ISSN 1883-1974 (Print)
ISSN 1884-0787 (Online)
National Institute of Informatics News

NII Today

National Institute of Informatics News

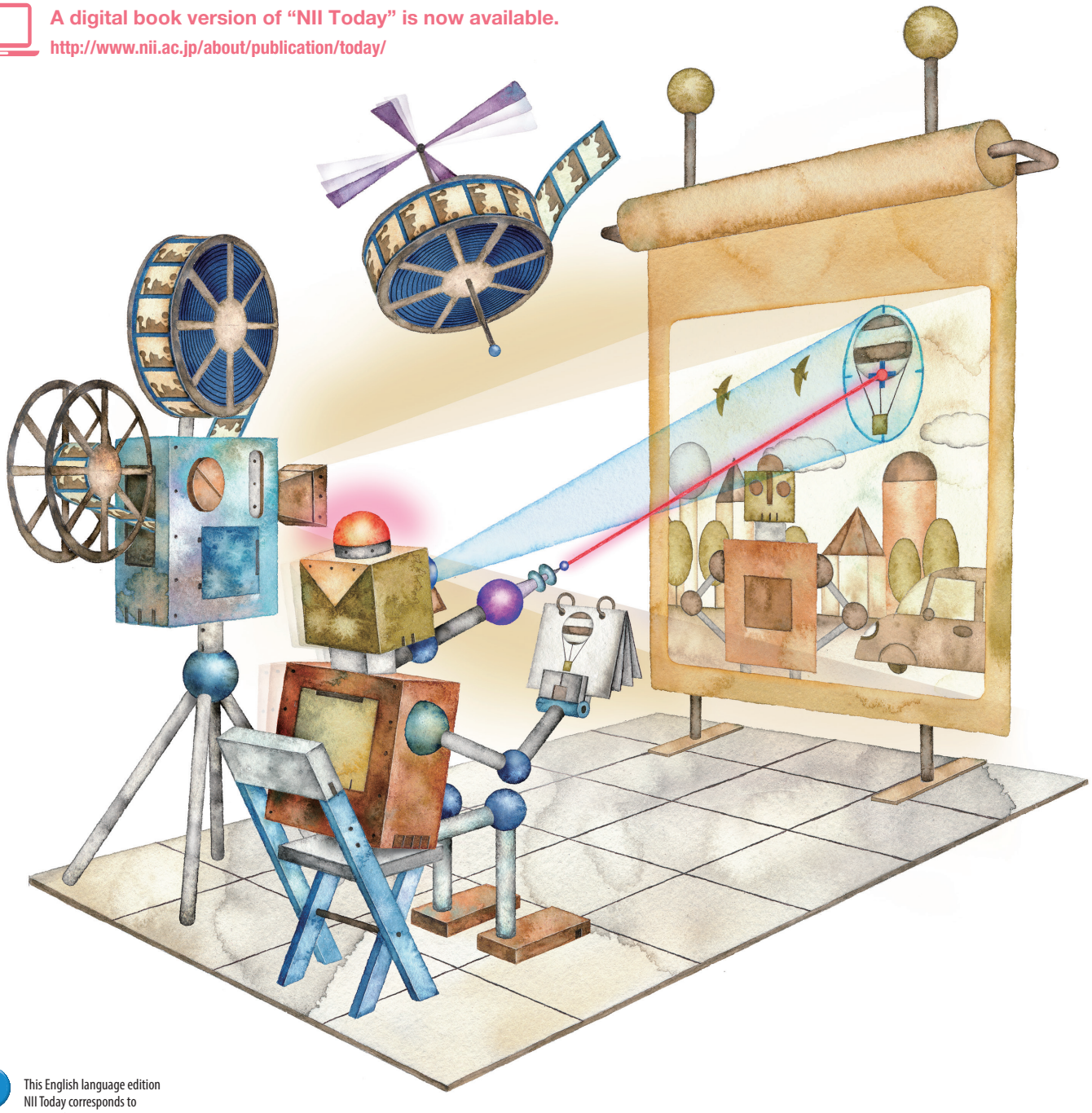
53
Apr. 2015

Feature

Images

New Trends Arising from Information Technology

A digital book version of "NII Today" is now available.
http://www.nii.ac.jp/about/publication/today/



This English language edition
NII Today corresponds to
No. 67 of the Japanese edition

NII Interview

Television is a Sensor that Perceives Trends in Society

Media such as television, radio, and newspapers are mirrors reflecting our society. They are also sensors that perceive trends. We spoke with Shin'ichi Satoh, a professor at NII engaged in research on systems to analyze the meaning of video, and Hidehiko Koguchi, Director of M DATA Co., Ltd.*, which provides, examines, and dissects TV metadata describing the broadcast content of television programs, about the use of television as a sensor, today and in the future.

* M DATA Co., Ltd. http://mdata.tv

Tsujimura: First, what kind of data do you collect and utilize?

Koguchi: Our company gathers the information of how and when something was shown on television. Television is a medium that integrates video, audio, and text. All of this information is captured as text to make it easy to handle.

