지능형 로봇 구동을 위한 제스처 인식 기술 동향

Survey: Gesture Recognition Techniques for Intelligent Robot

오 재 용, 이 칠 우*
(Jae-Yong Oh and Chil-Woo Lee)

Abstract : Recently, various applications of robot system become more popular in accordance with rapid development of computer hardware/software, artificial intelligence, and automatic control technology. Formerly robots mainly have been used in industrial field, however, nowadays it is said that the robot will do an important role in the home service application. To make the robot more useful, we require further researches on implementation of natural communication method between the human and the robot system, and autonomous behavior generation. The gesture recognition technique is one of the most convenient methods for natural human-robot interaction, so it is to be solved for implementation of intelligent robot system. In this paper, we describe the state-of-the-art of advanced gesture recognition technologies for intelligent robots according to three methods; sensor based method, feature based method, appearance based method, and 3D model based method. And we also discuss some problems and real applications in the research field.

Keywords : gesture recognition, intelligent robot, human motion analysis, human-computer interaction

I. 서론

지능형 로봇이란 인간과 비슷하게 시각, 청각 등의 감각 기관을 기반으로 자율적으로 판단하고 행동하는 독립형 자율구동 시스템을 말한다. 그동안 로봇은 인간의 일상생활보다는 산업용으로 널리 사용되어 왔으나 최근 컴퓨터 기술의 발달과 함께 인공지능, 마이크로프로세서, 제어, 센서 등의 관련 기술들의 발전에 힘입어 새로운 분야에서의 응용이 기대되고 있다. 최근 개최된 세계 최대의 로봇 박람회인 "ROBODEX" 에서도 그 변화가 두드러지고, 약 80여 종의 신기술을 적용한 인간 친화형 로봇이 전시되어, 4일간 7만명 정도의 관람객이 모일 정도로 로봇에 대한 관심이 높았다. 특히 최근 생활수준의 향상과 복지에 대한 사회적 요구가 커짐에 따라 일상생활에서 가사 혹은 엔터테인먼트 등을 목적으로 하는 생활지원 로봇들의 개발도 활발히 진행되고 있다. 인간의 생활을 보다 편리하게 하며, 유용한 정보를 신속히 전달하는 로봇, 이러한 로봇이 자연스럽게 일상생활의 일부가 되는 날도 멀지 않은 것이다.

인간과 닮은 로봇, 인간처럼 행동하는 로봇을 구현하기 위해서는 인간과 로봇간의 자연스러운 의사소통 수단의 확보가 가장 중요하다. 인간은 80% 이상의 정보를 시각을 통하여 획득한다고 알려져 있다. 다시 말해서, 시각정보는 일상생활에서 매우 많은 비중을 차지하며, 이를 통한 의사소통이 가장 자연스러운 것임을 알 수 있다.

인간의 눈과 대응하는 것이 로봇의 카메라이며, 로봇은 카메라를 통하여 입력되는 영상을 분석하여 외부 상황을 자율적으로 판단하게 되는 것이다. 또한 인간은 언어 이외에도 제스처와 같은 비언어적 수단을 이용하여 의사소통을 하며, 이러한 비언어적 의사소통 수단을 로봇이 이해한다면, 로봇은 인간과 보다 친숙한 대상이 될 수 있을 것이다. 이러한 요구에 의해 얼굴 인식을 비롯한 HCI(Human and Computer Interaction) 기술들이 활발하게 연구되고 있지만 아직 해결해야 할 문제점이 많은 실정이다.

본 논문에서는 지능형 로봇을 위한 기반 기술 중 인간과의 가장 자연스러운 의사소통 방법의 하나인 제스처 인식 기술에 대하여, 최근 연구 성과를 중심으로 요소 기술의 중요 내용과 응용 사례를 소개한다.

II. 제스처 인식 기술의 개요

'제스처'의 사전적 의미는 1)표현의 수단으로서 팔다리 또는 신체의 사용, 2)생각, 감정, 태도를 표현하거나 강조하는 신체나 팔다리의 움직임으로 정의된다[1]. 마찬가지로 HCI(Human and Computer Interaction)의 관점에서의 제스처의 의미도 무심코 행한 움직임이 아닌, 의미를 전달하는 움직임이나 기계와 컴퓨터를 조작하기 위한 모든 움직임을 일컫는다. 따라서 컴퓨터나 로봇이 자율적으로 인간의 행동을 분석하고 인지하는 기술을 제스처 인식 기술이라고 한다.

최근 인간과 정보 시스템 간에 자연스럽게 정보를 교환할 수 있는 지적 인터페이스에 대한 관심이 높아지고, Smart Environment, 웨어러블 컴퓨터, 유비쿼터스 컴퓨팅과 같이 제 4세대 정보기술의 중요성이 강조되면서, 제스처 인식은 컴퓨터 비전 연구자들의 많은 주목을 받고 있다. 제스처를 인식한다는 것은 인체 각 부위가 시간의 경과에 따라 어떻게 변화하는가를 자동으로 분석하고 그 변화를 추상적인 의미로 해석하는 것을 의미한다. 즉 동영상(Moving Image)으로부터 신체 영역을 추출한 다음 각 부분들이 하나의 의미를 갖기 위해 어떤 변화를 거치는지를 알아내는 것이다.

* 책임저자(Corresponding Author)
논문접수 : 2004. 6. 6., 채택확정 : 2004. 7. 18.
오재용 : 전남대학교 컴퓨터공학과(ojyong@image.chonnam.ac.kr)
이칠우 : 전남대학교 정보통신공학부(leecw@chonnam.ac.kr)
* 본 연구는 한국 과학재단 지정 전남대학교 "고품질 전기전자부품 및 시스템 연구센터"의 연구비 지원에 의해 수행되었음.

