# SVTE를 기반으로 한 가상 공간에서의 비디오 컨퍼런스 설계에 관한 연구

김태훈* · 김남효* · 김선우* · 최연성*

*군산대학교

# A study for video-conference architecture in Virtual Reality based on SVTE

Tae-hun Kim* · Nam-hyo Kim* · Sun-woo Kim* · Yeon-sung Choi*

*Kunsan National University

E-mail : thl79@kunsan.ac.kr

## 요 약

비디오 컨퍼런스는 시간적, 공간적으로 분산된 그룹의 의사소통 및 공동작업에 매우 유용하게 활용되고 있다. 이는 이동에 소요되는 시간 및 경제성에서 특히 강점을 가진다. 본 논문에서는 SVTE(Shared Virtual Table Environment)를 기반으로 하여, 3D 비디오와 Virtual Reality 기술을 결합한 차세대 비디오 컨퍼런스 시스템을 제안한다. 또한 서로 공유된 공간내에서, 평면상의 2D 화면보다 더욱 사실감 있는 3D 영상을 제공하는 비디오 컨퍼런스 시스템의 구조를 보이고, 제안된 시스템에서 고려되어야 할 영상처리 기법을 설명한다.

## 키워드

Video Conference, VR, SVTE, VIRTUE

## Ⅰ. 서 론

비디오 컨퍼런스는 10여년 전에 첫 번째 시스템을 선보인 후, 현재 전세계적으로 다양한 응용분야로 확대, 발전하고 있다. 최근 네트워크와 비디오 압축 기술이 급속도로 발전된 후로, 비디오 컨퍼런스는 더욱 넓은 응용분야로 다양하게 활용되고 있으며, 특히 원격강의, 원격회의 또는 원격 공동작업과 같은 다른 응용을 위한 현장감 있는 화면과 함께 고품질 오디오와 영상을 제공한다.

그럼에도 불구하고, 이러한 시스템은 자연적인 인간중심의 통신을 제공하는 데에는 여전히 한계가 있다. 이러한 시스템의 대부분은 모니터와 같은 디스플레이 장치를 기반으로 한 단일 화면이다. 이러한 영상은 회의 참가자에게 MPEG-2 또는 H.263과 같은 진부한 2D 비디오의 영상 및 음성과 같은 제한적인 정보만을 주게 된다. 결국 사용자의 몸짓이나, 정교한 움직임, 참가자의 시선 및 주위 환경의 보다 사실감 있는 소리와 같은 인간중심의 Face-to-Face 통신의 대부분은 재생될 수 없다. 이는 상당히 복잡하고도 어려운 기능이다. 사실, 그러한 특징을 지원한다는 것은 모든 참가자들의 개개인의 3D 위치로부터 얻을 수 있는 모든 정보를 얻어야 만이 가능하다. 이러한 어려움은 VR(Virtual Reality)의 개념을 통하여 해결할 수 있다. 이것은 가상 카메라 또는 가상 공간 오디오 렌더링과 같은 기능을 제공한다.

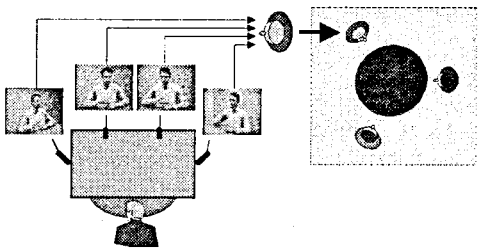
VR에 관한 개발은 최근 몇 년 이래 텔레통신에 관한 연구의 주요 이슈가 되었다. 이는 tele-immersion 기술의 진보를 이루었다. 지리적으로 분산된 사용자는 3D VR 세계에서 다른 참가자들과 함께 사실적인 상호작용을 가능하게 하였다. 특히, 비디오 컨퍼런스는 SVTE(Shared Viturual Table Environment) 내에서 더욱이 그 효과를 증대 시킬 수 있다.

