MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Surveillance System with Mega-Pixel Scalable Transcoder

Toshihiko Hata, Naoki Kuwahara, Derek Schwenke, Anthony Vetro

TR2007-008 January 2007

Abstract

This paper presents a video surveillance system that displays mega-pixel streams effectively, while traansmitting and processing the streams efficiently with limited resources such as bandwidth, computing power and display resolution. The proposed system stores high-resolution and high-quality video data and associated object metadata, which includes ROI (Region-of-Interest)information. To satisfy such resource constraints and display important parts in detail without missing the overall scene context, the stored images are efficiently transcoded in the compressed-domain based on the ROI information, display resolution and available bandwidth. Simulation results confirm the effectiveness of the proposed system in terms of objective measures and subjective evaluation.

SPIE Conference on Visual Communications and Image Processing (VCIP)

Surveillance System with Mega-Pixel Scalable Transcoder

Toshihiko Hata *[a], Naoki Kuwahara [a], Derek L. Schwenke [b], Anthony Vetro [b]
[a] Mitsubishi Electric Corporation, Amagasaki, Hyogo 661-8661, Japan;
[b] Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA 02139, USA

ABSTRACT

This paper presents a video surveillance system that displays mega-pixel streams effectively, while transmitting and processing the streams efficiently with limited resources such as bandwidth, computing power and display resolution. The proposed system stores high-resolution and high-quality video data and associated object metadata, which includes ROI (Region-of-Interest) information. To satisfy such resource constraints and display important parts in detail without missing the overall scene context, the stored images are efficiently transcoded in the compressed-domain based on the ROI information, display resolution and available bandwidth. Simulation results confirm the effectiveness of the proposed system in terms of objective measures and subjective evaluation.

Keywords: Surveillance, Mega-Pixel, Object-Based, Scalable, Transcoding, Region-of-Interest, JPEG2000, Streaming

1. INTRODUCTION

Recently, surveillance systems are beginning to employ mega-pixel cameras to identify the faces of people and license plates of vehicles. These mega-pixel cameras are expected to come into wide use in near future due to technological advances, such as mega-pixel sensor devices with wide dynamic range and low cost, as well as increasing market demands described above. But, there are some problems caused by increasing the bit rate of mega-pixel images, e.g., the bit rate of compressed video streams with SXGA is several times that of VGA. Multiple mega-pixel streams can be transmitted over wide bandwidth networks such as Gigabit Ethernet, but it is impossible to transmit them via narrow bandwidth or bandwidth-varying networks such as a wireless LAN and ADSL. Further, the overall images with mega-pixel cannot be displayed on low resolution displays and more computing power is required to decode and display them.

In our prior work, we have developed an object-aware video surveillance system based on JPEG 2000 [1] that is not only smart and friendly for users, but allows for transmission of the scene over limited bandwidth networks [2]. In this system, an image sequence is encoded and stored as a JPEG 2000 bitstream, and then the stored images are efficiently transcoded in the compressed-domain using a low-complexity adaptation technique. In one particular streaming mode, the ROI (Region-of-Interest) are transcoded with higher quality than the background to satisfy network constraints.

This paper considers some extensions of the system to process mega-pixel video streams efficiently and display them effectively. The extended system supports new display methods appropriate for mega-pixel surveillance video. Specifically, we include a mode for displaying ROIs in high-resolution and high-quality and an overall view of the scene with low-resolution and low-quality simultaneously. In this way, it is possible to observe each object in detail without missing the overall scene context. We also propose mega-pixel video streaming with JPEG2000 transcoder that uses a combination of three types of scalability: quality, resolution and position, to realize such effective display modes.

The rest of this paper is organized as follows. In the next section, we provide a brief overview of our system. Mega-pixel video streaming and JPEG2000 scalable transcoding are described in section 3 and 4 respectively. In section 5, experimental results are provided and concluding remarks are given in section 6.

2. SYSTEM OVERVIEW

Figure 1 shows a system overview. The main functions are outlined below.

- *Video encoding:* Encode mega-pixel images into JPEG2000 in network cameras and smart cameras.
- *Object extraction and metadata creation:* Extract objects over successive frames and create metadata such as their existing regions and movements in smart cameras and image processing servers. Our system is independent of any particular method for this function. We use an object tracking algorithm [4] and a face detection algorithm [5] to create the metadata for evaluation in section 5.

- *Data storage:* Link the mega-pixel video data and metadata together and record them onto hard disk drives. They are always recorded in an endless style and read out according to requests. The stored video has high spatial quality: image quality (SNR) and resolution, and high temporal quality: frame rate, independent of the object metadata since important scenes are required with high quality for detailed analysis and evidence.
- *Scalable transcoding:* Transcode the stored images efficiently in the compressed-domain based on the metadata according to object importance, user preference and available resources. The transcoder uses combination of three types of JPEG2000 scalability: quality, resolution and position to output video streams with limited bit rate.
- *Transmission and display:* Transmit a transcoded video stream with dynamic message exchanges between a client and a server. The received data is decoded and displayed on a screen. The decoded images are synthesized in some display styles.

