Journal of The Institute of Electronics and Information Engineers Vol.52, NO.7, July 2015 http://dx.doi.org/10.5573/ieie.2015.52.7.093
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2015-52-7-11

실감형 화상 회의를 위해 깊이정보 혼합을 사용한 시선 맞춤 시스템

(Eye Contact System Using Depth Fusion for Immersive Videoconferencing)

장 우 석*, 이 미 숙**, 호 요 성***

(Woo-Seok Jang©, Mi Suk Lee, and Yo-Sung Ho)

요 약

본 논문에서는 실감형 원격 영상회의를 위한 시스템을 제안한다. 원격 영상회의에서 카메라는 보통 디스플레이의 중앙이 아닌 측면에 설치가 된다. 이는 시선 불일치를 만들고, 사용자들의 몰입도를 떨어뜨린다. 따라서 실감형 영상회의에 있어서 시선 맞춤은 중요한 부분을 차지한다. 제안하는 방법은 스테레오 카메라와 깊이 카메라를 사용하여 시선 맞춤을 시도한다. 깊이 카메라는 비교적 적은 비용으로 효율적으로 깊이 정보를 생성할 수 있는 키넥트 카메라를 선택하였다. 하지만 키넥트 카메라는 비용적인 장점에도 불구하고 단독으로 사용하기에는 내제하는 단점이 많다. 따라서 스테레오 카메라를 더하여 각 깊이 센서 간의 단점을 보완하는 방법을 개발하였고, 이는 각 깊이 정보 간의 혼합 및 정제 과정을 통해서 실현된다. 시선 맞춤 영상 생성은 후처리를 통한 보완된 깊이 정보를 이용하여 3차원 워핑 기술을 이용하여 구현된다. 실험결과를 보면 제안한 시스템이 자연스러운 시선 맞춤 영상을 제공하는 것을 알 수 있다.

Abstract

In this paper, we propose a gaze correction method for realistic video teleconferencing. Typically, cameras used in teleconferencing are installed at the side of the display monitor, but not in the center of the monitor. This system makes it too difficult for users to contact each eyes. Therefore, eys contact is the most important in the immersive videoconferencing. In the proposed method, we use the stereo camera and the depth camera to correct the eye contact. The depth camera is the kinect camera, which is the relatively cheap price, and estimate the depth information efficiently. However, the kinect camera has some inherent disadvantages. Therefore, we fuse the kinect camera with stereo camera to compensate the disadvantages of the kinect camera. Consecutively, for the gaze-corrected image, view synthesis is performed by 3D warping according to the depth information. Experimental results verify that the proposed system is effective in generating natural gaze-corrected images.

Keywords : Depth camera, eye contact, gaze correction, stereo matching, videonferencing.

Ⅰ. 서 론

영상회의는 통신기술의 하나로서 어느 장소에 있던지 같은 장소에 있는 것처럼 대화 할 수 있게 하는 기술이다. 최근 들어, 영상회의 시스템은 텔레비전, 컴퓨터 그리고 휴대폰과 같이 일반인들의 접근이 쉬운 전자기기 기반으로 개발되고 있다[1]. 최근에는 기술 발달로

* 학생회원, *** 평생회원, 광주과학기술원 정보통신공학부
(Gwangju Institute of Science and Technology, School of Information and Communications)
** 정회원, 한국전자통신연구원
(Electronics and Telecommunications Research Institute)
© Corresponding Author(E-mail: jws@gist.ac.kr)
Received ; November 11, 2014 Revised ; April 2, 2015
Accepted ; March 26, 2015

큰 디스플레이를 기반으로 한 영상 시스템이 제작되고 있다. 하지만, 기존의 시스템에서는 카메라와 두 눈 사이의 거리 차이를 발생시켜 상호간의 시선 맞춤을 어렵게 한다[2]. 디스플레이의 크기가 큰 경우 화자가 디스플레이의 정면을 주시하더라도 카메라에 획득된 화자는 다른 곳을 쳐다보는 것처럼 영상이 생성되기 마련이다. 만약 상호간의 시선 맞춤을 하지 않고 대화를 하면 정확한 의사전달이 힘들고 서로의 이야기에 집중하기 어려워진다. 때문에 화자간의 시선 맞춤은 영상회의에서 가장 중요하게 다루어져야할 문제이다[3].

