IP 네트워크에서 시점전환이 가능한 고화질 다시점 방송 시스템

(View-switchable High-Definition Multi-View Broadcasting over IP Networks)

이 석 희 † (Seokhee Lee) 이 기 영 †† (Kiyoung Lee) 김 만 배 ††† (Manbae Kim)

한 충 신 †††† (Chungshin Han) 유 지 상 ††††† (Jisang Yoo) 김 종 원 †††††† (Jongwon Kim)

요 약 본 논문에서는 시점변환이 가능한 고화질 다시점 비디오 전송 시스템의 디자인 및 구현에 관해 소개한다. 다시점 비디오는 기존의 방송 시스템에 실감성과 몰입감을 증가시키는 차세대 방송 컨텐츠로 각광 받고 있다. 하지만 다시점 비디오 전송 시스템을 구축하기위해서는 고비용의 하드웨어 컴포넌트들이 필요하고 충분한 대역폭이 요구된다. 본 프로토타입 시스템은 고비용 및 전송 대역폭의 문제를 해결하기 위해서 소프트웨어 기반의 시스템 컴포넌트와 IP 멀티캐스트를 이용한 전송에 중점을 둔다. 또한 유동적으로 시점 및 사용자의 수를 확장할 수 있는 시스템 디자인을 고려한다. 본 논문에서는 소프트웨어 기반의 시스템 컴포넌트로 다중화와 역다중화의 구현과 얼굴 추적을 통한 시점전환 기법을 소개한다. 또한 선도망/연구망으로 구성된 실제 네트워크 상황에서 전송 시현을 통해 사용자에게 실감성과 몰입감을 제공할 수 있음을 보여준다.

키워드 : 다시점 비디오, 방송 시스템, 실감성, 몰입감, IP 멀티케스트

Abstract In this paper, we present a prototype of view-switchable high-definition (HD) multi-view video transmission system. One of the major bottlenecks for the multi-view broadcasting system has been the hardware cost and transmission bandwidth. The proposed system focuses on software-based design, transmission over IP multicast networks, and flexible system configuration to address aforementioned problems. In the proposed system, we implement software stereo HD multiplexing, demultiplexing and decoding, and take advantage of high-speed broadband convergence networks to deliver HD video in real-time. Moreover, the proposed system can be scalable and flexible in terms of the number of views. Furthermore, in order to display any multiview video on 3D display monitor, a face tracking system is integrated to our system. Therefore, users can watch the different stereoscopic video at its related locations.

Key words : Multi-view video, broadcasting system, immersion, IP multicast

· This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment) (IITA-2005-(C1090-0502-0022)).

† 비 회 원 : 광주과학기술원 정보통신공학과
shlee@nm.gist.ac.kr
†† 비 회 원 : LG 전자
kylee@lge.com
††† 비 회 원 : 강원대학교 전기전자정보통신공학 교수
manbae@kangwon.ac.kr
†††† 비 회 원 : 서울통신기술 (Seoul Commtech)
chungshin.han@samsung.com
††††† 비 회 원 : 광운대학교 전자공학과 교수
jsyoo@daisy.kw.ac.kr
†††††† 종신회원 : 광주과학기술원 정보통신공학과 교수
jongwon@nm.gist.ac.kr

논문접수 : 2006년 9월 14일
심사완료 : 2007년 6월 19일

1. Introduction

Even though there has been much interest in the multiview video, multi-view high-definition (HD) video systems have been neither technically nor commercially viable until recently. Following the rapid development of broadband Internet service, several companies have provided the service of multi-view 3D products that can be delivered over the Internet. Therefore, it is expected that multi-view 3D products will gain popularity in the near future. In spite of such rapid technical progress, the development of a multi-view video system dealing with HD resolution still poses many technical challenges in the fields of signal processing and transmission.

