신체 부분 포즈를 이용한 깊이 영상 포즈렛과 제스처 인식

(Depth Image Poselets via Body Part-based Pose and Gesture Recognition)

박재완*, 이칠우**

(Jae Wan Park, Chil Woo Lee)

요 약

본 논문에서는 신체 부분 포즈를 이용한 깊이 영상 포즈렛과 제스처를 인식하는 방법을 제안한다. 제스처는 순차적인 포즈로 구성되어 있기 때문에, 제스처를 인식하기 위해서는 시계열 포즈를 획득하는 것에 중점을 두고 있어야 한다. 하지만 인간의 포즈는 자유도가 높고 왜곡이 많기 때문에 포즈를 정확히 인식하는 것은 쉽지 않은 일이다. 그래서 본 논문에서는 신체의 전신 포즈를 사용하지 않고 포즈 특징을 정확히 얻기 위해 부분 포즈를 사용하였다.

본 논문에서는 16개의 제스처를 정의하였으며, 학습 영상으로 사용하는 깊이 영상 포즈렛은 정의된 제스처를 바탕으로 생성하였다. 본 논문에서 제안하는 깊이 영상 포즈렛은 신체 부분의 깊이 영상과 해당 깊이 영상의 주요 3차원 좌표로 구성하였다.

학습과정에서는 제스처를 학습하기 위하여 깊이 카메라를 이용하여 정의된 제스처를 입력받은 후, 3차원 관절 좌표를 획득하여 깊이 영상 포즈렛이 생성되었다. 그리고 깊이 영상 포즈렛을 이용하여 부분 제스처 HMM을 구성하였다. 실험과정에서는 실험을 위해 깊이 카메라를 이용하여 실험 영상을 입력받은 후, 전경을 추출하고 학습된 제스처에 해당하는 깊이 영상 포즈렛을 비교하여 입력 영상의 신체 부분을 추출한다. 그리고 HMM을 적용하여 얻은 결과를 이용하여 제스처 인식에 필요한 부분 제스처를 확인한다. 부분 제스처를 이용한 HMM을 이용하여 효과적으로 제스처를 인식할 수 있으며, 관절 벡터를 이용한 인식률은 약 89%를 확인할 수 있었다.

■ 중심어 : 신체 부분 포즈 ; 제스처 인식 ; HMM(Hidden Markov Model)

Abstract

In this paper we propose the depth-poselets using body-part-poses and also propose the method to recognize the gesture. Since the gestures are composed of sequential poses, in order to recognize a gesture, it should emphasize to obtain the time series pose. Because of distortion and high degree of freedom, it is difficult to recognize pose correctly. So, in this paper we used partial pose for obtaining a feature of the pose correctly without full-body-pose.

In this paper, we define the 16 gestures, a depth image using a learning image was generated based on the defined gestures. The depth poselets that were proposed in this paper consists of principal three-dimensional coordinates of the depth image and its depth image of the body part.

In the training process after receiving the input defined gesture by using a depth camera in order to train the gesture, the depth poselets were generated by obtaining 3D joint coordinates. And part-gesture HMM were constructed using the depth poselets. In the testing process after receiving the input test image by using a depth camera in order to test, it extracts foreground and extracts the body part of the input image by comparing depth poselets. And we check part gestures for recognizing gesture by using result of applying HMM. We can recognize the gestures efficiently by using HMM, and the recognition rates could be confirmed about 89%.

■ keywords : Body-partial Pose ; Gesture Recognition ; HMM(Hidden Markov Model)

* 학생회원, 전남대학교 전자컴퓨터공학과
** 정회원, 전남대학교 전자컴퓨터공학과
본 연구는 교육과학기술부와 한국연구재단의 지역혁신인력양성사업으로 수행된 연구결과임

접수일자 : 2015년 11월 25일 게재확정일 : 2016년 04월 03일
수정일자 : 2016년 03월 22일
교신저자 : 이칠우 e-mail : leecw@jnu.ac.kr

I. 서 론

카메라를 이용하여 자연스러운 동작을 획득하면서 이를 인터페이스로 사용하는 지능적 환경을 만드는 기술은 산업 기술 분야의 중요한 관심사라고 할 수 있다. 이제는 산업 기술이 인간 중심적인 환경을 제공하기 위한 중요한 기술로써 액션 인식을 필요로 하게 된 것이다.

그러므로 액션의 기본이 되는 포즈 영상의 특징을 잘 검출하는 것이 중요하며, 카메라 잡음, 가려짐, 겹침 등에도 불구하고 인간의 신체 포즈 영상을 검출하는 것이 가장 중요하다고 할 수 있다.

인간의 신체를 검출하는 방법 중, 보행자 검출에서 많이 사용되는 HoG(histogram of Oriented Gradient)[1]는 사람의 신체 영역을 특징 패턴으로 모델링하여 영상으로부터 동일한 패턴을 갖는 영역을 사람으로 검출하는 방법이다. 하지만 특징 패턴에서는 에지 정보만을 이용하기 때문에 검출 이외의 용도에서 부적합하다는 단점을 가지고 있다. 또한 본 논문에서의 입력 영상의 특징이 밝기 값이며 밝기 변화에 따라 형태가 변하는 영상에서는 강건한 특징 값을 추출하기 어렵기 때문에 본 논문에서는 HoG를 사용하지 않는다.

최근에는 사람 영역을 부분적으로 모델링하여 검출함으로써 신체의 가려짐이나 신체 형태의 변화에 강인한 알고리즘들이 개발되어 오고 있다. 이러한 신체 부분 기반 사람 검출 방법은 전체 포즈를 사용하는 방법에 비해 다음과 같은 장점을 가진다.

