Journal of the Ergonomics Society of Korea
Vol. 25, No. 2 pp.135-146, May 2006

멀티모달 인터랙션을 위한 사용자 병렬 모달리티 입력방식 및 입력 동기화 방법 설계*

임 미 정[1] · 박 범[2]

[1]아주대학교 미디어학과 / [2]아주대학교 산업공학과

Design of Parallel Input Pattern and Synchronization Method for Multimodal Interaction

Mi-Jung Lim[1], Peom Park[2]

[1]Department of Media, Ajou University, Suwon, 443-749
[2]Department of Industrial Engineering, Ajou University, 443-749

ABSTRACT

Multimodal interfaces are recognition-based technologies that interpret and encode hand gestures, eye-gaze, movement pattern, speech, physical location and other natural human behaviors. Modality is the type of communication channel used for interaction. It also covers the way an idea is expressed or perceived, or the manner in which an action is performed. Multimodal Interfaces are the technologies that constitute multimodal interaction processes which occur consciously or unconsciously while communicating between human and computer. So input/output forms of multimodal interfaces assume different aspects from existing ones. Moreover, different people show different cognitive styles and individual preferences play a role in the selection of one input mode over another. Therefore to develop an effective design of multimodal user interfaces, input/output structure need to be formulated through the research of human cognition. This paper analyzes the characteristics of each human modality and suggests combination types of modalities, dual-coding for formulating multimodal interaction. Then it designs multimodal language and input synchronization method according to the granularity of input synchronization. To effectively guide the development of next-generation multimodal interfaces, substantially cognitive modeling will be needed to understand the temporal and semantic relations between different modalities, their joint functionality, and their overall potential for supporting computation in different forms. This paper is expected that it can show multimodal interface designers how to organize and integrate human input modalities while interacting with multimodal interfaces.

Keyword: Multimodal interaction, Human-computer interaction, Human modality I/O design

*본 연구는 정통부주관 유비쿼터스컴퓨팅네트워크(UCN) 사업단 지원을 받아 수행되었음.
교신저자: 박 범
주 소: 443-749 경기도 수원시 영통구 원천동, 전화: 031-219-2428, E-mail: ppark@ajou.ac.kr

1. 서 론

멀티모달 인터페이스는 인간의 제스처, 시선, 손의 움직임, 행동의 패턴, 음성, 물리적인 위치 등 인간의 자연스러운 행동들에 대한 정보를 해석하고 부호화하는 인지기반 기술이다. 멀티모달 인터페이스는 인간과 컴퓨터 인터랙션 과정에서 동시 여러 모달리티의 입출력을 허용하며 다수의 모달리티의 조합과 입력신호 통합해석 등을 통해 상호 의사교환을 한다. 모달리티(Modality)란 인터랙션 과정에서 사용되는 커뮤니케이션 채널을 의미한다. 현재 시스템에서 휴먼-컴퓨터 인터랙션은 한 번에 한 가지 사용자 입력을 허용하기 때문에 하나의 오브젝트가 활성화되면 다른 오브젝트들은 모두 비활성화되어 두 개 이상의 오브젝트의 동시적 입력이 불가능하다. 그러나 멀티모달 인터페이스에서는 청각이나 촉각 등의 모달리티 입력은 동시 입력이 가능하여 여러 프로세스의 동시적 제어가 가능하다. 본 연구에서는 멀티모달 병렬 입력을 효율적으로 설계하기 위해 W3C(월드와이드웹컨소시엄), ETSI(유럽표준화기구)에서 발표한 멀티모달 시스템 프레임워크, 요구사항, 관련 기술에 대한 자료를 분석하고 여러 해외논문과 저널에서 발표된 모달리티들 특징들을 수집하였다. 그리고 수집된 자료를 바탕으로 멀티모달 인터랙션 과정에서 필요한 모달리티 결합 방법과 입력문법, 동기화 방법 등에 대해 설계하였다.

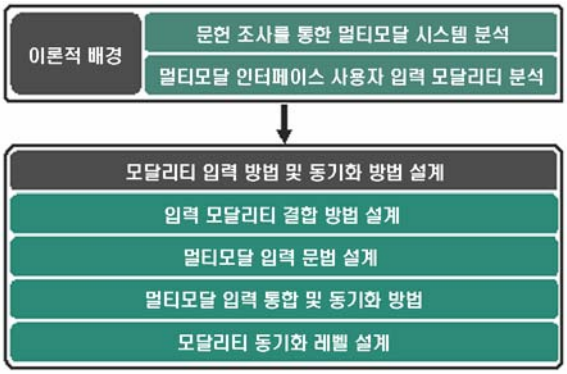

그림 1. 연구 절차

2. 연구배경

멀티모달 인터페이스는 MIT Media Lab.과 Microsoft, HP, Intel 등에서 연구 진행 중이며 음성인식, 표정인식, Haptic Interface, Body Tracking, DataGlove 기반 제스처 인식 기술 등을 중심으로 연구가 진행되고 있다. 기존 인터페이스는 한 번에 한 가지 입력만을 허용하며 일단 하나의 오브젝트가 선택되거나 활성화되면 다른 오브젝트들은 모두 비활성화되어 입력이 불가능하기 때문에 사용자의 작업 효율성에 한계가 있었다. 그러나 멀티모달 인터페이스에서는 청각과 촉각 등의 모달리티 입력들을 동시에 받아들일 수 있으며 모달리티 정보들을 분석하여 여러 개의 입출력 형태로 통합하여 이를 어플리케이션 프로세스에 병렬적으로 적용 가능하다. 멀티모달 인터페이스는 기존 직렬 입력방식과 달리 동시에 여러 다차원적인 입력 채널을 허용하기 때문에 기존 인터페이스와 구현상에서 본질적으로 차이가 난다.

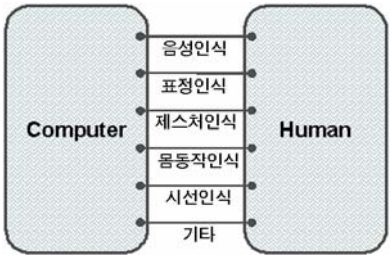

그림 2. 멀티모달 인터랙션 병렬 입력

그러나 멀티모달을 이용한 사용자 입출력방식에 관련한 형식화 설계방안이나 출력 메타포에 관해서는 아직 많은 부분에서 연구 보고되지 않은 실정이다. 현재까지 개발된 멀티모달 사용자 인터페이스는 주로 특정 어플리케이션의 일부 기능을 지원하기 위해 사용된 입출력 형식일 뿐 보편화될 수 있고 범용성 있는 입출력 표준 형식은 제시하지 못하였다. 이러한 이유는 아직 멀티모달 인터페이스가 대중들에게 보편화되지 않은데다가 새로운 기술들이 계속적으로 등장하고 있어 현재까지 사용자 입출력방식을 표준화하기에 어려움이 있기 때문인 것으로 사료된다.

