ISSN 1883-1974 National Institute of Informatics News

No.43 Oct. 2012

This English language edition NII Today corresponds to No. 57 of the Japanese edition

NII Today

National Institute of Informatics News

FEATURED TOPIC | **Multimedia Sensing**

Pioneering Future Communication with

In the last ten years or so, the field of multimedia sensors such as cameras and microphones has seen both technological innovations and decreased costs. With these sensors becoming more ubiquitous throughout society, a legitimate question is what kind of new values do they bring us. I spoke with Associate Professor Gene Cheung and Chief Editor (Professor by Special Appointment) Yoh'ichi Tohkura, about the latest developments in this issue and what the future holds for these technologies and where they will be applied.

Collecting Large Amounts of Multimedia Data and Processing in Alignment with the Characteristics of Our Five Senses

Yukiko Motomura: First of all, what is Multimedia Sensing?

Gene Cheung: Well, all data obtained by sensors consist of signals, whether they are photographs taken by cameras or sound clips recorded by microphones. Our view is that Multimedia Sensing is what creates new values by processing these kinds of signals. Signals are the raw state; once they are processed, they become information.

Yukiko Motomura: The word for "information" in modern Japanese (*jōhō*) is written using the characters *jō*, which generally translates to "feeling," and *hō*, meaning "report." In the past, however, *jō* was written another character with the same pronunciation which generally means "the state of something." In line with this, the word "information," as it used to be written, referred purely to the concept of making known the state of something. It is said that the *jō* meaning "feeling" did not come to be used until people began to add their own analysis of the information. So, in other words, looking at the meaning of "information" as we know it today, you can collect data which expresses the state of something, but no matter how much you have, it has no meaning if it is not first processed effectively.

Yoh'ichi Tohkura
Professor by Special Appointment, NII
NII Today Chief Editor

Gene Cheung: That's exactly right. We now have what's called "Big Data," which is made up of data from multimedia sensors. If this data is not processed, however, it is nothing but noise. The issue for us lies in processing large amounts of data.

Yukiko Motomura: We have been hearing the term "Big Data" a lot recently. People say that you can create new values by finding breakthroughs in processing huge quantities of data which were previously un-processable. But where exactly are these breakthroughs in processing large amounts of data with Multimedia Sensing?

Gene Cheung: The input sources of multimedia data are things like cameras and microphones. As this kind of data is consumed by human's five senses, it can also be processed in alignment with the characteristics of each of those senses. When looking at a single image, for example, a person's line of sight might travel all over the place. When a viewer focuses in on one spatial area of an image, he can perceive great details, but other spatial areas away from his gaze point can become blurry and yet the viewer would not be able to notice.

Yukiko Motomura: In other words, what we need to do is carefully process only the portions which catch the eye?

Gene Cheung: That's right. Our eyes are attracted to things that flash or move. If we focus processing in alignment with the characteristics of the five senses, it is possible to process even very large amounts of data effectively. (Refer to pages 8-9.)

Yukiko Motomura: Associate Professor Cheung, may I ask which research theme has most caught your interest recently?

Gene Cheung: Well, my speciality is the compression and transmission of visual data. Although we can now watch video even on the Internet, we can only experience it in 2-D format at this point in time. Viewers cannot be immersed in the visual experience, nor can they feel any physicality of the objects in the video scene. I am currently researching visual data compression technologies for the purpose of recreating such immersive viewing experiences in 3-D. (Refer to pages 10-11.)

Gene Cheung
Associate Professor, Digital Content and Media Sciences Research Division, NII

The Issue of How to Process Large Quantities of Collected Data

Yukiko Motomura: Are there any other methods for processing data effectively?

Gene Cheung: Well, when you process large amounts of data, speed is important. Both the speed and efficiency of processing are enhanced by algorithms that pinpoint, out of a given set of data, only the instances of input that we as humans perceive as valuable and subsequently eliminate unnecessary operations/calculations. These days, there are surveillance cameras everywhere. While very cheap cameras can record images, the resolution is not very high, but afterwards, images can be processed to increase resolution. This can lead to new applications.

Yukiko Motomura: So processing the material

Multimedia Sensing Technologies

Yukiko Motomura

Deputy General Manager, Science and Environment News Department, The Mainichi Newspapers Co., Ltd.

A graduate of the Educational Studies Department at Kyushu University, Yukiko Motomura joined The Mainichi Newspapers in 1989. In 2001, she moved to the Science and Environment News Department. She began a series of articles that investigate the state of science technology/expertise in Japan, while painting a picture of the present state of scientists in a compilation called "Rikei Hakusho," or "The Scientific Whitepaper." In 2006, she received the first "Science Journalist of the Year" award from the Japanese Association of Science & Technology Journalists (JASTJ). She has also authored books such as "Rikei Shiko," or "Scientific Thinking."

effectively can make video images even clearer and sharper?

Gene Cheung: Yes. When you exploit the temporal correlation of successive images, the image resolution can be increased. Inter-view correlation in images taken from many angles can also be exploited, so that novel images from virtual views can be synthesized. The issue is how to create a complete image using a diverse collection of data.

Yukiko Motomura: You said a moment ago that it is important to pinpoint the five senses. Was it not possible to perform well-modulated processing with past technologies?

Gene Cheung: Well, take computers for example. They can only register images as simple arrays of numbers. People are different; they immediately see and understand things with ease. For instance, we can be certain that over there we have a person, a desk and whatever else.

Yoh'ichi Tohkura: When people look at an image, they know exactly what the picture shows. Machines, however, do not have this level of comprehension. In other words, it is very difficult for computers to assign a meaning to the objects shown.

Familiar Applications of Multimedia Sensing

Yukiko Motomura: How will our society change as a result of applying the technologies you have described today?

Gene Cheung: "Gait recognition" is an example that is easy to understand. For example, behavioral patterns and expressions can be automatically recognized when recorded on surveillance cameras. Thus, technology can report unfamiliar faces or activity, alerting us to suspicious activity.

