전역 및 부분 특징 정보를 이용한 제스처 인식
(Gesture Recognition using Global and Partial Feature Information)

이 용 재 † 이 칠 우 ††
(Yong Jae Lee) (Chil Woo Lee)

요 약 본 논문에서는 다중 혼합 특징 정보를 저 차원 제스처 심볼로 구성하여 제스처를 인식하는 알고리즘에 대해 기술한다. 기존의 기하학적인 특징 기반 방법이나 외관기반 방법에서는 팔, 다리의 위치나 몸의 형상 정보만을 특징 값으로 이용하기 때문에 유사한 신체 동작이나 신체 부위의 움직임에 따라 애매한 결과를 나타내었지만 제안한 방법은 신체의 어느 부위가 움직이는지를 나타내는 부분특징정보(partial feature information)와 전체적인 신체의 형상을 표현하는 전역특징정보(global feature information)를 이용함으로써 동작의 구분뿐만 아니라 유사한 동작을 인식할 수 있는 장점이 있다. 그리고 비교적 적은 계산량과 높은 인식률 때문에 감시 시스템이나 지적 인터페이스 시스템 같은 여러 응용 분야에 적용될 수 있다.

키워드 : 제스처 인식, 모션 인식, 부분특징정보, 전역특징정보, 영상 그룹핑, PCA, HMM

Abstract This paper describes an algorithm that can recognize gestures constructing subspace gesture symbols with hybrid feature information. The previous popular methods based on geometric feature and appearance have resulted in ambiguous output in case of recognizing between similar gesture because they use just the position information of the hands, feet or bodily shape features. However, our proposed method can classify not only recognition of motion but also similar gestures by the partial feature information presenting which parts of body move and the global feature information including 2-dimensional bodily motion. And this method which is a simple and robust recognition algorithm can be applied in various application such surveillance system and intelligent interface systems.

Key words : Gesture recognition, Motion recognition, Partial and global feature information, Image grouping, PCA, HMM

1. 서 론

인간은 일상생활에서 제스처(gesture), 표정과 같은 비언어적인 수단을 이용하여 수많은 정보를 교환한다. 예를 들어 얼굴표정, 손의 움직임, 시선 방향, 머리의 동작, 신체의 포즈 등이 대화 중에 상대편에게 많은 정보를 전달한다. 이러한 비언어적 수단을 정보기기와 인간의 대화에 사용한다면 보다 사용자 친화적인 인터페이스 실현이 가능하게 된다. 특히 근래에 들어 입는 컴퓨터(wearable computing), 지적 인터페이스(perceptual user interface), 유비쿼터스 컴퓨팅(ubiquitous computing)과 같이 제4세대 정보기술의 중요성이 강조되면서 인간의 행동에 대한 각종 인식은 컴퓨터비전 연구자들의 많은 주목을 받고 있다[1].

컴퓨터를 사용하여 동작을 인식한다는 것은 인체 각 부위가 시간의 흐름에 따라 어떤 모습으로 변화하는 가를 자동으로 분석하고 그 변화를 추상적인 의미로 해석하는 것을 의미한다[2]. 즉 동영상으로부터 신체 영역을 추출한 다음 특정 부분을 식별(identify)하고 각 부분들이 하나의 의미를 갖기 위해 어떤 변화를 거치는지를 알아내는 것이다. 그러나 인체는 고자유도(high degree of freedom)를 지닌 매우 복잡한 3차원 관절 물체(three dimensional articulate object)로 2차원의 동영상으로부터 인체부위를 안정적으로 분리해 내고 그 내용을 인식한다는 것은 매우 어려운 일이다. 또, 사람에

† 정 회 원 : 전남대학교 컴퓨터공학과
ufosklee@paran.com
†† 종신회원 : 전남대학교 컴퓨터공학과 교수
leecw@chonnam.ac.kr
논문접수 : 2005년 3월 29일
심사완료 : 2005년 8월 1일

따라 착용하는 의복이 다르므로 특징정보(feature information)를 안정적으로 추출하기가 어려워 많은 노력에도 불구하고 만족할 만한 결과를 얻기가 어렵다.

인체의 시간에 대한 형상 변화를 알아내는 가장 쉬운 방법은 인체 각 부위를 연결하는 관절 각의 변화를 관찰하는 것이다. 따라서 인간 동작 인식에 관한 초기연구에서는 인체 각 부위의 관절에 센서를 부착하고 센서로부터 얻어진 관절 각의 변화를 시간을 기준으로 패턴으로 분류하여 인식하였다[3]. 이 방법은 장치를 몸에 붙이는 과정이 복잡하고 초기 교정이 어려울 뿐만 아니라 센서와 컴퓨터간의 연결 케이블 때문에 자연스러운 제스처 입력이 불가능하여 현재는 거의 사용되고 있지 않다.

최근에 들어, 특징 점 추출의 어려움을 피하기 위해 광학적 마커(marker)를 신체의 중요 부위에 부착하고 카메라로 입력된 영상으로부터 이 마커들의 궤적을 추적하여 동작을 인식하는 방법들이 개발되었다[4]. 이 방법은 제스처를 인식한다기보다는 인체의 3차원 움직임을 각 관절 각을 기준으로 정확히 측정하여 그 결과를 수치적인 데이타로 표현이 가능하기 때문에 애니메이션이나 영화의 제작에 활용할 수 있는 매우 유용한 방법이다. 대개의 경우 미리 교정(calibration)된 여러 대의 카메라를 이용하여 각 마커의 3차원 위치를 측정하게 되고 신체의 일부분이 가려져 마커를 추정하기 곤란한 경우는 수 작업을 통해 마커의 위치를 지정하거나 입력하여 정확한 계산을 돕기도 한다. 이 방법은 정확한 측정을 위해 매우 유용한 방법이기는 하나 영상의 입력을 위해 특별히 제작된 스튜디오와 고가의 장비가 필요하므로 간단한 응용 소프트웨어를 제작하기에는 부적당하다. 또 감시 카메라의 경우와 같이 관찰대상에게 마커를 부착할 수 없는 경우 근본적으로 적용이 불가능하다.