깊이 정보 예측과 영상 합성은 3차원 비디오 시스템에서 중요한 기술이다[4]. 이러한 기술은 시선 맞춤뿐 아니라 다양한 응용에서 사용이 가능한 기술이다. 영상 합성이란 존재하지 않는 가상의 카메라 시점에 새로운 영상을 생성하는 방법으로, 3차원 콘텐츠의 발전으로 인하여 영상 합성 기술은 더욱 중요하게 되었다. 특히, 대상 시점에 색상 영상과 깊이 영상을 투영시켜 가상의 영상을 만드는 기술인 깊이 기반 렌더링은 가장 널리 사용되는 방법이다[5]. 깊이 영상은 카메라와 객체의 거리를 나타내고 일반적으로 색상 영상과 대응하는 화소의 깊이 정보를 제공한다[6].

깊이 정보를 획득하는 방법은 크게 두 가지로 나누어진다. 능동형과 수동형 센서를 이용하여 깊이를 예측하게 된다. 능동형 방식은 적외선, 레이저 그리고 빛의 패턴과 같은 물리적 센서를 사용하여 깊이 정보를 직접적을 측정한다. 깊이 카메라, 구조광 3차원 센서 그리고 3차원 스캐너는 능동형 방식으로 사용된다. 일반적으로, 능동형 센서는 수동형 센서보다 효율적으로 보다 좋은 품질의 깊이 영상을 제공한다. 하지만, 능동형 센서는 수동형 센서보다 작은 해상도의 영상을 제공하고, 수동형 센서에 비해 비싸다. 반대로 수동형 센서는 두 대 이상으로부터 촬영된 2차원 영상들을 사용하여 깊이 정보를 예측한다[8]. 스테레오 정합은 수동형 센서를 사용하여 깊이를 예측하는 가장 대표적인 방법이다[9]. 수동형 센서를 사용하는 장점은 시스템을 구축하는데 능동형 센서에 비해 적은 비용과 다양한 해상도의 깊이 영상을 제공한다. 하지만, 수동형 센서 또한 영상에 특징이 없거나 패턴이 반복되는 영역에서 잘못된 깊이 정보를 계산하게 된다.

최근 들어, 양안식 혹은 다시점 카메라와 깊이 카메라를 함께 이용하여 각 방법이 가지고 있는 장점을 통해 단점을 보완하는 방식인 혼합형 방식이 이용되어지고 있다[7, 10]. 하지만, 혼합형 시스템을 구축하기 위한 높은 비용 때문에, 일반적으로 소비자들에게 적합하지 않다. 최근 들어, 비용 부담을 줄이기 위하여 키넥트와 같은 적외선 구조광 패턴 카메라가 소개되었다[11]. 하지만 적외선 구조광 패턴 깊이 카메라를 사용하여 얻은 깊이 정보는 부정확한 깊이 정보와 많은 잡음과 폐쇄 영역을 포함한다.

본 논문은 다양한 깊이 정보 혼합을 이용한 시선 맞춤 방법을 제안한다. 새로운 시스템을 구축하여 장점은 강화하고 깊이 센서의 약점을 극복한다. 게다가, 깊이 융합과 영상 합성을 활용하여 시선 맞춤 영상을 만든다. 본 논문의 구성은 다음과 같다. Ⅱ장에서는 시선 맞춤을 위한 시스템을 소개하고, Ⅲ장에서는 시선 맞춤 알고리즘을 소개한다. Ⅳ장에서는 제안하는 방법의 실험 결과를 분석하고, Ⅴ장에서는 논문의 결론을 맺는다.

