Kinect 깊이 카메라를 이용한 실감 원격 영상회의의 시선 맞춤 시스템

정회원 이 상 범*, 종신회원 호 요 성*°

Real-time Eye Contact System Using a Kinect Depth Camera for Realistic Telepresence

Sang-Beom Lee* *Regular Member*, Yo-Sung Ho*° *Lifelong Member*

요 약

본 논문에서는 실감 원격 영상회의를 위한 시선 맞춤 시스템을 제안한다. 제안하는 방법은 적외선 구조광을 사용하는 Kinect 깊이 카메라를 이용해서 색상 영상과 깊이 영상을 획득하고, 깊이 영상을 이용해서 사용자를 배경으로부터 분리한다. 깊이 카메라로부터 획득한 가공되지 않은 깊이 영상은 다양한 형태의 잡음을 가지고 있기 때문에, 첫번째 전처리 과정으로 결합형 양방향 필터를 사용해서 잡음을 제거한다. 그 다음, 깊이값의 불연속성에 적응적인 저역 필터를 적용한다. 색상 영상과 전처리 과정을 거친 깊이 영상을 이용해서 우리는 가상시점에서의 화자를 3차원 모델로 복원한다. 전체 시스템은 GPU 기반의 병렬 프로그래밍을 통해 실시간 처리가 가능하도록 했다. 최종적으로, 우리는 시선이 조정된 원격의 화자 영상을 얻을 수 있게 된다. 실험 결과를 통해 제안하는 시스템이 자연스러운 화자간 시선 맞춤을 실시간으로 가능하게 하는 것을 확인했다.

Key Words : Eye contact system, gaze correction, depth camera, realistic telepresence, depth image-based rendering

ABSTRACT

In this paper, we present a real-time eye contact system for realistic telepresence using a Kinect depth camera. In order to generate the eye contact image, we capture a pair of color and depth video. Then, the foreground single user is separated from the background. Since the raw depth data includes several types of noises, we perform a joint bilateral filtering method. We apply the discontinuity-adaptive depth filter to the filtered depth map to reduce the disocclusion area. From the color image and the preprocessed depth map, we construct a user mesh model at the virtual viewpoint. The entire system is implemented through GPU-based parallel programming for real-time processing. Experimental results have shown that the proposed eye contact system is efficient in realizing eye contact, providing the realistic telepresence.

Ⅰ. 서 론

차세대 멀티미디어 컨텐츠인 3차원 비디오는 현실 세계를 재구성한 컨텐츠로부터 현실감 있는 느낌을 사용자에게 제공할 수 있기 때문에 많은 관심을 받고 있으며, 현재 사용하고 있는 2차원 비디오를 대체할 것으로 기대를 받고 있다. 색상 영상과 이에 상응하는 깊이 영상으로 구성된 3차원 비디오를 획득하는 방법은 수동 센서 기반 방법, 능동 센서 기반 방법으로 나눌 수 있다. 수동 센서 기반

* 광주과학기술원 정보통신공학과 ({sblee, hoyo}@gist.ac.kr), (°: 교신저자)
논문번호 : KICS2012-02-059, 접수일자 : 2012년 2월 11일, 최종논문접수일자 : 2012년 4월 10일

방법은 두 대 혹은 그 이상의 카메라로부터 획득한 2차원 영상의 상관관계를 유추함으로써 깊이 정보를 계산하는 방법이다. 대표적인 방법으로는 스테레오 정합 기술이 있다[1,2]. 능동 센서 기반 방법은 레이저, 적외선, 구조광 등과 같은 다양한 종류의 센서를 이용해서 3차원 장면으로부터 깊이 정보를 직접적으로 획득하는 방법이다. 깊이 카메라, 3차원 스캐너 등이 이 방법에 포함된다[3-5].

예전부터 능동 센서 기반 방법은 높은 정확도의 깊이 영상을 획득할 수 있는 대신 장비가 워낙 고가이다 보니 접할 수 있는 기회가 많지 않았다. 하지만, 최근 Kinect 깊이 카메라와 같이 저가임에도 불구하고 높은 성능을 보이는 카메라가 시중에 출시되면서 다양한 형태의 응용 분야에 많이 사용되기 시작했다[6]. 그로 인해, 능동 센서 기반 방법은 3차원 컨텐츠 저작 환경에서 가장 강력한 기술로 재평가 받기 시작됐다.