In ATTEST project, researchers developed a 2D compatible multi-view 3DTV system for broadcast environments [1]. MERL (Mitsubishi Electric Research Laboratories) proposed a multi-view 3DTV prototype system with real-time acquisition, transmission and autostereoscopic display [2]. NTT (Nippon Telegraph and Telephone Corporation) Cyber Space Laboratories proposed a free-viewpoint video communication system using a multi-view video compression [3]. ETRI (Electronics and Telecommunications Research Institute) developed an experimental testbed for 3DTV broadcasting systems that are compatible with current HDTV broadcasting infrastructure such as terrestrial and satellite DS-3 network [4]. In order to deliver natural viewing experience, ATTEST project proposed an efficient approach to generate 3D contents based on a depth camera [1]. The system of MERL can provide 16 view-points and 1024×768 pixels per viewpoint so that immersive and convincing 3D experience can be enjoyed without special glasses [2]. NTT provided a free-viewpoint viewer that generates a natural view from any arbitrary viewing positions and directions. To realize this, it is reported that Ray-Space interpolation and extrapolation methods are applied to decoded multiview video [3]. Moreover, ETRI already provided an experiment service of 3D HDTV broadcasting relay during 2002 FIFA World Cup Korea-Japan [4].

In this paper, we have built a prototype of view-switchable HD multi-view video transmission system with the following features:

- Software-based: We implement software-based stereo HD multiplexing, demultiplexing and decoding. One of the major bottlenecks for 3D broadcasting systems is hardware cost. The cost of a true-HD camera and its components such as HD encoders/decoders and HD-SDI interface boards can be reduced by software-based development.
- Transmission over high-speed IP networks: Our proposed system enables high-resolution video to be delivered in real-time with a minimal amount of lag over high-speed broadband convergence networks (BcN). Furthermore, a dedicated feedback channel from a client provides the client with interactive and personalized service.
- Flexible system configuration: Each video stream is compressed individually. If the property is combined with two aforementioned features (software-based components and transmission over IP networks) as mentioned above, the proposed system can be completely scalable and flexible in the number of acquired, transmitted, and displayed views.

The remaining outline of the paper is as follows. In Section 2, we introduce the proposed view-switchable 3D HD video system. Section 3 deals with implementation and demonstration of the proposed system. Finally, conclusions and future work are given in Section 4.

2. Proposed System

Our view-switchable HD multi-view video system is composed of view acquisition, multiplexing, transmission, demultiplexing, decoding, and 3D display with view adaptation as shown in Fig. 1. The acquisition stage captures stereoscopic HD videos, VP_1, ..., VP_N, each of which consists of left video, VP_{iL} and right video VP_{iR}, i = {1, ..., N}. The stereoscopic HD video streams are then broadcast on separate channels of a multicast network in the transmission stage. In the display stage, a receiver decodes the received videos and displays them on a 3D display monitor after demultiplexing. A view adaptation module being composed of a face tracker on the top of the monitor and face-tracking soft-

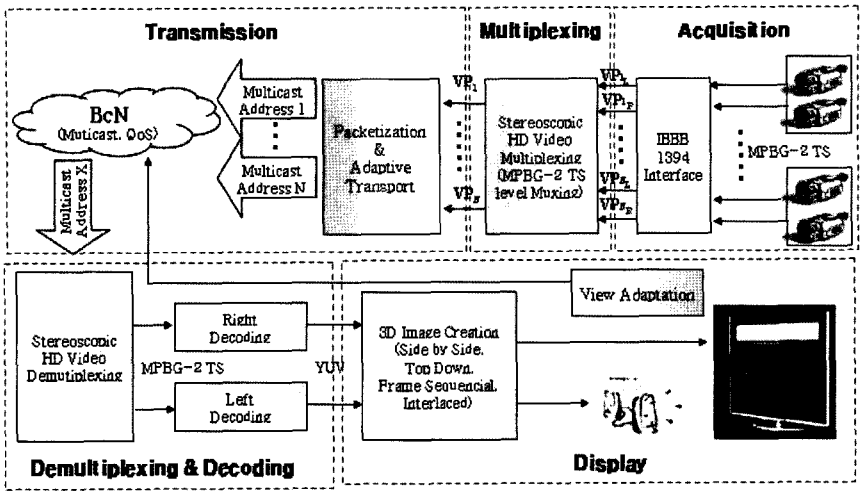


Fig. 1 System architecture

ware locate the position of a human face and displays its associated video chosen from multi-view videos.