첫 번째로 포즈 형태의 부분적 변형 및 신체의 부분적 가려짐에 덜 민감하며, 두 번째로 포즈 모델이 신체 영역의 중요 영역에서만 추출되므로 모델자체에 필요한 특징만 포함된다. 마지막으로 신체 부분 영역만 계산하므로 연산 시간이 줄어든다는 장점이 있다.

하지만 이러한 장점에도 불구하고 신체 부분 기반의 방법은 신체 부분을 선택하기 위한 별도의 알고리즘이 필요하며, 각 신체 부분 별로 전경과 배경을 분류할 수 있는 분류기가 필요하다는 단점이 있다. 그리고 신체 부분의 수가 많을 경우 오히려 전체 포즈 기반 방법에 비해 검색시간이 증가하는 문제점을 가진다. 따라서 이러한 문제점을 해결하기 위해 각 신체 부분 별로 사람의 특징을 잘 포함할 수 있는 최적의 신체 부분 영역을 선출할 수 있는 알고리즘이 필요하다.

본 논문에서는 HoG를 사용하지 않는 대신, 신체 부분 영상을 이용하는 방법 중 최근 가장 효율적인 방법인 포즈렛[2]을 사용한다. 포즈렛이란 사전에 설정된 사람 영상의 3D좌표를 이용하여 신체 부위별로 부분 영상을 구성하고 이를 학습시켜 가려짐에 강건한 사람 검출 알고리즘이다.

본 논문의 실험 영상은 키넥트를 통해 얻은 깊이 영상을 신체 부분에 따라 나누어 사용하고 있으며, 부분 영상을 생성하기 위해 신체 구조에 따라 분리하였을 뿐만 아니라 신체 부분의 형태 정보를 동시에 고려하여 특징으로 사용하기 때문에 깊이 영상 포즈렛이라고 부를 수 있다.

인간의 신체 부위가 포함된 영상에 포즈렛을 적용하면 신체 영역을 검출하면서 미리 정의된 3차원 관절벡터를 얻을 수 있다. 이처럼 포즈렛은 인간의 액션을 인식하기보다는 주로 영상 내의 신체를 검출하는 용도로 사용된다. 그렇기 때문에 포즈렛을 이용하여 액션을 인식하는 경우에는 포즈렛 모션 히스토그램(Poselet Motion Histogram)[3]이나 포즈렛 활성 벡터(Poselet Activation Vector)[4]를 이용하는 것이 일반적이다.

이런 경우 인간의 액션을 인식하기 위해 하나의 영상을 이용하는 방법을 선호하지만, 본 논문에서는 정의된 액션의 군집화된 영상을 사용하는 방법대신 순차적인 입력에 따라 얻을 수 있는 포즈렛 벡터들을 이용하여 액션을 인식하도록 하였다.

II. 깊이 영상 포즈렛

본 장에서는 서론에서 언급한 것처럼 부분 포즈를 이용하여 포즈를 비교할 것이다. 이에 앞서, 포즈의 전체 특징을 이용하는 연구와 포즈의 부분 특징을 이용하는 연구에 대해 알아볼 것이다.

최근 가장 각광받는 방법은 신체 부분을 이용한 포즈렛(Poselets)[2]이다. 포즈렛은 입력 영상에서 포즈렛을 찾기 쉬워야 하며 포즈렛의 검출을 통해 관절의 3차원 위치를 구성하기 쉬워야 한다는 조건이 필요하다. 이 조건을 실행하기 위하여 Bourdev[2]는 2차원의 영상에서 3차원의 관절 정보를 포함하고 있는 신체 부위의 명칭을 가지고 있는 'H3D'라는 새로운 데이터셋을 만들었다. 포즈렛은 PS(Pictorial Structure)[5]와 같은 부분 포즈를 이용한 이전까지의 연구에 비교했을 때, 신체 외형 정보와 신체 구조 정보를 결합하였다는 점에서 차이가 있다. 그림 1은 H3D 데이터셋에서 입력된 영상과 가장 유사한 형태를 가지고 있는 신체 부위를 검출한 결과이다.

그림 1. 입력 영상과 포즈렛[9](H3D 데이터셋)

포즈렛과 같이 신체 부분 영상을 이용하여 신체를 검출하거나 인간의 액션을 인식하는 연구는 다음과 같다.

Li[6]는 보행자를 검출하기 위해 신체 부분을 'head-shoulder', 'torso', 'legs'의 세 부분으로 나누었다. 그리고 신체 부분 중 알아내기 어려운 부분에 대해서는 여러 개의 분류를 갖게 한 뒤, or-node로 'Grammar Model'을 구성하여 신체를 구분하였으며, 입력 영상에 ABM템플릿과 HoG템플릿를 적용한 결과를 결합하여 보행자를 검출하였다.

Wang[7]은 액션을 인식하기 위해 신체 포즈를 다섯 개의 부분으로 구분하였다. 하나의 포즈 당 14개의 주요 관절만을 사용하여 포즈를 표현하였다. 논문 내에서는 모션 정보를 중요하게 생각하고 언급하고 있지만, 정작 액션을 정의하는 단계에서는 주요 포즈들을 군집해서 인식에 사용하였기 때문에 14개의 주요 관절로 이뤄진 포즈들의 중복을 전부 다 표현할 수 없다는 단점을 가지고 있다. 하지만 포즈와 모션 정보를 결합해서 인식하려는 의도를 알 수 있다.

Yang[8]은 신체를 검출하기 위해 부분 패치, 신체 부분보다 더 작은 단위의 패치를 이용해서 신체를 검출한다. 연구단계에서 고민해야 할 부분은 신체 부분을 분리하였을 때, 신체 구조에 따라 신체의 연결을 고려할 수 있어야 하는데, 일반적으로 부분 포즈를 이용해서 포즈를 검출하는 연구는 신체의 형태를 고려하지 않는 반면 이 연구에서는 신체의 연결을 고려하고 있다. 그렇기 때문에 포즈를 구성할 때, 자유도가 높은 부분 모델을 구성할 수 있다는 장점을 가지고 있다.

