pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2013.2.3.209

Effective Pose-based Approach with Pose Estimation for Emotional Action Recognition

Kim Jin Ok†

ABSTRACT

Early researches in human action recognition have focused on tracking and classifying articulated body motions. Such methods required accurate segmentation of body parts, which is a sticky task, particularly under realistic imaging conditions. Recent trends of work have become popular towards the use of more abstract and low-level appearance features such as spatio-temporal interest points. Given the great progress in pose estimation over the past few years, redefined views about pose-based approach are needed. This paper addresses the issues of whether it is sufficient to train a classifier only on low-level appearance features in appearance approach and proposes effective pose-based approach with pose estimation for emotional action recognition. In order for these questions to be solved, we compare the performance of pose-based, appearance-based and its combination-based features respectively with respect to scenario of various emotional action recognition. The experiment results show that pose-based features outperform low-level appearance-based approach of features, even when heavily spoiled by noise, suggesting that pose-based approach with pose estimation is beneficial for the emotional action recognition.

Keywords : Pose-based Action Features, Appearance-based Action Features, Action Recognition, Emotional Expression Recognition, Pose Estimation

자세 예측을 이용한 효과적인 자세 기반 감정 동작 인식

김 진 옥†

요 약

인간의 동작 인식에 대한 이전 연구는 주로 관절체로 표현된 신체 움직임을 추적하고 분류하는데 초점을 맞춰 왔다. 이 방식들은 실제 이미지 사용 환경에서 신체 부위에 대한 정확한 분류가 필요하다는 점이 까다롭기 때문에 최근의 동작 인식 연구 동향은 시공간상의 관심 점과 같이 저수준의, 더 추상적인 외형특징을 이용하는 방식이 일반화되었다. 하지만 몇 년 사이 자세 예측 기술이 발전하면서 자세 기반 방식에 대한 시각을 재정립하는 것이 필요하다. 본 연구는 외형 기반 방식에서 저수준의 외형특징만으로 분류기를 학습시키는 것이 충분한지에 대한 문제를 제기하면서 자세 예측을 이용한 효과적인 자세기반 동작인식 방식을 제안하였다. 이를 위해 다양한 감정을 표현하는 동작 시나리오를 대상으로 외형 기반, 자세 기반 특징 및 두 가지 특징을 조합한 방식을 비교하였다. 실험 결과, 자세 예측을 이용한 자세 기반 방식이 저수준의 외형특징을 이용한 방식보다 감정 동작 분류 및 인식 성능이 더 나았으며 잡음 때문에 심하게 망가진 이미지의 감정 동작 인식에도 자세 예측을 이용한 자세기반의 방식이 효과적이었다.

키워드 : 자세기반 동작 특징, 외형 기반 동작 특징, 동작 인식, 감정 인식, 자세 예측

1. 서 론

사람의 동작 인식은 컴퓨터 비전 분야에서 활발하게 연구되고 있는 주제로서 관련 기술은 HCI, 콘텐츠 기반의 비디오 색인, 지능형 감시와 요양시설 모니터링과 같은 응용 분야에서 실효성을 인정받고 있다. 동작에 대한 초기 연구는 관절 움직임을 이용해 신체 부위를 추적하는 것에 중점을 두었다[1-2]. 동작은 관절 자세의 연속이라는 정의에서 파생한 자세 기반의 동작인식 접근 방식은 가장 간단한 방법이지만 신체 관절 부위를 정확하게 추적해야 하기 때문에 사람의 자세를 정확히 추출해야 하고 실제 이미지 조건에서 처리해야 한다는 문제로 인해 최근에는 거의 주목을 못 받고 있다.

대신 동작 인식 대상이 특정 영화 시퀀스[3], 스포츠 중계 방송[4], 유투브 비디오[5]와 같은 자연스러운 이미지 시퀀스

† 종신회원 : 대구한의대학교 국제문화정보대학 모바일콘텐츠학부 부교수
논문접수 : 2012년 9월 3일
수 정 일 : 1차 2012년 10월 18일
심사완료 : 2012년 10월 20일
* Corresponding Author : Kim Jin Ok(bit@dhu.ac.kr)

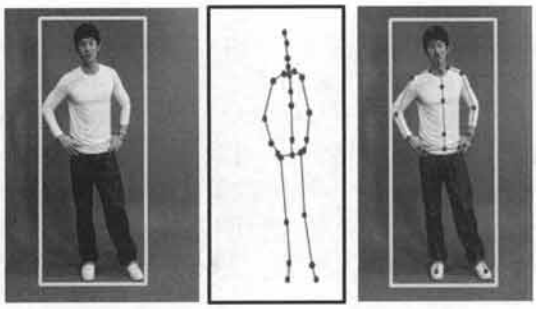
를 분석하는 형태로 바뀌면서 사람의 신체를 고수준으로 모델링하는 대신 점차 추상적이고 저수준의 외형 특징으로 동작을 분류하는 외형기반의 방법[6-9]을 주로 이용하고 있다. 외형기반의 방법은 고수준 특징 처리를 하지 않기 때문에 자세 예측을 해야 하는 어려움을 피할 수 있으며 외형 특징이 사람의 신체에만 한정되지 않기 때문에 배경과 같은 상황인식 정보를 반영할 수 있다는 장점이 있다. 또한 외형기반의 시스템은 사람의 외형, 복잡한 배경 분리, 여러 시점처리와 같이 클래스 간에 변화가 다양한 데이터를 처리할 수 있어 자세 예측이 어려운 이미지나 저 해상도 이미지에도 적용할 수 있다[7].

