로컬모션정보와 글로벌모션정보를 이용한 제스처인식

이현주[O] 이칠우
전남대학교
leehj@image.chonnam.ac.kr[O], leecw@chonnam.ac.kr

Gesture Recognition using Combination of Local and Global Information

Hyun-Ju Lee[O] Chil-Woo Lee
Dept. of Computer Engineering, Chonnam National University

요 약

본 논문에서는 입력 시퀀스의 각 영상으로부터 신체 영역을 분리한 후 신체 영역의 2차원 특징정보들을 이용하여 제스처를 자동적으로 인식할 수 있는 알고리듬을 제안한다. 먼저, 샘플 영상들로부터 구한 2차원 특징 벡터들의 통계적 정보를 주성분 분석법으로 분석하고 제스처 모델 공간을 구성한다. 입력 영상들은 미리 구성된 모델과 비교되어지고 각각의 영상은 모델 공간의 한 부분으로 심볼화되어진다. 마지막으로 심볼 시퀀스로 형상화되어진 영상 시퀀스는 은닉 마르코프 모델(HMM)을 이용하여 하나의 제스처로 인식된다. 우리가 이용하는 2차원 특징 정보는 대략적으로 신체의 어느 부분이 움직이는지를 알 수 있는 로컬 정보와 전체적인 신체 모션의 정보를 나타내는 글로벌 정보를 이용하는 것으로 실세계에서 적용하기 용이하고, 좋은 인식 결과를 얻을 수 있다.

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 일상 생활에서 제스처, 표정과 같은 비언어적인 수단을 이용하여 수많은 정보를 교환한다. 따라서 자연스럽고 지적인 인터페이스를 구축하기 위해서는 제스처와 같은 비언어적 통신 수단에 대한 연구가 매우 중요하다. 최근에 들어, 대규모 비디오 데이터베이스의 구축, 감시 시스템, 고 압축 통신 시스템의 구축을 위해 제스처 인식에 관한 연구가 활발히 진행되고 있다.

제스처를 인식한다는 것은 인체 각 부위가 시간 축에 대해 어떠한 형상 변화를 가지는가를 자동으로 알아내는 것을 의미한다. 그러나 인체는 매우 복잡한 3차원 관절 구조를 지니고 있어서 자동으로 제스처를 인식한다는 것은 매우 어렵다.

초기에는 인체 각 부위의 관절에 부착된 센서를 통해서 형상 변위 값을 입력하여 시공간적인 형상 패턴을 추출하고 제스처를 인식하였다. 이 방법은 장치를 몸에 붙이는 과정이 복잡하고 초기 교정이 어려울 뿐만 아니라 연결 케이블 때문에 자유스런 제스처 입력이 불가능하여 현재는 거의 사용되지 않는다.

따라서, 최근에는 센서를 이용하지 않고 비디오 카메라를 통하여 얻은 시각 정보로 제스처를 인식할 수 있는 많은 알고리듬들이 개발되어졌고 애니메이션 제작, 영화 제작과 같은 몇몇 응용 분야에서 사용되어지고 있다.

본 논문에서 우리는 입력 시퀀스의 각 영상으로부터 신체 영역을 분리한 후 신체 영역의 2차원 특징 정보들을 이용하여 제스처를 자동적으로 인식할 수 있는 알고리듬을 제안한다. 우리는 먼저, 샘플 영상들로부터 구한 2차원 특징 벡터들의 통계적 정보를 주성분 분석법으로 분석하고 제스처 모델 공간을 구성한다. 입력 영상들은 미리 구성된 모델과 비교되어지고 각각의 영상은 모델 공간의 한 부분으로 심볼화되어진다. 마지막으로 심볼 시퀀스로 형상화되어진 영상 시퀀스는 은닉 마르코프 모델(HMM)을 이용하여 하나의 제스처로 인식된다. 우리가 이용하는 2차원 특징 정보는 대략적으로 신체의 어느 부분이 움직이는지를 알 수 있는 로컬 정보와 전체적인 신체 모션의 정보를 나타내는 글로벌 정보를 이용하는 것으로 실세계에서 적용하기 용이하고, 좋은 인식 결과를 얻을 수 있다. 그림 1은 전체 알고리듬의 간단한 블록 다이어그램을 보여준다.

