2000年度 第13回 信號處理合同學術大會 論文集 第13卷 1號

연속 영상에서의 제스처 인식

이 현 주, 이 칠 우
전남대학교 컴퓨터공학과

Gesture Recognition on image sequences

Hyun-Ju Lee, Chil-Woo Lee
Dept. of Computer Engineering, Chonnam Univ.
E-mail : leehj@image.chonnam.ac.kr, leecw@chonnam.ac.kr

요 약

인간은 일상 생활에서 제스처, 표정과 같은 비언어적인 수단을 이용하여 수많은 정보를 전달한다. 따라서 자연스럽고 지적인 인터페이스를 구축하기 위해서는 제스처 인식에 관한 연구가 매우 중요하다. 본 논문에서는 영상 시퀀스의 각 영상들이 가지고 있는 정적인 양이 아닌, 영상과 이웃하는 영상들의 변화량을 수치적으로 측정하고 이를 주성분 분석법(PCA : Principal Component Analysis)과 은닉 마르코프 모델(HMM : Hidden Markov Model)을 이용하여 인식하는 방법을 소개한다.

1. 서론

컴퓨터 기술의 발달과 함께 정보 시스템이 복잡하게 되면서 인간과 정보 시스템 사이에 자연스럽게 정보를 교환할 수 있는 지적 시스템에 관한 관심이 날로 커지고 있다. 인간은 일상 생활에서 제스처, 표정과 같은 비언어적인 수단을 이용하여 수많은 정보를 전달한다. 따라서 자연스럽고 지적인 인터페이스를 구축하기 위해서는 제스처와 같은 비언어적인 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.

본 논문에서는 영상 하나 하나의 기하학적인 특징을 이용하지 않고 연속적인 영상들의 변화량을 특징 값으로 형상화하는 방법을 사용하였다. 이 특징 값들은 주성분 분석법이라는 통계적인 수법에 의해서 신체의 전체적인 외관 특징들을 표현할 수 있는 저차원 벡터 공간, 즉 파라메트릭 고유 공간에 투영된다. 투영된 점들은 클러스터링 알고리듬에 의해서 자동으로 분류하고 은닉 마르코프 모델을 이용하여 인식한다. 전체 시스템의 구성도는 그림 1과 같다.

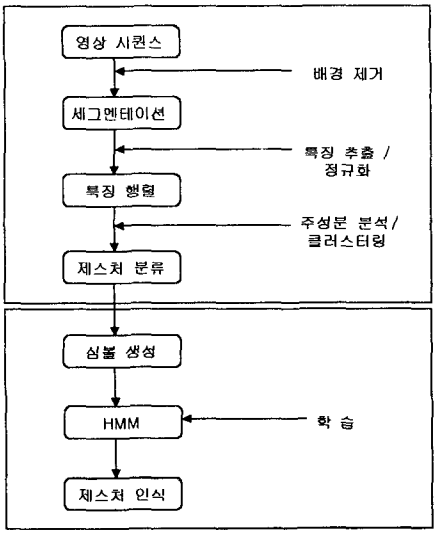

그림 1. 시스템 전체 구성도

2. 영상 시퀀스의 자동 분할

2.1 세그멘테이션

전경 영역(foreground region)을 배경으로부터 분리하기 위해서는 먼저 배경 모델을 생성해야 한다. 배경 모델(background model : BM)은 전경 영역을 포함하지 않은 영상 시퀀스로부터 계산되어지는 것으로 식(1)과 같이 표현되어진다.

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

여기서 $M(x,t)$는 화소 x가 시간 t에 의해서 갖는 최

소 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.

$$|M(x,t)-I(x,t)| > D(x,t)+C \quad \text{or}$$
$$|N(x,t)-I(x,t)| > D(x,t)+C \qquad (2)$$

여기서 $I(x,t)$는 입력 영상이고 C는 상수 값이다.

2.2 특징 추출

제스처들을 분류하기 위해 사용한 특징 값은 신체 영역의 가로축 길이(FeretX)와 세로축 길이(FeretY)의 비를 나타내는 페렛비(Feret_ratio), 무게 중심의 x좌표, 무게 중심의 y좌표, 조밀성(Compactness), 모멘트의 주축, 모멘트 주축의 수직인 축으로 총 6가지이다.