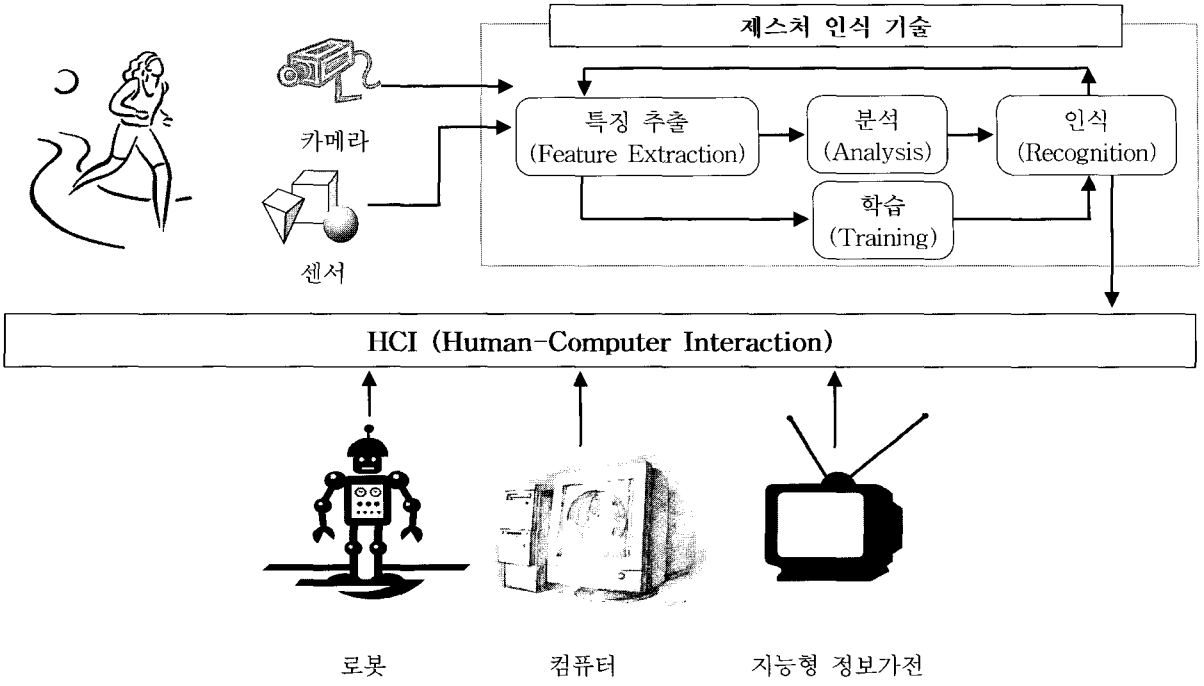


그림 1. 제스처 인식 기술의 응용.
Fig. 1. Application of Gesture recognition.

그러나 인체는 고자유도를 지닌 매우 복잡한 3차원 관절 물체로 동영상으로부터 인체부위를 안정적으로 분리해 내고 그 내용을 인식한다는 것은 매우 어려운 일이다. 또, 사람에 따라 착용하는 의복이 다르므로 특징정보를 안정적으로 추출하기가 어려워 많은 노력에도 불구하고 만족할 만한 결과를 얻기가 어렵다.

일반적인 제스처 인식 기술은 그림 1과 같이 특징 추출, 분석, 학습, 인식의 단계로 구성된다. 카메라 혹은 센서로부터 움직임 정보를 취득하고, 이로부터 제스처를 구별할 수 있는 특징정보를 추출한다. 이렇게 추출된 특징정보와 미리 학습된 모델 제스처를 비교하여 행위자가 어떤 제스처를 행했는지를 인식하게 된다.

인간의 제스처를 분석하고 인식하기 위하여 사용되는 방법은 특징 추출과 인식 방법에 따라 몇 가지로 분류될 수 있다. 첫째, 제스처를 정량화하는 가장 기초적인 방법으로 물리적인 센서를 이용하는 방법이 있다. 이 방법은 제스처 인식의 대상 즉, 인체에 광학식 혹은 자기식 센서를 부착하고 이로부터 획득되는 데이터를 분석하는 방법으로 제스처 인식의 초기에 많이 사용되었다. 그러나 물리적인 장치를 부착해야하며, 장비가 고가라는 단점 때문에 현재는 정확한 모션 데이터 측정을 필요로 하는 모션 캡쳐 분야 이외에는 많이 사용되지 않고 있다.

앞서 언급한대로 신체에 장비를 부착하는 방법은 사용이 번거롭고, 행동의 제약을 받기 쉽다. 이러한 이유에서 센서의 부착이 없이 카메라를 통해 입력되는 영상을 이용한 제스처 인식 방법이 사용되며, 다양한 영상처리 기술이 응용된다. 먼저, 손과 발, 몸통 등과 같이 제스처를 분석하는데 특징이 되는 신체 부위의 시공간적 궤적을 분석함으로써 제스처를 인식하는 특징 기반의 제스처 인식 방법이 있다[2].

또한 영상으로부터 추출된 신체 실루엣 영상을 PCA (Principle Component Analysis) 혹은 ICA(Independent Component Analysis) 등의 통계학적 방법을 이용하여 제스처를 분석하고 인식하는 방법도 있다[3][4].

또, 영상의 기하학적 특징을 이용하지 않고 영상 자체가 가지는 음영정보를 그대로 이용하는 MHI(Motion History Image), MEI(Motion Energy Image)의 방법도 소개되었다[5][6]. 이러한 영상 기반 제스처 인식 방법들은 인식 환경이 인식 대상으로부터 독립적이라는 장점이 있지만, 조명 조건 등 주위 환경에 많은 영향을 받으며, 정밀한 인식이 불가능하다는 단점을 가지고 있다.

이러한 불안정한 요인을 제거하기 위하여 영상으로부터 3차원 정보를 추출하여 이를 제스처 인식에 사용하기도 한다[7][8]. 2차원 영상이 갖는 제스처의 모호성을 극복할 수 있는 방법이기는 하지만, 계산량이 많고, 오류에 민감하기 때문에 실제 시스템에 적용하기에는 더 많은 연구가 필요하다.