Desai[9]는 포즈렛, 연결된 스켈레톤(FMP: flexible mixtures of parts), Visual phrase, 이 세 가지 방법의 장점을 결합하여 Phraselet이라는 신체 모델을 정의하였으며 FMP에서 필요한 부분 포즈 매칭 바운딩 박스를 필요로 하지 않는다. 그리고 Appearance model과 Spatial model을 결합하여 사용함으로써 부분 포즈를 이용하는 장점을 충실히 이행하고 있지만 특별히 포즈렛과 차이를 주지 않고 있다. 하지만 액션을 검출하기 위해서 하나의 영상을 이용하는 점이 단점이고 이를 보완하기 위해 주변 상황과 함께 Object를 이용하려 하였다.

Wang[10]은 계층적인 포즈렛을 제안하고 있으며 새로운 개념인 part-based methods와 exemplar-based methods를 적용하였다. 이 연구에서 정의하길, part-based 모델은 두 가지 중요한 요소로 구성되어 있다고 한다. 첫 번째는 각 신체 부분이 특징을 가지고 있어야 하고, 두 번째는 각각의 신체 부분은 신체에 맞게 정렬되어 있어야 한다. 계층적인 포즈렛에서 하위 포즈렛은 상위 포즈렛에 속할 뿐 어떤 관계요소를 보이지 않고 있으며, 단지 계층구조로 구성되어 있을 뿐이다.

이와 같이 포즈렛을 이용하여 신체를 검출하고 액션을 인식하기 위해 다양한 연구와 시도가 이루어지고 있다. 본 논문에서는 이러한 포즈렛의 개념을 이용하여 신체를 검출하고 시간 연속적인 순차적 포즈를 이용하여 액션을 인식하도록 할 것이다.

본 논문에서는 포즈렛과 마찬가지로 신체의 외형(Appearance Model)과 신체 구조(Configuration Model)를 이용하여 깊이 영상 포즈렛을 제안하였다. 학습 영상은 키넥트를 통해 얻을 수 있는 깊이 영상과 3차원 관절 좌표를 가지고 있는 스켈레톤을 매 프레임마다 저장하여 학습에 사용하였으며 신체 부분의 특징으로 사용한다. 신체 구조에 따라 깊이 영상 포즈렛에 이용하는 신체 관절은 다음 그림 2와 같다.

인간의 포즈를 구성하고 있는 신체 부분은 크게 머리, 몸, 팔, 손, 다리, 발 등으로 구분할 수 있으며, 서로 연결된 신체 부분은 포즈에 따라 특징을 나타낼 수 있다. 즉, 인간이 어떤 제스처를 취하거나, 어떤 포즈를 취할 때, 특정 신체 부분에서 특징을 나타내게 된다.

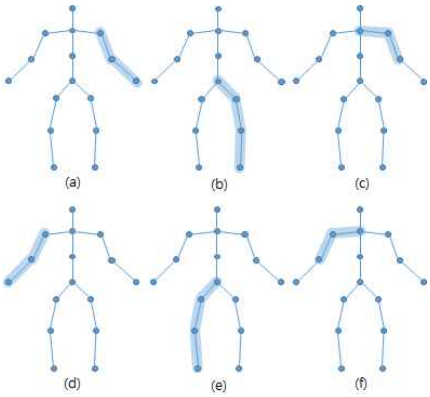

그림 2. 신체 구조에 따라 구분하는
신체 관절 기반 깊이 영상 포즈렛
(a: left arm, b: left leg, c: left upper body,
d: right arm, e: right leg, f: right upper body)

그러므로 인간의 신체 포즈에서 획득할 수 있는 관절의 상대적인 각도, 위치 등을 이용하여 신체 포즈에서 깊이 영상 포즈렛들을 찾아낼 수 있다.

본 논문에서 제안하는 깊이 영상 포즈렛은 제스처를 구성하고 있는 하나의 포즈를 신체 부분에 기준하여 분할한 것이다. 본 논문에서는 깊이 영상 포즈렛을 구성하기 위하여, 인간의 신체 관절을 위치에 따라 16개의 주요 관절로 정의하고 주요 관절에 레이블을 부착하였다. 그리고 하나의 신체 포즈에서는 그림 2와 같이 6개의 깊이 영상 포즈렛을 얻을 수 있다.

이러한 깊이 영상 포즈렛을 사용하면 신체 전신 포즈의 특징정보를 사용할 때보다 신체 부분 포즈의 특징정보를 사용하기 때문에 신체의 여러 움직임에 대해 정확한 특징을 구분할 수 있다. 그림 3에서는 본 논문에서 제안하는 깊이 영상 포즈렛과 그에 해당하는 3차원 관절 좌표를 그림으로 보여주고 있다.

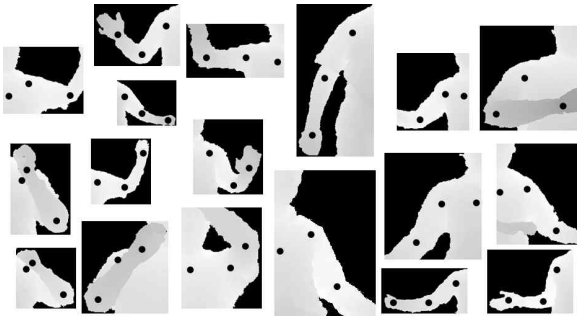

그림 3. 그림으로 표현되는 깊이 영상 포즈렛의 3차원 관절 좌표

Ⅲ. HMM을 이용한 제스처 인식

1. HMM

확률 이론에 바탕을 두고 있는 HMM은 관측되는 벡터 특징 사이의 시간 연관성이 존재하는 시계열 데이터 분류에 주로 사용되는 방법이다. 시계열 데이터는 음성 인식이나 필기체 인식 등과 같이 벡터 고유의 특징과 더불어 인접한 특징들의 연속적인 관계가 분류에 중요한 요소가 되는 데이터를 말한다. 그렇기 때문에 순차적인 데이터를 갖는 패턴을 분석하기 위해서는 시간 연속적인 데이터를 표현하는 방법과 데이터로부터 정보를 추출할 수 있는 모델이 필요하다.