하지만 자세기반 방식은 동작 인식 성능이 탁월하며 외형기반 동작 인식 방식의 문제점으로 대두되어 온 클래스 간의 적은 변화량에도 강건하다는 장점이 있다. 특히 3D 형태의 스켈레턴 자세를 분석할 경우 시점과 외형이 불변하기 때문에 동작을 취하는 사람 간에 변화가 크지 않아 외형기반 방법으로는 동작을 인식하기가 어려우므로 자세기반 방법을 적용하는 것이 적절하다. 그리고 인식 대상의 고수준 특징 정보를 이미 추출한 상태라면 자세 기반 방식이 동작 인식의 학습과정을 단순화시킬 수 있다는 점도 큰 장점이다. 지난 몇 년간 자세 기반 동작 인식 연구의 발전 과정을 보면 전처리 단계인 학습과정이 단순해 졌음을 알 수 있다[11-13].

따라서 외형기반의 특징을 이용하여 동작인식을 수행하는 연구 주류에서 동작인식의 대표적 방식인 자세기반과 외형기반 방식을 비교하여 동작 분류기가 비디오 데이터에서 추출한 저수준의 외형특징에서 필요한 정보만을 확인하는 것이 효과적인지를 검증하고 자세 예측을 통한 자세 기반의 방식에 대한 재검토가 필요하다.

본 연구는 Fig. 1과 같이 동작 인식의 대표적 두 가지 접근 방식인 외형기반과 자세 기반 방식의 비교 실험과 더불어 두 방법의 특징을 조합한 방법도 실험하여 동작 인식 기술에 가장 효과적인 접근 방식을 제시하고자 한다. 실험에서는 사람의 다양한 동작 중 인식대상 동작에 집중하기 위한 정확한 인식 도메인을 설정하는 것이 필요하므로 감정 표현 동작을 대상으로 동작 인식 접근 방식에 대한 실험을 수행하도록 한다.

연구에 필요한 자세 기반의 특징은 감정을 드러낸 동작을 형성하는 서로 연관된 3D 관절 정보를 이용하고, 외형기반의 특징은 신체 모델링 대신 비디오 데이터에서 직접 외형 특징을 추출하여 이용한다. 추출한 자세기반의 특징과 외형기반의 특징 집합 그리고 두 가지 특징을 조합하여 단일 시스템 상태로 만든 조합 특징에 각각 동일한 동작 분류기[14]를 적용하여 그 결과를 고찰함으로써 감정 표현 동작 인식을 위한 자세기반, 외형 기반, 조합 방식 총 세 가지 접근 방식의 인식 성능을 비교하고 자세 예측을 통한 자세기반 방식의 효과를 제시한다.



(a) Appearance (b) Pose (c) Combination

Fig. 1. Recognition methods of emotional posture

2. 관련 연구

동작을 인식하는 초기 연구는 이미지 프레임에서 관절 위치를 찾아 자세의 특징을 시퀀스로 연결하여 처리하는 것에 중점을 두었다. 자세 특징 정보는 일반적으로 움직이는 빛 디스플레이[2], 동작 포착시스템[1], 세그멘테이션[15]을 통해 획득했다. 시퀀스 그 자체는 예제 정합[2][16]이나 HMM과 같은 상태 공간 모델로 분류하였다. 이 연구들은 자세기반의 방식으로 분류한다.

실루엣이나 시각적 볼록점[17-18]을 이용하여 신체 전체를 단일 개체로 모델링하는 방법에 대한 연구도 제시되는데 신체 실루엣 특징으로는 신체 부위의 개별 해석이 어렵기 때문에 실루엣 기반의 동작 인식 연구들은 외형기반의 접근 방식으로 분류한다.

최근의 동작 인식 연구에서는 자세 기반 방식의 관절체 추적이나 세그멘테이션 과정의 복잡함을 피하기 위해 가버 필터와 광류 분석을 통해 포착한 저수준의 외형 특징과 큐보이드[6], 3D 해리스 코너, 3D 헤시안[19]과 같은 시공간적 관심 점 특징들을 이용한 외형 기반 접근 방식이 주류를 이루고 있다. 이 방법들은 물체 검출에 사용된 2D 방법을 확장하여 동작을 인식하는 객체 검출법으로서 대상의 신체 부위를 다양한 크기의 관심 점으로 검출한 후 해당 점에 대한 특징 설명자를 계산하고 분류하여 동작 코드북에 할당한 다음 BoW(Bag of Word)표현을 통해 동작 카테고리를 결정한다[5-6].

