저차원 특징 공간에서 HMM을 이용한 제스처 인식

이용재, 이칠우
전남대학교 컴퓨터공학과

Gesture Recognition Using HMM on Feature Subspace

Yong-Jae Lee, Chill-Woo Lee
Dept. of Computer Engineering, ChonNam National University

요 약

본 논문에서는 연속적인 인간의 제스처 영상을 저차원 제스처 특징 공간과 HMM 이용하여 인식 할 수 있는 방법에 대해 소개한다. 일반적으로 제스처 공간에서 모델 패턴들과 매칭하기 위해서는 모든 모델 영상과 연속적인 입력영상들간의 거리평가로 인식을 수행하게 된다. 여기서 제안한 방법은 연속성을 가진 모델영상들을 HMM로 구성하여 포즈들의 시공간적 특성을 매칭에 이용하였다. 이 방법은 동작의 구분뿐만 인식결과를 학습에 이용할 수 있는 장점이 있다.

1. 서론

최근 인간의 제스처를 인식하기 위한 연구가 활발하게 진행되고 있으며 인식방법에 대한 여러 가지 접근법이 시도되고 있다. 영상을 이용한 제스처 인식방법은 크게 몇 가지로 나뉘어 지는데 그중 대표적인 것으로는 인간의 몸에 직접 마커를 붙여 인식하는 방법과 카메라로 촬영된 인간의 움직임 영상을 분석하여 인식하는 방법이 있다. 특히 연속적인 영상을 제스처 인식에 이용하는 방법은 데이터 글로브(Glove)같은 장치를 직접 몸에 부착하지 않고도 동작을 분석할 수 있고, 복잡한 계산이나 기하학적 특징점들이 필요하지 않다는 장점 때문에 많이 활용되고 있다.

시각기반의 제스처 인식 방법은 크게 2차원 영상기반과 3차원 정보를 이용하는 방법으로 구분된다. 2차원 영상을 이용하는 방법에서 Haritaoglu [1],[2]는 정확한 세그멘테이션을 통해 얻어진 실루엣 영상을 분석하고 인식에 이용하였다. 이 방법은 배경모델을 구성하여 배경과 인간 신체 영역을 분리하고 추적할 뿐 아니라 인간 동작의 주기성을 이용하여 감시시스템에 적용에 용이하였다.

Watanabe [3]는 저차원 특징 공간을 이용하여 실시간 제스처 인식시스템을 구현하였다. 이 방법은 고유 공간에서 모델 제스처영상과 입력 제스처 영상을 비교하여 동작을 구분하고 빠르기와 크기를 평가할 수 있는 방법이다.

Andrew D. Wilson [4]는 머리와 손의 위치에 대한 3차원 정보를 HMM으로 구성하여 스테레오 기반 가상 시스템을 구현하였다.

본 논문에서 제안하는 방법은 주성분 분석을 이용한 외관기반 인식법(Appearance-Based Recognition)과 확률 모델 구성 방법인 HMM(Hidden Marcov Model)을 이용하여 제스처를 인식하는 것이다. 이 방법은 제스처 영상의 형상정보를 저차원 특징 공간으로 투영하여 연속적인 심볼로 구성한 다음, 은닉 마르코프 모델방법을 이용하여 제스처를 인식하는 방법이다. 각각의 포즈는 독립적인 영상의 공간적 특징을 나타내기 때문에 포즈간 매칭시 손, 발의 구분이 필요하며 비슷한 포즈일 경우 다른 동작으로 오인하는 문제가 발생한다. 이는 모델 영상들의 연속적인 관계를 HMM을 이용하여 모델을 구성하면 자연스런 동작간 매칭을 할 수 있다. 먼저, 제스처 공간(GS : Gesture Space)과 템플리트 제스처 점들의 집합을 정규화 된 영상으로부터 얻게 된다. 정규화 된 영상은 안정적인 인간의 행동 패턴을 얻기 위해 배경과 각 사람간의 차이를 제거한 실루엣 영상을 사용하게 된다. 여기서

구해진 실루엣 입력 영상은 의미 있는 고유값을 가진 GS로 투영한다. 마지막으로 투영된 영상을 클러스터링 방법을 이용하여 특징 심볼로 구성한 다음 은닉 마르코프 모델을 이용하여 제스처를 인식하게 된다. 그림 1.은 본 논문의 전체 구성도를 나타내었다.