3. MEGA-PIXEL VIDEO STREAMING

The extended system provides the following methods described in [2] for mega-pixel images or images down-sampled by the transcoder.

- *Frame by frame streaming (FFS)* gives higher quality to ROIs than a background in a frame-by-frame manner.
- *Successive ROI with occasional background streaming (RBS)* gives higher frame rate to ROIs than a background.
- *Mosaic streaming (MS)* superimposes successive ROI images on a background in a mosaic style.

We propose new streaming methods appropriate for mega-pixel surveillance video.

Multiple camera display with digital PTZ (MCPTZ)

Recently, surveillance cameras with more than SXGA (1280x960 pixels) are available in the market, which are not suitable to display on lower resolution devices or while multiple video feeds are being monitored. In the case when multiple images are displayed on a screen in a surveillance system with multiple cameras, down-sampling transcoding is essential for mega-pixel images. The transcoder should change both resolution and quality because the correct choice in quality depends on human perception and monitoring purpose. Human beings can perceive less quality for low-resolution images and moving pictures. Further, high quality is not expected when images are displayed in very low resolution, even if some quality degradation is evident.

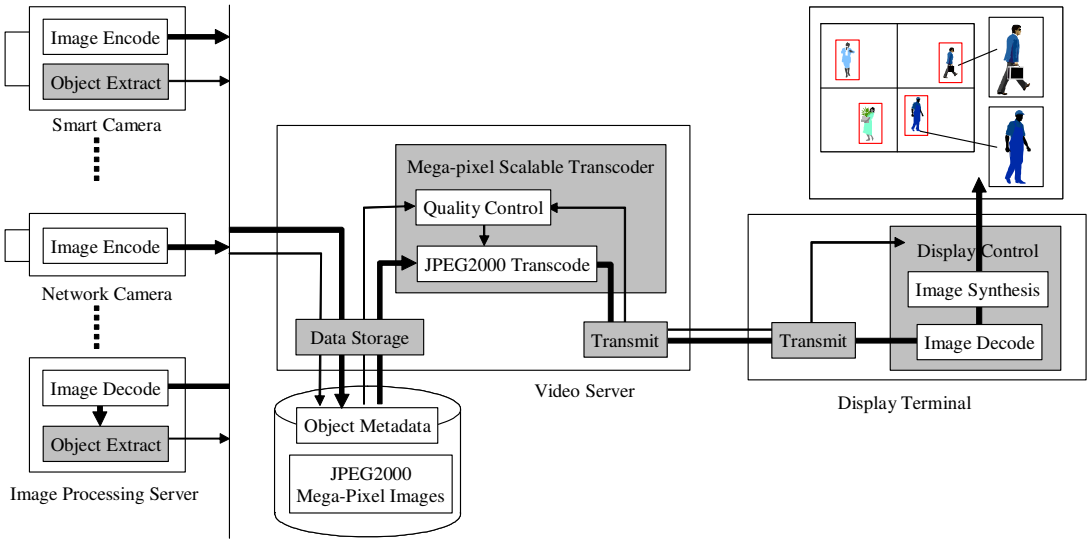

Fig. 1. Video Surveillance System with Mega-Pixel Scalable Transcoder

Figure 2 shows an example of multiple camera display with digital PTZ: pan, tilt and zoom. Nine cameras with 1/16 (1/4x1/4) resolution are displayed on a screen in (a). When something unusual is found in camera 1 and it is selected on the screen, camera 1 with 1/4 (1/2x1/2) resolution is displayed in a window which size is four times as the other cameras, as shown in (b). In (c), a 1/4 part of camera 1 with the original resolution is displayed in the same window because a mega-pixel image cannot be displayed as it is. Any part of camera 1 can be seen by a digital PTZ operation.

Multiple mega-pixel streams are transcoded into streams with low resolution, low quality, and sometimes for a part of an image in the server. This is done to decrease the total bit rate and facilitate the display of mega-pixel streams under restricted system resources.

Overall with low-resolution and ROI with high-resolution (OLRH)

OLRH displays ROIs with high-resolution and high quality, as well as an overall image with low-resolution and low-quality as shown in Figure 3. In this way, it is possible to observe ROIs in detail without missing the overall scene context. The total bit rate also decreases as in the MCPTZ application.

ROIs can be specified by both the metadata created by the image recognition described in section 2 and user interaction. OLRH with the metadata may cue a user to look at objects when they appear and the user does not need monitor the screen attentively all the time. It is also easy and precise to specify the ROI on the overall image interactively for PTZ control. Multiple ROIs are displayed at the same time on the same screen or on different screens though it is impossible with a combination of a fixed camera and a PTZ camera.