동작 인식에 사용자 스터디와 같은 피실험자의 판별력을 이용하는 연구도 시도되고 있다[20-21]. 두 사람의 검출주체를 설정하고 동작인식 임무를 부여하는 방법으로 크게는 외형기반 동작인식 방식에 속하지만 사람의 자세 특징에 주목한다는 점이 고유한 외형기반 방식과는 다른 점을 보인다. Latev의 연구에서는 실제 환경과 비슷한 영화를 대상으로 사람의 자세를 예측하여 동작을 인식하는 방법을 제안하였다[3].

동작을 통해 드러나는 감정을 인식하는 연구도 다양하게 제시되고 있다. 신체는 얼굴보다 훨씬 많은 자유도가 있어서 신체의 전체 모양은 관절 동작에 따라 다양하게 변하기 때문에 신체 표정의 인식 자체가 어렵다. 하지만 컴퓨터 비전과 기계 학습 관련 연구[22-23]에서 동작에 나타나는 낮은 수준의 시각적 단서를 통해서도 높은 성능으로 감정 카테고리 결정을 이루어냄으로써 동작으로 그 사람의 의도를 추론하여 컴퓨팅 환경이 인간에 적절한 서비스를 제공할 수

있음을 입증하고 있다.

이와 같이 동작을 대상으로 하는 연구에서는 외형, 자세 특징을 이용해 동작 인식을 처리하고 있으며 동작으로 드러난 감정을 인식하는 방법도 제시되고 있으나 여전히 연구 방법의 영리한 단순화를 통해 신속하고 정확하게 동작을 인식하는 기술의 발전이 지속적으로 필요한 상태이다.

본 연구는 복잡한 전처리과정이 필요했던 기존 자세 기반 방식을 개선한 예측 기술을 적용하여 감정을 표현하는 동작 인식을 빠르고 효과적으로 처리하는 방법을 제안한다. 그리고 동작 인식의 대표적인 두 가지 흐름인 자세기반과 외형 기반 동작 인식 방법 그리고 두 방법의 특징을 조합한 조합 방법을 비교함으로써 자세 예측을 이용한, 전처리과정이 필요 없는 자세 기반 감정 동작인식의 새로운 방향을 제시하고자 한다.

3. 연구 방법

3.1 동작 특징

본 연구에 적용한 색상, 시공간적 기울기 등의 외형기반의 특징과 관절거리, 평면, 자세 등의 자세기반 특징을 추출한 다음 이를 학습하고 분류하는 방법에 대해 설명한다. 동작 특징과 학습에 적용한 기호는 Table 1과 같다.

Table 1. Sign description

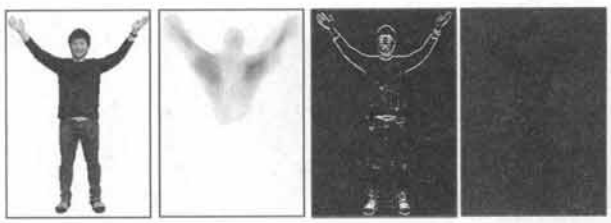
Sign	Contents
Tree T	Tree constructed random sampling patch set in training sequence
Patch Set A_i	3D spatio-temporal patch set sampled from action track
$I_i = (I_i^1, I_i^2, ..., I_i^F)$	i: selected features from patch. F : feature channel number
c_i	Action label of emotional class$(c_i \in C)$
d_i	temporal moving distance from sequence to action center
P_i	3D location of i joint
V_i	3D velocity of i joint

1) 외형 기반 특징

저수준의 외형기반 특징을 이용하여 감정 동작 인식을 수행한 기존 연구에서는 다양한 시공간적 외형 특징들을 다수 이용하고 있으나 Fig. 2의 색상, 광류, 시공간 기울기의 대표적 저수준 특징들이 기존 외형기반 방식연구의 특징들을 포괄하기 때문에 본 연구에서는 Fig. 2의 외형 특징을 이용한다. Fig. 2의 (a)는 실험실 공간 내 색상 특징이고 (b)는 x,y 좌표상의 광류 밀도이다. (c)는 x,y 좌표상의 공간 기울기 특징이고 (d)는 시간 기울기 특징이다.