III. 센서 기반 제스처 인식

인간의 제스처는 3차원 공간에서 매우 복잡한 구조를 가지고 있기 때문에 그 움직임을 수치적으로 정량화하는 일은 매우 어렵다. 그러나 변화량을 측정할 수 있는 부위에 물리적인 센서나 마커를 부착하면 움직임의 위치와 방향을 정확히 추출할 수 있다[9]. 인간의 제스처를 수치적으로 표현하는 작업을 모션 캡처라고 하고, 이를 위한 장비는 작동방식에 따라 크게 기계식, 자기식, 광학식으로 분류할 수 있다. 주로 자기식과 광학식 방법을 많이 사용하며, 손, 발, 팔꿈치 등과 같이 동작의 주가 되는 신체 부위에 기구를

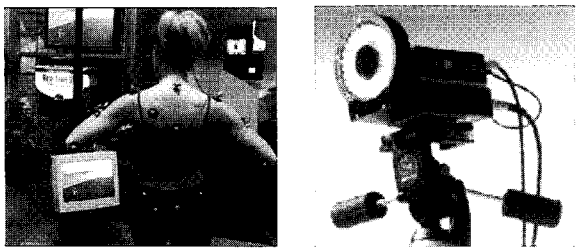


그림 2. 광학식 모션 캡쳐 장비.
Fig. 2. Optical motion capture device.

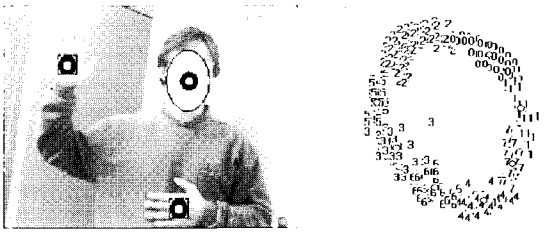
그림 3. 데이터 글로브(Data Glove).
Fig. 3. Data glove.

부착하여 데이터를 얻어낸다. 기계식 혹은 자기식 방법의 경우 센서를 연결하는 장치들을 모두 몸에 부착해야 하므로, 동작의 제약이 있을 수 있다. 그러나 광학식의 경우 별다른 장비 없이 적외선 카메라에 반응하는 마커를 몸에 부착하고, 그 마커의 궤적을 추적함으로서 정확한 모션 데이터를 추출할 수 있기 때문에, 캐릭터 애니메이션이나 컴퓨터 그래픽스 분야에서 최근 많이 사용되고 있는 추세이다. 광학식 모션 캡쳐 시스템의 경우 다른 마커에 가려서 마커가 보이지 않는 마커들 간의 중첩(Occlusion) 문제가 발생하여 3차원 좌표를 얻는 것이 불가능하고, 이 때문에 많은 후처리 과정을 필요로 하게 되며, 실시간 처리가 불가능하게 되거나 모션 캡쳐 성능을 떨어뜨리는 요인이 되기도 한다.

광학식 모션 캡쳐장비 이외에 그림 3과 같은 데이터 글로브(Data Glove)도 제스처를 획득하기 위한 장비로 사용된다[10]. [10]에서는 파킨슨(Parkinson's disease) 환자의 손 제스처에 나타나는 증상을 판단하기 위하여 데이터 글로브를 사용한다. 간단한 손 제스처 인식을 위하여 레이저 빔이 사용되기도 한다[11]. [11]은 PDA(Personal Digital Assistants)와 같은 휴대용 장치에 레이저 빔을 이용한 제스처 인식 장치를 설치하고, 키보드나 터치스크린이 아닌 손가락 제스처로 명령을 입력하는 시스템이다.

IV. 영상 기반 제스처 인식

앞 절에서 언급한 센서 기반의 방법들은 신체에 많은 장비들을 부착해야 한다는 큰 단점을 가지고 있다. 특히 기계식 혹은 자기식 방법의 경우 케이블의 길이에 따른 행동의 제약을 받게 된다. 이러한 문제점들을 해결하기 위하여 센서를 부착하지 않고 카메라를 통해 입력되는 영상을 분석하여 제스처를 인식하는 방법이 제안되었다. 객체 추출 및 추적 등의 영상처리 기술을 제스처 인식에 응용함으로서,



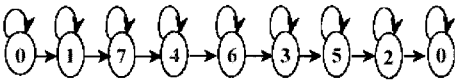


그림 4. 신체 특징점의 궤적을 이용한 제스처 인식.
Fig. 4. Gesture recognition using feature points.

환경의 제약을 완화시키고, 보다 자연스럽게 인간과 컴퓨터 간에 의사소통을 할 수 있게 되었다. 영상 기반 제스처 인식 방법은 크게 특징기반, 외관 기반, 3차원 모델 기반의 3가지로 분류될 수 있다.

1. 특징 기반 제스처 인식

간단한 제스처의 경우에는 복잡한 파라미터를 구하지 않고도 에지, 윤곽선, 특징점의 위치 등의 정보를 이용하여 쉽게 구별할 수 있다. 이 방법에서는 명확한 특징 추출의 여부가 관건이며, 추출된 특징을 이용하여 제스처 모델을 구성하고 입력 영상과 제스처 모델의 비교를 통하여 제스처를 인식한다. 그러나 특징으로 사용되는 에지 및 윤곽선 정보가 주위 환경에 매우 민감하기 때문에, 일반적으로 환경에 대한 제약조건과 함께 사용된다.

에지나 윤곽선 이외에 제스처의 특징으로 손이나 발, 얼굴 등의 위치 정보도 사용된다[2](그림 4). 왜냐하면, 인간은 비언어적 정보를 전달하는 수단으로 손, 발, 얼굴 등의 움직임을 많이 사용하고 있기 때문이다. 이러한 개념에서 [2]에서는 제스처를 시·공간상에서 정의되는 특정 부위 움직임 조각의 집합이라고 가정한다. 먼저 얼굴과 양손의 공간상 위치를 dynamic k-means 클러스터링 방법을 이용하여 자동으로 분류한 뒤, 분류된 클러스터의 시간에 따른 상태 천이를 Finite State Machine(FSM) 알고리즘을 이용하여 제스처를 학습하고, 인식하게 된다.

