Deep Photo Rally: Let's Gather Conversational Pictures

Kazuki Ookawara, Hayaki Kawata, Masahumi Muta, Soh Masuko, Takehito Utsuro, Jun'ichi Hoshino

Deep Photo Rally: Let's Gather Conversational Pictures

Kazuki Ookawara[1], Hayaki Kawata[1], Masahumi Muta[2],
Soh Masuko[2], Takehito Utsuro[1], and Jun'ichi Hoshino[1]

[1] University of Tsukuba, Graduate School of Systems and Information Engineering,
1-1-1, Tennodai, Tsukuba-shi, Ibaraki, Japan
{okawara.kazuki, kawata.hayaki}@entcomp.esys.tsukuba.ac.jp,
utsuro@iit.tsukuba.ac.jp, jhoshino@esys.tsukuba.ac.jp
[2] Rakuten, Inc., Rakuten Institute of Technology,
Rakuten Crimson House, 1-14-1, Tamagawa, Setagaya-ku, Tokyo, Japan
{masafumi.muta, so.masuko}@rakuten.com

Abstract. In this paper, we propose an anthropomorphic approach to generate speech sentences of a specific object according to surrounding circumstances using the recent Deep Neural Networks technology. In the proposal approach, the user can have pseudo communication with the object by photographing the object with a mobile terminal. We introduce some examples of application of the proposal approach to entertainment products, and show that this is an anthropomorphic approach capable of interacting with the environment.

Keywords: Augmented Reality, Anthropomorphic, Deep Neural Networks.

1 Introduction

Humans can feel familiarity and connection to personified objects [1]. In order for humans to personify something, expressions interacting with the environment (effectance motivation) are considered important [2]. In recent years, the Deep Neural Networks technology makes it possible to handle complex information such as the state of an object and its surrounding environment.

In this paper, we propose an approach to anthropomorphize a specific object using the state-of-the-art technology of Deep Neural Networks. The anthropomorphization of an object in the proposal approach can be realized by speech expressions suitable for the state of the object and environment. A user can communicate with an anthropomorphized object by photographing the specific object with a mobile terminal. Finally, we introduce examples of application of the proposal approach to entertainment products, and show that it is feasible to realize an anthropomorphization approach capable of interacting with the environment.

2 User Interaction

Fig. 1 shows the relatedness between a user and an object through the proposal approach. In the proposal approach, it is possible to anthropomorphize a specific object the user photographed with a mobile terminal. The object generates speech sentences based on the state of the object and the surrounding environment. The user receives organically changing speech sentences based on the environment. Then the object is anthropomorphized, and the user can feel familiarity with the object.

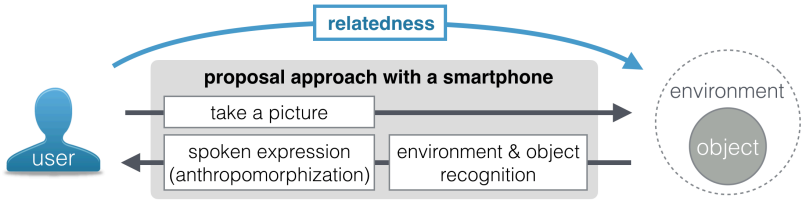

Fig. 1. The relatedness between a user and an object.

3 Anthropomorphization based on Environment Recognition

3.1 Recognition of environment and object state

The object recognition technology in the proposal approach adopts YOLO 9000 [3] which is the state-of-the-art technology capable of detecting an object at high speed.

Fig. 2 shows the network configuration of the actually adopted YOLO 9000. YOLO 9000 realizes extraction of input image features, detection of object regions, and estimation of object labels with a single Neural Network. Image features (information on the state of the object and environment) are extracted through multiple network layers. The image features are also used for speech generation described in Section 3.2. Object region coordinates and object label candidates are stored in the output layer. Multiple objects can be detected simultaneously by calculating the probability of the object candidates. For details of the model, see the document [3].

3.2 Speech generation for anthropomorphization

The technology of Neural Network (i.e. Google NIC [4]) which automatically generates syntax from images has been proposed recently. In this paper, we adopt Google NIC as a speech generation technology for an object based on environment recognition.

Fig. 2 shows a simplified network configuration of Google NIC. This network is constructed very simply: it is realized by network layers having such roles as image feature extraction, feature conversion, and syntax generation. The proposal approach aims to make a speech generation process more efficient by diverting the image features extracted by YOLO9000. By calculating the probability of the estimated word string, it becomes possible to generate a speech sentence reflecting the state of the object and environment. For details of the model, see the document [4].

