상황 인지 방법을 이용한 지능형 제스처 인터페이스

오재용, 이칠우
전남대학교 컴퓨터정보통신공학과
지능영상미디어연구실
ojyong@image.chonnam.ac.kr,leecw@chonnam.ac.kr

Intelligent Gesture Interface Using Context Awareness

Jae Yong Oh, Chil Woo Lee
Dept. of Computer Engineering, Chonnam National University
Intelligent Image Media Lab.

요 약

본 논문에서는 상황 인지(Context Aware)를 이용한 제스처 인식 방법에 대하여 기술한다. 기존의 인식 방법들은 대부분 제스처의 개별적인 의미를 중심으로 제스처를 분류하는 방법을 사용한다. 그러나 이러한 방법들은 인식 알고리즘을 일반화하는데 있어서 다음과 같은 문제점들을 가지고 있다. 첫째, 인간의 모든 제스처를 제한된 특징으로 모호하지 않게 구별하기 어렵다. 둘째, 같은 제스처라 할지라도 상황에 따라 다른 의미를 내포할 수 있다. 이러한 문제점들을 해결하고자 본 논문에서는 확률 기반의 상황 인지 모델을 이용한 제스처 인식 방법을 제안한다. 이 방법은 제스처의 개별적인 의미를 인식하기 전에 대상의 상황을 추상적으로 분류함으로써 행위자의 의도를 정확히 파악할 수 있다. 본 방법은 시스템의 상태를 [NULL], [OBJECT], [POSTURE], [GLOBAL], [LOCAL]의 5 가지 상태로 정의한 뒤, 각 상태의 천이를 바탕으로 대상의 상황을 판단한다. 이러한 상황 정보에 따라 각 상태에 최적화된 인식 알고리즘을 적용함으로써 지능적인 제스처 인식을 수행할 수 있으며, 기존 방법들이 갖는 제스처 인식의 제약을 완화 시키는 효과가 있다. 따라서, 제안하는 제스처 인터페이스는 자연스러운 상호 작용이 필요한 지능형 정보 가전 혹은 지능형 로봇의 HCI 로 활용될 수 있을 것이다.

Keyword : Bayes' rule, Context-aware, Gesture interface

1. 서 론

최근 인간의 제스처에 대한 관심이 높아지고, 컴퓨터 시스템이 급속도로 발달하면서 인간 친화적 인터페이스와 같은 분야에 인간의 제스처를 기반으로 하는 기술들이 다각도로 응용되고 있다. 특히 가전제품과 같은 일상적인 기기에도 지능형 인터페이스 기술이 적용되면서, 컴퓨터와 인간 간의 자연스러운 의사소통 수단이 중요한 이슈가 되고 있다.

본 연구는 한국 과학재단 지정 전남대학교 "고품질 전기전자부품 및 시스템 연구센터", "네트워크 휴머노이드 기술 개발사업"의 연구비 지원에 의해 수행되었음.

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.

또한 인간은 언어 이외에 제스처와 같은 비언어적 수단을 이용하여 의사소통을 하며, 이를 이용하여 보다 자연스러운 의사소통이 가능하다. 따라서 인간-컴퓨터 간의 인터페이스에 제스처와 같은 비언어적 수단을 사용함으로써 보다 자연스러운 상호작용이 가능한 것이다. 그러나 인체는 매우 복잡한 3 차원의 관절 구조를 가지고 있을 뿐만 아니라 신체 부위에 따라 각기 다른 의미를 표현