[2]에서 s_k는 FSM의 현재 상태라고 하고, t를 그 상태에서 머무른 시간이라고 할 때, 각 제스처의 의미 정보는 (1)과 같이 표현되며,

$$c = < s_k, t > \tag{1}$$

새로운 입력을 $\vec{x}$ 라고 하고, 비교 대상을 $\vec{\mu_s}$ 라 할 때, (2)와 같이 정의되는 Mahalanobis 거리를 기준으로 상태를 비교하여, 제스처를 인식하게 된다.

$$D(\vec{x}, S) = \sqrt{(\vec{x}-\vec{\mu_s})^T \Sigma_s^{-1} (\vec{x}-\vec{\mu_s})} \tag{2}$$

한편, FSM는 아래 조건에서 상태가 변하게 된다.

(i) $(D(\vec{x}, s_{k+1}) \leq d_{k+1})$ & $(t > t_l^{\max})$

(ii) $(D(\vec{x}, s_{k+1}) \leq d_{k+1})$ & $(D(\vec{x}, s_{k+1}) \leq D(\vec{x}, s_k))$ & $(t > t_i^{\min})$

(iii) $(D(\vec{x}, s_{k+1}) \leq d_{k+1})$ & $(D(\vec{x}, s_k) \leq d_k)$

이와 같은 방법은 계산량이 적고, 처리 속도가 빠르기 때문에 실시간 인식 시스템에 적합하지만, 주위 환경의 영향에 민감하고 특징점의 중첩(Occlusion)이 발생할 경우 인식이 어렵다는 단점이 있다.

머리, 손, 발과 같은 신체 특징점 이외에도 외곽선의 첨점(Peaks and valleys)을 이용하기도 한다. [12]에서는 입력되는 손 영상의 외곽선을 구하고 그 외곽선의 첨점정보를 이용하여 손 제스처를 인식한다. 이 방법은 매우 간단하고 처리 속도가 빠르기 때문에 비디오 게임 컨트롤과 같은 시스템에 응용될 수 있지만, 안정된 외곽선 추출을 위하여 주위 환경에 제약을 두었다.

2. 외관 기반 제스처 인식

앞서 언급한대로 특징 기반의 방법은 인식 결과가 주위 환경의 영향을 많이 받기 때문에 매우 불안정하다. 이러한 단점을 보완하고자 영상의 기하학적 특징을 이용하지 않고 영상 자체가 가지는 음영 정보를 그대로 이용하고자 하는 방법이 외관 기반(Appearanced based) 제스처 인식 방법이다. 실루엣(Silhouette) 이라고 표현되는 입력 영상의 음영 정보는 배경 이미지를 제거한 전경의 이진화 영상이며, 이를 분석함으로서 제스처를 인식할 수 있다.

대표적인 외관기반 제스처 인식 방법은 MHI(Motion History Image)다[5]. MHI는 (3)에서처럼 입력 영상 시퀀스에서 더 최근에 움직인 영역의 화소들을 더 밝은 값으로 나타낸 영상으로서, 이 영상을 다시 9개의 세부 영역으로 나누고, 각 영역에서의 히스토그램을 모델 제스처와 비교하여 제스처를 인식하는 방법이다.

$$MHI(x,y) = \begin{cases} \tau & \text{if } currnet\ motion\ at(x,y) \\ 0 & else\ \text{if } MHI(x,y) < (\tau - \delta) \end{cases} \quad (3)$$

여기서, τ는 현재 time-stamp를 의미하고, δ는 시간 간격을 나타낸다.

또한, 전체 영상을 특징으로 사용하는 방법으로 MEI (Motion Energy Image)가 있다[6](그림 6). MEI는 영상 시퀀스의 어디에서 동작이 일어나고 있는지를 표현하는 이진화 영상으로 (4)로 표현되어진다. 얻어진 MEI 영상의 Hu 모멘트를 계산하고 모델 제스처와의 거리를 비교함으로써 제스처를 인식한다.

$$E_\tau(x,y,t) = \bigcup_{i=0}^{\tau-1} D(x,y,t-i) \quad (4)$$

여기서 $E_\tau(x,y,t)$는 MEI(Motion Energy Image)를, $D(x,y,t)$는 이진화 영상 시퀀스를, τ는 시간 윈도우의 길이를 각각 나타낸다.

이렇게 MHI 혹은 MEI와 같이 일정 시간 동안 영상의 누적된 모션 정보를 이용하면 실시간으로 행위자의 제스처를 인식할 수 있다. 그러나 위 두 방법은 행위자가 한명이라는 가정을 하고 있으며, 행위자 이외에 움직이는 물체가 없다는 가정을 하고 있다. 또한 실루엣 영상을 이용하기 때문에 복잡한 모션은 인식이 어렵다는 단점이 있다.

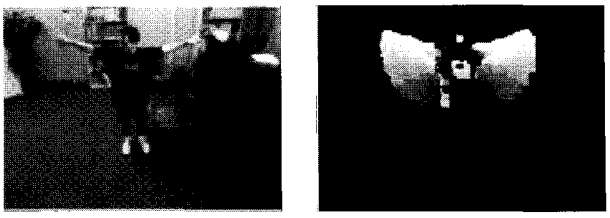


그림 5. MHI(Motion History Image).
Fig. 5. MHI(Motion History Image).

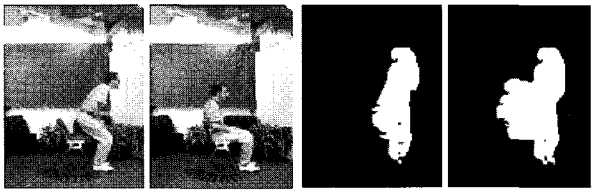


그림 6. MHI(Motion Energy Image).
Fig. 6. MEI(Motion Energy Image).

전체적인 움직임 특징을 이용하는 또 다른 예로 *Ross Culter's method*[13]가 있다. 이 방법은 영상 내 Blob의 Optical flow 특성을 이용하여 제스처를 인식하는데, 이 또한 영상 내 중첩영역이 있을 때는 정확한 인식을 수행할 수 없다는 단점이 있다.

