Cross-media Scene Analysis: Estimating Objects' Visuals Only from Audio

Go Irie, Hirokazu Kameoka, Akisato Kimura, Kaoru Hiramatsu, and Kunio Kashino

Abstract

Human beings can get a visual image of the surrounding environment from sounds they hear. Can we give similar capabilities to computers? In this article, we introduce our recent efforts in cross-media scene analysis applied to estimate the type, location, and visual shape of objects in a scene based only on sound sources recorded with multiple microphones.

Keywords: cross media, scene analysis, deep learning

1. Introduction

The success of deep learning has completely changed the framework of media processing and recognition. Deep learning has already delivered superhuman performance that can be applied to address many problems in image and audio recognition. More important is that although media processing technologies for different types of media (e.g., images, video, audio, sound, and language) had at one time mostly been studied and developed independently, they are now being looked at together as their frameworks are similar.

In this article, we introduce cross-media scene analysis technology that can predict image recognition results only from sound information. As the number of people who are interested in security/safety increases day by day, the importance of surveillance and crime prevention technology is increasing. Most such technologies achieve excellent performance by leveraging advanced image recognition techniques. However, such technologies are probably not applicable in very dark places or rooms that have many blind spots, or in private or public spaces where privacy is prioritized and cameras are prohibited.

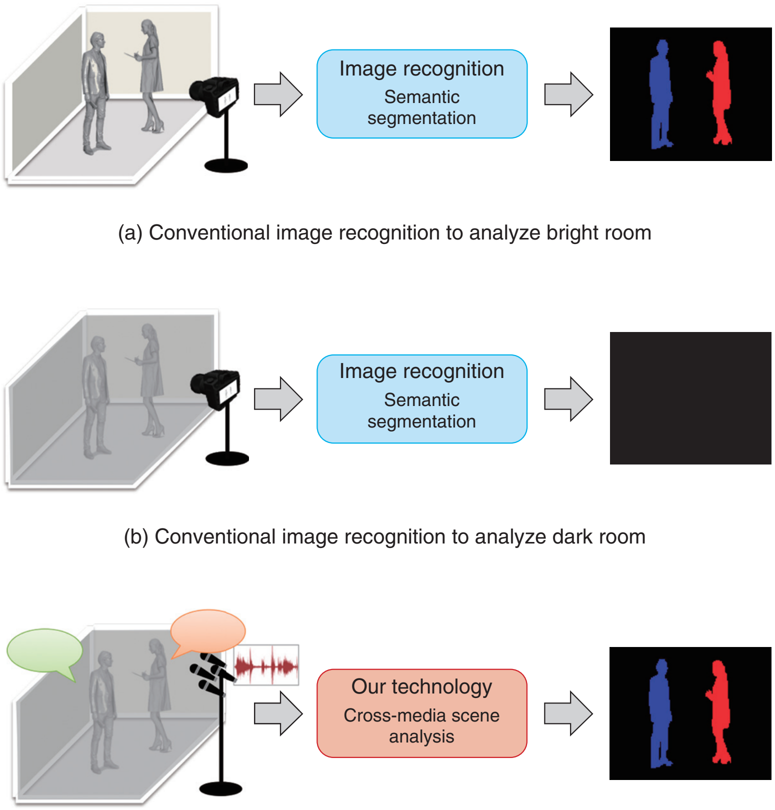
Our technology aims to provide visual recognition functionality without using any camera devices. This will make it possible to provide safe and comfortable monitoring even in those places where normal image recognition technologies cannot be used.

2. Cross-media scene analysis

Human beings recognize and understand their surrounding environments using their eyes and ears. More interestingly, we integrate and use these signals cross-sectionally to estimate the states of a scene. For example, when we are walking on a road and suddenly catch the sound of a car coming from behind us, we can infer how far away it is and possibly even what type of car it is without actually turning around and looking at the car.

What we aim to do here is to equip a computer with this kind of human ability. More specifically, the technology introduced in this article is designed to predict an image recognition result as if it had been photographed and recognized by a camera, using only the sound recorded by microphones.

An example of a use case of our technology is shown in **Fig. 1**. Suppose there are two people in a room and the objective is to automatically recognize them using a computer. The initial idea would be to



(a) Conventional image recognition to analyze bright room

(b) Conventional image recognition to analyze dark room

(c) Our cross-media scene analysis to analyze dark room

Fig. 1. (a, b) Conventional image recognition and (c) our cross-media scene analysis technology.

install a camera in the room and recognize them by using image recognition technology. If semantic segmentation technology is used, it is possible to detect not only the presence or absence of people, but also their locations and postures in silhouette form, as shown in Fig. 1(a). However, this approach would not work if the room was very dark or was a place where cameras are prohibited (Fig. 1(b)).