(a) Display nine cameras with 1/16 down-sampled images

(b) Display a selected camera with 1/4 down-sampled image

(c) Display a selected camera with a quarter of the original image and operate pan and tilt

Fig. 2. Multiple Camera Display with Digital PTZ

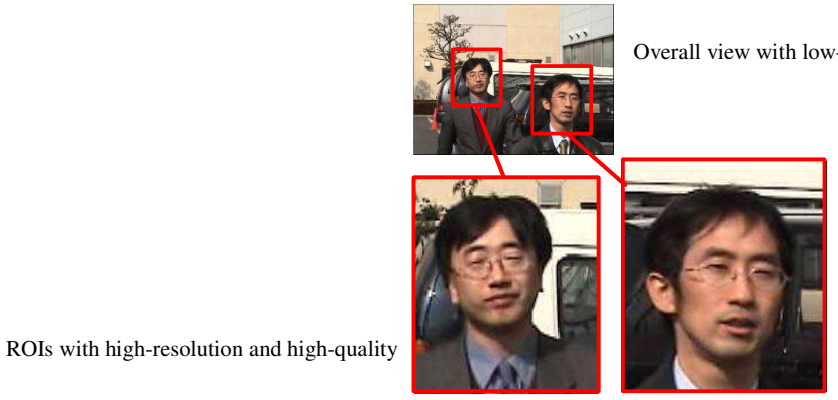

Fig. 3. Overall with Low Resolution and ROI with High Resolution

4. JPEG2000 SCALABLE TRANSCODING

The JPEG 2000 transcoding process supports reductions in both quality and resolution, is capable of adapting the quality of ROI and background, and supports additional functions such as cropping to achieve OLRH. The process itself is described in three main components including data analysis, quality control and ROI transcoder with byte manipulation as shown in Figure 4.

The data analysis module is responsible for extracting indexing information about the structure of the code stream. It is essentially a low-complexity parser that analyzes the packet header for each quality layer, resolution level and component. A multiple-dimensional array is used to store the packet information, which indicates the byte position, header length and body length for each packet. Since this partial decoding operates on the packet header only without performing entropy arithmetic decoding for code blocks, the computational complexity is very low.

The byte manipulation in the transcoder supports reduction of spatial resolution and quality layers, as well as a cropping functionality. For a given a set of ROI coordinates, the quality of the ROI could be increased by replacing packets at higher quality layers that are associated with the background of the scene with empty packets as defined by the JPEG 2000 standard. This is an effective method for reducing the rate of the overall code stream while retaining the quality of important objects and keeping the complexity low. To support OLRH, the transcoder must crop multiple ROI regions from the input image to create multiple output streams of ROI and backgrounds as need for the OLRH function. It is also worth noting that each components of the OLRH output is generated by analyzing the original code stream once, then transcoding it with different configurations to generate the each component. This simplifies the transcoder's logic, while allowing any number of streams/components to be generated from the input.

The number of quality layers for the background and ROI are determined by a quality control module. We have recently described a dynamic rate control algorithm that determines the quality layers based on target rate, buffer occupancy and ROI information [3]. In this work, the key functions have been extended to also work in the OLRH context.

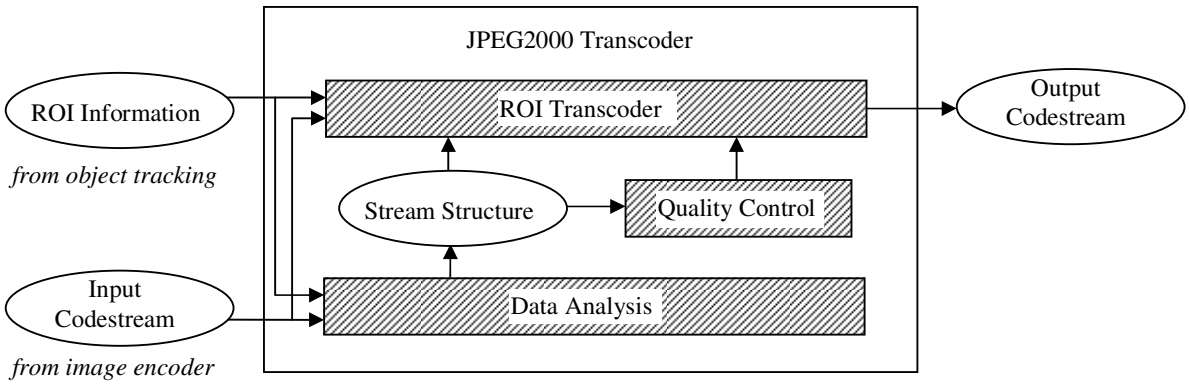

Fig. 4. Overview of JPEG2000 Transcoding System

5. EXPERIMENTAL RESULTS

5.1 Experimental Environment and Video Sequence

To confirm the effectiveness of the video streaming with transcoded images with low-resolution and low-quality and OLRH, we perform experiments with various combinations of scalability parameters. We evaluate subjective and objective image quality, subjective visual effectiveness, bit rate and computational complexity using several mega-pixel video sequences. Figure 5 and Table 1 show system construction and specifications of our evaluation environment respectively. Table 2 shows specifications of a video sequence and its example images with various quality are shown in Figure 6.

