전자통신동향분석 제20권 제2호 2005년 4월

인간-로봇 상호작용을 위한 제스처 인식 기술

Gesture Recognition for Natural Human-Robot Interaction

지능형 로봇 특집

김계경 (K.K. Kim) 인간로봇상호작용연구팀 선임연구원
김혜진 (H.J. Kim) 인간로봇상호작용연구팀 연구원
조수현 (S.H. Cho) 인간로봇상호작용연구팀 연구원
이재연 (J.Y. Lee) 인간로봇상호작용연구팀 책임연구원

목 차

인간과 로봇과의 자연스러운 상호작용을 위하여 시각을 기반으로 한 사용자 의도 및 행위 인식에 대한 연구가 활발히 진행되고 있다. 제스처 인식은 시각을 기반으로 한 인식 분야에서 핵심 기술 분야로 연구되어 왔으며 최근에는 로봇이 인간에게 자연스러운 서비스를 제공해 주거나 로봇의 동작을 제어하기 위해 연구되고 있는 분야이다. 본 고에서는 기존에 제어된 제스처 인식 기술과 최근 인간-로봇의 상호작용을 위한 제스처 인식 기술에 대하여 알아본다.

Ⅰ. 서론

인간과 컴퓨터간의 정보 이동에 중요한 역할을 해온 인간-컴퓨터 상호 작용(HCI) 기술은 핵심적인 정보 기술(IT) 분야에 속한다. 인간과 컴퓨터와의 상호 작용을 위하여 인간의 시각 기능을 컴퓨터에 이식하여 컴퓨터 비전에 기반한 사용자 의도 및 행위를 인식하는 연구들이 진행되어 왔다[1]-[12].

입력장비를 사용하지 않고 컴퓨터와 인터페이스하기 위한 인간-컴퓨터 상호 작용 기술로서 제스처 인식이 연구되어 왔으며 멀티모달 인터페이스에 대한 필요성이 증가됨에 따라 제스처 인식의 중요성도 함께 높아지게 되었다. 제스처 인식은 먼 거리나 잡음 환경에서 인간과 컴퓨터간의 정보 전달 수단이 될 수 있으며 더 나아가 로봇 동작 제어, 사용자에 대한 서비스 제공 및 게임, 가상 현실 응용 등에 활용될 수 있다. 최근 몇 년 동안 로봇의 동작을 제어하기 위한 방법으로 제스처를 인식하는 방법들에 대한 연구들이 많이 진행되었다[7]-[9].

제스처 인식은 카메라 영상 기반 제스처 인식과 3D 기반 제스처 인식으로 나눌 수 있다[6]-[12]. 전자의 카메라 영상 기반 제스처 인식은 하나 혹은 그 이상의 카메라를 이용하여 제스처 영상을 획득한 다음 영상 및 모션 등과 같은 특징 정보로부터 의미있는 제스처를 추출하여 인식하는 방법이다. 인간과 로봇과의 상호 작용을 위하여 카메라 영상에 기반한 제스처 인식 방법들도 제안되었다[7]-[9]. 로봇의 동작을 제어하기 위해 제안된 카메라 영상 기반 제스처 인식 방법에서는 인식 대상 제스처가 고정된 카메라 시점에서 확연히 구별이 가능하여야 한다. 한편, 후자의 3D 기반 제스처 인식 방법은 카메라로 획득한 입력 영상으로부터 3차원 데이터를 추출하여 제스처를 인식하는 것이다.

제스처 인식을 위하여 여러 가지 방법들이 제안되었으나 대부분의 연구가 극히 제한된 실험실 환경에서 이루어졌거나 소수의 실험 대상자들에 대해서 제스처 데이터를 얻고 인식 결과를 도출하여 왔다. 또한, 다양한 조명 환경 및 배경을 가지는 실세계 환경에서 제스처를 검출하여 인식하는 것은 상당히 어렵다. 특히, 시ㆍ공간에 따른 제스처 데이터의 다양성 및 제스처 동작자에 제한을 두지 않은 상황에서 연속 동작 제스처로부터 의미있는 단일 제스처를 추출하여 인식한다는 것은 그리 쉬운 문제가 아니다.