However, our technology can still be applied in such cases. Our technology uses audio information recorded by multiple microphones to deliver semantic segmentation results, instead of using a camera (Fig. 1(c)). If two people in the room are talking, their voices are recorded by the microphone arrays. With this sound information, our technology directly predicts the expected semantic segmentation result. To achieve this, we need to know (1) what kind of sound is coming from which direction, and (2) what kind of object (and its shape) is generating the sound. These details are respectively determined by signal processing and deep learning.

Imagine that a sound occurs at some location and is captured by multiple microphones. In this case, microphones closer to the source catch the sound earlier than the more distant ones. By analyzing this time difference of arrival, we can extract a directional feature that indicates the direction of the sound source. Furthermore, by analyzing the frequency information of the sound captured by each microphone, we can obtain a tonal feature that is useful to identify the type of the sound source (e.g., whether it sounds like a human voice or a train running on a track). With these features, we can determine the direction and the type of sound source. However, this information is not enough to recover a visual silhouette of the sound source representing the position and shape as compared to the semantic segmentation result.

Therefore, a deep neural network is used to estimate the type, shape, and position of the object. The overall setup is illustrated in **Fig. 2**. The neural network receives directional and tonal features as inputs

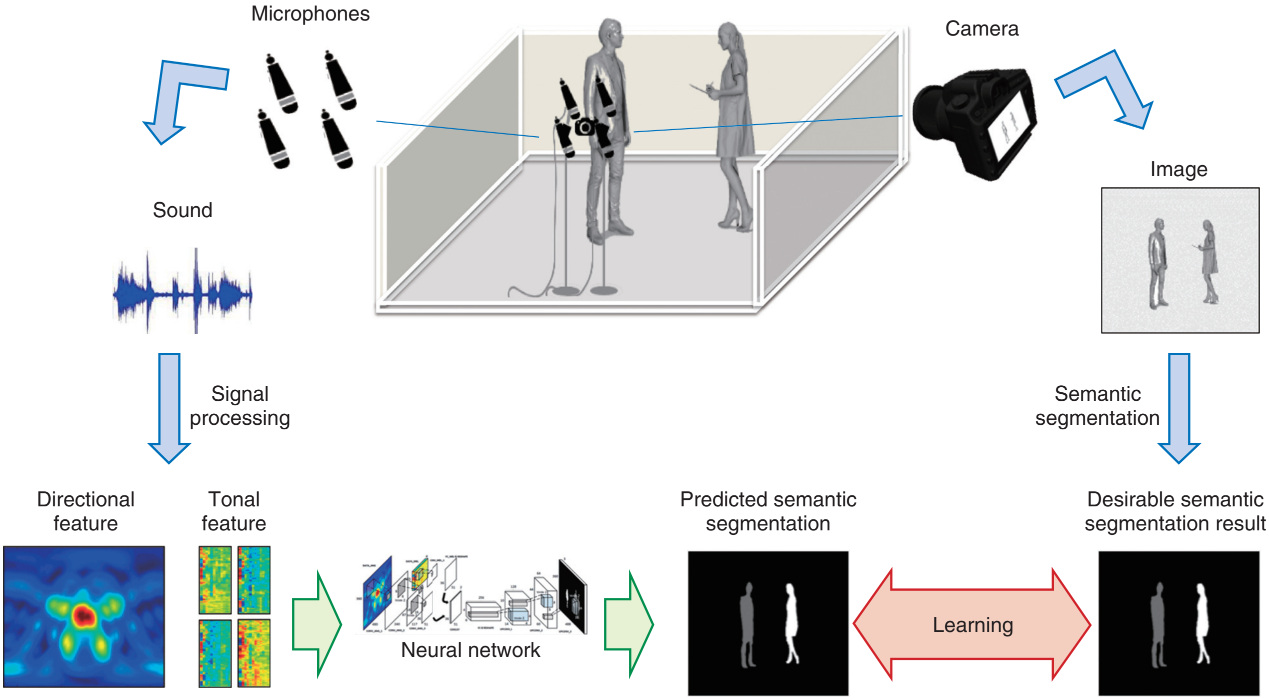


Fig. 2. Overview of cross-media scene analysis setup.

and is trained to output the semantic segmentation result directly from them. This is the core part of our technology that uses deep learning to convert media types from audio to visual. This is why we call our technology cross-media scene analysis. To train the network, we need to have a pair consisting of a desirable semantic segmentation result and corresponding audio features. Of course, a camera is needed to collect such data for training, but it is not necessary for actual recognition. With this process, we can build a basic mechanism to predict semantic segmentation results using only sound.

