제스처 공간에서 클러스터링 방법을 이용한 제스처 동작 평가

이용재, 이칠우
전남대학교 컴퓨터공학과

Gesture Motion Estimate Using Clustering Method on Gesture Space

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

요 약

본 논문에서는 저차원 제스처 특징 공간에서 연속적인 인간의 제스처 영상을 계층적 클러스터링을 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 모델영상들을 연속성을 가진 클러스터로 분류하여 입력 영상과 계층적으로 비교할 수 있으며 동작에 관한 구체적 정보를 얻을 수 있다. 이 방법은 매칭 속도와 인식률을 개선하고 인식결과를 학습에 이용할 수 있는 장점이 있다.

1. 서론

최근 연속적인 영상으로부터 인간의 움직임을 인식하기 위한 많은 접근이 시도되고 있다. 일반적으로 가장 많이 사용되는 방법 중에 하나는 손, 무릎, 발, 머리와 같은 신체의 일부에서 몇 가지 특징 점들을 검출하여 이용하는 기하학적 기반의 방법이 있다. 하지만 이러한 방법들은 특징 점들을 검출하기 위해 신체의 각 관절에 마커를 부착하거나 복잡한 계산이 필요하게 된다.

본 논문에서 제안하는 방법은 PCA(Principal Component Analysis)를 이용한 외관기반 인식법(Appearance Based Recognition)의 한 부류에 속한다. 이 방법은 인간의 제스처 영상을 저차원 특징 공간으로 투영하여 어떤 영상의 구체적인 특징을 고려하지 않고 인식할 수 있는 방법이다. 먼저, 제스처 공간(GS : Gesture Space)과 템플리트 제스처 점들의 집합을 정규화된 영상으로부터 얻게 된다. 정규화된 영상은 안정적인 인간의 행동 패턴을 얻기 위해 배경과 각 사람간의 차이를 제거한 실루엣 영상을 사용하게 된다. 그 다음, 입력 영상은 의미 있는 고유값을 가진 GS로 투영되게 된다. 마지막으로 모델 영상과 입력 영상간의 거리 비교를 통해 동작을 인식하게 된다.

여기서 입력 영상은 모든 제스처 모델과 유사도를 구해야만 된다. 그러므로, 실제 실시간 시스템에 적용하기 위해서는 매칭 처리 시간과 전체 모델 영상의 수를 줄여야만 한다.

본 논문은 효과적인 매칭을 위해 계층적인 모델 구성 방법을 소개한다. 제스처 공간에서 계층적으로 모델들을 클러스터링하여 구성한다. 그리고 각 클래스의 평균 영상을 대표 영상으로 이용한다. 모델 영상들은 서로 시간 변화에 따라 연관성이 있으므로 최적의 클래스 경계와 클래스의 수를 연속적인 방법으로 결정하게 된다. 분류 결과로부터 트리 구조로 된 모델이 구성되며 각 노드들은 클래스의 평균 영상이 된다. 입력 영상은 각 노드들과 매칭 하여 가장 가까운 거리의 클래스로 인식하게 된다.

이 방법은 매칭 속도와 인식률을 개선할 뿐만 아니라 인식 결과를 학습에 이용 할 수 있다는 장점이 있다.

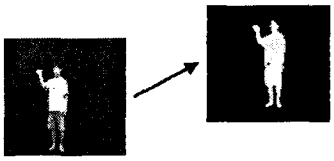

그림 1. 원 영상과 정규화 영상

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. 1는 원 영상과 이상에서 설명한 방법을 통하여 얻어진 이진 영상을 나타낸다.

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 λ 와 고유벡터 e 를 구한다.

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

$$X \triangleq [x_1-c, x_2-c, x_3-c, \ldots, x_N-c]^T \tag{2}$$

$$Q \triangleq XX^T \tag{3}$$

$$\lambda_i e_i = Q e_i \tag{4}$$

여기서 M은 한 영상의 픽셀 수이고 N은 전체 영상의 개수를 나타내는 정수이다. 고유치와 고유벡터를 구하는 데는 특이치 분해를 이용한다[2,3]. 특이치 분해를 이용하면 영상집합 X의 공분산 행렬에 대한 고유 벡터를 고유치가 큰 순서대로 구할 수 있다.

