가상 놀이 공간 인터페이스를 위한 HMM 기반 상반신 제스처 인식

HMM-based Upper-body Gesture Recognition for Virtual Playing Ground Interface

박재완, 오치민, 이칠우
전남대학교 전자컴퓨터공학과

Jae-Wan Park(cyanlip@image.chonnam.ac.kr),
Chi-Min Oh(sapeyes@image.chonnam.ac.kr), Chil-Woo Lee(leecw@chonnam.ac.kr)

요약

본 논문은 HMM기반의 상반신 제스처 인식에 대하여 연구하였다. 공간상의 제스처를 인식하기 위해서는 일단 제스처를 구성하고 있는 포즈에 대한 구분이 우선되어야 한다. 인터페이스에 사용되는 포즈를 구분하기 위해서 정면과 옆면에 설치한 적외선 카메라 두 대를 실험에 사용하였다. 그리고 각각의 적외선 카메라에서 하나의 포즈에 대한 정면 포즈와 옆면 포즈로 나눠서 획득한다. 획득한 적외선 포즈 영상은 SVM의 비선형 RBF 커널 함수를 이용하여 구분하였다. RBF 커널을 사용하면 비선형적 분류 포즈들간의 오분류 현상을 구분할 수 있다. 이렇게 구분된 포즈들의 연속은 HMM의 상태천이행렬을 이용하여 제스처로 인식된다. 인식된 제스처는 OS Value에 매핑하여 기존의 Application에 적용할 수 있다.

■ 중심어 : | 상반신 포즈 | 상반신 제스처인식 | 은닉 마르코프 모델 | 서포트 벡터 머신 | EOH |

Abstract

In this paper, we propose HMM-based upper-body gesture. First, to recognize gesture of space, division about pose that is composing gesture once should be put priority. In order to divide poses which using interface, we used two IR cameras established on front side and side. So we can divide and acquire in front side pose and side pose about one pose in each IR camera. We divided the acquired IR pose image using SVM's non-linear RBF kernel function. If we use RBF kernel, we can divide misclassification between non-linear classification poses. Like this, sequences of divided poses is recognized by gesture using HMM's state transition matrix. The recognized gesture can apply to existent application to do mapping to OS Value.

■ keyword : | Upper-body Pose | Upper-body Gesture Recognition | HMM(Hidden Markov Model) | SVM(Support Vector Machine) | EOH(Edge Orientation Histogram) |

I. 서 론

본 논문은 HMM 기반의 상반신 제스처 인식에 대하여 연구하였다. 어떠한 행동을 표현하는 인간의 제스처는 많은 포즈로 구성되어 있다. 그러므로 제스처를 인식하기 위해서는 그 제스처에 속하는 포즈를 우선적으

* 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2009년도 문화콘텐츠산업기술지원사업의 연구결과로 수행되었습니다.
접수번호 : #100702-006
접수일자 : 2010년 07월 02일
심사완료일 : 2010년 07월 26일
교신저자 : 박재완, e-mail : cyanlip@image.chonnam.ac.kr

로 정의하고 인식하여야 한다. 제스처와 시스템을 결합하여 사용자 인터페이스로 사용하기 위해 포즈를 정의하고 인식하는 과정은 제스처를 직관적으로 이해할 수 있도록 구성되어야 한다. 그러므로 본 논문에서 정의한 제스처는 양 손의 움직임을 토대로 한 직관적인 포즈 동작으로 구성하였다.

일단 상반신 제스처를 인식하기 위해서는 제스처를 구성하고 있는 상반신 포즈 동작을 구분할 수 있어야 한다. 상반신 포즈를 구분하기 위하여 다음과 같은 방법들이 있다.

신체의 포즈를 인식하기 위하여 주로 미리 정의한 구조적 포즈 모델을 이용하는 방법을 사용하는 방법은 particle filter를 이용하여 신체의 포즈를 인식하고 추적하고 있다. 파티클의 개수가 증가함에 따라 인식률의 성능은 높아지지만 그에 반해 시스템의 속도에 영향을 주기 때문에 적절한 파티클의 개수를 설정하는 것이 중요하다[1].