할 수 있기 때문에 컴퓨터가 이를 자동으로 분석하고 인식하는 것은 매우 어려운 일이다. 이 때문에 온 몸에 센서를 부착하고 이를 통해 얻어지는 데이터를 분석하는 방법을 사용하기도 한다[1,2]. 그러나 이 방법은 복잡한 장비를 몸에 부착해야 하기 때문에 시스템으로의 활용도가 매우 낮다. 이와 같은 이유에서 신체에 특별한 장치를 부착하지 않고 제스처를 인식하는 방법으로 카메라를 이용하는 방법이 많이 연구되고 있다. 대표적인 영상 기반 제스처 인식 방법은 MHI(Motion History Image)를 이용한 방법이다[3]. MHI 는 입력 영상 시퀀스에서 더 최근에 움직인 영역의 화소들을 더 밝은 값으로 나타낸 영상으로서, 이 영상 정보를 이용하여 제스처를 인식하는 방법이다. 이 밖에도 신체 특징점이나 실루엣 영상을 이용하는 방법[4,5]들이 제안되었지만, 이러한 방법들은 모두 시스템의 일반성이 결여되는 문제점을 가지고 있다. 즉, 미리 정의되어 학습된 제스처를 인식하는 데는 문제가 없지만, 정의되지 않은 제스처의 경우 인식하지 못하거나 잘 못 인식될 확률이 높다. 더욱이 인간의 모든 제스처에 대해서 시스템을 학습시킬 수 없기 때문에, 시스템에 종속된 제한적인 알고리즘이 되기 쉽다.

이와 같은 문제점을 해결하고자 본 논문에서는 상황 인지(Context Aware) 방법을 이용한 제스처 인터페이스 방법을 제안한다. 본 방법은 제스처의 개별적인 의미를 인식하기 전에 대상의 상황을 추상적으로 파악함으로써 행위자의 의도를 보다 정확히 분석할 수 있으며, 기존 인식 방법에 비해 인식 대상을 확장할 수 있는 장점이 있다.

본 논문의 2절과 3절에서는 확률 기반 상태 천이 모델 기반의 제스처 인터페이스 시스템에 대하여 설명한다. 4절에서는 제스처 인식 시스템을 예로 시스템의 적용 가능성을 제시하고, 마지막으로 5절에서는 결론을 정리하고 추후 연구 방향에 대하여 기술한다.

2. 시스템의 개요

본론에 들어가기 앞서 『상황(Context)』에 대한 명확한 정의가 필요하다. 상황은 “어떤 일의 모습이나 형편”의 사전적 의미를 가지고 있지만, 공학적으로는 “주변 조건의 변화”와 같이 매우 포괄적으로 사용되기도 하며, 특히 로봇 분야에는 로봇의 자기 위치와 같이 세부적인 정보를 의미하기도 한다. 본 논문에서는 상황을 『인간과 컴퓨터 사이에서 발생하는 행위자의 의도』라고 정의하며, 카메라를 의도 전달의 매체로 사용하기 때문에, 소리 혹은 기타 센서의 입력에 의한 상황은 정의에 포함하지 않는다.

본 논문에서는 표 1과 같이 제스처 인터페이스 시스템을 행위자 중심의 5개의 상태로 분류한다. 카메라를 통해서 입력되는 동작 영상 시퀀스는 그림 1에서와 같이 상태들의 연속적인 조합으로 표현될 수 있으며, 해당 상태 정보에 따라 행위자의 의도를 파악할 수 있다. 또한, 각 상태간의 천이는 학습을 통해 확률 모델로 표현되며, 그림 2와

