3D Shadow-Robust Visual Object Detection

Jui-Ting Wu and Chih-Hung G. Li*, *IEEE Member*

***Abstract*— When an Autonomous Mobile Robot (AMR) is deployed to work in different kinds of environments, the visual system of AMR is subject to uncontrolled illumination conditions. To enhance the adaptability of the AMR while performing object detection tasks, in this article, we proposed a deep learning framework that automatically enriches the training contents particularly in the effects of 3D shadows. The Convolutional Neural Network (CNN)-based object detector trained under the proposed framework then exhibits significantly improved robustness to strong illumination interferences. The proposed method and experimental results are reported.**

I. Introduction

Flexibility and adaptability are the two key aspects of the expected characteristics of the Autonomous Mobile Robot (AMR). When AMRs are deployed and work in the environments, they are expected to provide equally good performances under various environmental conditions (see Fig. 1). One challenging problem is the illumination condition, which may vary dramatically from place to place. Regarding the illumination effects, in addition to some typical lighting measures such as contrast, lumen, and chromaticity, shadows accompanying the objects are often a critical cause of detection failure. The irregular and uncontrolled geometry of the shadows makes it difficult to propose a general rule of image processing for the elimination of the adverse effect. Yet the nature of the shadow always places it near the object and often generates high contrast on the body of the object or the environment near it. When the pixel values are largely altered, the features of the target object become difficult to extract.

The deep learning framework provides engineers a path of training visual detection with data and achieving high success rates. The superior capability in generalization allows the deep learning framework to extract abstract features and thus exhibit high adaptability and flexibility in the works it performs. In this shadow interference problem, we proposed a deep learning approach by training the visual object detector with a large number of shadowed images. Simply put, we let the object detector be familiar with the possible appearances of the target object, should a strong light cast on the object and create a shadow near it. To automate the generation process of the shadowed images, we adopted a specific conditional GAN (Generative Adversarial Network) and trained it to be a generator of various shadow patterns. To our knowledge, this is the first time a deep learning framework was proposed to tackle the object detection problem under the strong influence of 3D object shadows.

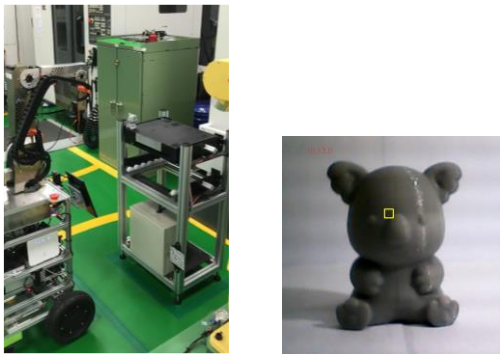

Fig. 1. An AMR is conducting object manipulation in the environment. The visual object detection system must show robustness on the shadow effect.

II. Related Works

A. Illumination Problems

The visual detection system may fail due to various factors, the most common one being the lighting and shadow. Under the influence of illumination, the performance of visual detection can be weakened. Among them, the 3D shadow of the object itself is particularly powerful in producing undesirable effects. Usually, a good lighting design is prepared, so sufficient contrast can be generated for the object to be detected by mark search and edge detection, so as to correctly identify the position of the object. However, good and consistent lighting design may not be practical for AMRs working in various environmental conditions. Some methods were developed to tackle the problem; e.g., [1] proposed a gradient-enhancing conversion method that produces a new gray-level image based on linear discriminant analysis. By updating the gray-level conversion vector dynamically, lane detection is performed using adaptive Canny edge detector, Hough transform, and the curve model-fitting method. [2] proposed a shadow detection and removal algorithm based on the normalized difference index (NDI) and morphological operation. The utilization of depth detection was also presented as an approach to reduce the shadow effect [3].

B. Visual Object Detection

Recently, general-purpose visual object detection is dictated by Convolutional Neural Network (CNN) –based

This work was supported by the Ministry of Science and Technology of the Republic of China, Taiwan under Contract No.: MOST109-2637-E-027-008-.