시ㆍ공간적으로 변하는 제스처의 형세를 고려하여 지식 기반 방법에 의하여 제스처를 인식하여야 한다고 논의되었다. 제스처를 인식하는 방법에 있어서는 기존에 많이 적용되었던 HMM, 신경망 및 특징 기반 통계적 방법들은 특정 제스처 모델을 미리 가정하여 인식하는 방법들이므로 동작자에 제한을 두지 않은 환경에서 다양한 제스처를 인식하지 못하는 문제점이 있다. 따라서, HMM이나 신경망을 컨텍스트 정보와 함께 사용하는 복합 제스처 인식기를 사용함으로써 기존 제스처 인식의 문제점을 해결하려고 하였다.

카메라 영상 기반 제스처 인식 기술을 인간-로봇 상호작용 목적으로 활용하기 위해서는 일반 범용 사용자를 대상으로 인식 대상 제스처의 종류를 10개 이내로 제한하여 복합 특징 정보 값 및 복합 제스처 인식기를 사용하여야 한다. 또한, posture 정보를 같이 활용하면 보다 나은 인식 성능을 보장할 수 있을 것이다. 인간-로봇간의 자연스러운 상호작용을 위한 제스처 인식 기술의 개발은 기계적이고 인위적인 컴퓨터 인터페이스 환경을 자연스러운 컴퓨팅 환경으로 대체할 수 있다는 데 그 중요성을 가지게 될 것이다.

Ⅱ. 기술별 현황

1. 카메라 영상 기반 제스처 인식

카메라 영상 기반 제스처 인식은 카메라로 사람 영상을 획득한 다음 의미 있는 제스처를 검출하고 제스처를 계속 추적하면서 제스처의 의미를 분석하여 인식하는 단계를 포함한다. 고정된 카메라 시점에서 확연히 구별 가능한 일곱 내지 아홉 가지 종류의 제스처를 인식 대상으로 하여 template

matching, 은닉 마르코프 모델(hidden Markov model) 또는 신경망을 이용하여 제스처를 인식하여 왔다. 이 방법에서는 인식 대상 제스처의 종류를 잘 선택하면 오히려 3D 기반 제스처 인식 방법보다 문제의 복잡성은 줄이고 더 나은 인식 성능을 보장할 수 있다는 장점이 있다. 그러나, 고정된 카메라 시점에서 구별이 안되는 제스처들이 존재하므로 인식 대상 제스처의 수가 극히 제한적이라는 문제점이 있다.

카메라 영상 기반 제스처 인식 방법은 trajectory 기반 인식 방법 및 history 기반 인식 방법으로 나눌 수 있다[6]. 전자의 방법에서는 의미 있는 제스처의 추적을 통해 생긴 제스처의 궤적을 특징 정보로 사용하여 HMM을 이용하여 제스처를 인식한다. 반면, 후자의 방법에서는 연속된 영상으로부터 제스처로 추출된 화소들을 모두 합하여 제스처 패턴을 구한 다음 통계적인 방법을 이용하여 제스처를 인식한다.

로봇의 동작을 제어하기 위하여 제안된 카메라 영상 기반 방법에서는 영상 시퀀스를 이용하여 동적 패턴 정합, 통계적 방법, 신경망 및 HMM을 이용하여 다섯 종류의 제스처를 인식하였다[7]-[9]. 또한, 영상 특징 정보 및 모션 정보를 이용하여 제스처를 검출하고 의미 있는 제스처 좌표를 추적한 다음 제스처 프레임 길이에 무관한 특징 벡터를 추출한 다음 여섯 종류의 제스처를 신경망을 이용하여 인식한 다음 로봇의 동작제어를 위하여 사용하였다[9].

일반적으로 카메라 영상 기반 제스처 인식 방법에서는 연속된 영상 프레임으로부터 피부색 등과 같은 영상 특징 정보 및 모션 정보를 구하여 제스처를 추출하는 데 이용하였다. 연속된 동작으로부터 제스처의 시작점과 끝점을 찾아내기 위하여 동작자 제스처의 위치값을 계속 추적하면서 의미 있는 제스처를 추출하였다. 시간 변화에 따른 제스처의 위치 값 및 각도 등을 이용하여 제스처 인식을 위한 특징 벡터를 추출한 다음 HMM이나 신경망 등을 이용하여 제스처를 인식하도록 하였다. 제스처 인식에 대한 성능 평가는 연속 동작 제스처에서 의미있는 제스처를 분할한 다음 인식하는 방법 및 단일 동작 제스처를 인식하는 방법으로 행해졌다.

