Journal of Korea Multimedia Society Vol. 7, No. 8, August 2004(pp. 1035-1045)

# 파라메트릭 제스처 공간에서 포즈의 외관 정보를 이용한 제스처 인식과 동작 평가

이칠우[†], 이용재[††]


## 요 약

본 논문에서는 저차원 제스처 특징 공간에서 연속적인 인간의 제스처 형상을 이용하여 제스처를 인식하고 동작을 구체적으로 평가하는 방법에 대해 소개한다. 기존의 HMM, 뉴럴 넷을 이용한 제스처 인식방법은 주로 인간의 동작 패턴을 구분할 수 있지만 동작의 크기 정보를 이용하기엔 어려움이 있다. 여기서 제안한 방법은 연속적으로 촬영된 인간의 제스처 영상들을 파라메트릭 고유공간이라는 저차원 공간으로 표현하여 모델과 입력 영상간의 거리 계산으로써 포즈뿐만 아니라 동작에 관한 빠르기나 크기와 같은 구체적인 정보를 인식할 수 있다. 이 방법은 단순한 처리와 비교적 안정적인 인식 알고리즘으로 지적 인터페이스 시스템이나 감시 장비와 같은 여러 응용 시스템에 적용 될 수 있다.


# Gesture Recognition and Motion Evaluation Using Appearance Information of Pose in Parametric Gesture Space

Lee Chil-Woo[†], Lee Yong-Jae[††]


## ABSTRACT

In this paper, we describe a method that can recognize gestures and evaluate the degree of the gestures from sequential gesture images by using Gesture Feature Space. The previous popular methods based on HMM and neural network have difficulties in recognizing the degree of gesture even though it can classify gesture into some kinds. However, our proposed method can recognize not only posture but also the degree information of the gestures, such as speed and magnitude by calculating distance among the position vectors substituting input and model images in parametric eigenspace. This method which can be applied in various applications such as intelligent interface systems and surveillance systems is a simple and robust recognition algorithm.


**Key words:** Gesture Recognition(제스처 인식), PCA(주성분 분석), Appearance-Based Recognition(외관 기반 인식)

※ 교신저자(Corresponding Author) : 이용재, 주소 : 광주시 북구 용봉동 300번지(500-757), 전화 : 062) 530-0258, FAX : 062)530-1809, E-mail : myfaith72@daum.net
접수일 : 2003년 9월 22일, 완료일 : 2004년 1월 26일
[†] 종신회원, 전남대학교 정보통신공학부 부교수
(E-mail : lcecw@chonnam.ac.kr)
[††] 준회원, (주)어플라이드비전텍 연구/개발부 근무
※ 본 연구는 한국 과학 재단 지정 전남대학교 "고품질 전기전자 부품 및 시스템 연구센터"의 연구비 지원에 의해 수행되었음.

## 1. 서 론

컴퓨터 기술의 발달과 함께 정보 시스템이 복잡하게 되면서 인간과 정보 시스템 사이에 자연스럽게 정보를 교환할 수 있는 지적 인터페이스에 관한 관심이 날로 커지고 있다. 인간은 일상생활에서 몸짓, 손짓, 표정과 같은 비언어적 수단을 이용하여 수많은 정보를 전달하기 때문에 이러한 비언어적 정보를 디

지털 정보로 변환하여 인터페이스 구현에 활용하면 사용자 친화적인 정보 시스템 구현이 가능하게 된다. 비언어적 대화 수단 중 인간이 자주 사용하는 제스처는 대화시의 정보 전달 뿐만 아니라 인간의 행동이 갖는 의미를 해석함에 있어 매우 중요한 단서를 제공하기 때문에 최근에 들어, 대규모 비디오 데이터베이스의 구축, 감시 시스템, 고압축 통신 시스템의 구현을 위한 연구가 활발히 진행되고 있다.

제스처를 인식한다는 것은 인체 각 부위가 시간축에 대해 어떠한 형상 변화를 가지는가를 자동으로 알아내는 것을 의미한다. 그러나 인체는 매우 복잡한 3차원 관절 구조를 가지고 있을 뿐 아니라 신체부위에 따라 각기 다른 의미를 표현 할 수 있기 때문에 자동으로 제스처를 인식하는 것은 매우 어렵다. 정확한 제스처 인식을 위해서는 인간의 손, 팔, 심지어 인간 신체의 동적이고 정적인 형상 변화를 기계에 의해서 측정할 필요가 있다. 또한 인간의 동작은 그 빠르기나 크기에 따라 같은 동작이라 할지라도 다르게 해석 될 수 있기 때문에 시간변화에 따른 형상의 변화량 또한 중요한 인식 요인으로 고려되어야 한다.

제스처 인식의 자동화에 관한 연구는 크게 물리적 장치를 이용하는 방법과 비디오카메라를 통해 얻어진 영상을 이용하는 방법으로 나누어 질수 있다. 물리적 장치를 이용한 방법 중 가장 대표적인 것은 데이터 글러브나 마그네틱 센서를 이용하는 방법이다[3]. 이 방법들은 인체의 각 관절 각도와 공간적 위치를 장치를 통해 직접 측정하기 때문에 정확한 3차원 데이터를 복잡한 계산 없이 얻을 수 있어 최근 가상현실 시스템이나 영화의 특수 효과 부분에 가장 많이 쓰이고 있다. 그러나 이 방법은 장치를 통해 얻어진 데이터를 전달하기 위해 전체 시스템과 케이블로 연결되어야 하기 때문에 자연스러운 인터페이스 구축에 많은 문제를 가지고 있다. 비록 연결선이 필요 없는 무선 장치를 이용한다 할지라도 사용자가 그런 특별한 장치들을 몸에 부착한다는 것은 부자연스러울 뿐만 아니라 응용 분야에 따라서는 불가능할 경우도 있다.

인체에 별도의 장치를 부착하지 않고 신체의 형상을 인식하는 방법으로는 다중 카메라 기반의 3차원 데이터를 이용하는 방법과 2차원 영상 시퀀스를 이용하는 방법이 있다. 3차원 데이터를 이용한 대표적인 방법으로는 여러 대의 카메라를 통해 입력된 영상을 3차원 데이터로 복원하여 신체부위의 3차원 이동궤적을 이용하여 인식하는 방법이 있다. 하지만 이 방법은 복잡한 계산과 적당한 파라메타 추출의 어려움으로 인식 분야 보다는 3차원 영상 복원이나 애니메이션 분야에 주로 이용되고 있다[1,3].