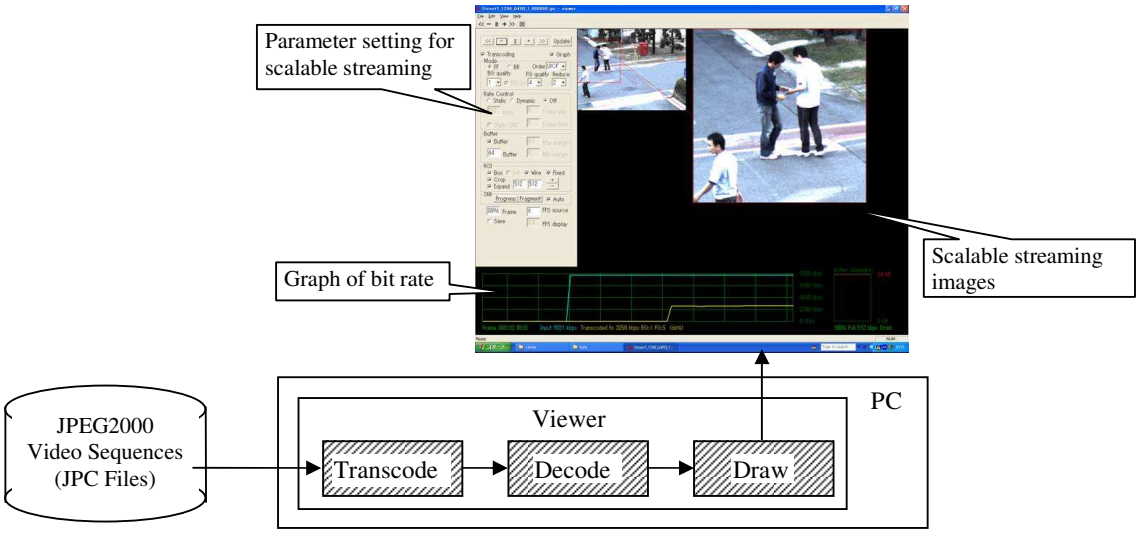

Fig. 5. System Construction of Evaluation System

Table 1. Specifications of Evaluation System

		Specifications
S/W	Application	Viewer
	JPEG2000 decoder	Kakadu Ver. 5.2 (Kakadu)
	OS	Windows XP Professional Ver. 2002, Service Pack 2
H/W	PC	Pentium 4 3GHz, 1GB main memory
	Graphics	NVDIA Geforce FX900, 1280x1024, 32bit color

Table 2. Specifications of Input Video Sequence

Original images	1280x960 pixels, 9.0fps, 348 frames, Pedestrians walking on a path
JPEG2000 encoding parameters	Quality layers = 4, Resolution levels = 3, Color = 4:4:4 Precinct = 128x128, 64x64, 32x32, 16x16, Code block = 64x64, Bit allocation = 1.0, 0.3, 0.15, 0.075 Data size = 150KB/frame, Bit rate = 10.5Mbps

(a) Original Overall Image

(b) Part of Original Image

(c) Part of Image with Q=3

(d) Part of Image with Q=2

(e) Part of Image with Q=1

(f) Part of Image with Q=0

Fig. 6. Example Images of Video Sequence

5.2 Evaluation of Images with Various Combinations of Resolution and Quality

The bit rate of transcoded streams with different combinations of resolution and quality is shown in Figure 7. The bit rate with the same quality layer decreases as the resolution decreases though the decreasing degree of bit rate is less than that of resolution. For example, the bit rate with quality layer = 3 is (10.5, 10.3, 5.6, 2.4) Mbps for resolution (1280x960, 640x480, 320x240, 160x120) respectively. The bit rate difference between 1280x960 and 640x480 with the same quality layer is very small. On the other hand, the bit rate of each quality layer in 1280x960 and 640x480 is proportional to the bit allocation for each quality layer. The bit rate ratio among different quality layers in 320x240 and 160x120 is less than that of the bit allocation. This may be because the quality layer 3 in higher resolution level includes a lot of quality data related to resolution such as image granularity. It is necessary to operate not only the resolution but also the quality layer for bit rate reduction.

Necessary quality (quality layer, resolution and frame rate) is dependent on monitoring purpose in surveillance context. In general, moving pictures are used to find suspicious people and unusual events. Image frames with high quality are required to observe faces and small things in detail and high frame rate is required to watch quick motion. Still images with resolution and quality higher than the moving pictures are necessary for very detail observation and evidence. The video sequence in Table 2 is encoded with multiple quality layers and resolution levels to satisfy various monitoring purposes and one bit is allocated to the highest quality layer to observe faces and small things in detail in 1280x960. Table 3 shows subjective quality evaluation of moving pictures. First we define monitoring purpose for each resolution from the surveillance context and evaluate images with various quality layers at four levels subjectively for the monitoring purpose of each resolution. “Fine” and “good” mean that the quality degradation is negligible or acceptable respectively. It should be noted that “fine” may include the case of having excessive quality for moving pictures because the quality layers provide a discrete set of rate points. “Conditionally usable” means that the quality can be used in some cases though the quality degradation is noticed. The appropriate quality for moving pictures is defined as a set of quality layers evaluated “good” in Table 3. We estimate an appropriate quality for 1280x960 between quality layers 2 and 3 because there is no quality layer with “good” evaluation.

