핵심-포즈 분포 기반 다중 시점에서의 휴먼 행동 인식

김선우 [a], 석흥일 [a,O], 이성환 [a,b]
[a] 고려대학교 컴퓨터학과
[b] 고려대학교 뇌공학과
{swkim, hisuk, swlee}@image.korea.ac.kr

Human Action Recognition in Various Viewpoints with a Key-Pose Distribution

Sun-Woo Kim[a], Heung-Il Suk[a,0], Seong-Whan Lee[a,b]
[a]Department of Computer Science and Engineering, Korea University
[b]Department of Brain and Cognitive Engineering, Korea University

요 약

휴먼 행동 인식은 크게 3D 모델 기반 방법과 템플릿 기반 방법으로 나눌 수 있다. 3D 모델 기반 방법은 휴먼의 포즈를 3D로 재구성한 뒤 특징을 추출하는 것으로 인식 정확도는 높으나 연산량이 많아 매우 비효율적이다. 반면 템플릿 기반의 방법은 간단하고 수행 시간이 빠르기 때문에 여러 논문들에서 채택되고 있다. 그러나 템플릿을 이용한다는 특성 때문에 시점, 행동 스타일의 변화 등에 따라 실루엣의 변화가 심해 인식 성능에 한계점을 가진다. 본 논문에서는 핵심-포즈들의 히스토그램으로 표현되는 핵심-포즈 분포와 광류의 변화를 이용하여 다중 시점에서의 휴먼 행동 인식 방법을 제안한다. 제안하는 방법은 IXMAS 데이터 셋을 이용한 실험에서 적은 수의 템플릿을 이용하면서도 평균 87.9%의 높은 인식률을 보였다.

1 서론

휴먼 행동 인식은 컴퓨터 비젼 분야에서 가장 활발한 연구 내용 중 하나로 비디오 서베일런스 시스템, 휴먼-컴퓨터 상호작용, 비디오 인덱싱, 스포츠 비디오 분석 등과 같은 다양한 분야에서의 응용이 가능하다. 행동 인식에서의 어려움은 시점에 따라 변화하는 외형, 휴먼마다 다른 행동 스타일, 카메라와의 거리 변화에 따른 휴먼 객체의 크기 등에 따른 문제들에 기인한다. 이러한 문제들을 해결하기 위해 많은 연구들이 수행되어 왔으며, 우수한 인식 성능을 보이고 있다[1,2,3]. 그러나 시점의 변화에 따라 달라지는 외형의 변화는 여전히 극복하지 못하고 있는 실정이다.

휴먼 행동은 시공간에서의 3차원 패턴으로 표현될 수 있으며, 이러한 패턴을 인식하기 위한 방법은 크게 두 가지로 나눌 수 있다. 첫 번째는 인체 구성 요소의 변화 특성을 추출하여 인식하는 방법이다. 이는 인체 구성 요소의 위치 변화를 추적하기 위한 초기화 과정이 필요하고, 가려짐과 같은 상황에서 추적에 실패할 경우 인식을 수행할 수 없다. 두 번째는 외형 정보를 이용하여 미리 정의된 템플릿과의 매칭을 이용하는 방법이다. 첫 번째 방법과 달리 초기화 과정이 필요하지 않으며, 고정된 시점에서는 매우 정확한 인식 성능을 보인다. 하지만 템플릿을 이용한다는 특성 때문에 시점, 행동 스타일의 변화 등에 따라 실루엣의 변화가 심한 경우 인식을 제대로 하지 못하는 문제점을 가지고 있다. 본 논문에서는 이러한 문제점을 해결하기 위한 방법을 제안한다.

Bobick과 Davis는 검출된 실루엣들을 2차원 영상에 누적하여 표시한 Motion History Image를 제안하였으며, 입력 비디오 영상과 미리 정의된 템플릿과의 비교를 통해 인식을 수행하였다[1]. 이는 행동 패턴을 간단한 연산만으로 표현하는 좋은 방법이나 배경의 변화나 행동의 기하학적 변화에 민감하다.