The information resembles the bibliographic data of a book. One searches for a target book based on data such as title, author, number of pages, and genre. Bibliographic data are data about a book, which is itself a collection of data, so bibliographic data are what we call "metadata".

Metadata must also be assigned to television video images. We express video images in text, and create a database of this text. Then, if we enter "Godzilla", for example, we can find out how and when Godzilla was broadcast in the past year. Television constantly transmits information that symbolizes trends in the world. When a viewer watches it, that behavior is fed back to the party who transmitted the information, and it then appears in an increasing number of situations. One can also say television is a sensor that captures this flow of information in the world.

Shin'ichi Satoh
Professor, Digital Content and Media Sciences Research Division, National Institute of Informatics (NII)

Tsujimura: Mr. Satoh, what kind of research are you involved in?

Satoh: About 20 years ago, I started studying image and video search technology. This was around the time when technology emerged for searching several thousand to tens of thousands of images for something similar, based on color distribution. At the time, however, it was only possible to search a limited group of images, such as only animals or only buildings. I thought that the ability to obtain desired information from television, which streams a diverse range of video images, would be useful technology.

First, I created a system for recording television broadcasts and stored 400,000 hours of video. I also developed a technique for extracting commercials, which were an obstacle to my research. It was also possible to determine which of the extracted commercials were the same. That being the case, it was possible to find out which and how many commercials were aired on which days of the week and in which timeslots, and by analyzing those patterns, the consumer strategies of companies became apparent. For example, there were almost no commercials for expensive products after the Lehman shock, but when the economy improved, the number of commercials for luxury cars increased. I thought this could be used as a sensor for society.

Koguchi: Television influences our lives as a source of information. Marketing people want to know how and when such information is mentioned on television. An idea on how we could provide a service to fulfill this need was the impetus for the company M DATA.

Meanwhile, I also dream of casting a net over the intersections of all information flows across the world to get a high-level perspective. For example, when a political issue in one country is reported in other countries, that issue can be judged from a different viewpoint. If you were able to view both reports together so that you could compare them, it would help give you the basis for a more neutral decision and provide a broader range of alternatives.

Information is always inherently biased. Being able to see how it is biased is a great advantage to people. It would be exciting if such "media of the media" were possible with metadata.

Satoh: How do you actually produce metadata?

Koguchi: A staff of about 80 experts work on a rotating schedule, 24 hours a day, 365 days a year, entering information while watching television. If we could automate it, that would be great, but even the latest audio-text technology cannot avoid producing a few percent of errors. So far, recognition and interpretation by humans is overwhelmingly faster than by machines.

Satoh: We are showered with the criticism, "Isn't it sufficient to have humans watch video and assign metadata?" We respond by saying that it is arduous to watch all video in large quantities, but that's what M DATA does, so it's a problem. (Laugh)

Hidehiko Koguchi
Director, M DATA Co., Ltd.

Koguchi: I think your research is very meaningful. A good supportive example is the Formula One car race broadcasting. For a sponsor company of the F-1, how many seconds the company's logo appears on TV and in what part of the screen are vital pieces of information to measure the advertising effect of sponsorship. If such information could be captured and analyzed exactly and automatically, it would be far more efficient than having humans do it.

Satoh: Our research on "object search" is continuing, and is ideal for logo detection. This kind of detection is becoming fairly possible through our instance search technology, which is competitive worldwide.

Continuing the previous thread, doesn't human subjectivity enter in when extracting metadata to create a news summary?

Koguchi: We have devised an input interface so that emotional judgment does not enter in. Because we extract subjects that have no room for subjectivity as much as possible, there is little fluctuation in data no matter who enters it.

Satoh: But there are cases where it is important to extract emotional information from a video. This is an area where machines have difficulty.

Koguchi: Although it is difficult to grasp directly, if "frequency and time of exposure" are analyzed, it is, for example, possible to capture the phenomenon of some issue being treated as a sensation.

Satoh: What do you think about using a wearable sensor to capture the stimuli a person watching television is receiving?

Koguchi: It would be difficult to commercialize anytime soon, but we would like to try it. What a person watching television posts on Twitter in real time and how empathy spreads is becoming a subject of monitoring. The approach of handling records of a person's life together with television metadata is also starting to be seen.

A Word from the Interviewer

The conversation continued to evolve with changing topics, extending beyond the allotted time. The impressions that remain with me are Japan's problems of "strong technical prowess but a lack of ideas to leverage it" and "the lack of a steadfast system of supporting basic research". As a result, we are lagging behind the United States in both the computer industry and image recognition technology. A revision of our R&D procedures is necessary.

Tatsuya Tsujimura
Editorial Writer and Senior Staff Writer, Kyodo News

Joined Kyodo News after graduating from the Second Department of Physics, Faculty of Science, Tohoku University, in 1984. After working at the Osaka City News Section, Otsu Bureau, Sapporo Bureau, and Kushiro Bureau, joined the Broadcast News Section and Science News Section at Kyodo News Headquarters.
Joint post as Senior Staff Writer since 2005. Broadcast News Section in 2006. Senior Staff Writer and Editorial Writer since 2007. Akita Bureau Chief in 2010. Present post since September 2013.

