ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

시공간 2D 특징 설명자를 사용한 BOF 방식의 동작인식☆

BoF based Action Recognition using Spatio-Temporal 2D Descriptor

김 진 옥[1]
JinOk KIM

요 약

동작인식 연구에서 비디오를 표현하는 시공간 부분 특징이 모델 없는 상향식 방식의 주요 주제가 되면서 동작 특징을 검출하고 표현하는 방법이 여러 연구를 통해 다양하게 제안되고 있다. 그 중에서 BoF(bag of features)방식은 가장 일관성 있는 인식 결과를 보여주고 있다. 비디오의 동작을 BoF로 나타내기 위해서는 어떻게 동작의 역동적 정보를 표현할 것인가가 가장 중요한 부분이다. 그래서 기존 연구에서는 비디오를 시공간 볼륨으로 간주하고 3D 동작 특징점 주변의 볼륨 패치를 복잡하게 설명하는 것이 가장 일반적인 방법이다. 본 연구에서는 기존 3D 기반 방식을 간략화하여 비디오의 동작을 BoF로 표현할 때 비디오에서 2D 특징점을 직접 수집하는 방식을 제안한다. 제안 방식의 기본 아이디어는 일반적 공간프레임의 2D xy 평면뿐만 아니라 시공간 프레임으로 불리는 시간축 평면에서 동작 특징점을 추출하여 표현하는 것으로 특징점이 비디오에서 역동적 동작 정보를 포착하기 때문에 동작 표현 특징 설명자를 3D로 확장할 필요 없이 2D 설명자만으로 간단하게 동작인식이 가능하다. SIFT, SURF 특징 표현 설명자로 표현하는 시공간 BoF 방식을 주요 동작인식 데이터에 적용하여 우수한 동작 인식율을 보였다. 3D기반의 HoG/HoF 설명자와 비교한 경우에도 제안 방식이 더 계산하기 쉽고 단순하게 이해할 수 있다.

☞ 주제어 : 동작 인식, 특징 단어장(BoF), 시공간 동작 특징 검출기, 특징 설명자

ABSTRACT

Since spatio-temporal local features for video representation have become an important issue of modeless bottom-up approaches in action recognition, various methods for feature extraction and description have been proposed in many papers. In particular, BoF(bag of features) has been promised coherent recognition results. The most important part for BoF is how to represent dynamic information of actions in videos. Most of existing BoF methods consider the video as a spatio-temporal volume and describe neighboring 3D interest points as complex volumetric patches. To simplify these complex 3D methods, this paper proposes a novel method that builds BoF representation as a way to learn 2D interest points directly from video data.

The basic idea of proposed method is to gather feature points not only from 2D xy spatial planes of traditional frames, but from the 2D time axis called spatio-temporal frame as well. Such spatial-temporal features are able to capture dynamic information from the action videos and are well-suited to recognize human actions without need of 3D extensions for the feature descriptors. The spatio-temporal BoF approach using SIFT and SURF feature descriptors obtains good recognition rates on a well-known actions recognition dataset. Compared with more sophisticated scheme of 3D based HoG/HoF descriptors, proposed method is easier to compute and simpler to understand.

☞ keyword : Action Recognition, BoF(Bag of Features), Spatio-temporal Action Detector, Feature Descriptor

1. 서 론

비디오에서 사람의 동작을 인식하는 연구는 동작 이미지 검색, 비정상적 행동 검출 모니터링, 사람과 컴퓨터 간 상호작용, 장애자의 움직임 패턴 분석, 사람의 동작을 인식해 반응하는 오락게임기 등 여러 분야에서 다양하게 응용되고 있다[1-4].

사람의 동작을 검출하고 인식하는 연구 방향은 대표적으로 모델기반 하향식 방식과 모델이 없는 상향식 방식으로 나뉜다. 모델기반 하향식 방식[5-7]은 움직이는 사람의 신체를 부분 영역화해 스틱 형태 등으로 미리 모델링한 후 인식 대상을 실루엣이나 경계상자 형태로 검출하고 인식 대상의 선택 동작을 저수준 특징으로 기술하여 모델과 비교한다.

[1] Department of Media Communication Design, Daegu Haany University, Gyeongsangbuk-do, 712-715, Korea
* Corresponding author(bit@dhu.ac.kr)
[Received 19 January 2015, Reviewed 28 January 2015(R2 26 March 2015), Accepted 18 May 2015]
☆ 이 논문은 2014년도 대구한의대학교 기린연구비 지원에 의한 것임

모델기반 방식의 단점은 중간 과업에 종속적이어서 분할(segmentation, 세그멘테이션)과 추적 기술을 별도로 적용해야 하므로 만약 우수한 분할방법과 추적 기술이 없으면 동작 인식의 정확성이 떨어져 특정 환경에서는 적용하기 어렵다. 예를 들어 Mokhber의 연구[8]에서 제시한 시공간 부피 기반의 모델기반 방식에서는 인식 대상의 실루엣 추출을 위해 인물의 신체 전체가 보이는 장면이 정지 상태여야 하고 배경을 정리하는 분할 방법이 필요하다. 이와 같이 모델기반의 인식 방식에 많은 제약이 있기 때문에 모델을 이용하지 않는 인식 방식이 제안되었다. 모델을 이용하지 않는 방식은 인식 대상이 있는 장면상태를 미리 설정하지 않고 저수준의 특징 분포를 통계분석적으로 처리한다. 따라서 모델이 없는 인식 방식은 인식 특징을 두드러지게 추출하여 표현하는 방법이 우수해야 한다.

모델이 없는 상향식 방식에서는 BoF(bag of features)를 이용하여 동작 특징을 추출하고 표현한 방법이 가장 우수한 인식 결과를 보이고 있다[9-10]. BoF는 텍스트 정보 검출에 이용해 온 BoW(bag of words)를 응용한 방식이다. BoW는 텍스트 문서를 표현하는 특징벡터를 단어 발생빈도 히스토그램으로 나타내는 것으로, 자주 나타나는 단어를 모아 단어장을 만들고 군집화하여 그 중 가장 판별력 있는 단어를 특징으로 선택하여 텍스트 검색의 핵심어로 이용한다[11]. BoF 방식은 BoW를 응용하여 비디오를 시공간 볼륨으로 간주하고 훈련 비디오 샘플에서 인식하고자 하는 동작의 두드러진 특징 영역을 특징점으로 검출한 다음 이를 BoW의 단어(word)를 대치한 부분 영역 특징 설명자로 표시한다. 그리고 부분 특징 설명자를 이용하여 전체 비디오 샘플에서 인식하고자 하는 특징 발생 빈도를 히스토그램으로 계산하여 인식을 수행한다.

이미지의 시각적 특징을 이용하는 BoF는 대상을 인식하는데 크기와 조명, 자세, 겹침에 강건한 장점이 있으므로[1] 특징 인식력이 좋은 BoF 방법을 확장하여 동작 인식에 적용한 연구가 다수 제시되었다[12-15]. 비디오에서 대상의 동작을 BoF로 표현하기 위해서는 움직임의 역동성을 나타내는 것이 가장 중요하다. 이에 대해 많은 연구에서 비디오를 시공간적 부피로 간주하여 3D 특징점 주변을 볼륨 패치로 설명하고 있다. 하지만 3D 특징점 검출과 3D 설명자 방식은 인식과정에서 계산 복잡도가 높아져 계산비용이 많이 들며 인식속도가 저하되는 단점이 있다[16-17].