HMM을 제스처 인식에 적용하기 위해서는 각 제스처 별로 학습을 수행해야 하며, 해당 제스처의 HMM 모델에 학습 결과를 적용하여야 한다. 그리하여 인식과정에서 인식하고자 하는 제스처와 HMM의 제스처 모델을 비교하여 가장 유사하다고 판단되는 가장 높은 확률을 보이는 제스처 모델을 최종 인식 결과로 출력한다.

HMM은 상태 전이 확률, 관측 확률, 초기 상태 확률 벡터, 이 세 가지 매개변수를 가지며 학습과정을 통해 매개변수의 값을 정하는 것이 HMM을 생성하는 과정이다. 그렇기 때문에 상태의 상태 전이 확률과 관측 벡터의 관측 확률을 잘 구성할 수 있는 HMM을 생성하여야 한다.

2. MM(Markov Model)과 HMM

HMM을 제스처에 적용하기 위해서는 MM과 HMM의 차이를 인지해야 한다. MM은 현재 상태에서는 관측되는 벡터가 상태가 같기 때문에 관측벡터에 따라 상태 전이를 알아낼 수 있어서 순차적인 데이터를 쉽게 처리할 수 있다. 하지만 그림 4와 같이 MM은 제스처를 구성하는 순차적 데이터가 다양해질수록 MM이 증가할 수밖에 없다. 반면, HMM은 관측되는 벡터가 어떤 상태에서 관측되는지 알 수 없지만, 학습한 데이터에 의해서 관측 벡터와 상태 이동에 대한 확률값을 얻을 수 있다. 그리고 HMM은 차수를 미리 고정하지 않고 모델 자체가 확률 프로세스에 따라 적응하여 먼 과거의 관측이라도 현재에 영향을 준다. 그렇기 때문에 모델을 잘 구성한다면 다양한 제스처에 대해서도 처리할 수 있다는 장점이 있다.

그러므로 제스처를 인식하기 위해 HMM의 상태를 구성하는 경우에는 신체 포즈와 모델의 상태를 동일하게 설정하기보다는 현재 상태에서 여러 신체 포즈가 관측될 수 있도록 구성하여야 한다. 또한 관측을 통해 얻을 수 있는 벡터와 상태가 같지 않도록 시간 연속적인 제스처 시퀀스도 상태의 이동을 고려해야 한다고 가정할 수 있다.

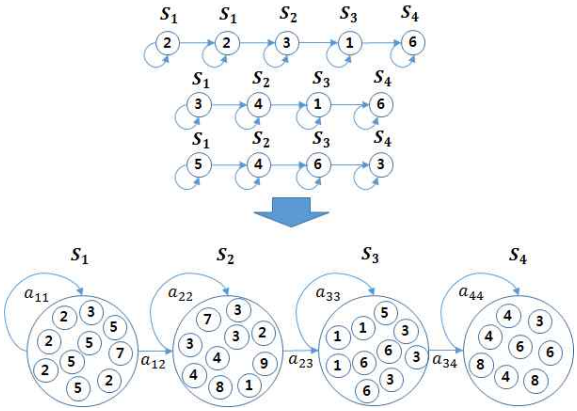

그림 4. 하나의 제스처 학습을 위한 MM(위)과 HMM(아래)

그림 4에서 보이는 숫자를 관측벡터라고 가정하였을 때, MM(위)은 각 상태에서 하나의 벡터만이 관측되기 때문에 관측벡터에 따라 상태의 이동을 알아낼 수 있는 반면, HMM(아래)은 벡터가 관측된다고 하더라도 어떤 상태에서 관측이 되었는지 알 수 없다. 그림 4에서는 S_1, S_2, S_3, S_4와 같은 상태들에는 비슷한 벡터가 군집되어 있다고 가정하였으며, 모든 상태에서 벡터가 관측되지 않지만, 유사한 형태를 가진 포즈가 군집되어 HMM의 상태를 구성할 수 있도록 HMM을 생성하여야 한다. 이는 인간의 제스처의 흐름이 시간의 연속성에 관계하고 있기 때문이다.

HMM의 상태 전이 모델은 좌우(Left-Right)모델과 어고딕(Ergodic)모델이 주로 사용된다. 그리고 상태가 왼쪽에서 오른쪽으로 전이되는 형태를 가진 좌우 모델의 경우에는 현재 상태로 회귀하는 부분의 존재 유무에 따라 LR모델과 LRB(Left-Right Banded)모델로 구분할 수 있다. Liu[11]는 LRB모델을 이용하여 비디오 속의 제스처를 인식하였다. 단지, 삼각형과 사각형, 두 개의 제스처만을 사용하였기 때문에 실험에 사용한 제스처가 충분하다고 할 수 없다. 그리고 삼각형과 사각형은 관측할 수 있는 방향벡터가 뚜렷한 차이를 보이므로 학습데이터가 충분하지 못하였다는 단점이 있다. Elmezain[12]

또한 LRB모델을 효율적이라고 주장하였으며, 모델을 구성하는 상태수를 조절하여 HMM을 생성하였다. 그렇기 때문에 LR모델을 이용하여 HMM을 생성할 경우에는 각각의 제스처마다 다른 LR모델을 적용해야 한다는 불편함이 존재한다. Kita[13]는 언어 인식을 수행하기 위하여 어고딕모델을 이용하였다. 9개의 말뭉치(corpus)를 상태로 설정하여 1700여개의 문장이 관측될 수 있는 어고딕모델을 구성하였다. 그리고 Kita가 사용한 어고딕모델은 시간의 흐름에 다소 독립적인 특징을 가지고 있다. Kumar[14]는 얼굴을 인식하기 위해 pseudo-2D 어고딕모델을 이용하였다.