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

여기서 T_1 는 고유벡터의 개수를 조정하는 임계치이며, 제스처 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 = 16을 이용했다.

2.3. 제스처 공간에서 거리평가와 인식

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

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

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.

$$m(m_1, m_2, \ldots, m_n) \tag{7}$$

3. 계층적 클러스터링을 이용한 구체적 동작 분석

3.1 계층적 클러스터링 구성

파라메트릭 제스처 공간에서 효과적인 매칭을 하기 위해 트리 구조로 된 클러스터링을 이용하여 계층적으로 모델 영상들을 구성한다.

Cm에서 모든 영상을 대표하는 각 클래스 $Cm(m=1,2,\ldots,M)$의 평균 영상을 만들기 위해 비슷한 특징과 인접한 파라메터를 가지는 영상들로 분류해야 한다. 하나의 제스처는 그림 2와 같이 R개의 영상으로 이루어지고 클러스터 Cm으로 나누어 진다고 하자. 그림 3에서 보여진 것처럼 이것은 tm ~ tm+1 의 범위 내로 제한된 Cm에서 이미지와 일치하는 t에 따라 구성된다. 하지만 평범한 분류 방법으로는 그림 4에서 보여진 것처럼 파라메터의 연속성을 설명할 수 없다. 이런 이유 때문에 제안된 방법에서는 우리는 고유공간에서 곡선을 이은 학습이미지들을 분류하여 Cm에서 t의 연속성을 유지한다. 계층적 모델 구성으로 식별과 최소 에러에 대해 최적인 클래스 경계를 결정하는 것이다. 그림 2에서 보듯이, 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 Cm 은 다음과 같이 표현된다.

$$1=k_1<k_2<k_3\ldots<k_M\leq R \tag{8}$$

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

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

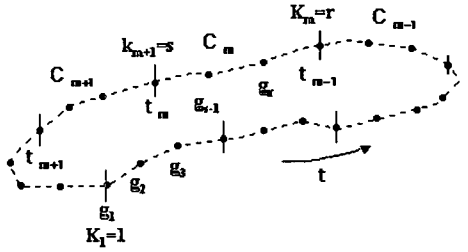

그림 2.제스처공간에서 모델 영상의 연속적인 분류

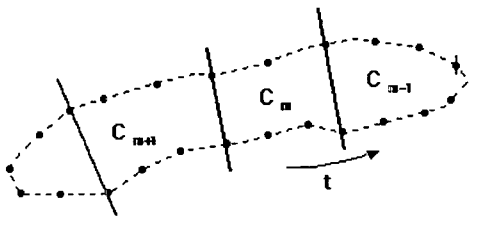

그림 3. 연속성을 고려하지 않은 분류

$$h_T=\frac{1}{R}\sum_{r=1}^{R}g_r\ ,\quad \sigma_t^2\frac{1}{R}\sum_{r=1}^{R}||g_r-h_T||^2 \tag{10}$$

Cm 은 gr을 포함할 확률 w_m을 가진다.

$$w_m=\frac{1}{R}\sum_{r\in C_m}1 \tag{11}$$

Cm에서 gr의 평균영상과 분산은 다음과 같다.

$$h_m=\frac{1}{Rw_m}\sum_{r\in C_m}g_r \tag{12}$$

경계 Km은 Cm의 분류에 의해 계산된다.

$$\sigma_m^2=\frac{1}{Rw_m}\sum_{r\in C_m}||g_r-h_m||^2 \tag{13}$$

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

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

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

$$\sigma_W^2+\sigma_B^2=\sigma_T^2 \tag{17}$$

결과적으로 식(14)는 다음 식과 동등하다.

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

이러한 이유 때문에 우리는 Km를 구할 수 있다.

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

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

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

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

여기서 R은 분류된 이미지의수이고 M은 클래스의 개수이다.