본 연구는 한국 과학 재단 지정 전남대학교 "고품질 전기 전자 부품 및 시스템 연구센터"의 연구비 지원에 의해 수행되었음.

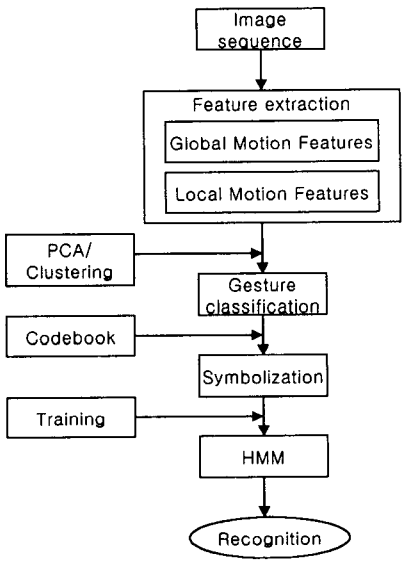

그림 1.제스처 인식과정의 블록다이어그램

2. 전처리와 특징추출

카메라로 얻은 영상 시퀀스는, 일반 환경에서 여러 가지 제스처를 취한 것이다. 따라서, 각 영상에는 제스처 인식에 필요 없는 많은 오브젝트들(배경)이 포함되어 있다. 각각의 영상에서 신체 영역에 해당하는 전경 영역을 배경으로부터 분리하기 위해서는 먼저 배경 모델을 생성해야 한다. 배경 모델은 전경 영

역을 포함하지 않은 영상 시퀀스로부터 계산되어지는 것으로 각 픽셀들이 조명 변화에 의해 갖는 최대 밝기 값, 최소 밝기 값, 최대 밝기 차이 값의 3가지 요소로 구성된다. 이는 조명으로 인해 생길 수 있는 밝기 변화는 무시하고 사람의 신체 영역에 해당하는 영역은 전경 영역으로 분리하는 기준이 된다. 배경 모델(background model : BM)은 식 (1)과 같이 표현되어진다.

$$BM=\{M(x,t),\ N(x,t),\ D(x,t)\} \tag{1}$$

여기서 $M(x,t)$는 화소 x가 시간 t에 의해서 갖는 최소 밝기 값, $N(x,t)$는 화소 x가 시간 t에 의해서 갖는 최대 밝기 값을 나타낸다. $D(x,t)$는 화소 x가 가질 수 있는 최대 밝기 차이 값을 나타낸다.

전경 영역은 식 (2)에 의해서 결정되어진다[1]. 즉 식 (2)를 만족하는 화소 x는 모두 전경 영역으로 세그멘테이션된다.

$$\begin{aligned} |M(x,t)-I(x,t)| &> D(x,t)+C \quad \text{or} \\ |N(x,t)-I(x,t)| &> D(x,t)+C \end{aligned} \tag{2}$$

여기서 $I(x,t)$는 입력 영상이고 C는 상수 값이다.

일단 신체 영역이 추출되어지면, 신체의 포즈와 모션을 효과적으로 표현할 수 있는 여러 가지 특징정보를 추출해야한다. 메릴랜드 대학에서는 MHI (Motion History Image)라는 Hu 모멘트 벡터를 특징정보로 사용하였다[10]. MHI[10]는 더 최근에 움직인 화소들이 더 밝은 값으로 할당되어지는 영상으로 모션 정보가 누적되기 때문에 우발적인 물체의 모션이 있을 경우 문제가 생긴다. 또한 완전한 모션 패턴이 항상 주어져야한다는 제약이 뒤따른다. 따라서 이러한 단점을 보완하고자 MHI 영상으로부터 신체 영역의 실루엣 윤곽선들을 구하고, 모션에 의해서 바뀌는 윤곽선들로부터 방향 정보를 얻어 사용하고 있다[11]. 그러나 이러한 방향 정보만으로 부분적인 신체 움직임을 구별하는데 한계가 있다. 따라서 본 논문에서는, 신체 전체의 모션 정보와 함께 신체의 어느 부분이 움직이고 있는지에 관한 정보를 보다 구체적으로 얻을 수 있는 비교적 간단한 방법을 제안하고자 한다.

