특집/가상 인물

인터렉티브 하이브리드 미디어 응용기술

-MPEG-4 SNHC를 중심으로-

김 형 곤
한국과학기술원 영상미디어 연구센터

요 약

최근의 멀티미디어 기술은 정보의 디지털화와 온라인화에 따라 가전, 컴퓨터, 통신 및 방송 기술이 융화되어 가는 추세에 있으며, 대화형의 하이브리드 멀티미디어 기술을 그 특징으로 하고있다. 하이브리드 멀티미디어는 컴퓨터 그래픽 및 미디(MIDI) 기술로 인위적으로 생성한 2D/3D 그래픽 및 음향을 실제의 자연적인 영상과 소리에 추가하여 합성하므로 생성된다. MPEG-4는 이렇게 인위적으로 합성되거나 자연적인 영상 혹은 음향 정보의 디지털 하이브리드 멀티미디어 부호화를 목적으로 하며, 활성화된 혼합 미디어의 내용기반 처리, 상호 동작 및 사용자의 쉬운 접근 등을 가능하게 한다. SNHC(Synthetic-Natural Hybrid Coding)는 기존의 수동적인 미디어의 전달뿐 아니라 실시간 처리가 가능한 인터랙티브 응용 분야까지 다루고 있으며, 통합된 시-공간 부호화 기법을 사용하여 시각, 청각, 2차원, 3차원 컴퓨터 그래픽스 등 다양한 형태의 표준 AV(Aural/Visual) 객체를 처리한다. 표준화는 주로 mesh-segmented 비디오 부호화, 구조물 부호화, 객체간의 동기화, AV 객체 스트림의 멀티플렉싱, 혼합 미디어 형태의 시-공간 통합화 등에서 이루어지게 되는데, 이는 궁극적으로 네트워크로 연결되는 가상 환경(Virtual Environment)에서 다수의 사용자가 서로 상호작용 할 수 있는 틀을 제공하는데 있다. 이러한 틀이 제공되면, 대화형 하이브리드 멀티미디어라는 새로운 형태의 정보를 사용함으로써 기존의 미디어로는 경험하지 못하는 다양한 응용과 서비스를 경험할 수 있을 것이다.

1. 서 론

최근의 멀티미디어 기술은 정보의 디지털화와 온라인화에 따라 가전, 컴퓨터, 통신 및 방송 기술이 융화되어 가는 추세에 있으며, 대화형의 하이브리드 멀티미디어 기술을 그 특징으로 하고있다. 대화형이란 사용자가 능동적으로 멀티미디어 제작에 필요한 요구 사항을 전달하여 멀티미디어 내용물에 반영되는 형태의 미디어를 의미하며, 하이브리드 멀티미디어란 자연적인 영상과 음향에 인간이 만들어낸 2차원 (2D) 혹은 3차원 (3D) 의 인위적인 합성 (Synthetic) 미디어가 혼합된 형태의 멀티미디어를 의미한다.

근래 들어 산업계를 보면 PC용 실시간 3차원 미디어 처리 프로세서가 흔하게 되었으며, 이를 이용하여 2차원/3차원 시각 및 청각 정보의 혼용이 보편화되기 시작하였다. 기술적인 주목 대상은 그래픽스 미디어, 3차원 영상과 장면의 항해를 위한 계층적 관리, 다해상도 영상이나 3차원 LOD(Level Of Degree) 및 3차원 점진적 전송 등의 Scalability, 3차원 모델 구조 및 표면 복사값 등의 고효율 부호화 방식, affine-warping 과 같은 실시간 비디오 효과 처리, 영상 기반 3차원 모델의 view 생성, 및 분산형 가상환경 등의 기술이다.

네트워크의 발달과 컴퓨터 그래픽스 등의 기술 진화는 가정과 사회에서 미디어의 생성 및 소비의 분산화를 가져오고 있고, 결과적으로 대화형 하이브리드 멀티미디어의 도래를 촉진하고 있다. 시각/청각 혹은 2D/3D 합성 그래픽스 미디어는 기존의 TV 나 PC의 기능으로는 처리할 수 없는 다양한 형태의 하이브리드 미디어로 융합되고, 인터랙티브 미디어에 의해 사용자에 의한 다양한 재구성 (composition) 기능으로 새로운 응용분야를 창조한다. 수동형 미디어가 실시간 여부에 관계없이 저작물에 대해 사용자의 제어가 불가능한 미디어를 말하는데 반해 인터랙티브 미디어는 사용자에게 저작물 내용이나 표시방법 등을 가공 할 수 있는 능력을 제공하게 된다. 이러한 발전은 비디오 휴대폰, 원격교육이나 게임과 같은 분산형 인터랙티브 실시간 응용 시스템 등에 널리 이용될 수 있는데, 여기서 실시간이란 사용자의 요구를 만족하도록 시간 지연이 거의 없거나 연속적인 부드러운 동작이 가능케 하는 주기적이고 결정적인 갱신을 하는 전송이나 표현을 말한다. 이

러한 실시간 응용을 위해서는 혼합(mixed) 미디어 부호화를 사용해야 하는데 이는 통신의 효율성을 향상시키고 수동 혹은 인터랙티브 통신에서 하이브리드 미디어가 전송율에 관계없이 네트워크나 platform을 사용할 수 있도록 한다. 또한 컴퓨터 그래픽스나 합성 음향에서 사용되는 A/V 객체를 재구성하는 능력을 제공하고 내용기반 응용 및 상호작용을 가능케 한다. 현재 비디오 부호화에 사용되는 Mesh 기반 부호화는 높은 압축율을 제공하며 저작물의 내용기반 가공 및 상호 작용성을 제공하는 중요한 방법중 하나이다.

하이브리드 멀티미디어는 컴퓨터 그래픽 및 미디(MIDI) 기술로 인위적으로 생성한 2D/3D 그래픽 및 음향을 실제의 자연적인 영상과 소리에 추가하여 합성하므로 생성된다. 이러한 대화형 하이브리드 멀티미디어라는 새로운 형태의 정보를 사용함으로써 기존의 미디어로는 경험하지 못하는 다양한 응용과 서비스를 경험할 수 있을 것이다.

대화형 하이브리드 멀티미디어 기술의 응용 예를 요약하면 다음과 같다.

- 얼굴 애니메이션과 언어 합성: 자동 응답기, kiosks, PC 등의 얼굴 에이전트
- 사람간의 미디어 회의: 멀티미디어를 사용한 가상 원격회의나 2D/3D 설계를 동반한 원격 협동 작업
- 멀티미디어 교육과 오락: 혼합 미디어를 사용한 지식 항해, 상호 작용을 갖는 애니메이션 응답기, 원격 교육 및 협동 원격 강의, 원격 진료 및 수술 교육 등의 의료용
- 상호작용을 갖는 멀티미디어 표현: 저작물 설계, 생산 및 서비스 데모, 사내 통신 및 promotion, 3D 모델과 영상 소리를 지원받는 원격 쇼핑, 가상 여행사, 가상 복덕방
- 디지털 A/V 미디어 작업: 탁상형 및 분산형 가상 스튜디오 및 Set
- 혼합 미디어를 사용하는 게임 및 교육: 분산형 가상 환경, 다수 사용 모의 시험 환경 및 게임 환경, 멀티미디어를 지원하는 컴퓨터 기반 훈련

2. MPEG-4 SNHC

대화형 하이브리드 멀티미디어의 새로운 기술의 요구에 따라 MPEG-4는 1998년 말까지 혼합 미디어 부호화의 1차 표준안을 제시하기 위해 작업을 진행중이다. 이 작업은 기존의 MPEG-1/2 에서 제공하는 모든 멀티미디어 서비스를 포함하며, 인터넷이나 이동 통신을 통한 멀티미디어 서비스로서 영상 전화, 영상 회의, 원격 진료, 원격 오락 및 가상 현실까지 그 대상을 확장하고 있다. MPEG-4는 이러한 기술을 바탕으로 최종 단말기 사용자, 서비스 제공자, 저작자 모두의 필요를 만족하는 도구들을 제공하는 것을 목적으로 한다. 단말기 사용자에게는 저작가에 의해 설정된 범위 안에서 저작물과의 인터랙션을 제공하고, 네트워크 서비스 제공자에게는 각각의 미디어에 필요한 서비스 질 (QoS)을 보장하는 방법과 이기종 간의 네트워크에서 전송을 최적화할 수 있는 신호방식을 제공하고, 저작자에게는 오늘날 독립적으로 사용되는 디지털 TV, 애니메이션 그래픽스, WWW pages 등의 것보다 훨씬 신축성 있고, 재사용 가능한 저작물의 생성을 가능케 하여준다. 이를 위하여 표준화되어야 할 내용에는 시각 및 청각 객체(Audio/Video Objects: AVO), 이들을 재구성해서 다양한 Scene 의 생성, AVO 데이터의 효율적인 multiplex와 동기화, 수신단에서의 인터랙션 방법 등을 포함한다.