본 연구는 비디오에서 인식 대상의 동작을 직접 2D 특징점으로 검출하고 이를 BoF로 표현하여 동작인식을 처리하는 방법을 제안한다. 제안 방식은 비디오 xy 공간 평면의 일반적 프레임뿐 아니라 시공간 프레임의 시간축을 따른 시공간 평면에서도 특징점을 검출하며, 검출한 특징은 설명자를 3D로 확장하지 않고 2D 설명자로 표현함으로써 비디오에서 복잡한 역동적 동작 정보를 단순하게 표현하여 수행비용은 낮추면서 높은 동작 인식율을 보인다.

그림 1은 연구에 필요한 BoF를 생성하는 단계를 설명한 것으로 동작 비디오의 xy, xt, yt 프레임 평면을 각각 선택하고 해당 평면에서 특징점을 검출한 다음 이를 동작 특징 설명자로 나타내 2D 상태에서 동작 역동성을 인식하는 과정이다.

본 논문은 2장에서 관련 연구를 설명하고 3장에서 제안 방법을 상세하게 제시하며 4장에서는 실험결과를 정리하고 5장에서 결론을 맺는 과정으로 구성하였다.

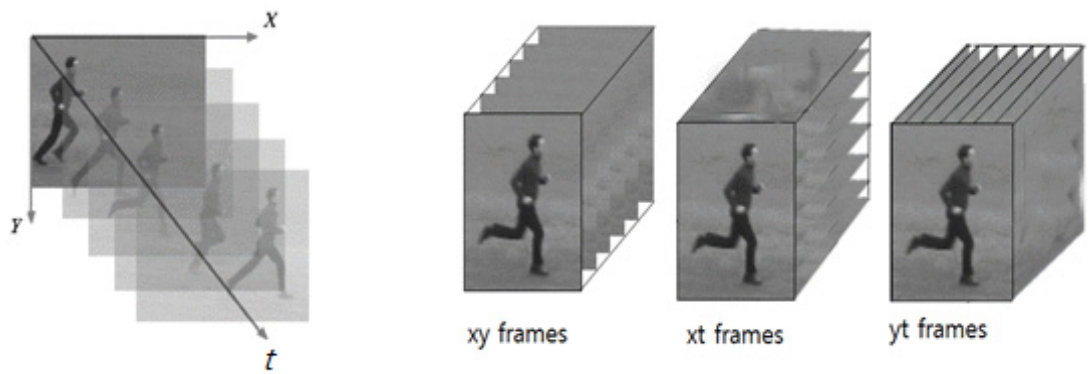


그림 1. 2D 기반 시공간 BoF 생성 단계
Figure 1. Generation Steps for 2D based Spatio-temporal BoF

2. 관련 연구

인식 대상의 특징점 검출에 BoF 방식을 적용한 기존 동작 인식 연구에서는 대부분 BoF로 특징점을 추출하고 특징의 표현에는 부분적 시공간 설명자를 이용하였다. Schuldt[12]의 연구에서는 시공간 특징점 알고리즘(Spatio-Temporal Interest Points, STIP)으로 특징점을 선택하고 선택한 특징점은 시공간 출력으로 설명하였다. Dollar[13]는 부분 선형 필터를 이용하여 표시한 특징점 주위에 큐보이드를 정의하고 정규화한 픽셀 값, 밝기 기울기, 윈도우 광류 등으로 큐보이드를 설명하여 특징을 표현하는 방법을 제안하였다. 시각적 인식 특징 중 밝기 기울기가 가장 나은 특징점 효과를 보였으며 이 특징들이 Nieble과 Liu의 연구[14-16]에서 사용되었다. Niebles의 연구[15]에서는 Dollar연구[13]의 밝기에 대한 기울기 특징을 단순한 BoF 표현으로 나타내 실험하였으며 특징 판별 분류과정에서 한 가지 변별 특징대신 여러 가지 특징을 통합하여 인식율을 높였다. 또 다른 BoF 방식[17]은 특징점 표현을 담당한 SIFT 설명자를 3D 공간으로 확장하였으며, 3D 시공간상의 특징점을 표현하기 위해 고유 SIFT 설명자에 시간정보를 추가하였다.

Ning의 연구[18]에서는 부분 설명자가 MAX 연산을 이용한 3D 가버필터 뱅크에 반응하는결과에 기반한 방법을 이용하였다. 특징점에 이웃하는 9개 방향에서 방향의 정량화를 통해 BoF 히스토그램을 생성하였으며 특징점의 공간 선택 대신에 슬라이딩 윈도우로 범위를 정한 패치에서 해당 특징을 계산하였다. Laptev[19]는 STIP 특징점으로 구축한 BoF 표현 방법을 제안하여 특징점 주위에 시공간 볼륨을 구축하고 방향 기울기(HoG)과 광류(HoF)를 히스토그램으로 계산하여 설명자를 설정하였다.

Lie의 연구[20]에서는 Dollar의 연구[13]와 유사한 밝기 기울기 특징을 이용하였다. 상호정보의 최대치(MMI, Maximization of Mutual Information)를 사용하여 $k-$평균(k-means) 알고리즘으로 큐보이드에 군집으로 출력한 것을 통합하는 새로운 방법을 제안하였다.

이상과 같은 기존 연구에서 제안한 BoF 기반의 특징 추출과 표현 방식들은 2D 특징 설명자를 3D 시공간으로 확장한 방식으로써 동작 정보를 직접 포착하였지만 대부분 3D 시공간 처리의 문제점인 계산 비용 및 복잡도가 높은 단점을 보인다.

동작 고유의 역동성은 여러 동작을 구분짓는 단서가 되므로 본 연구에서는 시간 축에 포함된 역동적 정보를 관측할 때 해당 특징 표현을 복잡한 3D로 확장하는 것은 피하면서 시공간 차원이 형성한 평면에서 BoF를 이용한 특징 검출과 특징 설명자를 이용한 동작 인식 방식을 제시한다. 이는 사람의 동작을 통해 감정과 의도 인식과정을 연구해 온 저자의 사전 연구[21][22]에서 동작 특징을 어떻게 포착하는 것이 가장 효과적인가에 대한 질문에서 출발한 것으로써 간단한 처리과정으로 시공간 동작 특징을 포착하고 2D 비디오 고유의 특성을 활용하면서도 동작 인식율은 높인다는 점에서 기존 연구와의 차별성을 보인다.