신체의 포즈를 템플릿을 이용하여 구분하고 있는 방법은 입력영상과의 템플릿 비교를 하기 위하여 포즈 후보영역에서 영역별 가중치를 이용하여 학습된 포즈를 인식하고 있다[2].

다중 카메라를 이용하여 획득한 영상에서 Haarlets[5]를 이용하여 포즈를 인식하는 방법은 포즈로 추정되는 전경의 실루엣 영상을 LDA[6]와 ANMM[4]을 이용하여 포즈로 구분한다. 하지만 ANMM은 속도가 느리기 때문에 Haarlets을 기반으로 하는 integral image를 ANMM의 component로 사용한다[3].

이러한 연구들은 입력 영상에서 포즈를 추출하기 위해 배경과 전경을 분리하고 전경의 실루엣을 이용하고 있다. 실루엣은 이용하는 방법은 2차원 평면 공간에서는 직관적인 형태를 얻을 수 있지만 3차원 입체 공간에서의 정보를 알기 힘든 단점이 있다. 그러므로 본 논문에서는 두 대의 카메라를 이용하여 정면과 옆면의 영상을 획득하고 x,y,z좌표 정보를 결합하여 포즈를 추출한다.

영상에서 피부색으로 신체의 일부를 검출하는 것은 어려운 일이다. 피부색과 유사한 배경 또는 객체가 존재할 수 있고 조명의 변화 또한 큰 영향을 끼치기 때문이다. 그리고 영상으로부터 신체의 완전한 3차원적 구조를 인식하기 위해서는 x,y축 이외의 z축의 정보를 이용해야 한다.

깊이 영상을 추출하기 위한 방법으로는 스테레오 카메라를 이용하거나 TOF 카메라를 사용한다. 스테레오 카메라는 해상도에 따라 깊이 정보를 획득할 수 있는 거리가 달라지는 단점이 있고 TOF카메라는 가격이 비싸다는 단점이 있다.

본 논문에서는 빠른 인식속도, 자유로운 해상도와 인식거리를 위해 일반 CCD 카메라에 적외선 필터를 장착하여 사용하였다. 적외선을 사용하는 방법은 적외선 광원과 적외선 카메라를 이용하므로 구조물에 크게 구애받지 않는 한, 손쉽게 영상을 얻을 수 있는 장점이 있다.

본 논문에서는 전경과 배경을 분리할 때, 적외선 영상을 사용하고 전경(상반신 포즈)의 실루엣 영상에서 EOH와 SVM을 이용하여 포즈 영상의 특징을 추출하고 포즈 영상을 구분한다. 그리고 구분된 포즈 영상을 HMM의 상태 천이 확률을 이용하여 제스처로 인식한다. 제스처는 연속적인 포즈의 결합으로 이루어져 있으므로 시계열의 순차적인 데이터의 연속성을 효과적으로 인식 가능한 HMM을 사용한다.

그림 1. 각 포즈에 해당하는 정면 포즈와 옆면 포즈 영상 (위 : 정면포즈 , 아래 : 옆면포즈)

II. 실험 내용

2.1 전경 분리

본 논문에서 제안한 시스템은 각 포즈에 대하여 정면과 옆면의 학습데이터를 50개씩 학습하였다.

[그림 1]에서와 같이 적외선 영상에서는 원하는 영역만을 쉽게 획득할 수 있으므로 간단한 배경제거만을 통하여 포즈 영상을 획득할 수 있다. 가상 놀이공간에서 획득한 포즈 영상은 [그림 1]에서와 같이 각각 정면 포즈와 옆면 포즈로 구성된다. 이렇게 분리된 포즈 영상을 이용하여 평면적인 실루엣에서 얻을 수 없는 깊이 정보를 대신하게 된다.

2.2.1. EOH (Edge Orientation Histogram)

본 논문에서는 ROI(Region of Interest)영상이 포함하고 있는 36개의 방향에 분포한 에지의 양을 표현하는 EOH 특징을 포즈 정보로 사용한다. EOH 특징은 총 36개의 차원으로 구성되며 총 36개의 에지방향을 가지고 있다.