2) 자세 기반 특징

자세 기반 특징을 이용할 때 가장 큰 문제점 중 하나는 의미상으로 유사한 동작이 수치상 반드시 유사하지 않다는 것이다[15][22]. 이를 고려하여 직접 동작 도메인상의 자세



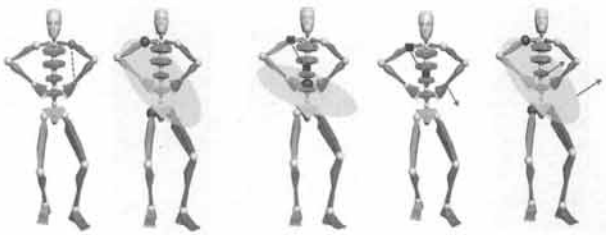
(a) (b) (c) (d)

(a) Color, (b) Optical Density, (c)Spatial Gradient, (d) Temporal Gradient

Fig. 2. Appearance based features

특징 모두를 대상으로 전체 관절 관계를 모두 설명해야 하지만 그럴 수 없어 본 연구에서는 정지 자세와 걷기 자세 그리고 짧은 감정 자세 시퀀스에서 특정 관절간의 기하학적 관계를 설명한 3D 스켈레톤 관절 관계를 자세 특징으로 이용하고 이 특징으로 자세를 예측하여 동작 클래스를 분류한다. 관계적 자세 특징은 [24]에서 소개된 것으로 모션캡쳐 데이터를 추출한 후 인덱싱하여 이용한다.

Fig. 3은 감정 자세를 취한 3D 스켈레톤 관절로 (a)는 두 관절간의 유클리디안 거리 특징을 나타내며 (b)는 평면 특징으로 붉은 색의 관절과 세 개의 검은 관절 점을 연결한 관절 평면간의 거리이다. (c)는 정규 평면 특징으로 (b) 평면 특징과 같으나 검은 사각형의 두 관절 방향으로 정규화한 평면 특징이다. (d)는 속도 특징으로 검은 사각형의 두 관절 방향에서 푸른 점 관절의 속도성분을 의미하며 (e)는 정규 속도 특징으로 검은 사각형의 세 관절이 설정한 평면에 정규화한 관절의 속도 성분이다.



(a) (b) (c) (d) (e)

(a) Joint Distance, (b) Plane, (c) Normalized Plane, (d) Velocity (e) Normalized Velocotiy

Fig. 3. Posture based features

자세 기반 특징은 스켈레톤 관절체간의 관계를 이용하기 위해 t 시간에 j_i 관절의 3D 위치와 속도를 $p_{j_{i,t}} \in \mathbb{R}^3$ 와 $v_{j_{i,t}} \in \mathbb{R}^3$으로 설정한다. Fig. 3(a)의 관절거리 특징 F^{jd}는 식 (1)과 같이 t_1과 t_2시간의 관절 j_1과 j_2 간의 유클리디안 거리로 정의한다. 동작 특징은, 자세 특징 량의 순간치가 아니라 국소적인 시간 구간에서의 평균값을 사용하면 동작타이밍의 미묘한 시간차이에 대응해서 동작을 확실하게 분류할 수 있다.

$$F^{jd}(j_1, j_2; t_1, t_2) = \| p_{j_1,t_1} - p_{j_2,t_2} \| \quad (1)$$

만약 $t_1 = t_2$이면 F^{jd}는 단일 자세에서 두 관절사이의 거리이고 $t_1 \neq t_2$이면 F^{jd}은 시간상 분리된 관절간의 거리를 인코딩한다. Fig. 3(b)의 평면 특징 F^{pl}는 식 (2)와 같이 정의된다.

$$F^{pl}(j_1, j_2, j_3, j_4; t_1, t_2) = dist(p_{j_1,t_1}, < p_{j_2,t_2}, p_{j_3,t_2}, p_{j_4,t_2} >) \quad (2)$$

여기서 $< p_{j_2}, p_{j_3}, p_{j_4} >$는 $p_{j_2}, p_{j_3}, p_{j_4}$ 로 연결된 평면을 의미하며 $dist(p_j, < \bullet >)$는 p_j 점에서 $< \bullet >$ 평면까지의 거리이다. Fig. 3(c)의 정규 평면 특징 F^{np}는 식 (3)과 같다.

$$F^{np}(j_1, j_2, j_3, j_4) = dist\,(p_{j_1,t_1}, < p_{j_2,t_2}, p_{j_3,t_2}, p_{j_4,t_2} >_n) \quad (3)$$

$< p_{j_2}, p_{j_3}, p_{j_4} >_n$ 은 p_{j_4}를 통과하는 정규 벡터 $p_{j_2} - p_{j_3}$가 있는 평면을 의미한다.

Fig. 4(d)의 속도 특징 F^{ve}는 식 (4)와 같이 t_2시간에 $p_{j_2} - p_{j_3}$ 방향을 따라 가는 v_{j_1,t_1} 벡터의 성분으로 정의한다.

$$F^{ve}(j_1, j_2, j_3; t_1, t_2) = \frac{v_{j_1,t_1} \bullet (p_{j_2,t_2} - p_{j_3,t_2})}{\| (p_{j_2,t_2} - p_{j_3,t_2} \|} \quad (4)$$

Fig. 3(e)의 정규 속도 특징은 식 (5)와 같이 t_2시간에 $p_{j_2}, p_{j_3}, p_{j_4}$ 로 연결된 평면의 정규 벡터 방향에서 v_{j_1,t_1} 벡터의 성분으로 정의한다.

$$F^{ve}(j_1, j_2, j_3; t_1, t_2) = v_{j_1,t_1} \cdot n'_{< p_{j_2,t_2}, p_{j_3,t_2}, p_{j_4,t_2} >} \quad (5)$$

여기서 n'은 $< \bullet >$평면의 단위 정규벡터이다. 이 특징들은 자세를 표현하는 패치 집합 $\{I_i = (P_i, V_i, c_i, d_i)\}$을 샘플링하여 랜덤 포레스트 학습 프레임워크에 포함시킨다. P_i와 V_i는 스켈레톤 관절의 위치와 가속도의 연속 프레임이다.

