시청각인식을 이용한 실감형 태보 게임

A physical Tae-bo game using audio-visual recognition

원 혜 민, 유 재 권, 신 지 예, 이 경 미
덕성여자대학교 컴퓨터학과 지능형멀티미디어연구실

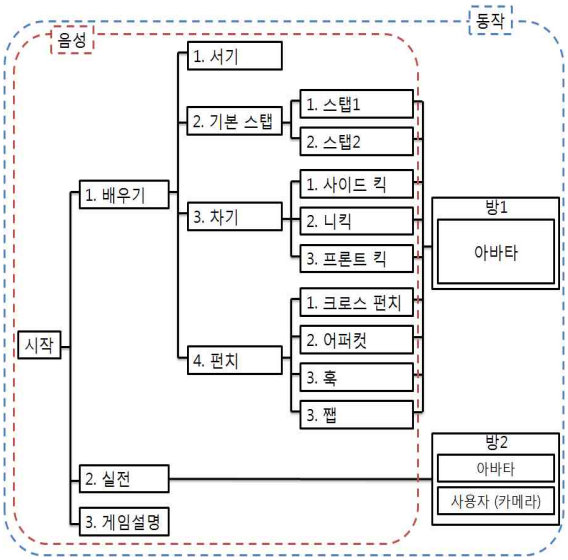
Hye-Min Won, Jae-Kwon Yoo, Ji-Yea Shin, Kyoung-Mi Lee
Duksung Women's University, Dept. of Computer Science, Intelligent Multimedia Lab.

I. 서론

국내의 컴퓨팅 산업과 게임 산업의 기술은 비약적인 발전을 거듭하여 세계적으로 높은 시장 점유율을 보이고 있다. PC 게임 콘텐츠는 키보드와 마우스의 단순 조작을 시작으로 게임 사용자와의 자연스러운 상호작용을 위해 동작과 음성을 이용한 실감형 게임이 게임 시장의 핵심이 되어가고 있다. 본 논문에서는 동작과 음성을 이용하여 즐길 수 있는 실감형 태보 게임을 제안한다.

Ⅱ. 태보 게임에서의 시청각 인식

그림 1은 본 논문에서 제안하는 실감형 태보게임의 순서도를 보여주고 있다. 실강혐 태보게임은 메뉴화면(Shell interface)에서 동작 인식과 음성 인식 기술을 사용하고, 게임 내부 사용자 인터페이스(In-game user interface)에서는 '서기', '스탭', '펀치', '차기' 동작을 이용해 게임을 진행할 수 있도록 개발하였다.



▶▶ 그림 1. 인식기반의 실감형 태보게임 순서도

1. 음성 인식

음성 인식은 컴퓨터가 음향학적 신호를 텍스트로 매핑시키는 과정으로, 인식된 결과는 명령이나 제어, 데이터 입력, 문서 준비 등의 응용분야에서 최종 결과로 사용될 수 있다. 본 논문에서는 체감형 게임을 진행하기 위해 음성인식을 사용한다. 사용된 단어는 숫자로 '일', '이', '삼', '사'와 명령어인 '시작', '취소', '모두 취소'이다. 사용된 단어들의 음성 DB는 ETRI에서 제공하는 숫자열 단어와 명령어 단어들이며, 인원은 남자 20명, 여자 20명으로 총 40명의 화자로 구성되었다.

음성 DB를 활용해 음성 인식을 하기 위해 Cambridge 대학이 개발한 HTK를 사용한다. HTK는 HMM 기반의 음성 인식기를 구현하는 사실상의 표준 도구로써, 세계 대부분의 연구기관과 학교에서 사용되고 있다. 게임의 진행에 맞도록 각각의 단어들의 집합을 훈련하고, HTK를 이용한 음성인식기를 구축한다. HTK는 FFT와 LPC를 모두 지원하지만 이 실험에서는 FFT 기반 log spectra로 유도된 MFCCs를 사용한다. 20개의 채널을 사용하여 추출한 12개의 켑스트럼 계수를 추출하고 MFCC는 39차 MFCC를 이용하여 특징추출을 한다. 음성 모델을 훈련하는 과정에서는 HMM을 사용한다. 태보 게임에 사용된 음성 인식은 1.5m 정도 거리에서 음성인식이 진행되기 때문에 잡음 처리가 필요하다. 잡음 처리는 강인한 음성 인식 기술의 하나인 모델 파라미터 변환 기법 중 Moreno가 제안한 VTS(Vector Taylor Series) 알고리즘을 이용하여 주어진 잡음 환경에서 실험하였다[1].