3. 동작 인식 프레임워크

제안 동작 인식 프레임워크에서는 그림 2와 같이 비디오 시퀀스상의 동작을 시공간 BoF로 표현하는 과정이 핵심이다. 비디오상의 동작 시공간 특징을 시각적 특징 단어로 정량화하여 단어 전체를 발생빈도 히스토그램으로 나타낸 후 인식 동작의 특징 발생빈도를 비교하여 인식하는 것이다. 이를 위해서는 동작별로 발생빈도가 높은 시각 특징을 선택해 나타내는 k개의 특징 단어장이 필요하다. 동작 특징 단어장 구축과 동작 인식 과정은 다음과 같다.

동작 특징 단어장을 구축하기 위해서는 a) 훈련 데이터에서 동작별로 빈번하게 나타나는 특징을 시공간 특징점으로 선택, 검출하여 b) 검출된 시공간 특징점 외 이웃 영역까지를 포함하여 특징공간으로 참조하도록 특징 설명자로 나타내 표현한다. c) 특징 분포만으로 비디오의 동작을 이해할 수 있도록 특징 설명자를 $K-$평균으로 군집시켜 특징 단어장을 생성하고 최적의 단어장 크기를 결정한다[22]. $K-$평균으로 군집시킨 동작 특징 단어 수 k는 k차원의 동작 특징 벡터가 된다. d) 동작별 특징 설명자를 유클리디안 거리를 이용해 가장 가까운 $k-$평균 군집화된 특징 단어장의 단어와 대응시킨다. e) 비디오를 BoF 표현으로 나타내는 히스토그램을 만들기 위해 단어장 인덱스 전체에서 각 단어의 발생빈도를 계산한다. 계산된 빈도 수를 이용하여 BoF 히스토그램을 정규화한다. f) 비디오별 전체 BoF 표현이 만들어지면 동작 클래스에 분류기를 적용하여 동작을 분류, 인식한다. 동작 분류는 SVM을 이용하여 식 (1)과 같이 수행한다.

$$K(H_i, H_j) = \exp\left(-\frac{1}{2A}\sum_{n=1}^{k}\frac{(h_{in}-h_{jn})^2}{h_{in}+h_{jn}}\right) \qquad (1)$$

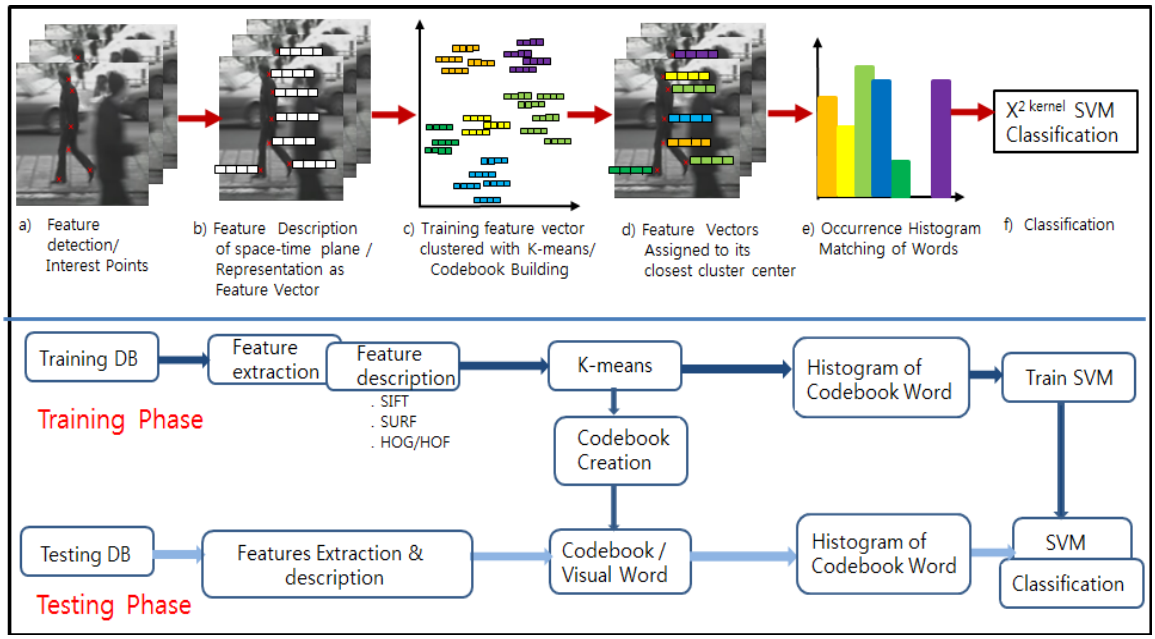


그림 2. 제안 동작인식 프레임워크

Figure 2. Proposed action recognition framework

$H_i = \{h_{in}\}$, $H_j = \{h_{jn}\}$은 특징 단어 발생의 빈도 히스토그램이고 k는 특징 단어장 크기이다. A는 모든 훈련 샘플의 평균 거리 값이다. 동작의 다중 클래스 분류는 다중교차 검증 방법을 이용하여 처리한다. 제안 동작인식 프레임워크에서 학습과 테스트 과정의 각 단계는 그림 2의 하단에 표현하였다.

3.1 이미지 전처리

동작 인식에 필요한 대상 정보를 획득하고 처리 속도를 빠르게 하기 위해서는 이미지 전처리 단계에서 시공간 정보 분할 수행이 필요하다. 이를 위해 정적 배경 유무에 관계없이 비디오의 인접 프레임 간 변화정도를 차분 값으로 계산하고 원 영상 프레임에서 시간상으로 떨어진 프레임까지의 시공간 기울기 차분 정보가 담긴 특징점을 표시한다. 전처리를 통해 차분 정도가 표시된 비디오는 움직임 정보를 포착할 뿐 아니라 동작 정지 형태와 역동적 움직임 모두를 포함하게 된다. 차분 동작을 이용한 분할 방법은 움직이는 카메라로 찍은 배경이 급변하는 비디오에는 적용하기 힘들다는 단점이 있으나 동작인식 데이터 대부분은 배경 변화가 크지 않은 동영상이므로 프레임간의 시간차를 이용한 변화 정도를 분할하는 방법을 적용하였다. 이미지 전처리 결과는 그림 3의 전처리 과정에서 확인할 수 있다.

3.2 BoF 구축

BoF는 특징의 발생 분포도를 반영하는 히스토그램을 이용하여 출력 값을 표시하는 가장 단순한 방법이다. 이 방법의 핵심은 미리 정의한 특징 단어 시퀀스 구축, 즉 단어장$\{w_j\}(1 \le j \le k)$의 구축으로 단어장 크기 k는 동작 특징 단어들을 나타낸다. 각 특징 단어에 대해, 미리 정의한 단어와 특징 간의 거리를 계산하여 최소 거리 ($j_d = \arg\min_{1 \le j \le k} D(d, w_j)$, $D(*)$는특징거리함수)를 구한 다음 특징 단어를 최소거리의 단어에 할당한다. 모든 특징 단어가 단어장에 포함되면 히스토그램 $H = \{h_j\}(1 \le j \le k)$으로 표현하여 단어장이 얼마나 많은 특징 단어를 포함하는지 빈(bin)으로 나타낸다. 즉, 히스토그램의 빈(bin)이 BoF의 출력이 된다. 그리고 이 빈(bin)을 다른 특징들의 히스토그램과 연결하거나 합쳐서 동작인식을 수행하는 분류기의 입력 값으로 이용한다. 동작 인식에 BoF 기법을 적용하면 구축해 놓은 특징

점을 동작 특징 설명자와 결합하여 여러 종류의 동작 비디오에 이용할 수 있다.