2차원 영상 정보를 이용하는 방법으로는 에지나 특징 점과 같은 기하학적 특징 정보를 이용하는 방법과 형상 정보를 이용하는 방법이 있다[1,2]. 기하학적 특징 정보를 이용하는 방법으로는 신체 영상의 윤곽선, 코너, 모멘트와 같은 특징정보들을 형판이나 모델로 구성하여 입력 영상과의 단순 차이를 통해 인식하는 방법이 있다. 하지만 이러한 방법들은 여러 가지 복잡한 동작에 대한 특징 추출이 어렵고 많은 전처리 과정이 필요하게 되어 계산 속도와 인식률이 떨어지게 된다. 특히 인간의 동작은 일정한 패턴을 얻기 어렵기 때문에 단지 영상을 통하여 기하학적 특징을 인식에 이용하는 것은 많은 어려움이 있다[8,9].

최근엔 HMM이나 뉴럴 넷을 이용하여 제스처를 인식하려는 시도가 많이 이용되고 있다[10,14,15]. 이 방법들은 시간의 변화에 따른 동작의 특징을 심볼화하여 확률 모델을 구성한다. 입력동작을 인식하는 방법은 모델로 구성된 동작과 매칭 수행 시 가장 높은 확률을 가지는 동작을 해당 모델로 인식한다. 포즈가 아닌 동작을 인식한다는 의미에서 동작을 분류하거나 구분하는데 효과적이지만 자동적으로 최적의 확률 모델을 구성하기가 어렵고 구체적인 동작 정보 즉 빠르기나 크기를 인식하는 데는 많은 어려움이 있다.

특징 추출이 어려운 인식에는 외관 기반 인식 법들이 주로 사용되고 있다. 이런 외관 기반 인식의 특징은 영상 전체를 통계적 입력을 위한 일련의 데이터로 간주하여 얼굴 인식에 적용할 경우 표정, 시선방향의 변화에도 비교적 안정적인 인식 결과를 얻고 있다[11]. 이 방법은 특정한 영상의 분석과 특징 추출 없이 영상 자체의 휘도 치 정보 또는 신체 형상을 나타내는 영상 집합에 대해 통계적으로 높은 확률 값을 가지는 밝기 정보와 위치 정보를 주성분 분석법을 이용하여 저차원 공간으로 표현함으로써 그 공간에서 각 모델 심볼과 입력 심볼 간의 거리 정도를 비교하여 매칭을 시도하는 방법이다[4-7,16,20]. 복잡한 전 처리 없이 영상자체의 형상이나 음영정보를 이용하기 때문에 수행속도가 매우 빠르고 비교적 안정적인 인식결과를 얻을 수 있을 뿐만 아니라 실시간

처리에도 효과적이다. 하지만 인식대상 간의 가림 현상 즉 겹침이 일어나거나 조명의 변화가 일어날 경우 인식에 어려움이 따른다[17,18].

일반적으로 카메라를 통해서 얻어진 인간의 동작 영상을 인식 하는 데는 몇 가지 어려운 점이 있다.

첫째는 2차원 영상의 신체 형상은 서로 다른 동작이라 할지라도 유사한 모습으로 보일 수 있기 때문에 안정적인 모델 구성이 어렵다는 것이다. 특히 팔과 다리는 동작의 의미를 전달하는 가장 중요한 부위지만 가장 많은 자유도를 지닌 신체부위 이기 때문에 이를 구분하여 모델의 특징을 추출하기란 더욱 어려운 문제인 것이다. 이 때문에 신체 형태에 대한 폭, 높이, 방향, 원형도, 모멘트 등과 같은 기하학적인 특징을 이용하거나 얼굴을 기준으로 손과 다리의 위치를 추정하여 인식 하는 방법들이 제안되었다. 하지만 단지 기하학적 특징으로 동작의 의미를 추출하는 것은 더욱 애매한 결과를 나타낼 수 있으며 칼라 정보를 이용하여 얼굴, 손, 다리 영역을 안정적으로 분할하는 데는 많은 어려움이 있다.

둘째는 인간의 동작은 얼굴의 표정과 같이 감정을 표현하는 중요한 수단이기 때문에 동작의 빠르기나 크기 또한 인식의 대상이 되어야 한다는 점이다. 하지만 각각의 포즈 변위 값을 특징으로 하는 여러 인식 방법들은 단지 동작 전체에 대한 프레임 간 특징의 차이를 인식에 이용할 뿐 변위의 정도 즉 얼마나 변하고 있는가에 대한 정보를 얻기에는 한계를 가지고 있다. 이를 해결하기 위해서는 물리적 센서를 신체에 부착하여 그 움직임 정보를 측정하거나 다양한 빠르기의 동작을 모델로 구성해야만 한다.

본 논문에서 제안하는 방법은 카메라로부터 연속적으로 촬영된 인간의 제스처를 파라메트릭 고유공간이라는 저차원적 공간으로 표현하여 식별할 뿐만 아니라 동작에 관한 구체적인 정보를 얻을 수 있어 보다 효과적이다. 먼저 인식에 이용하는 연속 제스처 영상들을 분할하여 정규화하고 주성분 분석법(Principal Component Analysis)이라는 통계적인 방법을 이용하여 매칭에 기준이 되는 고유공간을 구성한다. 모든 영상들은 이 공간에서 제스처의 내용에 따라 위치가 변하는 하나의 점으로 표현되며 실제 모델영상과 입력 영상은 이 공간 내에서 서로의 거리를 측정함으로써 비교가 가능하다. 구체적 제스처 정보를 얻는 방법은 연속적으로 입력영상이 공간상에 투영될 때 투영된 점(벡터)들 사이의 거리를 계산함으로써 실제 제스처의 빠르기나 크기를 인식할 수 있게 된다[8,22]. 제안한 방법은 복잡한 전 처리를 통해 기하학적 특징을 추출하지 않고 신체 영상의 형태를 의미 있는 압축된 벡터형태로 표현하여 비교적 간단한 처리로써 안정적인 결과와 구체적 제스처 정보를 얻을 수 있으며 애매한 인식결과가 발생할 경우 그 결과를 학습을 통하여 모델로 재구성 할 수 있다는 장점이 있다.

최근 이러한 알고리즘을 이용하여 인간의 동작을 인식하는 시스템은 여러 응용분야에서 다양하게 이용 될 수 있다. 게임이나 오락분야 에서는 사용자들이 자신의 신체를 이용하여 가상의 기구나 장비들을 제어 할 수 있으며, 사용자의 동작에 따라 가상의 캐릭터를 조작하여 게임을 즐길 수 있다[12,21]. 그리고 가상의 스포츠 강사가 사용자들의 동작을 분석하고 점검하여 가장 적절한 동작으로 수정하고 학습시키는데 적용 할 수도 있다. 이런 응용으로 사용자들은 더욱 실감나게 게임이나 스포츠를 체험 할 수 있다. 또한 보안이나 감시를 요하는 장소에서 어떤 돌출 행동이나 위험한 동작들을 자동적으로 찾아내어 적절한 대응을 유도 할 수 있다.