이차원 영상으로부터의 특징 추출의 한계성을 극복하기 위해 Efros 등은 입력 프레임 시퀀스에서 검출한 휴먼 객체의 실루엣을 시공간으로 구성된 3차원 볼륨으로 표현하였으며, 광류의 변화를 이용한 볼륨 매칭 방법을 제안하였다[4]. 그러나 이는 노이즈, 외형 변화, 자기 가려짐 현상에 취약하다는 단점을 가지고 있다. Ahmad와 Lee는 인체 변화의 지역적 특성과 전역적 특성을 광류로 표현하여 시점의 변화에 강인한 인식 방법을 제안하였으나[5], 여러 대의 카메라를 활용해야 한다는 한계점을 가지고 있다. Kim과 Cipolla는 비디오 볼륨 텐서에 정준상관분석(Canonical Correlation Analysis) 기법을 적용한 새로운 형태의 인식 방법을 제안하였다[6]. 그러나 카메라의 위치에 대한 사전 정보를 알고 있어야 하는 제약사항이 있다. 또한, Weinland 등은 여러 대의 카메라에서 획득한 실루엣 영상들을 3차원 Visual Hull로 표현하고, 은닉 마르코프 모델을 인식기로 사용함으로써 시점 변화에 강인한 방법을 제안하였다[3]. 이 방법은 4절의 실험 부분에서 본 논문에서 제안하는 방법과 비교 평가될 것이다.

기존의 대부분의 연구에서는 정확한 인식을 위해 휴먼이 카메라에 정면으로 향하고 있거나, 여러 대의 카메라를 사용해야 한다는 제약 사항이 있다. 현실적인 환경에서의 인식을 위해 다양한 시점 변화에 강인하고, 휴먼과 카메라와의 상대적 위치 변화에도 강인한 인식 방법이 필요하다. 본 논문에서는 다양한 시점에서의 행동 인식을 위해 사전에 정의한 여러 개의 시점에서 행동 패턴에 대한 템플릿을 생성하였으며, 효율적이고 빠른 매칭의 수행을 위해 핵심-포즈 분포를 이용한 인식 방법을 제안한다.

논문의 구성은 다음과 같다. 2절에서는 핵심-포즈를 통한 행동 특성 표현 및 포즈 분포와 광류를 이용한 행동 인식 방법을 설명한다. 3절에서는 제안하는 방법과 기존 방법과의 성능 평가를 위한 비교 실험을 수행하고, 4절에서 결론을 맺는다.

2 행동 인식

2.1 핵심-포즈 추출

핵심-포즈는 서로 다른 휴먼 객체들마다의 행동 스타일의 문제를 해결하기에 좋은 방법이다[7, 8]. 기존의 핵심-포즈 추출 방법은 광류의 변화[7], 움직임 에너지[8]와 같은 움직임 정보에 기반하였다. 그러나 이러한 방법들은 전처리 과정에서의 오차로 인해 발생하는 의미 없는 움직임 정보들을 그대로 반영하는 단점을 가지고 있다. 따라서 본 논문에서는 실루엣에 기반한 핵심-포즈 추출 방법을 제안한다.

핵심-포즈를 추출하기 위해 우선, 첫 번째 프레임을 핵심-포즈 셋 K에 포함시킨다. 새로운 입력 프레임에 대해서 현재 설정된 핵심-포즈들과의 유사도를 평가한다. 이때, 유사도는 Hu 모멘트[9]를 이용하여 다음과 같이 계산한다.

$$I(A,B)=\sum_{i=1}^{7}\left|\frac{1}{m_i^A}-\frac{1}{m_i^B}\right|$$

A와 B는 서로 다른 영상이며, $m_i^A = sin(h_i^A)\cdot log(h_i^A)$, $i=\{1,2,\ldots,7\}$이다. h_i^A는 영상 A에서 추출된 Hu 모멘트 값을 의미한다.