그림 1은 하이브리드 미디어에 의하여 대화형 멀티미디어가 어떻게 사용되는가에 대한 예를 보여준다. 화면은 2차원 혹은 3차원의 AV 객체들로 계층적으로 구성되어진 Scene Description 에 의해 표현되며, 단말기에 내재된 Compositor 에 의해 화면에 재구성된다. 최 하위 계층의 기본 AV 객체는 자연 혹은 합성된 음향, 영상, 문자, 그래픽스 등 다양한 형태를 가지며, 이들의 애니메이션을 포함한다. 그림에서 사람, 탁자, 칠판, 지구본, 사람의 목소리 및 다른 객체에서 나는 소리 등은 기본 AV 객체들이고, 말하는 사람에 해당하는 비디오 객체와 음성에 해당하는 청각 객체는 하나의 상위 객체를 이루며 이를 복합 객체라 한다. 전체적으로 하나의 장면은 그림 2에서 나타낸 계층적인 구조로 표현되며, 각 객체는 서로 독립적으로 부호화되어 전송된다. 이러한 AV 객체를 사용하여 사용자는 다음과 같이 장면을 재구성 할 수 있다.

1. AV 객체를 주어진 좌표 시스템의 임의의 위치에 설정
2. 기본 AV 객체를 그룹화 하여 임의의 복합 객체의 구성
3. 스트림 데이터를 AV 객체에 적용하여 애니메이션이나 텍스쳐 등의 속성제어
4. 사용자의 보는 관점이나 듣는 위치를 임의로 설정 가능

이러한 장면 재구성의 개념은 그 구조 및 객체의 기능성을 VRML (Virtual Reality Model Language)에서 빌린 것이다. 일반적으로 사용자는 저작자에 의해 구성된 형태의 장면을 보게되나, 저작자에 의하여 허용된 범위 내에서 사용자는 다음과 같은 새로운 차원의 기능을 선택할 수 있다.

1. 영상 장면의 보는 시점 및 듣는 위치의 선택 (장면에서 항해 할 때)
2. 필요한 청각 정보만 선택 및 해제 기능
3. 영상 내에서 물체의 위치 변경 기능
4. 영상 스트림의 시작, 멈춤 지정 기능
5. 다중 언어가 지원되는 경우 원하는 언어 선택 기능

MPEG-4는 자연적이거나 합성적인 영상 및 음향을 부호화할 때에 내용 기반 부호화를 수행한다. 멀티미디어의 내용을 객체 단위로 나누어 처리할 수 있게 함으로써 사용자의 요구에 의해 멀티미디어 내용물의 형태를 조작할 수 있고 또 원하는 표시가 가능해져서 사용자 위주의 대화형 멀티미디어가 가능하게 한다. 따라서 기존의 멀티미디어가 사용자가 수동적으로 받아들이기만 하는 멀티미디어라면 MPEG-4에 의한 멀티미디어는 사용자가 능동적으로 참여하고 조작할 수 있는 멀티미디어 표준을 제시한다고 할 수 있다.

사용자에 의한 대화형의 멀티미디어 기능을 제공하기 위해서는 양방향 정보가 필요하게 되며, 이의 효율적인 부호화 방식이 중요해 진다. MPEG-4에서는 객체 기반 영상 부호화 시에 SNHC (Synthetic - Natural Hybrid Coding) 기술을 추가하여 이 문제를 해결하려고 한다. 기존의 MPEG-1이나 MPEG-2가 자연적인 영상 및 음향의 부호화를 다루었던 반면, MPEG-4의 자연-합성 하이브리드 부호화는 자연적인 미디어 정보와 컴퓨터에 기반을 둔 인위적인 합성 미디어 정보를 혼용하는 대화형 멀티미디어를 구현하는 것을 주된 목적으로 한다. 또한, 작은 대역폭에서도 효율적으로 양방향 멀티미디어 데이터 전송이 이루어지도록 하기 위하여 합성 영상 및 소리의 부호화를 별도로 고려하여 그 효율성을 향상시킨다.

하이브리드 미디어의 효율적인 압축을 위해 현재 SNHC에서 중심적으로 제안하고 있는 분야는 다음으로 요약된다.

- 합성된 얼굴, 몸통 및 이의 애니메이션을 위한 변수화 된 묘사
- 문자, 그래픽스 등의 미디어 통합
- 관측점의 변화를 반영하는 텍스쳐 부호화
- 텍스쳐 매핑을 갖는 정적 및 동적 선틀(wireframe) 부호화