36방향의 에지를 얻기 위하여 하나의 영상에서 다음과 같은 에지의 x축과 y축에 대한 에지영상 G_x 와 G_y 를 추출한다.

$$G_x = Sobel(I_{ROI}),\ G_y = Sobel(I_{ROI}) \tag{1}$$

그리고 아래의 식을 통하여 픽셀 (i,j) 의 에지의 방향과 크기를 계산한다.

$$\theta_{i,j} = \arctan\left(G_x(i,j)/G_y(i,j)\right) \tag{2}$$

$$m_{i,j} = \sqrt{G_x(i,j)^2 + G_y(i,j)^2} \tag{3}$$

하나의 포즈당 두 영상을 얻으므로 총 72개의 특징을 통합하여 각 영상 ROI의 EOH 특징에 해당하는 각 픽셀 (i,j) 의 에지크기 $m_{i,j}$ 를 각 방향성별로 모두 합산을 통해 추출한다.

2.2.2. SVM (Support Vector Machine)

SVM은 1995년에 Vapnik에 의해 제안되었고 VC(Vapnik-Chervonenkis)이론에 근간을 두고 있으며 뛰어난 일반화 성능을 보여준다. SVM은 원래 이원 패턴인식, 문제를 해결하기 위한 방법으로서 두 데이터 집단 사이의 거리를 최대화하는 서포트 벡터를 이용하여 입력 데이터가 어느집단에 속하는지 결정할 수 있는 일반화 성능이 좋은 분류기이다. 기존의 경험적 에러 최소화 기법인 다층신경망과 비교하여 학습에 필요한 파라메터의 일부가 자동으로 결정된다.

선형분리 가능한 이진 클래스의 경우, 입력값들을 다른 클래스 간의 데이터 사이를 최대거리로 분리할 수 있는 초평면(Hyperplane)을 찾기 위해서 고차원의 특징공간(Feature space)으로 변환시킨다.

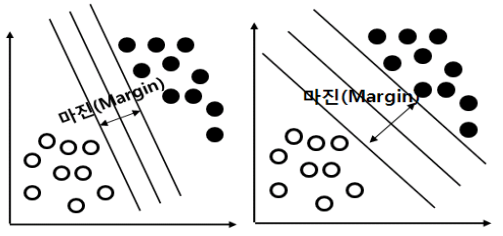

그림 2. 선형분리 가능한 데이터
(a) 마진의 거리가 짧음
(b) 마진의 거리가 최대

두 클래스 군집을 선형 분리하는 초평면과 가장 가까운 점을 'Support Vector(SV)'라고 한다. SV와 결정 평면간의 거리를 '마진(Margin)'이라고 한다.

SVM은 마진을 최대화하는 최적의 초평면을 찾는다. [그림 2]는 선형분리 가능한 데이터에 대한 초평면의 예를 보여준다. (a)는 마진의 거리가 작은 경우이고, (b)는 마진의 거리가 최대인 경우이다.

최대의 마진을 가지는 초평면을 구하기 위해 식(4)의 최소값을 구해야 한다. 동시에 데이터를 분류하기 위한 식(5)를 만족하여야 한다.

$$\tau(w) = \frac{1}{2}\|w\|^2 \tag{4}$$

$$y_i \cdot \left((w \cdot x_i) + b\right) \geq 1,\ i = 1,\cdots,l \tag{5}$$

식(5)의 제약조건을 만족하면서 식(4)의 최소값을 구하는 문제는 Lagrange Multiplier를 사용한 식(6)으로 표현된다.

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{l}\alpha_i(y_i\cdot((x_i\cdot w)+b)-1) \quad (6)$$

선형적으로 분리가능하지 않은 경우, 커널함수(Kernel function)을 사용하여 특징공간(Feature space)에서 선형분리를 수행한다.

[그림 3]은 선형분리가 불가능한 데이터에 대한 초평면의 예를 보여준다. 이 데이터를 커널함수를 사용하여 특징공간으로 이동한다. 본 논문에서는 포즈 영상들의 ROI를 정규화하였지만 선형분리가 불가능하므로 RBF(Radial Basis Function) 커널함수를 사용하여 포즈들을 분리하였다.