현재까지 설정된 핵심-포즈들과 입력 영상에서 추출한 실루엣 X 간의 유사도가 미리 설정해 놓은 임계치 θ보다 작은 경우에 입력 영상의 실루엣을 포함하는 새로운 핵심-포즈 셋 $\hat{K}$을 생성한다.

$$\hat{K}=\begin{cases}K\cup X & \text{if } I(X,B)\le\theta,\ \forall B\in K\\ K & \text{otherwise}\end{cases}$$

유사도의 평가를 모든 핵심-포즈들에 대해 수행하는 이유는 서로 다른 행동들에서 유사한 포즈들이 나타날 수 있기 때문이다. 예를 들면, 발차기 행동에서 발을 뻗는 행동과 반대로 발을 바닥으로 내려놓는 행동은 순서만 다른 뿐 구성되는 포즈가 동일하다.

각각의 시점에서 선택된 행동들에서 추출된 핵심-포즈들을 이용하여 핵심-포즈 셋을 구성한다. IXMAS[8]의 행동 데이터베이스에서 검출된 핵심-포즈가 그림 1에 나타나 있다.

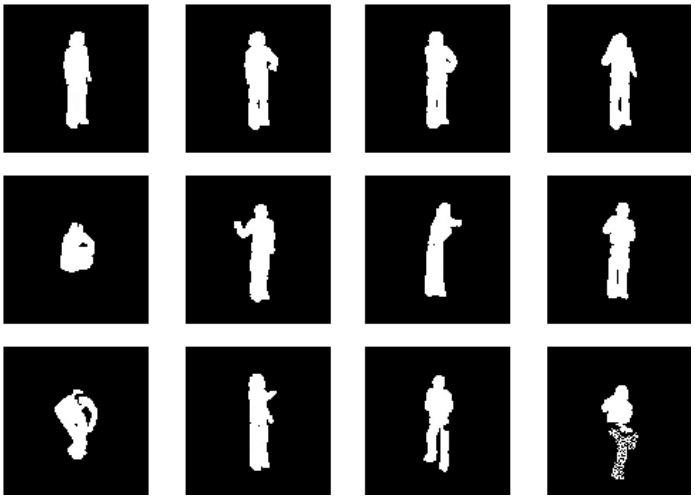


그림 1. IXMAS DB 중 시점 1에서 촬영된 행동들에 대한 핵심-포즈 셋

2.2 특징 추출

입력 영상에서 실루엣을 추출한 뒤, 포즈의 특징을 표현하기 위해 R-transform 방법[10]을 적용한다. R-transform은 Radon transform[11]을 기반으로 하는데 Radon transform은 이차원 영상에 나타난 패턴을 여러 각도에서의 직선들에 사영하여 생기는 히스토그램들의 누적함수로 정의된다. 이는 크기, 회전, 변형에 강인하나 기하학적 변형에 약한 특성을 가진다. 반면 R-transform은 이를 보완하여 기하학적 변형에도 강인하다는 특성을 가지며 다음과 같이 정의된다.

$$R_f(\theta)=\int_{-\infty}^{\infty}T_{Rf}^2(\rho,\theta)d\rho$$

여기서 T_{Rf}^2는 f의 Radon transform을 의미한다. 그림 2는 두 개의 실루엣에 대해 Radon transform과 R-transform을 적용한 결과 영상을 보이고 있다.

