Media Scene Learning: A Novel Framework for Automatically Extracting Meaningful Parts from Audio and Video Signals

Akisato Kimura†*, Hirokazu Kameoka, and Kunio Kashino*

Abstract

We describe a novel framework called Media Scene Learning (MSL) for automatically extracting key components such as the sound of a single instrument from a given audio signal or a target object from a given video signal. In particular, we introduce two key methods: 1) the Composite Auto-Regressive System (CARS) for decomposing audio signals into several sound components on the basis of a generative model of sounds and 2) Saliency-Based Image Learning (SBIL) for extracting object-like regions from a given video signal on the basis of the characteristics of the human visual system.

1. Learning for understanding scenes from media

Humans easily and naturally analyze the surrounding audio and visual scenes acquired by their ears and eyes and understand what is happening. Imitating this mechanism and implementing it on computers has been one of the most important research issues for several decades. With recent progress in hardware and software, several specific technologies such as clean speech recognition and face detection are becoming increasingly used in practice. However, a computer's ability to recognize and understand audio and visual scenes is generally far worse than a human's ability, in spite of the long history and importance of this research. Meanwhile, we have to note that this ability is not inherited but learned not only for computers but also humans, except for some basic functions provided by sensory organs such as eyes and ears. In fact, previous psychological studies [1], [2] indicate that most functions for understanding scenes are acquired posteriori through the human development process. This finding implies that learning plays an important role in humans as well as in computers understanding scenes from media such as audio and video signals.

† NTT Communication Science Laboratories
Atsugi-shi, 243-0198 Japan

2. Morphemes and their extension to media morphemes

2.1 Morpheme: a unit for understanding the meaning of text

In contrast to the case of audio and video signals, a lot of advanced technologies for understanding text information have already been developed. Internet search engines are one of the most successful applications that utilize those technologies. We can instantly find desired web pages just by entering relevant text information.

A typical procedure for understanding text information is shown in **Fig. 1** on the left. One of the fundamental and significant technologies is morphological analysis, which decomposes a sentence into small

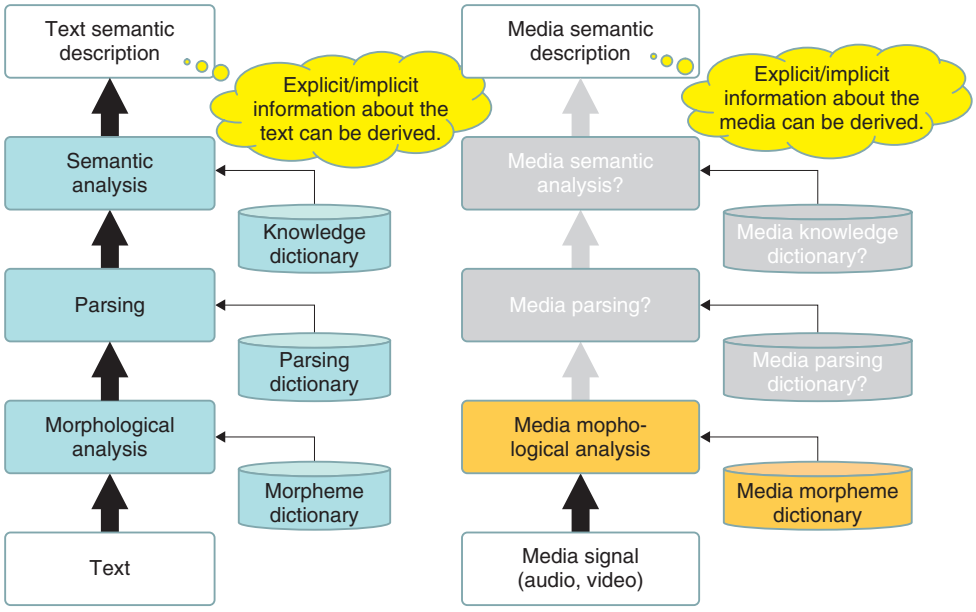

Fig. 1. Correspondence between text semantic analysis and media scene understanding.

parts called morphemes, such as verbs and nouns. Morphological analysis can conveniently provide indices to each web page registered in a database, where a morpheme is used as an index. These indices help us to obtain web pages relevant to a given set of keywords in almost the same way as finding a specific word in a dictionary.

