Digital Media

Hiroshi Fuju†

Abstract

This article presents some examples of video content delivery and two-way video communication services and describes current trends in the digital media technology supporting these services. It also describes NTT's research efforts in subjects such as video coding and image processing and introduces our vision of business services that combine high-quality secure video delivery with high added-value processing.

1. The rising quality of video media

1.1 Trends in digital media

The overall global population of broadband users topped 100 million for the first time in 2003 (source: ITU, 9/7/04), and the number of fiber-to-the-home (FTTH) subscribers had reached 1,689,000 by the end of September 2004 based on an increase of nearly 150% over the previous six-month period (Multimedia Research Institute, 11/4/04). NTT has therefore set its sights on expanding the number of optical access subscribers to 30 million by 2010 (NTT, 11/10/04).

In the field of broadcasting, terrestrial digital broadcasting is expected to be accessible in 37 million Japanese households (i.e., 79% of households) by 2006 (Ministry of Internal Affairs and Communications, 12/1/04). Similarly, in the field of domestic appliances, 2.17 million DVD recorders were shipped in 2003 (an increase of 289% on the previous year), resulting in total recorded sales of ¥160 billion (255% up on the previous year) (Multimedia Research Institute, 6/10/04). A questionnaire aimed at consumers revealed that the three digital domestic appliances most likely to appear at the top of their wish lists are now DVD recorders (38.3%), LCD or plasma TVs (37.7%), and digital cameras (25.9%) (Multimedia Research Institute, 6/10/04) (LCD: liquid crystal display, plasma: plasma display panel (PDP)).

On the other hand, with regard to changes in the leading broadband products, ADSL (asymmetric digital subscriber line) peaked in 2004 and has started to decline slightly, while the number of FTTH contracts is continuing to increase and is expected to exceed 14 million during 2008. There is also a significant trend towards higher quality in the market for color TVs. CRTs (cathode ray tube sets) accounted for over 90% of the 9.6 million sets sold in 2001, but this market share is continuing to decline. In 2009, it is expected that CRT sales will have fallen to 1.5 million while sales of LCD and plasma TVs will have increased to about 8.3 and 1.2 million, respectively.

1.2 2005 International Consumer Electronics Show

The leading products showcased at this year's Consumer Electronics Show (the largest to be held since the event was first held in 1967) included products such as DVD equipment, PVRs (personal video recorders), digital cameras, cellular phones, flat-panel displays, organic electroluminescent displays, naked-eye (spectacle-free) 3D displays, Blu-ray products, HD (high definition) DVD equipment, home servers, VoIP (voice over Internet protocol) products, portable media players (MPEG-4/AVC players), image printing equipment, devices using

† NTT Cyber Space Laboratories
Yokosuka-shi, 239-0847 Japan
E-mail: fuju.hiroshi@lab.ntt.co.jp

miniature hard disk drives (HDDs), H.264/MPEG-4 AVC equipment (see section 3.3), and hi-tech automobiles. Four products attracted particular interest:

- IPTV (Internet protocol TV)
- PLC (high-speed power line communication)
- UWB (ultrawideband communications)
- HDTV (high-definition television) flat-panel displays

1.3 Efforts at enhancing media quality

Under these circumstances, there is clearly a strong demand for higher quality in video-related services, such as:

- Improved quality of videophones and video conferencing
- Introduction of HDTV in terrestrial digital broadcasting and consumer equipment
- Video transmission services based on super-high-resolution (SHR) displays

Each of these services involves transferring enormous amounts of high-quality content across networks that have limited bandwidth, so it is important to use the latest video compression techniques such as H.264 instead of MPEG-2.

As an illustration of the sort of picture quality involved, a standard TV picture currently consists of 480 interlaced horizontal scanning lines, whereas an HDTV picture has 1080 and SHR has at least four times the resolution of HDTV. By enhancing the techniques used to compress, display, and process high-quality images and video signals, it should become possible to make services on broadband networks such as video delivery. The immediate tasks for achieving this include the following:

- Developing scalable video coding techniques and ultrahigh compression control techniques for H.264 video coding at the core of next-generation video compression standards
- Improving the ability of compressed video to tolerate electronic watermarking and improving the quality of mobile electronic watermarking leading to the creation of new markets
- Developing video processing techniques that can provide added value to video services such as applying human profiling and tracking the movements of humans in multi-camera environments.