인간 동작 인식을 위한 가장 이상적인 방법은 인식대상에 마커를 붙이지 않고 자연스럽게 사람이 눈으로 대상 인간을 관찰하듯 비디오카메라로 입력된 영상만을 분석하여 그 의미를 알아내는 것이다. 일반 비디오 영상으로부터 제스처를 인식하는 방법은 크게 3가지로 나눌 수 있다. 첫 번째 방법은 신체의 특징 점을 추출하여 그 특징 점의 운동을 시간에 대한 패턴으로 구별하여 인식하는 방법이다. 이 경우 대개 영상이 주어지면 전처리 과정을 거쳐 특징 점을 추출하고 이 특징 점을 추적하여 포즈(pose)를 추정하고 이 포즈의 시간적인 변화를 동작의 의미로 해석하는 방법이다[5-7]. 이 방법은 앞서 기술한 바와 같이 복장이나 카메라의 시선방향에 따라 특징 점이 달라지고, 복잡한 움직임의 경우 특징 점들이 서로 겹치기 때문에 특징 점 추출이 어려워 고도의 인식은 불가능하다. 따라서 머리, 양 손 끝, 양 발의 위치 등 가장 안정적으로 추출할 수 있는 특징 점들을 이용하여 제스처를 추정하는 경우가 많다[5,7]. 두 번째 방법은 인체의 형상 모델을 이용하는 방법이다[4,8]. 미리 일반화된 인체의 2차원 혹은 3차원 모델을 만들어 놓고 그 모델을 적당한 가설에 의해 변형시켜 2차원 영상 공간으로 투영한 다음 입력된 영상과 생성된 영상을 비교하여 인체의 자세를 인식하는 방법이다. 이 방법은 형상 모델을 어떻게 변형시킬까에 대한 탐색공간이 커져 계산 량이 많아지는 단점이 있다. 따라서, 원통과 2차원 곡면과 같은 단순 기하구조를 이용하여 계산 량을 줄이는 방법이 많이 연구되고 있다[4]. 세 번째 방법은 영상에 나타나는 인체의 전체적인 모습(whole bodily appearance)을 여러 가지 매개변수(parameter)를 이용하여 표현하고 그 변화를 해석하여 동작을 인식하는 방법이 있다[9,10]. 이 방법은 특징 점을 추출하지 않고 대개의 경우 인체 부분을 그림자영상(2치 영상)으로 변환하여 의복에 의한 영향을 제거한 후, 그 시계열 영상이 갖는 통계적 특성이나 기하학적 정보를 분석하여 동작을 인식하는 방법이다. 인체의 복잡한 특징점을 추출하지 않고 전체의 외관을 하나의 분석 대상으로 삼기 때문에 잡음과 환경변화에 영향을 받지 않아 구현이 쉽고 인식결과가 안정하다는 장점이 있으나, 인체의 동작의 상세한 정보를 인식할 수 없다는 단점이 있다.

외관기반 동작 인식의 원리는 인간의 한 동작을 여러 가지 자세의 집합으로 보고 각 자세들이 시간적으로 올바른 순서대로 나열되어 있는 가를 판단하는 것이다. 따라서 연속영상으로부터 인체의 외관을 계산하여 그 외관이 특정 동작에 소속된 자세인가 아닌가를 판단하고, 올바른 자세라면 정해진 순서에 맞는 가를 확인하면 동작인식이 가능하게 된다. 그러나 어떤 동작에 대해 연속영상으로부터 얻어지는 인체의 모습은 사람에 따라 다르고 또, 카메라의 시선 방향에 따라 달라지므로 안정되고 일반화된 인식방법을 개발하기는 매우 어렵다. 일반적으로 인식 결과를 안정시키기 위해선 시간적인 연속성을 반드시 고려해야 되며, 많은 경우 연속 자세를 상태변화로 표현한 HMM 인식 방법이 인식의 최종과정에 같이 사용된다.

외관기반 동작인식의 대표적인 방법으로 실루엣 영상의 시간적 변화를 하나의 영상으로 누적시켜 패턴화하여 사용하는 방법이 있다[10]. 이 방법에서는 동작의 시간적 위치 변화를 나타내는 MEI(Motion Energy Image)라는 2진 영상과 동작의 시간적 경과 정보(즉, 새로 추가된 화소일수록 휘도치가 밝고 시간이 지난 화소일수록 휘도치가 낮게 됨)를 표현하는 MHI(Motion

History Image)라는 두 개의 패턴을 만들어, 각 동작들의 영상에 대한 불변 모우멘트를 비교하여 동작을 인식하는 방법이다. 카메라의 시선 방향을 달리하여 촬영된 영상을 이용하여 모델 패턴을 만듦으로써 회전 불변의 인식이 가능하고 실루엣 영상을 이용하기 때문에 구현이 간단하다. 그러나 이 방법은 하나의 동작이 완전히 끝났을 때를 가정하여 동작모델을 만들었기 때문에 실시간으로 동작을 인식할 경우 동작도중에 있는 동작을 인식하기 어렵다는 단점이 있다. 또, 우발적인 변화가 동작 내에 일어났을 경우 오 동작을 일으키기 쉽다.

외관기반 동작인식의 가장 큰 장점은 신체의 특징 점을 추출하지 않기 때문에 매우 안정된 인식 결과를 얻을 수 있다는 것이다. 그렇기 때문에 이 방법의 가장 결정적인 단점은 인체의 특정 부위가 갖는 중요성을 전혀 고려하지 않고 있다는 점이 된다. 예를 들어 손을 흔들어 "바이 바이(bye bye)"를 나타내는 동작의 경우 인체 전체의 형상은 거의 변화가 없으나 손 부분에는 매우 큰 변화가 일어난다. 이 경우 인체의 다른 부분은 동작 인식에 전혀 영향을 끼치지 않으므로 손 영역만 추출하여 그 변화를 해석하면 동작을 인식할 수 있다. 단순한 외관기반 인식 방법을 사용할 경우 전체 형상에 대해 손의 변화란 매우 미미한 것이므로 한 손을 올리고 있는 것과 손을 흔드는 것은 구별하기가 매우 어렵다. 이것은 특정 동작의 경우 손, 머리, 발등과 같은 인체의 돌출부위, 즉 특징부위가 갖는 의미가 매우 크다는 것은 의미한다. 인간이 타인의 행동을 관찰 할 경우 대개 외관으로부터 개략적인 의미를 추정하고 필요한 경우 특정부위를 세심히 관찰함으로써 정확한 인식을 한다는 사실은 널리 알려져 있다. 그러나 영상으로부터 특정 부위를 추출한다는 것은 앞서 기술한 바와 같이 간단한 일이 아니다. 계산 량이 많아질 뿐만 아니라 추출결과가 불안정하여 오히려 인식에 장애가 되는 경우가 많다. 따라서 외관기반 인식 방법의 테두리를 벗어나지 않는 범위에서 인체의 전체 형상을 표현하는 정보와 특정 부위 정보를 결합하면 보다 정확한 동작 인식이 가능하게 된다.