이와 같은 방법으로 관절 거리, 평면, 정규 평면, 속도, 정규 속도관절 등 자세를 결정하는 특징만을 선택하여 감정자세를 설정하고 이를 이용하여 자세를 미리 예측함으로써 관절전체를 이용하거나 많은 자세 클래스 설정으로 인해 전처리 학습과정이 복잡했던 기존 자세기반 동작 인식과 차별화했다. 자세 예측은 5.2 자세 기반 실험내용에서 Fig. 6과 같이 수행한다.

3) 조합 특징

대표적 외형특징과 자세특징을 추출한 다음 각각의 특징을 이용해 동작 인식을 수행하는 것 과는 별도로 외형과 자세 기반 특징을 조합하여, 이 정보들을 이용한 조합 패치 $\{\mathfrak{I}_i = A_i, I_i\}$로 조합 특징을 이용한 동작 인식을 시도한다. A_i는 시공간상에서 샘플링한 외형 특징 패치이고 I_i는 자세를 시간에 대해서만 샘플링한 것이다. 학습단계에서 이진 테스트를 생성할 때 외형 특징을 사용할지 자세 특징을 사용할지는 임의로 결정하여 분류기가 가장 관련성이 높은 특징을 자동으로 선택하도록 한다.

3.2 학습과 분류

본 연구에서는 이미지에서 사람의 동작을 검출하고 분류하기 위해 픽셀 기반의 래스터이미지에서 기하학적 성분을 추출하는 방법인 허프(Hough)변환과 Fig. 4와 같이 랜덤 포레스트[14]를 이용한 집단 학습과 분류를 수행한다. 허프변환은 여러 가지 특징을 동시에 다루는데 장점이 있고 랜덤

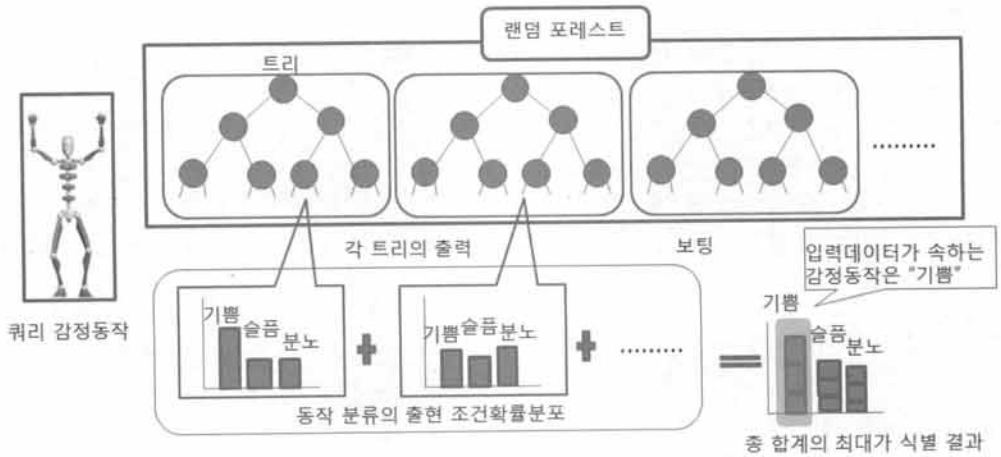


Fig. 4. Random Forest Example

포레스트 분류기는 적은 계산 양으로 높은 식별 성능을 얻을 수 있다. 특히 랜덤포레스트는 여러 가지 감정 자세 패턴에 대한 결정 트리를 만들어 보팅하면서 흔들림이나 잡음을 많이 포함하는 데이터에 대해서도 고정밀도의 판별이 가능하다.

이미지의 공간 분리단계에서는 이미지시퀀스에서 사람 검출을 위해 허프변환을 하고 랜덤포레스트를 학습하여 시퀀스 각 프레임에 대한 검출 추정치를 생성한다. 대상 검출 위치와 크기간의 강한 상관관계를 통해 추출한 결과는 트랙처럼 모이게 된다. 모인 동작은 시간상 대상의 불변 위치와 크기 표현 값을 구하기 위해 동작 트랙이라 부르는 큐보이드 표현과 매핑한다. 다음 동작 분류와 시간적 분리 단계에서는 정규화된 동작 트랙상의 동작 종류와 시공간 중심을 보팅한다.