2.1 Acquisition and Transmission

Video acquisition system consists of prosumer HD video cameras with MPEG-2 encoder and PCs with IEEE1394 ports. The video streams are compressed individually with a built-in MPEG-2 TS encoder. After acquiring the streams, a pair of left and right HD videos is interlaced and forms a stereoscopic video. The multi-view multiplexing is performed in Transport Stream (TS) layer level. A transport stream consists of a sequence of fixed sized transport packet of 188 bytes. Each packet comprises 184 bytes of payload and a 4-byte header. In the 4-byte header, 13-bit Packet Identifier (PID) distinguishes either left or right video stream. The both streams are multiplexing into a single stream by changing PIDs of a left transport stream.

After multiplexing video streams, timing synchronization is carried out and the synchronized streams are sent over a multicast network as shown in Fig. 2. A timing management module implements program clock references (PCR) that is stored in the adaptation field of a TS packet header. The transmission module adds RTP header into the TS packets to make RTP packets and then sends them to the network. An RTP packet of 1,140 bytes is composed of six TS packets.

After receiving an RTP packet, a receiver works for parallel demultiplexing and decoding. Fig. 3 illustrates the demultipexing procedure in detail. First, the receiver should demultiplex the received

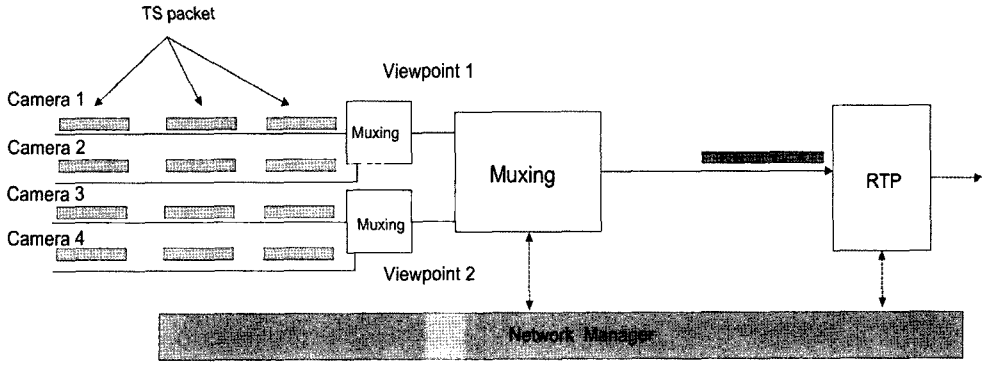


Fig. 2 View-switchable 3D HD transmission

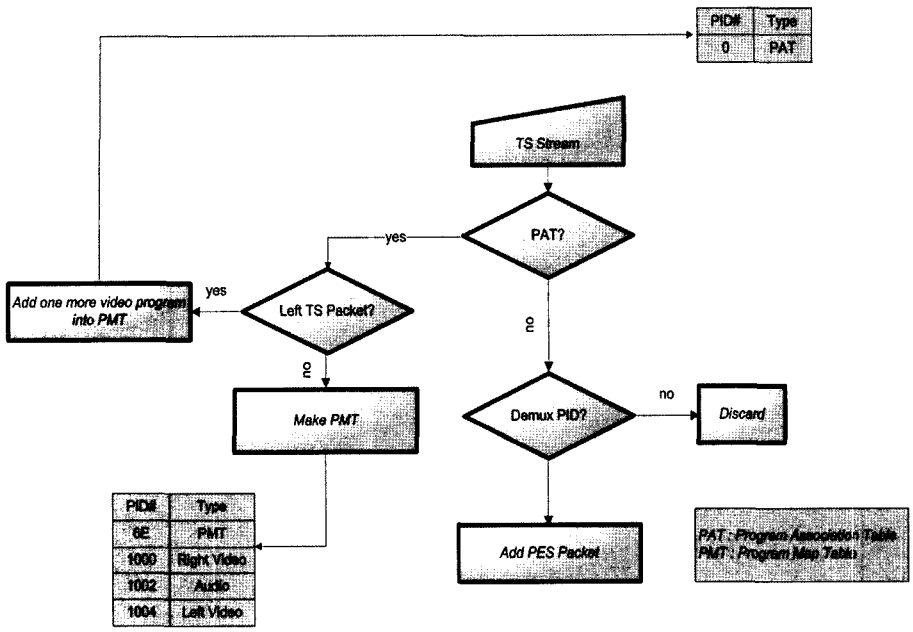


Fig. 3 Stereoscopic HD video dimultiplexing

stream. The muxed stream has two programs such as left and right cameras. The tables called program specific information (PSI) in MPEG-2 are made up for a description of the elementary streams which need to be combined to build programs. Program association table (PAT), transmitted with PID 0, contains a complete list of all programs in the transport stream along with the PIDs for the program map table (PMT) for each program. When making