*Ismail Haritaoglu*의 알고리즘은 인간의 신체를 6개의 영역(*Cardboard model*)으로 나누어 이를 분석하는 방법을 사용한다[14]. 이 시스템에서는 똑바로 선 상태에서의 제스처로 모델을 제한하여, 실루엣 영상의 볼록한 부분(*Convex hull*)을 이용하여 다양한 형태의 제스처를 인식할 수 있는 시스템으로, 실루엣 영상에서 수직 및 수평 히스토그램을 이용하여 대략적인 제스처를 찾아 낸 뒤, *Recursive Convex hull* 알고리즘과 위상적 분석을 통하여 최종 신체 특징점을 결정한다. 이 방법은 제스처 영상의 형상 정보와 세부 정보를 조합하여 사용하고, 신체 영역에 따른 계층적 분석방법을 사용했다는 점에서 주목 할만하다.

일반적으로 제스처 데이터는 고차원의 특성을 갖으며, 데이터의 특성상 직관적이지 못한 경우가 많다. 따라서 이를 효율적으로 분석하는 방법이 필요하며, PCA(Principle Component Analysis) 등의 통계학적 방법을 사용하기도 한다[3]. 그림 7에서와 같이 입력 실루엣 영상에서 추출된 특징벡터를 PCA를 이용하여 파라메트릭 제스처 공간을 구성한다. 입력되는 제스처는 이 공간에 투영되며, 모델 제스처와 거리를 비교함으로서 제스처를 인식하는 방법이다.

제스처 공간은 (5)와 같이 모든 평균 영상과 각 영상들과의 차를 이용하여 공분산 행렬 Q를 구하고, 특이치 분해(Singular Value Decomposition)등의 방법으로 Q에 대한 고유치(eigenvalue) λ와 고유벡터(eigenvector) e를 구할 수 있다.

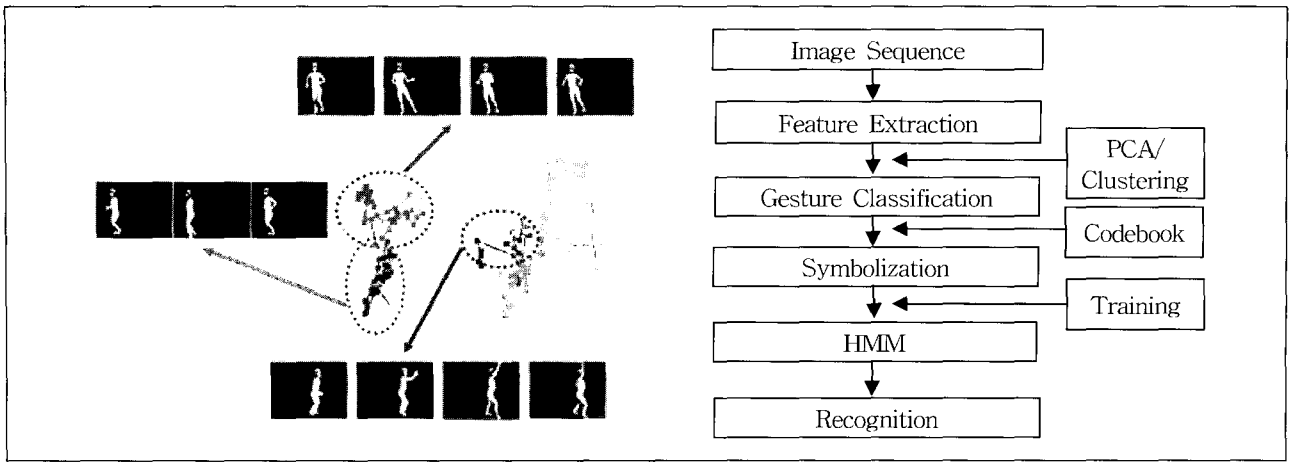


그림 7. 파라메트릭 제스처 공간으로의 투영과 외관 기반 제스처 인식 과정의 흐름도.
Fig. 7. Projection of gesture image into gesture space and Procedure of appearance-based gesture recognition.

$$Z=[z_1-c \;\; z_2-c \;\cdots\; z_M-c]$$
$$Q = ZZ^T \tag{5}$$
$$Qe_i = \lambda_i e_i$$
$$\zeta_i = E^T(z_i-c) \tag{6}$$

이 고유벡터들로 구성된 공간이 파라메트릭 제스처 공간이며, 고유벡터의 크기는 고유공간의 중요도를 의미한다. 따라서 고유벡터의 크기에 따라 고유벡터를 선택함으로써 고차원의 제스처 데이터를 저차원의 벡터를 이용하여 표현할 수 있다. 그러나 이 방법은 분석하고자 하는 데이터의 구조가 선형적이어야 한다는 전제조건이 있으며, 그렇지 않을 경우 큰 의미를 가지지 못하게 된다는 단점이 있다.

또한 [3]에서는 제스처를 시·공간상의 연속된 동작군으로 가정하고, 제스처의 시간적인 개념을 표현하기 위하여 은닉 마르코프 모델(HMM)을 이용한다. 이 방법은 모델 데이터를 파라메트릭 공간에서 클러스터링하고, (7)과 같이 각 제스처 상태로의 천이 확률을 계산하여 최대 확률값을 갖는 제스처로 인식하는 방법이다.

$$P(Y|\lambda_i) = \sum_i \sum_j \alpha_i(i)a_{ij}b_j(y_{t+1})\zeta_{t+1}(j) \tag{7}$$

은닉 마르코프 모델(HMM)은 은닉상태(Hidden status)와 관측가능상태(Observable status)로 이루어진 확률적 네트워크를 이용한 통계적인 인식 방법이며, 공간적인 개념과 시간적인 개념을 동시에 표현한다. HMM의 λ 는 다음과 같은 변수들에 의해서 표현된다. 상태천이 확률 a_{ij}는 HMM의 상태가 i로부터 j로 변화하는 확률을 의미한다. 그리고 확률 $b_{ij}(y)$는 출력 심볼 y가 상태 i로부터 j로 천이되면서, 관측될 수 있는 확률이며, π_i는 초기 상태 확률값을 나타낸다. HMM의 학습은 $\{\pi, A, B\}$의 파라미터들을 추정하는 것이며, HMM추정을 위해서 Baum-Welch 알고리즘, forward backward 알고리즘 등을 사용한다[15][16].