3. 깊이 영상 포즈렛을 위한 HMMs

본 논문에서는 제스처를 구성하는 신체 포즈를 다음 식 1과 같이 나타낼 수 있다.

$$\begin{aligned} G_i &= \{P_1, P_2, \ldots, P_n\} \\ &= \{P_t\} \qquad (1 \le t \le n) \end{aligned} \tag{1}$$

그리고 신체 포즈는 다음 식 2와 같이 6개의 부분으로 구분할 수 있다.

$$\begin{aligned} P_t &= \{MP_1^t, MP_2^t, \ldots, MP_6^t\} \\ &= \{mP_1, mP_2, mP_3, mP_4, mP_5, mP_6\} \\ &= \{LA^t, RA^t, LL^t, RL^t, LUB^t, RUB^t\} \end{aligned} \tag{2}$$

결국 제스처는 6개의 부분 제스처로 구분할 수 있으며 식 3과 같이 나타낼 수 있다.

$$\begin{aligned} &mG^1, mG^2, mG^3, mG^4, mG^5, mG^6 \\ &mG^1 = \{mP_1^1, mP_1^2, mP_1^3, \ldots, mP_1^n\} \end{aligned} \tag{3}$$

본 논문에서는 신체를 6개의 부분으로 구분하였으므로, 하나의 제스처를 입력하였을 때, 6개의 부분 제스처로 구분이 가능하다. 이를 각각 6개의 HMM의 관측 벡터로 사용하면, 신체 부분에 따라 구분된 6개의 HMM의 결과로 얻을 수 있는 값은 어떤 제스처로 속하는지에 대한 확률값이라고 정의할 수 있다. 그리고 확률의 최대값을 선택하면, 각각의 신체 부분 제스처들을 통해 제스처를 유추할 수 있으며 이는 식4와 같다.

$$\begin{aligned} &P(mG_i^1|\Theta_i^1), P(mG_i^2|\Theta_i^2), P(mG_i^3|\Theta_i^3), \\ &P(mG_i^4|\Theta_i^4), P(mG_i^5|\Theta_i^5), P(mG_i^6|\Theta_i^6), \\ &(1 \le i \le 16), \Theta = HMM \end{aligned} \tag{4}$$

그림 5에서는 왼팔을 올리는 동작을 보이고 있으며, 이 때 왼쪽 상반신부분만이 유효한 확률값을 얻을 수 있는 신체 부분 제스처라고 할 수 있다.

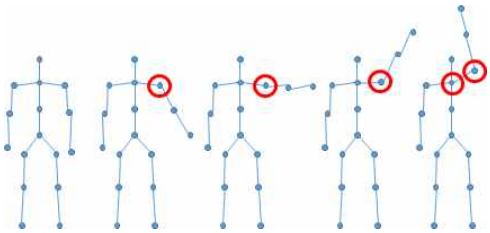

그림 5. 왼팔을 올리는 동작 (왼쪽 상반신만 각도가 변한다)

그리고 그림 6은 왼팔을 올리는 동작을 신체 구조에 따라 분리한 뒤, 각각을 표현한 것이다. 왼팔 부분(left arm)의 제스처는 움직임은 있지만 주요 관절의 변화된 값은 없기 때문에 IDLE상태로 표현된다.

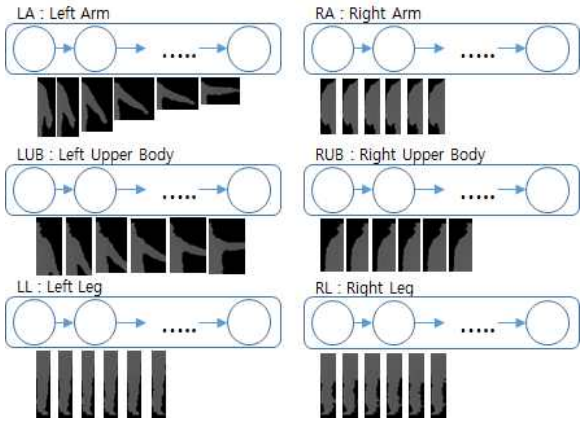

그림 6. 왼팔을 올리는 동작에서의 신체 부분 제스처

그리고 식 5는 신체 부분 제스처가 유효할 경우, 제스처에 대한 확률 값을 얻을 수 있으며, 유효하지 않은 제스처의 경우에는 IDLE상태로 표현되는 것을 보여준다.

$$\begin{aligned} &P(mG_i^1|\Theta_i^1) = IDLE, P(mG_i^2|\Theta_i^2) = IDLE, \\ &P(mG_i^3|\Theta_i^3) = G_1, \quad P(mG_i^4|\Theta_i^4) = IDLE, \\ &P(mG_i^5|\Theta_i^5) = IDLE, P(mG_i^6|\Theta_i^6) = IDLE, \\ &(1 \le i \le 16), \Theta = HMM \end{aligned} \tag{5}$$

즉, 식 5와 그림 6을 토대로 왼팔을 올리는 제스처에서 신체 부분 중 왼쪽 상반신 부분만이 유효한 값을 가지고 있다는 것을 알 수 있다. 이 제스처의 경우에는 신체 제스처의 특징을 보이는 신체 부분이 오직 왼쪽 상반신 부분이었기 때문에 베이즈 정리를 이용하지 않더라도 제스처를 유추할 수 있다. 하지만 각각 서로 다른 제스처 특징을 갖는 신체 부분 제스처들이 관측될 경우에는 HMM의 결과를 통합해야 한다. 그렇기

때문에 학습된 제스처와 비교하여 조건부 확률의 값을 곱했을 때, 가장 큰 값을 갖는 경우를 선택하여 제스처를 유추할 수 있다.

