the proposed semantic segmentation network, we simulated compound images with COCO-Stuff 10K data set. For this procedure, we assumed that the image was captured from the main camera. We selected four classes, and they are $things$, $ground$, sky, $structure$, and the other classes were considered as $background$.

Table 1 Performance of the proposed objectness estimation scheme with various sizes of single images

S (Pixel)	3	5	10	20	30	Baseline
Accuracy (%)	73.98	74.46	**77.14**	75.06	71.87	72.32

Table 2 Performance of the proposed semantic segmentation scheme with various sizes of single images

S (Pixel)	3	5	10	20	30	Baseline
Mean IU	0.421	0.427	**0.432**	0.425	0.416	0.364

We evaluated the proposed network with various single eye image sizes. In each case, the network structure is the same except the channel depth of the first layer which is $3 \times S \times S$. The comparison result is shown in Table 2. We can see that the best performance was achieved when the size of the single eye image is 10×10. To verify the merit of compound images compared to typical RGB images, we applied the proposed network to RGB images. For the fair comparison, we roughly cropped the RGB images to have the same visible regions compared to the compound images. After that, the cropped images were resized to the size of 210×210, which make an RGB image to have the same number of pixels compared to a compound image at $S = 10$, and every non-overlapping 10×10 patches were vectorized to make it suitable for the proposed network. The mean IU for this case was 0.364 which is much worse than the compound image cases.

ACKNOWLEDGMENT

This research was supported by a grant to Bio-Mimetic Robot Research Center funded by Defense Acquisition Program Administration and by Agency for Defense Development (UD130070ID).

REFERENCES

[1] E. Warrant and D.-E. Nilsson, *Invertebrate vision.* Cambridge University Press, 2006.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2015.

[3] H. Caesar, J. Uijlings, and V. Ferrary, "Coco-stuff: Thing and stuff classes in context," *arXiv preprint arXiv:1612.03716*, 2016.

[4] Everingham, Mark and Eslami, SM Ali and Van Gool, Luc and Williams, Christopher KI and Winn, John and Zisserman, Andrew, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98-136, 2015.