파라메트릭 은닉 마르코프 모델(PHMM)[17][18]은 표준 은닉 마르코프 모델(HMM)의 출력 확률 안에 전체적인 파라메트릭 변화를 포함함으로써 기존 알고리즘을 확장한 것이다. 이 방법은 선형 파라메트릭 은닉 마르코프 모델(Linear PHMM)과 비선형 파라메트릭 은닉 마르코프 모델(Nonlinear PHMM)로 나누어진다. 표준 은닉 마르코프 모델은 각 제스처 부류의 공간적 변화를 노이즈로 간주하는 반면, 파라메트릭 은닉 마르코프 모델(PHMM)은 각 부류에 존재하는 공간적 변화를 복원하여 구분함으로써 표준 은닉 마르코프 과정(HMM)보다 더 나은 성능 결과를 얻을 수 있다. 또한, 비선형 파라메트릭 은닉 마르코프 모델(Nonlinear PHMM)은 선형 파라메트릭 은닉 마르코프(Linear PHMM) 과정보다 더 많은 제스처를 모형화 할 수 있는 장점을 갖는다.

은닉 마르코프 시스템은 상향식 시스템 구성 방식으로 인해, 에러가 발생한 경우나 영상 특징이 빠진 경우 안정성에 문제가 생긴다는 단점이 있다. 또한 단일 오브젝트 여야 한다는 점 또한 다중의 오브젝트를 동시에 인식할 수 없다는 제약을 가진다. 그러나 활성화된 상태들을 색깔 있는 토큰으로 표시하는 방식으로 단점을 극복할 수 있다[19].

3. 3차원 시각정보 기반 제스처 인식

2차원의 영상정보는 3차원의 제스처를 표현하는데 한계가 있다. 앞 절에서 언급한 특징점의 중첩(Occlusion)과 같은 문제는 2차원 영상이 갖는 모호성 때문에 발생되는 것이며, 이를 해결하기 위하여 3차원 시각정보를 사용한다[5]. 3차원 시각정보는 스테레오 카메라 영상의 깊이(depth)정보를 이용하여 획득할 수 있다. [5]에서는 그림 8과 같이 깊이 정보를 이용하여 손 제스처 영역을 추출한다. 손 영상의 기하학적 형태와 움직임 궤적을 분석하여 제스처를 인식하여, 로봇에게 의사를 전달하는 수단으로 사용한다.

한편, 그림 9와 같이 머리와 양손의 3차원 데이터를 추출하고, 이를 제스처 인식에 응용하기도 한다[6]. 머리와 양손의 3차원 위치는 스테레오 기하를 기반으로 계산되며, 블랍(blob)의 모멘트 정보를 이용하여 회전각을 계산한다. 이렇게 계산된 데이터는 가상현실에서 인간-컴퓨터간의 인터페이스에도 응용되며, 시간상의 궤적을 이용한 제스처 인식 방법에도 응용이 가능하다. 그러나 카메라 보정 등의 사전 작업이 필요하며, 복잡한 3차원 계산 과정에서 오차가 발생할 수 있다는 단점이 있다.

그림 8. 깊이 정보를 이용한 영역 추출.
Fig. 8. Segmentation using depth information.

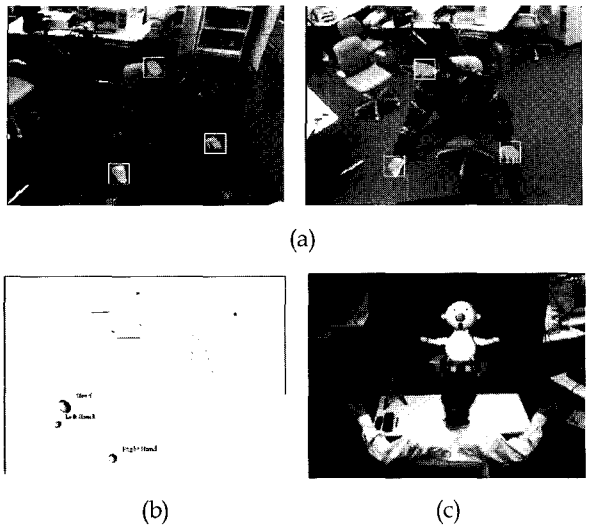


그림 9. 3차원 기반 제스처 인식 (a) stereo sequence (b) 3d estimation (c) Recognition.
Fig. 9. 3D based Gesture Recognition (a) stereo sequence (b) 3d estimation (c) Recognition.

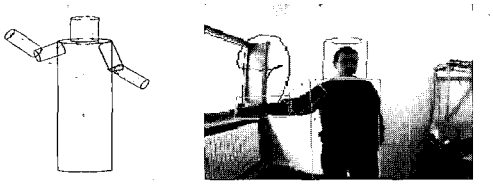


그림 10. 3차원 모델 기반 제스처 인식.
Fig. 10. 3D Model based Gesture Recognition.

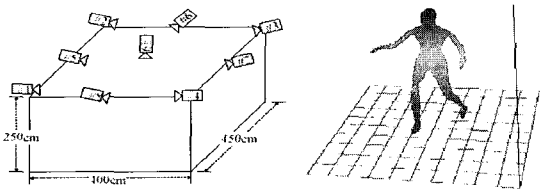


그림 11. Volumetric Model 기반 제스처 인식.
Fig. 11. Volumetric model based gesture recognition.

