로컬 모션정보와 글로벌 모션정보를 조합한 제스처 인식

오재용, 이칠우
전남대학교 정보통신공학과

Gesture Recognition in Video image with Combination of Partial and Global Information

Jae-Yong Oh, Chil-Woo Lee
Dept. of Computer Engeneering, Chonnam Nat'l University

요 약

본 논문에서는 일반적인 비디오 스트림에서 자동으로 인간의 제스처를 인식하는 알고리즘에 대하여 기술한다. 본 알고리즘은 입력된 비디오 영상으로부터 추출된 신체영역의 2차원적 특징 벡터를 사용하며, 주성분 분석법(*Principle Component Analysis*)을 통하여 모델 제스처 공간(*Model Gesture Space*)을 구성함으로서 제스처를 통계학적으로 분석/표현하며, 이 제스처 공간에서 새로 입력되는 영상을 같은 방법으로 투영시키고, HMM(*Hidden Markov Model*) 이론을 적용하여 심볼화함으로써 최종적으로 제스처를 인식하게 된다. 본 방법은 기존의 제스처 인식 방법들과는 달리 전체적인 영상 정보(*Global Information*)와 세부적인 영상 정보(Partial *Information*)를 조합하여 사용한다는데 특징이 있으며, 본 알고리즘을 통해 보다 정확하고 강건한 제스처 인식 기술을 실생활에 적용할 수 있을 것이다.

1. 서론

컴퓨터 비젼 분야에서 인간의 제스처를 자동으로 이해하는 시스템을 만드는 일이란 매우 어려운 일이다. 그러나, 최근 인간의 제스처에 대한 관심이 높아지고, 컴퓨터 시스템이 급속도로 발달하면서 인간 친화적 인터페이스, 대용량 데이터베이스 시스템, 지능형 감시 카메라 시스템, 고효율 통신 시스템 등과 같은 분야에 인간의 제스처를 기반으로 하는 기술이 다각도로 응용되고 있다.

인간의 제스처를 다루던 예전의 시스템들은 인간의 제스처 정보를 획득하기 위하여 몸의 손발이나 관절에 부착하는 센서류를 많이 사용했었다. 그러나 이러한 방법은 데이터 글러브와 같은 거추장스러운 장비를 몸에 부착해야만 했고, 더욱이 이들은 모두 긴 케이블로 시스템과 연결되어 있었기 때문에 매우 불편하였다. 이러한 불편함을 해소하기 위하여 시각 기반의 비접촉식(non-tactile method)방법을 사용하기 시작하여 애니메이션이나 영화에서 응용될 만큼 발전하게 되었다. 그러나 이러한 방법의 대부분은 아직도 여러 형태의 마커(*Marker*)를 몸에 부착해야만 하며, 영상정보 사용으로 인하여 발생하는 조명 조건의 민감성, 배경과 전경의 모호성 등의 문제점은 아직도 완전히 해결되지 않은 상태이다. 또한, 인간의 몸은 3차원의 매우 복잡한 다관절체로 이루어져 있기 때문에 움직임을 자동적으로 추적하고 분석하는 일은 매우 어려운 일이다.

영상에서의 인간의 제스처는 크게 두 가지 형태의 정보로 이루어져 있으며, 본 논문에서는 인간의 제스처를 움직임 전체의 변화를 표현하는 광역정보(*Global Information*)와 머리 손, 발과 같은 신체 부분의 움직임을 표현하는 상세정보(*Partial Information*)를 이용하여 표현한다. 인간의 제스처가 신체의 자세 변화라고 가정한다면, 특정 제스처에 대해서는 매우 쉽게 인식해 낼 수 있을 것이다. 의자에 앉는 제스처

본 연구는 한국 과학재단 지정 전남대학교 "고품질 전기전자부품 및 시스템 연구센터"의 연구비 지원에 의해 수행 되었음.