2. Impact of video coding technology

2.1 Video coding environments in 5–10 years

How will video coding technology change in the future? A few clues are shown below.

- Evolution of devices

According to Moore's law (which states that semiconductor integration density doubles every 18 months), the processing speed of personal computers (PCs) will have increased tenfold after five years. That corresponds to an increase from today's 2-GHz CPUs to 20 GHz. The capacity of recording equipment is also expected to increase tenfold.

- Evolution of networks

Meanwhile, Gilder's law states that network bandwidth doubles every six months. Accordingly, network bandwidth will have increased a thousandfold after 5 years, so for example the current 100-kbit/s networks should have reached a speed of 100 Mbit/s.

- Changing user requirements (sophistication level)

As devices evolve, the quality of the content increases and viewers develop more sophisticated tastes. The current level of 1-Mbit/s video will evolve in various ways—towards multi-viewpoint (interactive) TV and more natural images, from two-dimensional (2D) to three-dimensional (3D) and from RGB (red/green/blue) to multispectral images, and towards larger screen formats—resulting in larger quantities of data. For example, a transition from 2D to 3D would require 500 times as much data as for a 500 × 500 video image.

Assuming that current PCs can perform standard TV compression processing, then by 2009, the computational load will be almost 10 times that for processing standard TV in 2004 (**Fig. 1**). However, the ultimate dream of 3D multispectral SHR TV corresponds to 100,000 times as much processing, so it is quite evident that human desires remain several orders of magnitude off the scale.

2.2 Future video applications targeted by video coding technology

By developing technology that embodies the video-related demands listed below, it should be possible to create environments where SHR technology can be used to achieve super-real communication in a completely natural way. It will include:

(1) Being able to use video anywhere in the human environment ("video anywhere")

Environments where any terminal can access video from any network

(2) Changing video from something that is "shown" to something that is "experienced" ("supernatural")

Depicting objects so that they appear more realistic than the real thing, without the viewer being aware of the network or any delay

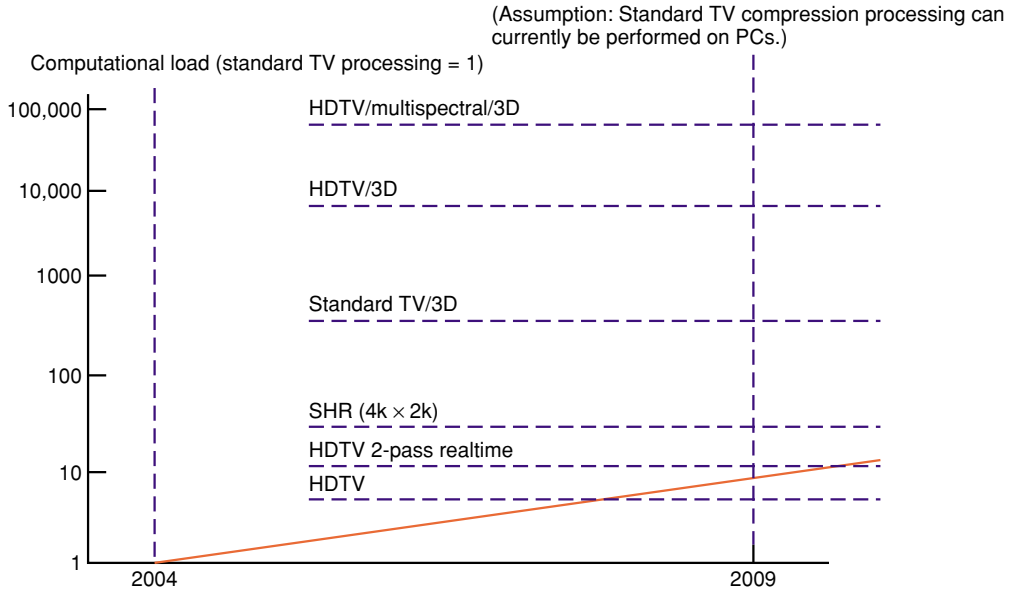

Fig. 1. Outlook for video coding environments after 5–10 years.