이들 특징 값은 그림 2에서 보여주는 것처럼 입력 영상이 n개가 들어왔을 때, 식 (3)에 의해 구해진 각 영상에 대해 계산되어진다.

$$I_t,\ I_t - I_{t+1},\ I_t - I_{t+2},\ \ldots\ldots\ I_t - I_{t+n-1} \qquad (3)$$

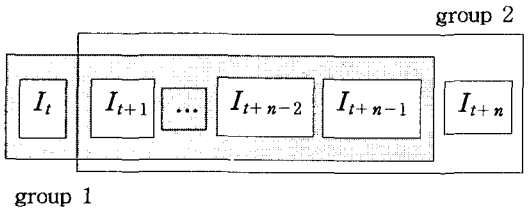

그림 2. 시간에 따른 영상 그룹화

그림 3. 시간에 따른 영상 그룹화의 예
(n=6일 때, 걷는 동작의 group 1)

즉, 시간에 따른 움직임의 변화량을 특징 값에 의해서 형상화한 것이다. 특징의 집합 x는 식 (4)와 같다.

$$x=[x_1, x_2, x_3, \ldots, x_N]^T \qquad (4)$$

여기서 N은 n개의 영상으로 구성된 그룹의 수로 전체 영상 시퀀스의 수가 T일 때 N은 T- n + 1의 값과 같다. 그리고 $x_i(i=1,\ldots,N)$은 M(= n × 6)개의 특징 값으로 구성된다. 그러나 특징의 집합을 그대로 사용하게 되면 수치적으로 동일한 단위를 가지고 있지 않기 때문에 정규화 과정을 거쳐야 한다.

2.3 주성분 분석

2.2절과 같이 특징 집합을 이용하여 신체의 전체적인 외관 특징들을 표현할 수 있는 저차원 벡터공간, 즉 파라메트릭 고유공간을 생성한다. 고유공간을 계산하기 위해서는 먼저 모든 특징 벡터에서 평균 벡터를 구하여 각 특징들과의 차를 구한다. 평균 벡터 c와 새로운 특징 집합 X를 식 (5)와 식 (6)과 같이 나타낸다[6].

$$c=(1/N)\sum_{i=1}^{N} x_i \qquad (5)$$

$$X \triangleq [x_1-c, x_2-c, x_3-c, \ldots, x_N-c]^T \qquad (6)$$

고유공간을 구하기 위해서는 M × N의 크기를 지닌 특징 집합 X를 식 (6)과 같이 계산하고 식 (7)을 만족하는 고유벡터를 구하면 된다[6]. 즉, 공분산 행렬 Q에 대한 고유치 λ와 고유벡터 e를 구한다.

$$Q \triangleq XX^T \qquad (7)$$

$$\lambda_i\, e_i = Q\, e_i \qquad (8)$$

고유치 분해를 위하여 특이치 분해(singular value decomposition)을 이용한다. 특이치 분해를 이용하면 특징 집합 X의 공분산 행렬에 대한 고유벡터를 쉽게 얻을 수 있다[6]. 이제 얻어진 고유공간에 평균 벡터 c에서 뺀 특징 집합 x를 모두 식 (9)을 이용하여 투영시킨다.

$$m_i=[e_1, e_2, e_3, \ldots, e_k]^T\,(x_i-c) \qquad (9)$$

2.4 클러스터링

파라메트릭 제스처 공간에 투영된 점들을 클러스터링 알고리듬에 의해서 분류할 때, 영상 시퀀스 안에 몇 개의 제스처 패턴들이 존재하는지를 알 수 없기 때문에 클러스터의 개수를 몇 개로 지정할 것인가에 관한 문제가 발생하게 된다. 본 논문에서는 다변량 분산분석법[7]을 이용하여 이런 문제를 해결하였다. 즉, 관찰 값들의 클러스터 내, 그리고 클러스터 사이의 흩어진 정도를 분산의 항목으로 측정하고 '클러스터 간의 분산'이 상대적으로 '클러스터 내 분산'보다 충분히 큰 클러스터 개수를 채택하면 되는 것이다. B를 클러스터 간 분산, W를 클러스터 내의 분산이라고 하면 이는 각각 식 (10), 식 (11)과 같다.

$$B=\sum_{i=1}^{g} N_i(\overline{X_i}-\overline{X})(\overline{X_i}-\overline{X})^T \qquad (10)$$

$$W=\sum_{i=1}^{g}\sum_{j=1}^{N_i}(X_{ij}-\overline{X_i})(X_{ij}-\overline{X_i})^T \qquad (11)$$

그리고 전체 분산 T, 클러스터 간의 분산과 전체 분산의 비(Λ)는 식 (12)와 식 (13)에 의해서 구할 수 있다.

$$T = B + W = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (X_{ij} - \overline{X})(X_{ij} - \overline{X})^T \quad (12)$$

$$\Lambda = \frac{|B|}{|B+W|} \quad (13)$$

여기서 작은 Λ값은 클러스터 내의 변동이 전체(클러스터 간)의 변동에 비해 상대적으로 작다는 것을 의미하며, 이는 클러스터간의 차이가 유의한지를 판정하는데 통계적 근거가 된다.

