SightX: Obtaining Information on a Scene by Pointing a Camera

Daichi Namikawa, Hiroya Minami, Haruno Kataoka, Motohiro Makiguchi, and Michio Shimomura

Abstract

This article describes the SightX communication service, which gives the user access to information without requiring the use of language. SightX allows users to obtain information about objects that they see in front of them when they do not know the name of the object or when they feel uncomfortable speaking into a device for information retrieval. With SightX, the user simply points the camera of a smart device at an object to retrieve information about it.

1. Introduction

The widespread use of smart devices and the development of mobile networks in recent years have created an environment that allows users to access information from anywhere at anytime. In the past, search services based on typed-in text keywords have been used widely, but recently, services that provide information after accepting voice input of questions have been introduced [1]. Nevertheless, when suitable keywords do not come to mind, or in situations where input by typing or speaking is inconvenient, it is not possible to make full use of the smart device and mobile network. To solve this problem, we are implementing a communication service called SightX that enables the user to access required information on a scene without having to use language input. The goal is to achieve our service vision of providing *services for media editing and conversion during real-time communication* (*RTC*).

2. Service overview

2.1 Concept

SightX enables users to obtain information about objects that they encounter by simply pointing the camera of a smart device* at the object when they do not know what it is called or when they feel uncomfortable using a speech interface. The user can thus retrieve information at any time without using language input. To ensure that this service can be used by various smart devices, old or new, complex processing is performed on the cloud by a high-performance server. This service incorporates the concepts of sight and exploration, as well as expansion of recognition and the field of view, as reflected in the name SightX.

2.2 Use scenarios

Here we introduce a few SightX use scenarios in detail, describe the information provided to the user, and present some specific examples. We begin with cases in which the name of an object is not known.

(1) Commercial products in department stores and specialty stores

This service can provide the user with basic information such as the name and price of products that have a form that is unfamiliar to the user and whose name or use is not known, such as kitchen appliances that have an elaborate design. The user need only point the camera at the product to obtain the information (**Fig. 1(a)**). For luxury goods such as wine and Japanese liquors, knowing the name and price is not enough to understand the product, so further details such as flavor and customer reviews are also provided (**Fig. 1(b)**).

* Smart device: A multifunction communication device such as a smartphone or tablet terminal

Fig. 1. Examples of using SightX.

(2) Persons whose names are not known or cannot be recalled

For famous people that appear on TV and in magazines and whose names are not known, the service provides a profile that includes the person's name, age, and affiliation when the camera is pointed at their face (**Fig. 1(c)**).

SightX can also be used in crowds, in vehicles, in public areas, or in other situations where it is inconvenient to use a voice interface to ask a question.

(3) Signs at tourist sites or shops

When the camera is pointed at a sign in a tourist site or in a shop, the gist of the information and the current location can be presented, along with relevant websites and other such information (**Fig. 1(d)**). For shop signs, information such as the type of shop and its reputation among customers, obtained through customer reviews, is provided.

Typically, restaurant review websites include the rating (score) of the restaurant, and a collection of customer opinions about the quality of the food, the service, the atmosphere, etc.

SightX thus provides a wide range of knowledge based on the object of interest and the user's situation.

3. Function configuration

The SightX functions are configured as described below (**Fig. 2**).

(1) Real-time video transmission function

Video taken by the user with a smart device is sent to a server on the cloud in real time.

(2) Video recognition function

The server on the cloud runs multiple video recognition engines that have different characteristics in parallel in order to recognize various kinds of objects and people in real-time video sent from a smart device and to respond to the user query, "What is this?"