Figure 8 shows the average PSNR of RGB components for images with various combinations of resolution and quality. For images with reduced resolution, the PSNR is calculated based on original images and up-sampled images. We use a Lanczos filter for the up-sampling. To capture the loss due to resolution conversion, we plot the PSNR of the reduced resolution images that have been up-sampled as indicated by the Q_L points in Figure 8. The PSNR decreases as the quality layer and the resolution decreases respectively. The difference between PSNR of images with different quality and the same resolution decrease as the resolution decreases. From this plot we could conclude that the high quality layer may not be needed for lower resolution images due to the small differences in measured quality.

As for computational complexity, Figure 9 shows average processing times of transcoding, decoding and drawing per frame with different combinations of resolution and quality. The average times of decoding and drawing with the same quality decrease as the resolution decreases. On the other hand, the transcoding time is relatively constant because the data analysis (parsing) is the dominant portion of the transcoding complexity and independent of the transcoding configuration [2]. It should be noted that the sum of decoding time and drawing time for images with the original resolution is over 100ms and these images cannot be displayed in real time even on a high performance PC. The resolution transcoding is essential for mega-pixel images.

The resolution has much influence for the processing times of decoding and drawing while the quality layer has much influence for the bit rate. It is important to select the resolution and quality carefully considering the monitoring purpose and the available system resource such as a network bandwidth and computing power.

From the above evaluations, the following scalable video streaming provides an effective multiple camera display to show the important parts in detail while showing multiple cameras on a screen under the restricted resources.

- State 1: Display moving pictures of multiple cameras with low-resolution and low-quality
- State 2: Display moving pictures of an important camera with higher-resolution and higher-quality than State 1
- State 3: Display a still image of a very important scene with highest-resolution and highest-quality