3.3 특징 설명자

본 연구는 프레임을 분리하여 BoF 방식으로 동작 특징점을 검출하고 이를 SURF(Speeded Up Robust Features)[24], SIFT(Scale-invariant Feature Transform) 설명자[25]로 나타내는 방식으로 동작인식을 수행하였다. 동작인식 수행 결과는 HoG/HoF 설명자[19]와 비교하여 테스트하였다.

설명자를 이용한 객체인식 방법은 영상에서 특징점을 검출하고 특징점 주변 모양을 특징정보로 나타내는 특징 설명자를 추출하여 동작 데이터집합의 특징설명자와 매칭하여 객체를 식별하는 과정으로 이루어진다. 특징 설명자는 입력영상과 대상과의 특징점 정합에 기준이 되는 성분으로 특히 SIFT 특징 설명자는 위치, 크기, 장소에 불변하며 빛의 변화뿐만 아니라 어파인 변환에 강건한 특징을 보여 컴퓨터비전 연구에서 탁월한 성능을 보였으며 이미지나 비디오에서 BoF 표현 특징점을 자연스럽게 이용하는 알고리즘으로 인정받고 있다. 하지만 속도가 느리기 때문에 이에 대한 대안으로 SURF 특징 설명자 알고리즘이 제시되었다. 이 방식은 높은 인식 성능을 보이면서도 단순하다는 장점이 있으며 낮은 계산 비용으로 SIFT와 비슷한 인식결과를 보인다. 이외, HoG/HoF 특징 설명자는 3D공간 기울기와 광류를 이용해 부분 동작과 외형에 대한 검출과 기술 기능을 개선한 방식이다. 본 연구에서는 시공간 BoF 특징점으로 동작 특징을 검출하고 이를 2D SIFT, SURF 설명자로 표현한 방법과 BoF 기반 헤시안 검출기를 이용한 HoG/HoF 설명자 간에 동작 인식 성능을 비교한다.

3.3.1 SIFT

SIFT 설명자는 변형, 회전, 크기에 불변하며 부분 겹침에 강건하게 특징의 위치와 크기를 결정한다. 2D SIFT 설명자는 정지 형상을 강조하며 기울기 방향에 가중치를 준 2D 공간 히스토그램으로 시간 도메인은 고려하지 않는다.

이미지 전체 스케일을 모두 반영하는 SIFT 특징은 입력 영상에 공간스케일인 σ 값을 증가시켜 가며 가우시안필터를 적용한 영상을 만들고 σ가 두 배가 될 때마다 영상을 1/2로 다운 샘플링하여 영상 만들기를 반복한다. 입력영상을 가우시안 스케일 공간의 피라미드 구조를 만든 다음에는 가우시안 차분(Difference-of-Gaussain, DoG) 함수를 계산한다. DoG 영상에서 x, y 및 t(시간) 축으로 인접한 26 지점보다 DoG의 절대값이 큰 극값 지점을 찾아 특징점 후보로 선택한다. 낮은 극값의 특징점 후보와 엣지 응답의 특징점 후보는 제거한다. 즉 샘플링된 극점이 아닌 하위픽셀 극점을 찾는다.

본 연구에서는 선택한 특징점을 통해 이웃 픽셀의 방향을 계산하여 방향 히스토그램을 만든다. 히스토그램의 기울기 크기는 히스토그램 빈(bin)을 계산하는 가중치로 이용한다. 방향 히스토그램을 통해 가장 큰 빈 값에서 주요 방향을 검출하며 특징점의 주변을 회전시켜 회전 불변치를 구한다. 이때 이웃영역의 크기를 16×16 픽셀로 설정한 다음 이 픽셀 전체 영역을 4×4의 하위영역을 나누어 하위영역 수가 16 이 되게 한다. 16 개 하위영역의 방향 히스토그램에는 8 빈을 할당하며 16 개 하위영역의 히스토그램이 128 특징차원벡터를 형성하여 이 벡터를 2D SIFT 특징 설명자로 이용한다.

3.3.2 SURF

SURF는 SIFT보다 고속으로 특징점을 검출하는, 하(Haar) 웨이블릿 응답을 이용한 특징설명자로 라플라시안 부호를 이용하여 매칭속도를 향상시켰다. SURF는 이미지 패치를 셀로 나누고 각 셀을 균일하게 샘플링한 하웨이블릿 결과 값을 더한 가중치 벡터로 나타낸 특징 설명자이다.

SURF 특징 설명자는 원본 영상의 원점에서 각 위치까지의 픽셀 값의 합을 저장한 통합영상을 이용하여 특정 사각형 태 픽셀 값의 합을 아주 빠르게 적분 계산한다. 영상의 크기를 줄이는 대신 필터의 크기를 키워 처리속도를 빠르게 하기 위해 적분 영상을 만든 후 웨이블릿 마스크를 이용하여 x, y, xy의 웨이블릿 정보를 옥타브와 레이어 개수만큼 만든다. 여기에 헤시안 연산을 하여 임계치를 넘으면 레이어를 포함한 이웃 26개 헤시안 정보와 비교한 후 가장 값이 큰 픽셀을 보간하여 특징점으로 추가한다. 이 특징점 주위 영역을 4×4 블록으로 나누고 각 블록을 5×5로 샘플링한 후 샘플에 대해 하 웨이블릿 응답과 $d_x, d_y, |d_x|, |d_y|$합을 구하여 $4\times4\times4$차원의 벡터로 특징 설명자를 나타낸다.

본 연구에서는 웨이블릿 응답으로 가장 수치가 큰 각도를 얻어 방향정보를 기반으로 이미지의 기울기 값을 계산하였다. 이 기울기를 64 - 128 개의 빈 히스토그램으로 표현해 크기와 회전에 강한 설명자 정보를 생성한다.

3.3.3 HoG/HoF

HoG/HoF 설명자는 Laptev[19]가 제안한 것으로 부분 동작패턴과 역동적 외형을 포착하기 위해 특징점 주변영역에서 공간 기울기의 히스토그램 HoG(Histogram of Gradient)와 3D으로 축적한 광류의 히스토그램 HoF (Histogram of Flow)를 조합하여 이용한 방식이다. 해리스 3D, 가버, 헤시안 등 특징점 검출기와 조합한 HoG/HoF는 샘플링 위치 주변을 부분 영역으로 설명할 때 샘플링 특징점을 x,y,t 로 포착하고 공간스케일은 σ 로 표시, 시간스케일은 τ 로 표시하여 $\triangle_x(\sigma)=\triangle_y(\sigma)=18\sigma$, $\triangle_t(\tau)=8\tau$ 로 설정한 다음 각 이웃하는 이미지 패치에 대한 볼륨 셀을 $n_x\times n_y\times n_t$ 셀 그리드로 나눈다[19]. 그리고 셀을 4개의 빈 기울기 방향 히스토그램과 5개의 빈 광류 히스토그램으로 표현하여 정규화한 다음 연결하여 형태와 동작 특징을 나타낸다. SIFT 설명자처럼 정규화된 히스토그램을 HoG/HoF 설명자 벡터로 연결하는 이 방식은 SIFT 설명자가 고려하지 않은 시간 도메인을 포함하고 있기 때문에 모양과 동작 단서를 모두 포착하는 장점이 있다. 또한 BoF 방식의 헤시안(Hessian) 검출기를 사용하여 특징 단어장을 구성하므로 본 연구의 BoF 특징설명자 방식과 비교 의의가 있다.