(3) Changing video from "moving pictures" to "movable pictures" ("tangible video")

Interactive environments that can be viewed from any angle, instead of just from a given angle

Human expectations of improvements in video technology are increasing faster than the speed of networks and processing devices, so it is only a matter of time before we arrive at an information or processing load explosion. Therefore, we need coding techniques that can provide stress-free natural communication with greater emotional satisfaction. Various techniques based on next-generation super-efficient coding techniques are described below.

2.3 Video anywhere: environment-adaptive coding

In the human environment, techniques such as the following are essential for allowing video to be transmitted or played back anywhere and at any time (**Fig. 2**):

- Coding techniques in which parameters such as the bit rate and image size can be changed flexibly according to various network/terminal conditions (scalable coding and intelligent error recovery)
- Encoding techniques that use self-organized optimal compression algorithms, and syntax-free coding for transmitting streams together with prerecorded programs (optimal compression/playback performed automatically without an encoding/decoding program)

Scalable coding is a coding scheme in which multiple bit streams with different bit rates can be extracted from a single bit stream. In this scheme, the data is divided into essential data that forms the picture (base layer) and optional data that provides greater quality (enhancement layers). The essential data must all be transmitted, but the optional data can be selectively cut or included to assemble and display video of various quality levels to suit the available communication bandwidth, terminal performance, and viewing environment. The basic procedure is as follows: the base and enhancement layers are first sent to a scalable server. Then, this server can choose how much data to send to each destination terminal. This allows various levels of image quality (e.g., high-, medium-, and basic-quality pictures) to be provided from a single source. Moreover, the quality level can be changed dynamically according to network congestion conditions, and playback can be continued without interruption at a suitable quality level on a different terminal when the user moves.

2.4 Supernatural: video coding that transcends reality

We are studying super-high-quality video coding from the following four aspects:

- Spatial resolution: How are things improved by moving from the standard TV resolution of 720 × 480 pixels to 8000 × 4000 pixels for SHR?
- Pixel depth: How about increasing the amount of data per pixel from the current value of 8 bits to