Ⅱ. 시스템 개요

제안한 시스템은 55인치 크기의 디스플레이를 사용하고, 사용자와 텔레비전의 거리는 약 2m~2.5m의 거리를 두었다. 그림 1은 제안한 시스템 구조를 보여준다. 제안한 시스템은 두 대의 스테레오 카메라와 한 대의 키넥트 카메라를 사용한다. 가상의 중간시점에서 효율적으로 객체를 합성하기 위하여 두 대의 색상 카메라는 각각 디스플레이의 중간 시점에 수평한 높이의 왼쪽과 오른쪽에 부착한다.

중간 시점에 합성을 할 때, 택스쳐가 없는 부분은 스테레오 정합으로는 정확한 깊이 정보를 얻기 어렵다. 게

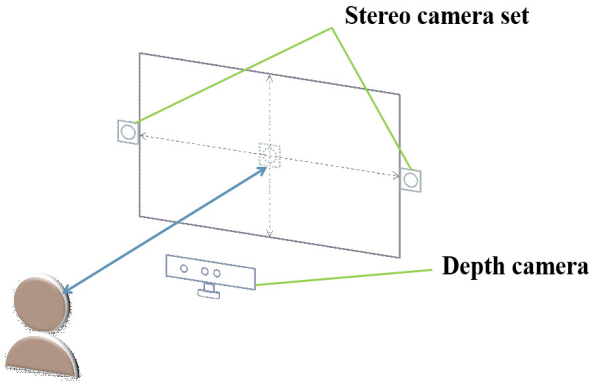


그림 1. 제안된 눈 맞춤 시스템
Fig. 1. Proposed gaze correction system setup.

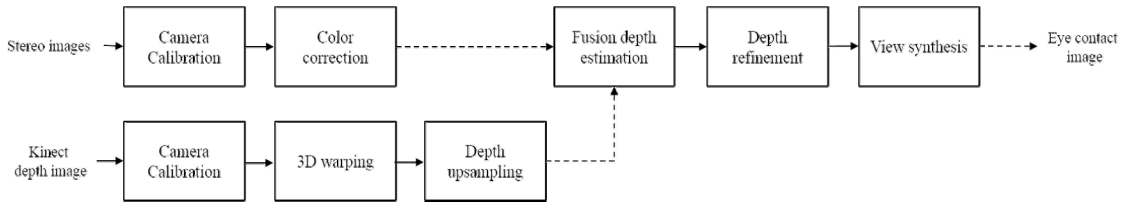


그림 2. 눈 맞춤 알고리즘을 위한 흐름도
Fig. 2. Overall framework of our algorithm for gaze correction.

다가 두 대의 카메라의 시차로 인한 폐쇄영역의 발생 또한 하나의 큰 문제점이다. 키넥트 카메라를 사용하여 스테레오 정합으로는 해결할 수 없는 문제점을 보완한다. 키넥트 카메라는 택스쳐가 없는 부분의 깊이 값을 보완하고 폐쇄영역을 줄임으로서 깊이 정보의 정확성을 높인다.

그림 2는 제안한 방법의 전체적인 흐름을 보여준다. 스테레오 카메라로 깊이 정보를 예측하는 동안, 키넥트에서 얻어진 깊이 정보를 활용하여 좌, 우 카메라에 3D 워핑과 업샘플링을 한다. 업샘플링된 깊이 영상은 스테레오 카메라로 예측한 깊이 정보와 함께 사용하여 좌, 우 깊이 영상의 정확도를 높인다. 하지만, 잘못된 정보로 인해 잡음이 생성되는데, 잡음을 제거하기 위하여 필터를 사용하여 깊이 정보를 개선한다. 그리고 이 깊이 정보를 이용하여 중간 시점에 합성 영상을 생성한다.

Ⅲ. 시선 맞춤 과정

1. 전처리

두 대 혹은 두 대 이상의 카메라를 사용하게 될 때, 카메라끼리 연관된 정보가 필요하다. 카메라 캘리브레이션을 통해 각각의 카메라의 연관된 파라미터를 얻는다[12]. 카메라 파라미터는 세 개의 요소로 구성되어 있는데, 내부 파라미터, 외부 파라미터, 이동 백터로 구성된다. 제안한 방법은 카메라 파라미터를 사용하여 키넥트 카메라에서 얻어진 깊이 정보를 3D 워핑하고, 스테레오 카메라로 예측한 깊이 정보와 합하여 시선 맞춤 영상을 합성한다.