3.4 시공간 프레임에서 역동적 동작 정보 수집

비디오를 BoF로 나타내는 가장 단순한 방법은 대상 비디오세그먼트에서 모든 프레임별로 특징점을 다수 선택하고 이 특징점 개수를 세서 고유 특징점 발생빈도 히스토그램을 만드는 것이다. 하지만 이런 방법은 시간 축을 따라 발생하는 동작 정보 등을 무시한다는 단점이 있다. 그래서 동작 정보를 포함시키기 위해 대부분의 연구자들은 2D 설명자를 3D 시공간으로 확장시키는 방법을 사용한다.

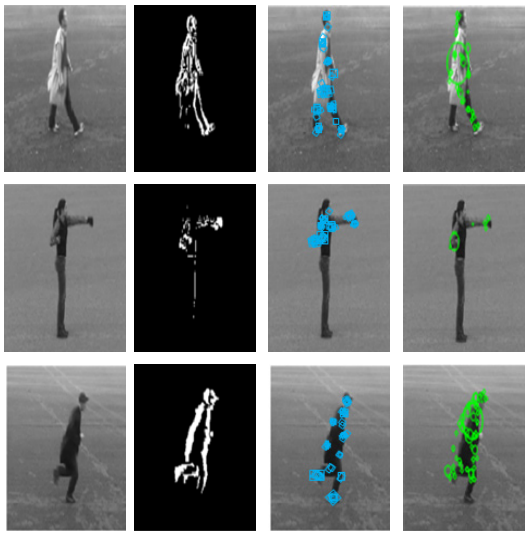
제안 방식에서는 2D 특징점 검출기와 특징설명자 알고리즘을 이용하여 비디오의 BoF 표현으로 일반적인 공간 프레임뿐만 아니라 시공간프레임이라 부르는 평면에도 적용하도록 한다. xy방향의 단순한 공간 평면은 비디오의 일반적인 프레임이다. 이 프레임을 축적하면 x,y 방향으로 확장 가능한 시공간 볼륨을 형성하여 시공간 프레임을 구성한다. 즉 시공간 프레임은 xt,yt처럼 시간 축과 공간 축 중 하나로 형성된 평면이 된다. 이 방법을 가정하는 것은 시공간 프레임에서 추출한 2D 설명자가 비디오에 포함된 역동적 정보를 포착할 수 있으며 역동적 정보는 반복해서 점진적인 변화과정을 감안할 수 있는 내용을 포함하기 때문이다. 역동적 정보는 동작을 인식하게 하는 기본 특징이 된다.

이 방법의 장점은 설명자를 일부러 3D로 확장을 할 필요 없이 기존 2D 기술을 이용하여 특징을 선택하여 처리할 수 있다는 점이다. 본 연구에서 적용한 BoF 구축 방법은 단순하다는 점도 장점이다. 데이터 차원감소는 PCA를 이용하여 수행할 수 있으며 특징벡터 정량화는 $k-$평균 알고리즘을 통해 처리하였다. 기본적인 BoF 구조는 변형하지 않았으며 동작 분류는 특징 단어 갯수를 조절할 수 있는 패널티 오류 파라미터를 설정한 선형 SVM을 통해 수행했다.

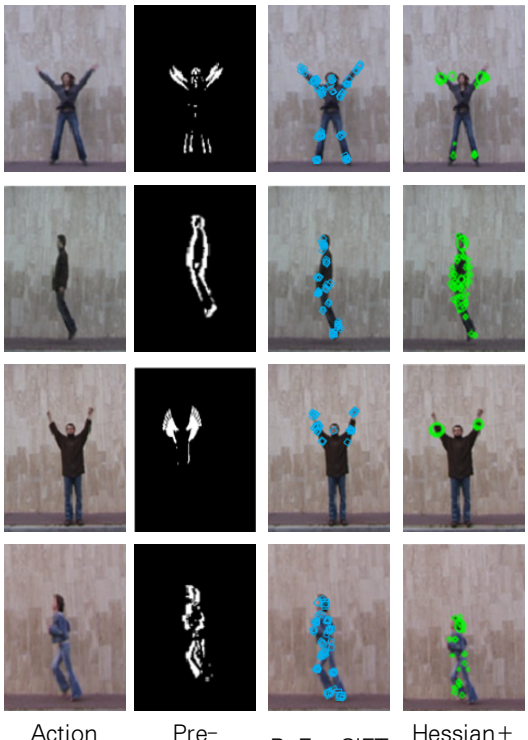
4. 실험 및 결과

4.1. 실험 데이터베이스

시공간 프레임에서 역동적 정보를 포착하는 제안 방식 기능을 평가하기 위해 그림 3과 같이 동작을 담은 Weizmann[25]와 KTH[11] 데이터집합을 이용하였다. 그림 3은 데이터집합의 이미지를 분할한 전처리과정을 거쳐 BoF와 SIFT, 헤시안 검출기와 HoG/HoF를 적용한 결과이다. BoF+SIFT는 BoF로 시공간 동작특징을 검출하고 SIFT로 특징을 설명하였으며 헤시안+HoG/HoF는 BoF 기반 헤시안 검출기로 동작특징을 검출하고 HoF/HoF로 특징을 설명한 것이다. 두 방식은 유사하게 특징점을 포착하였으며 동작이 단순한 경우에는 SIFT가, 복잡한 경우 HoG/HoF가 더 다양한 특징점으로 동작을 설명한다. Weizmann 데이터베이스는 짧은 비디오 세그먼트로 구성되어 있으며 9명의 인물이 10가지 동작을 취하는 내용이 담겨 있다. 굽히기, 점핑잭, 제자리 뛰기, 옆으로 뛰기, 달리기, 앙감질, 뛰면서 앞으로 가기, 걷기, 한손흔들기, 양손흔들기 동작이 포함되어 있다. Weizmann DB는 180×144 픽셀의 저해상도 상태로, 훈련을 위해 6명을 임의로 선택하고 남은 3명을 테스트하는데 이용했다. KTH는 6명의 인물이 복싱, 손뼉치기, 손흔들기, 조깅, 달리기, 걷기, 동작을 하는 내용으로 구성되어 있다. 160×120 픽셀로 공간 해상도를 낮게 샘플링하였고 16명의 동작은 훈련에, 나머지 9명의 동작은 테스트하는데 이용하였다.