Jui-Ting Wu is with the Graduate Institute of Mechatronics Engineering, National Taipei University of Technology, Taipei, Taiwan, R.O.C. (phone: +886-2-2771-2171; fax: +886-2-2776-4889; e-mail: rei870410@gmail.com).

Chih-Hung G. Li* is with the Graduate Institute of Manufacturing Technology, National Taipei University of Technology, Taipei, Taiwan, R.O.C. (phone: +886-2-2771-2171; fax: +886-2-2776-4889; e-mail: cL4e@mail.ntut.edu.tw).

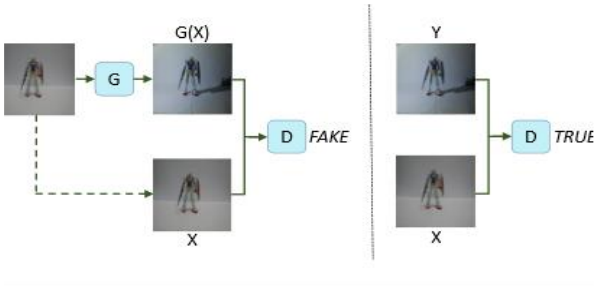

Fig. 2. Illustration of pix2pix GAN that is trained to superpose 3D shadow on an image of the object.

methods such as Yolov4 [4], SSD [5], FASTER R-CNN [6], Mask R-CNN [7], etc. To eliminate the need of manual annotation on the training images, OneShot [8], [9] was developed for precise localization of embedded objects. In the following test, we used OneShot as the visual detector for the evaluation of the proposed shadow-robust framework.

III. Proposed Method

When light beams cast on a 3D object, the part of the object facing the light becomes brighter and the back of the object becomes darker; in the meantime, shadows are formed behind the object on the ground. The strong illumination effects such as high contrast and large dark areas may confuse the visual detector and result in a detection error. Currently, many state-of-the-art visual detectors are based on the deep learning framework which obtains its superior generalization capability through the big training data. In other words, the cognition system is educated by the deliberately formed training set. If the scenarios subject to high illumination effects are not provided in the training set, it is very unlikely that the visual detector is capable of making the correct predictions under such effects.

To train the visual object detector to be shadow adaptable, we took the following approach. We acquired multitudes of photos of objects under various lighting conditions and used these images to enrich the training set of a CNN-based object detector. To automate the generation process of these shadow images, we proposed to adopt a special form of Generative Adversarial Network (GAN) [10] – pix2pix GAN [11] for learning to superpose the 3D shadow on a given image of an object. As shown in Fig. 2, pix2pix GAN is trained to generate the shadow pixels and superpose them onto the original image. During the training process, the discriminator of pix2pix GAN detects the fake images and makes the generator improve over time. Once the pix2pix GAN is well trained to generate the 3D shadow effects, it can be applied to the image of a new object to automatically produce an image of the new object under the same shadow effect. We then collect these images to train our object detector. Details of the shadow training process and the framework of the object detector are in the following.

A. *Training for 3D Shadow Adaptation*

To make the object detector adapt to different angles of light and shadow, we used pix2pix GAN as the image generation framework that produces shadowed images (see Fig. 3). Photos taken with a camera in the studio were used as the training images for pix2pix GAN (see Fig. 4). The camera used was Canon EOS-D60, with 6.3 million pixels and a maximum resolution of 3072×2048 pixels. The exposure time was set at 1/60 second and ISO-A. Photos of 50 different objects were obtained for each shadow patter; four shadow patterns were created with concentrated light casting at 45° and 90° on the left and 45° and 90° on the right.

For the training of each shadow pattern, the shadow images of 50 objects were input to pix2pix GAN. Each pattern training results in a template of a specific shadow pattern that can be applied to images of previously unseen objects for shadow creation. The computer for pix2pix GAN training has AMD Ryzen 9 3900X CPU with 4.1 GHz, 64 GB RAM, the graphics card is NVIDIA GeForce GTX 2080 TI, and the operating system is Windows 10. Averagely, it took around 12 minutes to complete training of a pix2pix shadow template.