Generating a 3D Shape Model in Real Time

Using an Inexpensive Consumer RGB-D Camera

NII Professor Akihiro Sugimoto is researching techniques for efficiently generating three-dimensional (3D) shape models based on data obtained by an RGB-D camera, which is capable of acquiring depth information in addition to color images. He developed a technique for efficiently representing a 3D shape model that requires dramatically less memory than conventional methods. Applying this technique to generating a 3D shape model of a person's face and expression and sending it from a smart phone to a social networking site is also being explored.

The era of easily manipulating real-time 3D shape data is coming

One impetus for Professor Akihiro Sugimoto embarking on research on generating 3D shape models from data obtained by an RGB-D camera was the fact that RGB-D cameras have become incredibly inexpensive. An RGB-D camera is a sensor capable of acquiring depth (D) to a target as well as color images (RGB). Until very recently, expensive range finders were primarily used as sensors for measuring depth. That situation changed a few years ago when inexpensive consumer RGB-D cameras became available, led by Kinect, introduced by Microsoft (USA) in 2010 as a peripheral device for game systems.

"It would be an overstatement to say that I wanted to conduct this research once Kinect came out. But I did think that I wanted to do something new that was affordable and more people could use," says Professor Sugimoto.

RGB-D cameras will become even smaller and less expensive in the future, and are expected to come into wide use in the form of a feature built into smart phones. Moreover, processor performance will be improved further and the graphics processing unit (GPU) for image processing will also significantly advance. That being the case, it will become possible to capture in real time not only a two-dimensional (2D) image sequence as a conventional camera does, but also a stream of 3D data. For example, using a smart phone with a built-in RGB-D camera will allow us to acquire 3D data, to generate a 3D shape model in real time, and then to post it immediately to a social networking site.

Actually, real-time processing of 3D shape models is already possible using Kinect and the GPU of a notebook PC. The same processing capability in a smart phone is expected in the near future.

Akihiro Sugimoto
Professor, Digital Content and Media Sciences Research Division, National Institute of Informatics (NII)
Professor, Department of Informatics, School of Multidisciplinary Sciences, the Graduate University for Advanced Studies

Manipulating 3D shapes by using 2D data

One particularly notable point about Professor Sugimoto's research is the fact that he developed a novel method for representing a 3D shape model. Through this technique, memory consumption is dramatically reduced and modeling of 3D shapes in real time and on a larger scale becomes possible. Low memory consumption is advantageous in not only processing but also transmitting data.

The technique for representing a 3D shape model primarily used today consists of partitioning 3D space into units called "voxels" and treating it as the set of voxels, that is, every point in 3D space is represented by a voxel. This method is wasteful because it allocates memory even for many voxels corresponding to space where nothing is present, which consumes an enormous amount of memory.

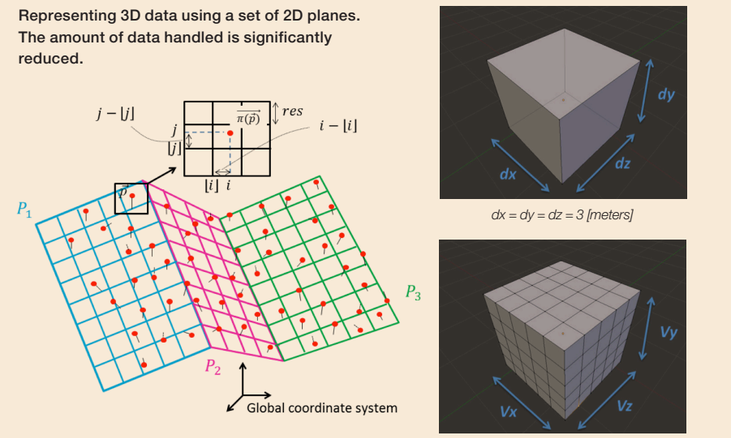
Professor Sugimoto's technique, on the other hand, represents 3D space using 2D planes. A 3D shape model is represented by making each 3D present point correspond to a point on a plane, and the information of how much the 3D point "deviates" from the corresponding point on the plane is kept as an image. It is sufficient to maintain data of a few 2D planes for a 3D shape model, rather than handling a huge number of voxels. A 3D shape model is manipulated by using 2D data, so to speak.

Memory consumption is dramatically reduced by this technique. If the size of a target that you want to model is large or if you want to scale it up, conventional methods quickly use up all memory and modeling can only be carried out in a limited range. With Professor Sugimoto's method, however, this is not the case. With his method, 3D shape modeling of a larger-scale target can be realized.

"Although it depends on how complex the environment is, with conventional methods, you can manage to model up to a 5 m^2 room, but using our technology, you can even model the space outside the room in real time," says Professor Sugimoto.

A variety of techniques studied over many years in the field of image processing can be applied to the 2D plane representation. For example, widely used image compression technology can also be applied to reduce required memory further.