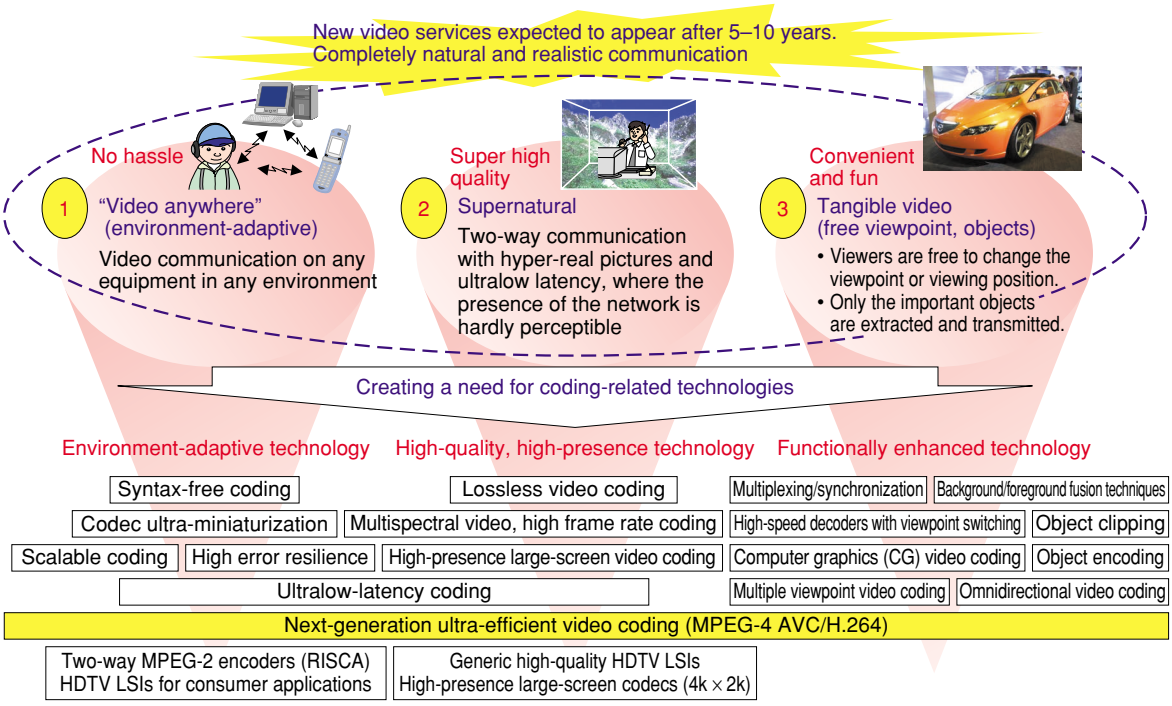

Fig. 2. Future of image-based communication and the required encoding techniques.

12–16 bits?

- Temporal resolution: Or how about increasing the frame rate from the current value of 30 frames per second (fps) to something like 60–300 fps?
- Chromatic fidelity: Providing multispectral hues by using more than three primary colors.

With regard to temporal resolution, the physiological limit of humans is thought to be about 150–200 fps, so our research is mainly concerned with frame rates of about 150 fps. Also, with regard to chromatic fidelity, we are participating in joint studies as part of a group called the Natural Vision Project.

We are also considering looking into "Super resolution" coding to create images that look even better than the original. This involves interpolation/processing techniques to generate an image of higher resolution than the originally captured image, that is intelligent decoding that can create "something" out of "nothing". For example, this sort of decoding could be used in applications such as:

- Signal restoration → semantic video restoration
- Computer graphics → natural image
- Black and white images → color restoration
- 2D images → 3D images
- Line-drawn animation → restoration of original image

2.5 Tangible video: 3D video coding

This is a service that breaks away from the notion that video is just a 2D medium. Specifically, it is a TV service that viewers can appreciate from any angle based on video obtained with multiple cameras and video content with 3D information including depth cues obtained from multiple cameras. This technology provides for greater audience participation and is regarded as an eye-catching type of next-generation video service.

3. Developed products and application examples

3.1 Flow of video coding techniques and their implementation in products

The work conducted by NTT Group into video coding techniques has a history of over 20 years and we have already developed numerous products. **Figure 3** shows the flow of video coding techniques and product developments.

"WarpVision" is a high-quality two-way video communication service based on MPEG-2, which provides an environment with the same video quality as ordinary TV by only software.