2.3 제스처 인식

제스처 인식 과정 단계에서는 인식된 포즈 영상들로부터 제스처를 인식한다. 본 논문에서는 제스처 인식을 위해 HMM을 사용하였다. HMM은 시간적으로 제약을 받는 정보의 구조를 모델링 하는데 뛰어난 모델이다. Simple Markov Model만으로 모델링하기 힘든 실세계의 문제를 통계적 매개변수로 접근할 수 있게 해준다. 상태 전이 매개 변수는 순차적인 일련의 사건 발생을 모델링한다. 그리고 관측 심볼 확률 분포는 각 사건의 특징을 유한개의 심볼로 대응시킨다.

HMM은 이러한 두 가지 확률 과정의 결합으로 이루어져 있고, 이 기준에 따라 생성된다. 생성된 HMM은 학습 데이터를 이용한 학습을 통해 적절한 제스처 모델을 구성한다.

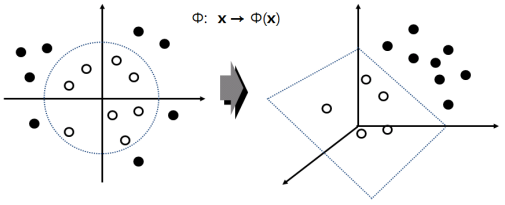

그림 3. RBF 커널을 사용하여 특징공간에 데이터를 투영

인식과정에서는 인식하고자 하는 제스처와 학습이 끝난 후 생성된 모든 HMM의 제스처 모델을 비교한다. 그리하여, 가장 유사하다고 판단되는 제스처 모델을 선택하고 결과를 확률로 나타낸다.

상태전이확률을 이용하여 각 샘플들을 이용한 전처리 단계는 잡음제거와 대표점 추출 과정을 거쳐 궤적을 체인코드로 변환하는 과정으로 각 포즈의 번호를 코드로 사용한다.

학습은 각 제스처별로 이루어지며, 해당 제스처의 HMM 모델에 학습결과를 적용한다. HMM의 학습 과정은 각 숫자별로 상반신 포즈를 이용하여 은닉 마르코프 모델을 구성하는 과정으로, EM알고리즘의 하나인 Baum-Welch 알고리즘을 이용한다. HMM의 인식 과정은 전처리 모듈을 통해 변환된 체인코드가 어느 숫자의 은닉 마르코프 모델에서 나타날 확률이 높은가를 판단하는 과정으로, 각각의 숫자 모델에 전향(Forward) 알고리즘을 적용하여 가장 높은 확률을 보이는 숫자 모델을 최종 인식 결과로 출력한다.

HMM은 아래와 같은 요소로 구성된다.

□ N: 상태의 수,
$S=\{S_1,S_2,....,S_N\}$: 상태의 집합,
q_t : 시간 t의 상태

□ M: 관측 심볼의 수,
$V=\{v_1,v_2,....,v_M\}$: 관측 심볼의 집합

□ A : $\{a_{ij}\}$: 상태 전이 확률 분포
$a_{ij}=P(q_{t+1}=S_j|q_t=S_i), 1\le i,j\le N$
: 상태 i에서 상태 j로 전이할 확률

□ $B=\{b_t(k)\}$: 관측 심볼 확률 분포,
$b_j(k)=P(v_k|q_t=S_j), 1\le j\le N, 1\le k\le M$
: j에서 심볼 v_k를 관측할 확률

□ $\pi=\{\pi_t\}$: 초기 상태 확률 분포,
$\pi_t=P(q_1=S_i), 1\le i\le N$: 초기 상태가 i일 확률

일반적으로 하나의 HMM은 $\lambda=(A,B,\pi)$로 표시된다. 주어진 모델과 관측열 $O=(O_1,O_2,....,O_\tau)$에 대해 생성확률은 식(7)과 같다.

$$P(O|\lambda)=\sum_{\text{for all } q}\left[\pi_{q_1}b_{q_1}(O_1)\prod_{i=2}^{\tau}a_{q_{i-1}q_i}b_{q_i}(O_i)\right] \quad (7)$$

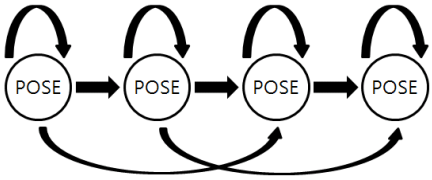

그림 4. Left-Right HMM 구조 모델

HMM의 응용에는 세 가지의 해결해야 할 문제가 있다. 즉, 평가, 해석, 그리고 학습에 있으며 이들은 각기 Forward-Backward 알고리즘, Viterbi 알고리즘 그리고 Baum-Welch 알고리즘으로 해결된다. 본 논문에서는 [그림 4]와 같은 Left-Right HMM구조 모델을 사용한다.