3. 관련 연구

3.1 QuickSet(1997)

QuickSet은 미국 Oregon Health & Science University (Department of Computer Science)의 CSCC(Center for Human-Computer Communication)에서 P. R. Cohen과 S. L. Oviatt을 주축으로 개발되었다. QuickSet은 지도 기반의 군사전략 시스템으로 원거리 다수 사용자들이 음성, 펜 입력을 이용하여 군사전략에 필요한 사물위치 지정, 행위 지정 등의 직접 조작(direct manipulation)이 가능하다. QuickSet 입력방식 펜/음성의 입력은 동시에 이루어지는 데

예를 들면 펜을 사용하여 화면에 항로를 그리면서 동시에 "Whisky four six following this route -이 항로를 따라 위스키46"라고 음성으로 말하는 것이다. 또는 원하는 위치를 펜으로 가리키며 "red T71 platoon - 적군 T71 소대" 그러면 화면에 원하는 오브젝트에 대한 표시가 나타나게 된다. 펜으로 나타내는 이미지 패턴 분석은 Neural Network과 HMM(Hidden Markov Model)을 이용하여 분석하였다.

3.2 MVIEW(1998)

MVIEW(Multimodal Tools for the Video Analyst)는 미국 SRI International에서 개발한 것으로 음성과 제스처를 이용하여 주변일대에 관한 실시간 영상 감시와 제어가 가능한 시스템이다. MVIEW 역시 군사 감시 어플리케이션으로 영상에 보이는 특정 물체의 경계를 펜으로 그려 선택하고 음성으로 명령을 하면 그 물체에 관한 정보들이 기록된다. 사용자의 음성명령과 동시에 해당 타임라인에 해당하는 비디오 클립 영상 분석이 실시되고 펜 입력 위치에 해당하는 좌표 값과 입력시간 카메라 타입 등을 읽어내어 정보기록파일을 만들게 된다. 이 기록파일들은 원격지에 있는 다른 사용자에게 전송되어 상호간의 협동작업 및 정보교환과 Object Tracking이 가능하다.

3.3 GIVEN(1993)

GIVEN(Gesture-based Interactions in Virtual Environments)의 약자로 Virtual Reality와 Dataglove 기반의 Hand Gesture 인식 시스템이다(Väänänen & Böhm(1993)). Hand Gesture는 크게 정적인 포즈와 동적인 포즈로 나누어지며 Neural Network 기반 소프트웨어에 의해 20가지 Hand Gesture들로 구분된다. 동적인 포즈의 경우에는 5가지 단계의 포즈를 연속적으로 취함으로써 하나의 동적인 포즈를 완성하게 되는데, 이때 5단계가 다 마무리되기 전에 다른 포즈를 취하면 시스템은 동적 제스처를 인식하지 못한다.

3.4 SPIDAR-8(2002)

Walariacht et al.은 AR환경에서 force feedback device를 이용한 양손 인터랙션 방법을 제안했다. 3차원 프레임 각 모서리에는 각각 2개의 wire가 연결되어 있으며 양쪽 4개의 손가락에는 각각 3개의 wire를 연결되어 있다(총24개). SPIDAR-8은 사용자가 힘을 가해 늘어난 wire 길이를 이용하여 양 손가락의 위치를 계산하였으며 양손의 위치를 3D interface 상에 정확히 구현하였다. 그러나 고정 프레임으로 인해 시스템의 mobility가 떨어졌으며 프레임 내 좁은 공간에서만 작업이 가능하였기 때문에 사용자의 행동의 제약이 따랐다.

3.5 Haptic and Vibrotactile Feedback using DataGlove

DiZio와 Lackner(2002)는 VR환경에서 피부자극을 이용하여 조작 시 정확한 포인팅 값을 유도하려 하였으며 결과적으로 VR Object들에 대한 정확한 포인팅 결과 값을 얻을 수 있었다. 또한 Klatzky와 Lederman(2003) 피부에 Vibrotactile Feedback을 이용한 인터랙션을 구현하였으며, Hauber(2002) 역시 Buzzer를 이용한 6DOF Tactile Pointer를 구현하였으며, 두 번째 버전으로는 압전기(壓電氣)를 이용한 Bending Actuator를 개발하였다.

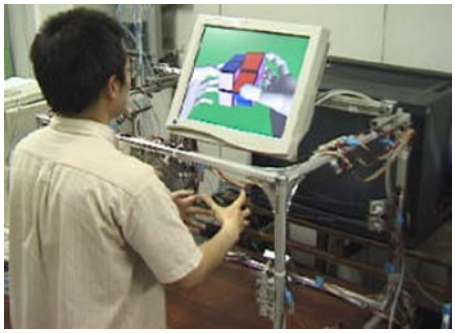

그림 3. SPIDAR-8 Multimodal Syste

3.6 Map-based System Using Speech and 3D Gestures

사용자가 맵 기반의 협동적인 인터랙티브 시뮬레이션을 생성하고 조절하도록 하기 위한 음성/제스처 멀티모달 시스템이다. 사용자는 Gloved Hand Gesture를 이용하여 지도의 직접 조작과 음성명령이 가능하다.

글러브는 그림 4와 같이 3차원의 x, y, z축을 가지며 x축을 기준으로 한 회전 값을 추가로 갖는다. 검지손가락이 가리키는 x축 방향의 벡터 연장선과 화면에서 교차하는 지점은 화면상 오브젝트 위치를 결정하는데 사용되며, y와 z축 이동 값은 오브젝트를 이동하고 명령을 주는데 사용된다. 데이터 글러브의 움직임에 대한 그래프는 그림 4에 나와 있는 것과 같으며 연속적인 손의 움직임을 articulation하기 위해서 움직임에 대한 회전각도의 미분 값을 사용하였다. 이는 현재 활성화된 오브젝트가 어떤 것인지 어플리케이션이 알 수 있도록 하는 데이터로 사용되었으며 회전 각도의 미분 값이 변하는 지점에서 오브젝트 음성명령적용이 가능하도록 하였다.

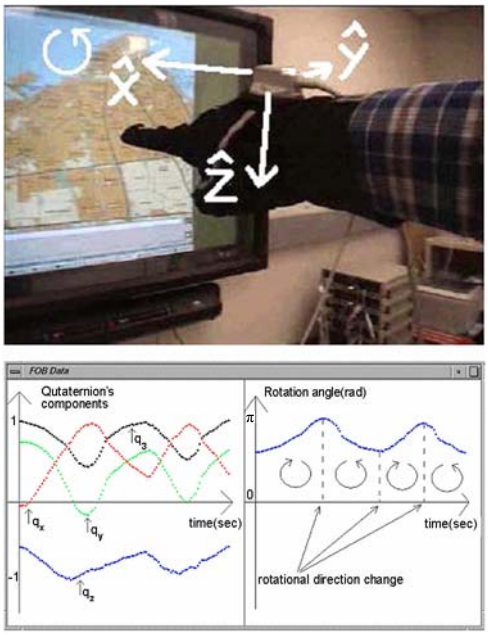

그림 4. Map-based Multimodal System

4. 입력 모달리티 분석

4.1 입력 모달리티

입력 모달리티란 인간이 컴퓨터에 입력하는 정보들의 설계를 말한다. 멀티모달 인터페이스를 구현함에 있어서 가장 중요한 것이 모달리티들 간의 결합방식이다. 어떻게 각각의 모달리티를 결합해야 인간의 수행능력을 향상시키면서 쉽게 컴퓨터로 의사전달을 할 수 있을까? 이를 설계하기 위해서는 먼저 각각의 모달리티들이 전달하는 정보의 특징들을 파악하고 이들이 각각 어떠한 정보를 전달하는데 더 효율적인지 비교평가가 필요하다. 그리고 결합된 모달리티들을 바탕으로 연속적이고 불규칙적이 멀티모달 입력을 체계화할 필요가 있다. 현재까지 멀티모달 인터랙션을 위한 입력 형식들이 여러 가지 제안되어 왔지만 아직까지 멀티모달 인터페이스를 상용화하기에는 시기상조라 입력 형태에 대한 표준안이 마련되어 있지는 않은 상태이다. 본 연구에서는 현재까지 연구되어 온 모달리티 사용에 관한 자료들을 바탕으로 멀티모달 입력방식을 형식화해 보고자 한다. 멀티모달 인터페이스에서는 기존에 마우스와 키보드의 입력방식이 아닌 보다 더 폭넓은 입력방식을 제안하고 있다. 표 1은 시각과 청각, 촉각 등에 해당하는 입력정보와 처리 장치에 대해 정리한 표다.