(a) Weizmann Dataset



(b) KTH Dataset

그림 3. 동작 데이터집합 예

Figure 3. Example of Action Datasets

4.2. 실험 과정

실험 시작 단계에서는 모든 프레임의 방향을 따라 BoF 특징점 검출 알고리즘을 적용한 다음 프레임집합을 다양하게 조합하여 비디오에 대한 BoF 검출과 특징 설명자를 구축한다. 이를 위해 xy 프레임에서 추출한 설명자가 동작의 역동성을 포착하도록 한 다음, 동작 인식결과 평가를 위해 프레임들을 모두 조합하여 프레임 집합별로 특징점 기술 알고리즘을 실험한다. 실험 과정은 그림 2에서 제시한 프레임워크에 따라 특징 단어장 크기 k를 설정하고 수행한다. 평면을 조합한 경우에는 모든 평면 집합에서 구한 BoF를 최종 BoF를 형성하기 위해 연결시킨다. 최종 BoF 표현의 차원 크기는 60-1,000이므로 모든 평면집합에서 구한 연결 히스토그램의 크기를 10-1000 간격으로 설정한다.

마진 폭과 분류 오류 사이의 타협점을 찾아주는 SVM 오류 패널티 변수 C 를 설정하고 특징 단어장 크기별로 적절한 크기를 찾는다. C 값은 10^{-10}과 10^{10}사이의 로그 크기로 설정한다. k 와 C 에 대한 최적의 인식율을, 관계없는 특징을 하나씩 제거해 나가는 리브원아웃(leave-one-out)방식의 5중 상호교차검증 과정을 통해 측정한다. 최적의 k 값과 초기 C 값을 찾은 다음에는 이전의 최적 값에서 C 값을 더 정교하게 찾는 과정을 수행한다. k 값과 C 값을 구한 다음 10회의 새로운 5중 교차검증을 다시 수행한다. 이 중 10개의 평균 인식율을 찾아 신뢰도 간격을 개선하는 평균값으로 이용한다.

4.3. 실험 결과

표 1은 KTH와 Weizmann DB를 대상으로 프레임들을 조합하여 BoF + SURF, BoF + SIFT와 헤시안 검출기+ HoG/HoF 설명자를 적용함으로써 획득한 최고 인식율이다. 신뢰도 95% 수준이다. SURF와 SIFT를 이용한 프레임 조합 대비 인식율을 나타냈다. 헤시안 검출기를 이용하는 HoG/HoF는 x,y,t 축을 이용한 볼륨 셀을 분리하여 처리하는 3D 기반 방식이므로 2D 개별 프레임 인식결과 대신 전체 인식율을 표시하였다.

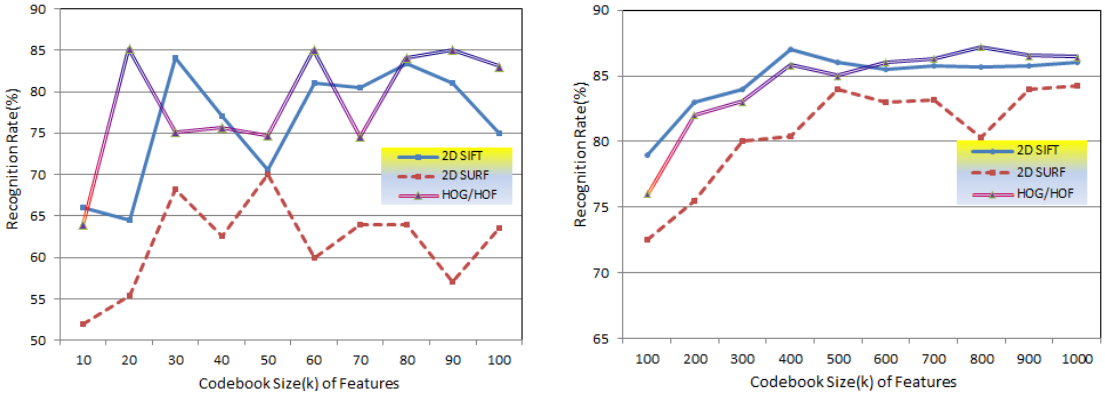
비디오를 BoF로 구축하는 데 시공간 한 개 프레임 정보를 이용하는 것이 xy 프레임에서 검출한 특징점에서 생성한 BoF 전체의 인식율을 평균 11% 개선함을 알 수 있다. 그리고 xt 또는 yt와 같이 시공간 프레임 중 하나를 xy프레임에서 가져온 특징점과 조합한 결과가 기준치보다 평균 16% 더 나은 성능을 보임을 알 수 있다. 그

표 1. 설명자별 인식률
Table 1. The recognition rates per different descriptors

DB	plane	BoF+SURF(%)	BoF+SIFT(%)	Hessian+ HoG/HoF(%)
KTH	xy	65±3	67±4	86±5
	xt	73±3	82±6	
	yt	74±4	89±3	
	$xy+xt$	80±3	84±3	
	$xy+yt$	80±2	90±3	
	$xt+yt$	79±2	85±4	
	$xy+xt+yt$	78±3	89±4	
Weizmann	xy	70±4	73±3	91±2
	xt	73±3	87±3	
	yt	77±2	90±3	
	$xy+xt$	86±2	89±1	
	$xy+yt$	87±2	91±3	
	$xt+yt$	85±2	89±2	
	$xy+xt+yt$	87±3	89±2	

렇지만 모든 프레임을 조합하는 것이 인식율 개선에 도움이 되는 것은 아니다. 표 1에 나타난 바와 같이 $xy+xt, xy+yt, st+yt, xy+xt+yt$ 조합에 따른 인식율은 통계적으로 큰 차이가 없다. 이 결과는 순수한 공간과 시간 프레임이 서로 보완역할을 하는 한편 다른 방향의 시공간프레임은 잉여 정보를 전달한다는 것을 의미한다. 프레임 중에서는 $xy+yt$조합이 최고의 인식결과를 보였다. 프레임 조합을 달리한 BoF + SURF, BoF + SIFT 설명자 인식율 결과는 거의 유사하였다. 결과적으로 KTH, Weizmann 데이터집합의 시공간 프레임에 여러 가지 설명자를 적용한 결과를 통해 비디오에서 역동적 동작 정보를 포착하는데 2D 특징점 설명자를 사용할 수 있음을 의미한다.

그림 4는 KTH과 Weizmann 데이터집합에 적용한 BoF의 단어장 크기 k에 따른 SURF, SIFT 및 헤시안검출기를 이용한 HoG/HoF 인식율 비교 결과이다. KTH 데이터집합에 대해 SIFT는 65, SURF는 30, HoG/HoF는 60이 최적의 특징단어 수이다. Weizmann 데이터집합에 대해서 SIFT는 400, SURF는 500, HoG/HoF는 400개가 최적의 특징단어 수이다. Weizmann 데이터집합이 KTH 데이터집



(a) KTH (b) Weizmann

그림 4. 특징단어 수에 따른 동작 인식율
Figure 4. Action Recognition rates according to feature words number

합보다 더 많은 특징단어 수를 필요로 한다. BoF에서 특징 수는 핵심 파라미터이므로 특징 수를 여러 가지로 변형하여 인식시킨 결과 특징 수가 늘어나면 SURF는 HoG/HoF, SIFT 방식보다 더 나은 인식율을 보였다. 이는 SURF가 HoG/HoF, SIFT보다 더 많은 특징점을 선택하고 조밀한 샘플링을 하기 때문이다.