Representing 3D data using a set of 2D planes. The amount of data handled is significantly reduced.



Carrying out "theory-oriented and application-oriented research together"

Professor Sugimoto has adopted a style of conducting theoretical research and applied research in parallel.

"Discrete geometry" is a research field Professor Sugimoto has been studying for many years. Errors caused by digitization, i.e., discretization errors, will always exist no matter how much resolution is improved. Discrete geometry is helpful to identify the theoretical bound in precision under the existence of discretization errors. As the resolution of digital images becomes higher, we tend to forget that there is always a limitation in resolution. "It is meaningless to pursue only precision in 3D modeling while forgetting about the precision of the original digital image," says Professor Sugimoto.

While undertaking this theoretical research, Professor Sugimoto is also engaged in the applied research of modeling 3D shapes using an RGB-D camera. He confides, "At the risk of it being misconstrued, I would say that, traditionally, researchers who placed importance on theory and researchers who placed importance on application tended to criticize each other. For myself, I think 'well, let's do them both.'" His unique technique for representing a 3D shape model using 2D data may be an idea born from this research style.

From generating 3D models of faces and expressions to entire city data

One of the targets that Professor Sugimoto is taking on in 3D shape modeling is the human face. If the 3D shape of a person's face can be modeled in real time, it is possible to more accurately convey subtle changes in expression. "For example, facial expression is really important when providing emotional care to an elderly person located in a remote region," says Professor Sugimoto. In such delicate communications, it is important to transmit data representing a person's expression more accurately.

As a prospect for the future, there is a trend toward handling 3D shape data of cities. "I think that the entire city of Tokyo will have been modeled in 3D in five years when the Tokyo Olympics are held," says Professor Sugimoto. While collection of this gigantic amount of 3D data progresses, 3D shape data can also be easily acquired using RGB-D cameras. It seems the day when people will use a diversity of 3D data in their daily lives is not far off.

(Interview/Report by Akio Hoshi)

Immersive communication evolving through “graph signal processing”

What are the new techniques for improving compression, interpolation, and denoising in images and video?

“Graph signal processing” technology is a relatively new technique used to efficiently analyze and process signals in social networks and sensor networks. We spoke with University of Southern California Professor Antonio Ortega, a pioneer in research on applying this technology to images and video, and NII Associate Professor Gene Cheung, who has been engaged in joint research with Professor Ortega for many years, to learn what kind of innovations graph signal processing technology will bring to “immersive communication” and to the world of images and video.

Antonio Ortega
Visiting Professor, National Institute of Informatics (NII)
Professor, University of Southern California

Gene Cheung
Associate Professor, Digital Content and Media Sciences Research Division, National Institute of Informatics (NII)
Associate Professor, Department of Informatics, School of Multidisciplinary Sciences, the Graduate University for Advanced Studies

Clarifying signal structure through “points” and “lines”

–– Can you provide a simple explanation of “graph signal processing”?

Ortega: A graph is a way to represent the structure of information using nodes and edges. As an example that is easy to understand, let’s think of users of a social networking site such as Facebook as “nodes” and the connections between users as “edges”.

Individual nodes possess a variety of attributes, such as gender, annual income, favorite music, field of interest, etc. These pieces of information are considered “signals” on each node. Nodes are connected primarily by “friendships”, but nodes with common “annual income level”, “favorite music”, “fields of interest”, etc., can also be linked. Edges can be weighted according to the degree of commonality or similarity of attributes. When a company wants to analyze a social networking site for the purpose of marketing, it can efficiently and effectively grasp market trends and customer attributes by appropriately designing the structure of nodes, edges, and edge weights according to a well-defined objective.

This is the same for a network of weather sensors. By arranging weather sensors in a mesh in a certain region, weather changes in that region can be understood by networking adjacent sensors with each other, collecting data such as temperature, humidity, and rainfall amount, and weighting the edges according to measured values and distance between sensors. Efficiently analyzing data by creating a graph made up of nodes and edges in this manner and examining the relationships between generated signals is the basis of graph signal processing.

Cheung: Applying the technique of graph signal processing to image processing is the topic of our joint research. In conventional discrete signal processing, audio is divided into equal time intervals, images are sampled at equal spatial intervals, and video is additionally divided into frames of equal time intervals. This processing is based on, so to speak, a uniformly organized “regular” data structure. In graph signal processing, on the other hand, signals having an “irregular” data structure can be used as the object. By freely creating appropriate graphs as necessary and interpreting the relationships between signal samples, new techniques not imaginable before can be developed.

Fig. 1 Example of depth image transform

Among pixels arranged in a 2D grid, those with high similarity to each other (in this case, those that are roughly the same distance from the camera) are given an edge weight of 1, and those that are very different (those far from the camera) are given an edge weight of 0. In so doing, some of the lines of the graph are disconnected, and the image can be split into two parts (images of the foreground and the background).