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

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.

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

3.2 제스처 공간에서 매칭과 인식

정규화 된 영상들의 고유공간과 인식을 위한 모델 영상들의 공간내의 위치가 정해지면 인식 과정은 매우 간단해진다. 먼저 평균영상에서 입력영상 y를 뺀 다음 고유공간에 식(24)과 같이 투영하여 제스처 공간의 위치를 계산한다.

$$z=[e_1,e_2,e_3,\ldots,e_k]^T(y-c) \tag{24}$$

제스처 인식은 입력 영상의 잡음이나 시스템의 착오를 감안한 임계값을 정해서 각 모델들과의 거리 중 식(25)에 의해 최소한의 거리를 갖는 모델을 선택함으로써 이루어진다.

$$d_1=\min\ \|z-S(M)\| \tag{25}$$

즉, 모델 $S(M)$과 입력영상을 제스처 공간에 투영하여 얻은 점 z 사이의 거리 d_1를 구하여 임계값 보다 작으면 입력 제스처는 그 점에 대응하는 제스처 영상으로 인식하면 된다.

3.3 구체적 동작 정보 추정

입력 영상에 대해 구체적 제스처 정보 즉 동작의 빠르기나 크기를 알기 위해서는 제스처 공간상에 연속적으로 투영된 입력 영상들간의 거리를 식(26)를 이용하여 계산한다.

$$\sum(I_i-I_{i+1})=\|Id_i\| \tag{26}$$

그림 5는 제스처 공간상에서 입력 영상의 제스처 정보 추정방법을 나타낸 것이다. 그림 5에서처럼 연속적으로 투영된 입력 제스처 영상의 거리가 제스처 모델 영상의 보다 거리 값이 크므로 ($\|Md_1\| < \|Id_1\|$) 입력 동작이 단위 시간 동안에 모델 동작보다 빠른 동작임을 추정 할 수 있다. 또한 연속적인 입력 동작에 대한 방향을 추정하기 위해서 식(27)을 이용하여 각 동작을 연결하는 모델 선분과 입력 선분과의 내적을 구하여 서로간의 방향의 일치정도를 평가한다.

$$\frac{Md\cdot Id}{\|Md\|\ \|Id\|} \tag{27}$$

여기서 P_i, D_i, A_i는 각 각 영상에 대한 포즈, 거리, 방향에 대한 평가 값을 나타낸다.

$$Y_n=\eta_1(P_i)+\eta_2(D_i+A_i) \tag{28}$$

여기서 η_1 , η_2 에 대해 포즈인식과 동작 정보의 가중치 비율은 6 : 4이다.

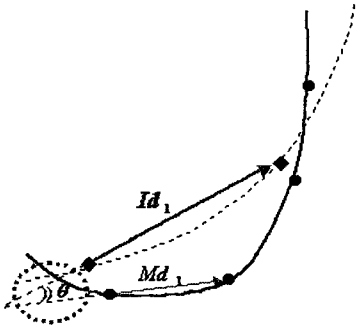

그림 4. 각 모델/입력 영상간 거리와 방향

4. 실험결과

실험에 이용된 제스처 영상은 한 사람의 간단한 맨손체조를 정상적인 속도로 수행한 것을 연속적인 영상으로 획득하

였다. 각 동작들이 서로 구분 될 수 있는 팔, 다리, 목, 허리 등의 구분 운동을 촬영하고 크기 정규화를 이용하여 100 × 100영상으로 변환하였다. 영상집합의 고유벡터를 계산한 후 재구성된 영상을 가장 복원하는 16차원의 벡터를 선택하여 제스처 공간으로 구성하였다. 따라서 100 × 100 = 10000차원의 이미지가 16차원으로 압축되는 효과도 거둘 수 있었으며 효과적인 모델 구성을 통해 실제 실시간 처리 시스템에도 적합하다는 것을 알 수 있었다. 그림 5와 6은 3차원 제스처 공간에서 각 영상들이 맵핑되는 결과를 나타냈다. 또한 입력 영상은 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. 그림 7.은 모델 구성별 매칭률을 나타내고 있다. k-NN을 이용한 매칭 방법이 가장 높은 인식률을 보였으나 비교 횟수가 가장 많았고 5개의 클래스를 모델로 구성한 경우 매칭율과 비교횟수에서 비교적 높은 인식률을 보였다.