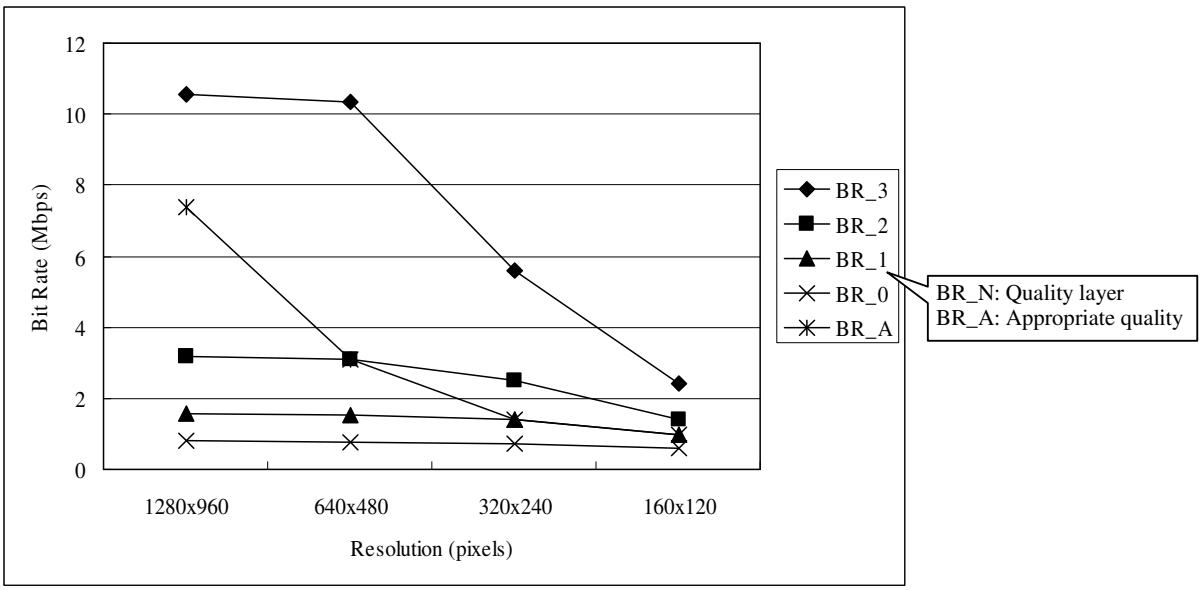

Fig. 7. Bit Rate of Images with Various Sets of Resolution and Quality

Table 3. Subjective Quality Evaluation of Moving Pictures and Compression Ratio

Resolution	1280x960	640x480	320x240	160x120
Monitoring purpose	Observe face, small thing and small behavior in very detail 340–480 pixels of 1.7m tall man	Observe face, small thing and small behavior in detail 170–240 pixels of 1.7m tall man	Observe medium size thing and behavior in detail 85–120 pixels of 1.7m tall man	Observe large thing and behavior 42–60 pixels of 1.7m tall man
Q=3	*Fine* CR=1.0 Less granularity as a still image than the original	*Fine* CR=0.98 Almost the same quality of a still image as the lossless	*Fine* CR=0.53 Almost the same quality of a still image as the lossless	*Fine* CR=0.23 The same quality of a still image as the lossless
Q=2	*Conditionally Usable* CR=0.30 Image degradation such as less texture, jaggy, unrecognized face, color change	*Good* CR=0.29 Less granularity than Q=3	*Fine* CR=0.24 Almost the same quality as Q=3	*Fine* CR=0.13 Almost the same quality as Q=3
Q=1	*Bad* CR=0.15 Increasing image degradation	*Conditionally Usable* CR=0.15 Image degradation is perceived	*Good* CR=0.13 Less granularity than Q=2	*Good* CR=0.09 Less granularity than Q=2
Q=0	*Bad* CR=0.07 Increasing image degradation	*Bad* CR=0.07 Increasing image degradation	*Conditionally Usable* CR=0.07 Image degradation is perceived	*Conditionally Usable* CR=0.06 Image degradation is perceived
Appropriate quality	Q=3-2, CR= 0.7	Q=2, CR=0.29	Q=1, CR=0.13	Q=1, CR=0.06

CR (Compression Ratio) is bit rate ratio of each stream compared to the stream with 1280x960 and Q=3.