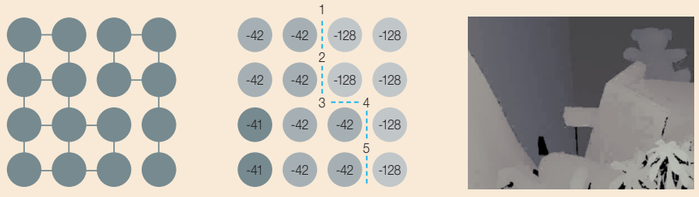
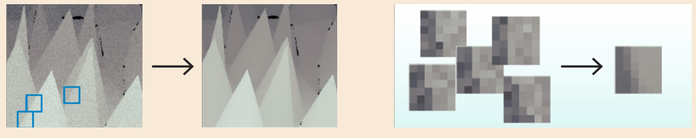


Fig. 2 Example of image denoising via graph signal processing

Noise is removed by averaging similar neighboring pixels. By designing a graph and giving it appropriate edge weights, noise can be removed to produce a clear image without blurring.



Improving indispensable immersive communication components: compression, interpolation, and denoising,

–– How is graph signal processing useful for research in immersive communication?

Cheung: Research in immersive communication aims to produce an interactive visual experience as close to reality as possible. For example, in video conferencing between remote locations, if the person facing the display screen looks over the shoulder of the person being captured, he can see the background behind that person. If we expand upon this capability, we expect that, for example, sports video viewers will be able to freely “change seats” and watch from an angle they like.

To realize this, video must be shot by multiple cameras, and the required viewpoint images must be synthesized instantaneously while predicting changes in the viewer’s line of sight and point of view. Since the amount of data will be enormous, data compression is essential to efficiently transmit and process it. Interpolation is also crucial to compensate for missing image pixels due to low resolution or disocclusion. Denoising noise-corrupted images is necessary as well. Graph signal processing is anticipated to be effective in all three areas of compression, interpolation, and denoising.

For example, to separate an object close to the camera and the background of an image, a graph is created by assigning large weights to the edges between nodes (pixels) having high similarity and assigning small weights to the edges between nodes with the greatest differences. When a graph Fourier transform is computed, the foreground and background can be efficiently separated according to the plus/minus sign of the lowest-frequency alternating-current (AC) component.

Applying this method to image compression, if we define a Fourier transform using a graph representing the structure of the image described above for one block, a compact representation of only low-frequency components can be obtained. Due to this sparse representation, the compression ratio can be increased. When we tried it with depth images used in video synthesis to produce realistic video, we demonstrated a high compression improvement of 30 to 40% over state-of-the-art image codecs.

Even if data from part of the image have been dropped, the dropped data can be interpolated by creating a low-frequency signal consistent with a derived graph structure, resulting in a smooth image. Denoising is performed by inserting a weighted average of similar portions in an image, where weights are pre-assigned according to derived image structure, resulting in removal of unwanted noise without blurring object boundaries.

The future of “graph signal processing” in images and video

–– What are your hopes for the future?

Ortega: The application of graph signal processing to images and video is at the research stage, but companies are paying close attention, particularly with regard to compression. Google and Microsoft are both involved in research on this topic. Enterprises will probably begin using their proprietary technology in 3 to 4 years, and may then face the prospect of standardization. Other applications are also emerging one after the other right now, and I anticipate that new applications will be discovered in the future as well.

Cheung: It is interesting that graph signal processing will bring about solutions to age-old problems surrounding image and video processing from a whole new viewpoint. By pushing ahead with research in close cooperation with Professor Ortega, I believe that we can take advantage of graph signal processing as an immensely powerful tool that can be applied to all of the issues involved in immersive communication, such as compression, interpolation, etc.

(Interview/Report by Masahiro Doi)

What new value does the collaboration between video and image analysis and natural language processing bring?

The horizon of semantic analysis seen through the harmony of video and language

That's **Collaboration**

Natural language processing researchers who have been analyzing text are now entering the field of video and image analysis. What are the commonalities and differences between these two types of researchers, who up to now have taken different approaches, and what direction should they head in? We spoke with Tohoku University Professor Kentaro Inui and National Institute of Informatics (NII) Associate Professor Yusuke Miyao, who are both engaged in research on natural language processing, and Professor Shin'ichi Satoh, who conducts research on video and image analysis at NII, and asked them to discuss the past, present, and future of semantic analysis from the two viewpoints.

Difficulties of machine analysis of videos and images

Inui: First, could you describe the evolution of research on video and image analysis?

Satoh: Ever since computers became available, there has been a huge demand for semantic analysis of video and images. Researchers were already studying it in the 1960s. At that time, they thought that since humans could do it easily, machines would be able to as well, and that analysis would become possible by programming human visual capabilities and defining rules for images. They quickly found out, however, that video and image analysis is enormously difficult. Even the simplest image could not be semantically analyzed without defining a dizzying number of rules. Because of this, there was a situation in the mid-80s where a large number of researchers once pulled out of the field.

Miyao: What was the turning point after that?