Yoh'ichi Tohkura: Take for instance the case of the Aum Shinrikyo members who were recently taken into custody. Even without the face of the suspect appearing on surveillance cameras, the characteristics of how the individuals walked were recognized and matched with 95% certainty to those of the suspect. Thus, if we record patterns related to things such as how members of a certain organization walk or act, then we can distinguish between them and individuals who do not have the same characteristics, and the need to confirm things like personal identification disappears. This technology is already at the stage in which it can be put into practical use.

Yukiko Motomura: Okay. Lastly, Professor Tohkura, what do you see on the horizon with respect to the kind of future and possibilities that Multimedia Sensing opens up for us?

Yoh'ichi Tohkura: We have reached the point where the development of sensors has been advanced considerably, making all kinds of information attainable. Sensors are not just devices like microphones and cameras; devices like cellular phones and car navigation systems are also types of sensors. Recent cellular phones and smartphones not only allow us to attain sound, still images and video images; they also collect things such as positioning information through the use of GPS and real-time transmission of text-based information via mediums like Twitter. This means that phones are multimedia sensors capable of so many things, and that most people are walking around with one of these sensors. Also, in terms of car navigation systems, it's almost as if people have been replaced by cars in a sense, as cars also command many of the same features of other sensors. Right after the Great East Japan Earthquake, we could pinpoint the location of cars through transmissions from the navigation systems, allowing us to understand the status of road recovery in the region. This can be considered a good example of how car navigation sensor networks have been used effectively.

Yukiko Motomura: What would you say are the technological issues, if there are any?

Yoh'ichi Tohkura: Well, the first issue would be how to take diverse multimedia data and analyze it in a horizontal and time-oriented manner so that it can be both turned into value and expressed. To put it in other words, the issue is developing methods that appeal to the five senses using ways in which the value is most easily understood; whether the medium is images, written words, sounds or any other. As I have expressed here, the concepts of crime-prevention and disaster-reduction provide greater values of safety and security to society.

The second issue would be dealing with "Big Data" in Multimedia Sensing. In terms of units of information, this kind of data is said to be in the petabyte range (quadrillion ranges). There are still obstacles to overcome with respect to the amount of data in terms of both the number of calculations and in terms of algorithms.

The third issue would be figuring out by whom and for what purpose data obtained from sensors will be used. This is a big problem related to data security and the rights of both the transmitting parties and the parties obtaining the data. This is an issue in which society as a whole must reach a consensus. I think that if we overcome this obstacle, we can use it as a starting point for general research to begin expanding new fields of application for Multimedia Sensing.

****

A Word from the Interviewer

Multimedia Sensing is still a concept that is unfamiliar to the general public. However, I now have the impression and understanding that it involves the process of combining various pieces of technology. This includes the performance of each terminal and the networks that connect them together, as well as methods of application. It is also necessary to give consideration to privacy protection. I would personally like this to be a general technology that forms around linking people in various fields together. It was only fifty years ago that we were overjoyed just to talk with someone far away on the telephone. That has since transformed into video, and in the years to come, it could mean life-like 3-D communication. I felt as if I was being shown a piece of a bright future.

NII Special 1

Sensing Sound: The New Sound Signal

As the use of surveillance cameras throughout cities becomes more popular, they are being used more and more in many fields of application, such as tracking persons of interest. In this sense, visual images have become the current agent of sensing technologies. On the other hand, sound information has also played an important role in sensing technology in much the same way as visual information. We had the chance to hear about the newest in sound sensing technologies from Associate Professor Nobutaka Ono. He researches Microphone Array technologies, in which it is possible to apply such technologies in preparing minutes from meetings and in securities systems by using multiple microphones to pick up only certain sound signals.

Aiming for the Ears of a Machine that Distinguishes Sounds Spatially

Most animals have two sensory organs which they use to distinguish between sounds – their ears. Animals use both ears to compare the strength and timing of the signals they pick up and to assess the direction from which these sounds are approaching, as well as to be aware of sounds approaching from a particular direction. This sense of hearing is crucial for wild animals as it allows them to detect enemies and hunt for food in complete darkness, where their sense of sight is rendered useless. Associate Professor Nobutaka Ono (Principles of Informatics Research Division) is researching Microphone Array technologies that achieve sensing functions similar to the sense of hearing described above.

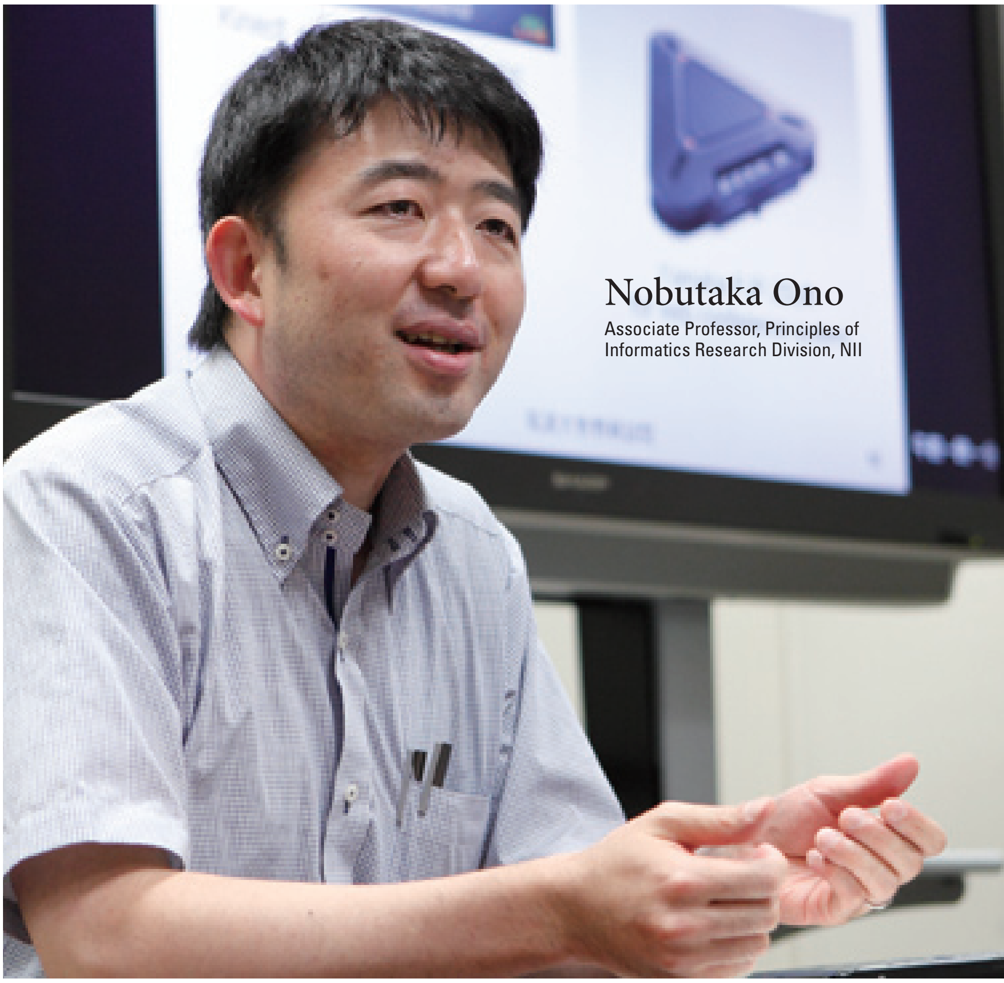
Associate Professor Ono says that in the Microphone Array he is researching at NII, the word "array" refers to "something that has been lined up." In other words, lining up multiple microphones creates superior sound sensing functions that can differentiate between sounds spatially, with the goal of application in areas such as artificial systems and robots.