본 논문에서는 연속적인 영상 시퀀스에서 나타나는 신체 모습, 즉 외관(appearance)을 간단한 2차원 기하 정보로 표현하고, 이 정보를 다차원 특징의 양으로 결합한 뒤 시간에 따른 신체의 형상 변화를 축소된 특징 매개변수 공간에서 인식하는 방법에 대해 기술한다. 입력 영상은 전처리 과정을 통해 실루엣 영상으로 변환되며 이 실루엣 영상의 인체 영역에서 추출된 특징량은 주성분 분석법(PCA : Principal Component Analysis)이라는 통계적인 수법에 의해 인체의 외관 특징들을 표현할 수 있는 저 차원 벡터 공간, 즉 파라메트릭 고유 공간으로 투영된다. 각 동작들은 이 공간 내에서 순차적으로 연속된 점의 궤적으로 표현되고, 미리 학습된 모델 궤적과 입력영상의 궤적을 비교함으로써 동작인식이 이루어지게 된다. 아울러 보다 안정된 인식을 위해 최종 인식 결과를 HMM을 사용하여 판단하였다.

본 논문에서 제안한 방법의 특징은 인체 형상의 전체적인 모습을 나타내는 전역특징정보(global feature information)와 부분특징정보(partial feature information)를 결합하여 외관 기반 동작인식의 인식 결과를 향상시켰다는 점이다. 즉, 기존의 외관기반 동작인식 방법에 신체의 외부로 돌출 하는 부분특징 정보를 결합함으로써 기존 방법의 단점을 개선하였다. 또, 인체 형상을 표현하기 위해 사용된 특징량들은 인체영역의 2차원 특징을 나타내는 매우 간단한 값들로 계산이 쉬우며, 부분특징정보를 결합하기 위해 인체의 특징점을 직접 추출하는 것을 피하고 실루엣 영상으로부터 간단히 인체영역의 부분별 특징정보를 구하여 사용하였다.

본 논문의 구성은 다음과 같다. 2절은 입력 영상에서 배경과 신체 영역을 분리하기 위한 전처리 과정을 설명하고 3절에서는 분리된 신체 영역으로부터 부분특징정보와 전역특징정보를 추출하고 시간에 따라 특징을 집단화(grouping)하여 동작의 모션 정보를 얻는 방법을 설명한다. 4절에서는 고차원의 2차원 특징 벡터의 차원을 줄일 수 있는 주성분 분석법에 대해서 기술하고 5절에서는 은닉 마르코프 모델을 이용하여 학습, 인식하는 방법을 설명한다. 마지막으로 6절의 결론에서는 제안하는 알고리즘을 이용하여 얻은 실험 결과와 문제점 분석 및 이후의 연구방향 등에 대해 기술한다.

2. 전처리

2.1 분할(Segmentation)

카메라를 통하여 얻은 영상 시퀀스는, 일반 환경에서 취득한 것으로 영상에는 제스처 인식에 필요 없는 많은 오브젝트들(배경)이 포함되어 있다. 그런데, 제스처 인식에 필요한 것은 신체 영역(전경)이므로 우선 배경과 신체 영역을 분리하는 작업이 필요하고 이를 위해서는 먼저 배경 모델을 생성해야 한다. 그러나 조명의 밝기가 일정하지 않고 수시로 변하기 때문에 같은 카메라로 일정 시간 동안 똑같은 배경을 촬영한다고 할지라도, 모두 동일하지 않아 안정적인 배경 모델을 얻는데 어려움이 따른다.

본 논문에서는 조명 변화로 인한 배경의 밝기 변화를 측정하기 위해 시간 요소(t)를 고려해서 일정 시간 T_1 동안 배경 영상 I_t을 취득한 다음, 영상 영역 R 내에 있는 각 픽셀(x)들의 밝기 값 $I(x)$들을 분석하여 조명이 가장 밝았을 때의 화소값 $M(x)$와 가장 어두울 때의 화

소값 $N(x)$을 얻는다. 결국, 이 두 화소값의 차이 $D(x)$는 조명의 변화로 나타날 수 있는 밝기의 임계치로, 이 3가지 요소를 이용해 배경 모델(Background Model : BM)을 구성한다. 이와 같은 내용을 수식으로 표현하면 식 (1)과 같다[5].

$$BM = \{M(x), N(x), D(x)\}_{x \in R} \quad (1)$$

$$M(x) = Max\{I_t(x), (1 \le t \le T_1)\}$$

$$N(x) = Min\{I_t(x), (1 \le t \le T_1)\}$$

$$D(x) = M(x) - N(x)$$

일단 배경 모델이 만들어지면, 이진 영상 $B(x)$는 식 (2)에서 보여주는 것처럼 입력 영상 $I(x)$와 가장 밝은 화소값 $M(x)$와 가장 어두운 화소값 $N(x)$의 차분 연산을 통해 얻은 차이 값이 임계치 $D(x)$ 보다 크면 255의 화소값을, 그 외에는 0의 화소값을 갖는다.

$$B(x) = \begin{cases} 255 & if\{(|M(x) - I(x)| or |N(x) - I(x)|\} > D(x) \\ 0 & otherwise \end{cases} \quad (2)$$

식 (2)는 조명으로 인해 생길 수 있는 밝기 차이는 무시하고 신체의 움직임으로 차이를 갖는 영역만 분리하는 기준이 된다. 그러나 이 방법 역시 조명으로 인해 그림자가 생기는 경우, 그림자도 전경 영역으로 분리되어 정확한 신체 영역을 얻을 수 없다는 단점을 가지고 있다.

2.2 잡음 제거

식 (2)의 결과로 얻은 이진 영상 $B(x)$에는 인간이 움직이는 동안에, 배경 모델에서 설정한 밝기 값의 임계치 $D(x)$를 벗어나는 갑작스런 조명의 변화로 인해 배경임에도 불구하고 전경 영역으로 분리되어 1픽셀의 작은 점들이 포함될 수 있다. 따라서 이 잡음을 제거하기 위해서는 한번의 침식(erosion) 모폴로지(Morpology) 연산을 수행하고 이 때 신체영역도 같이 줄어드는 현상을 막기 위해 팽창(dilation) 모폴로지(Morpology) 연산을 사용한다[16].