Moreover, morphological analysis plays an important role in bridging the gap between a sequence of characters and its meaning. Each morpheme is associated with some meaning, and its usage is also strictly defined: these are collected in a morpheme dictionary. The information obtained from this dictionary lets us accomplish high-level text processing such as parsing and semantic analysis to accurately capture the meaning of a given sentence.

2.2 Media morpheme: a component for understanding the meaning of audio and video signals

If we could achieve a procedure similar to morphological analysis for audio and video signals, that is to say *media morphological analysis*, it would be a significant step toward understanding media scenes. However, the problem is how to construct a media morpheme dictionary describing correspondences between *media morphemes* and their meanings because the definition of media morphemes has not yet been established.

To this end, we are taking another approach to media morphological analysis: we are trying to discover candidates of media morphemes by utilizing its significant characteristics as cues, which might be useful for acquiring and learning media morphemes from media. We call this framework Media Scene Learning (MSL). We focus mainly on two fundamental properties, shown in **Fig. 2**, to discover media morpheme candidates.

1) Repetition: If several signal elements frequently appear together, the set of elements can be considered to be a media morpheme.

2) Saliency: If a signal element is more salient than neighboring ones, the element can be considered to be a media morpheme.

As possible solutions to Media Scene Learning based on the above properties, we introduce two methods below: the Composite Auto-Regressive System (CARS) [3] for audio signals and Saliency-Based Image Learning (SBIL) [4], [5] for video signals.

3. Methods for MSL

3.1 CARS for audio MSL

One possible way to achieve MSL for audio signals is CARS [3], which is overviewed in **Fig. 3**. This method focuses on the first property, repetition: it decomposes a given audio signal into several pairs of a pitch and a tone.

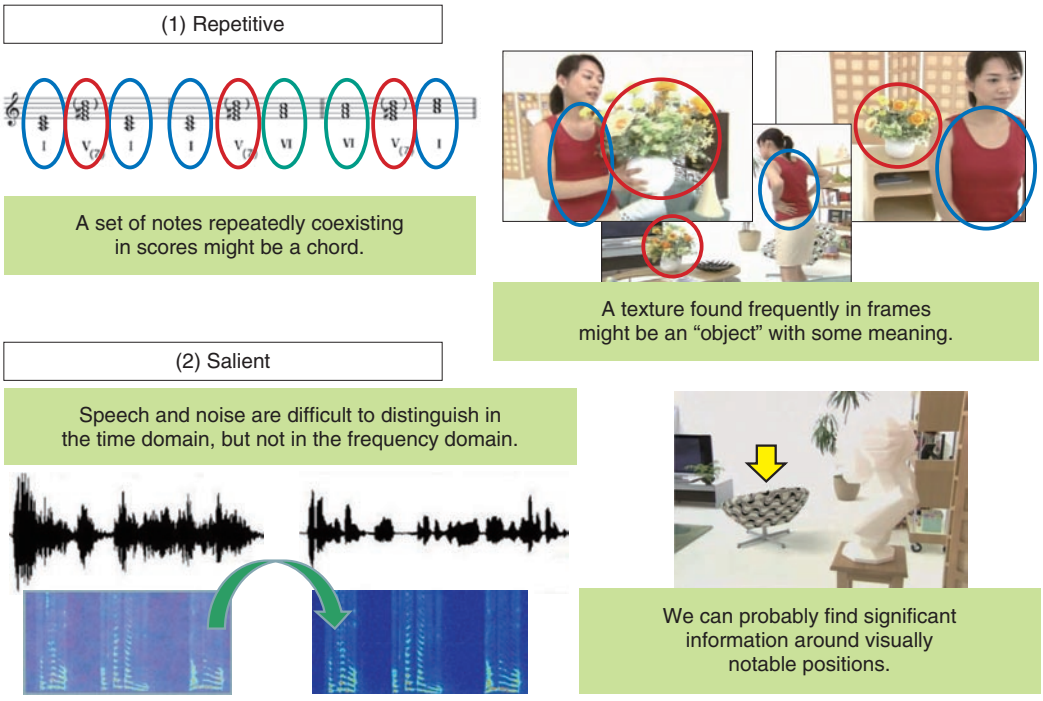

Fig. 2. Fundamental properties of media morphemes.