를 예로 들면, 서 있을 때는 위아래가 긴 직사각형 형태의 경계(*Boundary*)가 될 것이며, 앉을 때는 위아래가 보다 짧아진 정사각형 형태의 경계가 될 것이기 때문에, 두 팔의 움직임을 고려하지 않고도 이를 구분하는 것은 매우 쉬운 일이다. 그러나, 서있는 채로 손을 흔들고 있는 제스처의 경우, 손의 움직임이 매우 중요한 의미를 가지고 있게 된다. 따라서 보다 정확히 제스처를 정의하기 위해서는 이 두 가지 정보가 조합되어야 한다. 그러나 이들 두 움직임 요소는 다른 한 쪽에 독립 되도록 정의되기 매우 어렵기 때문에 적절히 조합되어 사용되어야 한다.

본 논문에서는 일반적인 비디오 스트림에서 자동으로 인간의 제스처를 인식하는 알고리즘에 대하여 기술한다. 비디오 영상으로부터 추출된 신체영역의 2차원적 특징 벡터를 사용하며, 기존의 방법들과는 달리 전체적인 영상 정보(Global *Information*)와 세부적인 영상 정보(*Partial Information*)를 조합하여 사용한다. 그림 1은 본 알고리즘의 개략적인 흐름을 나타낸다.

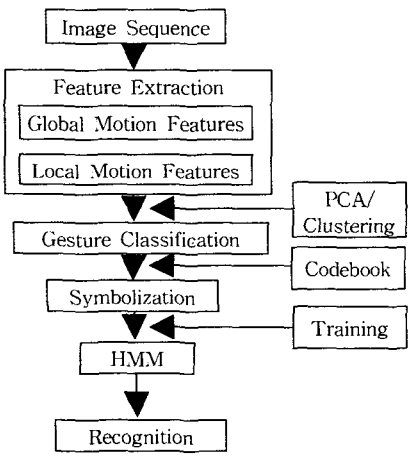

그림 1. 제스처 인식 과정의 흐름도

2. 전처리 및 특징 추출

일반적인 환경에서 획득된 비디오 이미지를 사용한다면, 사용되는 영상의 배경에는 수많은 물체들이 포함되어 있을 것이다. 이러한 배경 부분은 제스처 인식 과정에 불필요한 부분이며, 보다 정확한 특징 추출을 위해 배경이 제거된 영상, 즉 신체 영역만을 추출해야 한다. 전경(foreground)이라고 표현되는 신체 영역은 입력되는 영상에서 배경영상을 빼냄으로서 쉽게 추출할 수 있다. 본 논문에서는 식(1)과 같은 배경 모델을 이용한다.

$$|M(x,t) - I(x,t)| OR |N(x,t) - I(x,t)| > D(x,t) + C \quad (1)$$

위 식에서 각 기호가 갖는 의미는 다음과 같다.

M : 각 픽셀에 대한 최소 밝기 (Minumum brightness)

N : 각 픽셀에 대한 최대 밝기 (Maximum brightness)

D : 인접 프레임간의 최대 밝기 차이 (Maximum intensity difference of consecutive frame)

C : 입력 영상의 전체 밝기에 따른 상수

이와 같은 threshold 방법만으로는 조명 조건의 변화와 같은 상황에는 정확한 전경영상을 얻어내기 충분하지 못하므로, 본 논문에서는 영역 기반 노이즈 제거 방법 (*Region-based noise cleaning method*)를 추가적으로 사용한다. threshold를 수행한 뒤 수축 및 팽창 과정을 반복 수행함으로서 전격으로 추출된 1픽셀 단위의 작은 노이즈들을 제거할 수 있게 된다.

영상의 광역정보(*Global Motion Information*)를 이용하는 방법중 가장 널리 알려진 방법은 *Motion History Image (MHI)*[1]이다. 이 방법은 인간의 움직임을 분석하기 위해 시간에 따른 움직임의 누적 정보를 사용한다. 근접한 프레임간의 움직임 차이를 누적시킨 실루엣 영상이 만들어지고 이를 모델 영상과 비교함으로서 제스처를 인식할 있게 된다. 그러나 이러한 방법은 신체 영역의 윤곽만을 사용하기 때문에 윤곽 안쪽에서 생긴 제스처에 대해서는 인식이 불가능하며, 움직임의 속도에 따른 제스처의 구분이 모호하다는 단점이 있다. 또한, 움직임 정보를 누적해야 하므로, 움직임을 추적하기 어려운 떨어지는 물체에 대해서는 적용이 어렵다. 전체적인 움직임 특징을 이용하는 또다른 예로 *Ross Culter's method*[2]가 있다. 이 방법은 영상내 Blob의 Optical flow 특성을 이용하여 제스처를 인식하는데, 이 또한 영상내 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[3]. 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 분석 방법을 사용했다는 점에서 주목할만 하다.

