22 Recognizing types and shapes of objects from sound

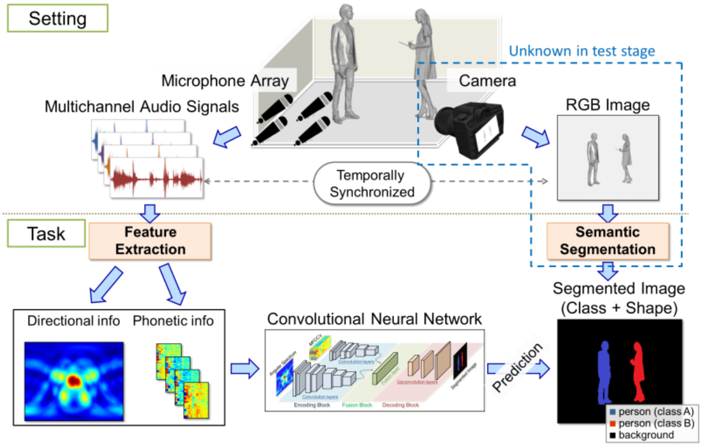
- Crossmodal audio-visual analysis for scene understanding -

Abstract

Sounds provide us with vast amounts of information about surrounding objects and scenes and can even remind us visual images of them. Is it possible to implement this noteworthy ability on machines? We addressed this task and developed a crossmodal scene analysis method that can predict the structures and semantic classes of objects/scenes from auditory information alone, i.e., without actually looking at the scene. Our approach uses a convolutional neural network that is designed to directly output semantic and structural information of objects and scenes by taking low-level audio features as its inputs. An efficient feature fusion scheme is incorporated to model underlying higher-order interactions between audio and visual sources. Our method allows users to visually check the state of the scene even in a case where they cannot or do not want to use a camera. Our method will contribute to expanding the availability of monitoring applications in various environments.

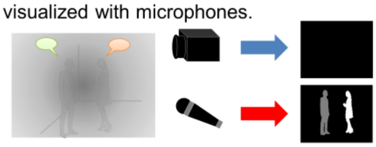
Crossmodal Scene Understanding

Predicting structures and semantic classes of objects from multi-channel audio signals



Visualizing scenes where photographing impossible or prohibited

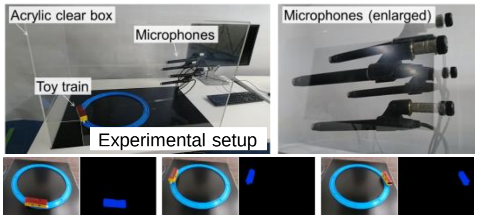
Dark rooms or spaces with high privacy levels can be visualized with microphones.



Cannot be photographed

Can be visualized from sounds

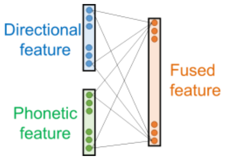
Demonstrated recognition of limited number of object classes possible so far



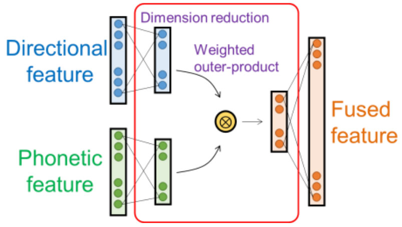
Features

Feature fusion layer for efficiently modeling higher-order interactions between audio and visual sources

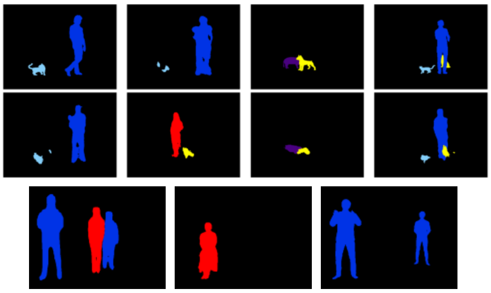
Typical fusion scheme has prohibitive number of parameters!



Our approach reduces the number of parameters by considering weighted outer-product of lower-dimensional features



Recognition of various classes of objects from real sound sources possible



References

[1] G. Irie, M. Ostrek, H. Wang, H. Kameoka, A. Kimura, T. Kawanishi, K. Kashino, "Seeing through sounds: predicting visual semantic segmentation results from multichannel audio signals," in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019.

[2] H. Wang, G. Irie, H. Kameoka, A. Kimura, K. Hiramatsu, K. Kashino, "Audio-based Semantic Segmentation based on Bilinear Feature Fusion," Meeting on Image Recognition and Understanding (MIRU), 2018.

Contact

Go Irie Email: cs-liaison-ml at hco.ntt.co.jp
Recognition Research Group, Media Information Laboratory

Innovative R&D by NTT
Open House 2019