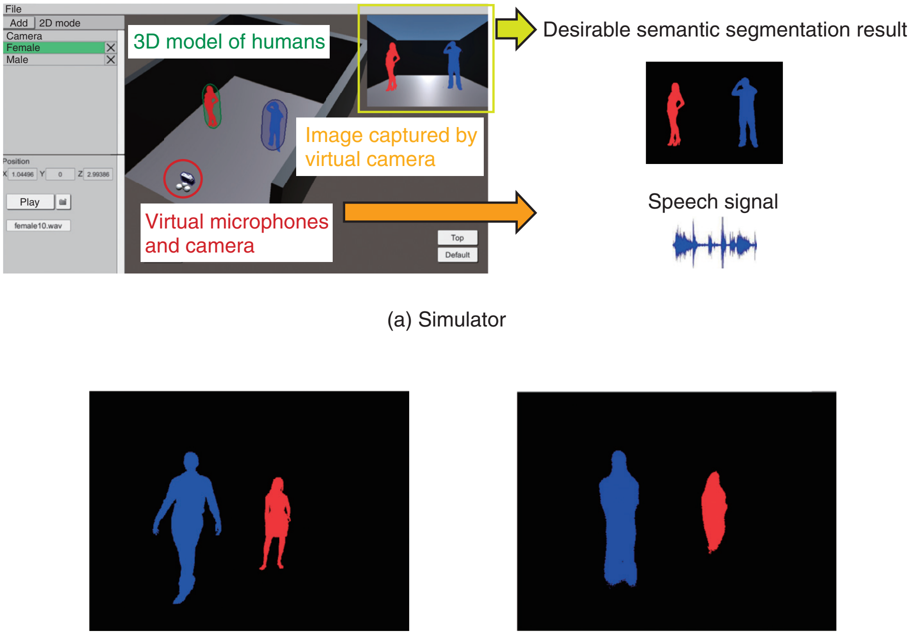
3. Proof of concept

We have conducted various experiments to evaluate how well our technology can predict actual semantic segmentation results from sound. We describe a few examples here.

First, we introduce an experiment using simulation data that was conducted to verify the feasibility of this technology in an ideal situation. We developed a simulator that can reconstruct a virtual room and generate conversational voices of people speaking and their corresponding semantic segmentation result simultaneously (**Fig. 3(a)**). We can freely change the size of the room, arrange people (three-dimensional models), the virtual microphones and camera, and simulate sounds recorded by the microphones taking the reverberation and echo of the room into account. At the same time, we can also simulate the image of the room taken by the camera and obtain the corresponding semantic segmentation result of the scene. Hence, we can train our neural network and measure how accurate the semantic segmentation result predicted by our technology is by comparing it to the desirable one.

We show an example of the desirable semantic segmentation result and the result predicted with our technology in **Fig. 3(b)** and **(c)**, respectively. Although the predicted result does not accurately recover details of posture and shape, we can see that the distance and rough shape can be successfully predicted by our technology.

Next, we describe an experiment done on real sounds. The task was to estimate the position and orientation of a toy train based on its running noise. Our evaluation setup is shown in **Fig. 4(a)**. It consisted of a toy train running on a circumferentially connected rail and four microphones connected to a personal computer in which our technology was installed. The entire setup was covered by a clear


File
Add
2D mode
Camera
Female
Male
Position
Play
female10.wav
Top
Default
3D model of humans
Image captured by virtual camera
Virtual microphones and camera
Desirable semantic segmentation result
Speech signal

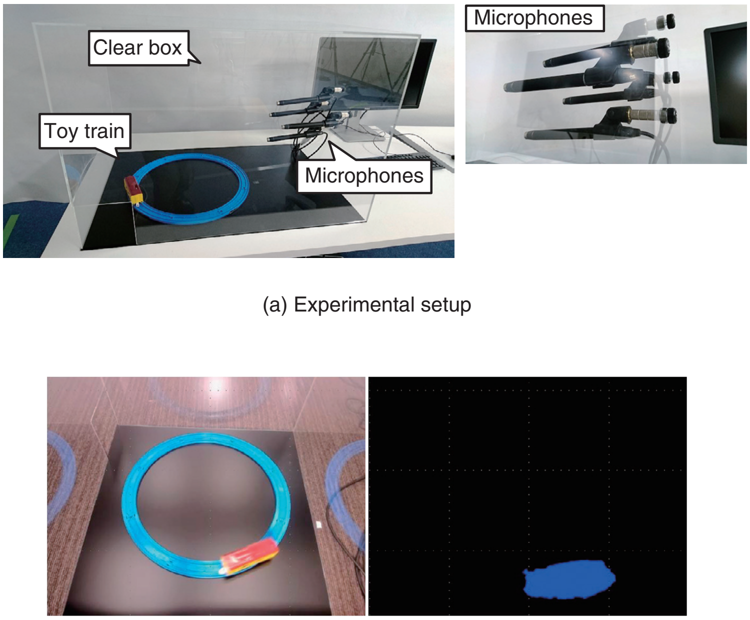
(a) Simulator

(b) Desirable semantic segmentation result

(c) Semantic segmentation result predicted by our technology

3D: three-dimensional

Fig. 3. Our simulator and predicted semantic segmentation result obtained by our technology.