본 논문에서는 1)너비, 2)높이, 3)무게중심의 수평성분, 4)무게중심의 수직성분, 5)Compactness, 6)Direction of first moment 의 6가지 제스처 영상의 형상 정보를 사용한다. 제스처가 부분 움직임과 전체 움직임의 시간에 따른 변화임을 고려할 때, 움직임의 변화 영상 (Motion Histroic Image)은 제스처 인식을 위한 많은 정보를 포함하고 있다. 또한, 서론에서 언급한 바와 같이 정확한 제스처 인식을 위해 보다 부분적인 움직임 정보를 추가적으로 사용한다. 보다 세분화된 신체 영역을 사용한다면 제스처를 보다 정확하게 표현할 수 있지만, 본 논문에서는 신체 영상을 4개의 세부 영역으로 나누고, 그 세부 영역의 무게 중심, 무게 중심의 변화량, 면적 정보를 특징으로 사용한다. 그림 2는 처리과정의 예를 나타낸다.

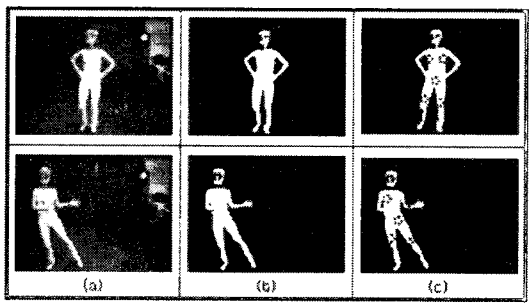

그림 2 전처리 과정 (a) 입력 영상, (b) 로컬모션을 위한 서브 영역, (c) 로컬 영역의 무게 중심

또한 본 논문에서는 시간에 따른 영상 그룹화 방법을 소개하며, 기본 개념은 식(2)와 같다.

$$F_t^{t+n-2} \qquad (3 \le t \le T,\ 0 \le n < 2) \qquad (2)$$

식(2)에서 T는 영상 시퀀스의 총 길이이며, 비디오 입력으로부터 추출된 이진 영상은 이미지 그룹화 및 특징 추출 과정을 거치게 된다. 시간 t 에서 I_{t-2}, I_{t-1}, I_t 가 하나의 그룹으로 분류되고, 각각의 영상에 대해서 18개의 특징 데이터 (6개의 전체 신체 영상 정보, 12 개의 부분 특징 정보)를 추출하게 된다. F_t 를 시간 t 에서의 특징집합이라고 가정하면, 식 (2)에서 F_t^t 는 시간 t 에서의 영상의 특징 집합이고, F_t^{t-1} 은 F_t와 F_{t-1}의 차이를 의미한다. 따라서 한 그룹의 전체 특징 정보는 총 54개가 되며, 그림 3은 이와 같은 시간에 따른 영상 그룹화 알고리즘을 도식화 한 것이다.