The history of the development of sound recognition technologies is very long, and there are many researchers continuing their studies at present. Associate Professor Ono is paying particular attention to technologies that can use signal processing to emphasize only certain sounds. For instance, there are often multiple voices heard at one time; although we can easily pick them up with our sense of hearing, it is not possible to distinguish between them when they appear mixed up in a recording. Microphone Array technology makes it possible to extract voices and to correctly specify sound positioning, using signal processing of recorded sounds obtained by multiple microphones that are lined up.

If one of the functions of our sense of hearing is to recognize these sounds and voices, another important role is to alert us to the presence of any dangers around us. The Microphone Array technology can even be utilized in this respect. One unique example of actual application is the Shot Spotter, a system developed by an American start-up company. This system involves multiple microphones set up throughout city streets which can alert both security companies and the police if any abnormal sounds (such as gunshots) are detected. Such authorities are immediately alerted after the location of the event is determined. You can essentially think of this as a surveillance camera that works via sound. Associate Professor Ono thinks that security by means of sound will be become very effective in the years to come.

Extracting Only Desired Sounds from a Jumble

Associate Professor Ono is, currently researching a technology called Blind Source Separation. The term "blind" in this case refers to the unknown direction of a sound. In the method traditionally used when broadcasting centers needed to clearly record the voices of specific individuals, a microphone with an extremely strong directionality was placed near the individuals being recorded. Blind Source Separation technology allows for editing and processing that can be performed



Nobutaka Ono

Associate Professor, Principles of Informatics Research Division, NII

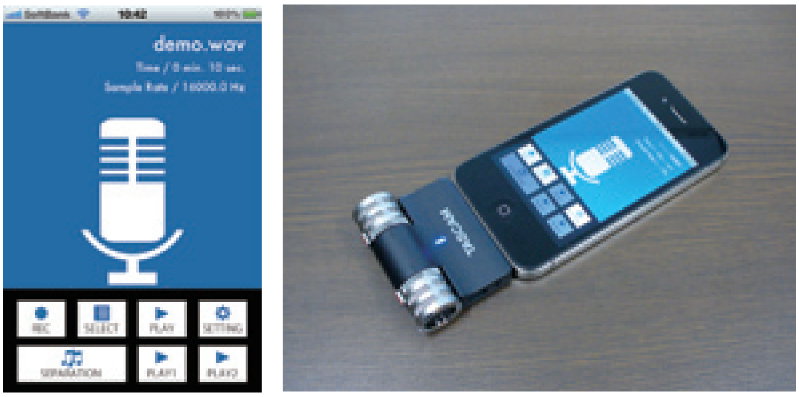
Processing Technology Growing Popular

freely after separating desired sounds from the jumble of unspecified sounds which are recorded with multiple microphones.

Imagine that you are recording your daughter's piano performance at her school play, and the sound of a person sneezing shows up in your recording. This technology is convenient, even in moments like this, when only the sounds you choose can be picked out after the fact.

There has been research in the past on technologies that use computers to separate multiple, overlapped sound source recordings. This, however, required time for calculation and operation processing. By developing high speed algorithms, Associate Professor Ono has succeeded in cutting the 50 instances of calculations needed previously down to about 10 instances. This success has also made it possible to install sound source separation applications on devices like smartphones.

■Blind Source Separation App for iPhone



With an iPhone equipped with a stereo microphone, you can simultaneously record the voices of people on both your left and right sides. Pushing the button labeled "Separation," immediately separates the signals of mixed up sounds so that each voice can be played back independently. It is expected that this device will be applied for uses in various fields, such as live recordings and preparing minutes for meetings.

Unsynchronized Microphone Arrays Break New Ground in Sound Sensing

Until now, it was crucial in Microphone Arrays that recorded signals were completed and synchronized in order to either estimate the position of sounds or to extract only certain sounds from sound signals recorded with multiple microphones. For instance, in order to investigate the direction sounds came from, it was necessary to line up two microphones and to detect sounds travelling at 340 meters per second using in units of time equivalent to 1/10000th of a second. Therefore, there was a need for a converter to convert sounds into digital signals to make perfect time synchronization possible. The cost, however, was high, greatly hindering the number of microphones (recording channels) in systems that could actually be used.