3. 영상 군집화(Grouping)를 통한 모션 히스토리 생성

신체 모션은 신체 형상 전체가 갑작스럽게 변하는 전역 모션과 머리, 손, 발과 같이 주된 신체 부위의 일부분만 변하는 부분 모션으로 구분할 수 있다. 이와 같은 사실을 실험을 통해 분석하고 증명할 수는 없지만, 모션을 신체 포즈의 변화라고 가정했을 때 인간의 다양한 동작들을 관찰함으로써 쉽게 이해할 수 있다. 예를 들어, 사람이 앉는 동작을 취했을 때 이 동작은 서 있는 포즈에서 앉은 포즈로 변하는 것으로, 서 있는 포즈는 수직으로 긴 직사각형 형태를 갖고 앉은 포즈는 거의 정사각형의 형태를 갖게 된다. 즉, 앉는 동작은 두 팔의 움직임을 고려하지 않았을 때, 직사각형의 형태에서 정사각형의 형태로의 포즈 변화로 해석 할 수 있다. 그러나, 서서 손을 흔드는 동작과 같은 경우에는 서 있는 동작과 비교했을 때 손의 위치와 움직임이 매우 중요한 의미를 갖게 된다.

그런데, 전역 모션과 부분 모션은 서로 영향을 미치기 때문에 따로 분리해서 생각한다는 것은 매우 어려운 일이다. 따라서, 이 두 가지 모션 정보를 결합하여 모션 히스토리 정보를 얻고 이를 인식에 사용함으로써 보다 안정적인 인터페이스 구현이 가능하다.

3.1 모션 특징 값 추출

시각적인 방법으로 얻은 영상으로부터 특징 값을 추출하여 신체의 포즈와 모션을 효과적으로 표현하고 인식하는 연구는 전역 모션 정보로부터 특징을 추출하는 방법과 머리, 손, 발과 같이 신체의 특정 부위에 의미가 있는 부분 모션 정보로부터 특징을 추출하는 방법으로 나눌 수 있다.

전역 특징 정보를 이용하는 방법으로, MITLab.에서 사용했던 MHI(Motion History Image)와 MEI(Motion Energy Image)가 있는데 이는 휴(Hu)모멘트 벡터를 특징정보로 사용한다[10]. 이 방법은 모션정보가 누적되기 때문에 우발적인 물체의 모션이 있을 경우 오 인식될 수 있고 완전한 모션 패턴이 항상 주어져야 한다는 제약이 뒤따른다. 따라서 이러한 단점을 보완하고자 MHI 영상으로부터 신체 영역의 실루엣 윤곽선을 구하고, 모션에 의해서 바뀌는 신체의 윤곽선들로부터 방향 정보를 얻어 사용하고 있다. 그러나 이러한 방향 정보만으로는 움직이고 있는 신체 부위를 구별하는데 한계가 있다.

전역 특징 정보를 사용하는 방법의 또 다른 예로 로스 커틀러(Ross Cutler)의 방법이 있다[12]. 이 방법은 광류(optical flow)를 이용해 모션 블랍(blob)들을 분할(segmentation)하고, 블랍의 수와 블랍의 상대적인 크기 및 움직임의 방향을 이용하여 규칙기반(Rule-based) 기법으로 제스처를 인식하는 것이다. 그런데 이 방법은 블랍의 수와 움직임의 방향이 제스처를 인식하는데 큰 기여를 하기 때문에 모션 블랍들이 서로 겹쳐지는 경우 잘못된 인식 결과를 초래할 수 있다.

이스마엘(Ismail Haritaoglu)은 카드보드 모델(card-board model)을 이용하여 신체 영역을 6가지로 분리하고 그 영역들을 추적한다[5]. 그러나 카드보드 모델은 서있는 사람에 국한되어 적용할 수 있으므로 다양한 포즈에 대해서 적용하기 어렵다. 고스트(ghost) 시스템에서는 다양한 포즈를 갖는 신체 부위를 레이블링

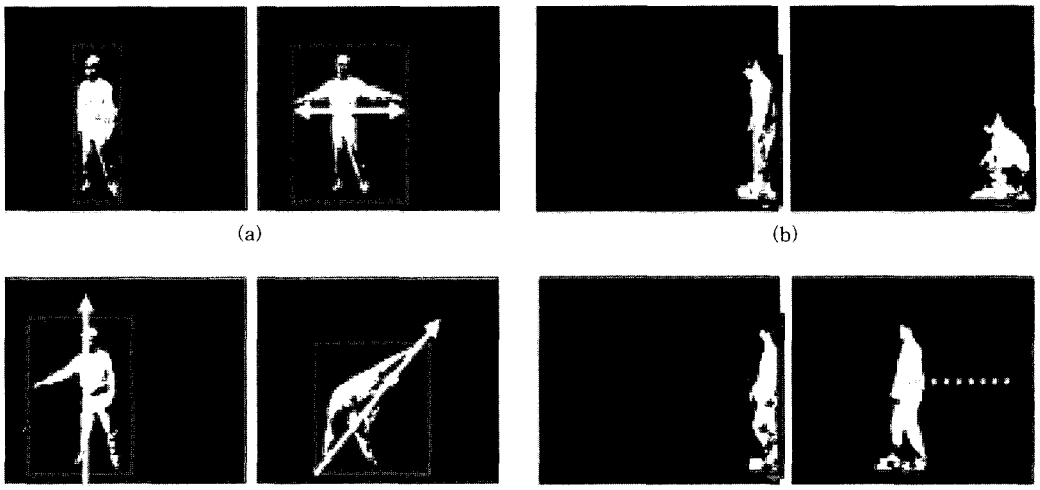

그림 1 동작에 따라 의미있게 변하는 특징 정보. (a)양팔을 벌리는 동작: 신체 영역의 가로축 길이(width), (b)서 있다 앉는 동작: 신체 영역의 세로축 길이(height)와 무게 중심의 y 좌표, (c)옆구리 운동: 모멘트의 주축, (d)걷는 동작: 무게 중심의 x 좌표가 많은 변화를 보이는 것을 확인할 수 있다.

(labeling)하기 위해 볼록집합(convex hull)과 위상학적인 형태를 분석한다[6]. 실루엣을 검출한 후 이를 이용해 수평 방향과 수직 방향으로 투영된 히스토그램의 유사성을 추정하고 특징 점들에 대해 반복적인 볼록집합 알고리즘을 적용한다. 그리고 볼록(convex) 점들은 위상학적인 해석을 통해서 구체적인 신체 부위로 구분 된다. 이 방법은 제스처 혹은 행동을 인식한다는 것보다는 포즈를 인식하는 기법으로, 전역 특징 정보와 부분 특징 정보를 조합함으로써 신체 부위를 해석하는 계층적 접근법이라는 점에서 중요한 의미를 갖는다.