a PMT with left PSI, the left video ES is added into the PMT. The final PMT has two video and one audio streams. As shown in Fig. 3, PMT is made from right PSI. According to the PID, each packetized elementary stream (PES) is made. From demultiplexing process, the receiver makes two decoding threads for left and right video streams. In the proposed system, we reduce system resource by ignoring the left audio part. The two videos are continuously decoded when timing information like PTS and DTS is scheduled. The decoding is performed separately.

2.2 Display with User Interaction

Given multiple views, a user can select a view-point by moving a face, as shown in Fig. 4. According to the face location, its associated stereoscopic video is displayed. The input image obtained from a face tracking camera on the display device is RGB color image and its size is 320 × 240. The face region extraction algorithm is applied after converting RGB into YC_bC_r format. Only the part of skin color is extracted from images, to make a histogram for the range of skin color values of C_b and C_r. The selected range of C_b and C_r used in our algorithm is expressed in Eq. (1).

$$f(x,y)=\begin{cases}0 & if\,(137<Cb<152)\cap(123<Cr<137)\\ 255 & otherwise\end{cases} \quad (1)$$

In order to eliminate noise, an opening morphological operation is first performed on segmented data as in Fig. 5(a) [5]. Next, in the horizontal scan, the number of zero pixels is computed and pixel values less than a threshold value are converted to 255. In the experiment, the threshold value is set to the half of a maximum value. After finishing the scanning of all the rows, Fig. 5(c) shows an output image.

In order to eliminate incorrect skin regions surrounding the face, it is needed to decide whether a detected candidate skin color region is a face area. For this, characteristic points of a face are detected through geometric features and templates of the face. Since a viewer wears polarized glasses, the eye detection is not an easy task. To overcome