2. 3D 기반 제스처 인식

3D 기반 제스처 인식 방법에서는 손의 궤적 정보만으로 구별하기 어려운 제스처를 인식하기 위하여 한 차원의 데이터를 입력에 추가하여 제스처를 인식하도록 하였다. 그러나, 전자의 2차원 데이터를 이용한 제스처 인식 방법과 비교해 볼 때 인식 대상이나 방법이 크게 다르지 않고, 오히려 추가된 한 차원의 입력 데이터 정보로 인하여 문제의 복잡도를 더 증가시켜 제대로 인식 결과를 내지 못하는 경우가 많이 발생하고 있다.

3차원 제스처 인식에 대한 연구는 MIT에서 가장 활발히 연구하여 왔다. 스테레오 블럽(blob) 추적기를 이용하여 3차원 손의 궤적 데이터를 입력으로 하여 HMM을 이용하여 양손의 제스처를 인식하는 방법[10], 3차원 데이터를 이용하여 웨이브 제스처와 포인팅 제스처를 상태 기반 접근 방법에 의하여 분할하는 방법[11] 및 3차원 입력을 통해 4개의 단순한 오른손 제스처를 분석하는 방법[12] 등이 제안되었다.

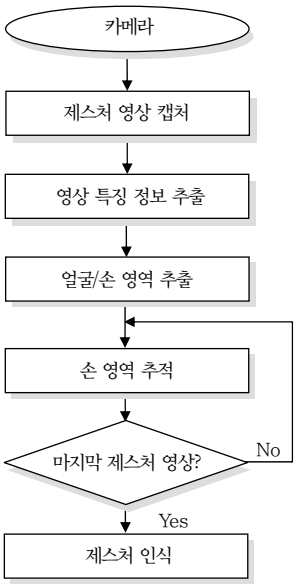
Ⅲ. 로봇 동작 제어를 위한 제스처 인식

III장에서는 로봇 동작 제어를 위한 카메라 영상 기반 제스처 인식 방법에 대하여 구체적으로 서술하였다. 카메라 영상 기반 제스처 인식에 대한 전체 흐름도는 (그림 1)에서와 같다.

1. 손의 위치 검출

가. 영상 특징 정보를 이용한 손의 위치 검출

연속된 영상 프레임에서 사람의 피부색 정보를 이용하여 사람의 얼굴 영역과 손 영역을 추출하도록



(그림 1) 카메라 영상 기반 제스처 인식

한다. 사람의 얼굴 영역과 손 영역 이외의 다른 영역에서도 피부색에 해당되는 후보 영역들이 추출된다. 이러한 가상 피부색 영역들은 피부색 영역들에 대한 블럽 정보, 즉 블럽의 크기, 위치 및 형태를 분석하여 사람의 얼굴과 손의 위치를 정확히 추출하도록 한다. 사람의 피부색을 나타내는 얼굴 및 손 영역을 추출하기 위하여 (1)을 이용한다.

$$C_t^b(x,y) > \alpha \times C_t^r(x,y) \,\&\, C_t^b(x,y) > \beta \times C_t^r(x,y) \,\&$$
$$C_t^g(x,y) > \gamma \times C_t^r(x,y) \,\&\, C_t^g(x,y) > \delta \times C_t^r(x,y) \,\&$$
$$\varepsilon < H(x+w\times y) < \eta \,\&\, \omega < S(x+w\times y) < \xi \qquad (1)$$

여기서 $C_t^{\,r}$, $C_t^{\,g}$ 및 $C_t^{\,b}$는 각각 t번째 프레임 영상의 r, g, b값을 나타내며 H와 S는 $C_t^{\,r}$, $C_t^{\,g}$ 및 $C_t^{\,b}$로부터 구한 H(Hue)와 S(Saturation) 값이다. (1)에서 α 등의 각 파라미터 값들은 실험에 의해 구해진 값들이다. (1)을 만족하는 화소 (x,y)는 피부색 영역으로 간주한다.

(1)에 의해 구해진 피부색 영역들에 대한 블럽들에 대하여 블럽의 크기, 위치 및 형태 정보를 해석하여 얼굴 영역을 추출하도록 한다. $B_t^{\,i}$를 t번째 프레임 영상에서의 i번째 블럽이라고 할 때 이 블럽의 가로 대 세로의 비 AR와 블럽의 크기 및 위치에 대한 조건이 (2)를 만족하면 얼굴 영역 블럽으로 간주한다.

$$T_1 < AR\{B_t^{\,i}\} < T_2 \,\&\, |B_t^{\,i}| > T_3 \qquad (2)$$

여기서 T_1, T_2 및 T_3는 실험에 의해 구해진 값들이다. (1)과 (2)를 이용하여 구해진 얼굴 영역을 기준 영역으로 하여 손의 영역을 추출하도록 한다.