이절 이후의 논문 내용은 다음과 같다. 2절에서는 제스처를 인식하기 위한 전 처리 과정인 영상의 정규화와 제스처 분할을 설명하고 정규화 된 영상으로 주성분 분석법을 이용하여 제스처 공간을 구성하는 방법과 모델과 실제 영상간의 상관관계에 대해 설명한다. 또한 공간상의 효과적 매칭을 위한 특징 모델 구성 방법에 대해 설명한다. 3절에서는 공간상에서 입력영상과 모델영상간의 매칭 방법과 입력 영상으로부터 구체적 동작 정보를 추정 할 수 있는 방법에 대해 설명한다. 4절에서는 실험을 통해 제안된 알고리즘의 타당성을 확인하고, 마지막으로 5절에서 결론을 정리하고 앞으로 개선될 연구 방향에 대해 기술한다.

## 2. 제스처 공간의 표현과 구성

### 2.1 영상집합의 획득과 정규화

인간의 동작, 즉 자세변환은 너무나 많은 경우가 존재하기 때문에 본 논문에서는 신체부위의 구분 동작을 기본으로 하는 맨손체조의 7가지 동작을 인식 대상으로 하였다. 모델 영상은 7가지 동작을 정상적

인 속도로 수행하면서 획득하였고 참조 영상은 같은 동작을 빠르기를 바꾸어 가며 입력하였다. 그림 1은 인식 모델로 사용된 7가지 동작의 일부를 보여 주고 있다. 그림 1에서 알 수 있는 바와 같이 신체 부위가 서로 겹치지 않고 가장 정확하게 인식 할 수 있도록 동작이 잘 표현되는 방향에서 촬영 하였다.

이러한 영상의 집합을 $\hat{x}$ 라하고 식(1)과 같이 표현한다.

$$\hat{x}=[\widehat{x_1}, \widehat{x_2}, \ldots, \widehat{x_N}]^T \quad \text{식(1)}$$

여기서 $T$는 전치 행렬을 의미하고 N은 제스처 영상의 개수를 나타내는 정수이다.

제스처 영상집합의 정규화는 카메라로부터 연속적으로 얻어진 영상들을 모델구성에 필요한 상태로 만들기 위한 전 처리 과정으로 시각 기반 제스처 인식에 필수적인 처리 과정이다. 특히 일반 환경에서 연속적 영상들은 각 각 서로 다른 많은 잡음들을 포함하기 때문에 정규화처리에 의해 인식률이 크게 좌우 될 수 있다[4,11].

정규화 과정은 크기 정규화와 밝기 정규화 그리고 위치 정규화 과정으로 나누어진다. 크기와 밝기 정규화를 거친 영상들로부터 미리 촬영된 배경과의 차분에 의해 인체 부위를 구한 다음 2진화 하여 인식에 필요한 제스처 영상을 얻는다.

일반적으로 카메라로부터 획득된 영상은 640×480의 큰 영상이므로 계산 량과 잡음을 줄이기 위해 영상을 일정한 크기로 정규화 한다. 본 논문에서는 영상의 크기를 100×100으로 정규화 시켰다. 또한 특별한 조명을 이용하지 않았기 때문에 각 영상마다 잡음이 많이 포함되어있다. 이를 제거하기 위해 밝기 정규화를 수행하였다. 밝기정규화 방법에도 여러 가지가 있으나 본 논문에서는 메디안 필터를 이용하여 노이즈를 제거하고 평활 화를 수행하여 영상들이 동일한 밝기 범위 내에서 들어가도록 하였다[12].

영상의 정규화는 제스처 객체를 효과적으로 분할하기 위한 전 처리이다. 인간의 신체만을 분할 한 영상을 얻기 위해선 식(2)에 표현한 것처럼 각 입력 영상 집합을 미리 촬영한 배경 영상과의 차를 이용한다.

$$X(i,j)=\sum_{i=0}^{N}\sum_{j=0}^{M}[Img_i(i,j)-Img_b(i,j)] \quad \text{식(2)}$$

이 때, $Img_i(i,j)$는 입력영상이고, $Img_b(i,j)$는 배경영상이며 $X(i,j)$는 입력 영상에서 배경 영상을 뺀 값으로 이루어진 영상이다. 이렇게 구해진 차 영상을 각 개인 간의 동작의 차이와 배경의 노이즈를 제거하고 동작성만을 강조하기 위해 2진화 처리한다. 2진화된 영상은 단순히 배경과 뺀 차분 영상이므로 약간의 잡음과 실제 동작을 제대로 표현하지 못하는 결과 영상을 나타내었다. 이를 해결하기 위해 2진화된 객체를 가장 효과적으로 표현하는데 주로 이용되는

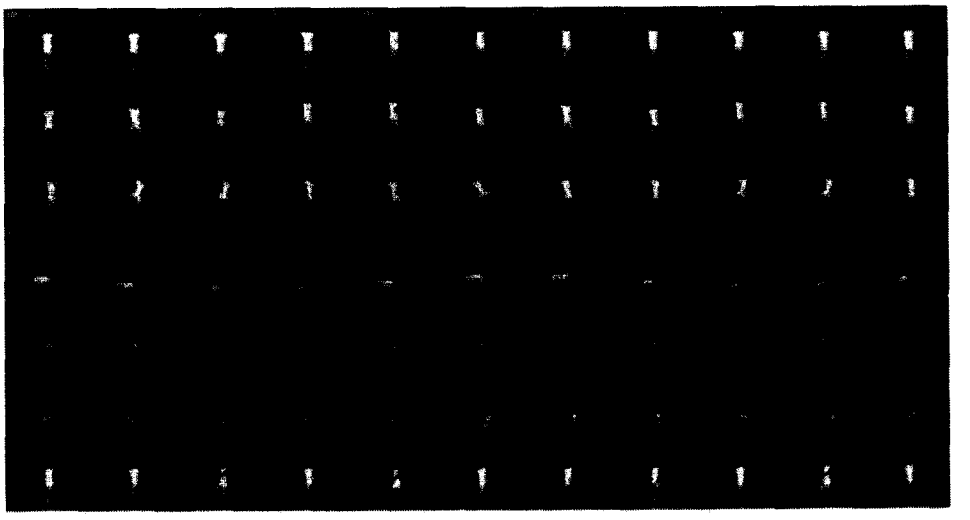


그림 1. 실험에 사용된 7종류의 맨손체조 동작.
(위로부터 팔 운동, 팔 벌려 뛰기 운동, 다리 벌리기 운동, 팔 굽혀펴기 운동, 쪼그려 뛰기, 다리 펴기 운동, 걷기)

Closing과 Opening 처리를 한다. 최종적으로 구해진 2진 영상을 위치변화에 안정적인 인식을 위해 제스처 객체의 무게 중심을 화면의 중심(50,50)으로 이동시킨다. 그림 2는 이렇게 만들어진 2진 영상들을 나타낸 것이다.