Satoh: A change happened in the 1990s, and an example of the results of that change was facial recognition. Researchers abandoned the conventional method of training a computer on where the eyes are and where the mouth is, and instead collected a large amount of facial learning data. They had the computer memorize, at minimum, "This is a face," by machine learning. This big-data approach bore fruit, and the foundation of facial recognition functionality of digital cameras in the 2000s was laid. After that, it was thought that a computer could be made to recognize anything in the world as long as a huge amount of learning data was prepared, and a gigantic database called "ImageNet" was built in 2010, which as of now houses about 14 million images based on approximately 22,000 concepts. Additionally, through the emergence of "deep learning" in 2012, video and image analysis has been qualitatively improved. However, one issue in analysis is how to select the "concepts" that serve as its basis.

Inui: I see. Can you give us more details?

Satoh: In performing image analysis, it was thought to be sufficient to assign relationships between images and symbols (i.e., concepts), but it is difficult to choose the concepts. If the number of concepts exceeds 10,000, a parent–child relationship emerges between concepts. Under the concept "vehicle", there are "car" and "airplane", and that type of definition must also be accurately set. Additionally, we must correctly define whether the association of meaning and the association of appearance coincide, such as "an image of an eagle flying" and "an image of a jet flying". We are at a stage where research for preparing the definitions of this diversity of concepts and improving the precision of image recognition is progressing. On the other hand, computers are still most clumsy when given an unknown image and asked, "What is this?" For example, it is difficult for a computer to look at an image alone and answer whether it is a dog or a cat, but if given the premise that it is a dog and then asked, "What breed of dog is this?" it can answer correctly with fairly high accuracy. It is a matter of trial and error to find the path to a solution.

Shin'ichi Satoh
Professor and Director Digital Content and Media Sciences Research Division, National Institute of Informatics (NII)

Kentaro Inui
Visiting Professor, National Institute of Informatics
Professor, Graduate School of Information Sciences, Tohoku University

Inui: Is it difficult even through the use of deep learning?

Satoh: There's still a ways to go. With certain data sets, it is now possible to analyze images with fairly good accuracy. However, when we try to dissect which type of learning was responsible for this recognition capability, for example, we know that the computer did not successfully recognize a "house" due to the shape itself but by looking at the surrounding shrubbery. This is because there is an infinite variety of shapes of houses, and from the computer's point of view, a highly accurate answer is more likely if it looks at the shrubbery. In short, we have not reached the point of analysis in the true sense even though the results look good.

Natural language processing has traced the same history as video and image analysis

Miyao: Although it's a very simple idea from the natural language side, are you taking the approach of using language models and contextual information?

Satoh: Use of contextual information can be cited as one method, but that itself has to be systematically created. As a result, it is no longer an approach used in research on image analysis. Researchers are undeniably aware that the "holy grail" relentlessly sought by researchers in image analysis is to provide only images as learning data and obtain accurate analysis results from unknown images.

Yusuke Miyao
Associate Professor, Digital Content and Media Sciences Research Division, National Institute of Informatics (NII)
Associate Professor, Department of Informatics, School of Multidisciplinary Sciences, the Graduate University for Advanced Studies

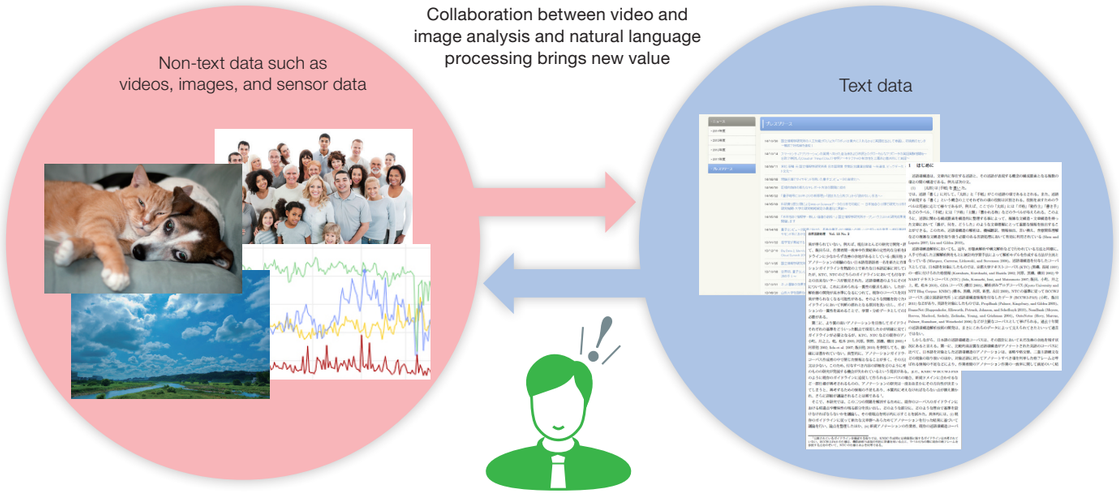
Miyao: That's surprising. One would think that, from the natural language side, it would be better to employ any data that can be used.