클러스터링을 통해 영상 시퀀스가 여러 개의 제스처로 분류되어지면, 각 제스처 시퀀스들은 심볼 시퀀스로 형상화되어지고 HMM의 입력으로 사용한다.

3. HMM을 이용한 제스처 인식

3.1 은닉 마르코프 모델

은닉 마르코프 과정은 다음의 5개 요소로 정의되어진다.

S : 상태의 유한 집합 ; $S=\{s_i\}$

Y : 출력 심볼의 집합

A : 상태 천이 확률의 집합 ; $A=\{a_{ij}\}$

a_{ij} : 상태 s_i에서 s_j로 천이할 확률

$\sum_j a_{ij} = 1$

B : 출력 확률의 집합 ; $B=\{b_{ij}(k)\}$

b_{ij} (k) : 상태 s_i에서 s_j 로 천이할 때 심볼 k를 출력할 확률

$\sum_k b_{ij}(k) = 1$

Π : 초기 상태 확률의 집합 ; $\Pi=\{\pi_i\}$

π_i : 초기 상태가 s_i일 확률

$\sum_j \pi_j = 1$

3.2 제스처 모델의 학습

은닉 마르코프 모델의 각 제스처 모델{ π, A, B }은 Baum-Welch 알고리듬에 의해서 추정되어지고 식(14)-식(15)에 의해서 계산된다[2].

$$\xi_t(i,j) = \frac{P(s_t=i,\ s_{t+1}=j,\ Y|\lambda)}{P(Y|\lambda)} = \frac{\alpha_t(i)\, a_{ij}\, b_j(y_{t+1})\, \beta_{t+1}(j)}{P(Y|\lambda)} = \frac{\alpha_t(i)\, a_{ij}\, b_j(y_{t+1})\, \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)\, a_{ij}\, b_j(y_{t+1})\, \beta_{t+1}(j)} \quad (14)$$

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \quad (15)$$

여기서 $\xi_t(i,j)$ 는 시간 t에서는 상태 i, 시간 t+1에서는 상태 j일 확률이고 $\gamma_t(i)$는 전체 관측 시퀀스와 λ가 주어졌을 때, 시간 t에서 상태 i일 확률을 나타낸다. 식 (14), 식 (15)를 이용하여 제스처 모델은 식 (16), 식 (17), 식 (18)으로 추정되어진다.

$$\overline{\pi_j} = \gamma_1(i) \quad (16)$$

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (17)$$

$$\overline{b_j(k)} = \frac{\sum_{t=1,\ s.t.\ o_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \quad (18)$$

3.3 제스처 인식

입력의 심볼 시퀀스(Y)가 주어지면 모델 λ_i에 대한 확률 값은 forward 변수인 $\alpha_t(i)$와 backward 변수인 $\beta_t(i)$를 이용하여 식 (19)와 같이 구하고 가장 높은 확률 값을 갖는 모델로 인식하게 된다.

$$P(Y|\lambda_i) = \sum_i \sum_j \alpha_t(i)\, a_{ij}\, b_{ij}(y_{t+1})\, \beta_{t+1}(j) \quad (19)$$