랜덤포레스트를 이용한 보팅을 수행하여 시공간적 특징 패치와 동작 간의 매핑을 학습하고 포레스트의 각 트리는 동시에 여러 감정 동작 클래스를 판별하도록 학습한다. 학습 후에는 트리의 리프 노드 결과 집합을 클래스의 공유특징인 판별 코드북으로 간주하고 쿼리 동작이 주어졌을 때 이 판별 코드북을 통해 동작을 판별 분류한다. 랜덤포레스트의 각 트리 T는 학습시퀀스에서 임의로 샘플링한 패치 집합$\{A_i = (I_i, c_i, d_i)\}$으로 만들어진다.

트리에서 리프가 없는 노드는 이진 테스트를 저장하고 학습하는 동안 노드 테스트를 무작위로 선택하여 학습 패치의 출현 확률 분포 값이 최대가 되도록 노드를 분할한다. 리프가 생성되어 트리의 최대깊이에 도달하거나 패치가 거의 남지 않을 때까지 이 과정을 반복한다. 트리의 리프들은 리프 $L(p_c^L)$에 이르는 클래스 당 패치의 비율과 패치의 해당 이동거리 벡터 (D_c^L)를 저장한다.

특히 외형 특징에 대한 노드 테스트는 식 (6)과 같이 오프셋 τ가 있는 특징 채널 f의 p와 $q \in \mathbb{R}^3$(시공간 패치 내) 위치에 있는 픽셀들을 비교하여 수행한다.

$$t(f;p;q;\tau) = \begin{cases} 0 & \text{만약 } I^f(p) < I^f(q) + \tau \\ 1 & \text{아니면} \end{cases} \tag{6}$$

자세 특징에 대해서는 3.1절 2)에서 도출한 자세 패치를 샘플링하여 식 (6)의 이진테스트를 식(7)과 같이 변형하여 수행한다.

$$\begin{aligned} & t(f; j_1, \ldots, j_n; t_1, t_2, t_3; \tau) \\ & = \begin{cases} 0 & \text{만약 } F^f(j_1, \ldots, j_n; t_1, t_2) < \tau \\ 1 & \text{그렇지 않으면} \end{cases} \end{aligned} \tag{7}$$

여기서 $f, j_1, \ldots, j_n, t_1, t_2, \tau$는 각각 임의로 선택한 자세 기반의 특징 종류, 관절, 분류 시간, 임계치를 의미한다.

이미지시퀀스 내 동작을 분류하기 위해 테스트 트랙에서 촘촘하게 패치를 추출하여 포레스트의 모든 트리를 통과하도록 한다. 리프가 없는 노드에 저장된 이진 테스트에 따라 패치는 분할되고 도달한 리프에 따라 p_c에 비례하여 동작 레이블과 각 클래스 c의 시간 중심에 보팅을 실시한다. 즉, $p \in \mathbb{R}^3$일 때 감정 동작 클래스 c에 대한 단일 패치 집합 $A_c(p)$의 분류는 식 (8)와 같이 트리 T에서 모든 패치의 출현 조건부확률 $\rho(A_c(p)|I(q))$를 통합하여 패치를 보팅함으로써 이미지 시퀀스 동작이 어느 감정 동작 채널에 속하는지 판별한다.