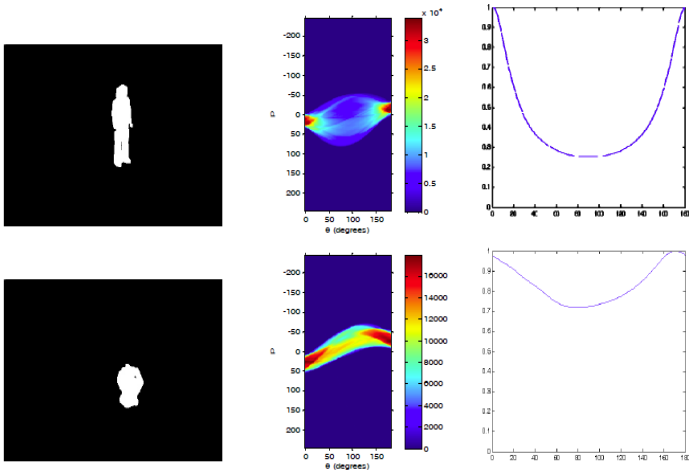
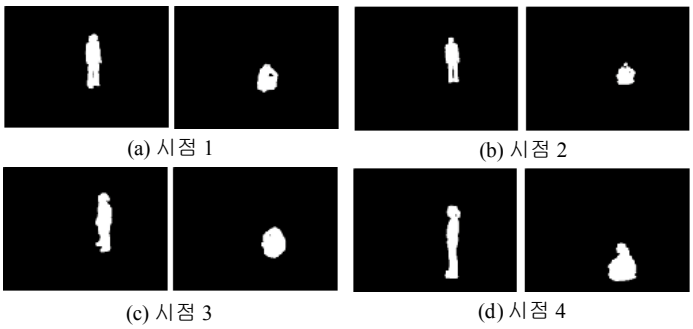


그림 2. Radon transform과 R-transform 결과 예: (왼쪽) 입력 실루엣, (가운데) Radon transform 결과, (오른쪽) R-transform 결과

2.3 행동 인식

2.3.1 시점 결정

템플릿 매칭을 이용한 행동 인식에서 시점을 결정하기 위해 기존 연구에서는 입력 비디오 시퀀스의 첫 프레임[12] 또는 입력 프레임 전체를 각 시점에 대해 정의된 템플릿들과 비교하였다[7,8]. 하지만 한 프레임으로 시점을 결정하기에는 신뢰도가 떨어지고 모든 프레임을 다 고려하기에는 계산량이 많다는 단점을 가지고 있다. 이에 본 논문에서는 비디오 영상에서의 첫 다섯 프레임과 마지막 다섯 프레임을 이용하여 각 프레임마다 시점을 결정한 뒤, 투표(voting) 방법을 이용하여 이런 문제점을 해결한다. 투표 결과 두 개 이상의 시점에 대해 동일한 표가 나올 경우, 처음과 마지막 프레임에서 한 프레임씩 추가하여 다시 결정을 하도록 한다. 그림 3은 '앉기'에 대해 서로 다른 시점에서의 포즈 셋을 보여주고 있다.



(a) 시점 1 (b) 시점 2

(c) 시점 3 (d) 시점 4

그림 3. '앉기'에 대한 서로 다른 시점에서의 포즈 셋

2.3.2 포즈 분포

입력 비디오 시퀀스에 대해 선택된 시점에 대한 템플릿들과의 매칭을 통하여 인식을 수행한다. 그러나 입력 실루엣과 템플릿 간의 일대일 비교를 수행하는 기존 방식은 적은 수의 프레임들에서 발생한 오차의 누적값으로 인해 인식에 실패하는 원인이 된다. 이를 해결하기 위해 각 행동을 구성하는 핵심-포즈들의 분포를 이용한다. 각 프레임에서의 실루엣을 사전에 학습된 핵심-포즈들 중 하나로 할당하고, 이를 이용하여 입력 비디오 시퀀스에 대한 핵심-포즈 히스토그램을 생성한다. 또한 휴먼 객체마다 서로 다른 행동 수행 시간을 감안하여 이를 전체 프레임수로 정규화시키며, 이를 핵심-포즈 분포라 정의한다. 입력 시퀀스에 대한 핵심-포즈 분포를 각 행동들에 대해 사전에 학습된 분포와 비교함으로써 행동을 인식한다. 그림 4는 포즈 분포의 생성 과정을 보이고 있다.

제안하는 핵심-포즈 분포는 행동 수행 과정에 나타나는 전체적인 포즈 구성을 분석하는 방식으로 소수의 프레임들에서의 매칭이 실패하더라도 전체적인 구성에는 큰 변화가 없다는 점에서 기존의 일대일 매칭 방법보다 매칭 실패 시 발생하는 인식 오류를 극복할 수 있다는 장점을 가진다.