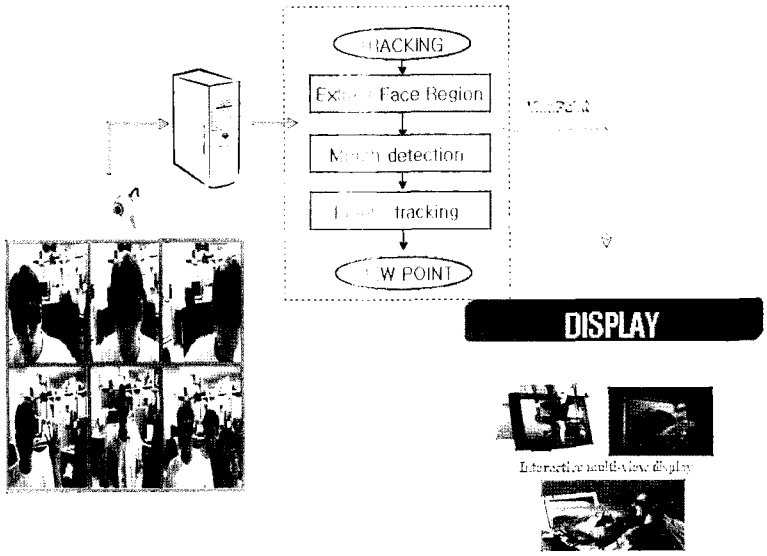


Fig. 4 Face tracking for the selection of a viewpoint

(a) (b) (c)

Fig. 5 Detection of a face region: (a) color segmentation, (b) noise removed, and (c) face region

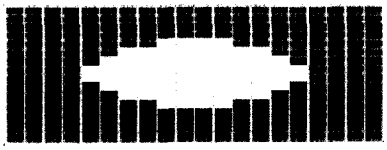


Fig. 6 Mouth template with 50 × 16 (in pixels)

this, a template for a mouth as shown in Fig. 6 is designed.

After extracting two characteristic points of a face, mouth and polarized glasses, the location and distance of them are measured. When a large error between the measured values and geometric features of the face occurs, it is declared as a candidate skin color region. After the face region is extracted, the extracted face region is defined as a key viewer. For face tracking, a method of using color based statistical color modeling and deformable templates are used [6]. The face region shows different features from a background area other than a face region. Main color of the face region is composed of skin and facial hair colors. On the contrary, the background region is composed of other different colors. That is, in Eq. (2), when the energy value is a minimum, the location value of a template is assumed to be the location of a face.

$$f(R) = \sum_{r \in R} \log\left[\frac{p(x_r|w_2)}{p(x_r|w_1)}\right] \quad (2)$$

Here, r is a pixel of the template region R and x_r is a pixel value. $p(x|w_i)$ is probability that x is in region w_i, where w_1 and w_2 mean foreground and background, respectively. A face region with a minimum energy of f(R) is searched while adjust-

ing the deformable template region size. The face tracking algorithm uses the statistical characteristic of face color by combining the deformable template matching method. Therefore, it can support a robust and correct viewpoint selection even under complicated background image.

The information on the user's face position changes into multicast join/leave message to receive the stereoscopic HD video which he or she wants to watch. There exists an unavoidable delay during view-point switching because of the network and processing latencies of the multicast group management protocol. In order to compensate the delay, a viewpoint prediction scheme that calculates a viewpoint after the delay is needed at the receiver side. Then, the receiver joins the multicast channel providing the video stream of predicted viewpoint in advance. For a short period of time, the receiver receives previous and predicted viewpoint streams at the same time. If the prediction is correct, the receiver leaves the multicast channel providing the previous viewpoint stream. On the other hand, it stops receiving the video stream from the multicast channel corresponding to the predicted viewpoint.

3. Experiments

As shown in Fig. 7, our system is composed of a highend PC, a laptop, four HD cameras, and two 3D monitors. Workstation 530MP with dual Xeon 1.7 GHz CUP, 1 GB memory, Gigabit network card, four IEEE1394 ports, and Linux OS is used on the sender side. A receiver system is equipped with Dell D800 laptop with Pentium M 2.0 GHz, 1 GB memory and Windows XP OS.