- Picture size: VGA (640 × 480)
- Frame rate: 30 fps
- Latency: 200 ms or less

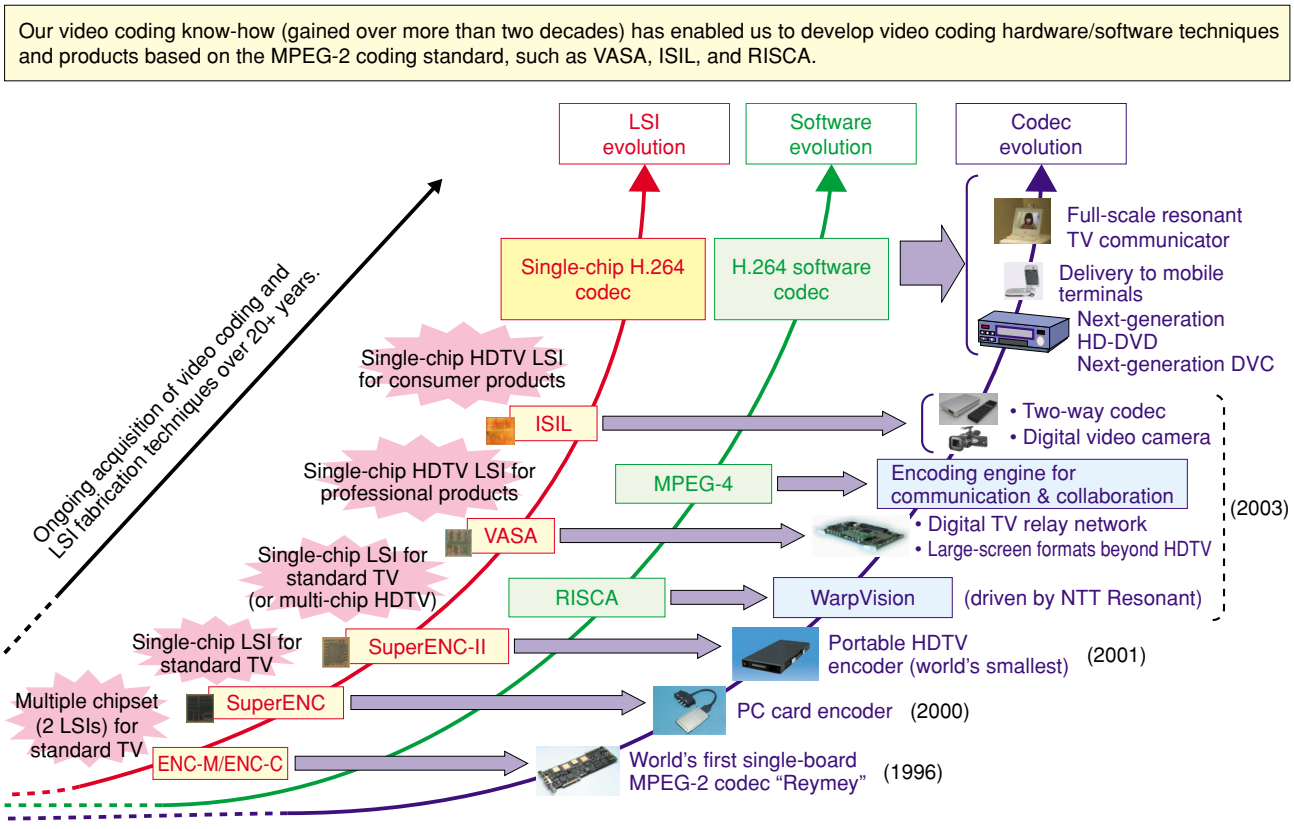

Fig. 3. Evolution of video coding technology and products.

• Transfer rate: Up to 10 Mbit/s

Originally, two LSI chips were needed to handle standard TV signals, but with current "VASA" technology we have made it possible to encode and decode HDTV signals in a single chip. The VASA MPEG-2 HDTV codec chip (or LSI) is the world's first professional device of this type. It was awarded the Prime Minister's Prize at the 33rd Japan Industrial Technology Grand Prix. VASA let us downsize HDTV encoders and reduce the size of the equipment from roughly 10U (1U = 1 × 19 inches) in 1995 when LSIs were unavailable to just 1U in 2001 using the SuperENC MPEG-2 video encoder LSI. Then in 2003 we used the VASA chip to miniaturize the HDTV encoder to the size of a postcard. In the future we hope to produce a chip-sized encoder.

The single-chip MPEG-2 HDTV codec LSI, "ISIL", is the world's first LSI for use in consumer equipment. It is a tiny chip (about 20 mm across) with low power consumption. This chip has already been incorporated into consumer products such as HDTV digital video cameras. At present, we are using this ISIL chip to develop a stand-alone codec (ISIL-BOX) for two-way video communication that works without a PC. By connecting a domestic camera and TV (monitor) to the ISIL-BOX and hooking it up to an optical broadband network, you can perform two-way communication without the need for a PC environment. This system provides standard TV quality with a latency of 200 ms and implements a natural communication environment in which the user's own image is displayed together with the image of the called party.

3.2 SHR codec

We are also working on the provision of rich video services that make full use of broadband networks by employing an SHR video codec. For example, we are working on an SHR video system where video from a 4K × 2K camera is compressed and transmitted by an SHR codec using multiple VASA chips and then displayed using an SHR projector. This type of system should make it possible to provide SHR live services with better resolution than HDTV so that audiences can feel as if they are actually watching sports events such as soccer, rugby, or baseball matches, or

indoor events such as music concerts.

