Voice-Input Multimedia Information Retrieval System Based on Text-to-image GAN

Rintaro Yanagi†, Ren Togo‡, Takahiro Ogawa‡, Miki Haseyama‡
†*Graduate School of Information Science and Technology, Hokkaido University*
‡*Faculty of Information Science and Technology, Hokkaido University*
N-14, W-9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan
{yanagi, togo, ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

***Abstract*—In this paper, we develop an integrated multimedia information retrieval system. By utilizing text-to-image Generative Adversarial Network and image-to-text model, the developed system enables users to retrieve an objective content utilizing voice as an input with high accuracy. Experimental results show the effectiveness of the developed system.**

I. Introduction

With the recent availability of a large variety of personal devices, users can easily access to multimedia contents and upload them through various Web services. Following this accessibility, the number of multimedia contents has been increasing. To utilize these large amounts of data effectively, retrieving multimedia contents is one of the most important tasks.

If we realize a retrieval system that can retrieve objective contents easily and accurately, it can be considered that we can utilize its system in various way. For instance, we not only easily retrieve objective multimedia contents but also describe what we want to say to others with showing actual examples. When we describe interesting parts of a movie to others, we can convey these interests by showing actual parts of the movie. It can be considered that such examples can be realized easily with the recent growth of IoT and increase of devices that can easily access multimedia data utilizing voice as an input (such as Google Home [1], Alexa [2] and so on).

However, the integrated retrieval system that can retrieve objective contents easily and accurately utilizing voice as an input is not still developed. Then we develop the above-described retrieval system utilizing text-to-image Generative Adversarial Network (GAN) [1]. What we want to achieve is illustrated in Fig. 1. The developed system consists of three steps, voice analysis, query preparation and retrieval. As a preparation, the developed system generates texts from candidate contents utilizing image-to-text model. First, the developed system receives a sentence as a form of voice. The developed system translates its voice to a sentence utilizing google cloud speech to text API[3] and utilizes its sentence as an input query. Next, the developed system generates an image from the input query sentence utilizing text-to-image GAN. Then the developed system calculates visual similarities and textual similarities based on the generated image and the generated texts parallelly as described in [2]. Finally, the developed system combines the above-described visual similarities and textual similarities, and extracts the most similar content. Summarizing the above, the developed system utilizes voice as an input and outputs a multimedia content that is related to the input voice. This system enables users to retrieve multimedia contents easily and accurately.

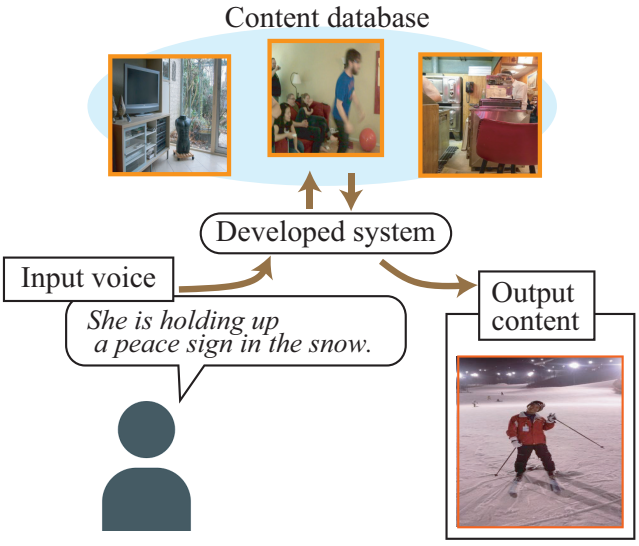

Fig. 1. Illustration of the scenario that we try. From input voice, we retrieve an objective content from a database.

II. Our Multimedia Information Retrieval System

Components and overview of the developed system are shown in Fig. 2. As a preparation, we generate texts that represent each candidate content. First, we make input voice into a sentence utilizing google cloud speech to text API and define it as a query sentence.

Next step consists of two flows, visual similarity calculation and textual similarity calculation. In visual similarity calculation, we generate an image utilizing text-to-image generation network based on AttnGAN [1]. Then we calculate cosine similarities between visual features extracted from the generated image and candidate contents. Here, we utilize the outputs of the third pooling layer of Inception-v3 [3] as visual features. On the other flow, we calculate cosine similarities between textual features extracted from the query sentence and the generated texts. Here, we utilize Sent2Vec [4] for the textual feature extractor.

Finally, we combine similarities calculated by visual features and textual features and present a content that has the highest similarity among all of the candidate contents to user.

III. Experimental Results

To evaluate retrieval performance of the developed system, we conducted a qualitative evaluation focusing on image retrieval task. From a test set of Common Objects in Contexts (COCO) dataset [5]

This work was partly supported by the MIC/SCOPE #181601001.

[1]https://store.google.com/product/google_home

[2]https://www.amazon.com/Amazon-Echo-And-Alexa-Devices/b?ie=UTF8&node=9818047011