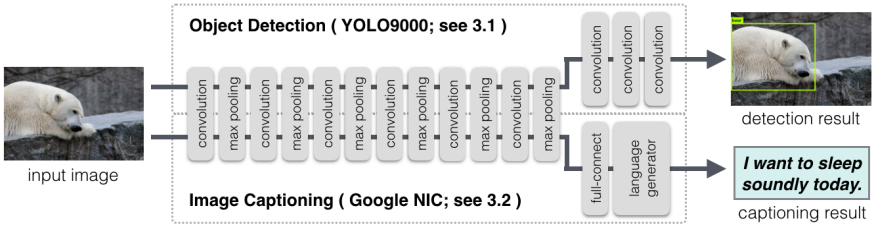

Fig. 2. The proposal approach for anthropomorphization.

3.3 Building a network model

To build a network model in the proposal approach, you need to prepare an image containing an object to be recognized and a speech sentence of the object. You also need to prepare a set of object labels and area coordinates.

The proposal approach begins with learning of the object detection model (YOLO 9000) to establish a network that extracts image features at the beginning. After that, it performs learning of the speech generation model (Google NIC) by diverting the image features extracted by the object detection model.

The application example of the proposal approach introduced in this paper uses a large scale image data set called Microsoft COCO (MSCOCO) [5] as learning data. MSCOCO is a data set with arbitrary object labels and captions provided for the image. This time, we have changed the captions to speech sentences of the image with arbitrary label combinations, and made it learn them.

4 Application Example

4.1 Photo rally at the zoo

The Photo Rally System is shown as an entertainment product example in which the proposal approach has been augmented for a zoo (Fig. 3). In this system, you can photograph animals speaking based on the state of their own (i.e. eating, running) as well as their surrounding environment (i.e. people are gathered, staying near the water). You can also incorporate gamification factors such as creating your original animal encyclopedia, by using the object detection function.

4.2 Play house with plush doll

You can enjoy the augmented play house by anthropomorphizing the plush doll with the proposal approach. Fig. 4 shows the Play House System to which the proposal approach applies. Fig. 4 shows how the teddy bear is speaking according to the surrounding environment. When a user changes the surrounding environment, the teddy bear's speech also changes according to the environment. This system enables a user to have pseudo communication with a plush doll by operating the doll and its surrounding environment.

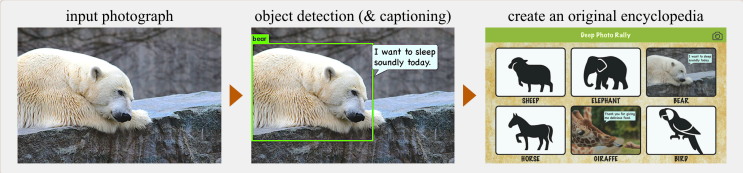

Fig. 3. The photo rally system with anthropomorphization.

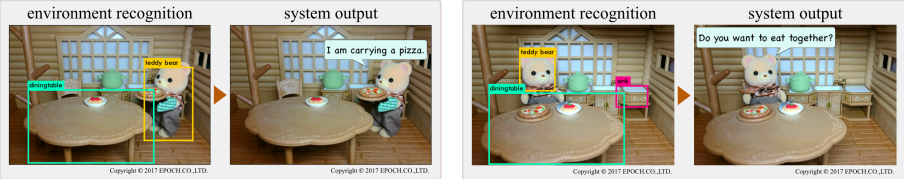

Fig. 4. The play house system with anthropomorphization.

5 Conclusion

In this paper, we have proposed an approach to anthropomorphize a specific object using the state-of-the-art technology of Deep Neural Networks. The anthropomorphization of an object in the proposal approach can be realized by speech expressions suitable for the state of the object and environment. A user can communicate with an anthropomorphized object by photographing the specific object with a mobile terminal.

The proposal approach adopts the Deep Neural Network technology, including object detection and image captioning. We can apply this technology to various products by utilizing the results. This paper could confirm the operation of Photo Rally at the zoo and Play House with plush doll as concrete examples. Photo Rally System is capable of creating an original animal encyclopedia by utilizing the object detection results. Play House System is capable of ensuring pseudo communication between a user and a plush doll by recognizing the environment of the doll. From here, we would like to examine the “effect of playing” in the example already described in this paper.

References

1. Waytz, A.: Social connection and seeing human. The Oxford handbook of social exclusion, pp. 251–256 (2013).
2. Epley, N., Waytz, A., Cacioppo, J. T.: On seeing human: a three-factor theory of anthropomorphism. Psychological review, 114(4), pp. 864-886, (2007).
3. Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger. arXiv preprint arXiv:1612.08242 (2016).
4. Vinyals, O., et al.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156-3164 (2015).
5. Chen, X., et al.: Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015).