Imaginary Soundscape : Cross-Modal Approach to Generate Pseudo Sound Environments

Yuma Kajihara[1, 2], Shoya Dozono[1], Nao Tokui[1]
[1]Qosmo inc.
[2]The University of Tokyo
yumakajihara@g.ecc.u-tokyo.ac.jp
{dozono, tokui}@qosmo.jp

Abstract

We propose a new interaction technique when viewing panoramic photos, that causes an experience of hearing pseudo environmental sounds, by using a cross-modal model between sounds and images. We developed a system, which extracts a cross-modal feature from an image of "Google Street View", and searches a sound file matching with the image. Our project can have us reconsider the concept of "soundscape" from the aspect of computational creativity. Users can walk around in Street View and listen to the "imaginary" soundscape at our website.

1 Introduction

Many types of research have studied to extract cross-modal features of latent space, especially features between sounds and images is one of the most popular. Aytar et al. [1] and Hong et al. [2] extracted cross-modal features using a model trained with a lot of videos, by unsupervised learning. They succeeded this task thanks to the excellent discriminating ability of pre-trained convolutional neural networks(CNNs) for images.

Since Schafer [3] proposed the concept of "soundscape", it has been important to reconsider the relationship between our living environment and sound surrounding it, among many kinds of designers and artists. "Imaginary Landscapes" [4], the series of electronic music composed by Cage, is the representative example. In deep learning research, Owens et al. [5] also focused on ambient sounds as supervisory signals for learning visual models.

In this work, we developed a system to generate a pseudo soundscape by inferring an environmental sound from an image of "Google Street View", using CNN based on "SoundNet" [1]. We show the overview of our system in Section 2 and discuss how to interpret a soundscape experience in terms of computational creativity, in Section 3.

It is possible to view this work at the following URL.

`http://imaginarysoundscape.qosmo.jp/`

2 System Overview

The overview of our proposed system is shown as Figure1. There are two main steps in order to generate an "Imaginary Soundscape" from a panoramic image.

1. To extract cross-modal features between audio and image from an image.
2. To search the best sound file matching with the extracted features.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

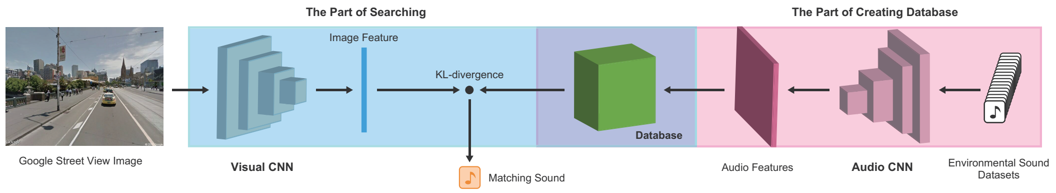

Figure 1: **The overview of our proposed system** : The part of Searching is run on a browser side.

In order to extract cross-modal features, we trained CNN for audio unsupervisedly, based on "SoundNet". The difference from the previous model is that we didn't use ImageNet CNN and trained so that the probabilistic distribution of the output tensor becomes close to Places365 CNN's [6].

Using this CNN, we created a database which includes a lot of files of environmental sounds and cross-modal features of them. While users look at panoramic images, visual-domain CNN extracts a feature from the image and a sound file, and a sound is chosen, which has the most similar distribution with the image feature.

In order to make the interactive sound search run on web browsers, we used lighter "SqueezeNet" architecture [7] instead of original Places365 CNN model.

3 Discussion

A sound selected by our system will be different if we use another visual CNN instead of Places365. When we tried using ImageNet CNN for extracting features from an image, sounds were chosen in response to local objects, such as a police car, a traffic light and etc ... We use Places365 CNN because it can capture more global features, but it would be interesting to train and evaluate the model for audios with various kinds of visual CNNs.

When we imagine something unknown, we associate it with something we have seen or heard. For instance, We can also think of environmental sounds or smell around unknown places while browsing "Google Street View". We believe a cross-modal model can enhance our imagination, by relating information we perceive with imaginary information we cannot directly perceive. Our system can be said to expand the experience of looking at panoramic photos by means of generating "Imaginary Soundscape".

In this work, we proposed a system to estimate environmental sounds from panoramic images, based on cross-modal CNN. The expression of our current system is limited to the number of sound files on the database. For overcoming that, we will seek the method to develop a generative, cross-modal model for high-quality environmental sounds.

References

[1] Y. Aytar, C. Vondrick & A. Torralba. SoundNet: Learning Sound Representations from Unlabeled Video. In *NIPS*, 2016.

[2] S. Hong, W. Im & H. S. Yang. Content-Based Video-Music Retrieval Using Soft Intra-Modal Structure Constraint. In *IEEE Transactions on Multimedia (TMM)*, 2017.

[3] R. M. Schafer. The Soundscape: Our Sonic Environment and the Tuning of the World, 1977.

[4] J. M. Cage. Imaginary Landscapes, No.1 - 5, 1939 - 1952

[5] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman & A. Torralba. Ambient Sound Provides Supervision for Visual Learning. In *ECCV*, 2016.

[6] B. Zhou, A. Lapedriza, A. Khosla, A. Torralba & A. Oliva. Places: A 10 million Image Database for Scene Recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[7] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally & K. Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. In *Arxiv*, 2016.

Supplementary Materials

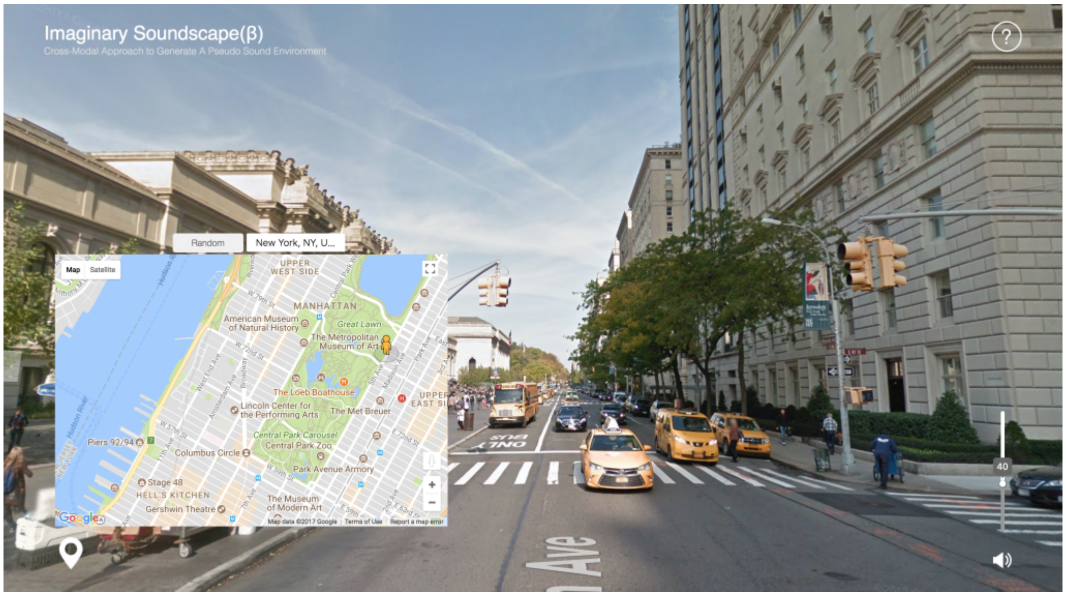

Figure 2: The project page of our work