Here is what Associate Professor Ono says about this:

I began this research with the idea that synchronization via signal processing after recording sound with multiple unsynchronized microphones attached to mobile phones and computers can give us the ability to guess the position of sound sources. The more microphones there are, the greater the ability to both emphasize sounds and to speculate on them. Fortunately, with mobile terminals such as smartphones becoming more popular and IC recorders becoming more affordable, it has gotten much easier to perform research in this field.

Various Applications to Expect in the Smartphone Era

Research on Unsynchronized Microphone Arrays will result in various ways this can be applied to smartphones. One of these uses is keeping minutes at meetings.

Associate Professor Ono says:

Imagine that meetings were recorded using smartphones owned by each of the parties present at the meeting. During the meeting, there may be static noise or overlapping voices when people talk at the same time. After the meeting has concluded, however, by uploading these recorded signals to an internet server, time is automatically synchronized on the server-side so that each person's statements are emphasized and a compilation of minutes is automatically prepared. This is the kind of system that we are shooting for.

In music-related fields, people also make use of these capabilities in the recording of instrumental performances. Traditionally, when there was a need to make corrections after a musical performance was recorded, it was necessary to record the sound sources of each individual instrument (such as pianos and guitars) so that the sounds would not be mixed up. There may be a point in the future when we are able to correct and process the sounds of instruments, almost as if the recording had been mixed by a professional in a studio, by just lining up smartphones and recording with them by means of an unsynchronized microphone array and then separating sounds after the fact.

Furthermore, applications for meteorological information systems are also being investigated.

Associate Professor Ono gives his views on this subject:

It is conceivable, for instance, that when lightning occurs, there could be an observational system where smartphone applications pick up the sound of lightning, and this information is collected and processed on a server to estimate where lightning will occur. There are currently plans to carry out preliminary experiments at fireworks festivals by using multiple smartphones to gather recorded data from different locations. These experiments will show the areas where the fireworks produced noise and the locations where the recording parties were at the time.

Compared to synchronized signal technology, Unsynchronized Microphone Arrays still require much work in terms of signal separation and are still in the developmental phases of research. It is, however, extremely easy to increase the number of microphones. As such, we would like to make the most of the current situation in which devices can be deployed freely while aiming for actual use in a variety of fields. (Associate Professor Ono)

In this era, where nearly everyone carries at least one smartphone or microphone-affixed device, there is no doubt that the possibilities of sound sensing technologies which make use of networks will grow further and further.

(Written by Hitoshi Asakawa)

Analyzing Television Broadcasts as

Television broadcasts are an important source of information for viewers, while at the same time they possess the function of a sensor that observes movements in society, culture, economics and so forth. For example, news programs are sensors for what's going on in the world, and by means of watching commercials it is possible to analyze economy trends and business activities. This is a report on front-line research at NII, which aims to analyze the rich content only available on such a medium and use television broadcasts as sensors.

Shin'ichi Satoh

Professor, Digital Content and Media Sciences Research Division, NII

Extracting Information Useful to People from Vast Television Broadcast Records

It was in the year 2000 when the attention of Professor Shin'ichi Satoh (Digital Content and Media Sciences Research Division), who is involved in research on visual information at NII, began to be drawn to television broadcasts as a research subject. The trigger for this research was when he began to think that if required information can be searched for and extracted from recorded programs, we can spin out new value.

That system became fully operational in August of 2001. When it was first put into operation, archives were created with close captioning for the program "News 7," a news program produced by NHK. After that, and along with the lowering prices of recording media, the system was upgraded and expanded from August 2009. On this occasion, a framework was put into place so that all programming aired (24/7) on the 7 broadcasters in Tokyo could be recorded 24 hours a day.

Professor Satoh looks back on this in the following way:

The original idea of turning television broadcasts into sensors was born when I had the chance to be a part of the Feasibility Study (Study of Cost/Effect) for Cyber-Physical Systems, a 2011 project at the Ministry of Education, Culture, Sports, Science, and Technology. It was a trial to analyze occurrences in the real world using sensors in cyber-space such as surveillance video footage and electric grid usage condition data for the entire country, among other things. While working on that project, I began to wonder if we could apply the system to television broadcasts as well.

Discovering New Value from Television Broadcast Analysis Results

Professor Satoh says that there are various approaches to detecting video images; one of which is looking for repeating patterns:

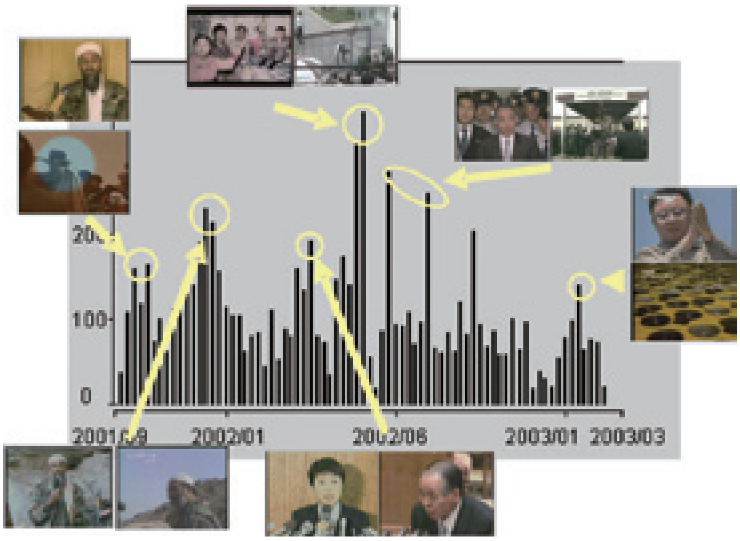
When Japan beat Tunisia at the 2002 Japan/Korea World Cup, the image of Tsuneyasu Miyamoto of Team Japan wearing his face protector was shown almost daily. Traditionally, in the field of image analysis, it was difficult work to mechanically analyze what was being shown in an image. However, with repeating television broadcasts it is very easy to process and extract. By analyzing the frequency of image repetition and the time-frames and occurrence patterns, one can understand which topics society is talking about.