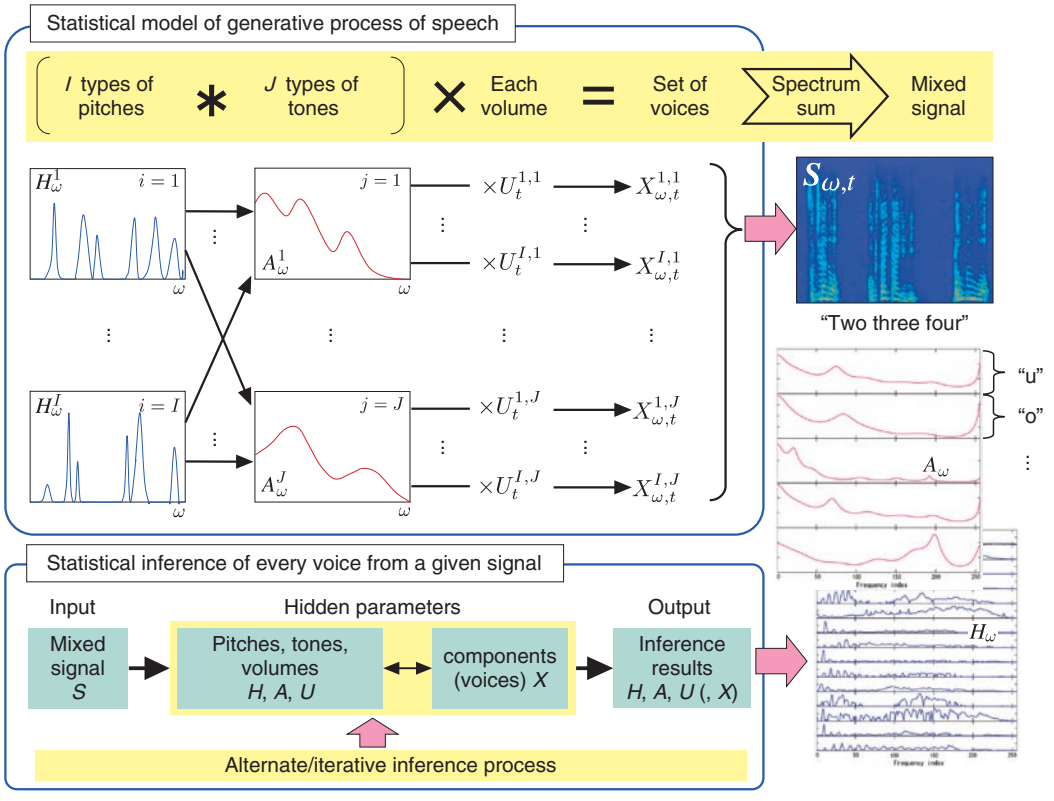

Fig. 3. Overview of CARS.

CARS represents an audio signal using a source filter model taking into consideration the process used to generate the audio signals. This model assumes that an audio signal is composed of a mixture of filtered sources, where every source and filter corresponds to a pitch and a tone, respectively, as shown in Fig. 3. The problem is to select as few sources and filters as possible to represent a given audio signal well enough. To do this, CARS tries to discover frequently and simultaneously appearing pitches and tones in a given audio signal by using the Expectation Maximization (EM) algorithm, a standard approach for iterative statistical inference. With the help of the EM algorithm, we mathematically formulated the problem with the following notation: sources $\boldsymbol{H}$, filters $\boldsymbol{A}$, volumes $\boldsymbol{U}$, component signals $\boldsymbol{X}$, and audio signal $\boldsymbol{S}$ to be analyzed. We derived a solution composed of the following alternating steps: 1) derive sources $\boldsymbol{H}$, filters $\boldsymbol{A}$, and volumes $\boldsymbol{U}$ for fixed component signals $\boldsymbol{X}$ and 2) derive component signals $\boldsymbol{X}$ for fixed sources $\boldsymbol{H}$, filters $\boldsymbol{A}$, and volumes $\boldsymbol{U}$. We also confirmed the basic operation of the above mathematical solution via several experiments using speech signals [3].