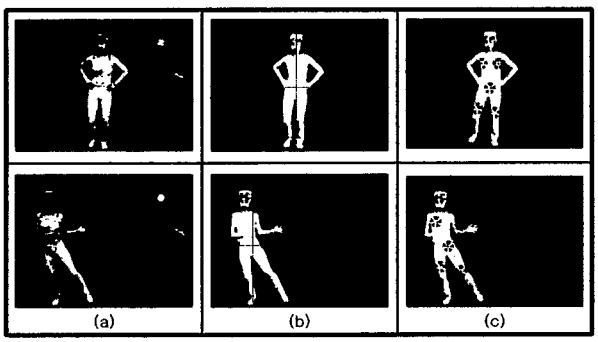

그림 2. 초기 과정의 결과. (a)입력 영상, (b)로컬 모션을 위한 서브 영역, (c)로컬영역의 무게 중심

전체적인 신체 모션을 형상화하기 위해 6가지 특징 정보; 1) 신체 영역의 가로축 길이(Width), 2)세로축 길이(Height), 3)무게 중심의 x좌표, 4)무게 중심의 y좌표, 5)조밀성(Compact-ness), 6)모멘트의 주축 값을 추출한 후, 이들 특징 값의 시간적인 변화량을 계산한다. 이와 같은 특징들의 변화량은 전체적인 외형의 변화만을 설명해 줄뿐, 신체의 어느 부분이 움직이고 있는지는 알 수가 없다. 그러나, 제스처는 때때로 부분적인 신체 모션의 변화에 따라 많은 차이를 보일 수 있다. 그래서, 우리는 정확한 제스처 인식을 위해 신체의 어느 부분이 움직이고 있는지에 관한 로컬 모션 정보가 필요하다. 이를 위해 손이나 발의 정확한 위치를 계측하려면, 너무도 많은 계산 량이 필요하다. 따라서 우리는 매우 간단한 특징 데이터(신체의 서브 영역의 무게 중심 좌표, 서브 영역의 면적)를 도입하여 이 문제를 해결하고자한다. 신체 영역은 무게중심을 기준으로 4개의 서브 블록으로 나누어지고 각각의 영역에 속하는 블랍들의 중심 좌표와 면적의 변화량을 살펴봄으로써 로컬 모션 정보를 얻을 수 있다. 그림 2는 지금까지 설명한 초기 과정의 간단한 예를 보여준다.

우리가 제스처 인식에서 고려해야하는 중요한 점은 모션 히스토리 정보(특징 값들의 시간적인 변화량)이다. 이는 행동이나 제스처가 연속적인 신체의 일부 또는 전체적인 움직임으로 이루어지기 때문이다. 본 논문에서는, 제스처의 시간적인 변화량을 강조하기 위해 영상을 그룹핑하는 방법을 제안한다. 그 개념은 다음과 같이 표현되어질 수 있다.