As part of Professor Satoh's research, he is also conducting experiments in which all broadcasted television commercials are recorded. Anything found within the large amount of recorded video data that is 15 or 30 seconds long, and is repeated throughout in the same format, is defined as a commercial.

By developing very high speed algorithms, we have made it possible to analyze three years' worth of commercials from five different channels in only one week. While advertising agencies also perform this kind of analysis, they perform much of the work manually, and therefore it usually takes at least a month to complete the analysis. The NII system speeds up this process, and as a result, we are able to make determinations; for instance, knowing that over a one-week time frame, on-air times for automobile manufacturer-related commercials are concentrated on Saturday late at night, we can also extrapolate that these commercials are probably geared towards businessmen, who are the viewers most likely to be watching during that time-slot. Or, for instance, knowing that the number of candy products or sweets commercials shown during daytime slots has decreased dramatically since 2011, we could guess that economic conditions may be affecting the sweets industry. Complete records of commer-

Sensors

■ Discovering New Value from Multimedia Archives



An analysis of important short patterns that repeatedly appear from stock news video footage spanning 10,000 hours; by looking at the time-slot and frequency of appearance, we can bring out the hidden value of information that regular television viewers cannot see.

cials over a long period of time can serve as a very effective sensor to indicate economic conditions and trends.

The Promotion of Comparisons with Multiple Mediums and Inter-field Joint Research

We can gain an awareness of what kind of information the television stations are attempting to transmit if we analyze news programs. For example, something that became a subject of debate after the Great East Japan Earthquake was whether or not the television stations are actually giving us information that is accurate. Professor Satoh conducted fix-point observations of news programs and compared them to a massive amount of Twitter posts, a highly significant source of information on the Web. Through this comparison, he wanted to know if the information we get on television contained any biases, as well as why things which get so much attention on the Web are not broadcasted on television. He has researched the degree in which mediums are linked by looking at data obtained from sensors.

Furthermore, in terms of Professor Satoh's project, he is also conducting research with respect to information analysis on the Web. This is the Ministry of Education, Culture, Sports, Science, and Technology -sponsored "Development of Societal Analysis Software and Construction of a Media-Web Analysis Base." Together with Professor Hayato Yamana from Waseda University, Associate Professor Masashi Toyoda and Professor Masaru Kitsuregawa of Tokyo University, he has been working towards research results through the collection of data from the Web over the span of 10 years.

Professor Satoh adds that in research on the theme of the Noda Administration, he is analyzing if there are any differences between the perceptions of how the behavior and movements of the prime minister are reported on through the extraction of text, images and video footage from both Web archives and television broadcasts. In this way, through putting together information from the Web and broadcasted video images and analyzing, that we are looking for possibilities in terms of more deeper analysis on what is happening in society.

Challenging the Possibilities of Enhancing Value of Television Broadcasting

What kind of technological steps are there with respect to this research that makes television into a sensor? Professor Satoh discusses this below:

I began research into analyzing what is depicted in visual information about forty years ago and I can say that it is exceedingly difficult for a computer to understand things the way we do as human beings. Right now, technology for recognition of facial information is moving forward. However, currently no computer program exists that can be given an unknown image and tell us with certainty what it is. At this point in time, it has been around-the-clock trial and error for many researchers with respect to methods for recognition of an infinite amount of existing concepts relating to an image that can represent various expressions.

By means of large scale analysis of television broadcasts in which diverse visual expressions exist together with an infinite number of concepts, Professor Satoh thinks that we can achieve a breakthrough in terms of meaning-analysis technology for images. If significant progress can be made in research with respect to semantic analysis of images, the viewing public can be provided with the information they need when they need it.

In the future, by being able to provide simultaneously information on other channels and comparisons with web pages to viewers, what was once a passive television viewing experience can be made into one in which viewers will have access to a panoramic of information. If this achieved, broadcasters will be able to provide more quality content and viewers will be living in a world where they can enjoy a higher level of quality information. (Professor Satoh)

Professor Satoh likes to watch television. He is certain that there is potential to provide superior content. He wants his research to help restore television so that video will have long term value, instead of ratings oriented programming that is here today and gone tomorrow.

Professor Satoh's ultimate aim is to have computers automatically understand all television broadcast content the humans consume and understand. When that becomes a reality it could bring about the birth of a sort of "television concierge," which could be the ultimate "video information assistant." I cannot wait to see that day.

(Written by Hitoshi Asakawa)

New Value Created by a Shift from “Who Is

Imagine a day when we will be able to manipulate information on a display just by looking at it, when public facilities will automatically monitor potential causes of hidden dangers and problems. The technologies of gaze estimation and gaze guidance, currently being developed through collaborative research between NII and the University of Tokyo, will exert a creative influence on information design in the future. These technologies analyze the mechanics of visual attention and read the behavior of humans using vast amounts of information collected by media sensors. We asked two researchers about the current state and future outlook of the research and development of these technologies.

The World’s First Research and Development of Technologies that can Sense Human Gaze

First of all, could you tell us how your collaborative research began?

Sugimoto: I first met Professor Sato more than ten years ago when I was participating in a research project. Later, after I moved from Kyoto University to NII in 2002, I started collaborative research with Professor Sato on a different project in a specific area of informatics. The research we conducted for this project was on a topic related to gaze estimation—how to recognize the body and hand gestures of people and rapidly measure their facial movements using information from cameras and sensors embedded in the environment. In 2009, our research proposal was selected by the Japan Science and Technology Agency as a CREST Project, and we began doing collaborative research on how to apply the “estimation and guidance of a gaze in everyday living environments” to the construction of an information infrastructure.

“Gaze estimation” is still an unfamiliar term. What kind of research does it entail?

Sato: This research tries to figure out what is going on inside a person’s mind and what their next action or state of consciousness will be by predicting what they will focus on next. A method for sensing the human gaze is needed to conduct this kind of research. At first, researchers used cameras mounted on subjects’ heads or embedded in their environment to sense facial direction and eye movement. This method works effectively in a controlled environment such as a research lab, but it is not practically useful for estimating people’s gazes when they are walking down the street or going about their business in other everyday environments. These kinds of situations require high-precision measurement technology and complicated tuning. At this point, we turned our attention to the application of “visual saliency” for a gaze estimation technology that would be easier to use.