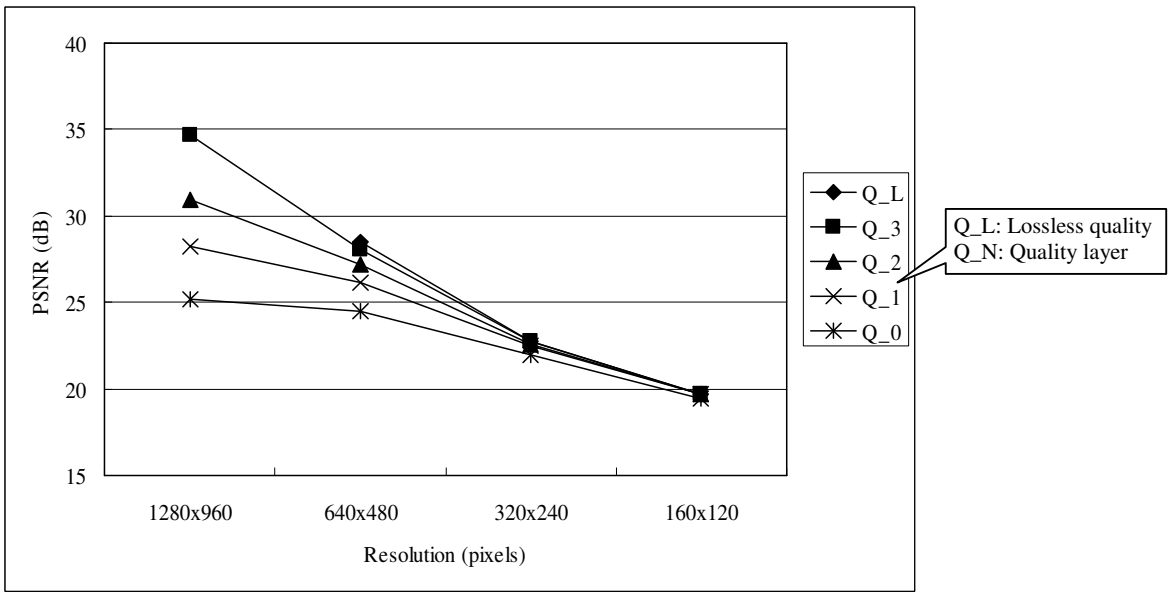

Fig. 8. PSNR of Images with Various Sets of Resolution and Quality

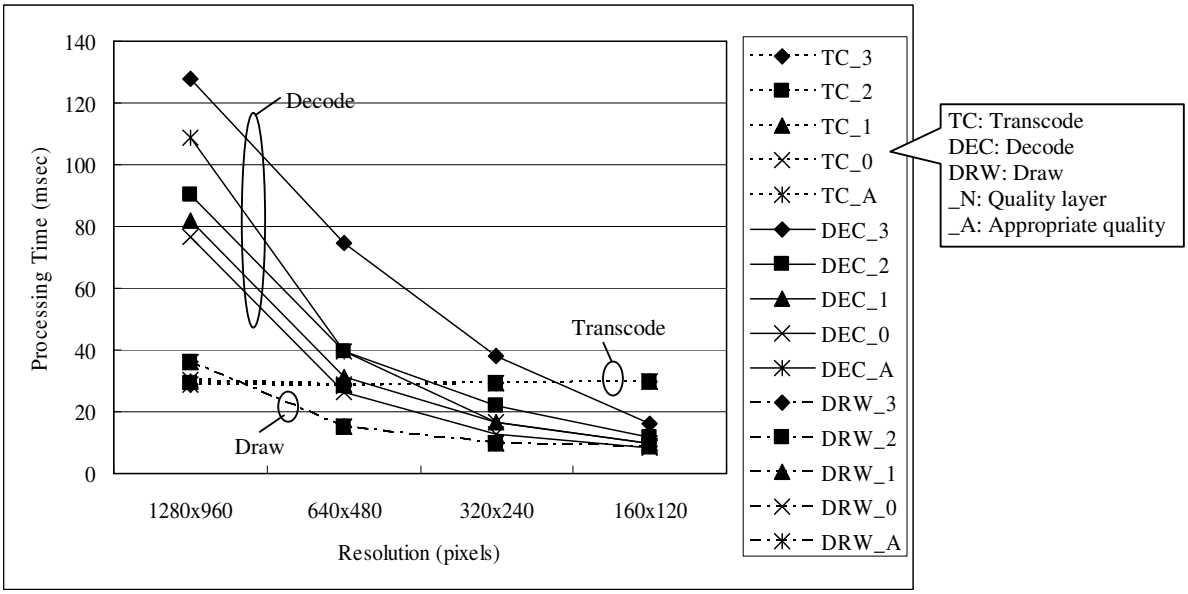

Fig. 9. Processing Times of Images with Various Sets of Resolution and Quality

5.3 Evaluation of OLRH

We evaluate subjective visual effectiveness, bit rate and processing times of OLRH with various combinations of resolution and quality. Examples of the bit rate and processing times per frame are shown in Figures 10 and 11 respectively. The quality layer of ROI is three. The bit rate, decoding time and drawing time decrease as the total resolution of ROIs and an overall image decreases while the transcoding time is almost constant as in Figures 7 and 9.

OLRH provides fine visual effectiveness and efficiency described in Section 3 by setting ROIs with quality = 3 and no down-sampling and an overall image with quality = 1 and 1/16 down-sampling or quality =0 and 1/64 down-sampling for the video sequence in Table 2. An overall scene with 1/4 down-sampling disturbs an observation because of large eye movement. It is important to note that the bit-rate and complexity are much less than that of the original mega-pixel streaming as shown in Table 4 and the OLRH images can be displayed in real-time.

We implement two methods to specify ROIs: user interaction on the overall image and the metadata extracted by a face detection algorithm [5]. It is very easy and precise to specify ROIs in the user interaction while it is usually difficult to manipulate a camera with mechanical PTZ over a network because of communication delay. It is very useful to use the metadata but it is noisy when the size and position of ROI changes often and quickly. The face detection module detects faces for each frame, but does not identify them over successive frames. We need higher level metadata processing.

Table 4. Bit Rate and Processing Time of OLRH

ROI	Overall	Bit Rate (Mbps, %)	Processing Times (ms, %)		
			Transcode	Decode	Draw
512x512, Q=3	320x240, Q=1	3.46, 32.9	31.8	45.9, 35.9	6.54, 18.2
512x512, Q=3	160x120, Q=0	2.68, 25.4	31.9	37.8, 29.6	6.22, 17.3
256x256, Q=3	320x240, Q=1	1.89, 18.0	31.2	26.7, 20.9	3.86, 10.7
256x256, Q=3	160x120, Q=0	1.11, 10.5	31.9	18.1, 14.1	3.74, 10.4

Bit rate and processing times per frame and their ratio compared to the original streams