표 1. 입력 모달리티와 입력 디바이스

Modality	I/O Data	Devices
Motion Pattern	Typing Handwriting Touch/Pointing Pushing and Clicking Gloved Hand Gesture Body Movement Hand Movement	Keyboard Stylus/Mouse DataGlove Tablet/TouchPad RollerBall LaserPointer Camera Magnetic Tracker
Visual	Lip Movement Eye Movement Body Movement Hand Movement Arm Movement Facial Expression Free Hand Gesture	Eye-Tracker Scanne Camera
Acoustic	Speech	Microphone Speech Recognizer
Tactile Force	Hand Pressure	Haptic Devices
Biometric	Fingerprint Brain Activity	Card Reader EEG/EMG

4.2 입력 모달리티 특징

입력 모달리티들의 특징들을 분석하면 다음 표 2와 같다. Oviatt의 연구에서 제안한 바와 같이 모달리티들을 Active Mode와 Passive Mode로 구분하였다.

5. 입력 모달리티 결합

멀티모달 인터랙션을 하기 위해서 사용자는 각각의 입력 모달리티들을 언제 어떻게 결합하여 사용할지 알고 있어야 한다. 사용자가 결합하여 사용하는 모달리티들은 사용자가 원하는 정보들을 충분히 컴퓨터에 전달할 수 있어야 하며, 각각의 모달리티들이 나타내는 정보들은 서로 상호보완적이여야 한다. 만약 모달리티들이 전달하는 정보가 중복이 되거나 상호보완적이 아닐 경우 사용자는 모달리티 입력에 혼란을 겪게 되거나 전달하고자 하는 정보를 효율적으로 전달하기 힘들게 될 것이다. 본 연구에서는 모달리티들의 심상을 크게 공간적 심상과 언어적 심상으로 구분하고 각각의 심상을 갖는 모달리티들을 결합하여 Blended input 형태를 제안하였다. 이러한 결합 형태는 특정 어플리케이션에서 이루어

표 2. 입력 모달리티 특징

Modalities	Mode	Remarks
Handwriting	Active	짧은 글 입력 시 사용, 문법 또는 입력형식 필요. 긴 글 입력 시 비효율적
Touch Pointing	Active	공간 좌표 표시, 정밀한 포인팅 가능
Hand Gesture	Active	Hand writing과 사용불가. 제스처 부호화. 대강의 위치포인팅. 원격제어
Speech	Active	문장 입력에 효율적. 눈에 보이지 않는 사물을 묘사하는데 사용 가능
Body Movement	Passive	행동 패턴을 이용하여 인지. 사용자의 행동추적을 통한 상황인식 가능
Lip Movement	Passive	음성의 보조 모달리티 사용 가능
Eye Movement Eye-Gaze	Passive	오브젝트 선택 시 사용자의 시야범위를 이용하여 시야 클리핑, 선택 가능한 오브젝트의 n-best list 추출
Arm Movement	Passive	Gloved Hand Gesture 보조 모달리티, 적은 량의 정보 전달(약 2bit)
Facial Expression	Passive	음성의 보조 모달리티 사용 가능 개인성향에 따른 인식 수준 차이
Brain Activity	Passive	무구속적인 인터랙션 가능, 다른 입력 모달리티들과 병렬적으로 사용할 경우 인터랙션 자체가 뇌파를 변화시킴. 물리적인 멀티모달 결합에 어려움.
Fingerprint	Passive	사용자 분별 및 인증에 사용

지는 공간작업과 언어작업의 상대적인 작업비율에 따라 각 심상정보를 전달하는 모달리티 사용 빈도수도 비례하여 나타날 것이다.

5.1 공간적 심상과 직선적 배열

인지심리학에서 '이중부호론(dual-code theory)'이라 하여 잘 알려진 이론(Bower, 1972; Paivio, 1971)은 공간적 그리고 직선적 지식 표상과 관계한다. 이중부호론은 공간적 부호를 시각적 모달리티에, 직선적 부호를 언어적 모달리티에 결부시킨다. 이중부호론은 두 종류의 표상을 두 모달리티와 관련시키는 데 있어서 상당한 논란의 대상이 되었다.

산타(Santa, 1977)의 실험에서는 이러한 공간적 표상과 직선적 표상의 차이를 잘 설명해 준다. 도형 조건에서 피험자들은 세 도형들이 두 개는 위에, 한 개는 아래에 위치해 있는 공간 배열을 학습했다(그림 5, 그림 6). 피험자들은 배열을 학습한 후 배열은 제거되었고 즉시 여러 감사 배열들 중 하나를 제시받았다. 피험자들의 과제는 검사 배열이 학습 배열과 같은 공간 형태를 가지지는 않더라도 동일한 요소들을 포함하고 있는지 검증하는 것이었다.

산타는 이 실험에서 도형들은 공간적으로 배치할 경우 판단이 더 빠르고 단어들을 직선적으로 배열할 경우 판단이 더 빠르다는 결론지었다. 그리고 도형정보는 공간위치에 따라 부호화되어 저장되며 단어정보는 직선적 순서에 따라 부호화되어 저장되는 경향이 있다는 것을 알 수 있었다.

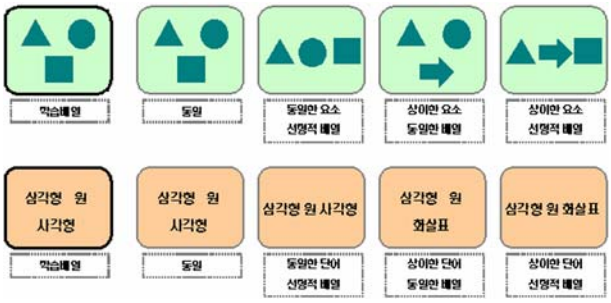

그림 5. 산타(Santa, 1977)의 실험

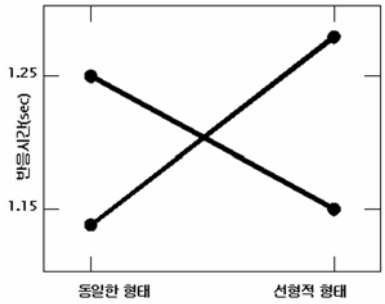

그림 6. 산타의 실험 반응시간 결과

산타의 실험에서 살펴보면 공간적 심상과 직선적 배열은 원래의 자극에서 추상화되어 원 자극의 몇몇 구조만을 인지과정 속에 보존한다. 공간적 심상은 공간상 대상들의 위치를 보존하며 직선적 배열은 사물들의 서열을 보존한다. 공간적 심상과 직선적 배열은 작은 단위들이 보다 더 큰 단위 내에 청크(chunk)가 되는 위계 구조로 다시 부호화될 수 있다. 예를 들어 피험자가 아래의 도형 배열이 사람의 얼굴과 같이 생겼다고 인식했다면 도형의 배열은 낱개가 아닌 전체적인 하나의 얼굴의 형상으로 기억될 것이다. 마찬가지로 사용자가 여러 개의 단어들을 보고 하나의 심상을 떠올렸다면 이것 역시 하나의 청크에 저장 가능할 것이다. 이처럼 공간적 심상과 직선적 심상을 이용한 이중부호화를 통해 제한된 인지 처리용량을 더 효율적으로 사용할 수 있다.