Sugimoto: To put it simply, “visual saliency” is what makes certain things stand out and draw our attention. The visual elements that tend to catch people’s attention, such as contrast in color or brightness and certain kinds of movement, have been modeled to some extent. If variables such as the characteristics of the retina and the direction a person is walking in are taken into account, these models can be used to predict “what the person will look at next.”

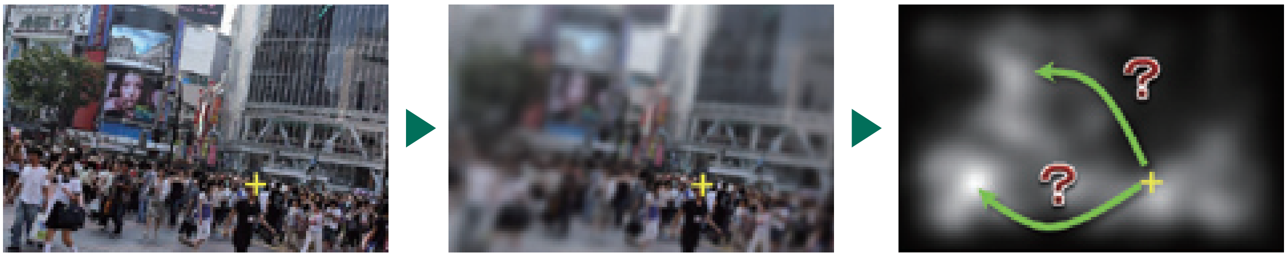
Sato: Up until this point, no one had thought of estimating a gaze based on prompts from the environment. In other words, gaze estimation is grounded in a shift in thought from focusing on the people who are doing the looking to what is being looked at, and it is the world’s first research to have taken its cue from visual saliency.

Controlling Gaze through the Application of Visual Saliency

It sounds like the behavior of “looking” has been researched from various angles.

Sugimoto: There are two main models for the human gaze. One is the bottom-up model. This model accounts for natural, reflexive gaze behavior, such as when the attention is drawn to flashes of light. The other model is the top-down model. This model is closely linked to a person’s

■The Process of Gaze Estimation Using Visual Saliency Models



A. The technology predicts where a gaze currently focused on a man walking by (marked with a “+”) will go next.

B. The eye will be susceptible to stimulation from factors such as luminance and color at the center of the gaze and brightness in the area around it. These variations (distinctive characteristics) in the person’s field of view will be taken into account.

C. After models that account for the high visual saliency of people’s faces and eyes and certain kinds of movements are considered, along with the distinctive characteristics in the person’s field of view, the sequential path of the gaze is predicted. Researchers are currently working on improving the accuracy of prediction.

Looking" to "What Is Being Looked At"

Akihiro Sugimoto
Professor, Digital Content and Media Sciences Research Division, NII

Yoichi Sato
Professor, Institute of Industrial Science, The University of Tokyo

recognition of objects based on the gaze behavior that has been cultivated by his or her experiences. The visual saliency model that applies depends on where the gaze is focused. For example, when our attention naturally shifts to objects such as a person's face or eyes, or a car, it is due to the top-down model of visual saliency. And when the attention is drawn to color or brightness, the bottom-up model applies. By building upon these models, we will be able to further increase the accuracy of gaze estimation.

And could you tell us what kind of research you are doing for your other research topic, gaze guidance?

Sato: Gaze guidance is the concept of controlling the human gaze. The use of visual saliency is essential for it to be carried out in a natural manner. For example, we can subtly draw a person's attention to an object we would like him to focus on by adjusting its color or brightness as appropriate, based on its visual saliency. We have also been doing research for the top-down visual saliency model, which involves people's faces and eyes, by first drawing the subject's attention to a robot's face and then trying to guide their gaze through joint attention. Through the use of models with high visual saliency in a variety of different environments, we can get people to look where we guide them in a natural manner. Such gaze guidance technology can be used for a variety of purposes.

For Example, What is Occurring at an Intersection Where There are Frequent Accidents

What are some of the possible applications for gaze estimation and gaze guidance technologies in society?

Sugimoto: Say there is an intersection where there have been a lot of car accidents. While there are various potential causes for the frequent accidents, an examination of the intersection from the perspective of visual saliency may show that the presence of billboards and designs that distract drivers is to blame. The removal of distractions that can lead to accidents is one of the possible applications of gaze guidance. There is no problem, of course, with informational signs that alert drivers to the possibility of accidents and have a high degree of visual saliency, as long as they are installed in safe locations. I think gaze estimation and gaze guidance will greatly expand the potential for information design in the public sector.

Sato: When we first started this research, we were thinking about its potential applications in schools and nurseries. We were studying gaze guidance methods that would draw students' attention to the blackboard during class and the creation of environments that would naturally direct people's attention to areas that would otherwise be difficult to see. These technologies can, of course, be applied to this kind of information design, but I think they also have a lot of potential applications in advertising and other areas in the field of business. For example, billboard advertisements could be designed using the concept of gaze guidance, and cameras embedded in the environment could measure each advertisement's "gaze rate," providing feedback for the further improvement of the design.

Designing a New Society and Way of Life Centered on "Gaze"

Please tell us about the future prospects for your collaborative research.

Sugimoto: The goal for our research is "to know people." Gaze estimation and gaze guidance are research areas that help us better understand human behavior and improve it. Information design for society and the public sector is another major theme in our work, and we hope to promote applied research that is involved with people's everyday lives.

Sato: Display devices using gaze reaction models are also being developed. In these models, information is arranged in the appropriate position to guide the user's gaze and can be manipulated by the direction of the gaze alone. This technology has not yet been implemented, but I believe that information design centered on "gaze" will change our future way of life.