Ⅳ. 실험 내용

본 논문에서는 입력된 제스처를 HMM을 이용하여 인식한다. 입력된 제스처는 6개의 신체 관절 기반 포즈로 분할되었다고 가정하면, 각각의 신체 관절 기반 포즈를 관측하기 위해 각각의 제스처마다 6개의 HMM을 구성하였다. 이 때, 6개의 HMM이 같은 제스처를 결과로 보이면 상관없지만, 각각 다른 제스처를 결과로 보였을 경우, 베이즈 정리를 이용하여 제스처를 유추한다. 베이즈 정리 식은 다음 식 6과 같다.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{\text{우도} \times \text{사전확률}}{\text{사전확률}} \quad (6)$$

16개의 제스처를 정의하였을 때, 신체 제스처에서 신체 부분을 따라 분리된 신체 관절 기반 포즈는 각각의 제스처에 속할 확률을 가지게 된다. 즉, 본 논문에서는 베이즈 정리를 이용해 확률을 획득한 후, 가장 높은 확률을 보이는 제스처를 선택하는 것이 효율적이라고 판단한다. 베이즈 정리를 통해 사후확률이 최대값을 가질 때, 제스처라고 선택할 수 있지만, 본 논문에서는 6개의 신체 관절 기반 포즈를 통합하는 결과를 얻어야 하기 때문에, 베이즈 정리 식을 다시 쓰면, 다음과 같이 식 7로 나타낼 수 있다.

$$G = \arg\max_i P(G_i|mG_i^1, mG_i^2, mG_i^3, mG_i^4, mG_i^5, mG_i^6) \quad (7)$$
$$= \arg\max_i \prod_{k=1}^{6} \frac{P(mG_i^k|G_i)P(G_i)}{P(mG_i^k)}$$

학습된 제스처와 비교하여 조건부확률의 값을 곱했을 때, 가장 큰 값을 갖는 경우를 선택하여 제스처를 유추할 수 있으며, 신체 동작에 따라 중복되는 제스처를 줄이는 것도 가능하다. 식 7에서 mG_i^1부터 mG_i^6까지는 신체 부분을 관절 기반 부분 포즈로 분할했을 때 관측되는 신체 관절 기반 부분 포즈를 의미한다. 그리고 i는 정의한 16개의 제스처에서 가장 높은 확률값을 나타내는 제스처를 의미한다.

본 논문에서는 학습을 위해 키넥트를 이용하여 깊이 영상을 획득하였다. 키넥트를 사용하면 깊이 영상을 획득함과 동시에 3차원 관절 좌표를 획득할 수 있다. 그리고 2장에서 서술한 것처럼 신체 구조에 따라 영상을 분할하였다.

본 논문에서는 전부 12명의 사용자 제스처를 수집하였으며, 1명의 사용자가 제스처를 각각 10번씩 수행하였다. 그리하여 4명의 사용자 제스처는 HMM의 초기 매개변수들을 설정하기 위해 사용하였다. 그리고 나머지 8명의 사용자 제스처는 제스처 인식의 실험데이터로 사용하였다.

본 논문에서는 클러스터의 수 K를 6개에서 11개까지 변화시켜가면서 군집상태를 확인하였으며, 클러스터의 개수를 이용하여 HMM의 상태 수를 결정하였다. K-평균 클러스터링 과정은 다음과 같다. 입력받은 데이터들에서 초기의 K 개수만큼의 클러스터를 지정하였으며, 나머지 데이터들에서 가장 가까이 있는 중심을 찾는 작업을 반복적으로 수행한 뒤, 각각의 데이터들을 클러스터로 할당하고, 클러스터가 할당된 후에 할당된 클러스터의 중심이 다시 계산된다. 이 과정은 클러스터의 중심이 일정한 값으로 유지될 때까지 반복하였다.

K의 개수가 6일 때, 1명의 사용자가 Lift Left Arm 제스처를 5번 수행한 학습 데이터를 이용하여 K-평균 클러스터링을 수행한 결과이다. Lift Left Arm 제스처에서 유효한 신체 부분 제스처는 왼쪽 상반신(LUB)인데, 특징 벡터는 그림 7과 같이 뚜렷한 각도 변화를 나타내고 있다.

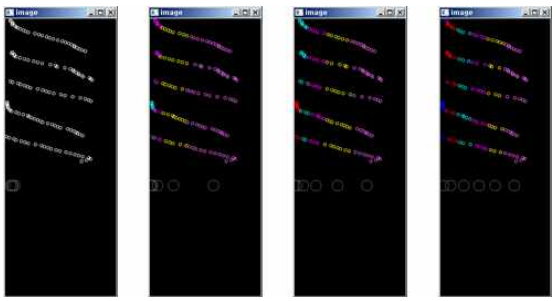

그림 7. K=6일 때, 왼팔을 올리는 동작에서의 왼쪽 상반신의 특징 벡터 시퀀스 (5번 수행)

실험 과정에서는 학습 과정과 마찬가지로 키넥트를 이용하여 깊이 영상 시퀀스를 획득하였으며, 학습된 깊이 영상 포즈렛을 적용하여 매 프레임의 부분 영상 특징을 획득하였다.