각 제스처는 시작포즈를 가지고 있고 제스처에 해당하는 상태천이확률이 무효할 경우에는 "IDLE" 상태를 가지고 있으며, 상태천이확률이 유효한 시점부터 각 제스처에 해당하는 포즈를 전부 만족할 경우 제스처가 종료된 것으로 판단한다. Left-Right HMM 구조 모델은 포즈 사이에 잡음이 생겨도 미리 학습한 상태천이확률이 유효하므로 실시간 인터페이스로 사용하기에 무리가 없다.

[그림 6]는 학습과정을 나타낸 것이고, [그림 7]는 HMM을 사용한 제스처 인식 과정을 보여준다.

III. 실험 결과

본 논문에서는 총 7개의 제스처를 정의하였고 제스처에 필요한 포즈의 개수는 노이즈로 인식되는 포즈를 포함하여 제스처마다 3~6개로 정의하였다.

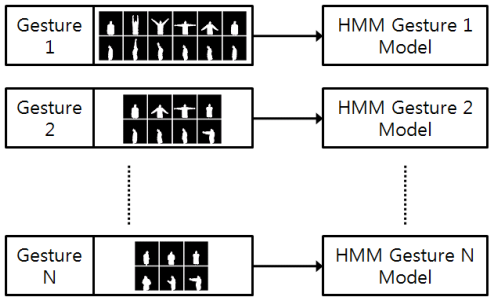

그림 6. 제스처 학습 과정

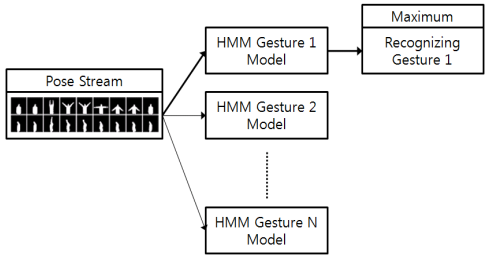

그림 7. 제스처 인식 과정

그림 5. 제스처 인식을 위한 포즈 스트림의 예 (각 스트림 당 위 : 정면 포즈 , 아래 : 옆면 포즈)

표 1. 제스처를 구성하는 포즈의 개수에 따른 인식률 (인식률이 0인 경우는 제스처에 필요한 최소 포즈개수에 미달한 제스처 임)

구성 포즈 개수 / 제스처 종류	3 pose	4 pose	5 pose	6 pose
gesture 1	0	0	92%	83%
gesture 2	0	86%	87%	53%
gesture 3	0	84%	79%	46%
gesture 4	92%	85%	64%	27%
gesture 5	74%	53%	38%	14%
gesture 6	92%	95%	85%	52%
gesture 7	94%	96%	79%	61%

제스처마다 포즈영상의 수에 따른 인식률은 [표 1]과 같다. 제스처4의 경우는 제스처를 구성하는 포즈의 에지 방향이 비교적 비슷한 포즈에서는 인식률이 좋지 않았다.

각각의 제스처는 11개의 기본 포즈를 이용하여 구성된 제스처이므로 각 제스처는 동일한 포즈가 중첩될 수 있다.

동일한 제스처라고 할지라도 제스처에 포함되는 포즈의 개수가 많을수록 인식률이 낮아지는 결과를 보였다. 그러므로 각 제스처에 필요한 적절한 포즈들을 이용하여 학습하는 과정이 필요하고, 가능하다면 구분이 가능한 포즈들을 되도록 많이 정의하여 포즈가 겹치지 않아야 한다는 것을 알 수 있었다.