Sugimoto: Gaze technology could also be used in conjunction with the multimedia sensing and immersive visual communication technology that Cheung Gene at NII has been working on. One of its applications to multimedia sensing is the sensing of information about people and environments and the output of information connected to gaze estimation and gaze guidance. Research on gaze estimation will also help improve technology that predicts the gaze of the person you are talking to and adjusts her 3D image to compensate for shifts in gaze, producing the same sense of realism as if she were right there in front of you. I hope that Professor Sato and I, and Associate Professor Gene, can all take advantage of the findings of our respective studies and think of ways we can put the technologies of gaze estimation and gaze guidance to good use.

**For more information on Associate Professor Gene's research findings, see page 2 and page 10.*

(Written by Yoshikazu Takahashi)

Immersive Visual Communication Will Allow

Sometimes you cannot help but wish that you could shake hands with a colleague on the other side of the world. Enter "immersive visual communication," a technology that makes startlingly realistic communication possible, even across great distances. Image data recorded by multiple cameras is processed, transmitted and synthesized, allowing users to capture a 3D scene from any angle. It is an advanced form of communication made possible by multimedia sensing.

The networked system can reproduce a scene with the same sense of realism as if it were happening right in front of your eyes.

This technology allows users to make eye contact as they converse via video conference. Hospitalized patients can enjoy watching a sports game from any angle they like while lying in bed. Immersive visual communication produces a 3D visual environment that appears to viewers as if it were unfolding right in front of their eyes, even if they are separated by long distances. Gene Cheung, an associate professor in the Digital Content and Media Sciences Research Division at NII, has been researching this technology in collaboration with Antonio Ortega, a professor at the University of Southern California.

Associate Professor Cheung says, "From our perspective, the imaging technology that is currently referred to as 3D is more like 2.5D. What Professor Ortega and I are trying to achieve is a 3D effect that is truly realistic—video that looks 3D no matter what angle it is viewed from, whether from the side or from the opposite direction. An immersive 3D visual experience can be reproduced by recording video of an object from every angle with multiple cameras and then selecting and combining the necessary images."

Gene met Professor Ortega about fifteen years ago when Gene was a graduate student, and they began their joint research five years ago while Gene worked in a corporate lab. Professor Ortega now visits Japan twice a year as a visiting professor at NII. The two professors usually share the results of their research via video conference as they collaborate across the globe.

Antonio Ortega
Professor, University of Southern California

Gene Cheung
Associate Professor, Digital Content and Media Sciences Research Division, NII

Depth image compression technology will make free viewpoint 3D video a reality.

There are three technical processes that make immersive visual communication possible. First, large amounts of image data recorded by numerous cameras are encoded and compressed. Next, as the data is transmitted over the network, the impact of packet loss on video quality is kept to a minimum. Finally, each viewer' s eye gaze and head movements are predicted, and requests are made to the sender side to transmit the necessary data. You can carry on a conversation with a distant colleague while maintaining constant eye contact, thanks to real-time adjustments that compensate for shifts in gaze when she moves her head, and so on. The compression technology used for the image data is particularly important in this process.

"Depth images are the key to achieving a 3D visual effect," Prof. Ortega explains. "When regular images are recorded, one surface of the object is captured, but the depth of the object cannot be captured. In our research, we use special cameras to record depth images as well as regular images. Through the synthesis of these images, we can make it so that you can see the actual shape of an object from any viewpoint. To do this, the image data we take must be compressed so that it can be transmitted over the network. For depth images, however, the compression methods that work well for regular images are not so effective, so we are currently researching optimal compression methods that won' t produce severe distortions in the final synthesized image the viewer sees."

Thanks to other efforts, such as the design of systems that prevent packet loss (the loss of data during transmission) by optimizing data redundancy, and research on technology that predicts each person' s gaze and continually takes images in synchronization with the other person' s movements (see pp. 8-9), researchers are steadily moving toward making their vision of immersive visual communication a reality.

Users to View 3D Video from Any Angle

The Synthesis of Regular Images and Depth Images

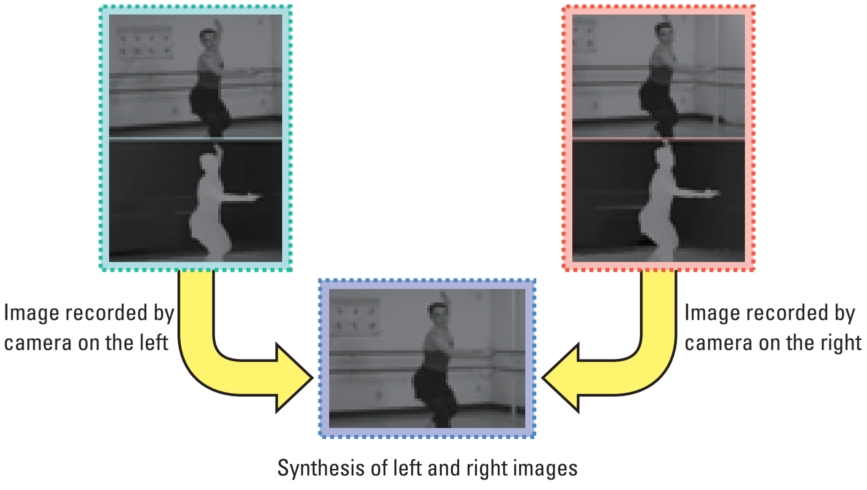
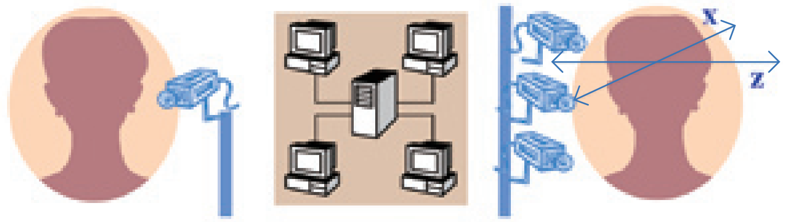


Image recorded by camera on the left
Image recorded by camera on the right
Synthesis of left and right images

The regular images are above, and the depth images are below. Combining the left and right images reproduces a 3D visual effect in which the depth of an object can be captured and viewed from different angles.