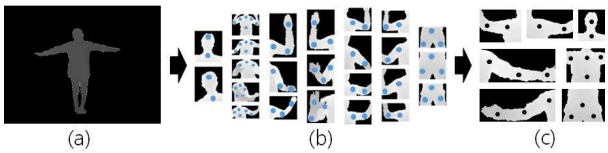

그림 8. (a) 입력 영상, (b) 깊이 영상 포즈렛, (c) 깊이 영상 포즈렛에서 입력 영상과 유사한 부분 영상과 3차원 깊이 영상 좌표

부분 영상의 특징을 획득한 뒤, 부분 영상의 시퀀스에 HMM을 적용하면 신체 부분 포즈의 부분 제스처를 인식할 수 있다. 각 신체 부분 시퀀스에 제스처에 해당하는 HMM을 적

용하면 6개의 신체 부분의 부분 제스처를 얻을 수 있다. 본 논문에서는 유효한 신체 부분의 부분 제스처와 IDLE 상태의 신체 부분 제스처를 구분하여 본 논문에서 제안하는 HMM과 베이즈 정리를 적용하였을 때, 연산량을 효과적으로 줄이면서 신체 제스처를 인식할수 있어야 할 것이다.

본 논문에서는 정의한 제스처를 인식하기 위해 수집한 사용자 제스처 뿐만 아니라, MSR Action 3D 키넥트 데이터셋을 이용하여 인식률을 시험한다. MSRAction 3D 데이터셋은 총 20가지 제스처로 구성되어 있으며, 10명의 사람이 각각의 제스처를 2번에서 3번 동작시킨 스켈레톤의 3차원 관절 좌표로 구성되어 있다.

V. 실험 결과

본 논문에서는 16개의 제스처를 정의하였으며, 키넥트를 이용하여 획득한 스켈레톤 정보를 이용하여 HMM을 구성한 뒤, 실험에 사용하였다. 각 HMM에서 얻어진 결과는 베이즈 정리를 이용하여 제스처를 인식할 수 있었다. 실험에 사용한 데이터는 HMM의 상태 개수가 8개일 때 인식률이 가장 좋았다. 그렇기 때문에 HMM의 상태 개수가 8개일 때, 정의한 제스처의 인식률을 표로 나타내면 표 1과 같다. 관절의 변화가 크게 달라지지 않은 왼쪽 무릎 올리기(15)나 오른쪽 무른 올리기(16)동작은 인식률이 저조한 것을 알 수 있다.

표 1. 상반신 부분 제스처와 하반신 부분 제스처 인식률

제스처		인식률
상반신	왼팔 들어 올리기	91%
	왼팔 접기	93%
	왼팔 돌리기	81%
	오른팔 들어 올리기	83%
	오른팔 접기	85%
	오른팔 접기	86%
	양 팔 들어 올리기	91%
	양 팔 접기	85%
	양 팔 들고 상체 왼쪽 기울기	86%
	양 팔 들고 상체 오른쪽 기울기	86%
하반신	왼발 들어 올리기	83%
	왼발 옆으로 들어 올리기	83%
	왼쪽 무릎 들어 올리기	85%
	오른발 들어 올리기	93%
	오른발 옆으로 들어 올리기	85%
	오른쪽 무릎 들어 올리기	78%

정확도(CDR)를 구하면 85.875%로 약 86%의 인식률을 보이는 것을 알 수 있다. 본 논문에서는 12명의 사용자 제스처를 수집하였으며, 각각 10번씩의 제스처 동작을 인식에 사용하였다. 본 논문에서는 신체 관절의 각도를 특징으로 사용하고 있으므로 각도 벡터에서 두드러진 특징 변화를 보이는 제스처는 인식이 순조로운 것을 알 수 있다.

그리고 키넥트를 이용하여 제작된 MSR Action 3D 데이터셋을 적용하여 65.35%의 인식률을 얻었다. 선행 연구에서는 HMM을 사용하였을 때, 63%의 인식률을 보였으며, 본 논문에서는 액션의 성분에 따라 각각 다른 인식률을 보였다. MSR Action 3D 데이터셋은 567개의 깊이 맵으로 구성되어 있으며 20개의 액션을 10명의 사용자가 각각 2번에서 3번 저장한 결과를 신체의 주요 관절 정보인 20개의 관절로 구성된 스켈레톤 정보를 가지고 있다. 본 논문에서는 2명의 191개의 액션을 학습에 사용하였으며, 8명의 476개의 액션을 실험에 사용하였다. 그리하여 8명이 2번 혹은 3번씩 20개의 액션을 사용하였다고 가정했을 때, 1개의 제스처마다 24번의 횟수를 가진다고 할 수 있다.

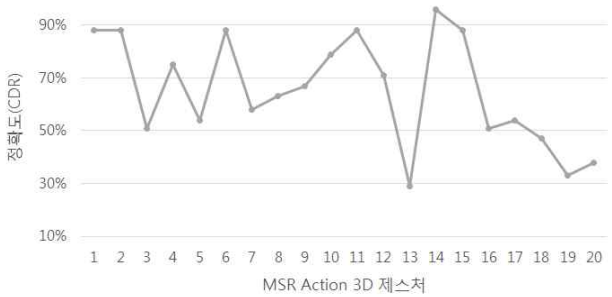

그림 9. MSR Action 3D 제스처 데이터셋을 적용한 인식률

본 논문에서는 신체 관절의 각도를 특징으로 사용하고 있으므로 각도 벡터에서 두드러진 특징 변화를 보이는 마이크로 제스처는 인식이 순조로운 것을 알 수 있다. MSR Action 3D 데이터셋은 3개의 Action 클래스로 나눌 수 있는데 MSR Action 3D 데이터셋에서 허리 부분의 관절을 사용하는 제스처(13:bend, 17:tennis swing, 18:tennis serve, 19:golf swing, 20:pickup&throw)는 인식률이 저조하였다. 반면에 본 논문에서 제안하는 관절 벡터의 회전 정보로 표현이 가능한 제스처들(1:high arm wave, 2:horizontal arm wave, 6:high throw, 10:hand clap, 11:two hand wave, 14:forward kick, 15:side kick)은 높은 인식률을 보이고 있다.