(3) Knowledge processing function

On the basis of the results obtained by the video recognition engine, information about the recognized object is generated to present to the user. Here, too, multiple types of knowledge processing engines run in parallel so that a wide range of information can be

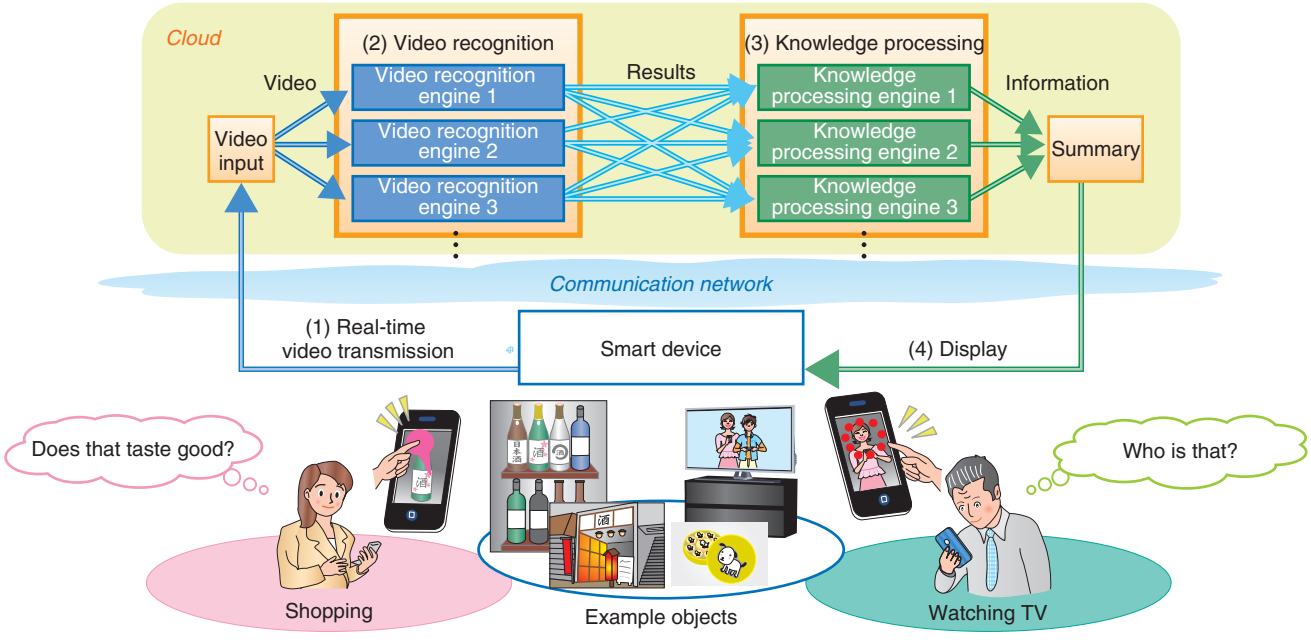


Fig. 2. Function configuration of SightX.

generated.

(4) Information display function

The information generated by the knowledge processing engine is received by the smart device and displayed on its screen.

In this way, the only functions the user's device require are for acquiring video in real time, sending and receiving data, and displaying the information on the screen. These functions are provided by older devices as well as the latest devices. Furthermore, the video recognition function and the knowledge processing function are configured from several kinds of video recognition and knowledge processing engines arranged in multiple layers on the cloud. That makes it possible to retrieve information on diverse recognition targets when the user simply points the device camera, without distinguishing between the services involved. Individual engines can also be added as the recognition targets and generated information are added or changed, and the targets are not limited to those mentioned in the use scenarios described in the previous section. For example, if there is a future need to recognize people's emotions, a facial expression recognition engine can be added.

4. Prototype implementation

We have put together relevant technology from the NTT research laboratories and other sources to implement a SightX prototype that allows actual testing of the use scenarios presented as examples in the previous section. Below, we describe the operation of the user interface and the video and knowledge processing engines that are used in the current prototype.

4.1 User interface

When the user points the camera at something they want to know about, information (e.g., a name) is superimposed on the device screen in the form of augmented reality (AR) icons. The user can tap on an AR icon to switch the information, such as from a name to a price. Pressing on an icon for a few seconds will open a tab window that presents more information (**Fig. 3**).