본 논문에서는, 먼저 전체적인 신체 모션을 형상화하기 위해 6가지 전역 특징 정보; 1)신체 영역의 가로축 길이(width), 2)세로축 길이(height), 3)무게 중심의 x좌표, 4)무게 중심의 y좌표, 5)조밀성(compactness), 6)모멘트의 주축을 추출한 후, 이들 특징 값의 시간적인 변화량을 계산한다. 이 6가지 특징 정보는 동작의 변화로 인해 신체의 형상 변화가 생겼을 때 의미 있게 변하는 특징 값들을 관찰하여 채택한 것들이다. 예를 들어, 걷는 동작의 경우 의미 있게 변하는 특징 값은 무게 중심의 x 좌표이고 서있다 앉는 동작은 신체 영역의 세로축 길이와 무게 중심의 y 좌표가 많은 변화를 보였다. 그리고 손과 발을 벌리는 동작은 신체 영역의 가로축 길이와 조밀성(compactness), 몸을 한쪽으로 기울이는 동작은 모멘트의 주축 값이 의미 있게 변함을 확인할 수 있다.

이들 특징 값들의 변화량은 전체적인 외형의 변화만을 설명해 줄뿐, 신체의 어느 부위가 움직이고 있는지는 보여주지 않는다. 그러나, 제스처는 때때로 신체 특정 부위의 모션의 변화에 따라 많은 차이를 보일 수 있다. 그래서, 우리는 정확한 제스처 인식을 위해 신체의 어느 부위가 움직이고 있는지 알 필요가 있고 이를 위해서는 부분 모션 정보가 필요하다. 이를 위해 손이나 발의 정확한 위치를 계측하려면, 너무도 많은 계산 량이 필요하다. 따라서 우리는 매우 간단한 특징 데이타(신체의 부분 영역의 무게 중심 좌표, 부분 영역의 면적)를 도입하여 이 문제를 해결하고자 한다.

신체 영역은 전역 특징으로 구한 무게중심을 기준으로 4개의 부분 영역으로 나눌 수 있고 각각의 영역에 속하는 블랍들의 중심 좌표와 면적의 변화량을 살펴봄으로써 부분 모션 정보를 얻을 수 있다. 그림 2는 지금까지 설명한 초기 과정의 간단한 예를 보여준다.

3.2 영상 군집화

행동이나 제스처는 연속적인 신체 부위 또는 전체적인 신체의 움직임으로 이루어지기 때문에 제스처 인식에서 고려 해야 하는 중요한 사항은 모션 히스토리 정보(특징 값들의 시간적인 변화량)를 구하여 이를 인식에 이용하는 것이다. 본 논문에서는, 3.1절에서 구한 특징 값들의 히스토리 정보를 구하기 위해 연속하는 3개의 영상을 하나의 그룹으로 간주하고 이들 특징 값들의 시간적인 변화량을 계산하는 영상 그룹핑 방법을 사용한다.

$$F(t) = F_g(t) + F_p(t) \quad (1 \le t \le T) \tag{3}$$

$$F'(y) = \{F(t-2) - F(t), F(t-1) - F(t), F(t)\}$$

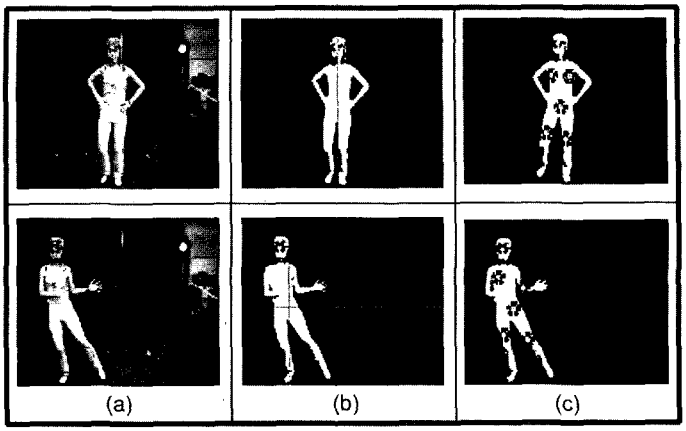

그림 2 입력 영상(a)이 주어지면, 배경으로부터 신체 영역을 분리한 후 전역 특징 정보의 하나인 전체 무게 중심을 기준으로 4개의 부분 영역으로 분리한다. (b)에서는 두 선이 만나는 지점이 바로 전체 신체 영역의 무게 중심의 위치이다. (c)는 각 부분영역의 블랍(blob)들의 무게 중심을 원으로 나타낸 것으로 이는 부분 특징 정보로서 대략적인 신체 부위의 위치를 나타낸다.

$$(3 \le t \le T, 1 \le y \le G) \tag{4}$$

식 (3)에서 F(t)는 T를 영상 시퀀스의 총 길이라고 했을 때 시간 t에서의 특징 집합을 나타내는 것으로, 신체 전체의 포즈 정보를 담고 있는 전역 특징 벡터 $F_g(t)$와 신체의 특정 부위의 모션 특징을 나타내는 부분 특징 벡터 $F_p(t)$의 합으로 표현할 수 있다. 따라서 $F(t)$는 18개의 특징 값(6개의 전역 특징 + 12개의 부분 특징)으로 구성 되고 시간 t에서의 영상 I_t는 이웃하는 영상 I_{t-1}, I_{t-2}와 함께 하나의 그룹으로 묶인다. 따라서 식 (4)에서 보여주는 것처럼 $G(=T-2)$를 전체 영상 시퀀스에서 얻을 수 있는 그룹의 수라고 할 때, y번째 영상 그룹의 특징 벡터 $F'(y)$는 식 (3)을 이용해 구한 특징 벡터 $F(t)$뿐만 아니라 $F(t-1)$, $F(t-2)$와의 차분을 통해 얻은 특징 값들의 변화양의 합으로 구성 된다. 결국, 한 그룹의 전체 특징 정보의 개수는 54가 되고 이와 같이 시간적 영상 군집화 알고리즘은 그림 3에서와 같이 도식적으로 보여준다.