We use four JVC GR-HD1 cameras which could generate video streams encoded with MPEG-2 MP@HL through IEEE1394 as shown in Fig. 8(a). The resolution of each stream is 1280×720 at 30 fps. Each stream consumes 19.2 Mbps bandwidth when it is delivered over IP networks. Therefore, the total network bandwidth needs to support at least 80Mbps in order to send muxed video stream

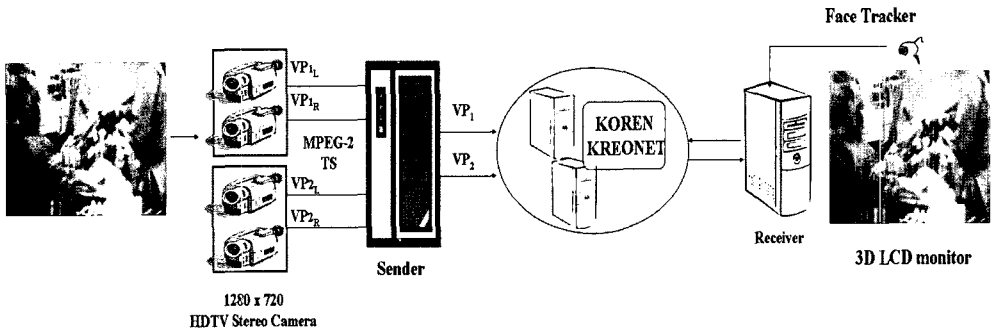
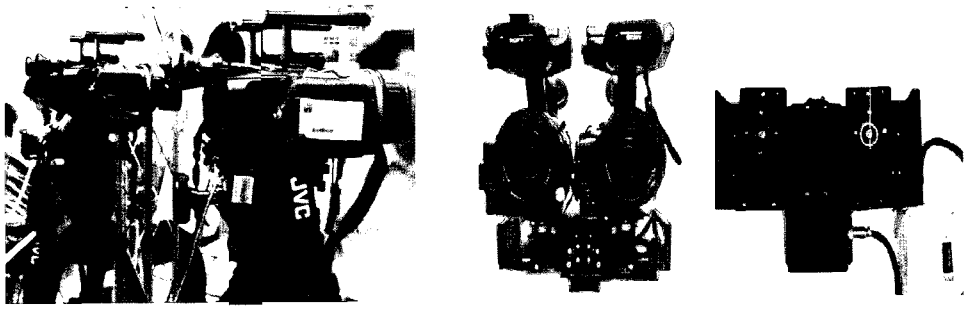


Fig. 7 Implementation overview of view-switchable 3D HD video system



(a) (b)

Fig. 8 Camera configuration for view-switchable HD muliview video system: (a) multi-view camera configuration and (b) cameras and mount for stereoscopic HD video

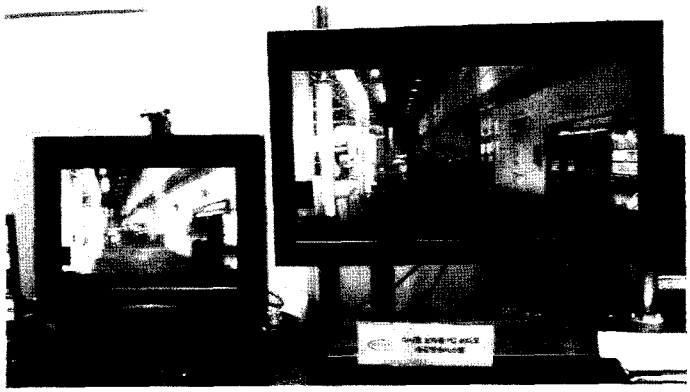


Fig. 9 3D HD video display monitors with a face tracking camera

and overheads. In order to generate a stereoscopic HD video, two cameras placed on the top of a camera mount are used as shown in Fig. 8(b).

Two 3D LCD monitors manufactured by Pavonine Inc. are used in our experiments as shown in Fig. 9. Our demonstrations were performed though KOREN (KOrea advanced REsearch Network) and KREONET (Korea Research Environment Open NETwork) during ITRC (Information Technology Research Center) Forum in Seoul, Korea on June, 2005.