5.2 모달리티 결합

표 3은 표 2를 바탕으로 결합 가능한 모달리티를 묶어 구분한 것이다. Speech는 Lip Movement와 Facial Expression, Eye Gaze 등의 Passive Input과 함께 Blended In-

put으로 결합될 수 있으며 이는 주로 언어적인 심상정보를 전달한다. Hand Gesture는 Arm Movement와 Eye Gaze, Hand Pressure 등과 결합하여 공간적인 정보를 전달한다. Speech는 Hand Writing과 기능 상 많은 부분이 중복되며 Gloved Hand Gesture도 Touch와 Hand Writing과 기능 상 많은 부분이 중복된다. 또 Gloved Hand Gesture는 Touch, Hand Writing과 조작 방법이 겹치기 때문에 함께 사용할 경우 많은 부분의 조작 과정이 겹치게 된다. 한편 Eye-Gaze와 Body Movement, Physical Location 등의 사용자 상황 정보는 사용자와 공간 간의 인터랙션(User-Space Interaction)이 가능하도록 돕는데 사용된다.

표 3. 사용자 입력 모달리티 조합

Input mode	Characteristics	Image	Modalities
Blended Mode	언어정보 전달	직선적	Speech + Lip Movement + Facial Expression + Eye Gaze
Blended Mode	공간정보 전달	공간적	Hand Gesture + Arm Movement + Eye Gaze
Active Mode	언어정보 전달	직선적	Hand Writing
Active Mode	공간정보 전달	공간적	Touch
Passive Mode	상황정보 전달	공간적	Eye Gaze, Body Movement, Physical Location

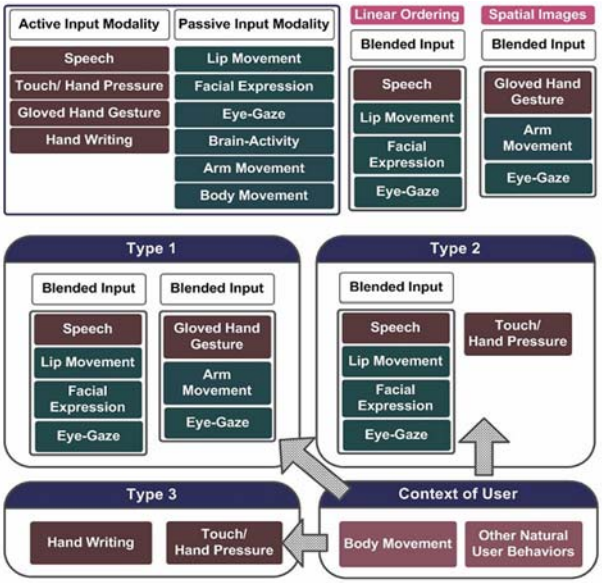

그림 7. 사용자 입력 모달리티 조합

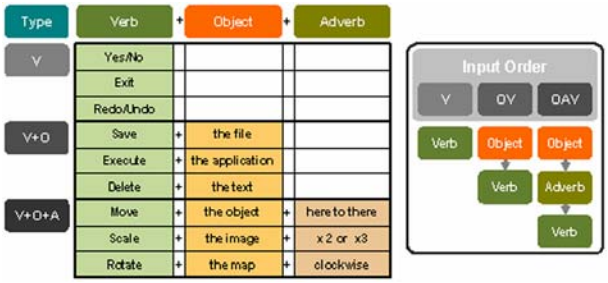

그림 8. 멀티모달 입력 문법과 입력 순서

6. 멀티모달 입력문법

모달리티 조합 방법 외에 고려해야 할 요소는 입력의 동시성이다. QuickSet, MVIEW 등 대부분의 시스템의 경우에서 디폴트 형태로 동시적 입력 방법을 채택하고 있으며, 사용자 옵션으로 비동시 입력을 허용하고 있다. 본 절에서는 동시적, 비동시 입력을 위한 멀티모달 입력문법 [OAV]형식을 제안한다. 모달리티로는 Speech와 Touch, Gesture를 사용하였다.

6.1 [OAV] 입력문법

[OAV] 형식은 그림 8 오른쪽 그림에서 나타낸 것과 같이 멀티모달 언어입력 순서를 바탕으로 붙여진 이름이다. O는 오브젝트(Object)를 뜻하며, A는 속성(Adverb), V는 행위(Verb)를 의미한다. 영문형식을 기준으로 멀티모달 언어의 문장구조를 크게 '동사(V)', '동사 + 목적어(V+O)', '동사 + 목적어 + 부사(V+O+A)'로 구분하였으며 각각의 모달리티 입력순서는 그림 8의 오른쪽 그림과 같다.

많은 종류의 명령어들이 객체(object)에 행위(action)를 적용시킨다. 예를 들어 글자의 크기를 변경하고자 하는 경우 사용자가 크기를 변경하고자 하는 글자를 선택한 후에 글자의 크기를 변경할 것이다. 포토샵과 같은 그래픽 툴에서는 행위에 대한 명령을 선택하고 객체를 선택하기도 한다. 예를 들면 이미지를 이동시키고자 할 경우 "Move" 툴을 선택하고 이동하고자 하는 객체를 이동한다. 이 경우에는 세밀한 조작을 필요로 하거나 정교한 수정을 해야 하는 경우 이미지 속성 값이 여러 번 적용될 수 있기 때문에 속성 값은 맨 나중에 지정하는 것이 더 효율적이다. 입력 순서는 가장 잦은 수정 빈도를 나타내는 명령을 가장 나중에 입력하고 가장 적은 빈도의 명령은 제일 처음에 입력하도록 하는 것이 바람직하다. 본 연구에서는 오브젝트 객체를 먼저 선택하고 후에 행위명령을 입력하는 형식으로 멀티모달 언어를 설계하였다. 만약 어플리케이션이 포토샵과 같은 그래픽 툴과 같이 잦은

'속성' 변경과 '동작' 명령을 수행하는 경우에는 입력 순서를 조정해야 할 필요가 있을 것이다. [VOA] 또는 [VAO]

6.2 Touch, Gesture, Speech를 이용한 멀티모달 입력 방식

그림 9은 Touch와 Gesture, Speech를 이용한 멀티모달 입력 예이다. Speech는 Name Tag 기반으로 작동하며, Touch는 메뉴클릭기반으로, Gesture는 이동커서 및 제스처 명령기반으로 작동된다. 만약 사용자가 '되돌리기(Undo)' 명령을 입력하고자 할 경우 입력해야 할 명령어 형식은 [V]에 해당할 것이다. 각 모달리티를 사용하여 '되돌리기' 명령을 입력하는 방법은 Touch를 사용하여 '되돌리기' 메뉴를 클릭하거나 Speech를 사용하여 '되돌리기' 명령을 입력하거나 Gesture는 '되돌리기' 명령으로 부호화된 Gesture를 입력할 수 있을 것이다.