표 2는 BoF와 SIFT를 조합해 동작 데이터집합을 인식한 결과로 표의 행은 예측 클래스의 정확도이고 열은 실제 클래스의 정확도를 나타낸다. KTH 데이터집합에서 가장 높은 인식율을 보인 동작은 걷기이며 조깅 동작이 가장 낮은 인식율을 보였다. Weizmann에서는 달리기와 옆으로 뛰기 동작이 가장 낮은 인식율을 보였다.

표 2. BoF와 SIFT를 이용한 동작인식율
Table 2. Confusion Matrix of Action Recognition with BoF and SIFT

(a) KTH

	boxing	clapping	wave	jogging	run	walk
boxing	91.7	3.5	0	0	4.5	1.3
clapping	3.8	92.2	1	0	3.0	0
wave	0.5	4.5	90.0	0	3.5	1.5
jogging	0	0	0	87.5	10.8	1.8
run	0	0	0	9.3	90.5	0.2
walk	3.2	0	0	4.5	0	93.3

(b) Weizmann

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	92.0	0	0	0	0	0	1.0	0	0	0
jack	0	92.2	1.0	0	0	0	0	0	2.8	0
jump	0	1.0	93.0	2.0	0	0	0	0	0	0
pjump	0	0	1.0	91.0	0	0	0	1.0	0	0
run	0	0	0	0	91.0	0	4.4	1.5	0	0
side	0	0	0	4.0	0	92.0	0	0	0	0
skip	1.0	0	0	0.6	0	2.4	92.0	0	0	0
walk	0	0	0	0	3.7	0	0	93.3	0	0
wave1	0	3.4	0	0	0	0	0	0	91.6	0
wave2	0	0	0	0	0	0	0	0	2.7	92.3

표 3은 제안 방식을 적용한 설명자들을 Weizmann 데이터집합에 적용하여 비교한 것으로 SIFT는 초당 4.6개의 프레임을 계산하여 가장 빠른 처리 속도를 보였다. SURF는 초당 1.6 프레임, HoG/HoF는 0.9 프레임을 처리하였다. 그림 4의 실험 결과에 따라 특징 단어장 크기는 최적의 인식률을 보인 상태의 크기와 같다.

표 3. 설명자별 처리 속도
Table 3. Processing speed per descriptors

	SURF	SIFT	HoG/HoF
Frames/second	1.6	4.6	0.9
word size	500	400	400

표 4는 Weizmann과 KTH 데이터집합을 대상으로 한 기존 연구에서 제안한 특징 설명자를 제안 연구와 비교한 내용으로써 동일한 동작데이터집합인 KTH와 Weizmann 데이터집합에 실험한 결과를 제시한 것이다. 전반적인 알고리즘 성능테스트비교가 어려워 기존연구 결과를 통해 제시된 인식 결과를 단순 비교하였다.

표 4. BoF 기반 설명자 인식율 비교
Table 4. Comparing of Recognition rates of BoF based approaches

(a) KTH

	Approaches	Recognition Rates(%)
Proposed	BoF+SURF	80±3
	BoF+SIFT	90±3
	Hessian + HoG/HoF	91±2
Previous	SURF[21]	80±4
	SIFT[22]	91±3
	HoG/HoF[23]	87±3
	Niebles et al.[15]	87.5
	Liu et al.[14]	88.6
	Scovanner et al.[16]	76.1

(b) Weizmann DB

	Approaches	Recognition Rates(%)
Proposed	BoF+SUFT	87±2
	BoF+SIFT	80±2
	Hessian + HoG/HoF	86±5
Previous	SURF[21]	81±3
	SIFT[22]	91±3
	HoG/HoF[23]	89±3
	Niebles et al[15].	90.0
	Liu et al.[14]	89.3
	Scovanner et al.[16]	82.6

Niebles의 성좌모델[14]은 Dollar 연구[13]의 인식 특징인 밝기, 시공간 큐보이드 기울기를 지형정보로 추가한 것이다. SIFT를 3D로 확장한 연구는 Niebles[15]이 제안했다. 이 방법을 제안 연구와 비교한 결과 기존 연구에서는 BoF 방식에 대한 인식율 개선 과정은 없었으며 다만 2D 설명자를 확장하여 시간 차원을 처리했다. 이 경우, 특징점 검출기와 설명자로 SURF를 적용한 제안 방식들은 전반적으로 낮은 인식율을 보였으며, SIFT를 선택했을 때 가장 높은 인식 결과를 보였다.

Niebles[15]와 Liu[16]의 결과는 본 연구에서 제시한 SURF 적용 결과에 비해 더 높은 인식율을 보였다. 이는 Liu의 방법이 특징점 검출 시 특징점 크기를 다양화하여 촘촘하게 선택하여 높은 인식 결과를 보였기 때문이다. 하지만 본 실험에서는 여러 가지 크기의 특징점을 고려하지 않았기 때문에 Liu와 Niebles의 SURF 적용방식이 더 낫다고 단정할 수는 없다. 또한 Niebles의 방법은 신체 실루엣으로 구축한 시공간 볼륨에서 구한 특징을 Dollar의 특징에 혼합했기 때문에 실제 순수한 BoF 방식의 모델이 아니기 때문에 단순 인식율 성능 비교로 제안 방식보다 우수하다고 보기 어렵다.

한편, 전통적인 xy 프레임과 yt 시공간 프레임을 같이 연결하고 여기서 BoF로 특징점을 검출하여 SIFT 설명자를 이용한 방식이 가장 높은 인식율을 보였다. 그리고 전반적으로 KTH 동작 데이터집합보다는 Weizmann 데이터집합을 이용한 인식결과가 더 높은 인식율을 보였다.

5. 결 론

본 연구는 역동적 동작 정보를 담고 있는 비디오에서 동작을 2D 기반의 시공간 BoF로 표현하여 단순하면서도 정확하게 사람의 동작을 인식하는 방법을 제안하고 있다. 시공간 차원이 형성한 평면에서 2D BoF를 이용한 특징점 검출과 특징 설명자를 결합하여 표현한 동작 인식 방식을 제시하여 대표적 동작 데이터베이스인 Weizmann와 KTH 동작 데이터에 적용한 결과 높은 동작 인식율을 보였다.

BoF를 이용한 특징점 검출과 2D 기반의 SURF 및 SIFT 알고리즘, 3D 기반의 HoG/HoF 설명자를 비교하여 실험한 결과 2D xy 프레임과 yt 시공간 프레임에 적용한 BoF 방식과 설명자 결합만으로 비디오에서 역동적 동작 정보를 수집할 수 있으며 역동적 측면이 핵심 역할을 하는 동작 인식에서 인식율을 개선할 수 있음을 알 수 있다. 전통적 프레임과 시공간 프레임 양쪽의 정보를 수집하는데 SIFT 설명자를 적용하여 구축한 일반적인 BoF 표현이 복잡한 프레임과 비교했을 때도 가장 높은 인식율을 보였다. 또한 3D기반의 HoG/HoF 설명자와 비교했을 때 제안 방식이 3D 기반 방식과 유사한 인식율을 보였으며 처리 시간은 오히려 빠름을 확인하였다.