최근에는 몰입형 디스플레이를 위한 차세대 방송의 핵심 기술로서 깊이 영상 기반 렌더링 (depth image-based rendering, DIBR) 기법이 각광을 받고 있다[7]. DIBR 기법은 색상 영상과 텍스쳐 영상의 각 화소에 대응하는 거리 정보로 이루어진 깊이 영상 (depth image)을 사용하여, 임의의 시점에서의 영상을 렌더링하는 기법이다. DIBR 기법은 다양한 멀티미디어 산업에서 사용되고 있는데, 그 가운데 주된 응용 분야는 원격 영상회의가 있다. 원격 영상회의란 원격의 화자와 사용자가 마치 옆에서 대화하는 듯한 느낌을 제공해주는 기술을 말한다.

원격 영상회의의 주된 쟁점 가운데 하나인 시선 맞춤 기술은 오랜 기간 동안 많은 연구기관 사이에서 뜨거운 이슈로 자리 잡았다. 시선 맞춤을 해결하기 위해 많은 알고리즘들이 제안됐지만, 여전히 이 문제점은 쉽게 해결이 되지 않았다. 하지만, DIBR 기법을 사용하는 최근의 연구들을 통해 실감 원격 영상회의 구현이 가능해졌다. 최근의 연구들은 다수의 카메라를 디스플레이 주변에 배치시킨 다음, 깊이 정보를 탐색하는 방법을 사용했다[8,9]. 기존의 방법들은 깊이 정보를 예측한 다음, 시점 합성을 통해 시선을 조정했지만, 복잡한 하드웨어를 구성해야 하며 시간이 상당히 오래 걸리는 단점이 있다. 특히나, 깊이 탐색 기술의 성능이 장면의 환경에 민감하기 때문에, 안정적인 합성 결과를 기대하기 어렵다.

본 논문에서는 실감 원격 영상회의를 위해 깊이 카메라를 이용함으로써 시선 조정을 가능하게 하는 시스템을 제안한다. 제안하는 시스템은 적외선 구조광 패턴을 사용하는 Kinect 깊이 카메라를 통해, 실용적이고 안정적인 시선 맞춤을 가능하게 하는 것이 주 목표이다. 이 시스템은 디스플레이의 상단에 깊이 카메라 한대만 설치하는 상당히 간단한 구조를 가지는 장점을 가지고 있다.

본 논문은 다음과 같이 구성된다. 2장에서는 제안하는 시선 맞춤 시스템의 개요를 소개하고, 3장에서는 깊이 영상의 화질을 향상시키기 위한 전처리 기술들을 설명한다. 4장에서는 화자간 시선 맞춤을 위한 정면시점 영상합성 기술에 대해 기술하고 5장에서 실험결과를 통해 제안하는 시스템의 성능을 분석한 다음 6장에서 결론을 맺는다.

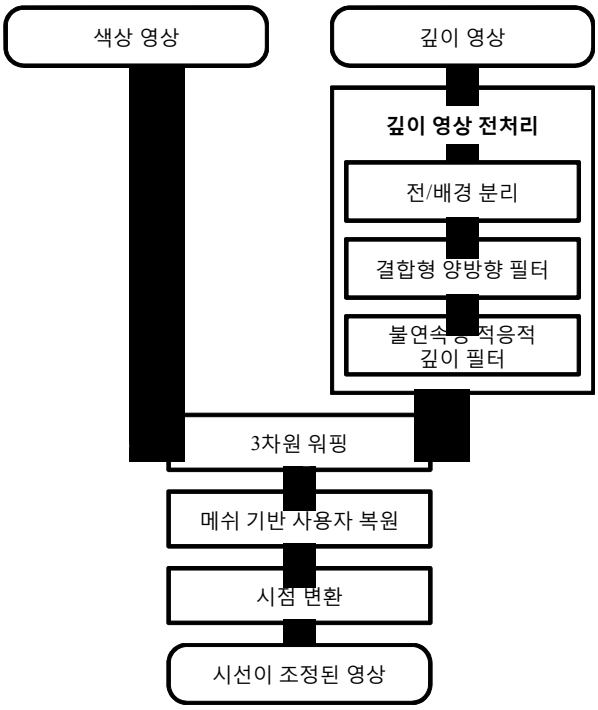


그림 1. 제안하는 시스템의 블록 다이어그램
Fig. 1. Block diagram of the proposed system

Ⅱ. 시스템 개요

그림 1은 시선 맞춤 영상을 생성하기 위한 제안하는 시스템의 구조를 나타낸다. 우선적으로, 깊이 카메라는 색상 영상과 깊이 영상을 동시에 획득한다. 깊이 영상의 전처리 과정에서는 첫번째로 단일 화자를 배경과 분리해낸다. 깊이 카메라는 카메라 자체의 센서 잡음과 구조광 패턴의 투사부와 수신부가 달라서 발생하는 폐색 영역으로 인해 전체 장면에 대해서 깊이값을 획득하지 못한다. 그렇기 때문에 결합형 양방향 필터를 이용해서 깊이 카메라가 획득하지 못한 영역에서의 깊이값을 채운다. 가상시점 합성과정에서 발생하는 비폐색 영역을 줄이