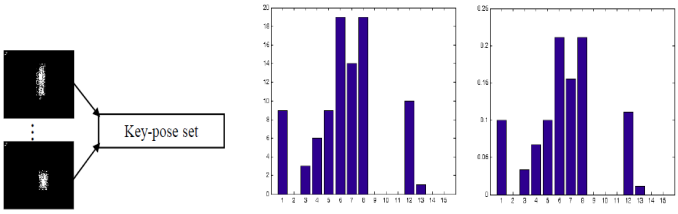


그림 4. 핵심-포즈 분포의 생성: (왼쪽) 핵심-포즈 매칭, (가운데) 핵심-포즈에 대한 히스토그램, (오른쪽) 정규화된 핵심-포즈 분포

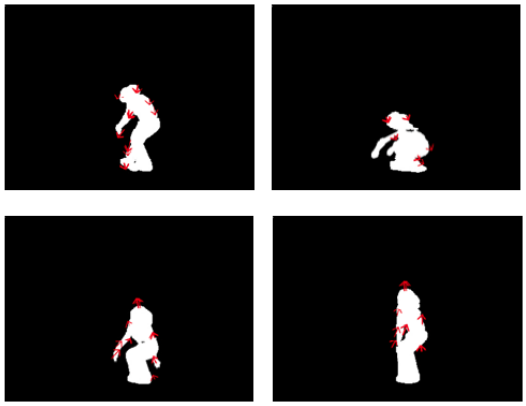


그림 5. '앉기'와 '일어서기'에 대한 광류 예

2.3.3 광류에 기반한 모션 정보

핵심-포즈 분포를 이용한 행동 인식 방법은 매칭에서의 오류에 강인하다는 특성을 가지지만, '앉기'와 '일어서기'와 같이, 행동을 구성하는 핵심-포즈 분포가 유사한 경우에는 인식을 하지 못하는 단점을 가지고 있다. 그러나 이는 행동에 대한 모션 정보를 이용하여 해결이 가능하며, 모션 정보를 표현하기 위해 광류를 이용한다.

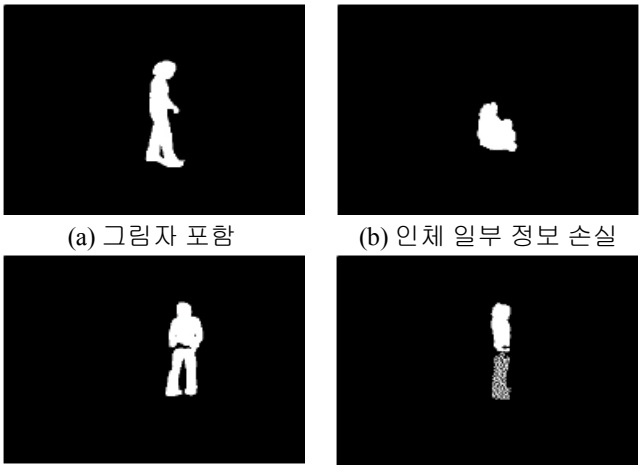
광류는 모션 정보를 잘 나타내기는 하지만 잡음에 취약하다는 단점이 있다. 이를 극복하기 위해서 각 행동에

의 주 움직임을 설정해 놓고, 입력 행동 시퀀스 역시 주 움직임에 해당하는 광류 정보만을 이용한다. 주 움직임을 선택하기 위해 광류의 방향을 46°~135°, 136°~225°, 226°~315°, 316°~45°와 같이 크게 4가지 방향으로 나누었다. 그림 5는 '앉기'와 '일어서기'에 대한 광류를 보이고 있다. 입력 시퀀스에서 획득한 모션 정보를 포즈 분포와 비슷하게 히스토그램으로 표현하여 각 행동들에 대해 미리 학습된 히스토그램과 비교하여 행동을 결정한다.