$$F_t^{t+n} \quad (1\le t\le T-2,\ 0\le n\le 2) \tag{3}$$

식 (3)에서 F는 특징 집합을 의미하고 T는 영상 시퀀스의 총 길이를 나타낸다. 그리고 n+1은 하나의 그룹으로 묶을 영상의 개수, t는 영상 시퀀스에서 그 영상의 위치를 뜻한다. 따라서 세 개의 이웃하는 영상들로부터 얻은 특징 집합이 하나의 시간적 그룹으로 간주되어진다. 우리는 일단, 입력 비디오 시퀀스로부터 신체 영역이 포함된 이진 영상들이 얻어지면, 영상 그룹핑과 특징 추출 과정을 수행한다. 다시 말해 시간 t에서, I_t, I_{t+1}, I_{t+2} 가 하나의 그룹으로 묶여지고 그 때 각각의 영상으로부터 18개의 특징 정보(6개의 전체 외관 특징 정보, 12개의 부분적인 신체 특징 정보)를 얻을 수 있다. 그래서, 한 그룹으로부터 신체 형상의 특징뿐만 아니라 그 특징들의 변화량을 계산하여 모션 정보를 계산할 수 있다. F_{imgt}를 시간 t에서의 특징 집합이라고 정의하자. 식 (3)에서 F_t^t은 시간 t에서의 영상의 특징 집합(F_{imgt}), F_t^{t+1} 은 F_{imgt}와 F_{imgt+1}의 차이를 표현한 것이다. 결국 한 그룹의 전체 특징 정보의 개수는 54가 된다. 그림 3은 시간적인 영상 그룹핑 알고리듬을 그림으로 보여준 것이다.

3. 주성분 분석법과 제스처 공간

주성분 분석법(PCA)은 고차원의 입력 데이터 집합의 차원을 명백하게 줄일 수 있다. 그러나 몇 가지 이유로 데이터의 시공간적인 구조를 찾는데 적합하지 않을 수 있다. 대표적인 이유로, 이 방법은 원래의 데이터 구조가 선형이라는 것을 가정한다. 원래의 데이터가 비 선형이면, 주성분 분석을 통해 분석한 결과는 그 의미가 약하다. 하지만 우리 방법에서는 가능한 한 선형적인 2차원 데이터를 특징으로 취급하였기 때문에 비교적 분석 결과가 양호하므로 입력 특징 벡터의 차원을 줄이기 위해 주성분 분석법을 채택하였다. 주성분 분석법을 적용하기 전에 특징의 집합은 수치적으로 동일한 단위를 가지고 있지 않기 때문에 정규화 과정을 거쳐야 한다.

특징 값들은 식 (4)와 같이 표현되어질 수 있다.

$$x=[x_1,x_2,\cdots,x_N]^T \tag{4}$$

여기서 $N(=T-2)$은 부분적인 신체 모션의 특징 값과 전체적인 외관의 모션 특징 값으로 구성된 그룹의 수를 의미한다.

식 (4)와 같은 특징 집합을 이용하여 신체의 전체적인 외관

특징과 부분적인 모션 특징들을 표현할 수 있는 저차원 벡터 공간, 즉 파라메트릭 고유공간을 생성하고 이를 제스처 공간이라 부른다. 우리는 그림 4로부터 영상 시퀀스가 몇 개의 성분 벡터에 의해서 잘 구분됨을 알 수 있다.