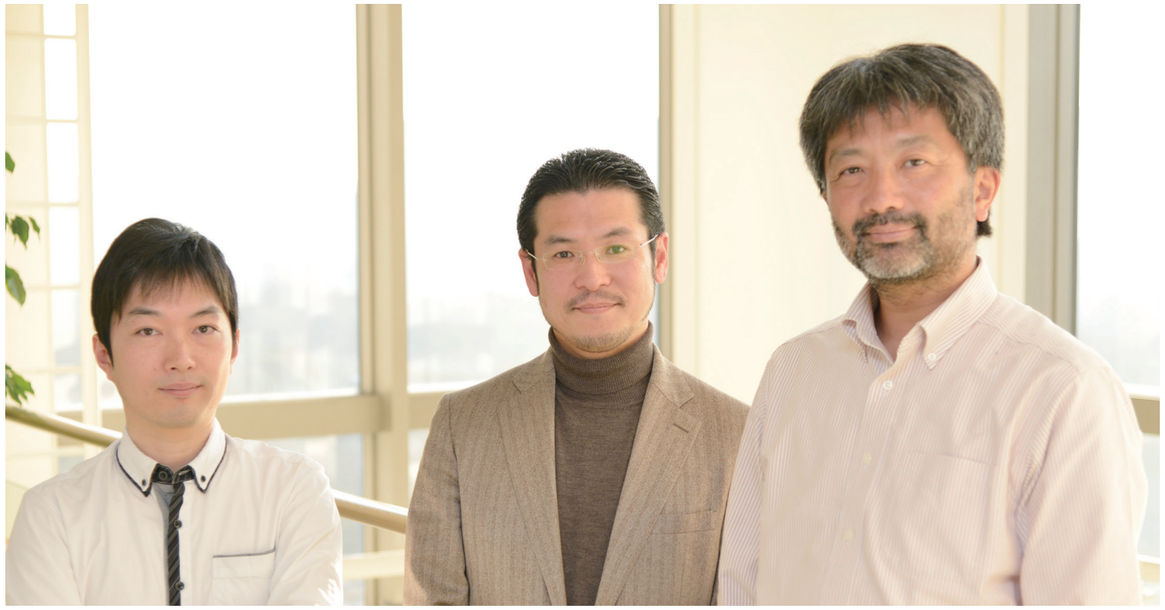
Inui: Natural language processing is also tracing the same history as video and image analysis. To use the analogy of interpreting cyphers using computers during the Second World War, at the dawning of the machine translation era, there was the optimistic view that translation would also be easy using computers. In reality, however, it was found to be extremely difficult. Subsequently, just like video analysis, syntax analysis and shallow semantic analysis models were created by using large-scale language data and by feeding data accurately described by humans to computers for learning in the early 1990s, and this was very successful. In recently years, the amount of text data has increased phenomenally due to the advent of the web and social networking sites, and there is a trend toward extracting a variety of language knowledge and world knowledge from this plethora of data. If there are a huge number of words and phrases, by combining them it may be possible to incorporate the "common sense" of humans into language analysis, and consequently for machines to be able to compensate for omissions in conversation in various situations.

Video and image analysis is also indispensable to the evolution of natural language analysis

Inui: In the meantime, there is something about which we are always chomping at the bit. Natural language processing tries to obtain knowledge only from the world of symbols, whereas the problem deliberated in the field of artificial intelligence (AI) is symbol grounding, namely, how to associate a symbol with its meaning in the real world. If we deal with only symbols, there is a possibility that true intelligence will not be realized. Human intelligence is amassed during interaction with a multitude of external relationships. For example, a human mother teaches her child, "There's an elephant," and the child recognizes that "That's an elephant." A wide range of intelligence is built through such interactions with the outside world. On the natural language processing side, there is also a constant debate about whether the essence of intelligence is in interaction with one's environment, with the pros and cons of dealing with symbols alone in the back of our minds. One of the critical elements in this interaction with the external environment is video and images. By handling them paired with language, it may be possible for a machine to relive the experiences of humans, and with time, this may lead to advancements like AI.

Miyao: I also feel that knowledge obtained from a large amount of text data and knowledge obtained from completely different media like images do not completely overlap, and may actually capture different things. In other words, I feel there may be knowledge that cannot be obtained from text alone. Right now we are entering the stage where computers are being made to perform semantic analysis using wide-ranging knowledge and common sense, and new findings will undoubtedly be obtained not only by language processing, but by combining it with video and image processing. Actually, there are also resources that can make use of each other. This has motivated my interest in video and image analysis.





Inui: From the natural language side, I think that video and image analysis could become one of the next mainstays of research. Just like Professor Miyao said, the parts that the two have in common will probably become research topics in the future.

Setting tasks with appropriate evaluation criteria is essential

Satoh: The problem of symbol grounding in assigning symbols to video and images has also been raised on our side as well. In video and image recognition, the objective is just to associate objects or situations in a video or image with symbols, and we thought it unnecessary to go as deeply as the problem of semantics, such as symbol grounding. But we now understand that the problem of semantics arises everywhere, such as ambiguity when creating learning data sets by hand, expanding a symbol lexicon to a practical scale, and assigning parent–child relationships between symbols. In the end, there are numerous situations where we will not be successful without delving deeply into the problem of semantics. On the other hand, we can say that natural language processing, although beginning with symbols, seems to head toward the same destination but from the opposite direction as video and image analysis side. Meanwhile, I think that if appropriate tasks are set, the two lines of research will converge to produce new results.