[3]https://cloud.google.com/speech-to-text/

978-1-7281-3575-5/19/$31.00 ©2019 IEEE

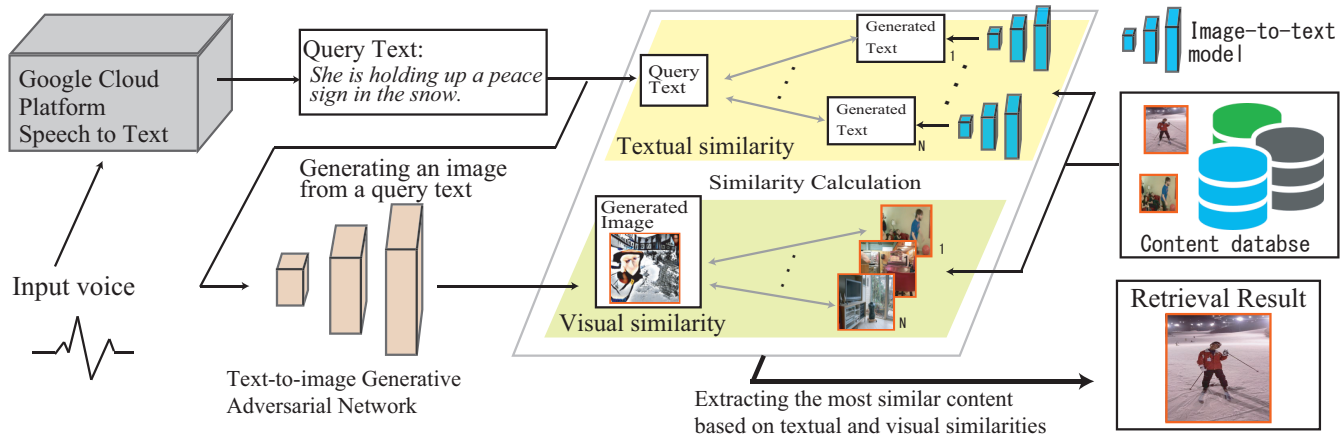

Fig. 2. Overview of the developed system. From input voice, we calculate a query sentence and generate an image utilizing text-to-image GAN. Then we calculate textual and visual similarities. Finally, we extract the most similar content based on these similarities.

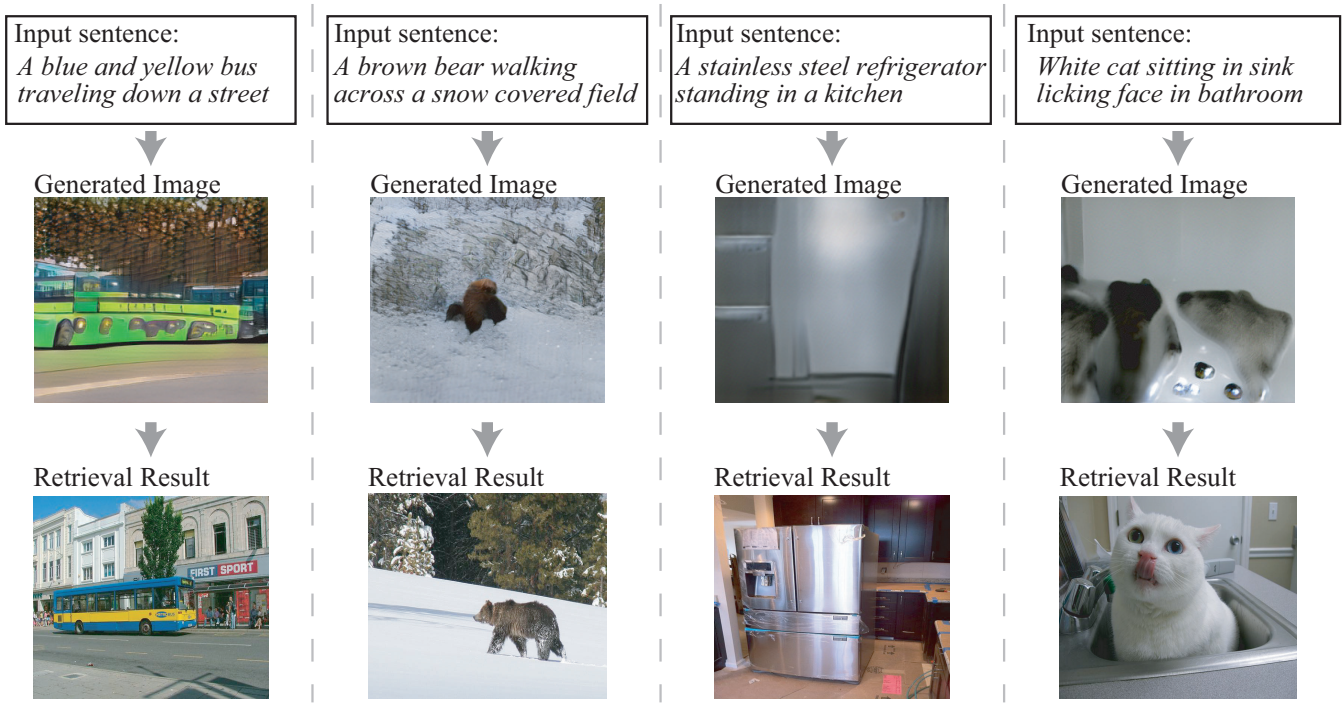

Fig. 3. Retrieval Results of the developed system. Each result is retrieved based on each input sentence and the generated image.

that contains daily scene images and descriptions for each image, we randomly selected 3,000 images and constructed database. Then we retrieve an image from the database utilizing the description annotated to the image in the database as a query. We show the above-mentioned retrieval results in Fig. 3. We can see that the developed system can retrieve objective contents.

IV. Conclusions

In this paper, we develop a novel retrieval system that utilizes text-to-image GAN. By calculating visual features and textual features, we realize multimedia information retrieval with high accuracy.

References

[1] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1316–1324.

[2] R. Yanagi, R. Togo, T. Ogawa, and M. Haseyama, "Scene Retrieval Using Text-to-image GAN-based Visual Similarities and Image-to-text Model-based Textual Similarities," in *Proceedings of the 2019 IEEE 8th Global Conference on Consumer Electronics, GCCE 2019; (Submitted)*.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2818–2826.

[4] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features," *arXiv:1703.02507*, 2017.

[5] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of the IEEE European Conference on Computer Vision*, 2014, pp. 740–755.