5. 결론

본 논문에서 매칭 효과를 향상시키기 위해 파라메트릭 고유공간 방법에 대한 계층적 모델을 구성하는 새로운 접근을 제안했다. 제안된 트리 구조 매칭 방법은 인식률의 감소 없이 매칭효과를 향상 시켰다. 또한 제스처의 구분뿐만 아니라 제스처의 구체적인 정보 즉 동작의 빠르기나 연속적인 방향 등의 정보를 얻어내는데 유용하다는 것을 알 수 있었다. 이러한 구체적인 동작정보를 분석하여 입력 동작에 대해 동작의 정도를 알 수 있어 좀 더 정교하게 동작을 분석 할 수 있었다. 그러나 동작도중 몸통과 손발이 겹쳐서 가려진 경우나 모델 영상과 입력 영상의 모습이 현저히 달라지는 경우는 인식에 어려움이 따르고 복잡한 배경에서는 사람 영역만을 세그멘테이션 하는데 어려움이 있었다. 이러한 문제점을 해결하여 보다 안정적인 제스처 인식 알고리듬을 개발할 계획이다.

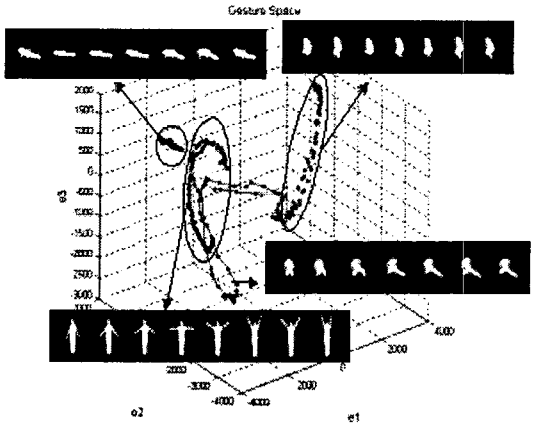

그림 5. 다중 모델 영상 맵핑

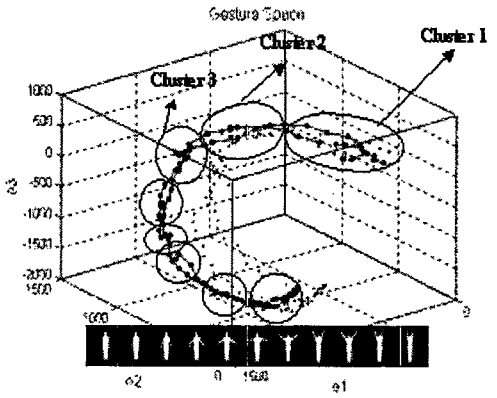

그림 6. 클러스터링된 모델 제스처 영상과 입력영상의 맵핑

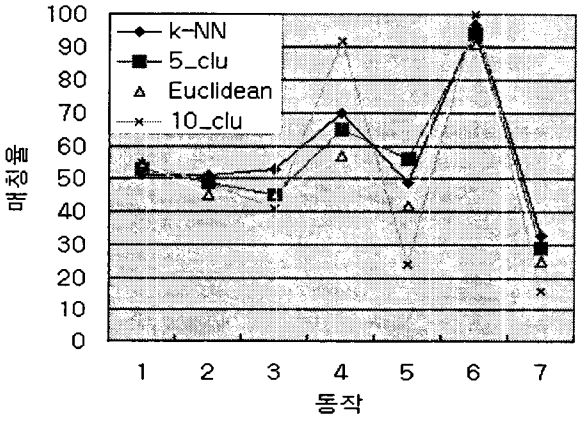

그림 7. 모델 구성별 매칭률

[참고문헌]

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

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

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

[4] 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,