일반적으로, 영상 보정은 스테레오 정합을 하기 위한 필수적인 요소이다. 영상 보정은 영상의 에필폴라 라인을 평행하게 만드는 방법으로, 평행한 에필폴라 라인이 임의의 동일한 평면 위에 위치하게 한다[13]. 영상 보정을 통하여 수평선 위의 대응점을 찾을 수 있다. 비록 영상 보정 방법이 1차원의 평행한 정렬에서는 유용하나, 2차원 이상의 정렬에서는 문제가 된다. 영상 평면의 변환으로 인해 본래 영상에 왜곡이 생기게 된다[14]. 하지만, 제안한 방법은 수렴형 카메라 배열을 사용하기 때문에 영상 보정 과정이 생략되어, 영상의 왜곡이 발생하지 않게 된다.

2. 키넥트 카메라 처리

키넥트로 얻어진 깊이 정보는 스테레오 정합으로 얻어진 깊이 정보를 개선한다. 색상에 관련된 좌표로 깊이 정보를 3차원 워핑한다[15]. 3차원 워핑은 두 단계로 구성된다. 키넥트 깊이 정보가 3차원 공간으로 역투영되고, 다시 이 정보가 대상 시점으로 투영되는 과정이다. 그림 3은 키넥트의 깊이 정보 워핑을 보여준다.

다음으로, 키넥트 카메라에서 얻어진 깊이 정보를 업샘플링하여 저해상도의 영상을 보간한다. 그림 4는 두 영상의 해상도의 차이를 보여준다. 키넥트 카메라로 얻어진 깊이 정보는 고해상도의 깊이 영상을 얻을 때 깊이 카메라의 특성으로 인하여 어려움이 발생한다. 이러한 문제를 해결하기 위하여 결합형 양방향 업샘플링 방법(joint bilateral upsampling)을 사용하였다[16]. 이 방법

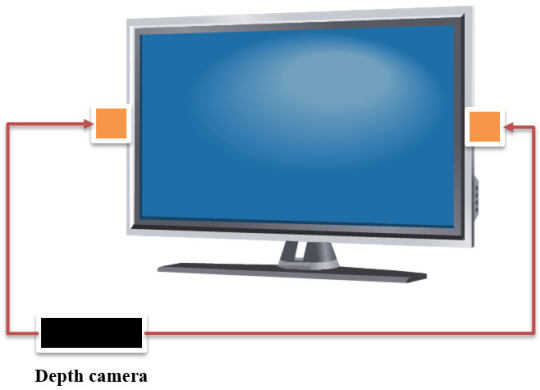


그림 3. 키넥트 카메라 워핑
Fig. 3. 3D warping for depth camera.

그림 4. 깊이 카메라 업샘플링
Fig. 4. Upsampling for depth camera.

은 고해상도 색상 영상과 저해상도 깊이 영상이 있을 때 색상 영상의 정보를 활용하여 깊이 영상의 해상도를 키우는 방법이다.

3. 혼합 깊이 영상 획득 및 시선 맞춤

본 논문에서는 전역적 스테레오 정합방법을 사용하여 좌, 우 색상 카메라의 깊이 정보를 예측한다. 키넥트에서 얻어진 깊이 정보는 전역적 에너지 함수에 포함되어 보다 정확한 깊이 값을 예측하도록 한다. 에너지 함수는 다음과 같은 식으로 표현된다.

$$E = \sum_{x,y} (E_{data} + E_{smooth}) \quad (1)$$

E_{data}는 데이터 항으로 색상의 유사성을 구하고, E_{smooth}는 평활화 항으로 주변 픽셀들과의 변위 변화도를 구한다. 본 논문에서는 키넥트에서 얻어진 깊이 정보를 데이터 항에 추가 하여 정확도를 얻는다. 다음에 나오는 식은 키넥트 깊이 정보가 더해진 데이터 항을 보여준다.