본 논문의 2장에서 우리는 VIRTUE(Virtual Team User Environment)라 불리우는 SVTE의 개념과 이를 이용한 가상회의를 선보인다. 또한 3장에서는 SVTE를 이용하여 3D 비디오와 VR 기술을 결합한 차세대 비디오 컨퍼런스 시스템의 구조를 보이며, 그 다음 4장에서는 이러한 비디오 컨퍼런스에서 고려되어야 할 배경분할, 헤드 트래킹, 참가자의 얼굴분석 및 가상 장면내에서의 복합영상처리 기술을 보인다. 그리고, 마지막으로 제안한 시스템의 문제점 전망으로 결론을 맺도록 하겠다.

## II. 개 념

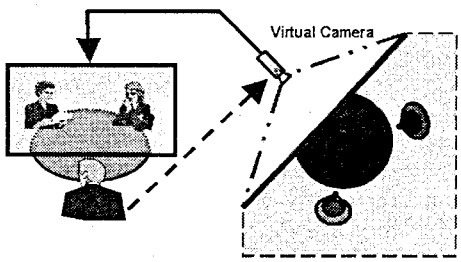
### 2-1 SVTE의 개념

SVTE(Shared Virtual Table Environment)는 가상의 테이블을 서로 공유하여 회의하는 환경을 말한다.[1] 이러한 공유된 가상 환경 개념의 기본적인 아이디어는 공유된 가상 환경 내에서 미리 정의된 장소에 위치한 참가자들의 3D 비디오 재생이다. 이러한 목적을 위하여 참가자들은 그림 1에서 보는 것과 같이 4개 이상의 다중 카메라에 의하여 영상을 수신하여 각각의 단말에 캡쳐되고, 수신된 영상은 3D 화면으로 보여질 수 있는 하나의 단일 이미지로 합성된다.



<그림 1. 3-그룹 회의를 위한 설정>

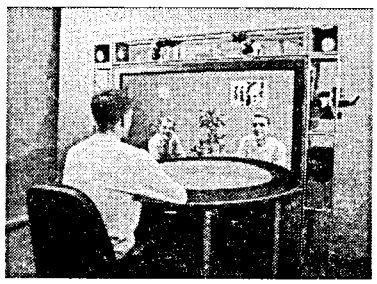
이때, 모든 참가자들의 3D 영상 객체는 가상적으로 공유된 테이블 주위에 그룹화된다. 이것은 각각의 각도에 적절히 수신된 영상으로 데이터를 합성하는 것으로 해결할 수 있다. 3-그룹 회의의 경우에는 등변의 삼각형으로부터, 4-그룹의 경우에는 정사각형의 영상으로부터 구할 수 있다. 이것은 미리 회의에 참가할 참가자들의 수에 따라 그 설정이 달라진다. 이러한 설정에 따라 SVTE는 장면 조합과 3D 비디오 스트림을 각각 달리 설계할 수 있다.




<그림 2. 가상 3D 회의 장면의 렌더링>

이러한 일반적인 환경 구성에 기반하여, 가상 회의 환경의 관점은 그림 2에 보이는 것과 같이, 가상 카메라를 사용하여 표현된다. 위치상으로 가상 카메라의 위치는 head tracker에 의하여 등록된 참가자들의 머리의 위치와 함께 동시에 이동한다. 그러므로 다중 뷰 캡쳐 장치, 가상 장면, 가상카메라와 같은 요소는 각각 다른 것들과 잘 조화된다. 그것은 모든 참가자들이 적절한 장면을 볼 수 있도록 한다.

이러한 요소들의 조화는 서두에 언급되었듯이, 참가자들을 보이게 하는데에 있어서 보다 현실감 있는 특성을 제공한다. 시선을 마주친다거나, 상대에게 제스처를 취한다거나하는 등의 누가 누구를 또는 누가 무엇에 포인팅하여 대화하는지를 판단할 수 있게 한다.