CARS can be applied to various kinds of situations in audio signal processing. One representative example is speech source separation, where an audio signal that includes a mixture of the voices of several people is decomposed into individual voices and they are restored. Our recent study [6] revealed the effectiveness of CARS for speech source separation.

3.2 SBIL for video MSL

One possible way to achieve MSL for video signals is SBIL [4], [5], which is overviewed in **Fig. 4**. This method extracts regions of interest as object candidates from a given video signal on the basis of visual saliency. The main features of SBIL can be summarized as follows:

1) Fully automatic extraction: A prior probabilistic density function (PDF) representing the possibility of object existence can be derived automatically via the algorithm for estimating saliency-based visual attention with our Bayesian model [8], [9]. Note that many existing methods, such as interactive graph cuts [7] for still-image segmentation, need some manually provided labels representing an object or a background. SBIL can convert such manual labels into saliency-based attention to achieve fully automatic segmentation. Feature likelihoods of objects representing the tendency of image features of objects can be derived by collecting image features of objects. Likewise, the feature likelihoods of backgrounds (i.e., non-object regions) representing the tendency of image features of backgrounds can also be derived by collecting image features of backgrounds. The prior PDF and feature likelihoods form a statistical model, and the prior PDF can be inferred to derive the segmentation result, in the same way as in interactive graph cuts.

2) Sequential region update: The segmentation result obtained from the previous frame includes significant information for identifying and localizing objects in the current frame since the location and image features do not change much within a short period of time. SBIL fully utilizes the above characteristics. In particular, the location of the previously extracted region is utilized and combined with the current result of visual attention estimation to derive the current prior PDF, and the distribution of image features in the previously extracted region is also combined with the feature likelihood obtained from the first characteristics to derive the current feature likelihoods. As a result, stable segmentation results can be obtained.

Many processes in SBIL can be converted to suit parallel processing, so SBIL can be accelerated if multi-core processors are available. We implemented SBIL on a consumer personal computer having a standard graphics processor unit (GPU) and achieved a speed nearly as fast as realtime processing [7], [10].

One promising application of SBIL is generic object localization and recognition. We developed a prototype system for generic object localization and recognition, as shown in **Fig. 5**. It first captures images from a webcam and extracts object-like regions by means of SBIL. It then retrieves its own database to find and provide registered information about the extracted regions. If there is no relevant information in the database, the system can ask users to provide some related information about the extracted regions. Note that SBIL's region extraction function plays an important role in identifying the target in question for the system. In the near future, we expect to further enhance this system by integrating the latest research on automatic image annotation and retrieval [11].

4. Prospects for MSL

In this article, we introduced a novel framework called MSL and two key methods, CARS and SBIL, for discovering media morpheme candidates without