기 위해서 우리는 깊이값의 불연속성에 적응적인 저역 필터를 적용한다.

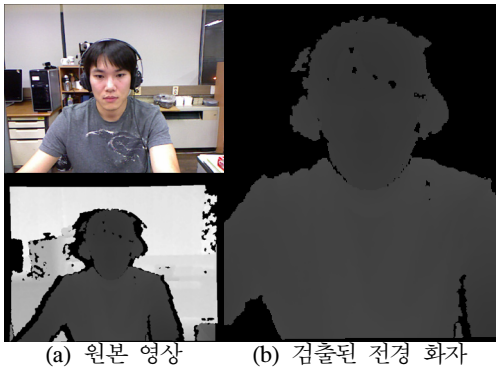
전처리 과정이 끝난 깊이 영상과 색상 영상을 이용해서, 화자를 표현하는 모든 화소들은 세계 좌표계로 투영된다. 그 다음, 세계 좌표계에 투영된 3차원 화소들은 삼각형 메쉬 형태로 구성된 화자의 3차원 모델을 복원하는데 사용된다. 원래의 시점에서 가상시점으로 시점을 변경한 다음 남아있는 빈 영역을 채우면 우리는 마침내 시선이 조정된 합성영상을 얻을 수 있게 된다.

Ⅲ. 깊이 영상 전처리 방법

3.1. 전경/배경 분리 방법

깊이 카메라 앞의 단일 화자를 인식하기 위해서 제안하는 시스템은 첫번째로 깊이 영상을 이용한 전경/배경 분리 방법을 사용한다. Kinect 깊이 카메라의 주된 특징 가운데 하나는 획득되는 깊이의 범위를 사용자가 임의로 설정할 수 있다는 것이다. 이렇게 가변적인 깊이 범위를 적절히 활용하면 특정 위치의 객체를 제거할 수 있다.

제안하는 시스템의 시나리오에서는 원격 영상회의를 위해 카메라에서 가장 가까운데 위치한 단일 화자만을 고려한다. 가장 가까운 화자 이외의 영역은 배경으로 간주하기 위해서 영상 내의 최소 깊이값을 찾은 다음, 깊이 범위를 1미터로 제한한다. 그림 2는 전경/배경 분리 결과를 보여준다. 그림 2(a)는 가공되지 않은 색상 영상과 깊이 영상을 나타내며, 그림 2(b)에서 알 수 있듯이, 제안하는 방법은 3차원 장면에서 깊이 범위를 제한함으로써 전경만을 검출해 낼 수 있다. 본 논문에서 깊이 영상은 0부터 255의 값을 가지도록 정규화 되어 있다.



(a) 원본 영상 (b) 검출된 전경 화자

그림 2. 전경/배경 분리 결과
Fig. 2. Result of foreground/background separation

3.2. 결합형 양방향 필터

깊이 카메라는 센서 잡음, 반짝이거나 어두운 색을 갖는 표면에서 깊이 검출 실패, 센서 송출부와 수신부의 시점 차이로 인한 폐색 영역 등의 카메라 자체의 문제점들로 인해 장면의 깊이값을 완벽하게 획득하지 못한다. 그림 2(a)의 깊이 영상에서 검정색으로 보이는 영역은 앞서 언급한 카메라 자체의 문제로 인한 것이다.

깊이값을 획득하지 못한 영역을 채우기 위해, 우리는 결합형 양방향 필터(Joint Bilateral Filter, JBF)를 사용한다[10]. 제안하는 시스템에서 JBF는 비어있는 깊이값을 채우기 위해 두 개의 Gaussian 분포, 즉, 색상 영상의 화소값 차이를 이용한 분포, 화소의 거리 차이를 이용한 분포를 사용한다. 깊이값 JBF는 다음과 같이 정의된다.