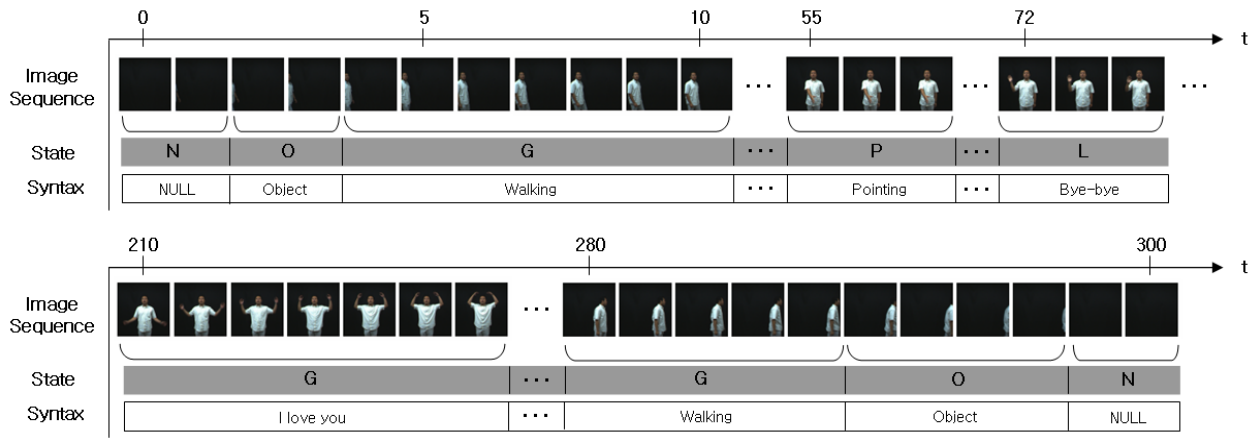

그림 1. 제스처 인식을 위한 시스템의 상태 정의
(N: Null, O: Object, P: Posture, G: Global Motion, L: Local Motion)

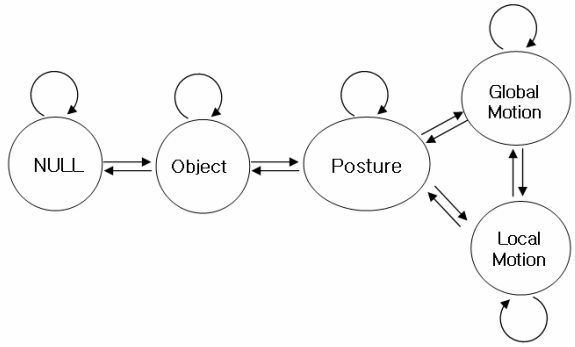

그림 2. 시스템의 상태 천이도

표 1. 시스템의 상태 정의

분류	정 의
Null	대상이 존재하지 않는 상태
Object	움직임이 있지만, 행위자로 판명되지 않은 상태
Posture	행위자가 존재하며, 움직임이 없는 정적인 포즈 상태
Global Motion	행위자의 큰 움직임에 의미가 포함되어 있는 제스처 상태
Local Motion	행위자의 지역적 움직임에 의미가 포함되어 있는 제스처 상태

같이 도식화 할 수 있다. 그림 3은 시스템의 전체 흐름도를 나타낸다. 첫 번째 단계로, 미리 얻어진 동작 영상 시퀀스를 이용해 학습 과정을 수행한다. 그 다음, 학습을 통해 얻어진 천이 조건과 특징 벡터의 가중치를 이용하여 시스템의 상태 천이를 결정하고, 각 상태에 따라 제스처 특성에 맞는 적절한 인식 알고리즘을 적용하도록 설계하였다.

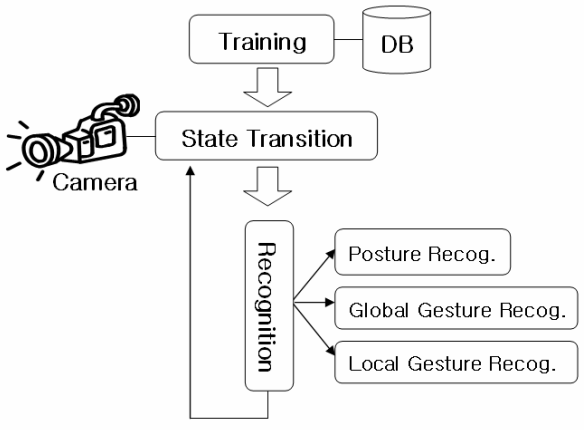

그림 3. 시스템의 흐름도

3. 확률 기반 상태 천이 모델

앞 절에서 언급한 바와 같이 시스템은 5개의 상태로 정의되며, 각 상태로의 천이 조건은 영상 시퀀스를 이용한 학습을 통해서 확률 모델로 표현된다.

3-1. 특징 추출

본 논문에서는 상태 천이를 결정하기 위한 특징으로 다음과 같은 영상 특징 정보를 이용한다.

a) f_1: 영상내의 움직임 정보
b) f_2: 행위자의 유무
c) f_3: 행위자와의 거리
d) f_4: 행위자의 움직임 정보
e) f_5: 얼굴과 양손의 이동 경로

시스템은 매 프레임 스테레오 카메라를 통해 입력되는 컬러 영상을 통해 행위자의 얼굴 및 양손 영역을 추적하고, 시차 영상을 이용하여 3차원 거리 정보를 계산하고, 특징 정보들을 추출한다.