2. 동작 인식

표 1. 실감형 태보 게임에 사용된 동작

구분	동작
시작	두 손 들기
취소	손 앞으로 내밀기
모두 취소	카메라를 손으로 막기
선택	한 손 들기
서기	서기

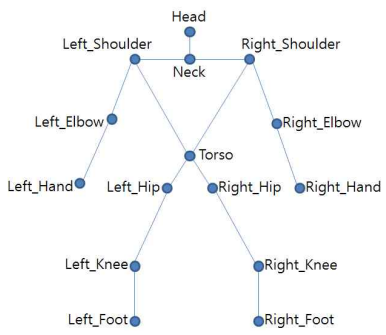
기본 스탭	양쪽 발 한 번씩 번갈아 뛰기
	한 발을 축으로 양쪽으로 번갈아 뛰기
펀치	크로스 펀치
	어퍼컷
	훅
	쨉
차기	프론트 차기
	사이트 킥
	니킥

동작 인식 인터페이스란 동작을 인식하여 기존의 키보드나 마우스와 같은 입력 장치를 대체하여 컴퓨터와 상호작용할 수 있는 기술을 말한다. 제안된 실감형 태보 게임에서는 크게 8개로 구분한 14가지 동작을 사용하여 게임을 진행했다(표 1). 게임의 메뉴 부분에서는 시작, 선택, 취소, 모두 취소를 위해 '두 손 들기', '한 손 들기', '손 앞으로 내밀기', '카메라 가리기' 등의 동작을 취하게 된다. 실제 태보 게임 단계에서는 트레이너인 아바타가 '서기', '스탭', '펀치', '차기' 동작을 보여주면, 사용자는 해당 동작을 취하면서 게임을 진행하게 된다.

2.1 동작 분할 (Gesture Spotting)

입력된 연속된 영상에서의 동작 인식은 연속된 동작의 인식을 어렵게 하기 때문에 의미있는 동작을 찾아내기 위한 동작 분할이 필요하다[3]. 동작 분할은 속도를 기준으로 하는 동작 시작 전과 후의 일정 시간 동안 멈추는 동작의 특징을 이용하는 방법과 동작의 시작과 끝에 많이 나타나는 동작의 특징을 이용하는 방법, 심한 굴곡 포인트를 가진 프레임을 기준으로 하는 방법 등이 있다. 본 논문에서는 이 세 가지 특징을 이용하여 연속으로 들어오는 영상에서 동작을 분할하고 그 분할된 부분에 대해서만 동작 인식을 실행한다.

2.2 동작 인식 (Gesture Recognition)



▶▶ 그림 2. Kinect에서 사용되는 인체 관절 모델

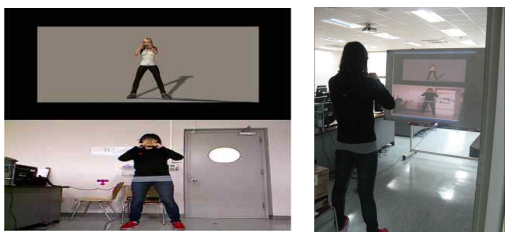
본 논문에서는 Kinect에서 사용된 관절 모델을 사용하여 인체 모델을 구성한다. 그림 2에서 보여지는 것처럼 인체 관절 모델은 사람의 15가지 관절로 이루어졌으며, 연속된 프레임에서 찾은 각각의 관절값을 이용하여 동작 특징을 추출하여야 한다. 본 논문에서는 분할된 동작 내 각 프렘임에서의 각각의 관절값으로 TSS(Temporal Self-Similaries)를 계산하였다[2]. 또한 계산된 TSS의 결과 값을 이용해 기계학습알고리즘인 SVM(Support Vector Machine)을 사용해 각각의 동작들을 분류하고 인식했다.

3. 멀티 모달 융합

멀티모달 입력 시스템은 사용자가 컴퓨터와 다양한 입력 장치(음성, 동작 등)를 통해 상호작용이 가능하게 한다. 이런 형태의 사용자 상호작용은 접근성이 향상될 뿐만 아니라, 자연스런 입력 장치를 가능케 해 사용자의 편의성과 유연성을 높일 수 있다.

우선, 동작 인식 게임을 실행하기 위해서는 사람을 검출하고 사람의 관절을 구분하여야 한다. 제안하는 게임에서는 kinect 카메라를 사용하여 사람을 검출하고 역기를 드는 자세인 'PSI' 자세를 취하면 사람의 관절을 구분하게 된다. 게임은 'PSI' 자세를 인식한 이후에만 다음 단계를 진행할 수 있도록 한다.

메뉴에서 음성 인식은 동작 명령에 의해 실행된다. 즉, 우선 특정 동작이 인식된 후, 음성 인식을 실행하게 된다. 예를 들면, 사용자가 양 손을 드는 동작이 인식된 후 "시작"이라는 음성이 인식되면, 다음 단계로 진행하게 된다.



(a) (b)

▶▶ 그림 3. 실감형 태보 게임 : (a) 게임 화면, (b) 실제 실행 장면

그림 3은 본 논문에서 개발한 '태보' 게임을 보여주고 있다. 그림 3(a)는 태보 게임의 화면으로 위에 트레이너 아바타를, 아래엔 게임 사용자의 모습을 담고 있다. 그림 3(b)는 실제 공간에서 사용자가 직접 실행하는 모습이다.

▌참 고 문 헌▐

[1] 홍미정, 이호웅 "잡음에 강인한 음성 인식을 위한 환경 파라미터 보상에 관한 연구", 한국 ITS 학회 논문지, 제5권, 제2호, pp.1-10, 2006.

[2] Imran N. Junejo, Emilie Dexter, Ivan Laptev, and Patrick P´erez, "View-Independent Action Recognition from Temporal Self-Similarities", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 1, pp.172-185, 2011.

[3] Hyun Kang, Chang Woo Lee and Keechul Jung, "Recognition-based gesture spotting in video games", Pattern Recognition Letters, Vol. 25, Issue 15, pp.1701-1714, 2004.