또한 사용자가 만약 세 번째 [OAV]형식에 해당하는 '지도를 2배로 확대시켜라'는 명령을 수행할 경우 사용자는 Touch나 Speech로 해당 오브젝트(지도)를 선택한 후 Speech나 Gesture, Touch 등으로 '2배'를 입력하고 마지막에 마찬가지로 Speech나 Gesture, Touch 등을 사용하여 '확대하라' 명령을 수행할 수 있을 것이다.

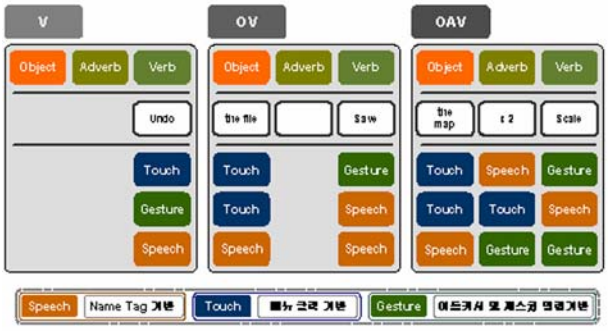

그림 9. 멀티 모달리티 입력 예(Touch, Gesture, Speech 사용)

6.3 Gesture, Speech, Touch 멀티모달 입력 비교

- Speech와 Touch는 오브젝트를 선택하기 용이한 반면 Gesture는 그렇지 못하다.
- Speech와 Touch를 사용하여 오브젝트를 가리키는 것은 '오브젝트를 선택하다'의 동사(Verb) 의미를 내포한다.
- Gesture의 경우 [Roll & Pitch정보]+[Gesture]를 이용하여 오브젝트 방향과 관련된 '속성(Adverb)' 정보와 '행위(Verb)' 정보를 동시에 나타낼 수 있다.
- 속성정보와 행위정보를 적용할 오브젝트가 한 개밖에 없는 경우에는 오브젝트를 선택하지 않아도 인터랙션 매니저에 의해 자동적으로 해당 오브젝트에 명령이 적용되도록 한다.
- Speech의 경우 음성인식기의 단어 인식 범위에 따라 오브젝트와 속성, 행동을 하나의 명령어처럼 입력할 수 있다.
- Touch와 Gesture는 동시 입력이 불가능하며 연속적으로 사용한다 하더라도 두 모달리티 모두 손을 사용하여 입력되기 때문에 사용자가 입력방식에 혼란을 겪을 수 있다.

6.4 Touch, Gesture, Speech를 이용한 Object Selection

Speech, Touch와 Gesture 멀티모달 입력을 이용한 Object Selection하는 방법들을 살펴본다. 사용자가 Speech, Touch를 사용하여 오브젝트를 선택할 경우 Speech이나 Touch 둘 중에 어느 하나를 사용하더라도 동일하게 효과적인 수행이 가능하다. 예를 들어 사물에 태그(다른 태그들과 사용에 있어 구별되어야 함)가 제시되면 태그는 음성인식 어휘를 할당하여 오브젝트 선택이 가능하며, 터치나 펜 입력을 사용하여서도 동일한 오브젝트 선택이 가능하다. 오브젝트에 태그(Tag)를 달거나 손으로 가리키는 구체화된 행동을 통해서 사물의 고유한 identification을 충분히 나타낼 수 있고 애매하거나 잘못된 참조를 피할 수 있다. 그러나 제스처 명령의 경우 자유로운 오브젝트 선택이 어렵다. 이유는 제스처는 사용자가 행동하기 원하는 작업을 기준으로 부호화되는데 온갖 오브젝트들을 제스처로 부호화하기는 힘들기 때문이다. 제스처를 사용하여 오브젝트를 선택하기 위해서는 새로운 방식의 메커니즘이 설계되어야 한다. 예를 들면 Roll & Pitch를 이용한 포인터 네비게이션 방법이 있을 것이다. 이 외에 Passive input modality들과 결합한 Blended input mode도 가능한데 Passive modality로 동공의 위치나 머리의 방향 움직임 값을 사용하여 오브젝트의 위치를 추적하고 해당 오브젝트 위에 포인터가 위치할 때 제스처를 입력하는 방법도 가능하다. Gesture 대신에 Speech를 사용해서 동공의 움직임과 Speech를 이용한 Object Selection도 가능할 것이다. 이러한 경우 시선을 고정시키고 "이것", "저것" 등과 같은 지시적 음성표현만을 이용해서도 오브젝트 선택이 가능할 것이다.

6.5 Touch, Gesture, Speech를 이용한 Menu Navigation

메뉴 선택을 하기 위한 멀티모달 입력을 형식화하는 것도

앞에서 설명한 Object Selection과 비슷하다. Speech와 Touch는 메뉴입력에 효율적으로 사용이 가능하지만 Gesture의 경우 새로운 입력 메커니즘을 필요로 한다. 예로 사용자가 Speech와 Touch를 이용하여 풀다운 메뉴를 선택하고자 한다고 가정하자. Speech로 풀다운 메뉴를 선택하는 경우 시각적으로 메뉴 표시를 보지 못할 것이기 때문에 사용자가 선택하려는 메뉴 옵션에 대한 사전지식이 있어야 한다. 만약 사용자가 메뉴에 대한 사전지식이 있는 경우 여러 개의 Menu Depth를 거치지 않고서도 한 번에 효율적으로 자신이 실행하고자 하는 메뉴를 선택할 수 있다. 음성은 계층메뉴구조 또는 Direct Menu Access와 관련해서 발생하는 행동제어 문제들을 해소시킬 수 있다. 극단적인 예를 들면 스크린과 사용자가 멀리 떨어져 있거나 손을 제대로 사용하지 못하는 경우 Speech는 Touch에 비해 탁월한 성능을 보일 것이다.

하지만 초보 사용자에게 음성은 펜과 거의 동일한 수행효과를 보이거나 그보다 못한 수행 효과를 보일 수 있다. 이유는 초보자는 Menu Depth 정보를 알지 못하기에 모든 메뉴를 일일이 다 확인해야 하고 Speech의 경우 Touch 입력에 비해 인식정확도가 떨어지기 때문이다. 초보자는 Direct Menu Access가 불가능하기 때문에 정상적으로 올바른 수행을 한다고 할 경우에도 Speech는 Touch과 거의 동일한 수행 효과를 보인다. 또 Speech의 경우 에러 발생 시에 이전 상태로 되돌리는데 Touch 입력과 비교해서 에러 발생확률이 더 많기 때문에 수행속도가 더 떨어질 수 있다. Touch는 사용자로 하여금 메뉴 옵션 경로를 더욱 더 분명하게 알도록 하기 때문에 초보 사용자들의 계층구조 메뉴 네비게이션 입력 매개체로 선호된다.