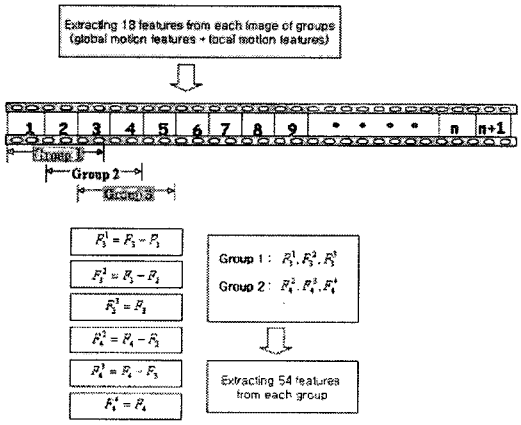

그림 3 시간에 따른 영상 그룹화와 특징 추출

3. 주성분 분석법과 제스처 공간

일반적으로 통계학에서는 다차원의 복잡한 데이터를 분석하기 위해서 주성분 분석법 (Principle Component Ananysis)을 사용한다. 앞서 언급한대로 본 논문에서는 입력 영상으로부터 추출된 다차원의 특징 벡터를 분석하는 방법으로 주성분 분석법을 사용한다. 그러나 이 방법은 분석하고자 하는 데이터의 구조가 선형적이어야 한다는 전제조건이 있으며, 그렇지 않을 경우 크 의미를 가지지 못하게 된다. 이러한 의미에서 본 논문에서 사용되는 특징벡터는 다차원의 선형 데이터 특성을 갖기 때문에, 주성분 분석법을 사용하기에 적합하다.

주성분 분석법의 첫단계로, 서로 다른 단위를 갖는 데이터들을 수치적으로 정규화를 수행하며, 특징값들은 식(3) 와 같이 표현된다.

$$x = [x_1, x_2, \Lambda, x_N]^T \qquad (3)$$

여기서 $N(=T-2)$은 로컬 및 글로벌 모션 정보의 특징값으로 구성된 그룹의 수를 의미한다. 식(3)과 같은 특징 집합을 이용하여 전체 신체 영상 정보와 부분 특징 정보들을 표현할수 있는 저차원의 벡터 공간, 즉 제스처 공간이 생성되며, 이 공간에서 저차원의 벡터를 이용하여 입력 영상을 분류하고 분석할 수 있게 된다.

4. 심볼을 이용한 제스처 인식

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. 입력영상이 그림 4에서처럼 클러스터링 알고리즘에 의해서 몇 개의 제스처 패턴들로 구분되어지면, 영상 시퀀스는 심볼 시퀀스로 형상화되고, 이를 HMM의 입력으로 사용한다. HMM 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* 알고리즘을 사용한다.

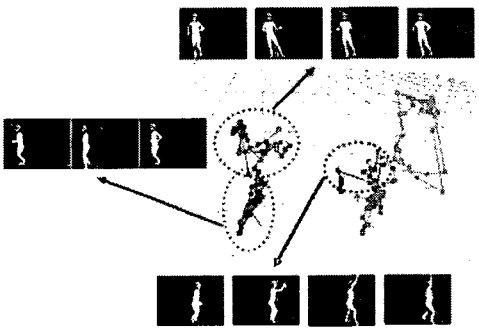

그림 4 영상 시퀀스의 제스처 공간으로의 투영

5. 실험과 결론

실험에 사용한 제스처 영상은 걷기, 앉기, 일어서기, 춤추기 등의 13가지 제스처 시퀀스를 사용한다. 이 영상 시퀀스는 320*240 해상도를 사용하며, 총 13개의 제스처 시퀀스를 모델로 구성하였다. 모델로 구성된 제스처들은 50개의 서로 다른 클러스터로 분류되어졌고, HMM을 통하여입력 결과에 대한 인식결과를 확인하였다.

본 실험에서 제안한 알고리즘이 테스트 영상들에 대해 비교적 높은 인식률을 보였다. 신체 특징점을 보다 정확히 세그멘테이션 한다면 인식률을 높일 수 있을 것으로 예상된다. 본 논문에서 사용한 제스처 인식 방법은 에지나 코너와 같은 기하학적인 특징정보가 아닌 모멘트, 신체영역 사이즈 및 무게중심을 이용하여 전체 영상의 변화량과 부분적인 특징 변화량을 사용함으로서, 간단한 행동들은 쉽게 인식 되며, 실생활에도 효율적으로 적용될 수 있다.

[참고문헌]

[1] James W. Davis and Aaron F. Bobick, The Representation and Recognition of Action Using Templates, CVPR, 1997.

[2] 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.

[3] 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, 1998.