$$E_{data}(x,y,d) = 0.5*(|I_L(x,y) - I_R(x',y',d)| + |d - d_{SLP}|) \quad (2)$$

$I_L(x, y)$는 좌 영상에서의 각각의 화소 값을 나타낸다. $I_R(x', y', d)$은 좌 영상의 한 점과 대응하는 우 영상의 한 점의 화소 값이다. d_{SLP}은 업샘플링된 키넥트 데이터 값이다. 깊이 정보는 다양한 깊이 정보의 변화가 있을 때도 정확한 값을 얻게 한다. 제안한 방법은 3D 워핑을 사용하여, 대응점인 $I_R(x', y', d)$을 찾는다[14]. 평활화 항은 주변 픽셀들과의 변위 변화도를 구하는데 아래의 식으로 표현된다.

$$E_{smooth} = \sum_{t \in N(x,y)} | d - d_t | \quad (3)$$

$N(x, y)$는 현재 화소에 이웃한 화소들을 나타낸다. 알고리즘은 계층적으로 수행하여 텍스처가 없는 영역에서도 좀 더 정확한 깊이 정보를 획득한다. 게다가, 잘못 정합된 화소와 불연속성에 대한 후처리를 하여 품질을 올린다[14].

혼합형 깊이 획득 방법으로 얻어진 깊이 정보는 영상 합성을 위해 사용된다. 영상 합성을 사용하여 시선 맞춤이 된 영상을 얻게 된다. 영상이 합성될 위치는 좌우 색상카메라의 카메라 파라미터의 중간값을 이용하여 구한다[27].

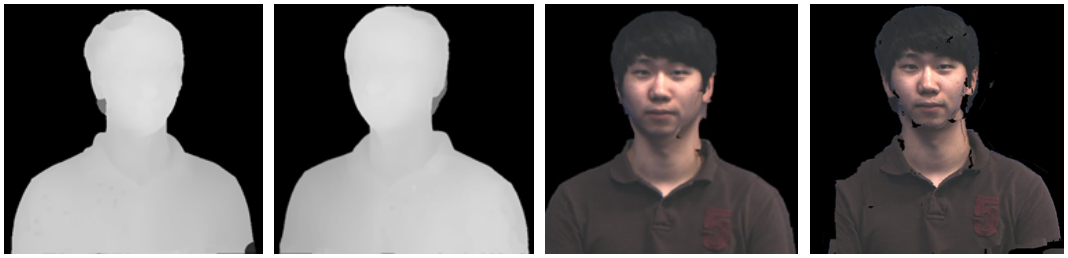
Ⅳ. 실험 결과

제안한 방법을 평가하기 위하여 네 가지 실험 영상을 사용하였다. 피 실험자들은 1920*1080의 해상도로 촬영되었고, 깊이 영상은 640*480의 해상도로 촬영되었다. 키넥트 카메라는 16-bit의 bin으로 거리를 표현한다. 정확한 정보를 얻기 위해 카메라 캘리브레이션과 색상 보정 방법을 사용하였다[26]. 그림 5는 원본 영상과 워핑된 깊이 영상에 JBU를 적용한 깊이 영상을 보여준다. 워핑된 깊이 영상의 경계는 비록 정확하게 보이지 않지만, 얼굴 중심으로는 정확한 깊이 정보를 얻은 것을 확인하였다. 그림 6과 그림 7의 결과는 혼합형 깊이 정보에 후처리를 통해서 잘못 처리된 깊이 정보를 보정한 결과 및 스테레오 매칭에 기반하여 시선맞춤을 한 결과를 보여준다[1]. 스테레오 매칭에 기반한 결과는 혼합형 깊이 정보를 이용하여 얻은 제안하는 방법의 결과와 비교하기 위해서 제시되었다. 결과를 통해서 제안된 방법이 깊이 정보를 더 정밀하게 표현하는 것을 볼 수 있다. 게다가 제안된 방법이 텍스쳐가 없는 얼굴 부분에서도 깊이 정보를 향상 시킨 것을 확인 할 수 있다. 특히, 텍스쳐가 없는 넓은 부분에서 이러한 결과를 더 잘 확인할 수 있다. 합성 결과를 보면 스테레오 정합 기반 방식은 텍스처가 부족한 부분에서의 정확한 영상을 생성하지 못했고, 두 영상간의 일치하는 점을 찾지 못한 부분에 대해서는 정확한 합성을 하지 못해서 중간 얼굴 부분과 어색하게 결합되는 것을 보여준다. 결과적으로 혼합형 깊이 예측 방법이 시선 맞춤 영상을 만드는데 매우 효과적이 라는 것을 알 수 있다.