3 실험 결과 및 분석

3.1 실험 데이터 셋

본 논문에서는 INRIA의 IXMAS 데이터 셋[8]을 이용하여 실험하였다. 10개의 행동에 대해 10명의 휴먼 객체들이 3번씩 각 행동을 수행하였으며, 서로 다른 시점에 위치한 4대의 카메라를 이용하여 동시에 촬영하였다. 10개의 행동은 '서기', '시계보기', '팔짱 끼기', '머리 긁기', '앉기', '일어서기', '손 흔들기', '주먹질', '발차기', '가리키기'이다. 각 입력 영상에 대한 실루엣은 원본 데이터와 함께 제공되나 그림 6에 보여진 바와 같이 그림자가 포함되거나 잡음이 섞이는 등 실제 응용에서 획득되는 실루엣의 형태와 유사하여 이를 그대로 사용하였다.



(a) 그림자 포함 (b) 인체 일부 정보 손실

(c) 조각난 실루엣 (d) 잡음이 섞인 실루엣

그림 2. 그림자 및 잡음이 포함된 실루엣 영상 예

3.2 인식 결과 및 분석

각 행동으로부터 핵심-포즈를 추출하고, 핵심-포즈 분포를 생성하기 위해 한 명의 휴먼 객체에서 획득된 30개의 비디오 클립을 이용하였으며, 나머지 9명에 대한 클립으로 인식 성능을 평가하였다. 인식 결과는 표 1에 나타나 있다.

제안한 방법에 대한 인식 성능은 평균 87.9%를 보였으며, 동일한 데이터 셋에 대해 Weinland 등[3]은 93%의 성능을 보였다. 이 수치는 Weinland 등의 논문에서 발췌한 값이다. 제안한 방법의 인식 성능이 상대적으로 낮기는 하지만 적은 수의 템플릿을 사용함에도 불구하고 높은 인식률을 보였으며, 수행 속도도 빠르다는 장점을 가진다. 반면 Visual Hull을 생성하여 3차원으로 행동을 표현한 뒤 인식을 수행하는 Weinland 등의 방법은 계산량이 많아 실 세계로의 응용에는 제약이 있다.

인식 결과를 살펴 보면 '앉기', '일어서기', '주먹질', '발차기' 행동에 대해 상대적으로 높은 인식률을 보였다. 이는 이들 행동들이 다른 행동들에 비해 실루엣의 변화가 크기 때문이다. 이에 반해 '시계보기', '팔짱 끼기', '머리 긁기'는 상대적으로 낮은 인식률을 보였다. 이는 외형의 변화가 다른 행동에 비해 작으며, 신체의 일부로 다른 신체를 가리는 현상이 많이 발생하여 행동의 특성을 충분히 표현하지 못하기 때문으로 분석된다. 여러 가지의 시점 중에서, 4번째 시점에 대한 비디오 클립에서 가장 높은 인식률을 보이는데 이는 행동을 취한 사람과 평행한 측면에 위치에 있기에 다른 시점에 비해 움직임이 좀더 뚜렷이 나타났기 때문이다. 하지만 4번째 시점에서 '손 흔들기'는 낮은 인식률을 보이는 데 이는 손을 흔들 때 몸통에 가려지는 현상이 발생하여 '서기' 행동으로 잘못 인식되는 경우가 많았기 때문이다.

표 1. 제안하는 방법에 대한 행동 인식 결과

	시점 1	시점 2	시점 3	시점 4	**Overall**
서기	81.5%	81.5%	85.2%	92.6%	**85.2%**
시계 보기	81.5%	81.5%	81.5%	92.6%	**84.2%**
팔짱 끼기	85.2%	85.2%	81.5%	85.2%	**84.2%**
머리 긁기	81.5%	85.2%	85.2%	85.2%	**84.2%**
앉기	88.9%	85.2%	88.9%	92.6%	**88.9%**
일어서기	92.6%	88.9%	85.2%	92.6%	**89.8%**
손 흔들기	85.2%	85.2%	88.9%	85.2%	**86.1%**
주먹질	96.3%	92.6%	96.3%	96.3%	**95.3%**
발차기	96.3%	88.9%	92.6%	96.3%	**93.5%**
가리키기	85.2%	85.2%	88.9%	92.6%	**87.9%**
Overall	**87.4%**	**85.9%**	**87.4%**	**91.1%**	**87.9%**