이렇게 처리된 영상집합을 식(3)으로 나타낸다.

$$x=[x_1, x_2, x_3, \ldots, x_N]^T \quad \text{식(3)}$$

## 2.2 정규화 된 영상 집합을 이용한 제스처 공간 구성

앞 절에서 설명한 방법을 통하여 얻어진 영상들을 이용하여 제스처의 전체적인 외관 특징을 표현할 수 있는 저차원 벡터 공간, 즉 파라메트릭 제스처 공간을 생성한다. 이 공간은 주성분 분석법(Principal Component Analysis)이라는 통계적 방법에 의해 만들어진다. 주성분 분석법은 앞에서 간단히 설명했듯이 각 제스처 영상집합들에 대한 픽셀 값들의 공간적 위치 값들이 주로 각 영상에서 어디에 분포하는가를 계산하여 확률 빈도가 높은 벡터 값들을 고유치 값에 비례하여 재구성하는 방법이다. 따라서 이 방법은 고유벡터(Eigenvector)와 고유치(Eigenvalue)를 계산하여, 제스처의 평균 모델을 구하여 이용한다. 고유벡터를 계산하기 위해서는 모든 영상에 대한 평균 영상을 구하여 각 영상들과의 차를 구한다. 평균 영상 $c$ 와 새로운 영상 집합 $X$를 식(4)와 식(5)와 같이 나타낸다.

$$c=(1/N)\sum_{i=1}^{N} x_i \quad \text{식(4)}$$

$$X \triangleq [x_1-c, x_2-c, x_3-c, \ldots, x_N-c]^T \quad \text{식(5)}$$

고유공간을 구하기 위해서는 M×N의 크기를 지닌 영상 집합 X를 식(6)과 같이 계산하고 식(7)를 만족하는 고유벡터를 구하면 된다. 여기서 M은 한 영상의 픽셀 수이고 N은 전체 영상의 개수를 나타내는 정수이다. 즉 $Q$에 대한 고유치 $\lambda$ 와 고유벡터 $e$ 를 구한다.

$$Q \triangleq XX^T \quad \text{식(6)}$$

$$\lambda_i e_i = Q e_i \quad \text{식(7)}$$

여기서 고유벡터 $e$는 공분산 행렬 Q의 열(Row) 벡터들에 대해 높은 확률 값을 가지는 벡터 순으로 전체 행렬을 재구성한 결과이다. 하지만 모든 열(Row) 벡터들은 서로 직교이기 때문에 특징 값이 중복 되지 않으면서 고유의 의미를 가진다. 따라서 모든 영상의 기준이 되는 모델은 단지 각 영상 픽셀들의 위치와 휘도치를 앞에서 소개한 통계적 수법으로 구성 할 수 있다[4,6].

본 논문에서는 고유치와 고유벡터를 구하기 위해 특이치 분해(Singular Value Decomposition)를 이용하여 고유 공간을 구성하였다. 특이치 분해를 이용하면 공분산 행렬을 구성하지 않고 단지 영상집합 $X$는 3가지의 행렬 집합으로 분해 된다[7]. 구해진 행렬 집합 중 공분산의 고유 벡터로 이용되는 행렬은 공분산 행렬 $X$와 크기가 일치하는 행렬이다. 마지막으로 특이치 분해 과정에서 나온 고유벡터를 고유치가 큰

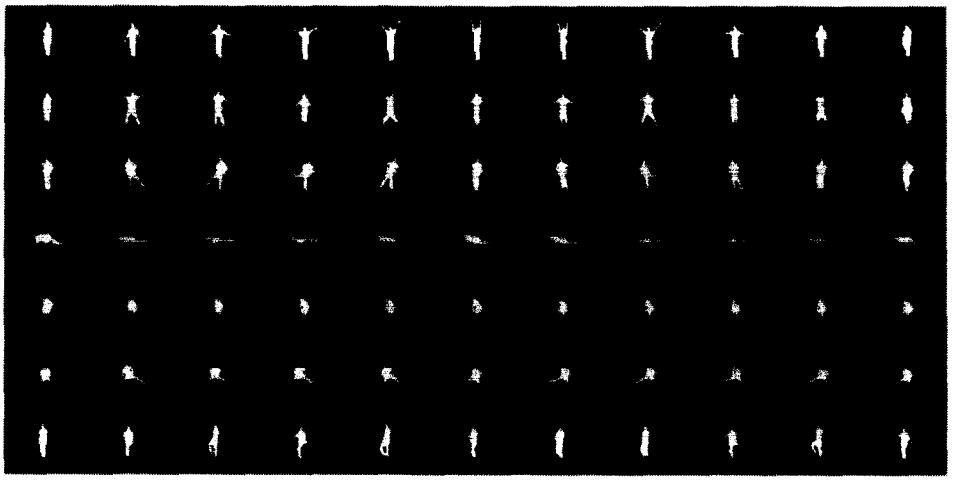


그림 2. 그림 1에 대해 크기, 밝기, 위치 등의 정규화를 통해 얻어진 영상집합.

순서대로 재구성한다. 각 고유벡터가 지닌 고유치의 크기는 그 고유벡터의 중요도를 의미하므로 그 고유 공간을 규정하는 중요 고유벡터를 식(8)을 이용하여 선택한다. 따라서 모든 고유벡터를 고유 공간 구성에 이용하지 않고 많은 영상을 대표할 수 있는 주성분의 벡터만을 이용 할 수 있다.

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{N}\lambda_i} \geq T_1 \qquad \text{식(8)}$$

여기서 $T_1$은 고유벡터의 개수를 조정하는 문턱치이며, 인식과 포즈 평가 시 이용되는 고유벡터 $\{e_i \mid i = 1,2,...,k\}$는 저차원 공간을 구성하기 위해 $(k \ll N)$을 만족하며 전체 영상에 대한 의미를 충분히 포함 할 수 있는 k = 16을 이용하였다. 이렇게 구성된 행렬 집합은 제스처 공간으로 표현할 수 있다.

### 2.3 제스처 공간에서 상관관계와 거리

2.2 절에서 얻어진 제스처 공간에 평균 영상 c에서 뺀 영상 집합 X를 모두 식(9)에서 나타낸 것과 같이 제스처 공간에 투영시킨다.

$$m_n = [e_1, e_2, e_3, \ldots, e_k]^T (x_n - c) \qquad \text{식(9)}$$

투영시킨 결과는 이산적인 점들로 표현되며, 이 점들은 하나의 영상을 의미하게 된다. 연속적인 점들은 서로 연관성이 많기 때문에 제스처 공간으로 투영시킨 결과는 서로 깊은 상관관계를 가진다[4,13, 16,19].