[20]에서는 피부색과 양말의 색을 이용하여 머리, 양손, 양발의 신체 특징점을 추출하고 이를 이용하여 가상의 3차원 아바타를 조종한다. 다른 접근 방법들과는 달리 정확한

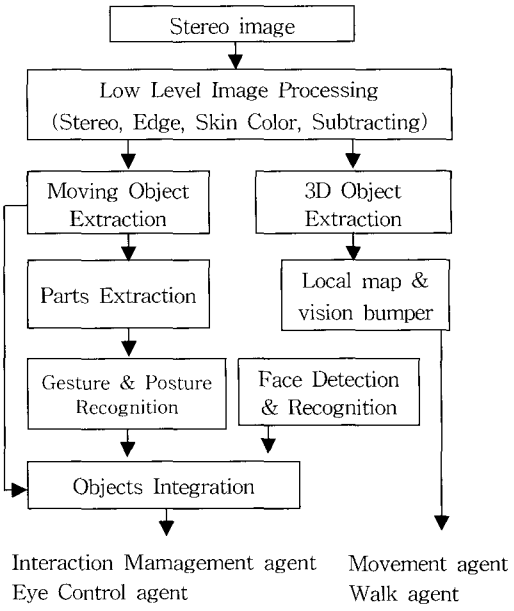


그림 12. ASIMO 비젼 시스템.
Fig. 12. Vision system configuration of ASIMO.

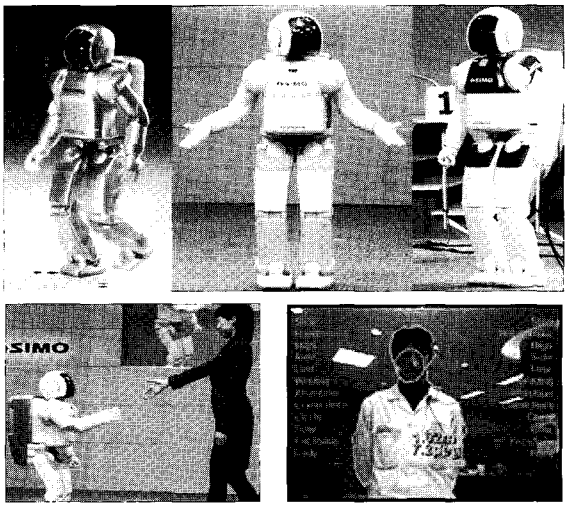


그림 13. HONDA 사의 ASIMO.
Fig. 13. HONDA ASIMO.

특징 추출을 위하여 여러 대의 카메라를 사용하여 특징점의 중첩(Occlusion) 문제를 완화하였다. 또한 역운동학(inverse kinematics) 알고리즘을 이용하여 팔꿈치나 무릎 등의 위치를 추정하기 때문에, 모션 캡쳐 시스템으로도 응용 될 수 있다. 그러나 세부적인 제스처가 포함된 복잡한 모션의 경우 인식이 불가능하며, 조명 변화 등의 이유로 얼굴, 양손, 양발의 특징점을 찾지 못하는 경우 오류가 발생하게 된다[20].

V. 기타 제스처 인식

제스처 인식에서 가장 어려운 작업은 제스처를 수치적으로 잘 표현할 수 있는 특징을 선택하는 일이다. 그러나 3차원의 다관절체로 구성된 인간의 움직임을 표현하는 일은 매우 어려운 일이며, 모든 데이터를 제스처에 사용할 수도 없

다. 이러한 배경에서 인간의 골격 모델을 기반으로 단순화된 3차원 모델을 생성하고, 이를 기준으로 움직임을 분석하는 연구도 진행되고 있다[21]. [21]은 그림 10과 같이 원통형의 단순화된 모델을 입력 영상과 비교하여 자세를 추정하고, 이 데이터를 바탕으로 제스처를 인식한다. 이 시스템은 양 손을 사용하는 9가지 정도의 간단한 제스처를 인식하여 HCI(Human Computer Interaction)로 응용된다. 그러나 이 방법은 단순화된 3차원 모델을 사용하기 때문에 정교한 제스처 인식에는 부적합하며, 영상 기반 시스템이 갖는 배경 및 조명에 따른 문제점을 가지고 있다.

3차원 제스처 정보를 획득하는 다른 방법으로 여러 대의 카메라를 이용한 체적(Volumetric) 모델을 사용하기도 한다[22]. [22]에서는 그림 11에서와 같이 여러 대의 카메라로부터 입력받은 영상을 합하여 3차원 모델로 복원함으로서 자세를 해석한다. 여러 각도에서 입력되는 영상을 사용하기 때문에 정교한 움직임을 추출할 수 있다는 장점이 있지만, 실시간으로 3차원 복원을 수행하는데 많은 계산량이 필요하다. 실제로 이 시스템에는 9 대의 클러스터링 된 컴퓨터가 사용되었다.

VI. 응용 예

현재 개발된 로봇 중 인간과 가장 흡사한 형태를 지닌 로봇은 HONDA社에서 개발된 "ASIMO"이다[23]. 인간처럼 자유로운 2족 보행이 가능하며, 대화자를 인식하고, 대화자의 간단한 제스처를 인식할 수 있도록 설계 되었다. ASIMO는 그림 12에서처럼 스테레오 카메라를 통하여 입력된 영상을 바탕으로 외부상황을 자율적으로 판단하고 이에 적당한 행동을 취한다. 입력된 영상을 자동으로 분석하기 위하여 Optical flow 알고리즘, snake 알고리즘 등의 다양한 영상 처리 기법이 사용되며, 얼굴인식과 간단한 제스처의 인식이 가능하다.

또한, ASIMO는 "손 흔들기", "악수하기" 등의 간단한 제스처의 인식이 가능하다. 인식 대상의 손 위치를 추적하고, 그 위치값과 확률 모델을 기반으로 제스처를 인식한다. 또한 스테레오 카메라로부터 얻어지는 깊이(depth)정보를 이용하여 "지시" 제스처를 인식하며, 해당 위치로 이동이 가능하다.