$$\vartheta(p, c) = \sum_{q \in C(p)} \rho(A_c(p)|I(q), T) \tag{8}$$

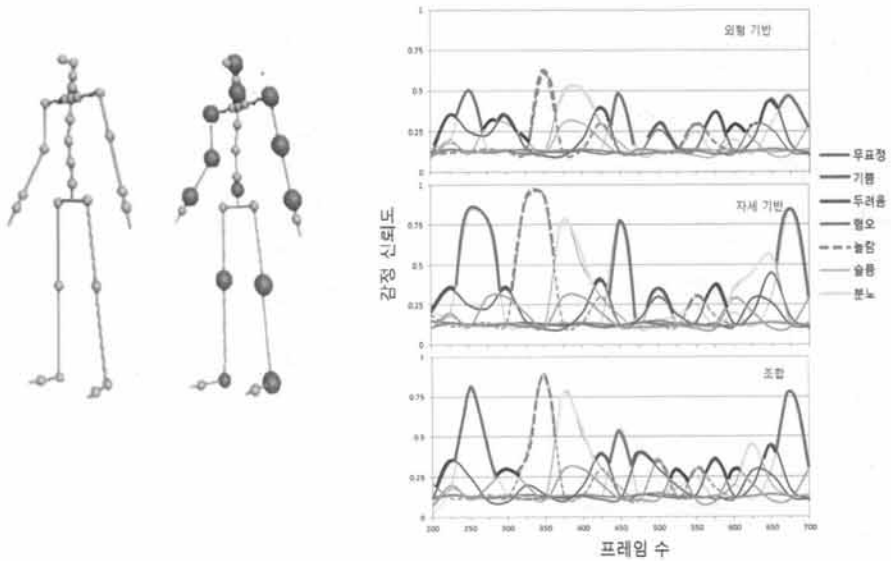
4. 실험 데이터

신체 특징을 이용한 동작 인식 비교를 위해 실내 모니터링 시나리오를 설정하고 20대 남녀 대학생을 실험자로 하여 다촛점 자세 표정 데이터를 추출하였다. 감정을 드러내는 동작을 포착하기 위해 동작자에게 실제 자연스러운 감정 자세를 취하게 했으며 외형기반의 특징 추출을 위해 배경의 변화를 일으키는 것들을 다수 제거한 채로 고정된 카메라를 통해 동작을 모니터링했다. 실험에서는 주어진 3D 관절 위치를 이용, 마커없는 모셥캡쳐 시스템으로 관절 위치를 자연스럽게 결정하도록 하여 마커에서 직접 측정한 값이 아닌 실제 자세 예측 결과를 이용한다.

각 자세 특징에 대해 모든 노드에 500번의 무작위 테스트를 생성하여 15개의 깊이가 있는 7개 트리를 학습시켰다. 정지, 걷기 두 가지 동작과 기쁨, 두려움, 혐오, 놀람, 슬픔 분노 6가지 기본 감정 동작으로 구축한 20개의 데이터집합에서 13개는 학습용으로, 나머지 7개는 테스트용으로 하여 동작 클래스 당 40개 정도의 인스턴스를 추출하였다. 랜덤포레스트의 출력을 각 동작의 신뢰도로 정규화하여 주어진 시간에 모든 동작의 값을 1로 요약했다.

외형기반의 특징으로는 배경 추출 방법을 이용하여 실루엣을 생성하고 동작 트랙에 연결된 바운딩 박스를 추출했다. 학습 단계에서는 15×15×5 크기의 1200 패치를 무작위로 선택하였다. 카메라 시점에 따라 각 포레스트를 별도로 학습하여 여러 시점에서의 출력 값을 결합하는 분류기 조합 방식을 이용하였다.

자세 기반의 특징에 대해서는 동작 인식에 상세한 자세 특징이 필요한지 분석하기 위해 Fig. 5(a), 5(b)와 같이 3D 스켈레톤의 관절 전체를 이용한 집합과 관절을 13개로 축소한 집합을 나누어 랜덤포레스트를 학습시켰다. 최적 수준의 자세를 예측하기 위한 자세 기반의 특징 실험을 위해 5 프레임의 시간에 학습 당 200개의 자세 패치를 이용하였다. 조합특징을 위해서는 외형기반 특징과 같은 실험환경을 설정하였다. 무작위로 선택한 이진 테스트를 생성할 때 절반은 외형 특징으로 나머지 절반은 자세 특징으로 하여 분류기가 최적의 테스트와 특징을 자동으로 결정하도록 하였다.



(a) Total Skeleton (b) Reduced Skeleton (c) Emotion Confidence

Fig. 5. Correlation between Skeleton-joint numbers

5. 실험 결과

5.1 외형 기반 특징

3.1절에서 설명한 외형 기반 특징을 이용하여 감정 자세를 분류한 결과 0.698의 분류 성능을 보였다. 정규화된 분류기 출력 예는 Fig. 5(c)와 같으며 Fig. 8(a)는 개별 클래스의 분류 결과이다. Fig. 8(a)에서 보여진 바와 같이 테스트한 여러 외형기반 특징 중 색상 특징은 43%로 기울기 36%, 광류 21%에 비교해 노드의 이진 테스트에서 대부분 다수 선택되었다.

Fig. 5(a)는 27개 관절 전체로 구성된 스켈레톤이고 (b)는 붉은 색의 13개 관절로 구성한 스켈레톤이다. (c)는 200개 프레임으로 구성된 0-7개의 에피소드에 대한 외형기반, 자세기반 그리고 조합 특징을 정규화한 감정 신뢰도이다. 동작 인식 신뢰도는 외형 기반보다 자세 기반의 특징을 이용한 동작 인식에서 더 높은 결과를 보였다. 특징을 더 정확하게 제공하면 외형기반의 특징보다 자세기반의 특징이 더 높은 판별력을 보임을 알 수 있다.

5.2 자세 기반 특징

모든 자세 기반의 특징은 외형기반의 특징에 비해 7-10% 더 나은 성능을 보였다. 전체 스켈레톤을 그대로 이용하여 테스트한 자세 기반의 특징 중 속도 특징과 평면 특징은 비슷한 인식 결과를 나타냈고 이 특징들은 관절거리 특징보다는 다소 높은 인식 결과를 보였다. 자세기반 특징에 따른 성능은 Table 2와 같다.