식(10)과 같이 각 제스처들은 서로 관계있는 연속성을 가진 점들의 집합으로 나타나게 된다.

$$m(m_1, m_2, \ldots, m_n) \qquad \text{식(10)}$$

제스처 공간을 계산하기 위해 이용된 영상 $x_1$과 $x_2$를 제스처 공간에 투영시켜 얻은 점들이 각각 $m_1$과 $m_2$ 라면 이 두 점의 사이의 거리와 두 영상간의 상관관계는 서로 밀접하다. 즉 거리가 가까울수록 두 영상은 닮은 영상이 된다.

공간상의 임의의 한 점을 $m_1$이라면 원래의 영상은 식(11)과 같이 복원 될 수 있다.

$$x_1 = \sum_{i=1}^{N} m_1 \, e_i + c \qquad \text{식(11)}$$

여기서 c는 전체 영상집합의 평균 영상이고 e는 전체 고유벡터를 나타낸다. $k$ 개의 제스처 공간에 투영한 영상은 식(12)로 표현된다.

$$x_1 \approx \sum_{i=1}^{k} m_1 \, e_i + c \qquad \text{식(12)}$$

두 영상간의 유사도는 식(13)과 같은 템플릿 매칭으로 표현 된다.

$$\|x_1 - x_2\|^2 = (x_1 - x_2)^T (x_1 - x_2) = 2 - 2x_1^T x_2 \qquad \text{식(13)}$$

이 식을 식(11)를 이용하여 나타내면 식(14)과 같다.

$$\|x_1 - x_2\|^2 = \| \sum_{i=1}^{k} m_1 e_i - \sum_{i=1}^{k} m_2 e_i \|^2 \qquad \text{식(14)}$$

이 식을 풀면 식(15)로 나타낼 수 있다.

$$\| \sum_{i=1}^{k} m_1 e_i - \sum_{i=1}^{k} m_2 e_i \|^2 = \| \sum_{i=1}^{k} (m_1 - m_2) e_i \|^2$$
$$= \sum_{i=1}^{k} \sum_{j=1}^{k} e_i^T e_j \times (m_1 - m_2)^2 = \| m_1 - m_2 \|^2 \qquad \text{식(15)}$$

여기서 $i = j$ 일 때 $e_i^T e_j = 1$이고 그렇지 않으면 0이 된다. 식(14)과 식(15)로부터 우리는 식(16)을 얻을 수 있다.

$$\| x_1 - x_2 \|^2 \approx \| m_1 - m_2 \|^2 \qquad \text{식(16)}$$

따라서 두 점간의 관계는 두 영상간의 관계와 매우 유사하다는 것을 알 수 있다. 즉 고유 공간에 투영된 점이 가까울수록 영상들은 높은 상관관계를 가진다는 것이 증명 되었다.

### 2.4 제스처 공간에서 특징 심볼을 이용한 모델 구성

이 절에서는 제스처 공간에서 여러 변화에 대해 안정된 특징을 가지는 모델을 선정하는 방법에 대해 설명한다. 일반적으로 시각적 학습 방법을 이용한 인식은 모든 외관들과 일치하는 영상들을 모델화 하여 등록한다. 그리고 입력 영상은 등록된 모델들과 비교하면서 인식을 위한 매칭이 수행된다. 효과적인 매칭을 수행하기 위해서는 매칭 횟수와 인식 속도를 개선시키는 것이 필요하다.

본 논문에서는 각 동작들의 모델들에 대한 특징들을 구하여 모든 모델들을 비교하지 않고 특징이 되는

일부의 특징 심볼들과 비교하여 인식률의 저하 없이 매칭 속도를 증가시키는 방법을 이용하였다.

그림 3은 $e_3$에서 시간 축에 대한 제스처 궤적을 나타낸 것이다. 이 궤적으로부터 반복된 곡선이 비슷한 형태를 가지며 국소적인 최대값과 최소값으로 표현되는 것을 알 수 있다. 하지만 제스처의 큰 변화에 대해 모든 차원에서 제스처를 정확하게 표현하는 곡선을 나타내지 않기 때문에 모델 영상을 가장 잘 표현할 수 있는 차원으로 재구성한다. 선택된 차원에 의해 구해진 모델 영상들 중 국소적으로 최대값과 최소값으로 나타난 모델 $S(m)$을 제스처 특징 심볼로 식(17)과 같이 구성한다.

$$S(m) \equiv [s_{m,1}, s_{m,2}, s_{m,3}, \ldots, s_{m,n}] \qquad \text{식 (17)}$$

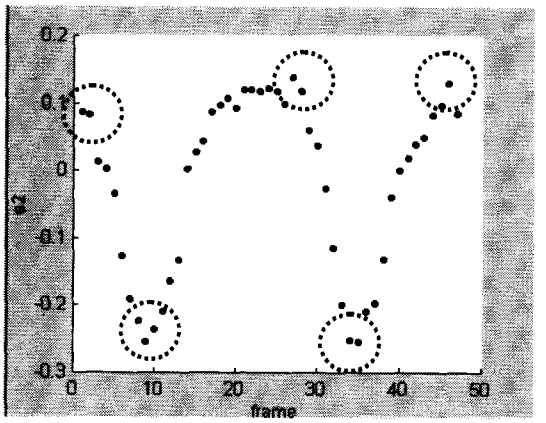



그림 3. 입력 영상의 e3 성분의 변화. 점선 원 내의 영상이 제스처의 변환 점을 나타낸다.

## 3. 제스처 인식과 동작 정보 추정

연속적으로 입력되는 영상은 2.3절에서 구성된 제스처 특징 심볼들에 의해 인식된다. 제스처 공간에서 입력 영상들은 선형적인 궤적을 그리게 되는데 이 궤적은 연속적인 입력 심볼로 표현된다. 이런 심볼들을 모델 심볼에서 제스처 특징 심볼을 구한 방법으로 제스처 입력 특징 중 첫 번째 심볼이 제스처 특징 심볼 내에 존재하면 심볼 매칭은 순차적으로 수행되게 된다. 연속적 입력 심볼이 완벽하게 모델 특징 심볼과 일치하면 일치된 모델 제스처로 인식하게 된다. 그림 4는 3차원 제스처 공간상에서 연속적인 입력 제스처와 모델 제스처 간의 맵핑을 나타내고 있다.