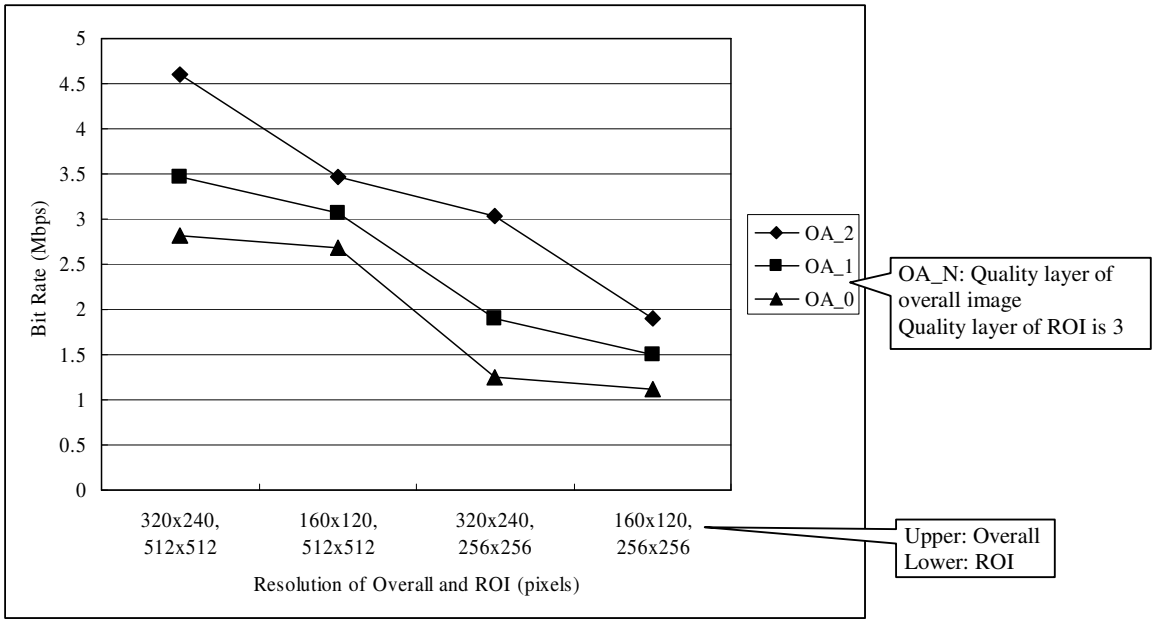

Fig. 10. Bit Rate of OLRH with Various Sets of Resolution and Quality

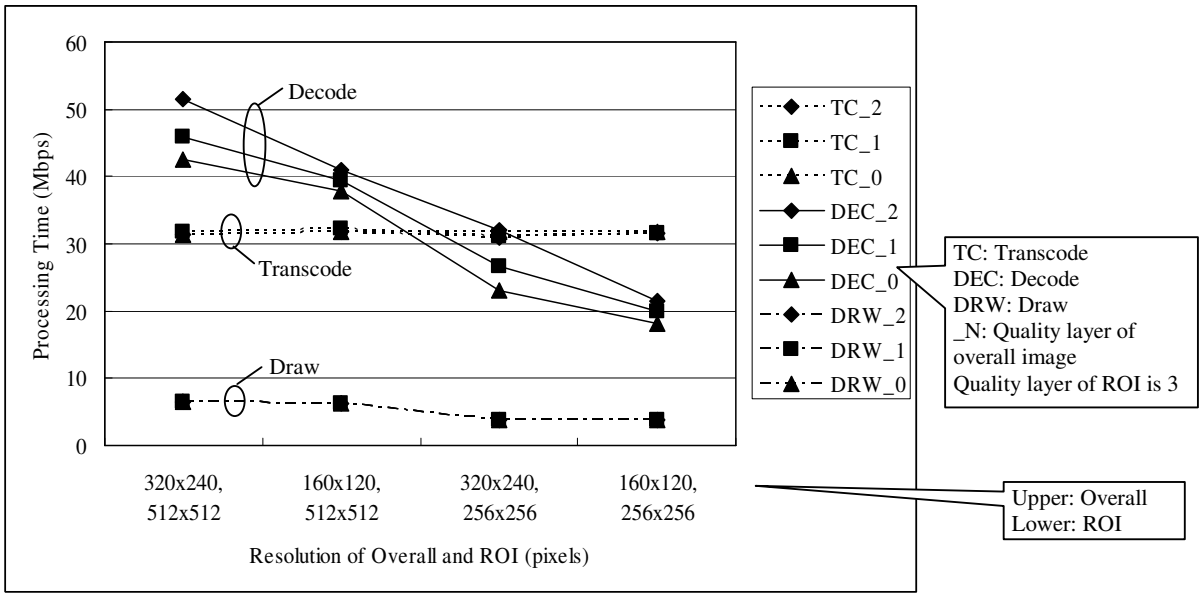

Fig. 11. Processing Times of OLRH with Various Sets of Resolution and Quality

6. CONCLUDING REMARKS

This paper presented a surveillance system that transcodes JPEG2000 mega-pixel images efficiently in the compressed-domain based on the ROI information as well as resolution and quality to display mega-pixel streams effectively with limited system resources. Two streaming methods are proposed. In MCPTZ, multiple cameras are displayed with lower resolution and lower quality at the same time and the resolution and quality of the selected cameras can be increased dynamically. This is done to decrease the total bit rate and facilitate the display of mega-pixel streams under restricted system resources. OLRH displays ROIs in high-resolution and high-quality and an overall view of the scene with low-resolution and low-quality simultaneously to observe each object in detail without missing the overall scene context. We also proposed JPEG2000 transcoder that uses a combination of three types of scalability: quality, resolution and position, to achieve such effective streaming methods. The experimental results showed the mega-pixel images should be transcoded with not only resolution but also quality to save bandwidth and computing power. OLRH is visually effective and consumes much less bit rate and computational complexity compared to direct processing of mega-pixel streams.

In terms of future work, we consider scalable progressive streaming and adaptive rate control using not only quality but also resolution.

REFERENCES

1. ISO/IEC 15444-1, "Information Technology – JPEG2000 Image Coding System – Part1: Core Coding System," March 2000.
2. T. Hata, N. Kuwahara, T. Nozawa, D. Schwenke and A. Vetro, "Surveillance System with Object-Aware Video Transcoder," IEEE International Workshop on Multimedia Signal Processing, Shanghai, China, November 2005.
3. D. Schwenke, A. Vetro, T. Hata and N. Kuwahara, "Dynamic Rate Control for JPEG 2000 Transcoding," IEEE International Conference on Multimedia & Expo, Toronto, Canada, July 2006.
4. F. Porikli and O. Tuzel, "Human Body Tracking by Adaptive Background Models and Mean-Shift Analysis," IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Graz, Austria, March 2003.
5. P. Viola, M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," International Conference on Computer Vision and Pattern Recognition, Vol. 1, pp. 511-518, December 2001.