(a) (b) (c) (d)

그림 5. 워핑된 깊이 정보의 업샘플링 결과.
(a) 좌영상 (b) 우영상 (c) 업샘플된 좌영상 (d) 업샘플된 우영상

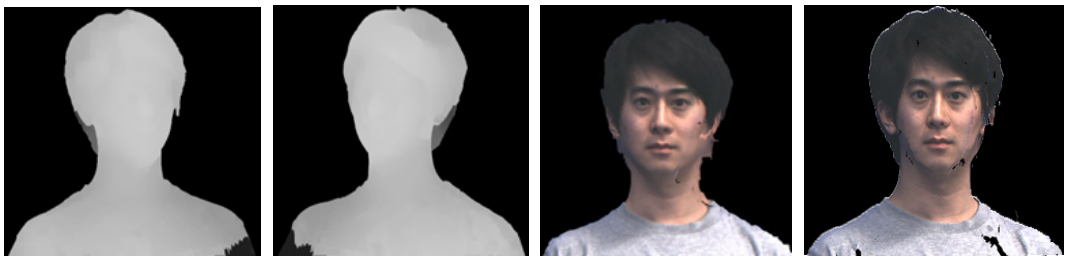
Fig. 5. Upsampling results of warped depth.
(a) Left image (b) Right image (c) Upsampled left image (d) Upsampled right image



(a) (b) (c) (d)

그림 6. 제안한 알고리즘 결과 1.
(a) 좌영상 (b) 우영상 (c) 시선 맞춤 합성 (d) 스테레오 정합 기반 합성

Fig. 6. Result 1 of the proposed method 1.
(a) Left image (b) Right image (c) Gaze-corrected view synthesis
(d) Stereo matching based view synthesis



(a) (b) (c) (d)

그림 7. 제안한 알고리즘 결과 2.
(a) 좌영상 (b) 우영상 (c) 시선 맞춤 합성 (d) 스테레오 정합 기반 합성

Fig. 7. Result 2 of the proposed method.
(a) Left image (b) Right image (c) Gaze-corrected view synthesis
(d) Stereo matching based view synthesis

V. 결 론

본 논문은 깊이 센서들에 내재하는 문제점을 해결하여 시선을 맞추는 방법을 제안하였다. 본 논문에서 가장 중요한 기여는 스테레오 비전에서의 오류 확률을 줄이는 경제적인 시스템의 구성하는 것이다. 최근에 많이 사용되는 대형 화면에서는 카메라들 사이에 정확한 일치점을 찾기 어려워서 정확한 깊이 정보를 얻기 힘들다. 이는 시선 맞춤 영상을 생성하기 어렵게 한다. 따라서 우리는 폐색 영역을 줄이는데 도움이 되는 키넥트 카메라를 스테레오 카메라 시스템에 포함하여 깊이 혼합 시스템을 구성하였다. 키넥트 카메라 정보는 자연스러운 시선 맞춤을 어렵게 하는 깊이의 비상세함을 보완하여 깊이 정보의 정밀도를 높인다. 제안하는 방법에서 시스템 구성으로부터 완전히 해결되지 못하는 문제는 깊이 예측 알고리즘에 의해서 해결하였다. 깊이 보정을 포함하는 혼합 깊이 예측은 각 깊이 센서의 약점을 보완한다. 제안하는 방법의 실험 결과는 시선 맞춤의 성능이 기존의 스테레오 카메라를 이용한 방법에 비해서 향상됨을 보여준다.

REFERENCES

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

[2] C. Kuster, T. Popa, J. C. Bazin et al., “Gaze correction for home video conferencing,” ACM Transaction on Graphics, vol. 31, no. 6, pp. 1-6, 2012.