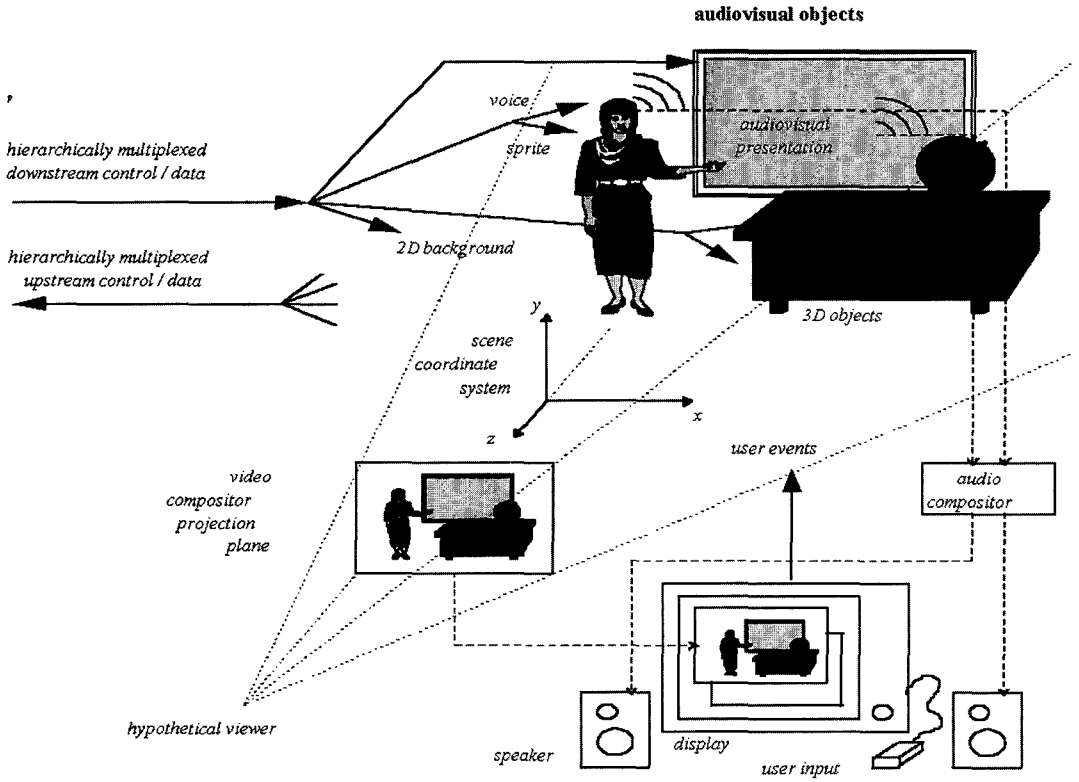

그림 1. 하이브리드 미디어를 이용한 대화형 멀티미디어 시나리오 예

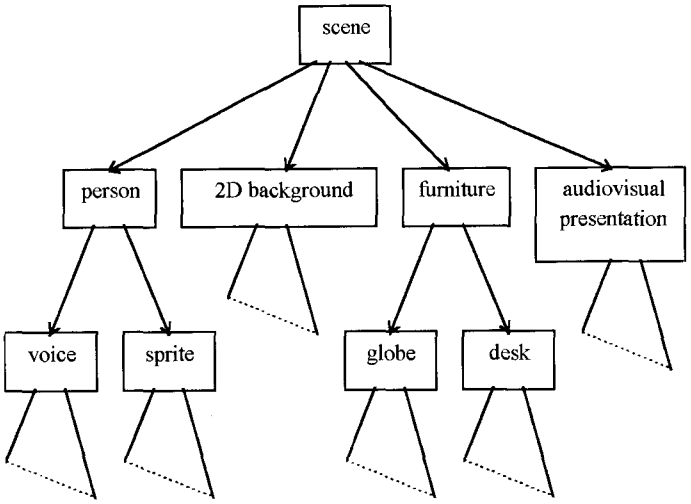

그림 2. 그림 1의 Scene을 나타내는 계층적 구조

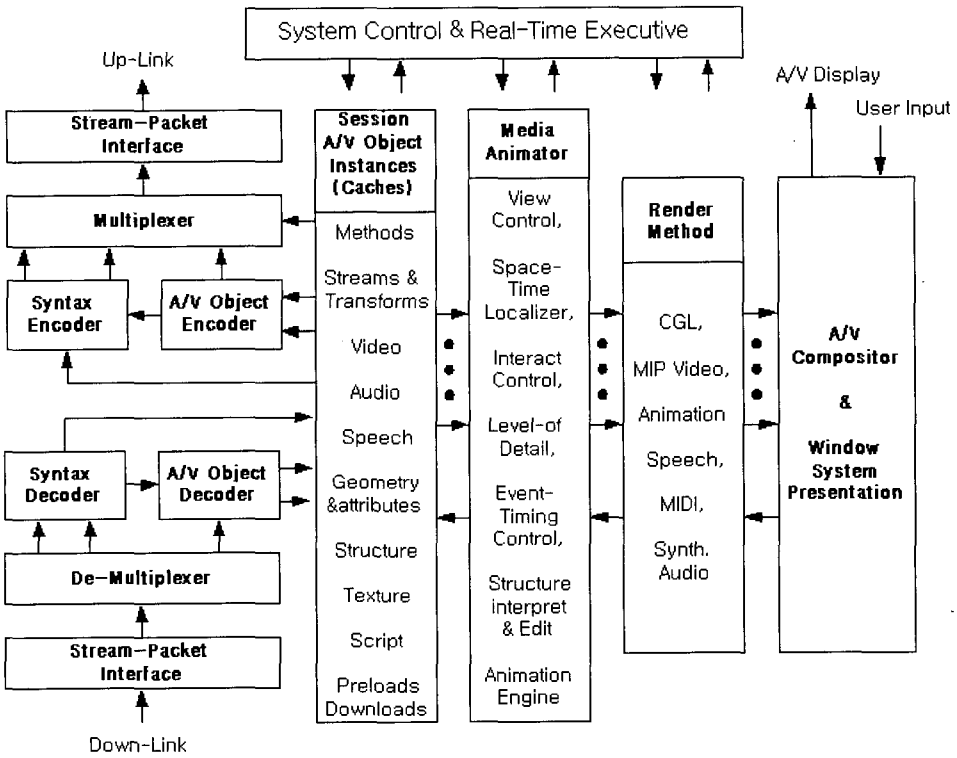

그림 3. MPEG-4 단말기의 기능 구성도

- 문자-언어 변환 (TTS) 합성을 위한 접속 방식
- 합성 음향의 부호화

사용자의 요구 중에서 보는 관점 및 듣는 위치의 변경이나 물체의 이동 등의 기능은 2차원의 자연 영상 및 소리를 사용하는 경우 대역폭의 제한과 2차원이라는 한계 때문에 처리하기가 불가능하였다. 그러나 컴퓨터 그래픽 기술 및 미디(MIDI) 기술을 이용하는 3차원 또는 2차원 합성 영상 및 소리를 사용하는 경우 사용자의 시점 및 객체의 위치를 3차원의 객체를 이용하여 변경하게 된다. 멀티미디어 데이터를 구성할 때 영상 및 소리 객체들을 자연 영상 및 소리를 이용하여 3차원의 데이터로 만드는 것은 계산상으로 어렵고 데이터가 크기 때문에 각각의 객체는 3차원의 모델을 이용하게 된다. 그림 3은 MPEG-4에 사용될 단말기의 기능 구성도를 나타낸 것으로 갖가지 A/V object들이 부호화, 복호화되어 분리되고 재구성되는 흐름을 보여주고 있다.

3. 말하는 얼굴 및 몸 동작 애니메이션

시각을 통한 의사 전달은 인간에게 널리 이용되는 방법으로 여러 가지 악조건에서도 의사소통을 가능케 한다. 실제 대부분의 사람은 얼굴을 볼 수 있는 경우 매우 시끄러운 상황에서도 의사 소통이 가능하며 이는 시각정보에 의한 것이다. 음성과 말하는 사람의 표정을 자동 해석하면 시각 언어, 표정, 제스처 등의 고급 정보를 제공할 수 있고, 이러한 해석으로부터 얻어지는 얼굴 특징 변수들은 언어 인식, 얼굴 애니메이션 및 저전송율 부호화 등에 광범위하게 이용될 수 있다.

MPEG-4 SNHC 의 기본 목적은 영화 제작, 인터랙티브 오락, 영상 및 음성을 포함하는 멀티모달 접속, 가정용 게임, 사람 사이의 통신 등 합성하거나 혹은 자연적인 말하는 얼굴 영상이 필요한 멀티미디어 응용에서 음성과 비디오 사이의 동기화를 위한 모든 기술을 제공하는 것이다. 얼굴 영상은 말하는 얼굴 모양과 이에 따른 음성이라는 두 가지 객체로 구성되며, 이들간의 동기화는 매우 중요하다. 비디오 프레임을 표시하거나 합성 얼굴을 생성하는 정확한 시간은 부호화 이전의 전처리 단계나 표시 이전의 후처리 단계에서 음성처리와 동기되어야 한다.

SNHC 가상환경에서 표시되거나 애니메이션 될 수 있는 AV 객체 중에서 가장 복잡한 형태가 가상 인물이다. 치아나 뼈와 같은 딱딱한 물체와 혀, 입술 및 피부와 같은 부드러운 물체가 복잡한 표면 정보와 3차원 다이나믹스를 가지고 복합적으로 작용하는 가상 인물은 인간의 다양한 감정 정보까지 표출 할 수 있어야 한다. 일반적으로 얼굴 애니메이션과 주변 환경을 나타내는 데이터의 부호화와 관련하여 네 가지 층의 정보를 생각할 수 있다. 가장 낮은 수준의 정보에는 오디오 데이터와 얼굴 및 몸통의 모델 파라메타 정보이며, 얼굴 애니메이션 시스템이 수신단에 상존하고 있는 경우 사용된다. 두 번째 계층은 가상환경과 얼굴의 정적 부분을 구성하는데 필요한 정적 물체 정보이다. 세 번째 층은 전송되는 파라메타들을 변형하는 얼굴 및 몸통 모델 프로그램을 포함한다. 가장 고위의 계층에는 비디오 및 텍스쳐 정보를 나타내며 많은 량의 정보가 포함

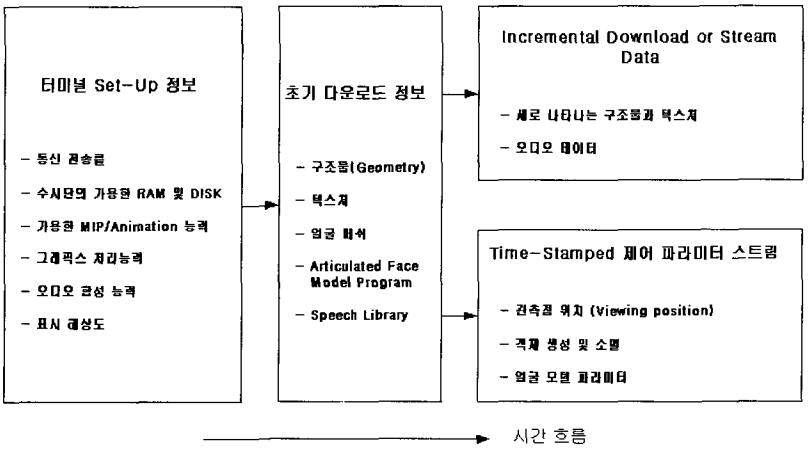

그림 4. SNHC 단말기 통신시 전송되는 데이터의 흐름

된다. 그림 4는 계층적 데이터 구조를 이용하여 SNHC 단말기가 데이터를 전달할 때의 흐름을 나타낸다. 스스로의 능력을 알고 있는 송신단은 수신단의 성능을 알아내어 가장 최적의 성능을 얻기 위해 사용가능 전송률, 수신단의 메모리 용량, 애니메이션 능력, 그래픽스 가속처리 능력, 오디오 합성 능력, 표시 해상 능력 등의 정보를 주고받아 단말기 설정을 완료한다. 이후, 초기화에 필요한 구조물, 텍스쳐, 얼굴 메쉬, 얼굴 모델 프로그램, 언어 라이브러리 등의 데이터를 전송한다. 이러한 초기화가 완료되면 오디오 데이터나 새로 노출되는 구조물이나 텍스쳐 정보와 같은 스트림 데이터나 증가형 다운로드 데이터와 관측점 위치 데이터, 객체 생성 및 소멸, 얼굴 모델 파라메타 등 time-stamped 된 제어 파라메타 스트림을 이용하여 수신단에 제어 및 갱신 데이터를 전송하게 된다.