향후에는 BoF 표현에 대한 특징 단어장 구축을 더 단순하지만 효과적으로 구축하는 방법을 심도있게 진행할 것이고 여러 동작이 섞인 복잡한 실제 동작 데이터베이스에도 제안 방법을 적용하여 동작 인식 평가 과정을 다각화할 계획이다.

참 고 문 헌 (References)

[1] R. Poppe, “A survey on vision-based human action recognition”, Image and Vision Computing, vol. 28. pp. 976-990, 2010.
http://doi:10.1016/j.imavis. 2009.11.014

[2] Q. V. Le, W. Y. Zou, S. Y. Yeung, A. Y. Ng, “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis,“ IEEE Conference Computer Vision and Pattern Recognition, 2011. pp. 3361-3368.
http://10.1109/CVPR.2011.5995496

[3] Shizhi Chen, YingLi Tian, Qingshan Liu, Dimitris N. Metaxas, “Recognizing expressions from face and body gesture by temporal normalized motion and appearance features,” Computer Vision and Pattern Recognition Workshops, 2011, pp.7-12.
http://10.1109/CVPRW.2011.5981880

[4] George Caridakis, Stylianos Asteriadis, Kostas Karpouzis, “Non-manual cues in automatic sign language recognition,“ Personal and ubiquitous computing, vol. 18, no. 1. pp. 37-46, 2014.
http://10.1007/s00779-012-0615-1

[5] J. K. Aggarwal, S. Park, “Human motion: Modeling and recognition of actions and interactions,” 3DPVT, IEEE Computer Society, 2004, pp. 640 - 647.
http://10.1109/TDPVT.2004.1335299

[6] T. B. Moeslund, A. Hilton, V. Kruger, “A survey of advances in vision-based human motion capture and analysis,” Computer Vision and Image Understanding,

vol. 104, no. 2, pp. 90 – 126, 2006.
http://doi:10.1016/j.cviu.2006.08.002

[7] A. K. Roy-Chowdhury, R. Chellappa, A. Bovik, S. K. Zhou, “Recognition of humans and their activities using video (Synthesis Lectures on Image, Video and Multimedia Processing)”, Morgan & Claypool Publishers, pp. 173, 2006.
http://10.2200/S00002ED1V01Y200508IVM001

[8] A. Mokhber, C. Achard, M. Milgram, “Recognition of human behavior by space-time silhouette characterization,” Pattern Recognition Letter, vol. 29, no. 1, pp. 81 – 89, 2008.
http://doi:10.1016/j.patrec.2007.08.016

[9] A. Oikonomopoulos, I. Patras, M. Pantic, “Spatiotemporal Localization and Categorization of Human Actions in Unsegmented Image Sequences”, IEEE Transactions on Image Processing. vol. 20, no. 4. pp. 1126-1140, 2011.
http://10.1109/TIP.2010.2076821

[10] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, C. Schmid, ″Evaluation of local spatio-temporal features for action recognition″, BMVC, 2009, pp. 499-502.
https://10.5244/C.23.124

[11] R. A. Baeza-Yates, B. A. Ribeiro-Neto, “Modern Information Retrieval.” ACM Press, Addison-Wesley, 1999. ISBN:020139829X

[12] C. Schuldt, I. Laptev, B. Caputo, “Recognizing human actions: a local SVM approach,” ICPR, 2004, vol. 3, pp. 32 – 36.
http://10.1109/ICPR.2004.1334462

[13] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, “Behavior recognition via sparse spatio-temporal features,” ICCCN, 2005, pp. 65 – 72.
htpp://10.1109/VSPETS.2005.1570899

[14] J. Niebles, F. Li, “A hierarchical model of shape and appearance for human action classification,” CVPR, 2007, pp. 1 – 8.
http://10.1109/CVPR.2007.383132

[15] J. Niebles, H. Wang, L. Fei-Fei, “Unsupervised learning of human action categories using spatial-temporal words,” IJCV, vol. 79, no. 3, pp.
299 – 318, 2008. http://10.1007/s11263-007-0122-4

[16] J. Liu, S. Ali, M. Shah, “Recognizing human actions using multiple features,” CVPR, 2008, pp. 9 – 18,
http://10.1109/CVPR.2008.4587527

[17] P. Scovanner, S. Ali, M. Shah, “A 3-dimensional sift descriptor and its application to action recognition,” MULTIMEDIA ‘07, 2007, pp. 357 – 360.
http://10.1145/1291233.1291311

[18] H. Ning, Y. Hu, T. Huang, “Searching human behaviors using spatial-temporal words,” IEEE International Conference on Image Processing, 2007, pp. 337 – 340.
http://10.1109/ICIP.2007.4379590

[19] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, “Learning realistic human actions from movies,” CVPR, pp. 1 – 8. 2008.
http://10.1109/CVPR.2008.4587756

[20] J. Liu, M. Shah, “Learning human actions via information maximization,” CVPR, 2008, pp. 21-30,
http://10.1109/CVPR.2008.4587723

[21] Jinok Kim, “A Study on Visual Perception based Emotion Recognition using Body-Activity Posture,” The KIPS Transactions, Part B, vol. 18, no. 5, pp. 305-314, 2011.
http://10.3745/KIPSTB.2011.18B.5.305

[22] JinOk Kim, “Agent’s Activities based Intention Recognition Computing”, Journal of Korean Internet Society, vol. 13, no. 2, pp. 87-98, 2012.
http://10.7472/jksii.2012.13.2.87

[23] T. M. Mitchell, “Machine Learning.” New York: McGraw-Hill, 1997.
ISBN:0070428077 9780070428072

[24] H. Bay, A. Ess, T. Tuytelaars, L. V. Gool, ″SURF: Speeded Up Robust Features,″ European Conference on Computer Vision, 2006, pp. 346-359.
http://10.1007/11744023_32

[25] D. Lowe, “Distinctive image features from scale-invariant keypoints,″ IJCV, vol. 60, no. 2, pp. 91-110, 2004.
http://10.1023/B:VISI.0000029664.99615.94

[26] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, “Actions as space-time shapes,” PAMI, vol. 29, no. 12, pp. 2247 – 2253, 2007.
http://10.1109/TPAMI.2007.70711

◐ 저 자 소 개 ◑

김 진 옥 (Kim Jin Ok)
1989년 성균관대학교 졸업(학사)
1998년 성균관대학교 대학원 정보통신공학과 졸업(석사)
2002년 성균관대학교 대학원 전기전자 및 컴퓨터공학과 졸업(박사)
2004년~현재 대구한의대학교 글로벌융합대학 시각미디어디자인학과 교수
관심분야 : 멀티미디어공학, 패턴인식, 영상처리, HCI
E-mail : bit@dhu.ac.kr