4. Conclusion and Future Work

In this paper, we presented a prototype of view-switchable HD multi-view video transmission system. The design of the proposed system focuses on software-based design, transmission over high speed IP multicast network, and flexible system configuration. As future works, we need to reduce random access delay between different viewpoint video streams and to increase the number of cameras for a general multi-view system.

References

[1] A. Redert, M. Beeck, C. Fehn, W. IJsselsteijn, M. Pollefeys, L. Gool, E. Ofek, I. Sexton, and P. Surman. "ATTEST: Advanced three-dimensional television system technologies," in *Proc. International Symposium on 3D Data Processing, Visualization and Transmission*, pp. 313-319, 2002.

[2] W. Matusik and H. Pfister, "3D TV: A scalable system for real-time acquistion, transmission and autostereoscopic display of dynamic scenes," *ACM Transactions on Graphics*, Vol.23, No.3, pp. 814-824, 2004.

[3] H. Kimata, M. Kitahara, K. Kamikura, and Y. Yashima, "Free-viewpoint video communication using multi-view video coding," *NTT Technical Review*, Vol.2, No.8, 2004.

[4] N. Hur and C. Ahn, "Experimental service of 3DTV broadcasting relay in Korea," in *Proc. SPIE ITCOM 3DTV, Video, and Display*, Vol.4864, pp. 1-13, 2002.

[5] J. Serra and P. Soille, "Mathematical morphology and its applications to image processing," *Series on Computational Imaging and Vision, Kluwer Academic Publishers*, 1994.

[6] R. Rao and R. Mersereau, "On merging hidden markov models with deformable templates," *in Proc. International Conference on Image Processing*, pp. 556-559, 1995.

이 석 희

2003년 2월 서강대학교 컴퓨터과학과(학사). 2005년 2월 광주과학기술원 정보통신공학과(석사). 2005년 3월~현재 광주과학기술원 정보기전공학부 박사과정. 관심분야는 Multimedia Delivery, Networked Virtual Environment, Teleoperation

이 기 영

2003년 8월 강원대학교 정보통신공학과(학사). 2005년 8월 광주과학기술원 정보통신공학과(석사). 2005년 9월~현재 LG 전자 주임연구원. 관심분야는 HD video delivery and display, 네트워크 보안

김 만 배

1983년 한양대학교 전자공학과(학사). 1986년 University of Washington 전기공학과(석사). 1992년 University of Washington 전기공학과(박사). 1992년~1998년 삼성종합기술원 수석연구원. 1993년 Georgetown University 의과대학 객원연구원. 1996년 University of Rochester 전기공학과 객원연구원. 1998년~현재 강원대학교 컴퓨터정보통신공학과 교수. 관심분야는 비디오신호처리, 영상통신, 입체영상처리, MPEG-21

한 충 신

2006년 2월 광운대학교 전자공학과(석사). 2006년 3월~현재 서울통신기술(Seoul Commtech) 연구원. 관심분야는 Codec, 3D

유 지 상

1985년 2월 서울대학교 전자공학과(학사). 1987년 2월 서울대학교 전자공학과(석사). 1993년 5월 Purdue 대학교 전기공학과(박사). 1993년 9월~1994년 8월 현대전자산업(주) 산전연구소 선임연구원. 1994년 9월~1997년 8월 한림대학교 전자공학과 조교수. 1997년 9월~현재 광운대학교 전자공학과 교수. 관심분야는 3D 입체 영상처리, 영상압축, 영상인식, 비선형 신호처리

김 종 원

1987년 서울대학교 제어계측공학과(학사). 1989년 서울대학교 제어계측공학과(석사). 1994년 서울대학교 제어계측공학과(박사). 1994년 3월~1999년 7월 공주대학교 전자공학과 조교수. 1997년 8월~2001년 7월 University of Southern California 연구 조교수. 2001년 9월~현재 광주과학기술원 정보기전공학부 부교수. 관심분야는 Networked Media Systems and Protocols focusing "Reliable and Flexible Delivery for Integrated Media over Wired/Wireless Networks"