3.1 3차원 (3D) 얼굴 표준 모델

3D mesh 얼굴 모델은 자연적인 2차원 얼굴 데이터와 3차원 합성 얼굴을 관계 지우는 가장 기본적인 것이다. 모델은 얼굴 표면 위의 많은 데이터 점과 이들을 연결하여 3각 면으로 표면을 근사화 하여 생성되고 이들의 애니메이션을 위하여 중요한 점들을 제어 점으로 정의한다. 3D mesh 의 데이터 파일 형태는 데이터 점들과 제어점 위치, 그리고 데이터 점들간의 관계 등을 포함 하여야한다. 3D mesh 가 준비되면 scaling, 임의의 축에 대한 회전등의 3차원 변환을 변환 matrix를 이용해 수행한다.

그림 5 에는 1038개의 삼각형 메쉬 꼭지점 및 1703개의 삼각 면 조각으로 이루어져 있는 3차원 얼굴 표준 모델을 나타내었다. 이러한 모델은 일반적인 사람의 얼굴 특징을 평균해서 나타내는 모델이며, 상황에 따라 그 모습과 표정 등의 변수를 바꿀 수 있어야 한다.

특정한 얼굴 모델은 얼굴 표현 변수인 FDP(Facial Definition Parameter)와 얼굴 동작 변수인 FAP(Facial Animation Parameter)에 의해 제어되고, 얼굴 모델의 표정 제어를 수행 할 수 있다.

3.2 얼굴 모양 지정 파라메타 (Facial Definition Parameter: FDP)

FDP는 얼굴 모델을 새로운 3D 메쉬 정보에 의해

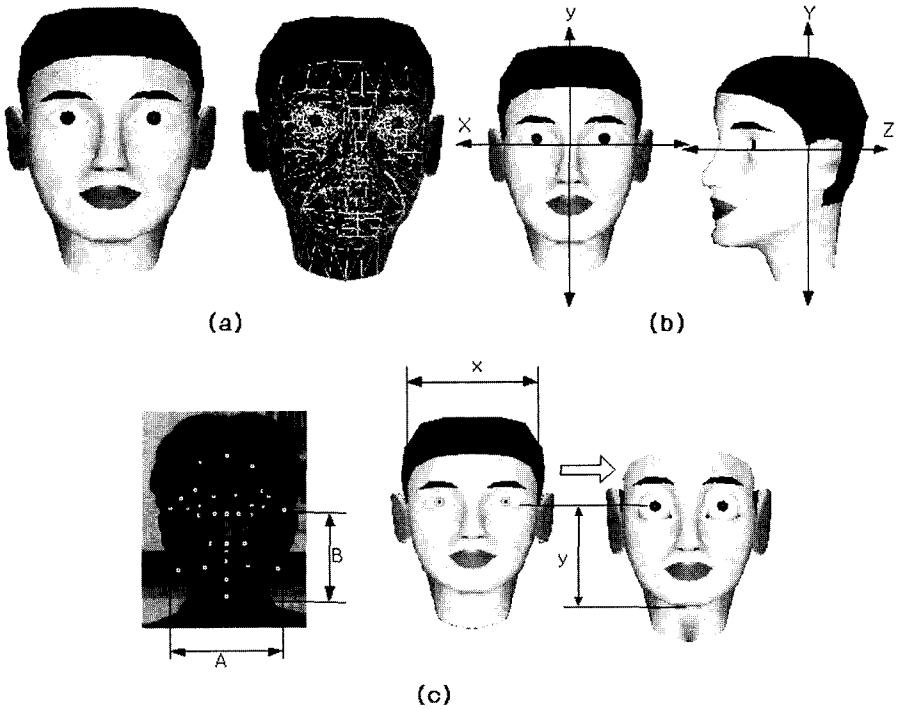

그림 5. 3차원 얼굴 표준 모델과 FDP 적용 과정

생성할 수 있게 하거나, 표준 얼굴 모델에서 주어진 일반적인 얼굴 모델을 특정한 개인의 얼굴 모델로 변형시키는 변수이며, 다음과 같이 구성되어 있다.

- 3D mesh (텍스쳐 사용시 텍스쳐 좌표 포함)
- 3차원 특징점
- 텍스쳐 영상 (선택사항)
- 기타 (머리카락, 안경, 연령, 성별) (선택사항)

FDP는 멀티미디어 전송 시에 처음 한번만 전송되어 표준모델을 변형시키고, 이후 FAP에 따라 얼굴 움직임 및 표정 및 입 모양이 변화될 수 있다. 3D mesh는 일반적인 사람 얼굴 형태를 가지고 있는 얼굴 모델이고, 3차원 특징점은 일반 얼굴 영상을 개개인의 얼굴에 유사하게 형태를 변경시키는 위치를 나타내는 점들이다. 3차원 특징점은 47개로 이루어지며 개개인을 구별하여 표현할 수 있는 얼굴의 중요한 위치를 나타낸 것으로 그림 6에 그 위치를 나타내었다.

그림 7에는 KIST 영상미디어 연구센터에서 수행한 FDP 연구 결과를 나타내었다. 개개인의 얼굴 특성을 표현하는 47개의 얼굴 특징점 파라미터(Facial Definition Parameter : FDP) 중 2차원 영상에서 추출 가능한 31 개의 특징점을 자동 추출한 결과를 두 번째 열에 나타내었다. 또한 추출된 2차원 얼굴 특징점들을 1038 개의 삼각형 메쉬로 이루어진 3차원 일반 얼굴 모델에 적용시켜 변형함으로써 개개인의 얼굴에 해당하는 모델을 자동 생성한 결과를 나타내었다.

3.3 얼굴 움직임 파라미터 (Facial Animation Parameter: FAP)

3차원 얼굴 모델이 FDP 에 의해 만들어지면 이의 움직임을 제어하기 위해 움직임을 표현하는 FAP를 이용한다. 얼굴 동작변수 FAP는 얼굴 움직임에 대한 연구에 근거하였으며, 근육 운동과 밀접히 관련되어 있다. 이들은 기본적인 얼굴 움직임을 모두 나타낼 수 있으며, 얼굴 표정, 언어 발음에 대한 입술 동기 등과 같은 기능을 제공하여 3D 얼굴 모델의 애니메이션을 구현하기에 충분하게 설계 되어있다.

FAP는 68개로 구성이 되며 각각의 FAP는 유사한 얼굴 내의 움직임으로 분류하여 10개의 그룹으로 형성된다. 10개의 그룹은 1) Visemes(입모양) 및 표정 Expression (2), 2) 턱 및 입 (16), 3) 눈 주위 (12), 4) 눈썹 (8), 5) 볼 (cheek)(4), 6) 혀 (5), 7) 머리 회전(3), 8) 바깥 입술 (outer lip) 위치 (10), 9) 코 (4), 10) 귀 (4)로 구성이 되어 좀더 효율적인 부호화를 이룰 수 있다. 괄호 속의 숫자는 해당 그룹에 포함되는 FAP 수를 의미한다. 각각의 그룹과 FAP 는 마스크를 이용해서 그 사용 여부를 결정하고 사용하지 않는 FAP 나 그룹은 그 값을 지정하지 않아도 된다.