An SHR codec has a resolution of about 4000 × 2000 pixels, which is equivalent to the size of digital cinema pictures. The uncompressed data rate of 4 Gbit/s decreases to 80–160 Mbit/s after MPEG compression. The size of the SHR codecs used to relay events such as the 2002 soccer World Cup (before the development of VASA) was 18U, but we have now reduced the codec size to 3U.

In the compression of large-format images, the screen is divided into quarters and processed by four separate VASA chips. The number of bits for each quarter is allocated adaptively, which means that instead of generating the same quantity of compressed data from each quarter, more bits are allocated to parts where there is more data to encode.

The compressed data rate of 80–160 Mbit/s is a target rate chosen to promote the popularity of SHR video from the viewpoint of network bandwidth and data storage capacity. That means it can be applied to high-speed network services provided by NTT, such as MegaLive and Metro Ether, and allows SHR video to be recorded on PC-based streaming recorders (costing less than $1000). This should make SHR video accessible to the general public.

3.3 Next-generation video coding technology: H.264/MPEG-4 AVC

Products currently under development include a single-chip H.264 codec LSI and software implementations of a scalable codec and an H.264 codec. H.264 is a technique for digitally coding moving images. It has been jointly standardized by the ISO MPEG and ITU-T (International Telecommunication Union Telecommunication Standardization Sector). It is often referred to as “AVC/H.264” or “H.264/AVC” to reflect this dual background. AVC is short for MPEG-4 Part 10: Advanced Video Coding (an extension to the MPEG-4 part 2) and H.264 is the name of an ITU-T standard (**Fig. 4**). Recently, in the field of terrestrial digital TV broadcasting, the demand for H.264 has been growing due to the adoption of this format for broadcasts called “single segment broadcasts” aimed at mobile terminals.

Technically, H.264 offers 2–4 times the coding efficiency of MPEG-2, but achieving this level of performance involves a much greater computational load, about ten times or more that of MPEG-2 (sometimes as much as 100 times). Compared with earlier techniques, the target compression rates for standard TV pictures are as follows:

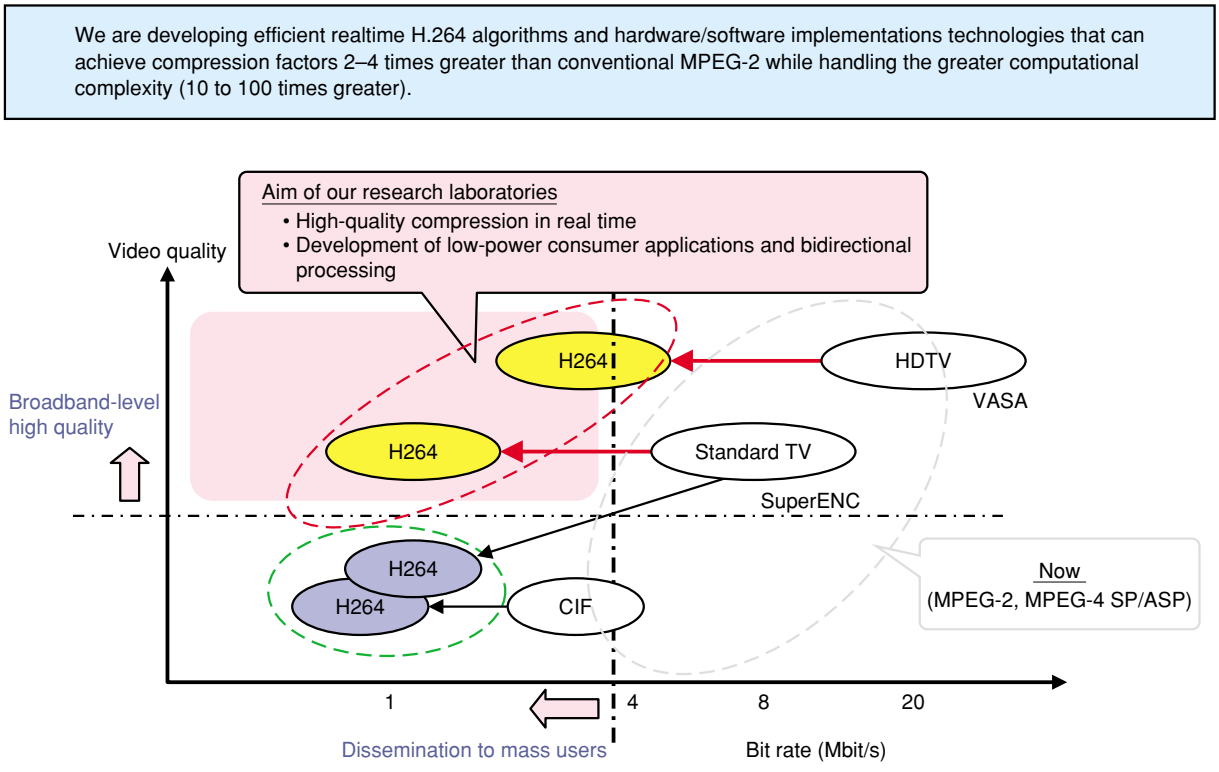

Fig. 4. MPEG-4 AVC/H.264 ultrahigh-compression coding.

MPEG-2: 4–6 Mbit/s
MPEG-4 ASP: 3–4 Mbit/s
H.264/MPEG-4 AVC: ≤2 Mbit/s

For HDTV pictures, the target rates are as follows:

MPEG-2: 15–30 Mbit/s
MPEG-4 ASP: 10–20 Mbit/s
H.264/MPEG-4 AVC: 8 Mbit/s

We are therefore currently engaged in the research and development of algorithms for compressing large volumes of data in real time while achieving super-high compression factors and of hardware and software to implement these algorithms.

3.4 Business applications

These video coding techniques can be used to create new business models by promoting a culture of video communication (**Fig. 5**). Potential applications include the following:

- Remote monitoring services
- Remote classroom services
- Medical and welfare services
- Consulting and counseling services
- Corporate communication services

We hope to provide high-quality realtime video communication services in these fields.

Application example 1: Keeping an eye on your children at the kindergarten

This is a system that allows parents to monitor their children at a kindergarten. Video images of the children are sent from a scalable server and received by parents at various locations with a bandwidth suited to the performance of each parent's terminal. So for example a home-based large-screen display could reproduce sounds and facial expressions with great clarity, while a mobile notebook PC could let parents away from home check up on what the children are doing and saying. Alternatively, the content could be delivered to a mobile phone with a small screen, which should at least let parents confirm that the children are there and listen in on what they are saying.

Application example 2: Monitoring from remote locations

This is a system that allows you to keep an eye on your house while you are out. The house is monitored