VII. 결론

본 논문에서는 로봇 구동을 위한 제스처 인식 기술의 최근 연구 동향과 그 응용에 대해서 기술하였다. 인간의 몸은 매우 복잡한 구조를 가지고 있어서 어떤 의미를 지닌 시·공간상의 패턴을 인식하기에는 여러 제약 사항들이 수반되며, 로봇과 같이 다양한 환경변수를 갖는 대상에의 적용은 더욱 어렵다. 그러나 제스처 인식 기술은 로봇과 인간의 자연스러운 의사소통을 위해서 반드시 해결해야 할 문제이며, 향후 지속적인 연구 개발이 필요한 분야이다. 로봇과 인간과 함께 생활하고, 더 나아가 로봇과 정서적인 교류가 가능해질 때까지 로봇 분야의 연구는 꾸준히 진전 되어야 할 것이다.

참고문헌

[1] V. I. Pavlovic, Rajeev Sharma, and Thomas S. Huang,"Visual Interpretation of Hand Gestures for Human-Computer Interpretation: A Review", *IEEE Transaction on PAMI,* vol. 19, no. 7, July 1997.

[2] P. Y. Hong, M. Turk, T. S. Huang, "Gesrure Modeling and Recognition using Finite State Machnes", *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on,* 28-30 March 2000.

[3] C.-W. Lee, H.-J. Lee, S. H. Yoon, and J. H. Kim, "Gesture Recognition in Video Image with Combination of Partial and Global Information", *in Proc. of VCIP 2003,* Lugano, July, 2003.

[4] C. BenAbdelkader, R. Cutler, L. Davis, "Motion-based recognition of people in EigenGait space", Automatic Face and Gesture Recognition, 2002. *Proceedings. Fifth IEEE International Conference on,* 20-21 May 2002.

[5] J. Davis, "Recognizing Movement using Motion Histograms", *MIT Media Lab. Technical Report* no. 487, March 1999.

[6] J. W. Davis, A. F. Bobick, "The representation and recognition of human movement using temporal templates" *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on,* 17-19 June 1997.

[7] X. Liu, K. Fujimura, "Hand gesture Recognition using depth data", *Automatic Face and Gesture Recognition, 2004. Proceedings. sixth IEEE International Conference on,* May 2004.

[8] A. Azarbayejani, C. Wren, and A. Pentland, "Real-time 3-D Tracking of the Human body", *MIT Media Lab. Technical Report* no. 374, May 1996.

[9] B. Li, H. H. "Articulated point pattern matching in optical motion capture systems", *Qinggang Meng; Control, Automation, Robotics and Vision, 2002. ICARCV 2002. 7th International Conference on,* vol. 1 , 2-5 Dec. 2002.

[10] Y. Su, C. R. Allen, D. Geng, D. Burn, U. Brechany, G. D. Bell, R. Rowland, "3-D motion system (data-gloves) : application for Parkinson's disease", *Instrumentation and Measurement, IEEE Transactions on,* vol. 52, Issue : 3, June 2003.

[11] W. StephanePerrin, A. Cassinelli and M. Ishikawa, "Gesture Recognition Using Laser-based Tracking System". *Automatic Face and Gesture Recognition, 2004. Proceedings. sixth IEEE International Conference on,* May 2004.

[12] J. Segen, S. Kumar, "Fast and accurate 3-D gesture recognition interface", Pattern Recognition, 1998.

Proceedings. Fourteenth International Conference on , vol. 1 , 16-20 Aug. 1998.

[13] R. Cutler, M. Turk, "View-based Interpretation of Real-time Optical Flow for Gesture Recognition", *Third IEEE International Conf. on Automatic Face and Gesture Recognition,* 1998.

[14] I. Haritaoglu, D. Harwood and L. S. Davis, "W4: Who? When? Where? What? A Real-time System for Detecting and Tracking People", *Third Face and Gesture Recognition Conference,* 1998.

[15] Y. Iwai, T. Hata and M. Yachida, Gesture Recognition based on Subspace Method and Hidden Markov Model, *IEEE,* 1997, pp. 960-966.

[16] T. Starner, A. Pentland, "Real-Time American Sign Language Recognition from Video using Hidden Markov Models", *ISCV,* 1995.

[17] A. D. Wilson, A. F. Bobick, "Parametric Hidden Markov Models for Gesture Recognition", *IEEE Transaction on PAMI,* vol. 21, no. 9, September 1999.

[18] A. D. Wilson, A. F. Bobick,,"Recognition and Interpretation of Parametric Gesture", *ICCV,* 1998.

[19] T. Wada, T. Matsyama, "Appearance Based Behavior Recognition by Event Driven Selection Attention", *CVPR,* 1998.

[20] H. Yoshimoto, N. Date, S. Yonemoto, "Vision-based real-time motion capture system using multiple cameras", Multisensor Fusion and Integration for Intelligent Systems, MFI2003. *Proceedings of IEEE International Conference on,* 30 July-1 Aug. 2003.

[21] J. Amat, A. Casals, "Stereoscopic System for Human body Tracking", M. Frigola, Modelling People, 1999. *Proceedings. IEEE International Workshop on,* 20 Sept. 1999.

[22] T. Wada, W. Xiaojun S. Tokai, T. Matsuyama, "Homography Based Parallel Volume Intersection", Computer Architectures for Machine Perception, 2000. *Proceedings. Fifth IEEE International Workshop on,* 11-13 Sept. 2000.

[23] Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki, K. Fujimura, "The intelligent ASIMO: system overview and integration", *Intelligent Robots and System, 2002. IEEE/RSJ International Conference on,* vol. 3, 30 Sept.-5 Oct. 2002.

오 재 용

2000년 전남대학교 컴퓨터공학과 졸업. 동대학원 석사(2002). 2002년~현재 전남대학교 대학원 컴퓨터정보통신공학과 박사과정. 관심분야 : 로봇 비전, 제스처 인식.

이 칠 우

1986년 중앙대학교 전자공학과 졸업. 동대학원 석사(1988). 동경대학 대학원 박사(1992). 1992년부터 1995년까지 일본 이미지정보과학연구소 수석연구원. 얼굴인식 및 제스처인식 등의 휴먼인터페이스 관련 연구수행. 1996년~현재 전남대학교 전자컴퓨터정보통신공학부 교수. 관심분야 : 디지털콘텐츠, 3차원 컴퓨터 비전, 컴퓨터그래픽스.