이 방법은 포즈간 매칭 시 발생하는 지역적 오류를 시간적 모델 방법을 통해 개선할 뿐만 아니라 인식 결과를 학습에 이용 할 수 있다는 장점이 있다.
본 논문은 다음과 같은 순서로 구성되어 있다. 먼저 2장에서는 제스처 공간구성에 필요한 영상을 획득하고 정규화에 관한 내용과 제스처 공간을 구성하는 방법에 대해 소개한다. 3장에서는 제스처 공간에 투영된 모델 영상들로부터 특징심볼을 구하여 은닉 마프코프 모델을 구성하는 방법과 제스처 인식에 관한 내용에 대해 설명한다. 4장에서는 실험 결과, 5장에서는 결론 및 향후 연구 진행 방향에 대해 고찰하고자 한다.

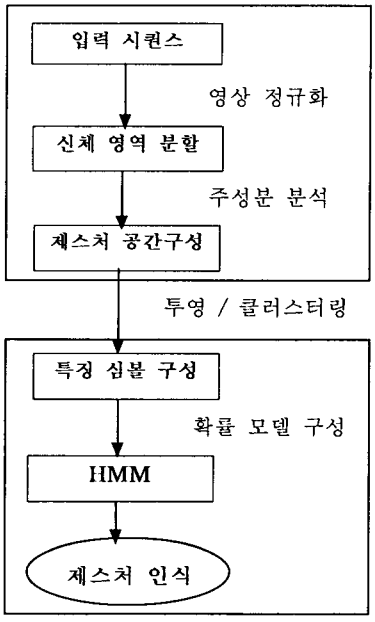

그림 1. 전체 구성도

2. 제스처 공간의 구성

2.1 제스처 영상집합의 정규화

이 절에서는 제스처 영상의 획득 방법과 정규화 및 벡터적 표현에 대해 설명한다. 먼저 카메라로 인간의 동작을 촬영한 다음 미리 촬영된 배경영상과의 차로써 제스처 영상만을 세그멘테이션 한다. 이 영상을 각 개인간의 동작의 차이와 배경의 노이즈를 제거하고 동작성 만을 강조하기 위해 이진화 처리한다. 구해진 이진 영상을 위치 변화에 안정적인 인식을 위해 화면의 중심으로 이동시킨다. 그림. 2는 원 영상과 이상에서 설명한 방법을 통하여 얻어진 이진 영상을 나타낸다.

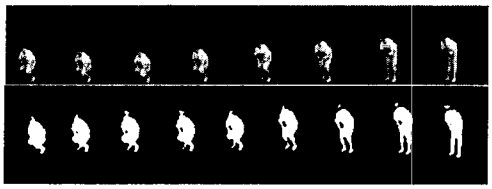

그림 2. 원 영상과 정규화 영상

2.2. 주성분 분석법을 이용한 제스처 공간의 구성

2.1절과 같이 정규화 과정을 거친 영상 집합을 이용하여 제스처의 전체적인 외관 특징들을 표현할 수 있는 저차원 벡터공간, 즉 파라메트릭 제스처 공간을 생성한다. 제스처 공간을 계산하기 위해서는 먼저 모든 영상 x_N 의 식(1)을 이용하여 평균영상 c 를 구하여 식(2)와 같이 각 영상들과의 차를 구한다. 여기서 M × N의 크기를 지닌 영상집합 X를 식(3)과 같이 계산하고 식(4)를 만족하는 고유벡터를 구하면 된다. 즉, 공분산 행렬 Q 에 대한 고유치 λ와 고유벡터 e를 구한다.

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

$$X \triangleq [x_1-c, x_2-c, x_3-c, \ldots, x_N-c]^T \qquad (2)$$

$$Q \triangleq XX^T \qquad (3)$$

$$\lambda_i e_i = Q e_i \qquad (4)$$

여기서 M은 한 영상의 픽셀 수이고 N은 전체 영상의 개수를 나타내는 정수이다. 고유치와 고유벡터를 구하는 데는 특이치 분해를 이용한다. 특이치 분해를 이용하면 영상집합 X의 공분산 행렬에 대한 고유 벡터를 고유치가 큰 순서대로 구할 수 있다.