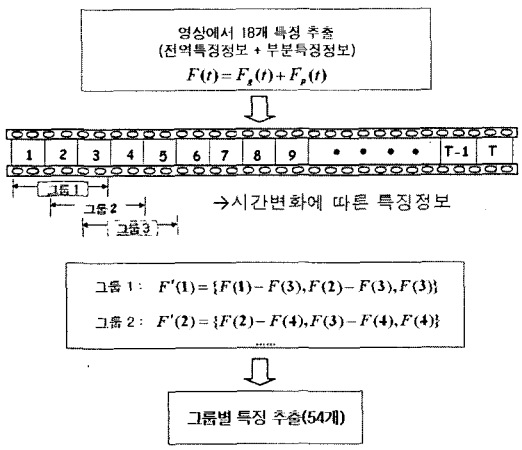

그림 3 특징 추출과 시간적 영상 군집화(grouping) 알고리즘

4. 주성분 분석법과 제스처 공간

주성분 분석법(Principal Component Analysis)은 고차원의 입력 데이타 집합을 저 차원의 의미 있는 데이타 집합으로 줄일 수 있다[13,14]. 제스처 영상 데이타의 경우 하나의 동작을 구성하는 프레임(frame)의 수가 많고 특징을 추출하기 비교적 어렵기 때문에 빠른 인식속도와 효과적인 특징 추출이 가능한 방법을 적용해야 한다. 따라서 4장에서는 3장에서 추출한 연속적인 동작의 선형적 특징을 이용하여 저 차원 벡터로 표현하는 방법에 대해 기술 한다.

3절에서 구한 특징 벡터 x는 식 (5)와 같이 표현될 수 있고 이 벡터의 고유공간을 계산하기 위해서는 먼저 모든 특징 벡터의 평균 벡터를 구하여 각 특징 벡터와의 차를 구한다[10]. 평균 벡터 c와 새로운 특징 집합 X는 식 (6)과 식 (7)과 같다. 그런 다음, 식 (8)을 만족하는 고유벡터를 구하기 위해 공 분산 행렬 Q에 대한 고유치 λ와 고유벡터 e를 구한다.

$$x = [x_1, x_2, \cdots, x_N]^T \tag{5}$$

$$c = (1/N)\sum_{i=1}^{N} x_i \tag{6}$$

$$X \triangleq [x_1 - c, x_2 - c, ..., x_N - c]^T \quad (7)$$

$$Q \triangleq X \cdot X^T \quad (8)$$

$$\lambda_i \cdot e_i = Q \cdot e_i \quad (9)$$

이 때, 고유치 분해를 하지 않고 특이치 분해(Singular Value Decomposition)를 이용함으로써 특징 집합 X의 공 분산 행렬에 대한 고유벡터를 쉽게 얻을 수 있다. 이렇게 얻어진 고유공간에 평균 벡터 c에서 뺀 특징 집합 X를 모두 식 (10)를 이용하여 투영시킨다[14].

$$m_i = [e_1, e_2, ..., e_k]^T (x_i - c) \quad (10)$$

이와 같이 얻어진 저 차원 벡터 공간, 즉 파라메트릭 고유공간을 제스처 공간이라 부른다[15].

이미 설명한 바와 같이 주성분 분석은 몇 개의 주성분 벡터를 유도하여 이를 통해 차원의 축소와 자료의 요약을 주목적으로 하고 있다. 따라서 전체 변이의 대부분을 적절히 설명하기 위하여 보유해야 할 주성분의 수를 결정해야 한다. λ_i는 i번째 고유값, p는 전체 고유값 개수라고 할 때 전체 분산 중 주성분 C_i가 설명할 수 있는 비율은 λ_i/p이다[14]. 우리는 처음 k개 주성분들이 설명할 수 있는 누적비율이 70% 이상일 때의 개수를 선택한다. 이를 수식으로 표현하면 다음과 같다.

$$(\sum_{i=1}^{k} \lambda_i / p) \times 100 \geq 70 \quad (11)$$

그림 4는 실험에 사용한 제스처들의 특징 집합으로부터 구한 고유치 개수에 따른 주성분의 누적 기여도를 보여주고 그림 5는 걷는 동작과 손 흔드는 동작이 포함된 영상 시퀀스를 제스처 공간에 투영한 결과를 나타낸 것이다.

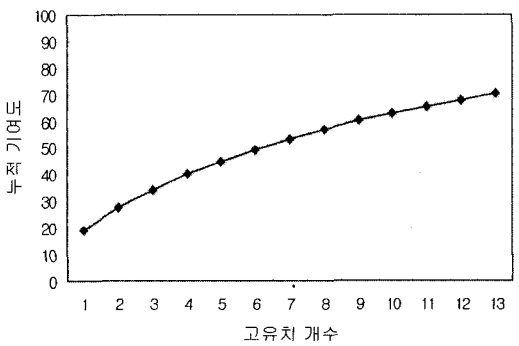

그림 4 실험에 사용한 제스처들의 특징 집합으로부터 구한 고유치 개수에 따른 주성분의 누적 기여도

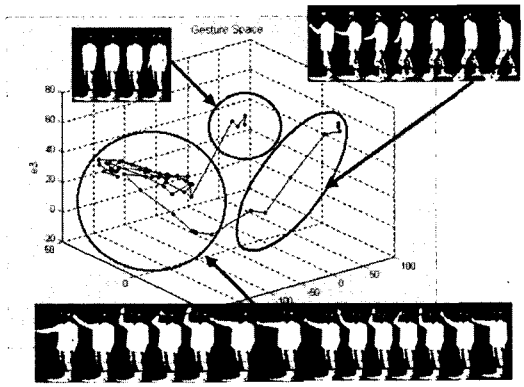

그림 5 걷기와 손 흔들기 동작이 포함된 영상 시퀀스를 제스처 공간으로 투영한 결과. 원으로 표시한 부분은 서로 비슷한 포즈를 가지는 영상들로 클러스터링 알고리즘으로 분리했을 때 같은 클러스터에 속하게 된다.

5. 은닉 마르코프 모델을 이용한 제스처 인식

파라메트릭 제스처 공간에 투영된 점(영상)들은 신체 포즈가 시간에 따라 선형적으로 변화한다는 것을 나타낸다. 이처럼 시간적으로 변화하는 데이타를 제스처 인식모델로 구성하기 위해 은닉 마르코프 모델(Hidden Markov Models)을 이용하였다[16,17].

먼저, 은닉 마르코프 모델의 입력은 심볼(symbol)의 집합이 된다. 그러므로 우리가 4장에서 구한 저 차원의 특징 데이타 값들을 심볼로 바꾸기 위해서 클러스터링(clustering) 알고리즘[18]을 이용해 몇 개의 제스처 그룹(Cluster)으로 나누고, 각 그룹에 대해 특정 심볼(숫자)을 할당한다. 그리고, 각 클러스터의 중심 좌표 값은 코드 북으로 저장되어 새로운 특징 값이 들어왔을 때 심볼을 할당하는 기준이 된다.