<그림 3. SVTE 내에서의 가상회의>

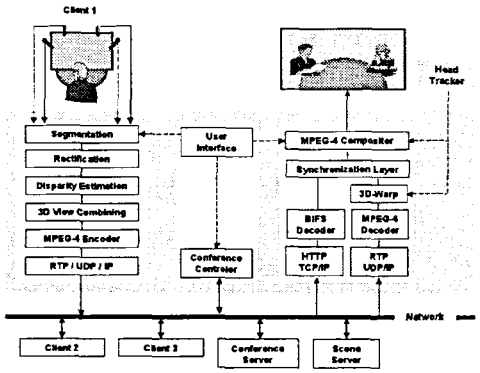
그림 3은 SVTE 내에서의 가상회의 장면이다. 화면 앞에 있는 실제 작업 책상과 확장된 화면공간에 참가자들의 영상을 seamless한 3D 영상으로 볼 수 있다.

### 2-2 VIRTUE(Virtue Team User Environment)

VIRTUE는 새로운 SVTE의 개념이다. 이것은 IST 프로젝트에 의해서 처음 제안되었다. 이것은 데이터와 장면의 전송과 표현을 위하여 최신의 멀티미디어 표준인 MPEG-4를 사용한다. MPEG-4 신호 프로토콜인 DMIF, 세련된 3D 영상 처리, MPEG-4 장면 서술언어인 BIFS의 사용법 등으로 시스템 구조는 미래의 다양한 확장과 몰입형(immersive)한 tele-collaboration 등의 개발에 관하여 다양한 유용성을 제공한다.

## III. 시스템 구조

그림 4는 3D 비디오 컨퍼런스 시스템의 시스템 구조를 보인다.[2] 먼저 각각의 카메라로부터 다중화면을 캡쳐한 이후에, 비디오 프레임은 배경으로부터 사람의 그림자를 분리한다. 이렇게 함으로써 참가자들은 가상 환경 내로 통합되고, 영상 객체에 의하여 새롭게 재생된다. 비디오 객체와 불규칙적인 컨텐츠는 MPEG-4에 의해 효과적으로 인코딩된다.




<그림 4. 3D 비디오 컨퍼런스 시스템의 구조>

제안된 시스템 개념은 MPEG-4 멀티미디어 표준의 몇몇 부분적인 특징들의 이점을 함께 가진다. MPEG-4는 임의로 그려진 영상 객체의 인코딩을 허용하고, 보조적인 평면 객체는 추가적인 픽셀 정보를 칼라 데이터로 보조하여 전송하는 것을 제공한다. 추가적인 정보는 불규칙적인 요소 비디오 영상의 동시 전송을 위하여 사용된다. 인코딩이 끝난 뒤에, 패킷은 RTP를 통하여 다른 참가자의 단말기로 전송된다. 동시에 터미널은 다른 회의 참가자로부터 영상 스트림을 수신하고, 다중 MPEG-4 영상 디코더와 함께 디코딩 된다. 화면에 그려진 영상 책체와 영상 데이터는 각각 다른 요소들과 동기화되고, SVTE 장면은 BIFS에 의하여 재생되고 통합된다.

결과적으로, MPEG-4 기술은 영상을 표현하는 데에 사용된다. 이 기술은 서버를 통하여 장면을 변환하여 전송하고 다른 사용자 이벤트를 변환, 합성하는데에 이용된다. 또한 MPEG-4 영상 객체는 영상기반 렌더링 기술을 이용하여 3차원으로 그려진다. 이것은 화면으로 그려지기 이전에 통합된다. head tracker로부터 현재 입력된 위치 데이터에 따라 올바른 표현거리는 계산되고 적용된 화면은 2D 영상과 같이 BIFS 장면내에서 삽입된다.

## IV. 영상 처리

제안한 시스템에는 다음에 설명하는 것과 같은 다양한 영상처리 기법이 필요하다. 이것은 MPEG-4를 기본으로 하고 있으며, 이를 실현하기 위한 최신의 장비 및 기술이 필요하다.

