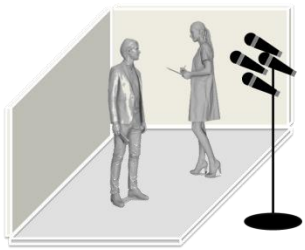
18 Estimating objects' visuals only from audio

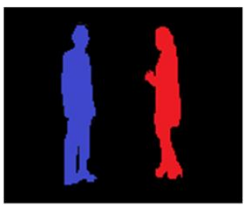
- Cross-media scene analysis -

Abstract

By bringing different senses together, people are able to achieve complicated mappings between natural sounds and visual images of their physical sources -- What about machines? We addressed this new task, which has never been done before, and attempted to develop a cross-media scene analysis method that can predict what objects are where in a scene from auditory information alone, i.e., without actually looking at the scene. Our method is based on a deep neural network (DNN) that is designed and trained to predict image recognition results based on acoustic features obtained through a microphone array. Our method allows users to visually check the state of the scene, even in cases where we cannot or do not want to use a camera. This property will contribute to expanding the opportunity to watch and track applications in living and public environments and monitor applications.



2. Predicting object types and locations based on sounds

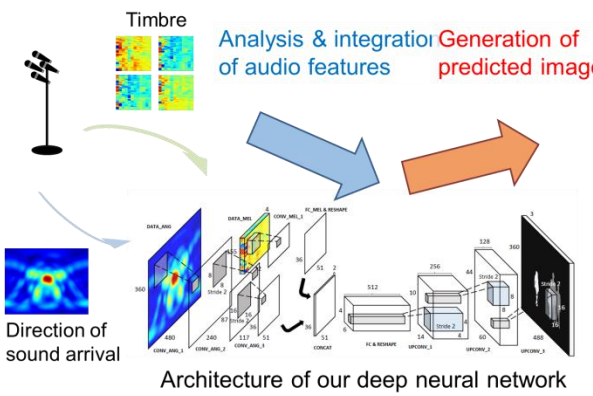


1. Recording sounds by microphone array

3. Visualizing predicted scene

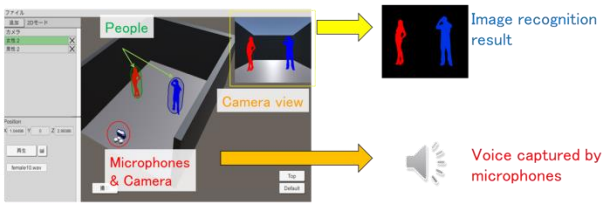
Cross-media DNN for visualizing object types and locations from acoustic features

Integration of audio features and image generation are performed through a single deep neural network.



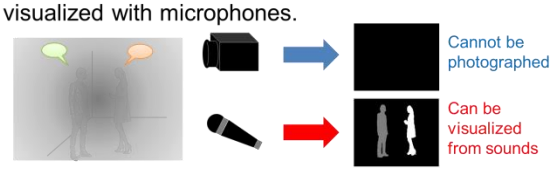
Architecture of our deep neural network

Artificial data augmentation for facilitating learning and validation



Visualizing scenes where photography is prohibited or impossible

Dark rooms or spaces with high privacy levels can be visualized with microphones.



References

[1] M. Ostrek, G. Irie, H. Kameoka, A. Kimura, K. Hiramatsu, K. Kashino, “Seeing through Sounds: Visual Scene Analysis from Acoustic Features,” *Meeting on Image Recognition & Understanding*, 2017. (in Japanese)

Contact

Go Irie Recognition Research Group, Media Information Laboratory