3-2. 학습

각 상태로의 천이 과정을 확률 모델로 표현하기 위해서 미리 획득된 동작 영상 시퀀스를 이용하여 학습 과정을 수행한다. 학습 영상 시퀀스는 영상 정보와 상태 정보를 동시에 가지고 있으며, 학습을 통해 상태 천이를 확률 모델로 표현할 수 있게 된다. 이러한 학습 과정을 통해서 각 상태에서의 영상 특징 정보의 확률이 계산되고, 각 특징의 가중치가 결정된다.

5개의 영상의 특징 정보를 f_i ($i = 1,2,3,4,5$) 라 하면, 임의의 특징 값을 갖는 특징 벡터 F 와 확률 $P(F)$는 식(1)과 같이 표현된다.

$$F = \{f_1, f_2, f_3, f_4, f_5\} \qquad (1)$$

$$P(F) = \prod_i P(f_i)$$

또한, 상태를 S_k ($k = 1,2,3,4,5$) 라고 하면, 각 상태에서의 영상 특징 정보의 확률은 다음과 같이 표현할 수 있다.

$P(F \mid S_1)$: Null 상태에서의 F 발생 확률
$P(F \mid S_2)$: Object 상태에서의 F 발생 확률
$P(F \mid S_3)$: Posture 상태에서의 F 발생 확률
$P(F \mid S_4)$: Global Motion 상태에서의 F 발생 확률
$P(F \mid S_5)$: Local Motion 상태에서의 F 발생 확률

한편, 상태 천이 확률 이외에도 학습을 통해서 각 특징의 가중치 정보를 얻을 수 있다. 이 정보는 임의의 상태에서 다른 상태로의 천이가 발생할 때의 특징 벡터 기여도를 의미하며, 상태 천이를 결정하는 과정에서 불필요한 특징의 영향을 줄이는 효과가 있다.

3-3. 상태 천이

본 논문에서는 상태 천이를 결정하기 위하여 베이즈 정리(Bayes' Theorem)를 사용한다. 학습과정을 통해 얻어진 각 상태에서의 특징 발생 확률(사전 확률)을 이용해 임의의 특징에 대해 각 상태로의 천이 확률(사후 확률)을 계산하여, 최대 확률 값을 갖는 상태로 천이를 결정하는 방법이다. 임의의 특징 벡터 F 에 대하여 각 상태로의 천이 확률은 식(2)와 같이 표현될 수 있으며, 최대 천이 확률을 가지는 상태로 천이가 발생한다.

$$P(S_k \mid F) = \frac{P(S_k)P(F \mid S_k)}{\sum_i P(S_i)P(F \mid S_i)} \qquad (2)$$

이때, 학습과정에서 얻어진 가중치 정보를 이용하여 특징 벡터의 가중치 w_i ($i = 0,1,2,3,4$)를 식(3)과 같이 재조정한다. 정리하면, 그림 4와 같이 시간 t의 상태 S_j^t에서 임의의 특징 벡터 집합인 F 가 입력되었다고 가정하면, 식(4)와 같이 각각의 상태에 대하여 가중치가 적용된 특징 벡터 F' 와 사전 확률 값을 이용하여 천이 확률을 계산한 다음 가장 높은 확률 값을 가지는 상태 S_k^{t+1}로 천이가 발생한다.

$$P(F') = \prod_i w_i P(f_i) \qquad (3)$$

$$k = \arg\max_j \; P(S_j^{t+1} \mid F') \qquad (4)$$

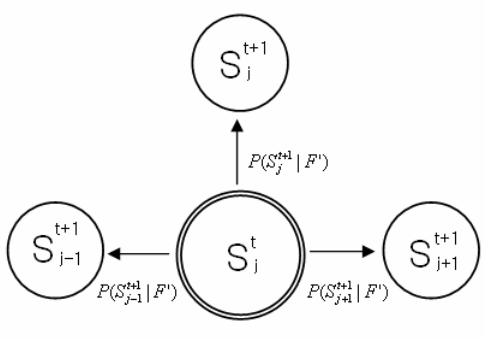

그림 4. 확률 기반 상태 천이

4. 동작 인식 시스템

서론에서 언급한 바와 같이 본 논문에서 제안하는 상황 인지 기술을 이용한 제스처 인식 방법은 각 상황에 적절한 인식 알고리즘을 적용할 수 있다는 점에서 효과적이다.