B. *Object Localization Framework*

For object localization, we used OneShot [8], [9], for inference of object coordinates. To train OneShot for any new object, one only needs to place the object at the center of the image frame and take a photo. During the training process, automated image resizing and sliding operations will be performed to generate thousands of images with the object located at different positions with explicit spatial coordinates. These images are also automatically annotated with the coordinates and assembled as the training set for the CNN-based OneShot. Here, we define the horizontal direction as

Fig. 3. Images of the objects with 3D shadows were captured in a studio. These images were used to trained pix2pix GAN to be a 3D shadow generator.

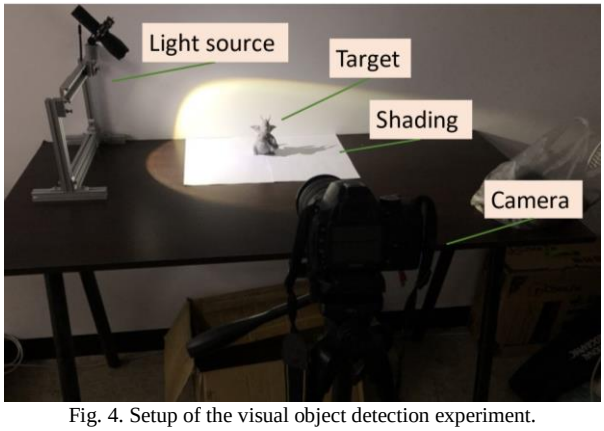

Fig. 4. Setup of the visual object detection experiment.

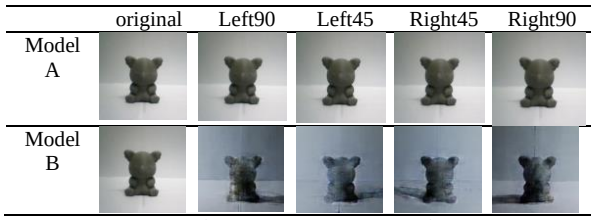

Fig. 5. Compositions of the training images used for object localization. Images for Model B include the ones generated by pix2pix 3D shadow templates, while Model A was trained with shadowless images.

the x-axis and the vertical direction as the y-axis and set the origin of the reference coordinates at the lower right corner with a maximum of (20, 20) at the upper left corner. Analogously, the center is (10, 10). Thus the target object can be detected with a resolution of 21 × 21 on the x-y plane. The architecture of OneShot is a light AlexNet and does not require the level of computation power like Pix2pix GAN, the computer used for OneShot testing has a 2.6 GHz Intel Core i7 CPU, 8 GB RAM, and a graphics card of NVIDIA GeForce GTX 1650 TI. The camera used for the object detection is the same as the one used for shadow generation by Pix2pix GAN, which is a 5.5 mm hose camera.

IV. Experiment

A. Model generation

As introduced in section III A, four shadow templates were obtained in pix2pix GAN, each one trained with images of 50 different objects. We then take a photo of the target object which was not included in the training of the shadow templates and applied the four shadow templates to the photo. The results are four images showing the target object under the four different shadow effects (see Fig. 5). We then combined the four shadowed images with the original image to form the basis image set for training of OneShot. The resulting object coordinate detector is denoted as Model B. We also created a benchmark model by training OneShot with the original shadowless images only. As shown in Fig. 5, Model A was trained with five sets of basic photos without shadows. Both Model A and Model B have a training set of 39,690 images.

B. Coordinate detection

To test the performance of object localization of the two models, hundreds of photos of the target object at different locations under various shadow effects were taken and used for testing. The LED flashlight was used to cast strong lights on the target object and create 3D shadows near the object. These images were then inputted into the two models for the prediction of the coordinates of the object.

To evaluate the prediction error, we divide the image into 21×21 grids; for each testing photo, the ground truth location of the object is manually identified and marked. The location predicted by OneShot was then compared to the ground truth; the distance between the predicted location and the ground truth was calculated.