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{N}\lambda_i} \geq T_1 \qquad (5)$$

여기서 T_1는 고유벡터의 개수를 조정하는 임계치이며, 제

스처 인식과 등장 인물의 포즈 평가 시 이용되는 고유벡터 $\{e_i \mid i = 1,2,..,k\}$는 저차원으로 구성하기 위해 $k \ll N$ 을 만족시킨다. 본 논문에서는 $k = 3$을 이용했다.

제스처 공간에 평균 영상 c에서 뺀 영상 집합 x를 모두 식(6)을 이용하여 투영시킨다.

$$m_n = [e_1, e_2, e_3, \ldots, e_k]^T (x_n - c) \qquad (6)$$

투영시킨 결과는 이산적인 점들로 표현되며, 이 점들은 각 영상을 의미하게 된다. 연속적인 점들은 서로 연관성이 많기 때문에 제스처 공간으로 투영시킨 결과는 서로 깊은 상관 관계를 가진다.

식(7)와 같이 각 제스처들은 서로 관계 있는 연속성을 가진 점들의 집합으로 나타나게 된다.

$$m(m_1, m_2, \ldots, m_n) \qquad (7)$$

제스처 공간에서 이산적인 점들은 모델로 구성되어 입력 제스처 영상과 거리 평가로 모델과 매칭을 수행한다. 하지만 단순히 가까운 거리의 포즈를 인식하게 되면 다른 동작의 같은 포즈끼리 매칭이되는 되어 매칭율에 영향을 주게 된다.

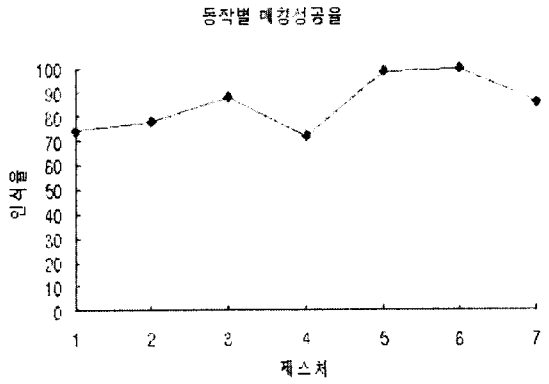

그림 3. 거리 평가를 이용한 동작별 인식 결과

구분	팔	뛰기	나리	앉기	다리 펴기	엎드리기	걷기
입력1	74	14	3	0	0	0	5
입력2	16	78	0	0	0	0	6
입력3	6	1	88	0	0	0	6
입력4	0	0	0	72	28	0	0
입력5	0	0	0	1	99	0	0
입력6	0	0	0	0	0	100	0
입력7	6	2	6	0	0	0	86

표 1. 동작별 인식 결과 테이블

그림 3은 거리 평가로 수행된 매칭 결과를 그래프로 나타낸 것이다. 표1은 동작별 매칭결과를 나타낸 것이다. 표1에서처럼 비슷한 포즈를 포함하고 있는 동작끼리의 잘못된 매칭을 수행한다는 것을 알 수 있다.

2.3 클러스터링을 이용한 특징 심볼 구성

파라메트릭 제스처 공간에서 효과적인 특징 심볼을 구성하기 위해 트리 구조로 된 클러스터링을 이용하여 계층적으로 모델 영상들을 구성한다.

Cm에서 모든 영상을 대표하는 각 클래스 $Cm(m=1,2,...,M)$의 평균 영상을 만들기 위해 비슷한 특징과 인접한 파라메터를 가지는 영상들로 분류해야 한다. 하나의 제스처는 R개의 영상으로 이루어지고 클러스터 Cm으로 나누어 진다고 하자. 이것은 $tm \sim tm+1$ 의 범위 내로 제한된 Cm에서 이미지와 일치하는 t에 따라 구성된다. 계층적 모델 구성으로 식별과 최소 에러에 대해 최적인 클래스 경계를 결정하는 것이다. 고유공간에서 학습이미지 gr의 개수인 R은 t의 순서에 따른 개수이고 Cm으로 분류된다. 첫 번째 경계를 g_1 이라 정하자. 그리고 $Km = r$ 은 g_r에서 정한 m 번째 경계를 표시한다. 경계 Km 과 클래스 Cm은 다음과 같이 표현된다.

$$1 = k_1 < k_2 < k_3 \ldots < k_M \le R \qquad (8)$$

$$C_m = [k_m, k_{m+1} - 1] (m = 1,2,3,\ldots,M) \qquad (9)$$

$K_{m-1} - 1 = R$. 모든 gr 평균 영상과 분산은 Km 에 대해 독립이다.