하지만 본 논문에서 정의한 신체 관절 기반 포즈는 6개의 관절 정보만을 사용하고 있어서 오픈 데이터셋에 적용하기에 아직 어려움이 있으므로 다양한 관절을 사용할 수 있어야 할 것이다.

VI. 결론

순차적인 포즈를 이용하여 제스처를 인식하는 연구의 경우에는 제스처를 구성하는 포즈가 연속적으로 입력될 때마다 이전

포즈와의 상태 전이 확률을 이용하여 제스처를 인식할 수 있으므로 포즈 인식의 중요도가 무척 높아진다. 하지만 인간의 포즈는 자유도가 높고 영상 프레임을 획득하면서 손실이 생기는 경우가 많으므로 포즈 인식은 쉽게 해결하기 어려운 문제이다. 그렇기 때문에 본 논문에서는 정확한 포즈 인식이 어렵다는 가정하에 신체의 부분 포즈를 기반으로 하는 신체 부분 제스처를 인식하였다.

깊이 영상 포즈렛을 이용하여 신체 포즈를 분리하는 과정에서는 실제로 크게 변화가 없는 신체 부분까지 부분 포즈로 분리하였기 때문에 제스처 인식률에 크게 영향을 끼치지 않거나 인식과정에서 불필요한 값을 더하는 신체 부분이 발생하였다.

본 논문에서는 신체를 신체 구조에 따라 작은 단위로 분할하여 깊이 영상 포즈렛을 제작하였으며, 가장 두드러진 특징을 보이는 6개의 신체 부분 포즈(LA: left arm, RA: right arm, LL: left leg RL: right leg, LUB : left upper body, RUB : right upper body)를 이용하여 결과를 보였다.

앞서 언급한 것과 같이 전신 제스처 대신 신체 부분 제스처를 사용하게 되면 제스처에서 움직임을 가지고 있는 신체 부분을 인식하는 것만으로도 제스처를 인식할 수 있다. 그렇기 때문에 신체 부분을 분할하고 유효한 부분 제스처만을 인식에 사용하는 것은 효율적이라고 할 수 있다.

향후에는 이 연구결과를 토대로 신체의 움직임을 잘 표현할 수 있는 신체 부분을 분할하여 정확한 포즈 인식과 함께 정확한 제스처 인식을 수행하도록 하겠다.

References

[1] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

[2] Bourdev, Lubomir, and Jitendra Malik. "Poselets: Body part detectors trained using 3d human pose annotations." Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009.

[3] Kraft, Erwin, and Thomas Brox. "Motion Based Foreground Detection and Poselet Motion Features for Action Recognition." Computer Vision--ACCV 2014. Springer International Publishing, 2015. 350-365

[4] Maji, Subhransu, Lubomir Bourdev, and Jitendra Malik. "Action recognition from a distributed representation of pose and appearance." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[5] Felzenszwalb, Pedro F., and Daniel P. Huttenlocher. "Efficient matching of pictorial structures." Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on. Vol. 2. IEEE, 2000.

[6] Li, Bo, et al. "Part-based pedestrian detection using grammar model and ABM-HoG features." Vehicular Electronics and Safety (ICVES), 2013 IEEE International Conference on. IEEE, 2013.

[7] Wang, Chunyu, Yizhou Wang, and Alan L. Yuille. "An approach to pose-based action recognition." Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013.

[8] Yang, Yi, and Deva Ramanan. "Articulated human detection with flexible mixtures of parts." Pattern Analysis and Machine Intelligence, IEEE Transactions on 35.12 (2013): 2878-2890.

[9] Desai, Chaitanya, and Deva Ramanan. "Detecting actions, poses, and objects with relational phraselets." Computer Vision–ECCV 2012. Springer Berlin Heidelberg, 2012. 158-172.

[10] Wang, Yang, Duan Tran, and Zicheng Liao. "Learning hierarchical poselets for human parsing." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[11] Liu, Nianjun, et al. "Understanding HMM training for video gesture recognition." TENCON 2004. 2004 IEEE Region 10 Conference. IEEE, 2004.

[12] Elmezain, Mahmoud, et al. "A hidden markov model-based continuous gesture recognition system for hand motion trajectory." Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. IEEE, 2008.

[13] Kita, Kenji, et al. "Automatic acquisition of probabilistic dialogue models." Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on. Vol. 1. IEEE, 1996.

[14] Kumar, S., D. R. Deepti, and Ballapalle Prabhakar. "Face recognition using pseudo-2D ergodic HMM." Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. Vol. 2. IEEE, 2006.

저 자 소 개

박재완(학생회원)

2007년 호남대학교 정보통신공학과 학사 졸업.
2009년 전남대학교 전자컴퓨터공학과 석사 졸업.
2016년 전남대학교 전자컴퓨터공학과 박사 졸업.

<주관심분야 : 휴먼제스처, HCI, 패턴인식 등>

이칠우(정회원)

1992년 동경대학교 대학원 전자공학과 졸업. (공학박사)
1996년 1월~현재 전남대학교 전자컴퓨터공학부 교수.
2002년 1월~2003년 2월 미국 NC A&T State University 방문교수.
2006년 3월~2008년 2월 정보통신부 자체평가위원.
2006년 3월~현재 전남대학교 문화콘텐츠기술연구소 소장
2009년 3월~현재 전남대학교 차세대휴대폰인터페이스연구센터(ITRC) 센터장.

<주관심분야 : 컴퓨터 비전, 지능형 휴먼 인터페이스, 디지털 콘텐츠, 컴퓨터그래픽스 등>