평행 이동을 포함하는 모든 변수들은 변수 단위 FAPU (Facial Animation Parameter Units) 로 표현되며, 그림 8에 나타내었다. FAPU 단위는 얼굴표정 및 언어 발음 시 FAP의 적용이 모든 모델에 대해서 일관된 방

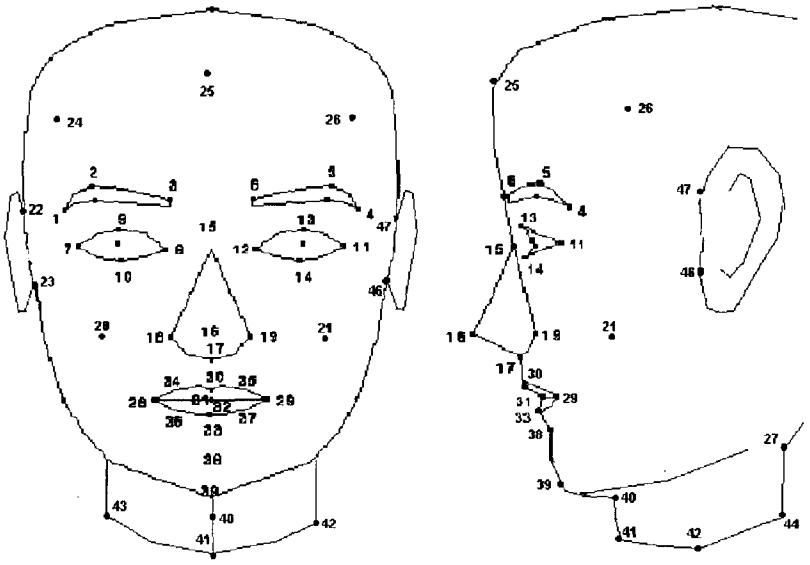

그림 6. 얼굴 모양 지정 파라미터(Facial Definition Parameter: FDP)의 3차원 특징점

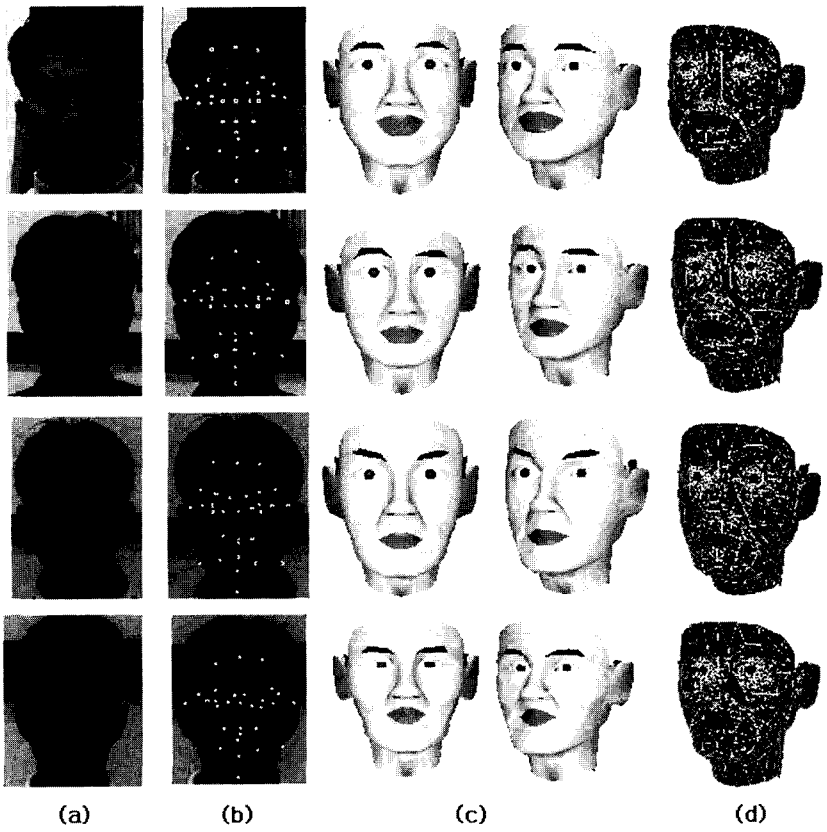

그림 7. 자동 추출된 FDP를 적용하여 개인의 특성을 반영한 얼굴 모델 생성 결과

식으로 적용되어 적절한 결과를 생성할 수 있도록 정의되어있으며, 눈 사이의 거리 (ESo), 눈과 코 사이의 거리 (ENSo), 입과 코 사이의 거리(MNSo), 입의 폭 (MWo)등의 핵심 얼굴 특징점 사이의 거리를 이용하여 단위가 결정된다. 이들 거리를 이용해 정의되는 ES, ENS, MNS, MW 단위들은 각각 해당 거리의 1/1024를 기본 단위로 한다. 예를 들면 ES는 ESo의 1/1024이 되는 것이다.

정의된 68개의 FAP는 그 이름, 간략한 내용, 사용되는 단위값 (FAPU), 값의 방향 (unidirectional or bi-directional), 양의 값이 나타내는 방향의 정의, scalable 부호화를 위한 group number, FDP subgroup number, 및 quantization step size를 포함한다.

각각의 FAP에 해당되는 값을 지정하고 조합하여 부호화하면 영상 내에 있는 사람의 얼굴의 움직임을 표현할 수 있게 된다. SNHC의 FAP에서 중요하게 제안하고 있는 부분이 사람이 언어를 발음할 때 입 주위의 자연스런 움직임이다. 따라서 입술, 혀의 움직임에

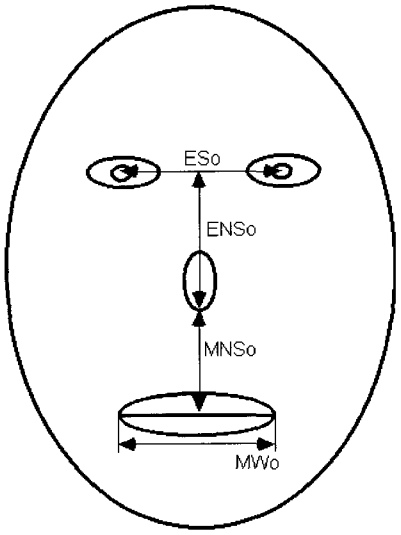

그림 8. 얼굴 애니메이션 변수의 단위 수정

그림 9. FAP 에 의한 얼굴 표정 합성 결과

많은 FAP를 할애하고 있다. 이러한 움직임이 자연스럽게 이루어져야 영상 회의나 영상 교육 등에서 중요시되는 자연스러운 TTS를 구현할 수 있게 된다. FAP를 이용하면 대부분의 자연적인 얼굴 표정을 표현할 수 있으며, 이들의 과장된 값들을 사용하면 일반적으로 사람에게서는 볼 수 없는 동작을 가능케 하여 만화에서와 같은 캐릭터의 움직임 제어에 사용 할 수 있다. 그림 9는 얼굴 영상에서 뽑아낸 FAP를 가지고 만들어진 캐릭터의 얼굴 표정을 합성한 결과를 보여주고 있다.

3.4 고급 FAP 와 모델의 언어능력

고급 얼굴 동작 변수들은 표정(expression) 변수와 언어 발음 시 입모양을 나타내는 viseme 변수의 두 형태로 나누어진다. 표정 변수는 얼굴 표정에 대한 고급 표현 방식으로써, 기쁜(Happy) 경우 입은 열리고 입 끝은 위로 당겨지며 눈썹이 이완되는 것과 같이 각 얼굴의 표정이 어떻게 animate 되는가를 설명한다. 그리고 해당 점들이 사전에 정의 된 FAPU 에 의하여 움직이는 방법들을 정의해 놓는다. 슬픔, 노여움, 두려움, 역겨움 및 놀람 등의 표정에 대해서도 이러한 방법으로 미리 정의하게 된다.

viseme 변수는 언어 발음시 발생하는 입모양을 표현하기 위한 고급 수준의 표현 방식이다. Phoneme들의 시각적 표현은 SAMPA 표준에 따르며 이는 컴퓨터가 읽을 수 있는 phonetic alphabet을 나타낸다. 현재 처음 단계로 아래와 같은 15 가지 Viseme Parameter FAP 가 포함되어 있다. viseme 선택 field 가 6비트로 정의되어 있으므로 확장하여 사용될 수 있다.

표 1. 현재 제안된 visemes 종류

Viseme_select	phonemes	example
0	none	na
1	p,b,m	put, bed, mill
2	f,v	far, voice
3	T,D	think, that
4	t, d	tip, doll
5	k, g	call, gas
6	tS, dZ, S	chair, join, she
7	s, z	sir, zeal
8	n, l	lot, not
9	r	red
10	A:	car
11	e	bed
12	I	tip
13	Q	top
14	U	book

다양한 모음은 다음과 같은 4개의 고급 파라미터로 표현 될 수 있다.

- 열린 입 높이 (Lip Opening Height : LOH)
- Y 축상의 턱 (Jaw on Y axis : JY)
- 열린 입 폭 (Lip Opening Width : LOW)

• 아래 입술 돌출 (Lower Lip Protrusion : LLP)

아래 표는 이러한 일반적인 모음을 나타내는 파라미터를 이용해서 중요한 모음을 나태내 본 결과이다. 즉, 각 모음은 아래와 같은 4개의 모음 고급 파라미터로 분해해서 나타낼 수 있다. 사용되는 단위는 MNS FAPU 단위이다.