The 3D Visual Effect Reproduction System



An image recorded by a single camera is only able to capture one surface of a given object.

By recording images of the object from every angle with many cameras and then combining the necessary images, it is possible to reproduce a free viewpoint 3D visual effect, in which the object appears three-dimensional no matter what angle it is viewed from.

As needs increase, a whole new world of visual communication is opening up.

How will immersive visual communication be implemented in the years ahead?

"Now, for example, when you watch a concert on an online video site, you can only watch it from a pre-determined point of view," says Gene. "Once our technology is implemented, you will be able to enjoy the concert in any way you like, as if you were actually there and could watch it from anywhere in the audience. You will even be able to change seats in the middle of the concert."

The technology also shows promise as a tool for educational or simulation purposes. For example, it could give viewers a 3D demonstration of what an operation is actually like, allowing many medical students to observe the same kinds of scenes if they were actually present at a surgery.

Gene also says that he has been contacted by a professor at Tohoku University who has mounted a 360-degree camera on a car and has been taking images of the areas struck by the Great East Japan Earthquake. The professor wanted to know whether the images could be converted into 3D and processed so that the current condition of the disaster sites could be shown from any viewpoint, and they have been discussing how they could implement his idea.

As a tool that will provide the foundation for communication in the next generation, immersive visual communication will be able to support various applications.

"Our technology, in combination with the needs of users, will give rise to applications that have never been thought of before," says Prof. Ortega. "The 3D conversion of the earthquake images is a good example. Synergy with the knowledge of users can create completely new forms of visual communication. Unimagined possibilities for the visual world of the future are opening up."

Collaborative research with global companies towards the next generation of standards has begun.

So, as 3D visual communication becomes more immersive, what directions will their collaborative research take?

"Raw data compression technology is certainly an important research area, but we think that 'compact representation' is a more fundamental problem," says Gene. "In compact representation, only the necessary elements are encoded, after applying appropriate data transforms, out of the large quantity of image data that is captured. We are currently doing research on this system and the standards for data selection."

Around the world, people are advancing research into technologies that enables immersive visual communication. The area that the Cheung-Ortega research team excels in is the compact representation of 3D geometry. This is a technology that makes appropriate adjustments to convert the coordinates of an object in three-dimensional space into coordinates on a display device, and compressing depth images is one of the techniques it uses. Global companies from the video conferencing and entertainment industries have shown interest in this research and become involved in collaborative studies. Once these technologies are implemented, new global markets will emerge, and the next generation of global standards for data transmission and 3D communication will be born. This is the path that is opening up before us.

The future of communication, where we will be able to enjoy an alternative reality created by 3D visual effects from the comfort of our living rooms, is just around the corner.

(Written by Yoshikazu Takahashi)

NII Essay

Color Is in the Eye of the Beholder: The Varying Perceptions of the Colors of the Rainbow

Imari Sato
Associate Professor,
Digital Content and Media Sciences
Research Division,
NII

What exactly is color? What we perceive as color is actually “light,” and the person who first realized this fact was the famous scientist Sir Isaac Newton. Newton revealed the true nature of color through experiments in which he passed a beam of white light through a triangular glass prism and observed how it separated into a spectrum of rainbow colors. At the same time, Newton noted the fact that light itself is colorless. In Optics, he writes, “For the Rays, to speak properly, are not colored. In them there is nothing else than a certain power and disposition to stir up a sensation of this or that Color.” So if the true nature of color is light, but light itself is colorless, what exactly does it all mean?

Even when there are rays of light that we see as red, it does not mean that the light itself has a color. For example, even if you see a red apple in front of you, it does not mean that the “red apple” actually exists. It is our human sense of sight through which we observe the apple that gives it its color. In other words, Newton knew that color was a perception produced by the stimulus of light.

We now know that the human visual system perceives color after it converts the visible range of light into the tristimulus values of “red,” “blue” and “green.” There are, of course, individual variations in human color sensors, and there is no guarantee that the color you see is the same as the one someone else sees. For example, there is no such thing as a universally shared concept of the colors of the rainbow. People in Japan generally believe there are seven colors in the rainbow, while people usually identify six colors in Britain and the U.S., five in Germany, and as few as four or three in other countries.

Interestingly, the real world appears quite different when observed through the light sensors of other organisms. For example, the male and female cabbage white butterfly both look yellow to us, but they look different from each other when viewed through the eyes of insects, which perceive ultraviolet light, enabling them to easily distinguish the male from the female. The vision of birds is said to be better than that of humans, allowing them to experience a richly colored world derived from combinations of four colors — the “red,” “green” and “blue” colors that humans perceive as well as “ultraviolet.” And despite the familiar red cape used in bullfights, cattle are actually unable to distinguish the color red. The red color is used in these fights to excite the audience. Red excites us, not the bull. The world of colors is truly fascinating—an endless source of wonder.

5colors!
6colors!
3colors!
4colors!
7colors!

Weaving Information into Knowledge
NII

NII Today No.43, Oct. 2012 This English language edition NII Today corresponds to No. 57 of the Japanese edition

Published by National Institute of Informatics, Research Organization of Information and Systems
http://www.nii.ac.jp/
Address: National Center of Sciences
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430
Chief editor: Yoh'ichi Tohkura Cover illustration: Makoto Komori Photography: Miyuki Kumagai
Production: Kosaido Co., Ltd.
Contact: Publicity Team, Planning Division, General Affairs Department
TEL: +81-3-4212-2164 FAX: +81-3-4212-2150
e-mail: kouhou@nii.ac.jp

On the Cover

Thanks to advances in multimedia sensing, realistic communication that engages all the senses is becoming possible. For example, videoconferencing is evolving from 2D to 3D. People who are physically far apart will be able to view 3D images of each other projected right in front of their eyes, from any angle. A future we have only seen in movies is already coming within our reach.