[3] J. Zhu, R. Yang, and X. Xiang, “Eye contact in video conference via fusion of time-of-flight depth sensor and stereo,” 3D Research, vol. 2, no. 3, pp. 1-10, 2011.

[4] Video and Requirement Group, “Call for Proposals on 3D Video Coding Technology ” N12036, ISO/IEC JTC1/SC29/WG11, 2011.

[5] L. Zhang, and T. Wa James, “Stereoscopic image generation based on depth images for 3D TV,” IEEE Transactions on Broadcasting, vol. 51, no. 2, pp. 191-199, 2005.

[6] C. Fehn, and R. S. Pastoor, “Interactive 3-DTV-Concepts and Key Technologies,” Proceedings of the IEEE, vol. 94, no. 3, pp. 524-538, 2006.

[7] E. K. Lee, and Y. S. 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, 2010.

[8] W. S. Jang, and Y. S. Ho, “Efficient disparity map estimation using occlusion handling for various 3D multimedia applications,” IEEE Transactions on Consumer Electronics, vol. 57, no. 4, pp. 1937-1943, 2011.

[9] D. Scharstein, R. Szeliski, and R. Zabih, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms.” pp. 131-140.

[10] S. Y. Kim, S. B. Lee, and Y. S. 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, 2006.

[11] L. Xia, C. C. Chen, and J. K. Aggarwal, “Human detection using depth information by Kinect.” pp. 15-22.

[12] Z. Zhang, “A flexible new technique for camera calibration,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 11, pp. 1330-1334, 2000.

[13] Y. S. Kang, and Y. S. Ho, “An efficient image rectification method for parallel multi-camera arrangement,” IEEE Transactions on Consumer Electronics, vol. 57, no. 3, pp. 1041-1048, 2011.

[14] W. S. Jang, and Y. S. Ho, “Efficient depth map generation with occlusion handling for various camera arrays,” Signal, Image and Video Processing, vol. 8, no. 2, pp. 287-297, 2014/02/01, 2014.

[15] Y. S. Kang, and Y. S. Ho, “High-quality multi-view depth generation using multiple color and depth cameras.” pp. 1405-1410.

[16] J. Kopf, M. F. Cohen, D. Lischinski et al., “Joint bilateral upsampling,” ACM Trans. Graph., vol. 26, no. 3, pp. 96, 2007.

[26] U. Fecker, M. Barkowsky, and A. Kaup, “Improving the Prediction Efficiency for Multi-View Video Coding Using Histogram Matching.” pp. 2-17.

[27] R. Hartley, and A. Zisserman, Multiple View Geometry in Computer Vision, Second ed.: Cambridge University Press, 2004.

저 자 소 개

장 우 석(학생회원)

2007년 전남대학교 전자정보통신공학부 학사 졸업.
2009년 광주과학기술원 정보통신공학부 석사 졸업.
2009년 9월~현재 광주과학기술원 정보통신공학부 박사과정

<주관심분야 :디지털 영상처리, 컴퓨터 비전, 실감 방송>

호 요 성(평생회원)

1981년 서울대학교 전자공학과 학사 졸업.
1983년 서울대학교 전자공학과 석사 졸업.
1983년 3월~1995년 9월 한국전자통신연구원 선임연구원
1989년 University of California, Santa Barbara Department of Electrical and Computer Engineering 박사 졸업
1990년 1월~1993년 5월 미국 Philips 연구소 Senior Research Member
1995년 9월~현재 광주과학기술원 정보통신공학부 교수

<주관심분야 : 디지털 신호처리, 영상신호 처리 및 압축, 디지털 TV와 고선명 TV, MPEG 표준, 다시점 비디오 부호화, 실감방송>

이 미 숙(정회원)

1991년 호서대학교 전자공학과 학사 졸업.
1993년 호서대학교 전자공학과 석사 졸업.
2001년 KAIST 전기전자공학과 박사 졸업.
2002년 2월~현재 한국전자통신연구원

<주관심분야 : 디지털 음성 코딩, 영상처리, 실감형 텔레프레즌스>