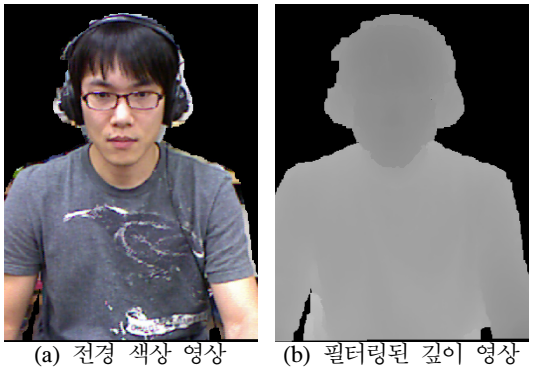
$$D(x,y)=\frac{\sum_{u\in u_p}\sum_{v\in v_p}W(u,v)\cdot D(u,v)}{\sum_{u\in u_p}\sum_{v\in v_p}W(u,v)} \tag{1}$$

$$W(u,v)=\begin{cases}0 & if\ D(u,v)=0\\ g_I(u,v)\cdot f(u,v) & otherwise\end{cases} \tag{2}$$

$$g_I(u,v)=\exp\left\{-\frac{|I(x,y)-I(u,v)|^2}{2\sigma_R^{\,2}}\right\} \tag{3}$$

$$f(u,v)=\exp\left\{-\frac{(x-u)^2+(y-v)^2}{2r^2}\right\} \tag{4}$$

여기서 $u_p=\{x-r,\cdots,x+r\}$, $v_p=\{y-r,\cdots,y+r\}$, r은 필터 반경을 나타낸다. 제안하는 방법에서 필터 표준편차와 반경은 σ_R=255, r=3으로 설정했다. JBF는 검정색으로 나타나는 영역에 대해서만 적용되며, 몇 번의 필터링



(a) 전경 색상 영상 (b) 필터링된 깊이 영상

그림 3. 결합형 양방향 필터링 결과
Fig. 3. Result of joint bilateral filtering

반복 과정을 거치면 비어있는 깊이값을 모두 채울 수

있게 된다. 그림 3은 결합형 양방향 필터를 적용한 결과를 보여준다. 그림 3(b)에서 알 수 있듯이, 비어있던 깊이값이 모두 채워진 것을 확인할 수 있다.

3.3. 불연속성에 적응적인 깊이 필터

그림 4는 가상 시점으로 3차원 워핑한 결과를 나타낸다. 그림 4에서 화자의 목 주변의 빈 영역은 가상시점에서 새롭게 드러난 영역이다. 이러한 영역을 비폐색 (disocclusion) 영역이라고 하며, 비폐색 영역에 대한 색상 정보를 가지고 있지 않기 때문에 이 부분을 채워주어야 한다.

그림 4. 화자의 3차원 워핑 결과
Fig. 4. 3D warping result of conferee

따라서, 제안하는 시스템에서는 3차원 워핑 이전에 깊이값에 적응적인 깊이 필터를 사용한다[11]. 객체 경계에서의 깊이값의 불연속성의 세기를 분석하고 필터링을 적용할 범위를 정하면, 깊이 영상의 변형을 최소화 할 뿐만 아니라 합성 영상의 화질 또한 향상시킬 수 있다. 불연속성에 적응적인 깊이 필터는 다음과 같이 정의된다.

$$D_{filtered}(x,y)=\alpha(x,y)\cdot D(x,y)+\{1-\alpha(x,y)\}\cdot D_{Gaussian}(x,y) \quad (5)$$

$$\alpha(x+u,y+v)=\begin{cases}\dfrac{|x-u|+|y-v|}{\delta(x,y)} & \text{if } |x-u|+|y-v|<\delta(x,y)\\ 1 & \text{otherwise}\end{cases} \quad (6)$$

$$D_{Gaussian}(x,y)=\sum_{v}\sum_{u}D_{original}(x-u,y-v)\cdot g(u,v) \quad (7)$$

$$g(u,v)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{|u-v|^2}{2\sigma^2}\right) \quad (8)$$

여기서 $D_{Gaussian}$은 필터링된 깊이 영상을 의미한다. u와 v의 범위는 $-D(x,y)\le u\le D(x,y)$, $-D(x,y)\le v\le D(x,y)$이다. 또한, 윈도우 크기는 필터의 표준 편차의 3배로 설정한다.