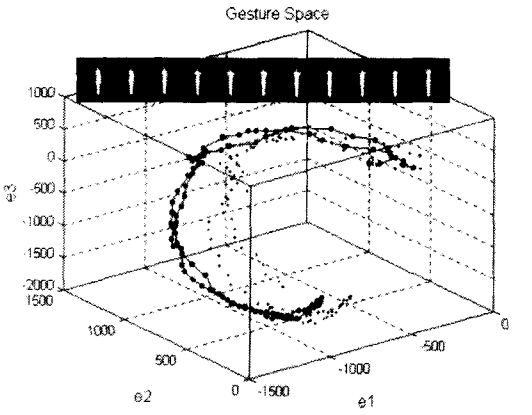



그림 4. 제스처 공간에 모델 영상과 입력을 영상 투영한 결과

### 3.1 입력 영상과 모델영상의 매칭에 의한 제스처 인식

정규화 된 영상들이 제스처 공간에 투영되어 인식을 위한 모델들로 정해지면 실제 인식에 필요한 처리는 매우 간단하다. 먼저 평균 영상에서 입력 영상 $y$를 뺀 다음 고유공간에 식(18)과 같이 모델로 구성된 제스처 공간상에 투영하면 된다.

$$z_n = [e_1, e_2, e_3, \ldots, e_k]^T (y_n - c) \qquad \text{식(18)}$$

구해진 $z_n$은 제스처 공간상에서 점들로 표현되는데 이러한 점들은 제스처 특징 심볼을 구한 것과 같은 방법을 이용하여 입력 특징 심볼로 구성하게 된다. 구성된 입력 특징 심볼들은 모델 특징 심볼들과 비교 하여 매칭이 이루어진다. 만약 $D_1$이 임계 치보다 적으면 입력 특징 심볼은 비교된 모델 특징 심볼 제스처로 인식하게 된다. 제스처 인식은 입력 영상의 잡음이나 시스템의 착오를 감안하여 임계 치를 정해서 각 모델들과의 거리 중 그림 5에서 나타낸 것과 같이 최소거리를 구하여 이루어진다.

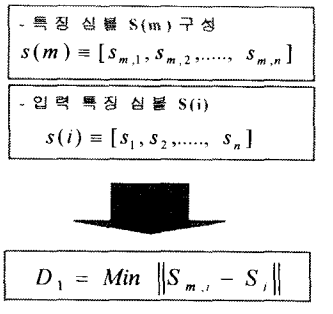



그림 5. 입력 영상과 모델 영상의 매칭 과정

### 3.2 연속적 입력 영상의 구체적 동작 정보 계산

입력 영상에 대해 구체적 제스처 정보 즉 동작의 빠르기나 크기를 알기 위해서는 제스처 공간상에 연속적으로 투영된 입력 영상들 간의 거리를 계산 하면 된다. 단위 시간에 동작의 이동의 합은 제스처의 빠르기를 나타내고 한 동작의 전체 이동거리의 합은 제스처의 동작 크기를 나타냄을 알 수 있다[8]. 따라서 단위 시간 내 투영된 입력 영상간 거리가 먼 경우 빠른 동작이며 한 동작의 전체 거리의 합이 클수록 동작의 크기도 크다는 것을 추정 할 수 있다. 그림 6은 제스처 공간상에서 입력 영상의 제스처 정보를 계산하는 방법에 대해 나타낸 것이다.

그림 6에서 보듯이 단위 시간당 입력 영상이 모델 영상보다 큰 거리 값을 가지므로 입력 영상은 모델 영상 보다 빠른 동작임을 알 수 있다. 입력 영상에 대해 동작의 빠르기나 크기를 알기 위해서는 제스처 공간상에 연속적으로 투영된 입력 영상들 간의 거리를 식(19)을 이용하여 계산한다.

$$\sum (I_i - I_{i+1}) = \| Id_i \| \qquad \text{식(19)}$$

여기서 $I$는 투영된 입력 영상을 나타내며 $Md_i$와 $Id_i$는 각 각 연속적으로 투영된 모델 영상들 간의 거리와 입력 영상간 거리를 나타낸 것이다.

그림 6에서처럼 연속적으로 투영된 입력 제스처 영상의 거리가 제스처 모델 영상 보다 거리 값이 크므로 ( $\| d_1 \| < \| d_2 \|$ ) 입력 동작이 단위 시간동안에 모델 동작보다 빠른 동작임을 추정 할 수 있다.

최종 인식 방법은 3.1절에서 기술한 방법과 이 절에서 이용한 방법을 결합하여 결정하게 된다. 먼저

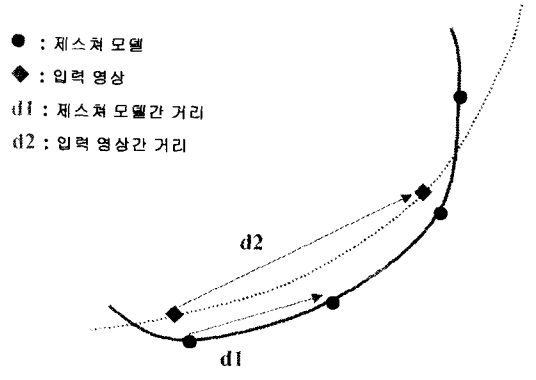



그림 6. d1<d2 이므로 단위 시간에 입력 제스처가 빠른 동작임을 나타냄

입력 영상과 최소 거리 값을 가지는 모델 동작을 해당 포즈로 인식하게 되며 그때의 각 포즈간 거리 값을 모델 영상간 거리 값과 비교하여 동작의 빠르기를 측정하게 된다. 여기서 $P_i, D_i$는 각 각 영상에 대한 포즈와 거리에 대한 평가 값을 나타낸다. 입력 영상 집합을 제스처 공간에 투영 하였을 때 투영된 모델집합과 최소가 되는 거리 값을 가지는 $P_i$를 해당 모델로 인식하고 연속 적으로 투영되는 입력 영상간의 거리 값을 구하여 모델 영상간 거리와 비교 하여 빠르기를 판단하게 된다. 만약 빠르기 측정 값 $D_i$이 에러를 고려한 임계 치 범위 내에 있으면 같은 빠르기로 인식하게 된다.

$$P = Min \| \sum_{i=1}^{k} Sm_i - \sum_{i=1}^{k} S_i \| \qquad \text{식(20)}$$

$$D = Min \| \sum (I_i - I_{i+1}) - \sum (MI_i - MI_{i+1}) \| \qquad \text{식(21)}$$

$$Y_n = \eta_1 (P_i) + (1 - \eta_1)(D_i) \qquad \text{식(22)}$$

여기서 $P$와 $D$는 각 각 투영된 입력 심볼과 모델 심볼의 최소거리의 값이고 $D$는 입력 심볼 간 거리와 모델 심볼 간 거리의 최소값을 나타낸다. 따라서 $Y$ 값이 최소가 될 때 입력 제스처를 해당 모델 제스처로 인식하게 된다.