IV. 결론

본 논문에서는 HMM 기반의 상반신 제스처 인식에 대하여 기술하였다. 사람마다 신체 부위의 크기가 다르므로 카메라를 통해 입력되는 각각의 포즈영상을 정규화하는 작업을 진행하였고, 정규화된 포즈영상을 SVM을 이용하여 분류하였다. 분류된 포즈영상은 HMM을 이용하여 제스처로 인식하였다. 인식된 제스처는 OS Value에 매핑하여 기존의 Application에 적용할 수 있었다. 분류된 포즈 영상 중 에지 방향이 비슷한 영상의 경우 같은 포즈 영상으로 구분되는 문제를 보였다. 그러므로 구분이 확실하게 가능한 포즈 영상의 획득과 강건한 포즈 인식 알고리즘이 필요할 것으로 생각된다. 그리고 앞으로 자유로운 환경에서 적용이 가능하려면 깊이 영상을 획득할 수 있는 Z-Depth 카메라 등을 활용하여 실시간으로 사용자의 의도에 맞는 인터페이스를 구현하여야 할 것이다.

참 고 문 헌

[1] 오치민, 정문호, 유범재, 이칠우, "개선된 챔퍼매칭 우도 기반 2차원 평면 객체 추적", 정보처리학회논문지 B, 제17-B권, 제1호, pp.37-46, 2010.

[2] M. Dimitrijevic, V. Lepetit, and P. Fua, "Human Body Pose Detection Using Bayesian Spatio-Temporal Templates," Computer Vision and Image Understanding, Vol.104, No.2/3, pp.127-139, 2006.

[3] M. Van den Bergh, E. Koller-Meier, and L. Van Gool, "Real-Time Body Pose Recognition Using 2D or 3D Haarlets," International journal of computer vision, Vol.83, No.1, pp.72-84, 2009.

[4] F. Wang and C. Zhang, "Feature extraction by maximizing the average neighborhood margin," In IEEE computer society conference on computer vision and pattern recognition, Vol.1-8, pp.1173-1180, 2007.

[5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," In IEEE computer society conference on computer vision and pattern recognition, Vol.1, pp.511-518, 2001.

[6] M. Van den Bergh, E. Koller-Meier, and L. Van Gool, "Fast body posture estimation using volumetric features," In IEEE visual motion computing, Jan, 2008,

[7] 박아연, 이성환, "이산 은닉 마르코프 모델에 기반한 3차원 전신 제스처 인식", 한국정보과학회 컴

퓨터 비전 및 패턴인식연구회 추계 워크샵 발표 논문집, pp.154-156, 2004.

[8] H. Kang, C. W. Lee, K. Jung, "Recognition-based gesture spotting in video games," Pattern Recognition Letters, Vol.25, pp.1701-1714, 2004.

저 자 소 개

박 재 완(Jae-Wan Park) 정회원

- 2007년 2월 : 호남대학교 정보통신공학과(공학사)
- 2009년 2월 : 전남대학교 전자컴퓨터공학과(공학석사)
- 2009년 3월 ~ 현재 : 전남대학교 전자컴퓨터공학 박사 과정

<관심분야> : 멀티터치 제스처, 상호작용 인터페이스

오 치 민(Chi-Min Oh) 정회원

- 2007년 2월 : 전남대학교 컴퓨터정보통신공학과(공학사)
- 2009년 2월 : 전남대학교 전자컴퓨터공학과(공학석사)
- 2009년 3월 ~ 현재 : 전남대학교 전자컴퓨터공학과 박사과정

<관심분야> : 객체추적, 제스처인식, HCI

이 칠 우(Chil-Woo Lee) 정회원

- 1986년 2월 : 중앙대학교 전자공학과(공학사)
- 1988년 2월 : 중앙대학교 전자공학과(공학석사)
- 1992년 : 동경대학교 대학원 전자공학과(공학박사)
- 1992년 ~ 1995년 : 이미지 정보과학 연구소 수석 연구원 겸 오사카 대학 기초공학부 협력연구원
- 1995년 : 리츠메이칸대학 정보공학부 특별초빙강사
- 1996년 ~ 현재 : 전남대학교 전자컴퓨터공학과 교수

<관심분야> : 컴퓨터 비전, 지능형 휴먼 인터페이스, 디지털 콘텐츠, 컴퓨터 그래픽스