Gesture를 이용하여 메뉴 네비게이션을 하기 위해서는 새로운 메커니즘이 필요하다. 본 연구에서는 '이동커서'를 이용한 메뉴 네비게이션을 제안한다. 알기 쉬운 예로 일반적인 윈도우 창에서 '탭(Tab)'키를 누르면 선택 가능한 버튼들이 돌아가면서 활성화되는 것을 볼 수 있을 것이다. 이것을 Gesture 입력에 적용하여 Gesture 입력이 한 번 적용할 때마다 Tab키와 같이 해당 메뉴들이 활성화되고 이 후 사용자가 2차로 선택 Gesture를 입력할 수 있도록 할 수 있을 것이다. W3C 표준화 자료에 의하면 인터랙션 레벨은 Session> Page> Form> Field> Event로 구분될 수 있다. 이는 수직적인 인터랙션 레벨이동을 의미하는데 사용자는 Gesture를 이용하여 각 Session 간, Page 간, Form 간 등의 수평이동이 가능하며 각 레벨 간 수직이동이 가능하다. 그림 10은 인터랙션 레벨에 따라 Gesture 메뉴 네비게이션 과정을 도식화한 것이다. Gesture1은 수평적 이동을 나타내며 Gesture2 & Gesture3은 수직적 이동을 나타낸다.

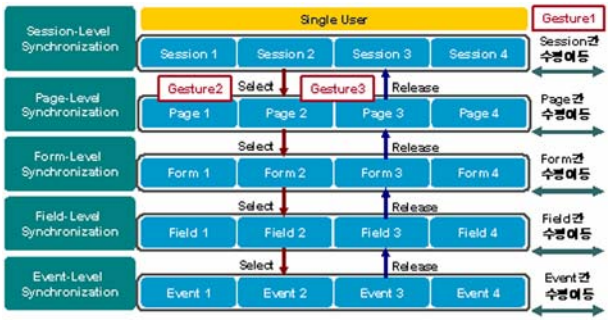

그림 10. Gesture를 이용한 메뉴 네비게이션

6.6 Touch, Gesture, Speech를 이용한 Object Manipulation

Touch와 Gesture, Speech를 이용하여 가상공간에서 오브젝트를 직접 조작하는 방식에 대해 살펴본다. 먼저 Touch를 사용하여 사물을 직접 조작하는 경우 윈도우 폼 안에서 이동하고자 하는 방향으로 손가락을 움직여 오브젝트를 이동시킬 수 있을 것이다. 또는 오브젝트의 각 모서리를 잡고 변경하고자 하는 크기만큼 오브젝트를 늘리거나 축소시키거나 회전시킬 수 있을 것이다. 마찬가지로 Gesture를 이용해서도 오브젝트의 효과적인 직접 조작이 가능한데 Roll과 Pitch, Static Gesture, Dynamic Gesture 등을 이용한 다차원적인 오브젝트 조작이 가능하다. Gesture의 경우 3D 가상환경에서 오브젝트 직접 조작에 탁월한 수행능력을 보인다.

Speech 명령을 사용하여 오브젝트를 디폴트 값의 크기와 위치를 갖는 오브젝트를 생성할 수도 있을 것이다. Speech 입력으로는 구체적인 공간 상에 위치나 크기를 정하기 어렵기 때문에 직접 조작의 입력 모달리티로는 적합하지 않다.

7. 입력 통합

멀티모달 입력 동기화 형식은 크게 Sequential multimodal input과 Simultaneous multimodal input, Composite multimodal input으로 구분되며 각각의 모달리티 통합 방법을 비교하여 표시하면 그림 11과 같다.

- Unimodal Input 단일 모달리티 입력.
- Simultaneous multimodal input 두 가지 이상의 명령어를 동시에 입력하여 어플리케이션의 병렬작업이 가능하다.
- Composite multimodal input Simultaneous multimodal input과 유사하지만 각각의 모달리티 정보가 결합하여 하나의 어플리케이션 명령을 수행한다.

• Sequential multimodal input 순차적으로 모달리티를 입력하는 것으로 두 개 또는 그 이상의 모달리티들이 합쳐져 하나의 명령어를 이룬다. 인터페이스 디자이너는 모달리티들 간의 결합이 가능한 time interval을 명시해주어야 한다.

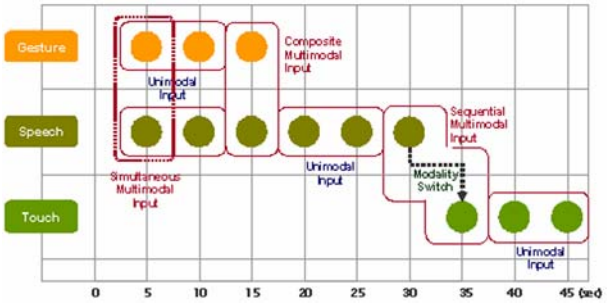

그림 11. 입력통합 및 동기화

8. Dual-Prompt

그림 12은 멀티모달 입력 동기화 레벨에 따라 인터랙션을 모형화한 것이다. 멀티모달 입력은 Simultaneous multimodal input을 이용한 병렬작업이 가능하므로 이를 수용할 수 있는 Dual-Prompt가 필요하다. Dual-Prompt는 병렬작업을 진행하기 위한 어플리케이션의 준비상태라 볼 수 있다. 병렬작업 과정에서 동시 처리되는 프로세스는 서로 종속관계가 아닌 상호 대등관계, 독립적인 관계여야 한다. 종속적인 프로세스의 경우 병렬로 신호를 입력할 경우 하부프로세스의 명령과 상위프로세스 명령 간에 충돌이 일어날 수 있기 때문이다. 가장 안정적으로 Dual-Prompt를 설계하려면 두 개 이상의 어플리케이션에서 Prompt가 동시에 작동하도록 할 수 있을 것이다. 경우에 따라 하나의 어플리케이션 상에서도 Page, Form, Field, Event 레벨에서의 Dual-Prompt가 가능하다.

따라서 본 연구에서는 Process가 이루어지는 Page-Level을 Dual-Prompt 처리 단계로 지정하였다. 사용자는 결과적으로 Simultaneous multimodal input을 사용하여 2개 이상의 Page에서 멀티모달을 이용한 병렬작업이 가능할 것이다. 하나의 Page에서도 프로세스가 독립적으로 처리된다면 Simultaneous multimodal input과 Dual-Prompt 사용이 가능할 것이다.

Simultaneous multimodal input의 경우 주의 자원(attention resource)을 효율인 분배가 가장 큰 문제가 된다. 이를 위해서 디자이너는 사용자가 양립적인 개념을 갖고 모

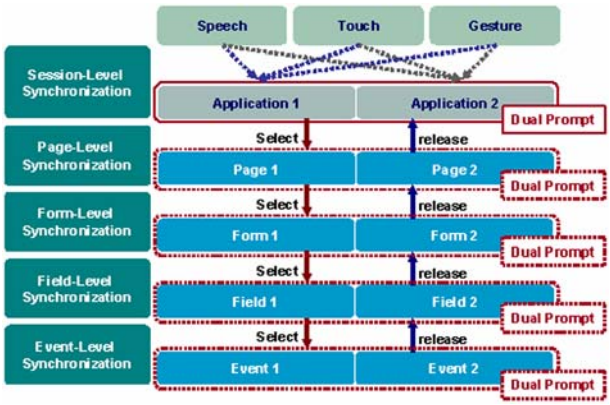

그림 12. 인터랙션 동기화 레벨과 Dual-Prompt

달리티 입력을 할 수 있도록 인터페이스를 설계해야 할 것이다. Dual-Prompt & Dual-Task 간에는 개념적인 양립성을 유지하고 있어야 하며, 사용자가 손으로 Image를 조작하면서 동시에 Speech로 음성을 입력하는 행위와 같은 동시적 병렬작업이 가능해야 할 것이다. 그림 13은 앞에서 설명한 입력 모달리티 레이어와 동기화 레벨, 입력 통합 방법 등을 이용하여 3차원 상 좌표상에 표시한 것이다. X축은 시간을 나타내며, Z축 양의 방향은 Multi-Layered Input, Z축 음의 방향은 모달리티 동기화 레벨, Y축은 Dual-Prompt(입력 명령의 수)를 나타낸다. Composite multimodal input과 Sequential multimodal input, Simultaneous multimodal input 각 입력통합 방법을 좌표 공간 내에 나타내었다.