표 2. 핵심 모음에 대한 고급 파라미터의 분해 값

VISEME	EXAMPLE	LOH(MNS)	JY(MNS)	LOW(MNS)	LLP(MNS)
a	car	410	369	40	-61
e	yes	369	246	40	-61
i	six	328	246	40	41
o	short	328	82	-163	102
u	put	328	82	-163	102

4개의 모음 고급 파라미터는 다시 기존의 FAP의 조합으로 표현된다. 예를 들면 아래 입술 돌출 (Lower Lip Protrusion) 변수인 LLP는 다음과 같이 분해되어 표시 가능하다. 여기서 x 값은 LLP 파라미터의 값을 나타내고, 기존의 저급 FAP 값은 주어진 값으로 곱해진 값을 갖는다. 각 고급 파라미터도 이와 같이 기존의 FAP의 조합으로 표현될 수 있다.

FAP number	FAP name	Intensity
6	stretch_l_cornerlip	1/2 * x
7	stretch_r_cornerlip	1/2 * x
53	stretch_l_cornerlip_o	1/2 * x
54	stretch_r_cornerlip_o	1/2 * x

MPEG-4에서는 고급 오디오 (> 15 KHz), 중급 오디오 (<15 KHz), 광대역 언어 (50 Hz - 7 KHz), 협대역 언어 (50 Hz - 3.6 KHz), 지능형 언어 (300 Hz - 3.4 KHz) 등의 소리 객체를 지원한다. 사람의 발음 소리인 지능형 언어의 경우 전송률이 2 Kbit/sec 근처이며, 이는 인간의 음성을 효율적인 전송이 가능한 문자를 이용해 보낸 후 인위적인 미디 기술로 음성을 합성하는 TTS(Text-to-Speech) 기술사용을 전제하고 있는 것이다. TTS에서는 사람이 언어를 구사할 때 나올 수 있는 모든 발음을 미리 정해 놓고 입력되는 문자에 따라 해당 발음을 나타내는 정보를 발음시의 사람 얼굴의 움

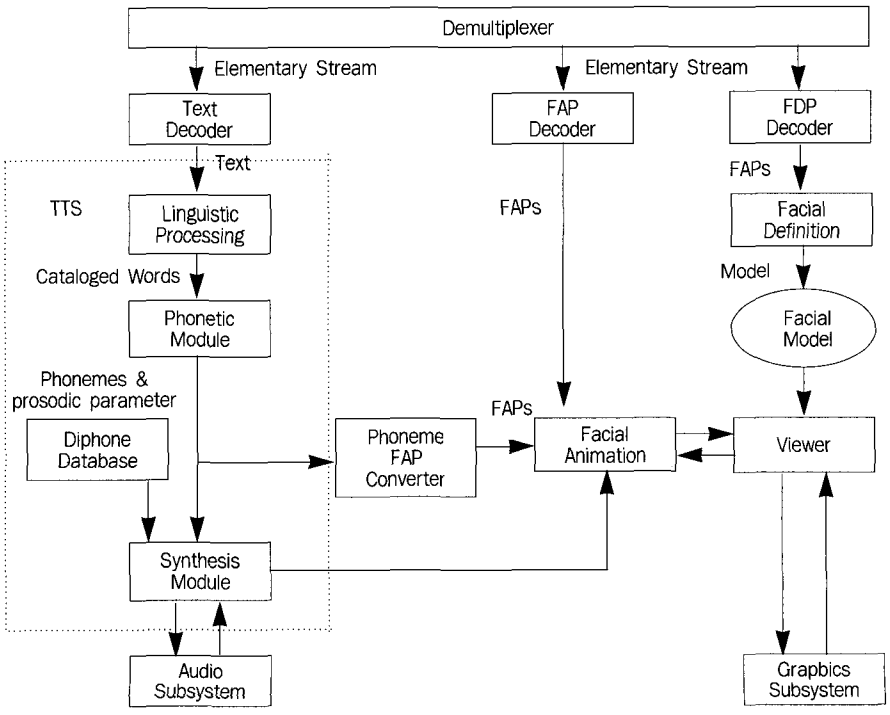

그림 10. 문자-음성 변환 (TTS) 기능을 갖는 얼굴 애니메이션 시스템 구성

직임에 맞추어 생성하게 된다. 여기서 대화형 멀티미디어의 한 기능인 여러 가지 언어 선택 기능을 추가하려면 TTS 디코더에 여러 가지 언어를 제공하는 기능을 추가하여 사용자가 원하는 언어로서 제공받을 수 있게 한다.

그림 10은 SNHC 제안하는 얼굴 애니메이션과 TTS의 구동 구조를 간략하게 나타낸 것이다.

3.5 몸 모양/움직임 파라미터 (BDP/BAP)

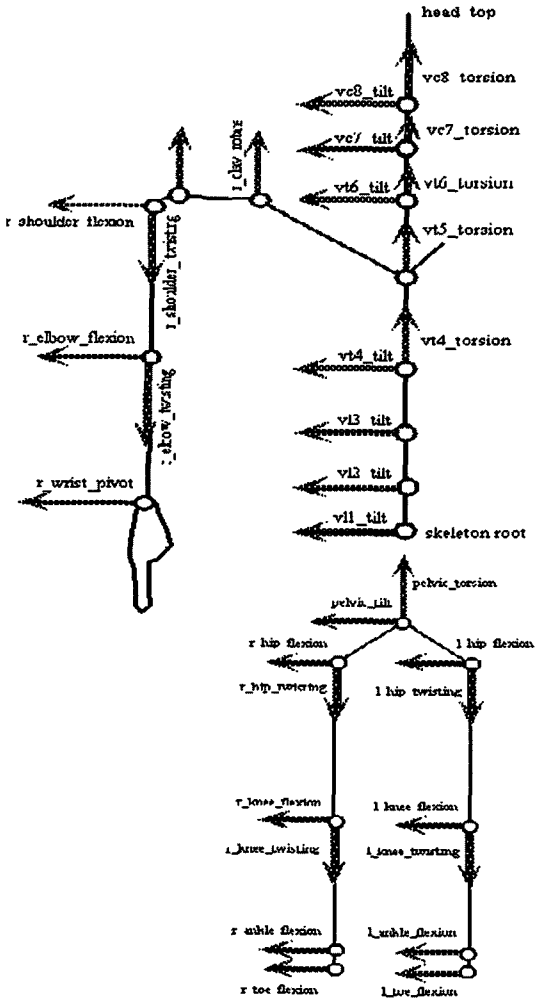

그림 11. BAP의 위치

BDP(Body Definition Parameter)는 다음과 같이 구성된다.

- Body surface geometry
- 3D reference points
- Texture images (optional)
- Attachment information of the geometry

Body surface geometry 는 일반 사람 몸 모델 표면의 기하학적 지정 값이며 3D reference points는 사람 몸 모델을 구성하는 몸의 위치를 나타내는 points이다. 이것에 의해 일반 사람 몸 모델이 개개인의 몸 모델이 된다. 이러한 BDP에 의해 영상내의 사람이 모델링 되면 BAP(Body Animation Parameter)를 전송하여 생성된 몸 모델을 영상의 지정된 움직임 좌표 내에서 움직이게 할 수 있다. BAP는 몸의 각 부분의 움직임을 지정하는 변수이며 169개로 구성되어 있고 169개의 BAP는 각각 0~225 까지의 값에 따라 움직임의 정도를 나타낼 수 있다. BAP도 역시 부호화의 효율성을 높이기 위해 Pelvis, Left leg1-2, Right leg1-2, Left arm1-2, Right arm1-2, Spine1-5, Left hand1-2, Right hand1-2, Global Positioning 의 유사한 움직임을 표현하는 19개의 그룹으로 나뉘어 다루어지게 된다. 그림 11은 BAP의 일부를 보여주고 있는데 여기에는 몸의 팔, 하체 및 몸통에 따라 움직임 방향이 정해져있다.

4. 분산 협동형 (Distributed Collaborative) 가상환경

SNHC 는 궁극적으로 네트워크로 연결되는 가상 환경 (Virtual Environment)에서 다수의 사용자가 서로 상호 작용 할 수 있는 틀을 제공하는데 있다. 이러한 틀이 제공되면, 여러 사용자가 현장감을 주는 오디오 및 비디오 객체를 통합하는 공통의 가상 환경에서 움직이면서 현실감 있는 상호 작용을 할 수 있다. 각 개인은 이러한 환경에서 가상 인물로 나타내 질 수 있으며, 서로 구별되고 대화하며 인식 할 수 있다. 이러한 환경은 게임과 원격 진료, 가상 원격 회의 등에 이용될 것이다.

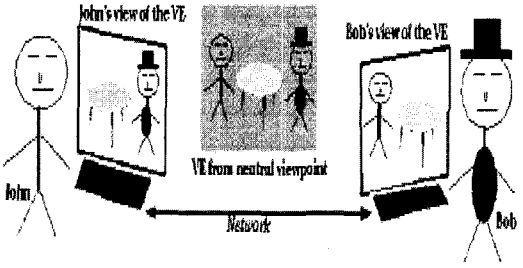

그림 12. 분산 협동 가상환경의 개념도

그림 12는 이러한 분산 협동 가상환경의 개념을 나타내 주는 그림이다.