### 4-1 전후방 분할

참가자들에게 보여지는 배경화면은 실제 참가자들의 배경화면과는 차이가 있다. 회의시에 참가자들이 공유하는 문서, 물건이 아닌 다른 것들은 실제 회의시에는 불필요하다. 따라서 참가자들의 인물은 하나의 영상 객체와 같이 seamless하게 가상 환경 내로 통합된다. 이것은 다소 정적으로 남아있다고 간주되는 배경화면으로부터 동적으로 움직이는 인물의 분할을 요구한다. 초기의 배경화면은 캡처되고 변화 검출 스키마는 현재 영상 테이타와 함께 움직임 검출 알고리즘을 이용하여 이전의 이미지와 비교하며 분할 마스크를 정하고 또한 수정된다.

이 기본적인 알고리즘은 속력과 품질을 향상시킨다. 또 다른 그림자 검출 도구를 이용하여 더욱 향상시킬 수도 있다. 이것은 보다 강건한 분할은 특별히 중요하다. 왜냐하면 그림자는 최상의 조명 상태 아래에서조차도 테이블을 항상 고려하는 것이 아니기 때문이다. 테이블은 배경과 같은 정적인 영상이지만, 인물의 그림자에 의하여 제외되지 않을 수도 있다. 그림 5는 이러한 과정을 나타낸다.

<그림 5. 좌 : 원본프레임. 중 : 그림자 검출없이 분할된 전경화면, 우 : 그림자 검출까지 분할한 전경화면>

### 4-2 깊이 분석

각각의 카메라에서 최종 회의 참가자에게 보여지는 3D 화면을 얻기 위해서, 캡처된 영상 객체의 깊이를 분석하는 과정은 필수이다. 깊이 분석은 객체의 이미지의 움직임으로 분석할 수 있다. 이 방법은 불균형 필드를 계산할 수 있는 hybrid block과 픽셀-귀납 방법으로 접근된다.

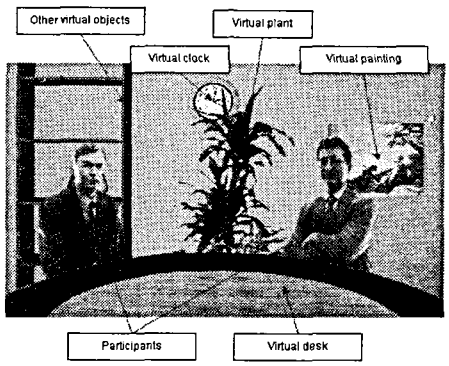
손, 팔, 또는 머리의 움직임에 대하여 깊이를 분석하는 데에는 몇가지 전처리 기술들이 적용된다. 가장 간단한 방법은 왼쪽에서 오른쪽 또는, 오른쪽에서 왼쪽으로 이미지를 매칭시키면서 정적인 필드를 동적인 필드의 이미지가 손상시키는 영역을 찾는 검사에 의하여 검출된다. 이러한 영역 내에서 신뢰할 수 없는 필드의 불규칙적인 값들은 적합한 예측기술로 효과적으로 채워진다. 불연속적인 손과 팔의 필드분할은 움직임과 피부색 정보를 이용하여 더욱 정화된다.

### 4-3 헤드 트래킹

화면상에 표현된 것의 장면은 보는 사람의 위치에 의존한다. 이것은 head tracking 모듈에 의하여 도달된 3D 공간 내에서 보는사람 위치에 대한 정확한 평가를 요구한다. 선택된 접근은 표면상의 특징을 추적하는 기기가 눈 위치를 찾는 것과 함께 공동으로 피부 색 분할하는 기술과 동일하게 추출된다. 두 영상 내의 카메라들이 얻은 2D 위치는 3D 화면에서 머리 위치의 정확한 계산을 위하여 사용될 수 있다.

**4-4 가상장면의 합성**

전송된 비디오 영상의 계산된 머리의 위치와 깊이는 원격 참가자들에게 충분한 정보를 복합적인 가상 화면으로 제공한다. 새로운 화면 표현에 관한 종합적인 알고리즘은 VIRTUE 프로젝트 내에서 이미 개발되었다.[3]




<그림 6. 컴퓨터 그래픽과 영상 객체를 포함하는 복합된 장면>

복합적인 영상처리의 가장 마지막에, 가상회의 장면은 구성되고 그림 6 내에 그려진 것과 같이 스크린 상에 디스플레이된다. 가상 장면은 3D 공간 내의 다각형의 수에 의하여 표현되어지고, BIFS 장면에 의하여 인코딩된다. 이러한 다각형 기반 장면 표현에서 참가자들은 2D 직사각형이 가상 테이블 주위에 위치하는 것에 의하여 대신된다.