인간의 모든 제스처를 하나의 분류 특성만으로 구분하는 것은 매우 어려운 일이기 때문에 잘못된 분류 특성을 선택하는 경우 시스템 종속적인 알고리즘이 개발되기 쉽다. 본 논문에서는 이러한 위험 요소를 줄이고, 일반화된 인식 알고리즘의 적용이 가능한 제스처 인식 시스템을 제안하고자 한다.

4-1. 포즈 인식

포즈(Pose)는 움직임 특성이 없으며, 행위자의 얼굴과 양손의 기하학적 관계에 의해 정해지는 인체의 정적인 자세를 말한다. 이러한 포즈는 얼굴과 양손의 위치 정보를 이용하여 행위자의 의도를 파악할 수 있는 특징이 있다.

본 논문에서는 포즈 인식을 위한 특징으로 전처리 과정에서 얻어진 얼굴과 양손 영역의 3차원 위치 좌표를 이용한다. 얼굴과 왼손, 오른손의 3차원 위치 좌표를 각각 $H(x_h, y_h, z_h)$, $L(x_l, y_l, z_l)$, $R(x_r, y_r, z_r)$ 이라 하면, 포즈의 특징 벡터는 식(5)와 같이 정의된다.

$$F_{pose} = \{D_l, D_r, N, A\} \qquad (5)$$

여기서,

$$D_l = H - L,\ D_r = H - R,\ N = D_l \times D_r$$
$$A = a\cos\left(\frac{|D_l| \cdot |D_r|}{D_l \bullet D_r}\right)$$

이렇게 추출된 특징은 식(6)과 같이 모델 포즈 집합 M 과의 비교를 통해 최대 유사도를 갖는 포즈 P_k ($P_k \in M$)를 선택한다.

$$M_i = \{D_{lM_i}, D_{rM_i}, N_{M_i}, A_{M_i}\} \quad (M_i \in M) \qquad (6)$$

$$R_i = (D_l \bullet D_{lM_i}) + \left\| A_i - A_{M_i} \right\| + (N_i \bullet N_{M_i}) \qquad (7)$$

$$k = \arg \max_i (R_i)$$

4-2. 로컬 제스처 인식

로컬 제스처는 양손의 지역적 움직임에 행위자의 의도가 포함되어 있는 제스처이다. 예를 들어 [손 흔들기] 제스처의 경우 손의 이동 경로 보다는 손 영역의 지역적 움직임에 제스처의 의미가 포함되어 있다. 이 경우 손 영역의 형태 혹은 크기의 변화를 이용하여 구별할 수 있으며, 본 논문에서는 형태 변화의 주기성을 이용하여 제스처를 인식한다.

4-3. 글로벌 제스처 인식

글로벌 제스처는 몸 전체의 큰 움직임에 행위자의 의도가 포함되어 있는 제스처이다. 특히 얼굴과 양손의 이동 경로에 많은 의미를 포함하고 있기 때문에, 본 시스템에서는 양손의 이동경로 정보를 모델 제스처와 비교하여 제스처의 의미를 파악하는 방법을 사용한다.

2차원 공간(X-Y)상의 이동경로를 매칭하는 방법은 지금까지 매우 다양하게 시도되어 왔다. 특히 필기체 인식과 같은 분야에 많이 이용되고 있지만, 제스처 인식의 경우 행위의 시작과 끝을 결정할 수 없기 때문에 효과적으로 적용되기 어렵다. 이는 인식의 수행 시점을 결정하는데 있어서 중요한 역할을 수행하며, "Gesture Spotting" 이라는 주제로 많은 연구가 진행되고 있다[6].