그림 5는 불연속성에 적응적인 깊이 필터링 결과를 나타낸다. 그림 5(a)에서처럼, 우리는 첫번째로 깊이 영상의 경계 정보를 추출하고 깊이 불연속성의 세기를 분석한다. 그 다음, 객체 경계 주변에서의 필터링 범위를 깊이값의 불연속성에 적응적으로 변화시킨다. 그림 5(b)는 필터링이 끝난 깊이 영상을 보여준다. 제안하는 시스템은 마침내 전처리가 끝난 깊이 영상을 얻을 수 있게 된다.

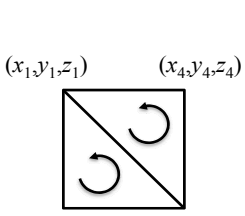
(a) 불연속성 영상

(b) 전처리된 깊이 영상

그림 5. 불연속성에 적응적인 깊이 필터링 결과
Fig. 5. Result of discontinuity-adaptive filtering

Ⅳ. 정면시점 영상합성 방법

제안하는 시스템에서 우리는 영상합성을 위해 화자를 삼각형 메쉬 모델의 형태로 표현하는 것에 중점을 둔다. 영상의 모든 화소들은 이러한 화자 모델 구성에 사용되며, 네 개의 이웃하는 화소들을 가지고 두 개의 삼각형을 생성한다. 네 개의 화소들은 3차원 워핑 과정을 통해 세계 좌표계로 투영이 되며 이들은 각각 x,y,z 좌표값을 가진다. 또한, 각각의 화소들은 각자의 색상 정보를 가지고 있으며, 삼각형 내부의 색은 각 꼭지점의 색상들로부터 선형 보간된 값으로 채워진다.



(a) 메쉬 구성방법

(b) 3차원 화자 모델

그림 6. 삼각형 메쉬 모델 구성 결과
Fig. 6. Result of mesh triangulation

그림 6(a)는 이웃하는 네 화소를 이용해서 삼각형 메쉬를 구성하는 과정을 나타내고 있다. 이 과정을 영상 전체 화소에 적용하게 되면 우리는 화자의 3차원 모델을 얻을 수 있다. 그림 6(b)는 3차원 모델의 확대된 부분을 보여준다.

3차원 모델 구성이 끝난 다음, 제안하는 시스템

은 시선이 조정된 영상을 합성하기 위해 가상 카메라의 위치를 변경한다. 다시 말해, 깊이 카메라의 광축과 원격의 화자의 시선을 맞춤으로써 원격의 화자에게 시선이 일치된 3차원 화자 모델을 보여줄 수 있게 된다. 최종적으로, 시선이 조정된 화자를 상대방에게 보여주게 되어 자연스러운 원격 영상회의를 가능하게 한다.

V. 실험 결과 및 분석

제안하는 시스템을 위해, 우리는 적외선 구조광 패턴을 통해 깊이값을 획득하는 Kinect 깊이 카메라를 사용했다. 깊이 카메라는 해상도 640×480에서 초당 30 프레임을 획득할 수 있으며 깊이 범위는 유동적으로 변한다. 하지만 사용자와 카메라 사이의 거리가 멀어질수록 깊이 센서의 정확도가 급격히 떨어지기 때문에 최대 깊이 범위를 1.5미터로 제한해서 실험을 진행했다.

(a) 색상 영상

(b) 전처리된 깊이 영상

(c) 시선이 조정된 화자 모델

그림 7. 시선 맞춤 결과
Fig. 7. Results of gaze correction

그림 7은 시선 맞춤 결과를 보여준다. 그림 7(a)와 그림 7(b)는 원본 색상 영상과 전처리가 완료된 깊이 영상을 나타낸다. 그림 7(c)에서 화자의 머리 주변에 경계 잡음이 발생했지만 화자가 정면을 바라보는 영상을 생성할 수 있었다. 또한, 깊이 영상을 이용해서 복잡한 배경을 쉽게 분리해낸 것을 확인했다. 그림 8은 화자의 얼굴을 확대한 영상이다. 그림 8(b)에서 알 수 있듯이, 그림 8(a)와 비교했을 때보다 자연스러운 시선으로 조정된 것을 볼 수 있었다.

(a) 원본 영상

(b) 시선이 조정된 영상

그림 8. 화자의 얼굴을 확대한 영상
Fig. 8. Zoom-in images of conferees' face

본 논문에서는 실시간 처리를 위해 그래픽 처리장치인 GPU 기반의 병렬 프로그래밍을 이용해서 시스템을 구현했다. GPU는 CPU와는 구조적인 차이로 수치 계산에 관련된 코어의 수가 월등히 많기 때문에 병렬 처리에 유리하다. 실제로 GPU 병렬 처리의 경우 단일 명령 복수 데이터 구조(single instruction multiple threads)를 지원하는데, 이것은 여러 화소에 대하여 동일한 명령을 줄 수 있어 영상처리에 적용이 용이한 장점이 있다. 병렬 프로그래밍을 이용해서 제안한 시스템을 구현한 결과, 제안한 시스템의 종합 수행 시간은 약 22.73 frame/s였다.