$$h_T = \frac{1}{R} \sum_{r=1}^{R} g_r \ , \ \sigma_t^2 \frac{1}{R} \sum_{r=1}^{R} ||g_r - h_T||^2 \qquad (10)$$

Cm 은 gr을 포함할 확률 w_m을 가진다.

$$w_m = \frac{1}{R} \sum_{r \in C_m} 1 \qquad (11)$$

Cm에서 gr의 평균영상과 분산은 다음과 같다.

$$h_m = \frac{1}{R w_m} \sum_{r \in C_m} g_r \qquad (12)$$

경계 Km은 Cm의 분류에 의해 계산된다.

$$\sigma_m^2 = \frac{1}{R w_m} \sum_{r \in C_m} ||g_r - h_m||^2 \qquad (13)$$

$$\lambda_M(k_1, \ldots, k_M) = \frac{\sigma_B^2(k_1, \ldots, k_M)}{\sigma_W^2(k_1, \ldots, k_M)} \qquad (14)$$

$$\sigma_W^2(k_1, \ldots, k_M) = \sum_{m=1}^{M} w_m \sigma_m^2 \qquad (15)$$

$$\sigma_B^2(k_1, \ldots, k_M) = \sum_{m=1}^{M} w_m ||h_m - h_T||^2 \qquad (16)$$

$$\sigma_W^2 + \sigma_B^2 = \sigma_T^2 \quad (17)$$

결과적으로 식(14)는 다음 식과 동등하다.

$$\eta_M(k_1, \ldots, k_M) = \sigma_B^2(k_1, \ldots, k_M)/\sigma_T^2 \quad (18)$$

이러한 이유 때문에 우리는 Km를 구할 수 있다.

$$\eta_M^*(k_1^*, \ldots, k_M^*) =$$

$$\max_{1 \le k_m \le R} \sigma_B^2(k_1, \ldots, k_M)/\sigma_T^2 \quad (19)$$

Km^* 은 식별뿐만 아니라 최소 에러에 대한 최적의 경계가 된다. 각 레벨에서 적당한 클래스 M^* 의 개수는 식(20)에 의해 결정된다.

$$Q(M) = \eta_M^* / \bar{\eta}_M^* \quad (20)$$

$$Q(M^*) = \max_{2 \le k_m \le R} Q(M) \quad (21)$$

여기서 R은 분류된 이미지의수이고 M은 클래스의 개수이다.

$$\bar{\eta}_M^* = 1 - \frac{(\frac{R}{M})^2 - 1}{R^2 - 1} \quad (22)$$

최적의 클래스 개수와 경계를 구하여 모델 영상을 계층적 클러스터로 구성하여 각 클래스의 중심이 되는 대표 모델을 아래와 같이 나타내었다.

여기서 m은 결정된 클래스의 개수를 나타낸다.

$$S(M) \equiv [s_1, s_2, s_3, \ldots, s_m] \quad (23)$$

구해진 클래스를 이용하여 HMM의 입력 특징 심볼로 이용하였다.

3. HMM을 이용한 제스처 인식

3.1 은닉 마르코프 모델

클러스터링을 통해 식 (23)과 같이 영상 시퀀스가 여러 개의 제스처로 분류되어지면, 각각의 클러스터 J_i는 코드북(code book)에 의해서 심볼로 형상화되어지고 이는 은닉 마르코프 모델의 입력으로 사용된다.

은닉 마르코프 모델 λ는, 상태 s_i에서 상태 s_j로 천이될 확률을 a_{ij}, 상태 s_i에서 시작할 초기 확률을 π_i, 상태 s_i에서 s_j로 천이할 때 심볼 y를 출력할 확률을 $b_{ij}(y)$로 정의된다. 각 제스처 모델 $\lambda_i(\pi, A, B)$는 보움-웰치(Baum-Welch) 알고리듬에 의해서 추정되어지고 식 (24)와 식 (25)에 의해서 계산된다[5].

$$\xi_t(i,j) = \frac{P(s_t = i, s_{t+1} = j, Y|\lambda)}{P(Y|\lambda)}$$

$$= \frac{\alpha_t(i)\, a_{ij}\, b_j(y_{t+1})\, \beta_{t+1}(j)}{P(Y|\lambda)}$$

$$= \frac{\alpha_t(i)\, a_{ij}\, b_j(y_{t+1})\, \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)\, a_{ij}\, b_j(y_{t+1})\beta_{t+1}(j)} \quad (24)$$