제스처들의 심볼 집합이 입력으로 들어오면 봄 웰치(Baum-Welch) 알고리즘에 의해 인식에 필요한 3가지 파라메터($\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}$)을 식 (12), 식 (13), 식 (14)로 추정할 수 있다. 상태 천이 확률 a_{ij}는 은닉 마르코프 모델의 상태가 i로부터 j로 변화하는 확률을 의미한다. 그리고 확률 b_{ij}는 출력 심볼 y가 상태 i로부터 j로 천이되면서 관측될 수 있는 확률, π_i는 초기 상태 확률 값을 나타낸다.

$$\overline{\pi_i} = \gamma_1(i) \quad (12)$$

$$\overline{\alpha_{ij}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (13)$$

$$\overline{b_{ij}}(k) = \frac{\sum_{t=1, s.t. y_t=k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \quad (14)$$

학습 과정을 통해 각 제스처에 해당되는 3가지 파라메터(π, A, B) 값들이 결정되는데 이를 제스처 모델이라 한다[11,17]. 인식하고자 하는 새로운 심볼 집합(Y)이 주어지면 은닉 마르코프 모델에서는 각 제스처 모델 λ_i에 대한 확률 값을 계산하고, 가장 높은 확률 값을 갖는 모델의 제스처로 인식하게 된다. 제스처 모델 λ_i에 대한 확률 값은 전방(forward) 변수인 $\alpha_t(i)$와 후방(backward) 변수인 $\beta_t(i)$를 이용하여 식 (15)와 같이 계산된다[17].

$$P(Y \mid \lambda_i) = \sum_i \sum_j \alpha_t(i) a_{ij} b_{ij}(y_{t+1}) \beta_{t+1}(j) \quad (15)$$

6. 실험 및 결론

사람의 정면에 위치한 비디오 카메라를 통하여 들어오는 256 계조의 흑백 영상을 입력 영상으로 사용하였다. 학습에 사용된 제스처는 그림 6에서 보여주는 것처럼 양팔 운동, 등배 운동, 팔 굽혀 엎드리기, 쪼그려 걷기, 한 손으로 흔들기, 다리 운동, 의자에 앉기, 걷기로

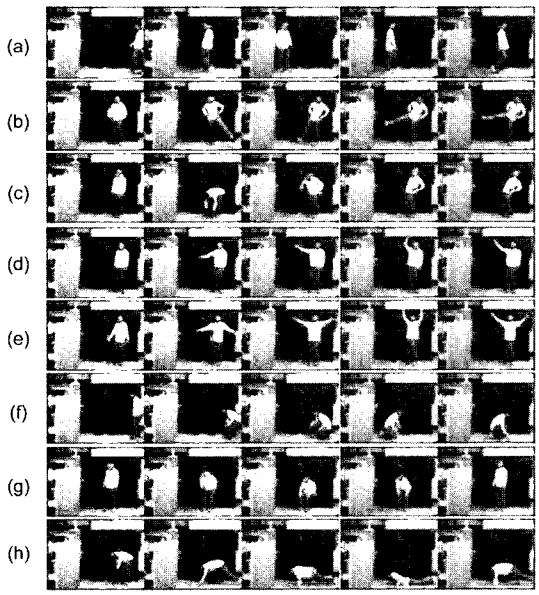

그림 6 실험에 사용된 영상시퀀스들. (a)걷기 (b)다리운동 (c)등배운동 (d)한 손으로 흔들기 (e)양팔운동 (f)쪼그려 걷기 (g)의자에 앉기 (h)팔굽혀 엎드리기

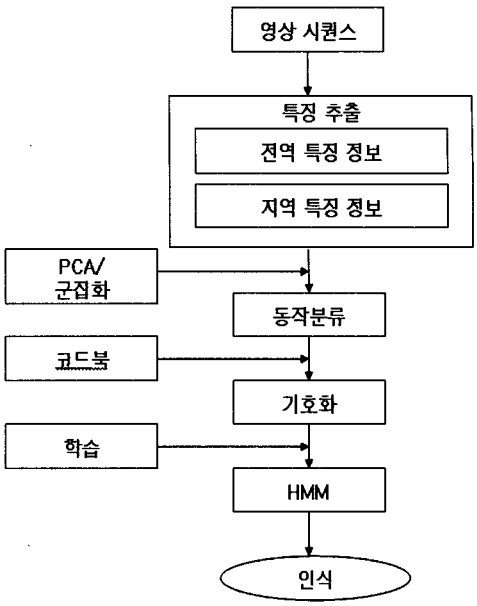

그림 7 전체 시스템 구성도

총 8가지이다. 영상의 크기는 320×240이고 각각의 제스처는 100프레임으로 구성되어있다. 따라서 한 사람 당 총 800프레임(8×100)의 영상이 사용하였고, 실험에 참가한 사람은 10명이므로 총 8000 프레임을 이용하여 모델 공간을 구성하였다. 인식을 위한 전체 시스템 구성 도는 그림 7에 나타내었다.

표 1은 전체적인 신체의 형상을 표현하는 전역특징정보만을 추출한 경우(방법 1)와 전역특징정보뿐만 아니라 신체의 어느 부위가 움직이는지를 나타내는 부분특징정보를 함께 추출한 경우(방법 2-본 논문에서 제안한 방법)에 대해 각각 동작별 인식률을 계산한 결과이다. 표 1에서 보는 바와 같이 전역특징정보만을 사용한 경우보다 부분특징정보를 함께 추출한 방법이 더 나은 인식

표 1 특징 정보별/동작별 인식률

구분	인식률	
동작	전역특징정보만을 사용 (방법 1)	전역특징정보 + 부분특징정보 (방법 2)
걷기	100%	100%
다리운동	50%	90%
등배운동	60%	90%
손흔들기	50%	70%
팔운동	50%	100%
기어가기	100%	100%
앉기	40%	90%
눕기	100%	100%
평균인식률	68.75%	92.5%

결과를 보임을 확인할 수 있었다. 이는 제스처가 신체 특정 부위의 모션의 변화에 따라 많은 차이를 보이기 때문이다.

방법 2의 경우, 모델을 구성했던 영상 시퀀스들과 모델과 동일한 속도로 동일한 동작을 취한 영상 시퀀스들에 대해서는 거의 대부분 올바르게 인식됨을 알 수 있었고 모델과 좀 다른 동작을 취한 영상 시퀀스들에 대해서는 일부(손 흔들기)가 다른 시퀀스(양팔 운동)로 인식하는 오류를 범하기도 하였다. 이는 본 논문에서 제안한 방법(방법 2)이 손이나 발의 정확한 위치를 계측하여 사용하지 않고 대략적인 위치 정보를 이용하기 때문에 모델로 구성된 동작과 다른 경우, 오 동작으로 인식된 것으로 간주된다.