4.2 Video recognition engine

This capability uses an object recognition engine that recognizes and matches objects in two dimensions, a face recognition engine that determines who a person is from an image of their face, and English and Japanese character recognition engines. The character recognition engine in particular can identify

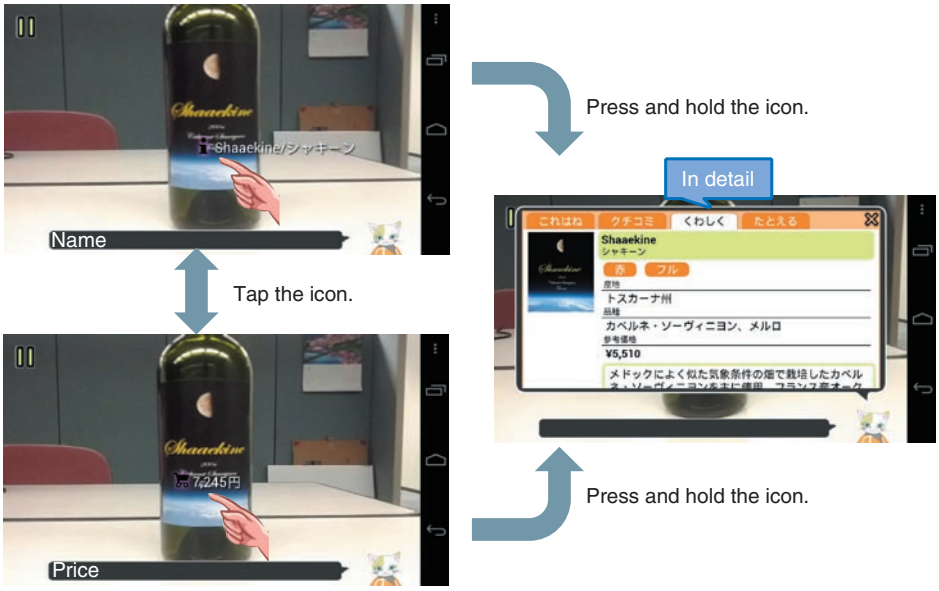


Fig. 3. Operation of user interface.

text sections within documents by taking the text column structure into account [3].

4.3 Knowledge processing engine

In addition to a search engine for gathering basic information such as the names of products and shops, there is a sensibility communication engine, which aids understanding by using life-log information such as purchase information and whether a product, shop atmosphere, or other thing seen for the first time corresponds to something in the user's past experience [4], [5]. There is also a word-of-mouth summary engine, which can analyze different kinds of evaluative information (known as word-of-mouth information) on a product or shop to create a short summary that can be displayed on the small screen of a smart device [6].

5. Conclusion and future development

We will continue to work towards implementing SightX as a communication service that enables users to access information without using language. We have so far implemented a prototype that allows users to get a feel for using SightX through actual experience and are promoting the visibility of the SightX service image.

In the future, we will cooperate with businesses in conducting field trials that give customers opportunities to actually use the service, and, in parallel with those trials, we will continue working on improving the elemental technologies of video recognition and knowledge processing and the application to wearable devices and other new devices.

For the trials, we first intend to increase the accuracy of the video recognition and knowledge processing functions and to stabilize the operation of the prototype system. The opinions from users and the overall knowledge obtained from the trials will serve as a basis for developing new use scenarios as well as providing feedback on the research and development (R&D) of the elemental technologies. Furthermore, the technical issues that are revealed will be used in discussing the direction of future R&D of the elemental technologies.

An example of a new elemental technology and device planned for incorporation is the application of three-dimensional (3D) object recognition [7], [8] in the video recognition engine. This technology enables recognition of 3D objects from a small number of training images, so it can be applied for uses that require recognition of various 3D objects. To make it possible to infer the user's circumstances or interests and select information that they truly want, we are also investigating the application of this technology to glasses-type camera devices that allow video to be captured from the user's actual viewpoint and field of view.

SightX features a high degree of generality due to the ability to flexibly change processing engines according to the addition of recognition targets or the information to be generated, or according to changes in either. There are probably use scenarios for solving various problems and tasks at places near the user that we cannot foresee. We would therefore like to implement communication services that have value in raising the visibility of the service vision by flexibly incorporating the opinions and requests of customers and businesses in expanding the range of application services.

References

[1] T. Yoshimura, "Shabette-Concier Service Realized by Natural Language Processing," Research Report of the Information Processing Society of Japan, 2012-SLP-93, pp. 1–6, 2012 (in Japanese).