나. 모션 정보를 이용한 손의 위치 검출

획득한 영상 시퀀스로부터 연속된 영상 세 프레임, 즉 t-1, t, t+1 프레임간의 화소 (x,y) 차가 임계치보다 큰 경우 모션이 발생한 화소로 간주하며 (3)과 (4)를 이용하여 모션 화소들을 구한다.

$$|f_t(x,y) - f_{t-1}(x,y)| > T_4 \qquad (3)$$

$$|f_{t+1}(x,y) - f_{t-1}(x,y)| > T_4 \qquad (4)$$

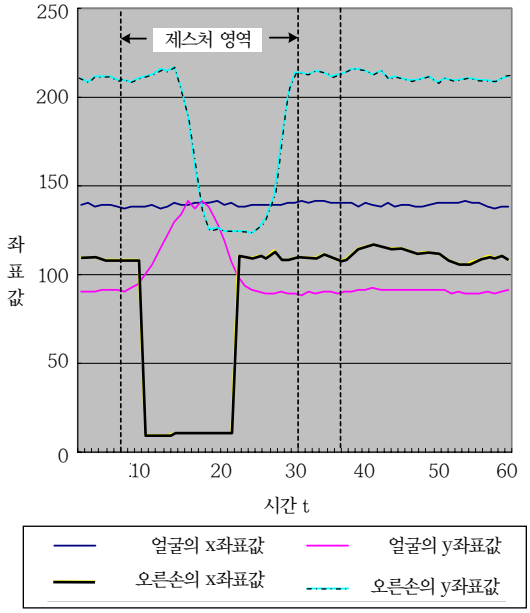
여기서, f_t, f_{t-1}, f_{t+1}은 t-1, t, t+1에 대한 각 프레임을 나타내며, T_4는 임계치를 나타낸다. 임계치는 대략적으로 10과 20 사이의 값을 가진다.

모션 정보와 영상 특징 정보를 이용하여 추출된 얼굴 영역과 손 영역은 (그림 2)에서와 같다.

(그림 2) 추출된 얼굴 영역 및 손 영역

2. 제스처 좌표 추적

연속 동작 제스처 프레임에서 의미있는 제스처 영역을 분리하기 위하여 연속된 프레임간의 얼굴 및 손 좌표 변화값을 계산한다. 이때 좌표의 변화값이



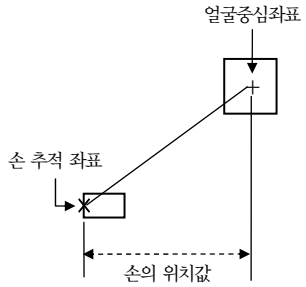
(그림 3) 한 동작의 제스처 분리

일정 임계치 이상이 되면 제스처 시작 프레임으로 간주하고 그 다음의 영상 프레임에서 계속 얼굴 및 손의 좌표값을 추적하도록 한다. 연속된 프레임간의 좌표 변화값이 일정 임계치 이하가 되면 제스처가 끝났다고 가정하고 제스처가 시작된 프레임부터 끝난 프레임까지를 의미있는 한 동작의 제스처 구간으로 간주한다. 제스처가 시작된 손의 좌표값 및 제스처가 끝난 손의 좌표값을 비교하여 더 정확한 제스처 영역이 추출될 수 있도록 한다. (그림 3)은 연속 동작 제스처 프레임에서 한 동작의 제스처를 분리한 것을 나타낸 것이다.

3. 특징 추출

연속 영상 프레임으로부터 추출된 얼굴 좌표값으로부터 얼굴의 중심점을 계산하여 얼굴을 기준으로 하여 상대적으로 움직인 손의 위치값을 계산하는 데 사용한다. 제스처 인식을 위한 특징 벡터는 얼굴을 중심으로 움직인 손의 상대 위치 값 및 손의 궤적값으로부터 추출하도록 한다. (그림 4)에서는 얼굴 중심점을 기준점으로 하여 손의 상대적인 위치값을 추출하는 것을 나타내었다.

(a) t번째 영상 프레임에서의 얼굴 중심점 및 손의 좌표 추출



(b) 손의 상대적 위치 값 추출

(그림 4) 손의 위치 값 추출

4. 제스처 인식

여섯 가지 유형의 제스처를 인식하기 위해 한 개의 은닉층을 가지는 MLP를 이용하였으며 개선된 backpropagation 알고리듬[13]으로 구현하였다. 사용된 알고리듬에서는 P개의 입력 패턴과 N개의 출력 뉴런 사이에서 목적값과 실제 출력값 사이의 오차 누적값을 계산하고 전체 오차값이 미리 정의된 임계치보다 작은 값을 가질 때까지 가중치를 계속 갱신하도록 하는 방법을 사용하였다.