우리가 실험에 사용한 데이타만 가지고는 입력되는 모든 영상을 분류해 내기란 무척 어려운 일이다. 따라서 우리가 사용한 방법이 보다 일반성을 갖기 위해 가능한 모든 영상을 수집하여 분석해 나가야 할 것이다. 그런데 우리가 취하고 있는 방법은 에지나 코너와 같은 기하학적인 특징을 이용하는 것이 아니고 면적, 영역의 가로 세로 비, 이동 량 등의 매우 추상적인 수치적 양을 이용하고 있다. 따라서 인간의 행동 전부를 인식할 수 없지만 예를 들어 공항 대합실에서의 행동의 분석, 또 학교 교실 내에서의 학생들이 행동 분석, 운동장에서의 행동 분석, 스포츠 분야에서의 행동 분석, 슈퍼마켓에서의 행동 분석 등 특정 좁은 분야에서 사용이 가능하다.

본 논문에서는 신체의 어느 부위가 움직이는지를 나타내는 부분(partial) 특징 정보와 전체적인 신체의 형상을 표현하는 전역(global) 특징 정보를 이용하여 모션 히스토리 정보를 얻고, 이를 이용하여 주성분 분석법과 은닉 마르코프 모델에 의해 제스처를 인식하는 알고리즘에 대해 기술하였다. 이 방법의 특징은 손이나 발의 정확한 위치를 계측하여 특징 정보를 얻는 것이 아니고, 영상에서 쉽게 계산이 가능한 특징 값들을 이용하여 구한 모션 히스토리 정보를 인식과정에 사용한다는 점이다. 따라서 많은 계산 량이 요구되지 않기 때문에 실 세계에서 구현이 용이하고 실시간 시스템 구축에 매우 적합하다. 그러나 영상 안에 여러 사람이 존재하는 경우와 사람이 물건에 가려진 경우에는 인식이 불가능하다는 문제점을 지니고 있다. 따라서, 앞으로는 이런 부분을 보완하기 위해 움직이는 사람들을 추적할 수 있고 여러 사람이 겹쳐지더라도 적절히 분할 할 수 있는 방법에 대해 연구해볼 계획이다.

참 고 논 문

[1] M. Weiser, "The Computer for the 21st Century," Scientific Anerica, Vol.265, No.3, pp.66-76, Sept. 1991.

[2] Vladimir I. Pavlovic, Rajeev Sharma, and Thomas S. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review," IEEE Transaction on PAMI, Vol.19, No.7, p.677, July 1997.

[3] 이인호, 박찬종, "모션캡쳐 기술의 현황과 응용 분야," 멀티미디어학회지, 3권, 1호, pp.38-48, 1999.

[4] D.M. Gavrila, L.S. Davis, "Towards 3D model-based tracking and recognition of human movement: a multi-view approach," Int. Workshop on Face and Gesture Recognition, Vol.0000162479, pp.272-277, 1995.

[5] Ismail Haritaoglu, David Harwood and Larry S. Davis, "W4: Who? When? Where? What? A Real-time System for Detecting and Tracking People," Third Face and Gesture Recognition Conference, pp.222-227, 1998.

[6] Ismail Haritaoglu, David Harwood and Larry S. Davis, "Ghost: A Human Body Part Labeling System Using Shilhouettes," International Conference on Pattern Recognition, 1998.

[7] J. Sherrah and S. Gong, "VIGOUR: A system for tracking and recognition of multiple people and their activities," Proc. ICPR, Barcelona Spain, Vol.1, pp.179-182, 2000.

[8] Vladimir I. Pavlovic, Rajeev Sharma, and Thomas S. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review," IEEE Transaction on PAMI, Vol.19, No.7, p.677, July, 1997.

[9] Takahiro Watanabe and Masahiko Yachida, "Real Time Recognition of Gesture and Gesture Degree Information Using Multi Input Image Sequence," ICPR, Vol.1, p.1855, 1998.

[10] James W. Davis, Aaron F. Bobic, "The Representation and Recognition of Action using Temporal Templates," MIT Media Lab Technical Report 402, 1997.

[11] Rafael C.Gonzalez, Richard E.Woods, "Digital Image Processing 2/E," Prentice Hall, pp.519-532, 2001.

[12] 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, pp.416-421, 1998.

[13] Turk. Matthew and Alex Pentland, "Eigenfaces for Recognition," Journal of Cognitive Neuroscience, Vol.3, pp.71-86, 1991.

[14] Hiroshi Murase and Shree K. Nayar, "Visual Learning and Recognition 3-D object from appearance," International Journal of Computer Vision, Vol.14, 1995.

[15] Shigeyoshi Hiratsuka, Kohtaro Ohba, Hikaru Inooka, Shinya Kajikawa, and Kazuo Tanie, "Sta-

ble Gesture Verification in Eigen Space," LAPR Workshop on Machine Vision Application, Vol.0000141262, pp.119-122, 1998.
[16] Yoshio IWAI, Tadashi HATA and Masahiko YACHIDA, "Gesture Recognition based on Subspace Method and Hidden Markov Model," Proc. of Intl. Conf. Intelligent Robots and Systems (IROS'97), Vol.2, pp.960-966, Grenoble, France, Sep. 1997.
[17] J.Yamato, J.Ohya, and K.Ishii, "Recognizing human action in time-sequential images using hidden markov models," ICCV, pp.379-385, 1992.
[18] 김기영, 전명석, "다변량 통계 자료 분석", 자유 아카데미.

이 용 재

1998년 호원대학교 전자계산학과 학사 2000년 전남대학교 컴퓨터공학과 석사 2002년 전남대학교 컴퓨터공학과 박사수료. 2004년 1월~9월 ㈜어플라이드비전테크 연구원. 2004년 10월~현재 ㈜삼성테크윈 정밀기기연구소 선임연구원. 관심분야는 컴퓨터비전, 제스처인식, 컴퓨터그래픽스

이 칠 우

1986년 중앙대학교 전자공학과 학사 1988년 중앙대학교 대학원 전자공학과 공학 석사. 1992년 동경대학 대학원 전자공학과 공학 박사. 1992년~1995년 이미지 정보과학 연구소 수석 연구원 겸 오사까대학 기초공학부 협력연구원. 1995년 리츠메이칸대학 특별초빙강사. 1996년~현재 전남대학교 공과대학 컴퓨터공학과 교수. 관심분야는 컴퓨터비전, 멀티미디어 데이타베이스, 컴퓨터그래픽스