4 결론

본 논문에서는 실루엣을 이용한 외형 변화와 광류를 이용한 모션 정보에 기반한 휴먼 행동 인식 방법을 제안하였다. Hu 모멘트[9]에 기반한 실루엣 간의 비교를 통해 핵심-포즈를 하였다. 또한 Radon transform[12]을 이용하여 포즈에 대한 특징을 나타냈으며, 입력 비디오 시퀀스에 대한 핵심-포즈들의 히스토그램으로 핵심-포즈 분포를 구성하여 인식을 수행하였다. 제안하는 방법은 템플릿과 입력 영상과의 일대일 매칭을 수행하는 기존 방법과 달리 핵심-포즈 분포 간의 매칭을 수행함으로써 매칭 과정에서 발생하는 오류로 인한 인식 성능 저하를

줄일 수 있다는 장점을 가진다.

IXMAS 데이터 셋을 이용한 실험에서 평균 87.9%를 보였으며, 이는 적은 수의 템플릿을 이용함에도 불구하고 높은 인식률을 보인 것이다. 적은 수의 템플릿을 사용하는 만큼 알고리즘의 수행 속도도 빠르다는 특성을 가진다. 하지만 실루엣을 이용하여 인식을 수행하기 때문에 포즈들이 신체 구성 요소 간의 가려짐을 포함하거나 행동이 몸통 내부에서만 일어나는 경우에는 인식에 실패하는 문제점이 있다. 이를 해결하기 위한 방안은 향후 연구로 남겨둔다.

Reference

[1] A. Bobick and J. Davis, "The Representation and Recognition of Action Using Temporal Templates," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3, 2001, pp. 257-267.

[2] P. Peursum, S. Venkatesh, and G. West, "Tracking-as-recognition for Articulated Full-body Human Motion Analysis," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, USA, June 2007, pp. 1-8.

[3] D. Weinland, E. Boyer, and R. Ronfard, "Action Recognition from Arbitrary Views using 3D Exemplars," *Proc. IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1-7.

[4] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance," *Proc. IEEE International Conference on Computer Vision*, Nice, France, Oct. 2003, No.2, pp. 726-733.

[5] M. Ahmad and S.-W. Lee, "Human Action Recognition Using Shape and CLG-Motion Flow from Multi-View Image Sequences," *Pattern Recognition*, Vol. 41, No. 7, 2008, pp. 2237-2252.

[6] T. Kim and R. Cipolla, "Canonical Correlation Analysis of Video Volume Tensors for Action Categorization and Detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 8, 2009, pp. 1415-1428.

[7] F. Lv and R. Nevatia, "Single View Human Action Recognition using Key Pose Matching and Viterbi Path Searching," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, USA, June 2007, pp. 1-8.

[8] A. Ogale, A. Karapurkar and Y. Aloimonos, "View-invariant Modeling and Recognition of Human Actions using Grammars," *Proc. IEEE International Conference on Computer Vision, Workshop on Dynamical Vision*, Beijing, China, Oct. 2005.

[9] M. Hu, "Visual Pattern Recognition by Moment Invariants," *IRE Trans. on Information Theory*, Vol. 8, No. 2, 1962, pp. 179-187.

[10] IXMAS DB: http://4drepository.inrialpes.fr/dataset.php?dspath=inria/ixmas

[11] S. Tabbone, L.Wendling, and J. Salmon, "A New Shape Descriptor Defined on the Radon Transform," *Computer Vision and Image Understanding*, Vol. 102, No. 1, 2006, pp. 42-51.

[12] S. Deans, *Application of the Radon Transform*, Wiley Interscience Publications, New York, Feb. 1983.