(a) Experimental setup

(b) Actual position and orientation of the toy train

(c) Semantic segmentation result predicted by our technology

Fig. 4. Experimental setup and predicted results by our technology.

acrylic box. An example of the semantic segmentation result predicted by our technology is shown in **Fig. 4(c)**. The blue region represents the silhouette of the train. Compared to the position and orientation of the real train **(Fig. 4(b)**), we can see that the estimated position and orientation are roughly consistent with it, even though it is a little blurry. This shows that our technology can be applied to sounds other than human voices.

4. Future development

We will continue working to improve and demonstrate our technology toward application to a more natural space. To date, we have successfully verified the technology in structured environments such as those reconstructed by our simulator or the clear box. However, we need to make our method more robust in order to apply it in more realistic situations where more complex types of noise exist. The types of objects that can currently be recognized by our method are limited (people and toy trains), so we will work to expand those and test for more diverse categories. We will continue to improve our technology to achieve secure monitoring with arbitrary media information suitable for the situation.

Go Irie

Senior Research Engineer, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.E. and M.E. in systems engineering from Keio University, Kanagawa, in 2004 and 2006, and a Ph.D. in information science and technology from the University of Tokyo in 2011. He joined NTT in 2006 and has been studying multimedia content analysis, indexing, and retrieval. During 2011–2012, he was a visiting researcher at Columbia University, NY, USA. He is a member of the Institute of Electrical and Electronic Engineers (IEEE) and the Institute of Electronics, Information and Communication Engineers (IEICE).

Hirokazu Kameoka

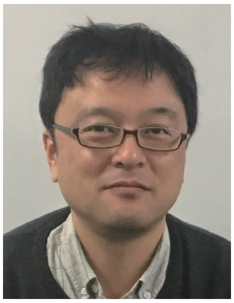
Senior Research Scientist, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.E., M.S., and Ph.D. from the University of Tokyo in 2002, 2004, and 2007. He is currently a Distinguished Researcher and a Senior Research Scientist with NTT Communication Science Laboratories, and an Adjunct Associate Professor with the National Institute of Informatics. From 2011 to 2016, he was an Adjunct Associate Professor with the University of Tokyo. His research interests include audio and speech processing and machine learning. He has been an Associate Editor for the IEEE/ACM Transactions on Audio, Speech, and Language Processing since 2015 and a member of the IEEE Audio and Acoustic Signal Processing Technical Committee since 2017.

Akisato Kimura

Senior Research Scientist, Research Planning Section, NTT Communication Science Laboratories.

He received a B.E., M.E., and D.E. in communications and integrated systems from Tokyo Institute of Technology in 1998, 2000, and 2007. He has been with NTT Communication Science Laboratories since 2007. He is engaged in work on pattern recognition, machine learning, and data mining. He is a board member of the Japanese Society for Artificial Intelligence (JSAI), a senior member of IEEE and IEICE, and a member of the Association for Computing Machinery (ACM) SIGMM (Special Interest Group on Multimedia) and SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining).



Kaoru Hiramatsu

Senior Manager, NTT Geospace Corporation.*

He received a B.S. in electrical engineering and an M.S. in computer science from Keio University, Kanagawa, in 1994 and 1996, and a Ph.D. in informatics from Kyoto University in 2002. He joined NTT Communication Science Laboratories in 1996, where he worked on the Semantic Web, sensor networks, and media search technology. From 2003 to 2004, he was a visiting research scientist at the Maryland Information and Network Dynamics Laboratory, University of Maryland, USA. He is a member of the Information Processing Society of Japan (IPSJ) and JSAI.

*He was a Senior Research Scientist, Supervisor, and Leader of Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories until June 2018.

Kunio Kashino

Senior Distinguished Researcher, Head of Media Information Laboratory, NTT Communication Science Laboratories.

He received a Ph.D. from the University of Tokyo for his pioneering work on music scene analysis in 1995. He is also an adjunct professor at the Graduate School of Information Science and Technology, the University of Tokyo, and a visiting professor at the National Institute of Informatics (NII). He has been working on audio and video analysis, recognition, and search algorithms. He is a senior member of IEEE and IEICE, and a member of ACM.