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

[5] Yoshio IWAI, Tadashi HATA and Masahiko YACHIDA, Gesture Recognition based on Subspace Method and Hidden Markov Model, IEEE, 1997, pp.960-966.

[6] Ismail Haritaoglu, Ross Cutler, David Harwood and Larry S. Davis, Backpack: Detection of People Carrying Objects Using Silhouettes, IEEE International Conference on Computer Vision(ICCV), 1999

[7] Takahiro Watanabe and Masahiko Yachida, Real Time Recognition of Gesture and Gesture Degree Information Using Multi Input Image Sequence, ICPR, 1998

[8] Shigeyoshi Hiratsuka, Kohtaro Ohba, Hikaru Inooka, Shinya Kajikawa, and Kazuo Tanie, Stable Gesture Verification in Eigen Space, LAPR Workshop on Machine Vision Application, 1998.

[9] Christian Vogler, Dimitris Metaxas, ASL Recognition Based on a Coupling Between HMMs and 3D Motion Analysis, ICCV, 1998.

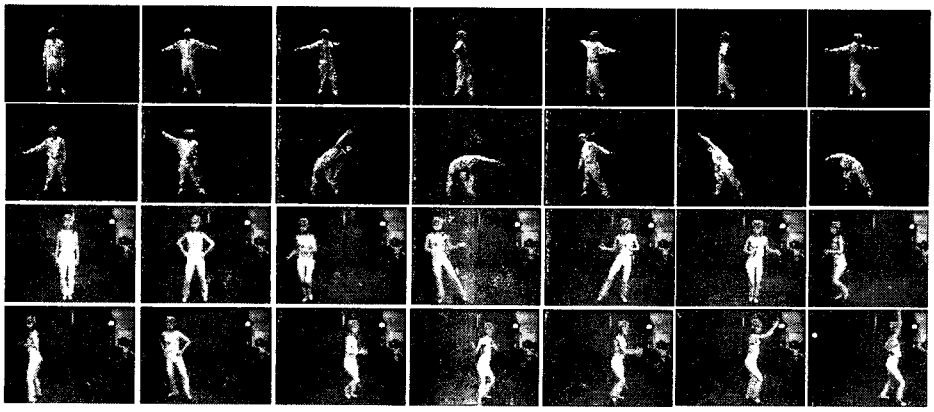

그림 5 실험에 사용된 제스처 이미지 시퀀스

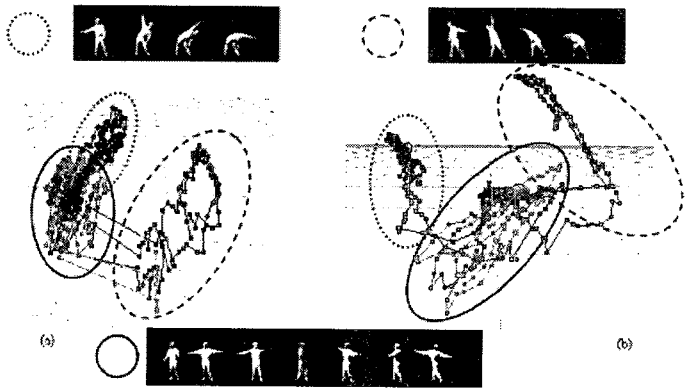

그림 6 (a) 글로벌 모션 정보를 이용한 제스처 공간, (b) 로컬 모션 정보를 이용한 제스처 공간

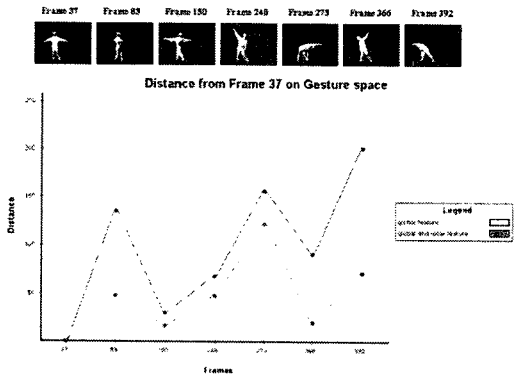

그림 7 제스처 공간에서의 거리 비교 (7~37 frame)