본 논문에서는 복잡한 제스처가 존재하지 않는다는 가정 하에 spotting 알고리즘 대신 간단한 큐 비교 알고리즘을 사용한다. 이 방법은 처리 속도에 상당한 이점을 가지며, 구현이 용이하다는 장점을 가지고 있다. 알고리즘의 기본 개념은 다음과 같다. N 개의 모델을 갖는 모델 집합 M 을 가정하자. 각 제스처 모델(G_j)에는 식(10)과 같이 양손의 이동 경로가 k 개의 방향 벡터로 표현되어 차례로 저장되어 있다.

$$Q_{\text{model}} = \{N_1, N_2, N_3, ..., N_k\} \qquad (8)$$

$$M = \{G_1, G_2, ..., G_j, ...G_N\} \qquad (9)$$

$$G_j = \{D_1^j, D_2^j, D_3^j, ..., D_k^j\} \qquad (10)$$

$$D_i^j = N_{i+1}^j - N_i^j \quad (N_{i+1}^j, N_i^j \in Q_{\text{model}})$$

한편, 길이가 l 인 입력 큐 I 는 일정 시간 간격으로 입력되는 방향 벡터의 집합이며, 식(12)와 같이 표현된다. ($l > k$)

$$Q_{input} = \{N_1, N_2, N_3, ..., N_l\}$$

$$I^j = N_{i+1}^j - N_i^j \quad (N_{i+1}^j, N_i^j \in Q_{input}) \qquad (11)$$

$$I = \{I^1, I^2, I^3, ..., I^l\} \qquad (12)$$

만약, 입력 큐 내에 행위자의 의도가 담긴 제스처가 포함되어 있다면, 이를 의미 있는 제스처로 인정하고, 모델 큐와의 비교를 위해 식(13)에서와 같이 각각의 제스처 모델 G_j 를 입력 큐 I 와 비교한 뒤, 최대 유사도를 갖는 제스처 G_k ($G_k \in M$)로 결정한다. 그림 5는 제스처 인식 시스템의 예를 보여주며, 그림 6은 큐 비교 방법을 이용한 글로벌 제스처 인식의 예를 나타낸다.

$$R_j = \max_m \left(\sum_{i=1}^{k} (I^{m+i} \bullet D_i^j) \right) \qquad (13)$$

$$k = \arg \max_j (R_j)$$

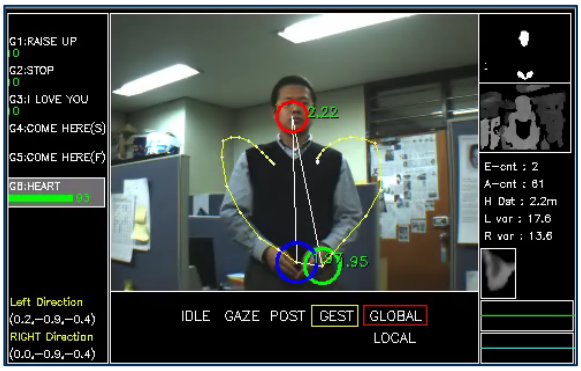

그림 5. 제스처 인식 시스템의 예.
(왼쪽은 모델 제스처와의 유사도, 노란색 선은 양손의 이동 경로를 나타낸다.)