로봇 동작 제어를 위한 카메라 영상 기반 제스처 인식 방법에 대한 성능 평가는 손의 위치를 추적하여 의미있는 제스처를 분리하여 추적하는 것과 여섯 가지 종류의 제스처에 대한 인식 성능을 평가하는 것으로 나누어서 행해졌다.

제스처 추적에 대한 실험에서는 영상 특징정보와 모션 정보를 이용하여 얼굴과 손의 위치를 추출한 다음 의미 있는 제스처를 분리하도록 하였다. 실험

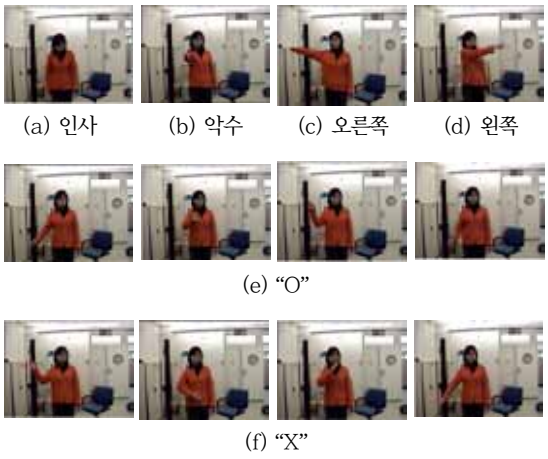
결과 84% 정도의 제스처 추적률을 나타내었다. 제스처를 추적하는 데 있어서 입력 영상에서 피부색과 비슷한 영역이 영상 전반에 나타난 경우 다른 영역을 손의 영역으로 잘못 추출하는 경우가 발생하였다. 또한, 의미있는 제스처의 시작점과 끝점을 제대로 판단하지 못하여 제스처 추적에 실패하는 경우도 발생하였다.

제스처 인식 성능을 평가하기 위하여 (그림 5)에서 나타낸 여섯 가지 종류의 제스처에 대해 인식실험하였다. 단일 동작의 제스처 데이터를 학습시킨 다음 연속 동작의 제스처 프레임으로부터 한 동작의 제스처를 분리한 다음 학습시킨 신경망을 이용하여 인식 성능을 평가하도록 하였다. (그림 5)는 여섯 가지 종류의 제스처, 즉 "인사", "악수", "오른쪽", "왼쪽", "O", "X"를 각각 나타낸 것이다.

제스처 학습 데이터베이스는 9명의 사람으로부터 얼굴 및 오른손의 x, y 좌표값을 추출하여 구현하였으며 이 데이터는 제스처 인식기를 학습하는 데 사용되었다. 또한, 제스처 인식 성능을 테스트하기 위하여 신경망 학습에 관여하지 않은 5명의 사람을 대상으로 얼굴 및 오른손 x, y 좌표값을 추출하여 실험하도록 하였다.

〈표 1〉은 여섯 가지 종류의 인식 대상 제스처에 대해 학습에 사용된 54가지 패턴의 제스처를 인식한 결과와 학습시키지 않은 38가지 패턴의 제스처를 인식해 본 결과를 나타낸 것이다. 〈표 2〉는 제스처 인식기의 성능 분석을 위하여 〈표 1〉의 인식 결과에 대한 confusion matrix를 나타낸 것이다.

〈표 1〉과 〈2〉의 결과로부터 "오른쪽", "왼쪽" 제스처가 "악수" 제스처로 오인식되는 경우가 발생하였다. 오인식된 제스처 데이터를 분석해 본 결과 사람에 따라 "악수"와 "오른쪽" 또는 "악수"와 "왼쪽"에 대한 제스처 패턴이 유사한 경우가 흔히 발생함을 확인할 수 있었다. 이러한 유사 제스처 패턴에 대한 오인식률을 줄이기 위한 방안으로 새로운 특징 벡터 추출 및 인식 방법이 모색되어야 할 것이다.