여기서 포즈인식과 동작 정보의 가중치 비율은 6 : 4이다.

## 4. 실험 및 고찰

그림 7에 제스처 공간을 이용한 제스처 인식과 동작 정보 추정의 전 과정을 보인다. 그림에서 알 수 있는 이 정규화 과정을 거친 모델 영상집합을 주성분 분석을 통해 제스처 공간을 구성하고 입력 영상은 모델 영상에서와 같은 전 처리 과정을 거친 후 제스처 공간에 투영하여 모델과 비교하게 된다. 모델과 투영된 입력 영상간의 거리 값을 계산하여 최소가 되는 모델을 입력 영상의 제스처로 인식한다.

실험에 이용된 제스처 영상은 7가지 맨손체조를 반복적으로 수행한 것을 카메라로 촬영하여 획득한 것이다. 각 동작들이 서로 구분 될 수 있도록 팔, 다리 등의 구분 운동에 대해 촬영하고 크기 정규화를 100 ×100 영상으로 변환하였다. 영상 집합의 고유 벡터를 계산한 후 재구성된 영상을 가장 잘 복원하는 16

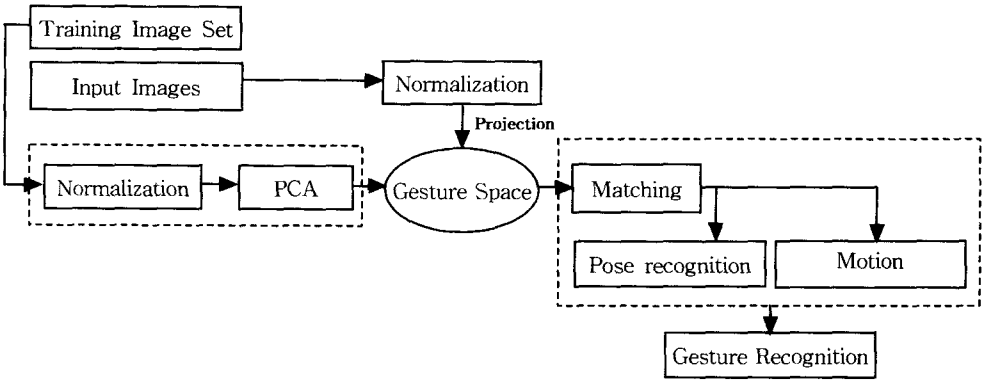



그림 7. 제스처 공간을 이용한 인식과정의 흐름도

차원의 벡터를 선택하여 제스처 공간으로 이용하였다. 따라서 100×100 = 10000 차원의 이미지가 16차원으로 압축되는 효과도 거둘 수 있었으며 실제 인식 처리 과정은 실시간 처리에도 적합하다는 것을 알 수 있었다[4-6,8]. 그림 8은 3차원 제스처 공간에서 각 영상들이 맵핑되는 결과를 나타냈다. 이 그림에서 모델영상의 투영된 각 포즈가 선형적인 궤적에 따라 배치되어 시간 축에 가까운 동작은 서로 깊은 상관관계를 가지고 있다는 것을 알 수 있다.

입력영상은 각 동작에 대해 임의적으로 빠르기와 포즈의 정도 차를 조정하여 실험에 이용하였다. 이러한 동작들은 모델과 비교하여 동작의 정도 차가 크고 속도가 빠른 동작에 대해 많은 거리 값을 나타내었고 비슷한 빠르기와 동작에 대해서는 비슷한 정도를 나타낸 다는 것을 알 수 있었다.

표 1과 2는 100프레임으로 구성된 입력 동작들에

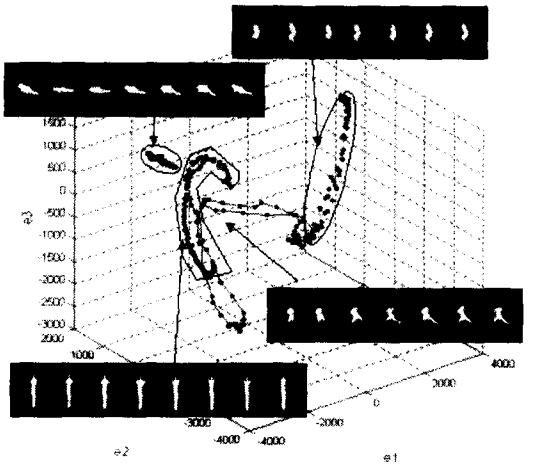


그림 8. 동작별 투영 결과

표 1. 동작별 인식 결과(인식 오류란 에는 오 인식 된 동작과 비율을 표시하였음)

| 입력 동작 | 인식성공률 (%) | 인식 오류 (%) |
|---|---|---|
| 팔 운동 | 90 | 팔 벌려 뛰기(8.0), 걷기(2) |
| 팔 벌려 뛰기 | 87 | 팔 운동(11), 걷기(2) |
| 다리 운동 | 100 | - |
| 앉아 뛰기 | 100 | - |
| 다리 펴기 | 100 | - |
| 팔 굽혀 펴기 | 100 | - |
| 제자기 걷기 | 98 | 팔 운동(1), 팔 벌려 뛰기(1) |

관한 인식률(%)을 나타낸 것이며, 각 입력 동작은 제스처 공간에 투영된 후 3.2절에서 설명한 방법을 이용하여 모델과 매칭을 수행 하였다.

보듯이 3차원과 16차원 벡터를 이용한 매칭 결과에서 팔운동과 팔 벌려 뛰기의 매칭에서 가장 오류율이 높게 나타났는데, 이는 두 동작 자체가 서로 유사한 형태를 가지고 있기 때문에 다른 동작에 비해 가장 유사 율이 높게 나타났다. 하지만 이러한 문제는 모델과 입력 동작의 매칭 도중 이전 프레임들의 매칭 결과를 참조하면서 매칭을 수행한다면 결과를 향상 시킬 수 있다.

## 5. 결 론

본 논문에서는 주성분 분석법을 이용하여 영상 집합에 대한 고유 벡터를 구하고 기여도가 큰 벡터만으

로 구성된 고유 공간을 이용하여 연속적인 인간의 동작 영상으로부터 제스처를 인식하는 방법을 제안하였다. 이 방법은 인체의 각 관절의 좌표를 복잡한 연산을 통해 계측하거나 기하학적인 특징 즉 에지나 모멘트와 같은 정보를 이용하지 않고 단지 인간의 2차원 실루엣 영상 그대로를 특징으로 이용하였다 [8,12]. 이 방법은 특징 정보를 구하기 위해 복잡한 전 처리 과정에서 나타나는 자동적인 임계 치의 설정 문제와 파라미터를 정확하게 추출해야 하는 문제를 피할 수 있으며 모델과 입력 영상간의 직관적인 매칭을 수행함으로써 안정적인 인식 결과를 얻는다. 또한 기존의 동작의 분류나 구분에 제한된 인식방법과는 달리 투영된 각 벡터들의 시간의 따른 상관관계를 이용하여 제스처의 구분뿐만 아니라 제스처의 구체적인 정보 즉 동작의 크기나 빠르기의 정보를 얻어내는데 유용하다는 것을 알 수 있었다.