Fig. 6. (a) Experimental results of Model A; (b) Experimental results of Model B. *Left*: ceiling light is on; *right*: ceiling light is off.

C. Experimental results

The experimental results are shown in Fig. 6. Each set of test results contain five categories of data pertaining to original light, 45° on the left, 90° on the left, 45° on the right, and 90° on the right. For each lighting category, 50 testing pictures were randomly selected for examination of the detection error defined in the previous section. Averages of the errors were calculated and displayed in Fig. 6.

It can be found that there are obvious differences in the performance between the two models. The average localization errors produced by Model B are clearly lower than Model A, particularly for those subject to strong lights. For the testing under the original lighting condition, both models render an average error of less than 2. When strong lights are introduced from various angles, the localization error by Model A jumped to around 8. On the contrary, the localization error by Model B was contained to under 5.

The result was not surprising, as Model A was trained with the original shadowless images; thus, its adaptability to the shadow effects is very limited. In contrast, Model B was trained with an image set containing the original photo and the ones reflecting a variety of shadow effects. One interesting thing worth noting is that the four basic photos with shadow effects were not real photos but the ones synthesized by a computer program. It verified the effectiveness of using pix2pix GAN for the automatic generation of 3D shadow effects on new objects.

The effect of the ceiling light is also demonstrated in Fig. 6. As one may find, when the ceiling light is on, both models exhibited slightly better performance. It is believed that the ceiling light provides more uniform lighting on the environment and the object; thus reduce the high contrast produced by the flashlight.

V. Conclusion

In an attempt to improve the performance of visual object localization subject to strong lighting effects. A deep learning framework was proposed, where 3D shadow templates were obtained by training pix2pix GAN and used for automatic generation of the training set of the CNN-based visual object detector. Our experimental evidence showed that the object detector trained with synthesized shadow images clearly performed better than the benchmark model trained without shadowed images. While strong lights from various angles deteriorate the localization accuracy, the proposed model successfully contained the localization error to a lower level. As the proposed model utilized shadow images synthesized by pix2pix GAN, the result verified the proposal of an automatic data augmentation process without manual effort for the collection of shadow images. It has the potential to be applied to various object detectors, including OneShot.

Acknowledgment

This work was supported by the Ministry of Science and Technology of the Republic of China, Taiwan under Contract No.: MOST109-2637-E-027-008-.

References

[1] H. Yoo, U. Yang, and K. Sohn, "Gradient-enhancing conversion for illumination-robust lane detection," *IEEE T. Intell. Transp.*, vol. 14, no. 3, pp. 1083–1094, 2013.

[2] O. Y. Agunbiade, S. M. Ngwira, T. Zuva, and Y. Akanbi, "Improving ground detection for unmanned vehicle systems in environmental noise scenarios," *Int. J. Adv. Manuf. Technol.*, vol. 84, pp. 2719–2727, 2016.

[3] S. S. Martínez, A. S. García, E. E. Estévez, et al., "3D object recognition for anthropomorphic robots performing tracking tasks," *Int. J. Adv. Manuf. Technol.*, vol. 104, pp. 1403–1412, 2019.

[4] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YoLo-v4: optimal speed and accuracy of object detection," arXiv:2004.10934.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. "SSD: Single shot multibox detector," arXiv:1512.02325.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2015, pp. 91–99.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," arXiv:1703.06870.

[8] C. G. Li, and Y. M. Chang, "Automated visual positioning and precision placement of a workpiece using deep learning," *Int. J. Adv. Manuf. Technol.*, vol. 104, pp. 4527 – 4538, 2019.

[9] Y. M. Chang, C. G. Li, and Y. F. Hong, "Real-time object coordinate detection and manipulator control using rigidly trained convolutional neural networks," in Proc. *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE 2019)*, 2019, pp. 1347–1352.

[10] L. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 1–9.

[11] P. Isola, J. Zhu, T. Zhou, A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.