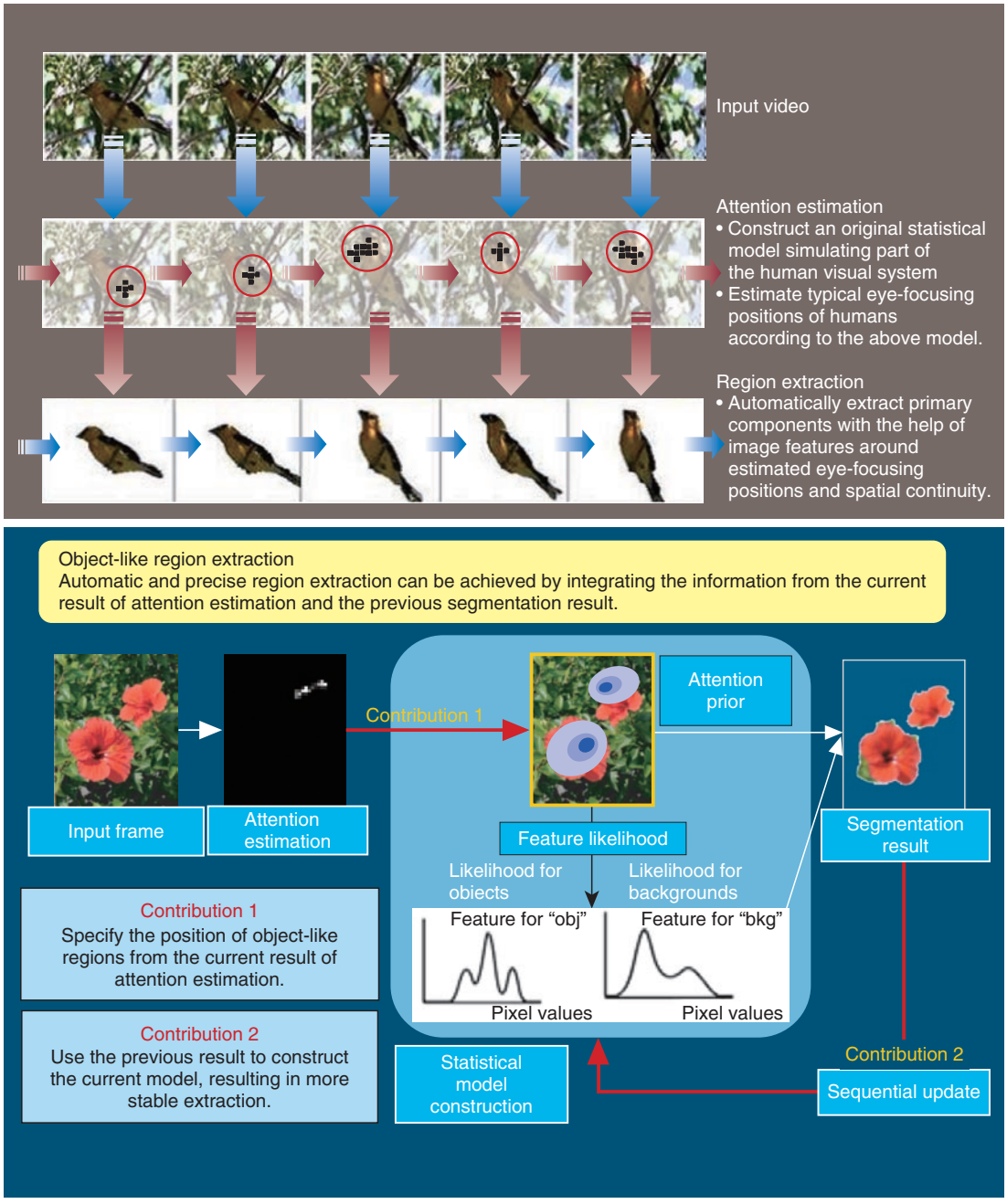

Fig. 4. Overview of SBIL.

any human supervision and prior knowledge. Further development of MSL might enable the creation of a media morphological dictionary in which every morpheme is connected to its meaning. Beyond that, we see prospects for media parsing to analyze the structure of media signals and media semantic analysis as the ultimate goal of MSL.

Computers have different functions from humans: some functions are possessed only by humans while others are possesses only by computers. This implies that when we eventually create computers that understand media scenes, they might use different techniques from humans. We will continue working on this research so that we get to see the birth of MSL computers in our lifetimes.