본 논문에서 이용한 매칭 방법은 입력 심볼과 모델 심볼의 단순 거리비교를 통해 최단 거리에 속한 제스처그룹을 해당 동작으로 인식한다. 하지만 7개의 동작들 중에서도 궤적들이 서로 얽혀있거나 교차되는 부분에 대해서는 애매한 결과를 나타내었다. 또한 동작도중 몸통과 손발이 겹쳐서 모델영상과 입력영상의 모습이 현저히 달라지는 경우는 인식에 어려움이 따르고 복잡한 배경에서는 사람 영역만을 분리하는데 어려움이 있었다. 따라서 다음 연구에서는 투영된 입력영상과 모델 영상에 대한 전체 궤적의 형태를 매칭에 이용하는 점과 하나의 포즈에 대해 다중 영상 즉 두 대 이상의 카메라로부터 입력된 영상을 모델로 구성하여 간섭과 겹침 문제를 해결하는 점이 연구 과제로 남는다.

## 참 고 문 헌


[1] Vladimir I. Pavlovic, Rajeev Sharma, "Visual Interpretation of Hand Gestures for Human Computer Interaction: A Review", IEEE Trans, Pattern Analysis and Machine Intelligent, Vol.19, No.7, 1997.

[2] Claudetter Cedras and Mubarak Shah, "Motion-Based Recognition: A Survey", Technical Report, Department of Computer Science, University of Central Florida, Orlando, 1994.

[3] 강호석, "동작 캡처 기술", The Magazine of the IEEK, Vol.25, No.2, pp.141-147, 1998.

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

[5] Shree K. Nayar, S.A.Nene and Hiroshi Murase, "Subspace methods for robot vision", IEEE, Trans, Robotics and Automation, Vol.12, No.5, pp.750-758, 1996.

[6] Shree K. Nayar, Simon Baker and Hiroshi Murase, "Parametric Feature Detection", Tech. Rep. CUCS-028-95, Dept. of Computer Science, Columbia University, 1995.

[7] William H, Saul A, Teukolsky, William T. bettering, and Brian P. Flannery. Numerical Recipes in C (Second Edition), Cambridge University Press 1992.

[8] Takahiro Watanabe and Masahiko Yachida, "Real Time Recognition of Gesture and Gesture Degree Information Using Multi Input Image Sequences", ICPR'98, Vol.2, p.185-188, 1998.

[9] Shigeyoshi Hiratsuka, Kohtaro Ohba, Shinya Kajikawa, Hidaru Inooka, Kazuo Tanie, "Stable Gesture Verification in Eigen Space", MVA '98 IAPR Workwhop on Machine Vision Applications, 1998.

[10] Andrew D. Wilson and Aaron F.Bobick, "Recognition and Interpretation of Parametric Gesture" M.I.T. Media Laboratory perceptual Computing Section Technical Report No. 421.

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

[12] Takahiro Watanabe, Chil-Woo Lee, Akitoshi Tsukamoto and Masahiko Yachida, "A Method of Real-Time Gesture Recognition for Interactive System", IEEE, Proceedings of ICPR, 1996.

[13] Hiroshi Murase and Michael Lindenbaum, "Partial Eigenvalue Decomposition of Large

Images Using Spatial Temporal Adaptive Method", IEEE Trans, Image Processing, Vol.4, No.5, 1995.

[14] Thad Eugene Starner, "Visual Recognition of American Sign Language Using Hidden Markov Models", Massachusetts Institute of Technology, Cambridge MA, Technical Report, 1995.

[15] Bavack Moghaddam, Alex Pentland, "Probavilistic Visual Learning for Object Representation", IEEE Trans, Pattern Analysis and Machine Intelligence, Vol.19, No.7, 1997.

[16] Shree K. Nayar and Hiroshi Murase, "Dimensionality of Illumination Manifolds in Eigenspace", CUCS-021-94, Technical Report, Department of Computer Science, Columbia University, New York, 1994.

[17] Hiroshi Murase and Shree K. Nayar, "Illumination Planning for Object Recognition Using Parametric Eigenspaces", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.16, No.12, 1994.

[18] Jeff Edwards and Hiroshi Murase, "Appearance Matching of Occluded Objects Using Coarse-to-fine Adaptive Masks", IEEE, 1997.

[19] Shree K. Nayar, S.A.Nene and Hiroshi Murase, "Real-Time 100 Object Recognition System", Technical Report, Department of Computer Science, Columbia University, New York, 1996.

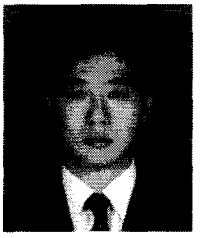
[20] Nan Li, Shawn Dettmer and Mubarak Shah, "Lipreading Using Eigensequences", Technical Report, Computer Science Department, University of Central Florida, Orlando, FL 32816, 1994.

[21] James William Davis, "Appearance-Based Motion Recognition of Human Actions", MIT Media Lab Perceptual Computing Group Technical Report, 1996.

[22] Sameer A. Nene and Shree K. Nayar "A Simple Algorithm for Nearest Neighbor Search in High Dimensions", Technical Report No. CUCS-030-95, Department of Computer Science Columbia University, New York, N.Y.10027, 1995.

**이 칠 우**

1986년 중앙대학교 전자공학과 학사
1988년 중앙대학교 전자공학과 석사
1992년 일본 동경대학교 대학원 전자공학과 공학박사
1992년~1995년 이미지 정보과학 연구소 수석연구원 겸 오사까 대학 기초공학부 협력연구원
1995년 리츠메이칸대학 특별초빙강사
1996년~현재 전남대학교 정보통신공학부 부교수
관심분야 : 컴퓨터비전, 멀티미디어 데이터베이스, 컴퓨터그래픽스, 의료 영상 처리



**이 용 재**

1998년 호원대학교 학사
2000년 전남대학교 컴퓨터공학과 대학원 석사
2002년 전남대학교 컴퓨터공학과 대학원 박사 수료
2004년 1월~현재, (주)어플라이드비전텍 연구/개발부 근무
관심분야 : 컴퓨터 비전, 제스처 인식, 의료 영상 처리