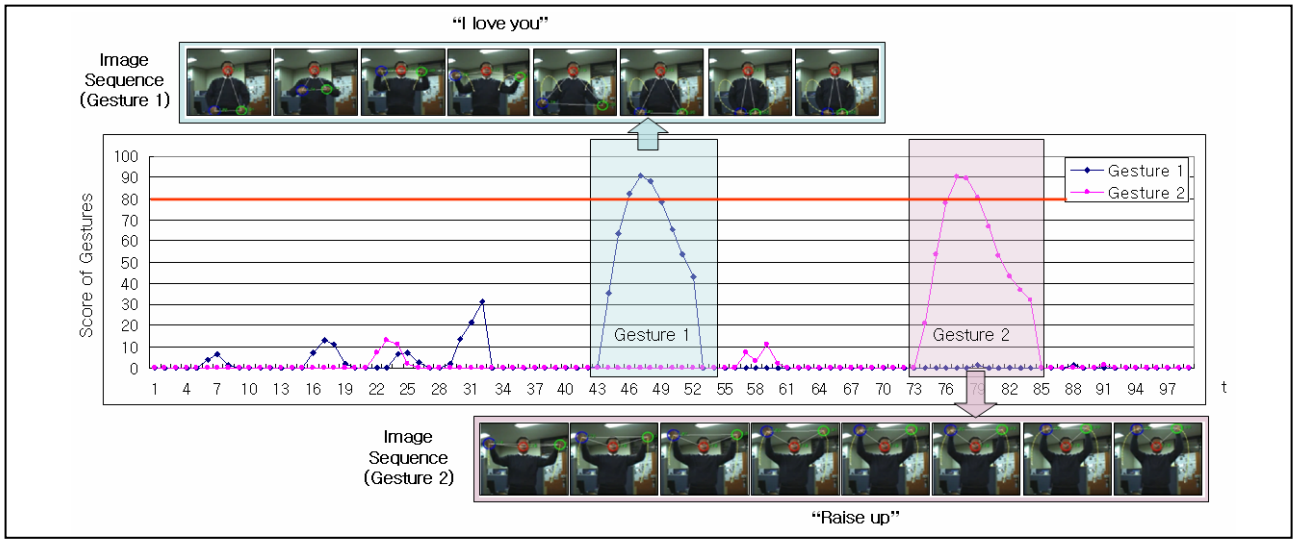

그림 6. 글로벌 제스처 인식의 예.
(그래프는 서로 다른 두 개의 연속된 제스처에 대한 각 모델 제스처와의 유사도를 나타낸다.)

5. 결론 및 향후 연구 방향

본 논문에서는 상황 인지를 이용한 제스처 인터페이스 방법에 대하여 기술하였다. 인간의 제스처를 분류하는데 있어서 개별적인 제스처 특징을 고려하는 대신 컴퓨터와 행위자 사이의 상황을 5 개의 상태로 정의하고, 이들 상태 천이를 이용하여 추상적으로 행동을 분류할 수 있었다. 또한, 각 상태에 적절한 인식 알고리즘을 적용함으로써 시스템의 효율을 높이고, 인식률을 향상시킬 수 있었다. 한편, 글로벌 제스처 인식 알고리즘에서는 spotting 알고리즘 대신 빠르고 간단한 큐 비교 알고리즘을 제안하였다.

그러나, 현재의 시스템에서는 다수의 행위자가 출현할 경우 문제가 발생한다. 두 명 이상의 행위자가 존재하는 경우 행위자의 얼굴과 양손 영역을 추적하지 못하거나, 행위자의 의도와는 상관없이 움직임 정보가 발생할 수 있기 때문이다. 이러한 문제는 효율적인 트래킹 알고리즘을 적용하거나, 인식 대상 선택 방법을 적용하면 해결될 수 있을 것이다.

본 논문에서 제안된 방법은 일반화된 제스처 인터페이스로 활용될 수 있으며, 앞서 언급한 문제점들이 해결된다면 일상생활에서 가전기기의 제어 혹은 지능형 로봇 등의 정보 시스템과 자연스러운 의사소통에 활용할 수 있을 것으로 생각된다.

[참 고 문 헌]

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

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

[3] James Davis, "Recognizing Movement using Motion Histograms", MIT Media Lab. Technical Report No. 487, March 1999

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

[5] Chil-Woo Lee, Hyun-Ju Lee, Sung H. Yoon, and Jung H. Kim, "Gesture Recognition in Video Image with Combination of Partial and Global Information", in Proc. of VCIP 2003, Lugano, July, 2003

[6] Ho-Sub Yoon; Byung-Woo Min; Jung Soh; Young-Iae Bae; Hyun Seung Yang, "Human computer interface for gesture-based editing system", Image Analysis and Processing, pp. 969 – 974, 1999