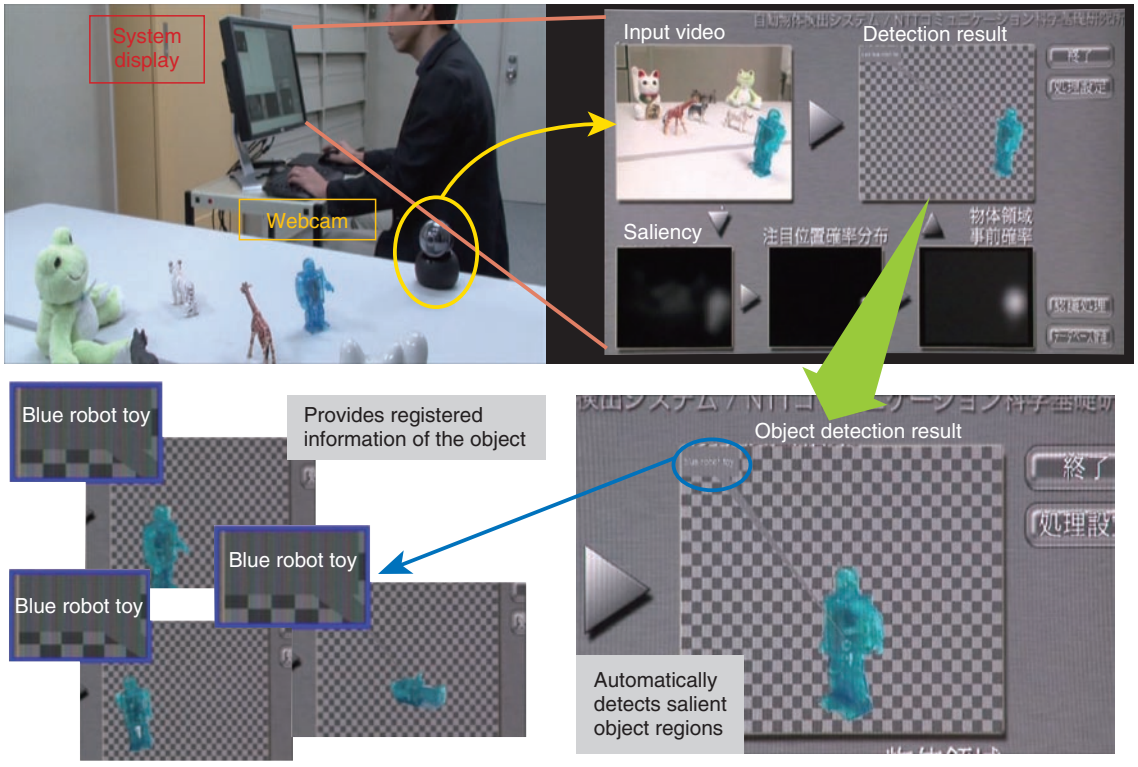

Fig. 5. Prototype system for generic object localization and recognition by SBIL.

References

[1] J. Piaget, "The Equilibration of Cognitive Structures: The Central Problem of Intellectual Development," University of Chicago Press, 1985.

[2] A. Kimura, K. Kashino, K. Fukuchi, K. Akamine, and S. Takagi, "Cognitive Developmental Approach to the Realization of Sophisticated Visual Scene Understanding," IEICE Technical Report, PRMU2009-144, December 2009 (in Japanese).

[3] H. Kameoka and K. Kashino, "Composite Autoregressive System for Sparse Source-filter Representation of Speech," Proc. of International Symposium on Circuits, Automation and Signal Processing (ISCAS), pp. 2477–2480, Taipei, Taiwan, 2009.

[4] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based Video Segmentation with Graph Cuts and Sequentially Updated Priors," Proc. of International Conference on Multimedia and Expo (ICME), pp. 638–641, New York, USA, 2009.

[5] K. Fukuchi, K. Miyazato, K. Akamine, A. Kimura, S. Takagi, J. Yamato, and K. Kashino, "Saliency-based Video Segmentation with Graph Cuts and Sequentially Updated Priors," IEICE Transactions on Information and Systems, Vol. J93-D, No. 8, pp. 1523–1532, August 2010 (in Japanese).

[6] H. Kameoka, T. Yoshioka, M. Hamamura, J. Le Roux, and K. Kashino, "Statistical Model of Speech Signals based on Composite Autoregressive System with Application to Blind Source Separation," Proc. of International Congress on Latent Variable Analysis and Speech Separation (LVA/ICA 2010), Lecture Notes in Computer Science 6365, pp. 245–253, 2010.