[2] N. Uchida, "Service Visualization to Achieve Faster Service Creation," NTT Technical Review, Vol. 11, No. 9, 2013. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201309fa1.html

[3] A. Miyata and K. Fujimura, "Document Area Identification for Extending Books without Markers," Proc. of the ACM Conference on Human Factors in Computing Systems (CHI 2011), pp. 3189–3198, Vancouver, Canada.

[4] R. Mochizuki, S. Eitoku, M. Motegi, T. Yagi, S. Muto, and T. Kobayashi, "Emotion Communication Model Using Life-logs: Effectiveness of Using Life-logs for Easily Understood Representation," Trans. of the Information Processing Society of Japan, Vol. 53, No. 1, pp. 30–38, 2012 (in Japanese).

[5] R. Mochizuki, T. Watanabe, and T. Kobayashi, "Emotion Communication Model Based on Life-Log Comparison: Mutual Understanding through Comparable Experiences," Proc. of the 12th International Symposium on Applications and the Internet (SAINT), 2012 IEEE/IPSJ, Izmir, Turkey, 2012.

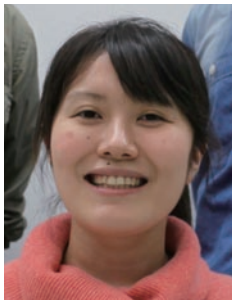
[6] H. Nishikawa, T. Hasegawa, Y. Matsuo, and G. Kikui, "Evaluative Text Summarization Model with Sentence Extraction and Ordering," Transactions of the Japanese Society for Artificial Intelligence, Vol. 28, No. 1, pp. 88–99, 2013 (in Japanese).

[7] H. Yabushita, J. Shimamura, and M. Morimoto, "A Framework of Three-Dimensional Object Recognition Which Needs Only a Few Reference Images," Proc. of the 21st International Conference on Pattern Recognition (ICPR 2012), TuPSBT2.16, pp. 1375–1378, Tsukuba, Japan.

[8] H. Yabushita, J. Shimamura, and M. Morimoto, "Three-dimensional object recognition based on spherical projection of obtained shapes and textures," Technical report of IEICE, PRMU, Vol. 111, No. 379, pp. 73–78, 2012 (in Japanese).

Daichi Namikawa
Researcher, Innovative Service Architecture Project, NTT Service Evolution Laboratories.
He received the B.E. and M.E. degrees from Kagoshima University in 2007 and 2009, respectively. He joined NTT in 2009 and is currently engaged in R&D of enhanced network services and systems. He is a member of the Information Processing Society of Japan (IPSJ).

Hiroya Minami
Manager, R&D Management, Planning Department, NTT Service Innovation Laboratory Group.
He received the B.E. and M.E. degrees from Seikei University, Tokyo, in 1994 and 1996, respectively. He joined NTT in 1996 and has been engaged in research on traffic technologies of communications systems, ubiquitous systems, and service delivery platforms. He received the Young Engineer's Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2003. He is a member of IEICE.



Haruno Kataoka
Researcher, Innovative Service Architecture Project, NTT Service Evolution Laboratories.
She received the B.E. and M.E. degrees from the University of Electro-Communications (UEC), Tokyo, in 2006 and 2008, respectively. She is currently a Ph.D. student at UEC. She joined NTT in 2008 and has been engaged in R&D of enhanced network services and systems. She received the IEICE Information Network Research Award in 2010, the Best Paper Award from the Japan Society of Security Management in 2010, and the Best Paper Award from the International Conference on Intelligence in Next Generation Networks in 2012. She is a member of IPSJ and IEICE.

Motohiro Makiguchi
Innovative Service Architecture Project, NTT Service Evolution Laboratories.
He received the B.E. and M.E. degrees from Hokkaido University in 2010 and 2012, respectively. Since joining NTT in 2012, he has been engaged in R&D of enhanced network services and systems. He is a member of IPSJ.

Michio Shimomura
Senior Research Engineer, Group Leader of Innovative Service Solution Group, Innovative Service Architecture Project, NTT Service Evolution Laboratories.
He received the B.S., M.S., and Ph.D. degrees in electronics and communication engineering from Waseda University, Tokyo, in 1988, 1990, and 1993, respectively. Since joining NTT in 1993, he has been researching network service systems. He is a member of IEICE.