**4-5 조명 표현**

자연적이고 균등하게 조절되지 않은 수신영상으로 인하여, 같은 장면이라도 각각의 카메라에 수신되는 영상은 그 밝기가 다양하다. 이러한 밝기의 다양함은 장면 내의 객체의 형상을 구분하는데에는 영향을 미치지는 않지만, 얼굴을 세밀한 부분을 표현하는데에 있어서는 큰영향을 미친다. 따라서 밝은 속성을 평가하고 제거하는 것은 상당히 중요하다. 대부분 이러한 것들은 전처리 과정에서 삭제될 수 있으나, 가상회의의 방 내에 미리 정의된 위치에 있다면 충분히 조절 될 수 있다.

**4-6 얼굴표현**

각각의 요소를 통합한 뒤의 가장 흥미를 끄는 부분은 3D로 표현된 얼굴 움직임과 수신된 2차원 이미지의 손상일 것이다. 이러한 이미지 손상은 영상 객체의 추가적인 기술에 의하여 보상된다. 또한 얼굴 표현을 위하여 3D 새로운 모델-기반 기법이 개발되었다. 모델-기반 기법은 구성 요소 및 주변 다른 것과의 상호작용시 복잡한 시스템을 수학 모델로 이용해 분석하는 접근방식이다. 개개인 사람의 움직임과, 형상, 색을 서술하는 매개변수로 나타내어진 3D 인간 헤드모델의 사용으로 표현한다. 이 모델 정보는 공간과 임시적인 이미지의 명암과 함께 동시에 개발된다.

## V. 결론 및 전망

우리는 공유된 가상 테이블 환경(SVTE)을 이용하여 몰입형 비디오 컨퍼런스를 위한 개념을 보이고, 이에 대한 설계 및 영상처리 기술을 살펴보았다. 이것으로 지역, 원격 사용자들은 서로 공유된 가상환경으로부터 비주얼하며 인터랙티브한 접근을 가능하게 한다. 이것은 이전의 VR-기반 가상환경 내의 접근에 대한 장점과 tele-cubicle[4] 시스템의 장점을 모두 가지고 있다. 실제와 가상회의 테이블 사이의 seamless한 전송은 회의 참가자에게 확장된 지각 공간의 표현을 제공할 뿐만 아니라 참가자간 눈짓, 몸짓, 행동등의 부드러운 표현을 3D 영상 객체를 사용함으로써 가능하게 한다. 또한 적합한 헤드-모델의 렌더링을 허용하는 MPEG-4 멀티미디어 표준을 사용함으로써 유연한 시스템 구조를 갖는다. 여기에 선보인 비디오 컨퍼런스는 곧 우리에게 보여질 자동 스테레오입체, 단일 또는 다중 사용자 디스플레이와 같은 몰입형 비디오 컨퍼런스와 같은 더욱 비주얼한 상호작용을 할 수 있는 시스템으로 확장될 것이다.

## 참고문헌

[1] Peter Eisert, "Immersive 3-D Video Conferencing : Challenges, Concepts, and Implementations", Proc. SPIE VCIP, Lugano, Switzerland, July 2003.

[2] P. Kauff and O. Schreer: "An Immersive 3D Video-Conferencing System Based on a Shared Virtual Environment", Proc. Int. Conference on Media Futures, Florence, May 2001.

[3] B. J. Lei and E. A. Hendriks, "Multi-step view synthesis with occlusion handling," in Proc. Vision, Modeling, and Visualization VMV'02, (Stuttgart, Germany), Nov. 2001.

[4] S. J. Gibbs, C. Arapis, and C. Breiteneder, "TELEPORT - Towards immersive copresence", Multimedia Systems, 1999.