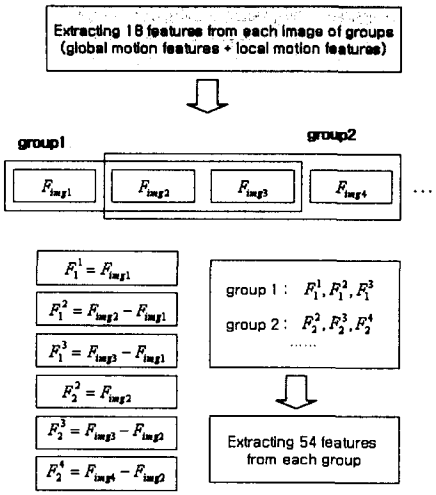

그림 3. 시간적인 영상 그룹핑과 특징 추출

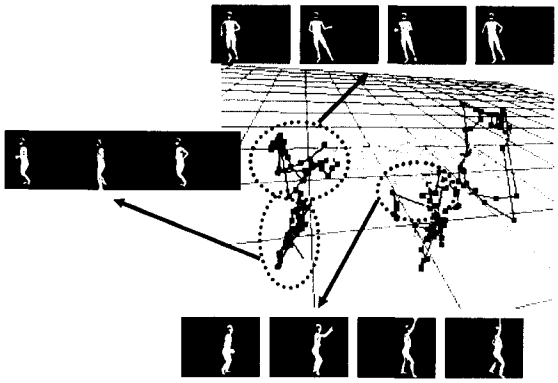

그림 4. 영상 시퀀스의 제스처 공간으로의 투영

4. 심볼을 이용한 제스처 인식

은닉 마르코프 모델(HMM)은 은닉 상태와 관측 가능한 상태로 이루어진 확률적 네트워크의 통계적인 인식 방법이다. 이 방법은 공간적인 개념과 시간적인 개념을 가진 데이터에 적합하다. 따라서 본 논문에서는 HMM을 이용한다.

영상 시퀀스들이 그림 4에서처럼 클러스터링 알고리듬에 의해서 몇 개의 제스처 패턴들로 구분되어지면, 각 제스처 시퀀스들은 심볼 시퀀스로 형상화되어지고 HMM의 입력으로 사용한다.

HMM λ는 다음과 같은 변수들에 의해서 표현되어진다. 상태 천이 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 확률, π_i는 초기 상태 확률 값을 나타낸다. HMM의 학습은 $\{\pi, A, B\}$의 파라미터들을 추정하는 것이다. 우리가 HMM의 추정을 위해 사용하는 알고리듬은 봄-웰치(*Baum-Welch*) 알고리듬이다.

5. 실험과 결론

실험에 사용한 제스처 영상은 걷는 동작, 앉는 동작, 일어서는 동작과 같이 일상 생활에서 우리가 매일 취하는 동작뿐만 아니라 맨손 체조에서 하는 다리 운동, 옆구리 운동, 제자리에서 걷는 운동, 댄스 등의 제스처를 사용하였다. 각 영상의 크기는 320×240을 사용하였고 총 13개의 제스처 시퀀스를 모델로 구성하였다. 그 결과 모델로 구성된 제스처들은 50개의 클러스터로 분류되어졌고 HMM을 통하여 입력 영상에 대한 인식 결과를 확인하였다. 실험을 통해서, 우리는 제안한 알고리듬이 테스트 영상들에 대해 비교적 높은 인식률을 보이고 특히, 우리가 신체 영역에 대해 보다 정확한 세그멘테이션 결과를 얻으면 더 좋은 인식률을 보임을 알 수 있었다.

본 논문에서 사용한 제스처 인식 방법은 에지, 코너와 같은 기하학적인 특징 정보가 아닌 모멘트, 신체 영역의 사이즈, 신체 영역의 무게 중심을 이용하여 전체 외관적인 변화량과 부분적인 신체 움직임의 변화량을 사용하였다. 그러므로 세밀한 모션 정보를 완벽하게 인식되어질 수는 없지만, 간단한 행동들 또는 제스처들은 쉽게 인식되어질 수 있다.

참고 문헌

[1]Ismail Haritaoglu, David Harwood and Larry S. Davis, "W4: Who? When? Where? What? A Real Time System for Detecting and Tracking People", International Conference on Face and Gesture Recognition, 1998

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

[3]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

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

[5]Shigeyoshi Hiratsuka, Kohtaro Ohba, Hikaru Inooka, Shinya Kajikawa, and Kazuo Tanie, "Stable Gesture Verification in Eigen Space", LAPR Workshop on Machine Vision Applicaion, 1998, 17-19

[6]이용재, 이칠우, "외관 기반의 파라메트릭 고유 공간을 이용한 물체인식", 정보과학회, 1999

[7]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, 1995

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

[9]Andrew D. Wilson, Aaron F. Bobick, "Parametric Hidden Markov Models for Gesture Recognition", IEEE Transaction on PAMI, Vol.21, No.9, September 1999

[10]James W.Davis, Aaron F. Bobic, "The Representation and Recognition of Action using Temporal Templates", CVPR, 1997

[11]James Davis, Gary Bradski, "Real-time Motion Template Gradients using Intel CVLib", ICCV, 1999