분산 협동형 가상환경은 최근 수년 동안 큰 연구 대상이 되었으며 다수의 실험 시스템이 구현되어 있다. 이들은 네트워크 사용 방법, 지원하는 사용자 수, 상호작용 능력 및 응용 분야 등에 따라 약간씩 차이가 있지만, 기본적인 원리는 동일하다. 즉, 각각의 단말기는 가상환경을 갖고 있어 사용자는 이를 통하여 그 속에서 움직이며 상호작용이 가능하다. 환경 속에서 발생하는 모든 사건들은 다른 단말기에 보내지게 되어 새로운 상태로 갱신되며 모든 사용자가 동일한 환경에서 있는 느낌을 준다. 사용자 자신도 가상 인물로 표현되어 가상환경의 일부분이 되어 서로 상호 작용한다.

분산 협동형 가상환경에서 가장 중요한 기능은 2차원, 3차원, 오디오, 비디오, 영상 등의 다양한 미디어 객체들의 통합 기능과 이들과의 인터랙션(interaction) 기능이다. 가상환경이 그 특성상 3차원이므로 기본 구성요소들도 모두 3차원 형태를 갖는다. 모든 object 들은 3D object 의 특별한 경우로 처리함으로써 이들과의 인식 및 상호작용을 통일된 방식으로 처리할 수 있다. 보는 관점을 사용자가 마음대로 바꾸면서 이러한 3차원 object 들을 보이게 하는 rendering 기술은 가상 환경의 기본 기능이다. 인터랙션에는 사용자가 마음대로 3차원 물체를 가상 공간에서 움직이게 하는 가장 기본적인 것에서부터 scripting 에 의해 물체의 행동(behavior)을 제어하는 복잡한 것도 있다. 3차원 오디오 표현 기술은 가상공간에서 오디오 object를 임의의 위치에 놓게 할 수 있는 기능을 제공한다. 이러한 소리 생성기는 그래픽스 형태로 표현될 수 있고 일반적인 3차원 object 와 같이 처리 될 수 있다. 비디오 object 는 가상 공간에서 가상 화면같은 2차원 object 혹은 다면체같은 3차원 object 위에 나타내 질 수 있다. Texture mapping 이라 불리는 이 방법은 사용자가 이동시킬 수도 있으며, object 의 형태를 변형시킴으로써 비디오에 특수 효과를 제공하기도 한다. 반투명한 비디오 object 는 환경과 혼합(blending) 되어 특수한 환경을 만들 수 있다.

SNHC 는 초기 단계에서 얼굴 및 몸 동작 애니메이션에 집중하였으며, 인간 표현에 대한 통합된 표현을 가능하게 하였다. 향후, 분산 협동형 가상환경에서의 가상인물은 주변에 사람이 있는가를 알 수 있는 인식 기능 (perception), 그 사람이 어디에 있는가를 파악하는 기능 (localization), 그 사람이 누구인지 식별기능

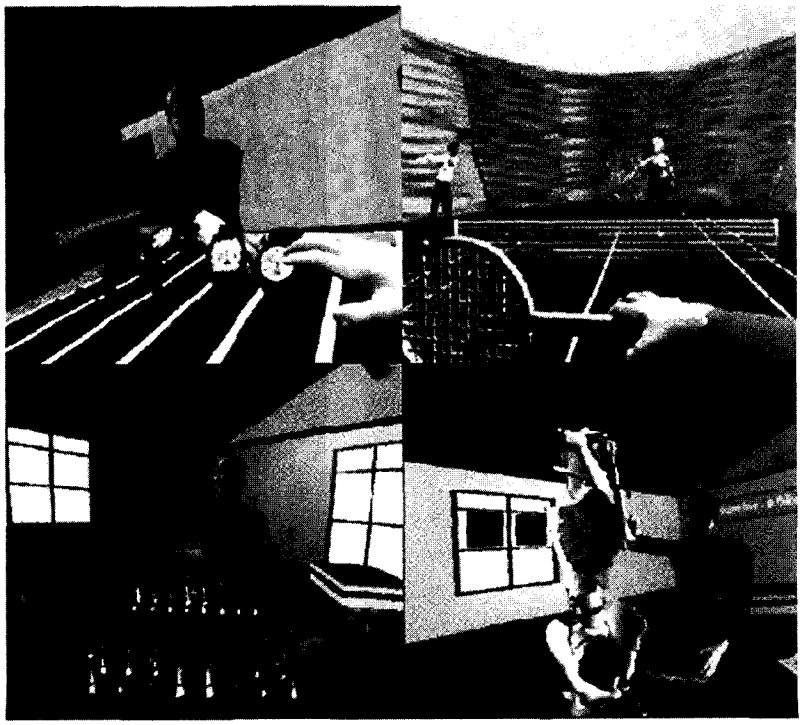

그림 13. 분산 가상 환경의 응용예
(상측 좌: 가상 쇼핑예, 상측 우: 가상 운동, 하측 좌: 가상 체스, 하측 우: 원격 의료 훈련)

(identification), 그 사람이 어디에 집중하고 있는지를 시각적으로 나타낼 수 있는 기능 (interest focus), 그 사람이 하고있는 행동을 시각적으로 나타낼 수 있는 기능 (action), 언어에 따른 입술 움직임, 얼굴 표정, 제스처 인식 등의 의사 소통 기능 등의 기능을 실생활에서 하는 것과 비슷하게 직관적으로 할 수 있어야 한다.

분산 협동형 가상환경의 응용과 표준화 방향은 크게 가상환경에서의 데이터 교환 방식과 3차원 물체의 압축 방식에서 검토될 수 있다. 데이터 교환 방식에는 클라이언트/서버 구조나, multicasting, multi-server 등 다양한 방법이 있을 수 있다. 사용자 수의 증가에 따라 증가하는 통신 량을 필요한 사용자에게만 제한하는 필터링 기법이 필요하게 되며, 이러한 문제는 응용 분야에 따라 결정되므로 SNHC 의 범주에서 벗어난다. SNHC 는 다운로드, 갱신, 오디오, object 상태 등의 네트워크를 통해 전송되는 내용물의 표준에 집중한다. 특히, 압축된 동적 상태의 데이터 표준화, object를 애니메이션 하거나, 특정 사용자에게 다른 사용자나 모델의 사건을 전달하기 위한 효율적 전송방법 등의 표준화는 매우 중요하다. 게임이나, 실시간 시뮬레이션, 분산 대화형 시뮬레이션 등에 사용되는 동적 파라미터 (dynamic parameter) 는 이와 관련이 있고, 동기화, scalability 등을 고려하여야 한다. 3차원 물체의 압축은 가상환경 모델이 사용자의 단말기에 어떻게 구현되는가에 따라 다양한 방법으로 사용된다. 간단한 경우, 사용자는 간단한 모델을 압축하여 다운로드 하여 사용할 수 있지만, 복잡한 경우 가상환경을 항해함에 따라 점차적으로 다운로드 받을 수도 있다. 그림 13은 분산 가상환경의 응용예를 보여주고 있는데 가상환경 하에서 사용자는 원격지에 있는 상대방이나 가상환경 내에 있는 object와 상호작용을 주고받는 것을 볼 수 있다.

5. 결 론

현재 세계 각국은 정보 선진화를 이루기 위하여 정보 고속도로를 구축 중에 있으며 이의 효율적인 활용은 매우 중요하다고 하겠다. 네트워크 상에서 만나는 가상 모임 (virtual meeting) 은 화상 회의와 비슷한 개념이지만 이는 얼굴 모델과 같이 물체의 모델을 기반으로 하여 효율적인 모양 변환 및 애니메이션을 사용하므로 값비싼 통신망의 대역폭을 효율적으로 사용할 수 있다. 더욱이 실제 영상만 사용하는 화상 회의에 비하여 실제 영상과 합성 물체를 동시에 제어하는 SNHC 는 향후 전개될 가상 세계에 대한 효율적인 접속 방법을 제공하게 된다. 가상모임 이외에도 이 기술은 만화 애니메이션이나 및 게임 등에도 이용 가능하다. 제작자는 캐릭터의 골격만 생성하고 이야기 속의 모든 움직임은 SNHC 기술을 이용하여 여러 가지 방법으로 다양하게 상황에 따라 생성할 수 있으므로 효율적으로 미디어를 사용할 수 있다.