[7] Y. Boykov and G. Lea, "Graph Cuts and Efficient n-d Image Segmentation," International Journal of Computer Vision, Vol. 70, No. 2, pp. 109–131, 2006.

[8] D. Pang, A. Kimura, T. Takeuchi, J. Yamato, and K. Kashino, "A Stochastic Model of Selective Visual Attention with a Dynamic Bayesian Network," Proc. of International Conference on Multimedia and Expo (ICME), pp. 1073–1076, Hannover, Germany, 2008.

[9] K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Real-time Estimation of Human Visual Attention with Dynamic Bayesian Network and MCMC-based Particle Filter," Proc. of International Conference on Multimedia and Expo (ICME), pp. 250–257, New York, USA, 2009.

[10] K. Akamine, K. Fukuchi, A. Kimura, and S. Takagi, "Fully Automatic Extraction of Salient Regions in Near Real-time," accepted by the Computer Journal, available on arXiv, a preprint server. http://arxiv.org/abs/1004.0085

[11] A. Kimura, H. Kameoka, M. Sugiyama, T. Nakano, E. Maeda, H. Sakano, and K. Ishiguro, "SemiCCA: Semi-supervised Learning of Canonical Correlations," Proc. of International Conference on Pattern Recognition (ICPR), pp. 2933–2936, Istanbul, Turkey, 2010.

Akisato Kimura
Research Scientist, Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E. and M.E. degrees in electrical and electronics engineering from Tokyo Institute of Technology in 1998 and 2000, respectively. He received the Ph.D. degree in communications and integrated systems from Tokyo Institute of Technology in 2009. He joined NTT Communication Science Laboratories in 2000. Since then, he has consistently developed core technologies for multimedia information retrieval and multimedia signal processing, most notably, time-series signal search (2000–2005), template matching (2003–2006), and visual attention estimation (2006–2010). His current research is mostly focused on next-generation visual scene understanding and the development of its core technologies using knowledge and techniques about pattern recognition, computer vision, image processing, machine learning, vision science, developmental psychology, robotics, and human-computer interaction. He received the Best Interactive Session Award at the Meeting on Image Recognition and Understanding (MIRU) in 2008 and 2010. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan and the Information Processing Society of Japan (IPSJ) and a senior member of IEEE.

Hirokazu Kameoka
Researcher, Media Recognition Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E., M.E., and Ph.D. degrees all from the University of Tokyo in 2002, 2004, and 2007, respectively. He joined NTT Communication Science Laboratories in 2007. His research interests include computational auditory scene analysis, acoustic signal processing, speech analysis, and music applications. He is a member of IEICE, IPSJ, and the Acoustical Society of Japan (ASJ). He received the Yamashita Memorial Research Award from IPSJ, the 20th Telecom System Technology Student Award from the Telecommunications Advancement Foundation in 2005, the Itakura Prize Innovative Young Researcher Award from ASJ, the 2007 Dean's Award for Outstanding Student in the Graduate School of Information Science and Technology from the University of Tokyo, the 1st IEEE Signal Processing Society Japan Chapter Student Paper Award in 2007, the Awaya Prize Young Researcher Award from ASJ in 2008, and the IEEE Signal Processing Society 2008 SPS Young Author Best Paper Award in 2009.

Kunio Kashino
Senior Research Scientist, Supervisor, Distinguished Researcher, and Leader of Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the Ph.D. degree from the University of Tokyo for his pioneering work on music scene analysis in 1995. He joined NTT in 1995. He has been working on audio and video analysis, search, retrieval, and recognition algorithms and their implementation. He has received several awards including the Maejima Award in 2010, the Young Scientists' Prize for Commendation for Science and Technology from the Minister of Education, Culture, Sports, Science and Technology in 2007, and the IEEE Transactions on Multimedia Paper Award in 2004. He is a senior member of IEEE. He is also a Visiting Professor at the National Institute of Informatics, Tokyo.