4. 실험 및 결론

실험에 사용한 제스처 영상은 걷는 동작, 앉는 동작, 일어서는 동작과 같이 일상 생활에서 우리가 매일 취하는 동작 뿐만 아니라 맨손 체조에서 하는 다리 운동, 옆구리 운동, 제자리에서 걷는 운동 등의 제스처를 사용하였다. 각 영상의 크기는 320×240을 사용하였고 총 8개의 제스처 시퀀스를 모델로 구성하였다. 그 결과 모델로 구성된 제스처들은 40개의 클러스터로 분류되어졌고 HMM을 통하여 입력 영상에 대한 인식 결과를 확인하였다. 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 일부가 다른 시퀀스로 인식하는 오류를 범하기도 하였다. 그 원인들을 분석해 본 결과 우리가 사용한 특징 값들은 전체적인 외형의 변화 량 만을 보고 있기 때문에 한 쪽 팔을 올리는 포즈와 다리 하나를 올리는 포즈들을 같은 포즈들로 분류한 것을 알 수 있었는데, 이는 앞으로 새로운 변화량 값을 특징으로 사용함으로써 보완해 나갈 계획이다. 우리가 실험에 사용한 데이터만 가지고는 입력되는 모든 영상을 분류해 내기란 무척 어려운 일이다. 따라서 우리가 사용한 방법이 보다 일반

성을 갖기 위해 가능한 모든 영상을 수집하여 분석해 나가야 할 것이다.

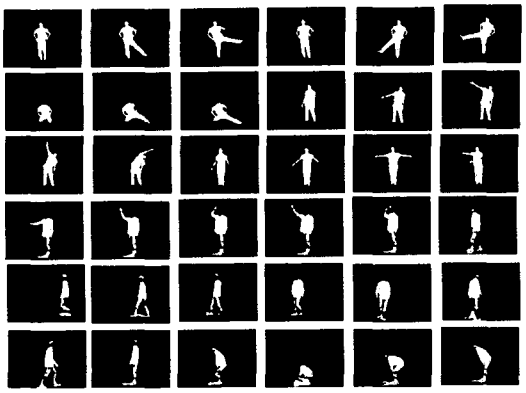

그림 4. 제스처 시퀀스의 일부 포즈들

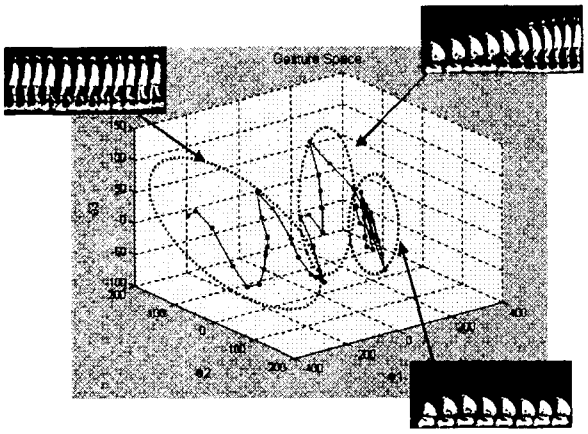

그림 5. 제스처 공간에서 입력 영상의 매핑
(걷다가 앉고 다리 운동 하는 시퀀스)

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, pp. 14-16

[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]김기영, 전명석, "다변량 통계 자료 분석", 자유 아카데미

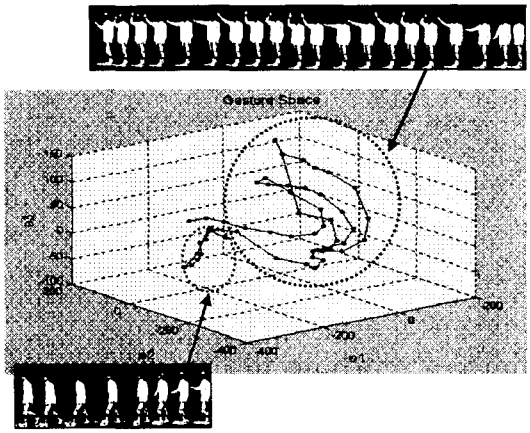

그림 6. 제스처 공간에서 입력 영상의 매핑
(손을 흔들다가 걷는 시퀀스)

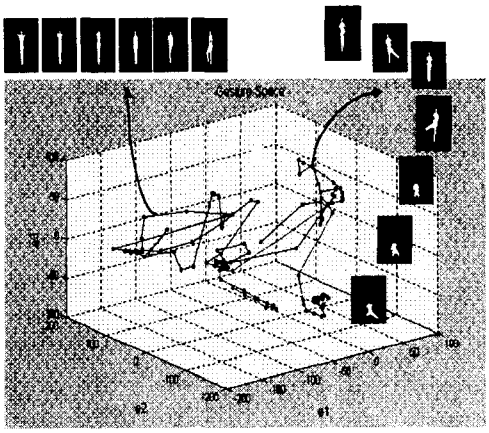

그림 7. 두 개의 영상 시퀀스를 제스처 공간에 투영한 결과