분산 협동형 가상환경은 다양한 미디어 object를 통합하므로 다음과 같은 다양한 응용 분야를 창출할 수 있다.

- 멀티미디어 객체를 교환 할 수 있는 가상 원격회의
- 3 차원 디자인을 포함하는 모든 종류의 협동 작업
- 다수 사용자가 참가하는 게임
- 부동산, 가구, 자동차 등 다중 미디어 지원이 필요한 원격 상거래
- 원격 진단, 가상 수술 훈련 등의 의료용 응용
- 분산형 미디어 통합을 필요로 하는 가상 스튜디오 및 가상 세트

MPEG-4 는 이러한 개념들을 실현 시켜주는 다양한 기술의 표준안을 제공한다. 현재 첫 단계로 얼굴 및 사람 몸의 표현과 이들의 합성, 미디어의 통합에 중점을 두고 있다. 현재 SNHC VM (Verification Model) 이 개발되어 얼굴과 몸의 부호화, 문자-언어 합성, 미디어 통합, SNHC 오디오 등에 대해 규격이 제안되었다. 또한, 최근에는 구조체 압축, 얼굴 및 몸의 애니메이션, TTS 합성 등에 많은 제안이 들어오고 있으며, MSDL 기반 SNHC 실험을 위한 소프트웨어 요소 개발이 예상되고 있다. 앞으로 이러한 표준안들이 확정되면 인터랙티브 하이브리드 미디어를 중심으로한 분산 협동형 가상환경 응용 분야가 더욱 널리 확대될 전망이다.

참 고 문 헌

[1] J. Shade, D. Lischinski, D Salesin, T. DeRose, J. Snyder, "Hierarchical Image Caching for Accelerated Walkthroughs of Complex Environments," ACM Computer Graphics Proceedings, Siggraph 96, New Orleans, pp. 75-90, August 1996.

[2] A. Finkelstein, C. Jacobs, D. Salesin, "Multi-Resolution Video," ACM Computer Graphics Proceedings, Siggraph 96, New Orleans, pp. 281-290, August 1996.

[3] H. Hoppe, "Progressive Meshes," ACM Computer Graphics Proceedings, Siggraph 96, New Orleans, pp. 99-108, August 1996.

[4] G. Taubin, J. Rossignac, "Geometric Compression Through Topological Surgery," research report, IBM Research, RC-20340 (#89924), 22 pages, January 1996.

[5] M. Deering, "Geometry Compression," ACM Computer Graphics Proceedings, Siggraph 95, Los Angeles, pp. 13-20, August 1995.

[6] S. E. Chen, "QuickTime VR - An Image-Based Approach to Virtual Environment Navigation," ACM Computer Graphics Proceedings, Siggraph 95, Los Angeles, pp. 29-38, August 1995.
[7] L McMillan & G. Bishop, "Plenoptic Modeling: An Image-Based Rendering System," ACM Computer Graphics Proceedings, Siggraph 95, Los Angeles, pp. 39-46, August 1995.
[8] J. Torborg, J. Kajiya, "Talisman: Commodity Real-time 3D Graphics for the PC," ACM Computer Graphics Proceedings, Siggraph 96, New Orleans, pp. 353-363, August 1996.
[9] L. Williams, "Pyramidal Parametrics," ACM Computer Graphics Proceedings, Siggraph 83, pp. 1-11, July 1983.
[10] F. Lavagetto, D. Arzarello, M. Caranzano, "Lip-readable Frame Animation driven by Speech Parameters", IEEE Int. Symposium on Speech, Image Processing and neural Networks, Hong Kong, April 14-16, 1994.
[11] B. Pinkowski, "LPC Spectral Moments for Clustering Acoustic Transients", IEEE Trans. on Speech and Audio Processing, Vol. 1, N. 3, pp. 362-368, 1993.
[12] B. P. Yuhas, M. H. Goldstein Jr. and T. J. Sejnowski, "Integration of Acoustic and Visual Speech Signal Using Neural Networks", IEEE Communications Magazine, pp. 65-71, 1989.
[13] Petajan, E. D., "Automatic Lipreading to Enhance Speech Recognition" , Proceedings Globecom Telecommunications Conference, pp. 265-272, IEEE, 1984.
[14] Petajan, E. D., "Automatic Lipreading to Enhance Speech Recognition", Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 40-47, IEEE, 1985.
[15] Petajan, E. D. and Brooke, N. M. and Bischoff, G. J., and Bodoff, D. A. "An Improved Automatic Lipreading System to Enhance Speech Recognition," in "Proc. Human Factors in Computing Systems," pp. 19-25, ACM, 1988.
[16] Goldschen, A., "Continuous Automatic Speech Recognition by Lipreading," Ph.D., George Washington University, 1993.
[17] Goldschen, A. and Garcia, O. and Petajan, E., "Continuous Optical Automatic Speech Recognition," Proceedings of the 28th Asilomar Conference on Signals, Systems, and Computers," pp. 572-577, IEEE, 1994.
[18] Cohen, Michael M. and Massaro, Dominic W., "Modeling Coarticulation in Synthetic Visual Speech," Models and Techniques in Computer Animation, Springer-Verlag, 1993.
[19] Parke, F. I., "A Parametric Model for Human Faces", Ph.D., University of Utah, 1974.
[20] Parke, F. I., "A Model for Human Faces That Allows Speech Synchronized Animation", Journal of Computers and Graphics, 1975.
[21] Parke, F. I., "A Parameterized Model for Facial Animation", IEEE Computer Graphics and Applications, 1982.
[22] Barrus J. W., Waters R. C., Anderson D. B., "Locales and Beacons: Efficient and Precise Support For Large Multi-User Virtual Environments", Proceedings of IEEE VRAIS, 1996.
[23] Carlsson C., Hagsand O., "DIVE - a Multi-User Virtual Reality System", Proceedings of IEEE VRAIS '93, Seattle, Washington, 1993.
[24] Macedonia M. R., Zyda M. J., Pratt D. R., Barham P. T., Zestwitz, "NPSNET: A Network Software Architecture for Large-Scale Virtual Environments" , Presence: Teleoperators and Virtual Environments, Vol. 3, No. 4, 1994.
[25] Ohya J., Kitamura Y., Kishino F., Terashima N., "Virtual Space Teleconferencing: Real-Time Reproduction of 3D Human Images", Journal of Visual Communication and Image Representation, Vol. 6, No. 1, pp. 1-25, 1995.
[26] Singh G., Serra L., Png W., Wong A., Ng H., "BrickNet: Sharing Object Behaviors on the Net", Proceedings of IEEE VRAIS '95, 1995.
[27] D. Thalmann, T. K. Capin, N. Magnenat Thalmann, I. S. Pandzic, "Participant, User-Guided, Autonomous Actors in the Virtual Life Network VLNET", Proc. ICAT/VRST '95, pp. 3-11.
[28] N. I. Badler, C. B. Phillips, B. L. Webber, "Simulating Humans: Computer Graphics Animation and Control", Oxford University Press, 1993.
[29] Boulic R., Capin T., Huang Z., Kalra P., Lintermann B., Magnenat-Thalmann N., Moccozet L., Molet T., Pandzic I., Saar K., Schmitt A., Shen J., Thalmann D., "The Humanoid Environment for Interactive Animation of Multiple Deformable Human Characters", Proceedings of Eurographics '95, 1995.
[30] Funkhouser T. A., "Network Topologies for Scalable Multi-User Virtual Environments", Proceedings of VRAIS '96, 1996.
[31] Kalra P., Mangili A., Magnenat Thalmann N., Thalmann D., "Simulation of Facial Muscle Actions Based on Rational Free Form Deformations", Proc. Eurographics '92, pp. 59-69, 1992.
[32] Jon Ostermann, "An Interface for the Animation of Human Heads from Text" , Contribution No. MPEG96/M1197, October 1996 Chicago Meeting of ISO/IEC JTC1/SC29/WG11.
[33] SNHC Ad Hoc Groups, "Draft Specification of SNHC Verification Model 1.0", Document No. MPEG96/N1364, October 1996 Chicago Meeting of ISO/IEC JTC1/SC29/WG11.
[34] SNHC Face/Body Ad Hoc Group, "Face and Body Definition and Animation Parameters", Document No. MPEG96/N1365, October 1996 Chicago Meeting of ISO/IEC JTC1/SC29/WG11.

필 자 소 개

조 재 문

- *1984. 2. 서울대 전기공학과 졸업*
- *1986. 2. 한국과학기술원 전자공학과 석사 졸업*
- *1991. 8. 한국과학기술원 전자공학과 박사 졸업*
- *현재. 삼성전자 멀티미디어연구소 수석연구원*