9. 인터랙션 메커니즘 설계

인터랙션 메커니즘은 멀티모달 입출력을 위해 정형화된 입력 패턴이라 볼 수 있다. 본 연구에서는 다음과 같은 인터랙션 메커니즘을 이용하여 멀티모달 입력방식을 형식화(Formulation)하고자 하였다. 그림 14는 홈네트워크 시스템 제어 시나리오를 바탕으로 한 3D 오브젝트 기반 멀티모달 인터페이스의 인터랙션 과정을 표시한 것이다.

- Dual Prompt: Simultaneous multimodal input을 위한 Dual-prompt로 이를 통해 모달리티 병렬 입력이 가능하다.
- Input Grammar: 모달리티 입력 순서와 모달리티 조합 방식 등에 관한 입력 규칙이다.
- Input Modality Switch: 사용자 모달리티 입력 전환으로 모달리티 입력에 에러가 발생하거나 입력 모달리티를 대체 모달리티로 전환해야 할 때 발생한다.
- Direct Menu Access: 음성과 제스처의 경우 트리구조

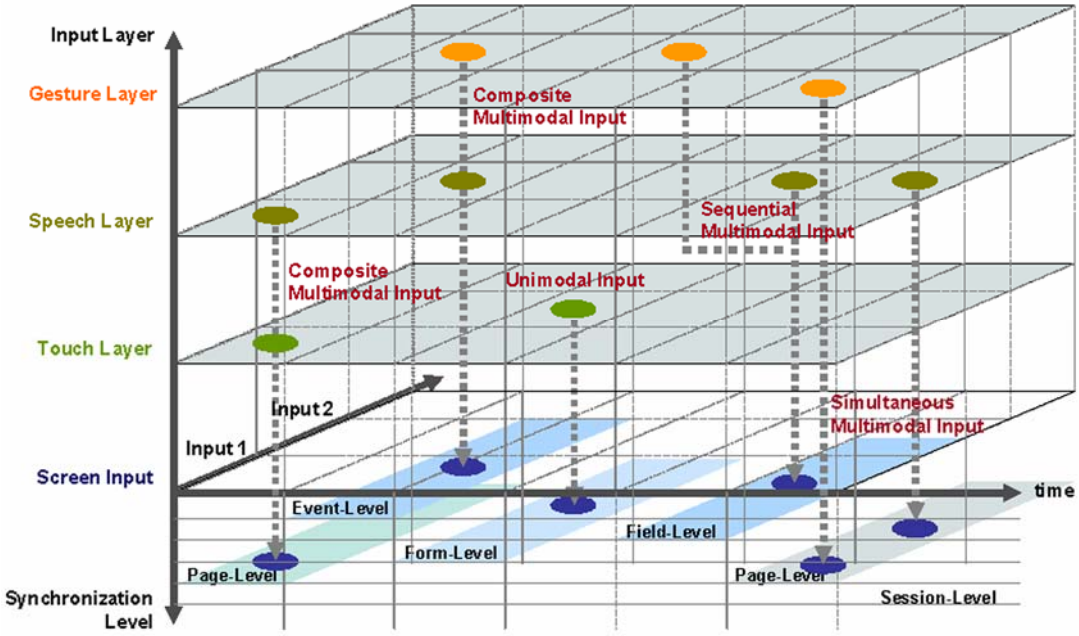

그림 13. 멀티모달 통합패턴과 동기화 레벨

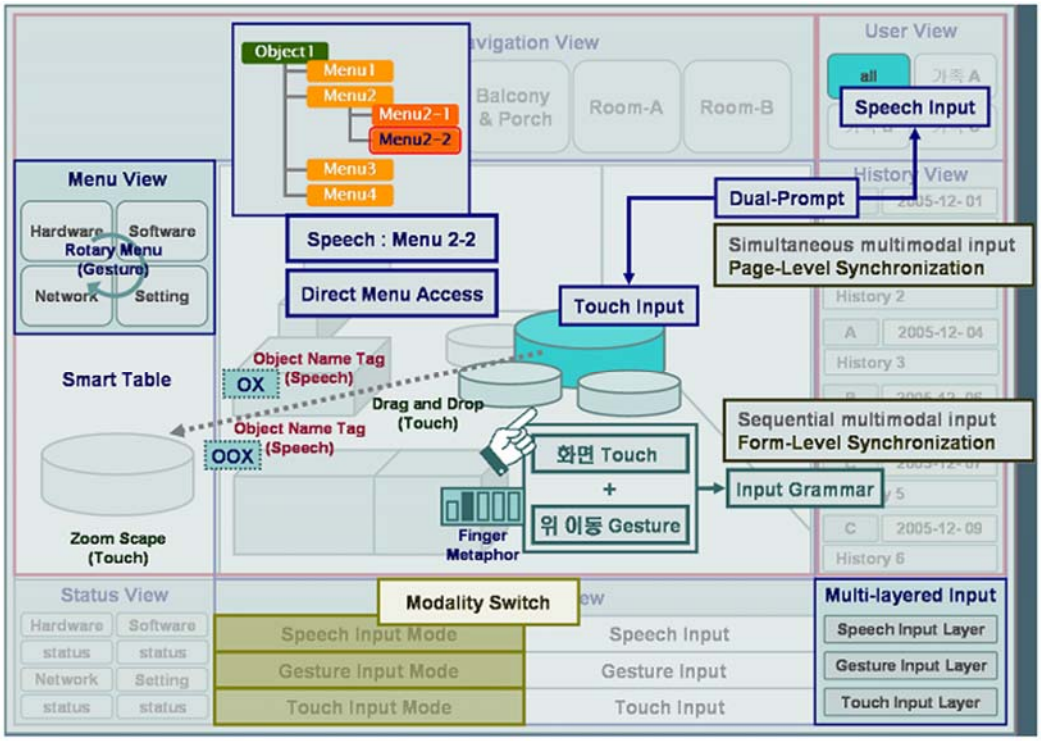

그림 14. 3D 오브젝트 기반 멀티모달 인터페이스 구현 화면

의 2차, 3차 Depth 메뉴의 직접적인 접근이 가능하다.

- Drag & Drop/ Zoom Scape: 오브젝트를 끌어다 놓으면 오브젝트의 크기가 확대되는 영역으로 3D 오브젝트 기반 인터페이스에서 일관성 있는 메뉴 선택과 화면레이아웃 제공이 가능하다.
- Rotary Menu: 제스처를 이용하여 메뉴를 선택하기 위

한 회전식 메뉴로 제스처는 활성화된 커서를 이동하는 제스처와 메뉴를 선택하는 제스처로 구분될 수 있다.