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \quad (25)$$

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(24), 식(25)를 이용하여 제스처 모델은 식(26), 식(27), 식(28)으로 추정되어진다.

$$\bar{\pi}_j = \gamma_1(i) \quad (26)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (27)$$

$$\bar{b}_j(k) = \frac{\sum_{t=1,\ s.t.\ o_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \quad (28)$$

3.2 제스처 인식

입력의 심볼 시퀀스(Y)가 주어지면 모델 λ_i에 대한 확률 값은 전방(forward) 변수인 $\alpha_t(i)$와 후 방(backward) 변수인 $\beta_t(i)$를 이용하여 식 (29)와 같이 구하고 가장 높은 확률 값을 갖는 모델로 인식하게 된다.

$$P(Y|\lambda_i) = \sum_i \sum_j \alpha_t(i)\, a_{ij}\, b_{ij}(y_{t+1})\beta_{t+1}(j) \quad (29)$$

4. 실험결과

실험에 이용된 제스처 영상은 한 사람의 간단한 맨손체조를 정상적인 속도로 수행한 것을 연속적인 영상으로 획득하였다. 각 동작들이 서로 구분 될 수 있는 팔, 다리, 다리 펴기, 뛰기, 걷기, 팔굽혀펴기 등의 구분 운동을 촬영하고 크기 정규화를 이용하여 50 × 50 영상으로 변환하였다. 영상집합의 고유벡터를 계산한 후 재구성된 영상을 가장 복원하는 3차원의 벡터를 선택하여 제스처 공간으로 구성하였다. 따라서 50 × 50 = 2500차원의 이미지가 3차원으로 압축되는 효

과도 거둘 수 있었으며 지역적 매칭 오류를 확률 모델 구성을 통해 비교적 정확한 동작별 매칭을 수행할 수 있었다. 그림 4는 3차원 제스처 공간에서 각 영상들이 맵핑되는 결과를 나타냈다.

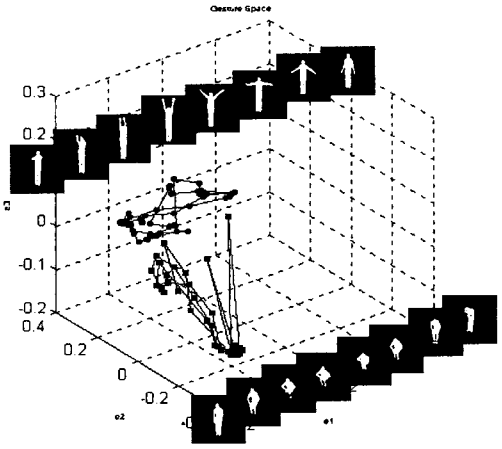

그림 4. 3차원 공간에서 두 입력 영상의 맵핑

5. 결론

본 논문에서 파라메트릭 고유공간 방법과 확률 모델 구성 방법인 HMM을 이용하여 제스처를 인식하는 방법을 제안했다. 제안한 방법은 포즈간 매칭 시 발생하는 지역적 오류를 시간적 모델 방법을 통해 개선할 뿐만 아니라 인식 결과를 학습에 이용 할 수 있다는 장점이 있다. 그러나 다른 포즈라도 비슷한 공간적 변화가 있을 경우에는 애매한 인식결과를 나타내었으며 잡음이 심한 배경에서는 사람 영역만을 세그멘테이션 하는데 어려움이 있었다. 이러한 문제점을 해결하여 보다 안정적인 제스처 인식 알고리듬을 개발할 계획이다.

[참고문헌]

[1] Ismail Harigagolu, Ross Cutler, David Harwood and Larry S. Davis, "Backpack:Detection of People Carrying Objects Using Silhouettes", ICCV99, Vol 2, 1999.

[2] 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

[3] Takahiro Watanabe and Masahiko Yachida, "Real Time Recognition of Gesture and Gesture Degree Information Using Multi Input Image Sequences", ICPR '98, Vol 2, p.185-1858,

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

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

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

[7] Toru Abe, Tomohiko Nakamura "Hierarchical Dic-tonary Constructing Method for the Parametric metric Eigenspace Method" MVA '98, IAPR Workshop on Machine Vision Applications, Nov, 17-19, 1998, Makuhari, Chiba, Japan

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