Inui: For task setting, for example, there was the problem of labeling. For the question "What is this a picture of?" it is relatively clear which is the correct interpretation because it is a matter of which label should be selected from among, for example, 10 categories, i.e., labels. On the other hand, if the problem is to generate a caption or summary, correct evaluation of video and image together with natural language has its difficulties. How should it be evaluated? It is difficult for research to proceed successfully unless evaluation criteria can be precisely established. It is also difficult to achieve results by simply exploring individual techniques, and researchers will not be motivated.

Satoh: As Professor Inui says, this topic is starting to emerge at the top conferences in video and image analysis. To advance to the next phase of research, appropriate evaluation criteria and common task settings must be established.

Miyao: On the other hand, there is an automated evaluation scale for machine translation called "BLEU", for example, which researchers in video and image analysis have recently started to use. But there is concern that their research will end up moving toward improving the accuracy rate based on this index. Of course, there is also the aspect that accuracy of machine translation is improving on account of this scale, but we must proceed with extreme caution.

Satoh: I agree. It is now becoming possible to recognize information provided from the external environment, including text, video, and images, and to analyze its meaning by computer. In the future, I expect that video and image search and automatic captioning will begin, and a variety of applications will emerge, such as monitoring and analysis by surveillance cameras. In the meantime, as previously mentioned, tasks must be set for breaking down individual problems that arise along the way and for reliably stepping up research. This is something that must be dealt with on both the video and image side and the natural language side. I think AI research will impact this as well some day in the future. I'm looking forward to it!

(Interview/Report by Hideki Itoh)

NII essay

The Historicity of Photographs

Asanobu Kitamoto
Associate Professor, Digital Content and Media Sciences Research Division, National Institute of Informatics

A photograph is a record of history. Because a photograph is a slice of the world today in printed form, from a future perspective it serves as a valuable record of the past. It may be only natural, but recently I have been thinking a little more deeply about the meaning of this.

A photograph is a record of a scene that will never return again. Something that made me keenly aware of this was the 1993 earthquake off the southwest coast of Hokkaido. I had traveled to Hokkaido's Okushiri Island two years before the huge earthquake and spent an enjoyable couple of days staying at an inn and gathering sea urchins. The television news coverage of the huge earthquake showed the village destroyed by a tsunami and engulfed in flames. The shock of realizing that I could never see that scenery again sparked my interest in an archive that would preserve it in photographs.

But the historicity of a photograph does not reside only in major events. Fragments of history recorded in photographs of our day-to-day lives also stir up memories of the past. There is a hugely popular account on Twitter called @HistoricalPics, with 2 million followers, but the only thing it does is tweet old photographs and their titles. The secret to its popularity is probably the fact that a photograph posted by chance vividly resurrects memories that have been totally forgotten, and from there a story is born.

However, the historicity of a photograph is, in fact, something deeper. I noticed this while developing the mobile app Memory Hunting. This application aids in taking a photograph having the same composition as an old photograph by overlaying the old photograph semi-transparently on the viewfinder of a camera. "Before" and "after" images having the same composition are useful as a medium to record changes from disaster to recovery.

But when you try it, taking a photograph having the same composition is fairly difficult. It is not enough for the photographer to simply stand in the same place; in fact, even the photographer's posture must be identical. To take a photograph with the same composition as a photograph taken while the photographer was squatting, you must squat as well. I noticed this by chance. Taking a photograph having the same composition is actually an experience of traveling back in time and laying your body over the position you think the photographer took in the past. If you have a photograph taken by a person who is now deceased, lay your own body in the same place. It will probably be an experience that stirs up poignant emotions.

In short, a photograph is also a medium that records historicity as the trace of a photographer's body. Even the photographs that everyone casually takes may become valuable to someone in the future. In thinking about the historicity of photographs, consider what kind of record you are leaving of yourself living through history.

Weaving Information into Knowledge
NII

National Institute of Informatics News [NII Today]

No. 53 Apr. 2015 [This English language edition NII Today corresponds to No. 67 of the Japanese edition.]
Published by National Institute of Informatics, Research Organization of Information and Systems
Address: National Center of Sciences 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430
Publisher: Masaru Kitsuregawa Editorial Supervisor: Ichiro Satoh
Cover illustration: Toshiya Shirotani Photography: Seiya Kawamoto, Yusuke Sato
Copy Editor: Madoka Tainaka Production: Nobudget Inc.
Contact: Publicity Team, Planning Division, General Affairs Department
TEL: +81-3-4212-2164 FAX: +81-3-4212-2150 E-mail: kouhou@nii.ac.jp

http://www.nii.ac.jp/about/publication/today/