- Modality Input Layer: 여러 모달리티 입력을 동시적으로 입력받기 위한 모달리티 입력 레이어 계층이다.
- Object Name Tag: 음성 인터페이스에서 오브젝트 선택을 위한 것이다. 필요 시에만 표시가 되도록 할 수 있다.
- Spatial Images: 화면을 직접 보지 않고서도 기억 속 메뉴 심상을 이용하여 제스처 메뉴 네비게이션이 가능하다.

10. 토론 및 향후 연구

본 연구은 멀티모달 인터페이스를 개발하기 위한 휴먼-컴퓨터 인터랙션 설계방안과 입출력 설계에 중점을 두었다. 현재까지 HCI적인 관점에서 멀티모달 인터페이스를 체계적으로 분석한 자료가 많지 않고 출판된 책도 거의 없기 때문에 멀티모달 인터페이스에 대한 논문들과 표준화 자료를 중심으로 멀티모달 인터페이스에 대한 설계방안을 제안하였다. 인간의 제스처와 동공의 움직임과 같은 연속적이고 불규칙적인 모달리티 입력을 처리하기 위해서는 형식화된 입력양식이 필요하다. 본 연구에서는 이러한 불규칙적인 모달리티 입력신호들을 구조화하기 위한 모달리티 조합방법과 입력문법, 입력통합과 동기화 방법들을 살펴보았다. 하지만 모달리티 성질 자체가 워낙 주관적인 성향을 띠고 있기 때문에 모든 사람들의 선호도와 사용방식에 맞는 멀티모달 입력 형식을 만들기는 어려울 것이다. 향후 이러한 '사용자 적합화' 문제를 해결하기 위해 자가학습형 에이전트와 Machine Learning과 관련한 소프트웨어 알고리즘 개선 사항들이 보충되어야 할 것이다. 향후 설계된 프로토타입을 보충하여 완전한 시뮬레이션 형태의 멀티모달 인터페이스를 제작해 볼 예정이다. 그리고 인간중심의 멀티모달 인터페이스를 구현하기 위해 인지실험과 사용성 평가를 수행하여 인터페이스를 수정, 보완해 나가야 할 것이다.

참고 문헌

Alan Wexelblat, "An approach to natural gesture in virtual environments," *ACM Transactions on Computer-Human Interaction(TOCHI)*, 2, 179-200, 1995.

Cohen, P. R. and Johnston, M., "QuickSet: Multimodal Interaction for Simulation Set-up and Control," in *Proceedings of the fifth conference on Applied natural language processing*, 20-24, 1997.

Adam Cheter, Luc Julia, "MVIEW: Multimodal Tools for the Video Analyst," in *Proceeding of IUI98*, 55-62, 1998.

Somsak Walairacht, "4 + 4 fingers manipulating virtual objects in mixed-reality environment," *Presence: Teleoperators and Virtual Environments*, 11, 2002.

Buchmann, S., Violich, M. and Billinghurst, A., Cockburn. "FingARtips: gesture based direct manipulation in Augmented Reality," In *Proceedings of the 2nd international conference on Computer graphics and interactive techniques in Australasia and SouthEast Asia (Graphite 2004)*, ACM Press, 212-221, 2004.

Dizio, P., Proprioceptive Adaptation and Aftereffects. In *Handbook of Virtual Environments*, 751-771, 2002.

Klatzky, R. and Lederman, S., Touch., In *Handbook of Psychology*, l.4, 147-176, 2003.

Andrea Corradini, Richard M. Wesson, Philip R. Cohen, "A Map-based System Using Speech and 3D Gestures for Pervasive computing," in *Proceedings of the 4th IEEE International Conference on Multimodal Interface*, IEEE Computer Society, 191, 2002.

Oviatt, S. L., *Multimodal Interfaces, Handbook of Human-Computer Interface*, Ed. By J.Jacko & A.Sears, Lawrence Erlbaum: New Jersey, 2002.

존 R.앤더슨 著., 李永愛 譯., *認知心理學*(*Cognitive psychology and its implications*), 乙酉文化社, 88-91, 1987.

Koons, D. B., Sparrell, C. J. and Thorisson, K. R., "Integrating simultaneous input from speech, gaze, and hand gestures. M.Maybury(Ed.)," *Intelligent Multimodal Interfaces*, 257-276, Menlo Park, CA: MIT, 1993.

Qiaohui Zhang, Atsumi Imamiya, Kentaro Go, Xiaoyang Mao, "Gaze and Speech Multimodal Interface," *Int. Conf. on Distributed Computing Systems Workshops(ICDCSW'04)*, 208-214.

Oviatt, S. L., Mutual disambiguation of recognition errors in a multimodal Architecture, in *Proc. CHI'99 Human Factors in Computing Systems Conf*., Pittsburgh, PA, 576-583, 1999.

Oviatt, S. L., "User-Centered Modeling and Evaluation of Multimodal Interfaces," in *Proc. of the IEEE*, 91(9), 1457-1468, 2003.

http://www.w3.org/TR/2003/NOTE-mmi-reqs-20030108/

http://www.w3.org/TR/mmi-framework/

Jennifer L. Leopold and allen L. Ambler, "Keyboardless Visual Programming Using Voice, Handwriting, and Gesture", in *Proc. of the 1997 IEEE Symposium on Visual Languages(VL '97)*, 28-35, 1997.

Hauptmann, A. G. and McAvinney, P., "Gesture with speech for graphics manipulation," *Int. J. Man-Machine Studies*, 38, 231-249, 1993.

Oviatt, S., DeAngeli, A. and Kuhn, K., "Integration and synchronization of input modes during multimodal human-computer interaction," in *Proc. Conf. human Factors in Computing Systems(CHI'97)*, Atlanta, GA, 415-422, 1997.

Cohen, P. R., Darlymple, M., Pereira, F. C. N., Sullivan, J. W., Gargan, Jr. R. A., Schlossberg, J. L. and Tyler, S. W., "Synergic use of direct manipulation and natural language," in *Proc. Conf. human Factors in Computing Systems(CHI '89)*, Austin, TX, 227-233, 1989.

Sharma, R., "Toward multimodal human-computer interface," in *Proc. IEEE*, 86(5), 853-869, 1998.

Anthony G. Greenwald, "A Reminder about procedures needed to reliably produce perfect timesharing: Comment on Lien, McCann, Ruthruff,

and Proctor," in *Journal of Experimental Psychology: Human Perception and Performance*, 31(1), 221-225, 2005.

◐ 저자 소개 ◑

❖ 임 미 정 ❖ sanctus50@hanmail.net
아주대 미디어학전공 석사
현 재: 아주대 HCI 및 인간공학 연구실
관심분야: HCI, 멀티모달 인터페이스, 가상현실, 3D 모델링

❖ 박 범 ❖ ppark@ajou.ac.kr
아이오와 주립대학 HCI전공 박사,
한국전자통신연구원 선임연구원
현 재: 아주대 산업공학과 정교수(HCI 및 인간공학 연구실),
(주)휴민텍 CEO
관심분야: 유비쿼터스 HCI, 인간공학, 텔레매디슨, 텔레매틱스

논 문 접 수 일 (Date Received) : 2006년 03월 03일
논 문 수 정 일 (Date Revised) : 2006년 05월 01일
논문게재승인일 (Date Accepted) : 2006년 05월 03일