VI. 결 론

본 논문에서는 깊이 카메라를 이용해서 시선 맞춤 영상을 생성하는 새로운 방법을 제안했다. 제안하는 시스템은 다양한 영상처리 기법들, 전경/배경 분리, 결합형 양방향 필터링, 불연속성 적응적 깊이 필터링 등을 사용했다. 깊이 카메라로부터 획득한 색상 영상과 전처리된 깊이 영상을 이용해서 화자는 삼각형 메쉬 기반의 3차원 모델로 복원됐고, 카메라의 광축과 원격 화자의 시선을 일치시킴으로써, 우리는 시선이 조정된 영상을 합성할 수 있었다. 실험 결과를 통해, 자연스러운 시선 맞춤 영상이 합성되는 것을 확인했으며, 깊이 카메라와 디스플레이만을 필요로 하는 제안하는 시스템의 특성상, 다양한 응용 분야에 사용될 수 있을 것으로 기대한다.

참 고 문 헌

[1] D. Sharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," IEEE Workshop on Stereo and Multi-Baseline Vision, pp. 131-140, Dec. 2001.

[2] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," SIGGRAPH'04, pp. 600-608, Aug. 2004.

[3] D. Scharstein, and R. Szeliski, "High-accuracy stereo depth maps using structured light," Computer Vision and Pattern Recognition Workshops, vol. 1, pp. 195-202, June 2003.

[4] S. Kim, S. Lee, and Y. Ho, "Three-dimensional natural video system based on layered representation of depth maps," IEEE Transactions on Consumer Electronics, vol. 52, no. 3, pp. 1035-1042, Aug. 2006.

[5] E. Lee and Y. Ho, "Generation of multi-view video using a fusion camera system for 3D displays," IEEE Transactions on Consumer Electronics, vol. 56, no. 4, pp. 2797-2805, Nov. 2010.

[6] L. Xia, C. Chen, and J. K. Aggarwal, "Human detection using depth information by Kinect," Computer Vision and Pattern Recognition Workshops, pp. 15-22, June 2011.

[7] Redert, M. O. Beeck, C. Fehn, W. IJsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, I. Sexton, P. Surman, "ATTEST: Advanced Three-dimensional Television System Techniques," International Symposium on 3D Data Processing, pp. 313-319, June 2002.

[8] O. Schreer, N. Atzapadin, and I. Feldmann, "Multi-baseline disparity fusion for immersive videoconferencing," International Conference on Immersive Telecomm., pp. 27-29, May 2009.

[9] S. Lee, I. Shin, and Y. Ho, "Gaze-corrected view generation using stereo camera system for immersive videoconferencing," IEEE Transactions on Consumer Electronics, vol. 57, no. 3, pp. 1033-1040, Aug. 2011.

[10] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," SIGGRAPH'07, pp. 96-100, Aug. 2007.

[11] S. Lee and Y. Ho, "Discontinuity-adaptive depth map filtering for 3D view generation," International Conference on Immersive Telecomm., pp. T8(1-6), 2009.

이 상 범 (Sang-Beom Lee) 정회원

2004년 경북대학교 전자전기공학부 졸업(학사)

2006년 광주과학기술원 정보통신공학과 졸업 (석사)

2006년~현재 광주과학기술원 정보통신공학과 박사과정

<관심분야> 3차원 TV, 실감방송, 3차원 비디오 부호화

호 요 성 (Yo-Sung Ho) 정회원

1981년 서울대학교 공과대학 전자공학과 졸업(학사)

1983년 서울대학교 대학원 전자공학과 졸업(석사)

1989년 Univ. of California, Santa Barbara, Dept. of Electrical and Computer Engineering.(박사)

1983년~1995년 한국전자통신연구소 선임연구원

1990년~1993년 미국 Philips 연구소, Senior Research Member

1995년~현재 광주과학기술원 정보통신공학과 교수

<관심분야> 디지털 신호처리, 영상 신호 처리 및 압축, 디지털 TV와 고선명 TV, 멀티미디어 시스템, MPEG 표준, 3차원 TV, 실감방송