(a) 인사 (b) 악수 (c) 오른쪽 (d) 왼쪽

(e) "O"

(f) "X"

(그림 5) 여섯 가지 종류의 제스처 유형

〈표 1〉 제스처 인식 결과

인식대상 제스처	학습 DB에 대한 인식률(%)	테스트 DB에 대한 인식률(%)
인사	100	100
악수	100	86
오른쪽	100	71
왼쪽	100	67
O	100	80
X	100	100
계	100	84

〈표 2〉 〈표 1〉의 인식 결과에 대한 Confusion Matrix

인식 / 인식대상	C[0]	C[1]	C[2]	C[3]	C[4]	C[5]	인식대상 제스처수	인식 제스처수
인사: C[0]	8	-	-	-	-	-	8	8
악수: C[1]	-	6	1	-	-	-	7	6
오른쪽: C[2]	-	1	5	-	1	-	7	5
왼쪽: C[3]	-	2	-	4	-	-	6	4
O: C[4]	-	-	-	-	4	1	5	4
X : C[5]	-	-	-	-	-	5	5	5
인식률 (%)	100	86	71	67	80	100	38	32

Ⅳ. 결론

본 고에서는 인간-로봇의 상호 작용을 위한 제스처 인식 방법에 대해서 살펴보았다. 카메라 영상 기반 제스처 인식 방법, 3D 기반 제스처 인식 방법 및 로봇을 제어하기 위한 카메라 영상 기반 제스처 인식 방법에 대하여 고찰하였다. 시ㆍ공간에 따른 제스처 데이터의 다양성 및 제스처 동작자에 제한을 두지 않은 상황에서 연속 동작 제스처로부터 의미있는 제스처를 추출하여 인식한다는 것은 그리 쉬운 문제가 아니다. 카메라 영상 기반 제스처 인식 방법을 사용하여 로봇의 동작을 제어할 수 있는 제스처를 인식할 경우에는 인식 대상 제스처는 고정된 카메라 시점에서 확연히 구별이 가능하여야 한다. 또한, 다양한 제스처 프레임 길이에 따른 오인식 및 제한되지 않은 동작자의 제스처를 연속 동작 프레임으로부터 추출할 때 발생한 오류를 개선하기 위하여 정규화된 특징 벡터 추출 방법, 구조적, 통계적인 복합 특징 벡터 추출 방법 및 HMM, 신경망과 같은 이종 인식기 및 컨텍스트 정보의 결합을 통한 복합 제스처 인식기 구현 등에 대한 연구가 계속 진행되어야 한다.

약어 정리

AR	Aspect Ratio
HMM	Hidden Markov Model
MLP	Multi-Layer Perceptron

참 고 문 헌

[1] A. Ali and J.K. Aggarwal, "Segmentation and Recognition of Continuous Human Activity," IEEE Detection and Recognition of Events in Video, 2001.

[2] A. Bobick, "Real Time Online Adaptive Gesture Recognition," *IEEE ICPR*, 2000.

[3] B. Li and H. Holstein, "Recognition of Human Periodic Motiona Frequency Domain Approach," *IEEE ICPR*, 2002.

[4] D. Ayers and M. Shah, "Recognizing Human Actions in a Static Room," *IEEE WACV*, 1998.

[5] A. Corradini, "Intergrated Dynamic Time Warping for Off-line Recognition of a Small Gesture Vocabulary," RATFG-RTS, 2001.

[6] K. Morrison and S.J. McKenna, "An Experimental Comparison of Trajectory-based and History-based Representation for Gesture Recognition," GW 2003, LNAI 2915, 2004, pp.152-163.

[7] A. Corradini and H.M. Gross, "Camerabased Gesture Recognition for Robot Control," IJCNN, 2000.

[8] Andrea Corradini, "Dynamic Time Warping for Off-line Recognition of a Small Gesture Vocabulary," RATFG-RTS, 2001.

[9] 김계경, 김혜진, 조수현, 이재연 "로봇 동작 제어를 위한 제스처 인식," IPIU, 2004.

[10] M. Brand, N. Oliver, and A. Pentland, "Coupled Hidden Markov Models for Complex Action Recognition," *IEEE Proc. Computer Vision and Pattern Recognition(CVPR)*, 1997, pp.994-999.

[11] A.F. Bobick and A.D. Wilson, "A Statebased Approach to the Representation and Recognition of Gesture," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.19, Issue 12, Dec. 1997, pp.1325-1337.

[12] G.S. Schmidt and D.H. House, "Towards Model-based Gesture Recognition," *Fourth IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, Grenoble, France, Mar. 26-30, 2000, pp.416-421.

[13] B. Kosko, "Neural Networks and Fuzzy System," Prentice-Hall, Englewood Cliffs, NJ, 1992.