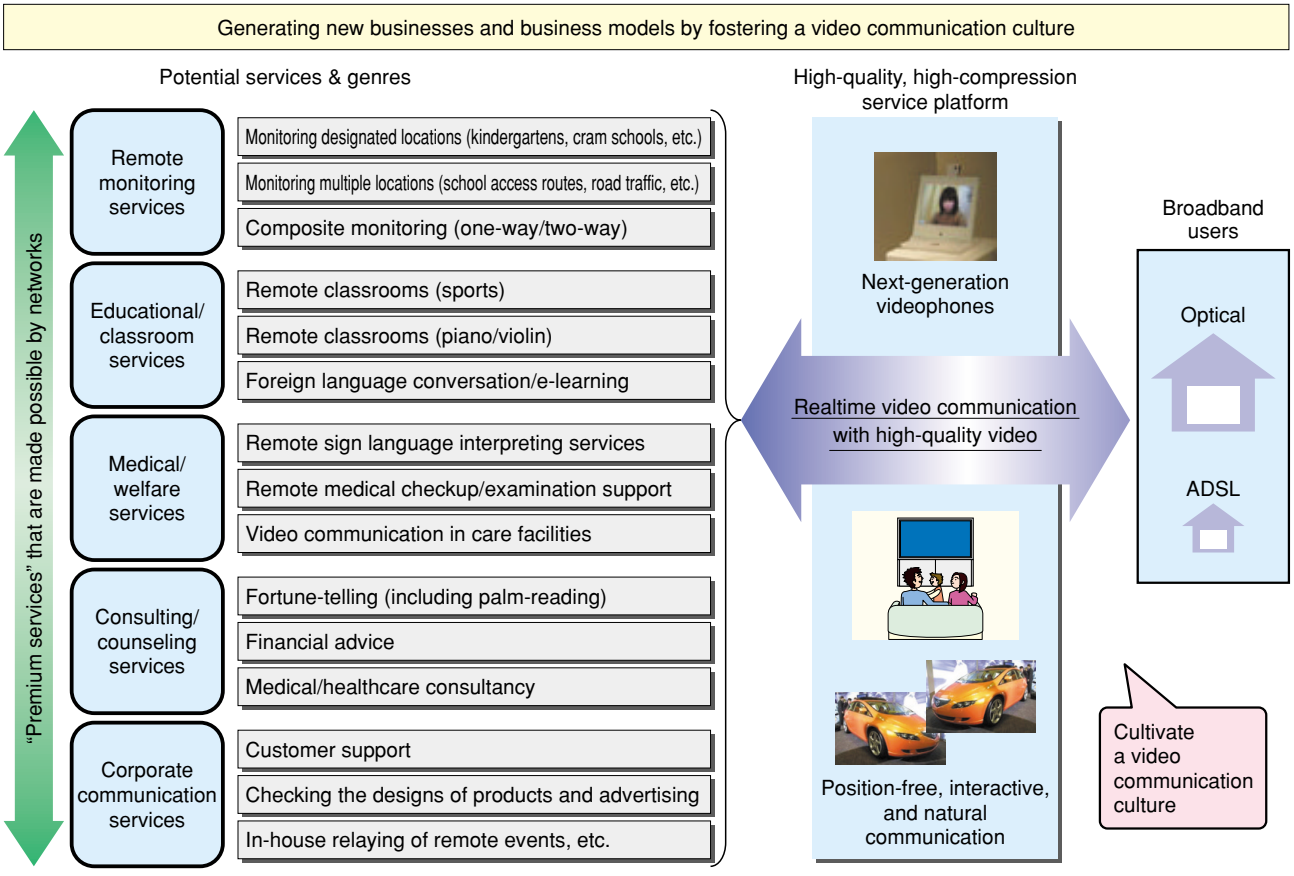

Fig. 5. Towards the incubation of video communication.

with cameras, and when changes occur in the images, they are reported to a terminal near you such as a cellular phone. It is also possible to process these images to, for example, identify an intruder's face and perform detailed checks.

Application example 3: Selecting which content to watch or listen to

In this system, video content with a selectable viewpoint is produced using video sources obtained from multiple cameras, and the viewer is left free to choose which video to watch and at what time. For example, a music lesson could be captured with multiple cameras to produce a single video that individual users can use to learn about different aspects of particular interest, such as the movement of a pianist's fingers across the keyboard, a guitarist's fingering technique, a violinist's posture, or the harmony produced by a group of players.

Application example 4: Translating text within a scene

This is a translation service that uses the camera functions of PDAs (personal digital assistants) and cellular phones. Foreigners visiting Japan can use it to take pictures of notices at train stations, restaurant menus, and the like and transmit the images to the service provider where they are analyzed. Then the meaning of Japanese words in the images can be conveyed either by translating them into words in the visitor's native language or by displaying pictures of the corresponding objects. We are currently concentrating on a system for translating Japanese into other languages, but we are also considering working on services for translating other languages back into Japanese due to the growing overseas availability of environments compatible with NTT Docomo's FOMA i-mode cellular phones.

Profile

■ Career highlights

Executive Research Engineer, Visual Media Communications Project, NTT Cyber Space Laboratories.

He received the B.E. degree in electronic engineering from Gunma University, Kiryu, Gunma in 1978. In 1978, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Public Corporation (now NTT), Tokyo, Japan. He is a member of the Virtual Reality Society of Japan.