관절거리와 평면특징, 관절거리와 속도 특징, 평면특징과 가속거리 세 가지 특징 모두를 조합한 인식 성능은 Table 2에 나타난 것과 같이 0.815로 가장 높은 결과를 보였다. 축소한 스켈레톤은 전체 스켈레톤과 비슷하거나 전체적으로

Table 2. Recognition performance of posture based features

Posture features	Total Skeleton (27 joints)	Reduced Skeleton (13 joints)
Joint distance (Fig. 3a)	0.777	0.733
Plane (Fig. 3b, 3c)	0.802	0.787
Velocity (Fig. 3d, 3e)	0.803	0.803
Joint distance and Plane	0.784	0.769
Joint distance and Velocity	0.800	0.774
Plane and Velocity	0.804	0.773
Total	0.815	0.776

약간 낮은 인식 결과를 보였다. 인식 성능 저하는 축소한 관절 수 때문이 아니라 Fig. 5(a), (b)의 척추와 고관절 스켈레톤에서 줄어든 관절의 분포로 인한 것으로 추정한다. Fig. 6(c)와 같이 여러 특징을 조합했을 때의 인식 성능은 현격히 개선되지 않았으며 오히려 가장 높은 인식 결과를 보인 단일 특징보다 특징을 조합한 인식결과가 더 낮기도 했다.

Table 3은 외형, 자세, 조합 방식으로 동작이 나타내는 감정을 인식한 결과 각각 0.605. 0.832, 0.798의 평균 인식 값을 보여준다. 두려움이 모든 경우에서 가장 인식율이 낮은 감정이었고 혐오와 슬픔이 자세 기반 특징을 이용한 인식 방식에서 가장 개선된 인식율을 보였다.

관절 거리 특징만을 이용할 때의 사례를 테스트하기 위해 Fig. 6처럼 같은 노드 위치에서 다른 감정 동작을 취하는 관절 특징을 실험하였다. Fig. 6은 포레스트의 노드에 할당된 이진 테스트로 선택한 관절로서 표시된 관절의 크기는 선택빈도를 나타낸다. 이진 테스트는 관계가 있는 자세 특징으로 두 개 관절을 이용하였다. 동작이 달라지면 이용하는 관절이 달라지기 때문에 서로 다른 관절 노드 위치를 이용하는 것은 합리적이지 않다. 즉, 기쁨 상태의 동작 인식에

Table 3. Emotional recognition rate of features

(a) Appearance-based

	normal	walking	Happy	Fear	Disgust	Surprise	Sad	Anger
normal	0.45	0.08	0.02	0.04	0.03		0.04	
walking		0.88	0.05		0.03	0.02		
Happy		0.08	0.72		0.04		0.11	
Fear		0.03	0.08	0.58	0.03		0.05	0.28
Disgust	0.01	0.04			0.51	0.02	0.36	
Suprise			0.04	0.02	0.32	0.73		0.12
Sad	0.03		0.04				0.52	
Anger		0.02		0.03		0.03	0.01	0.73

(b) Pose-based

	normal	walking	Happy	Fear	Disgust	Surprise	Sad	Anger
normal	0.84	0.13	0.01	0.04	0.01		0.04	
walking		0.81	0.01		0.03	0.02		
Happy		0.08	0.93		0.02	0.08	0.02	0.06
Fear		0.04	0.08	0.79	0.03		0.02	0.11
Disgust	0.01	0.04			0.65	0.02	0.08	
Suprise			0.04	0.27		0.92		0.07
Sad	0.01		0.05				0.68	
Anger		0.04		0.03		0.03	0.01	0.9

(c) Combination

	normal	walking	Happy	Fear	Disgust	Surprise	Sad	Anger
normal	0.82	0.14		0.02	0.01		0.01	
walking		0.83	0.01		0.01			
Happy		0.08	0.9		0.07	0.04	0.03	0.03
Fear		0.03	0.08	0.75	0.03	0.03	0.07	0.14
Disgust	0.01	0.04			0.7	0.04	0.16	
Suprise		0.02	0.02	0.02	0.32	0.87	0.04	0.09
Sad	0.01		0.03				0.66	
Anger		0.01	0.02	0.03		0.03	0.05	0.92

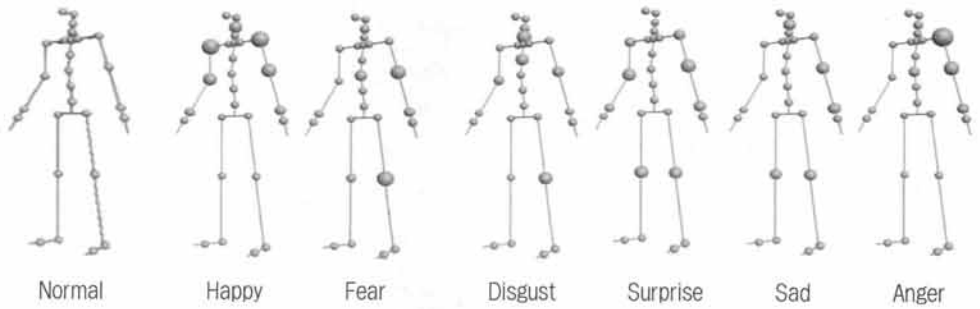


Fig. 6. Joint chosen for binary test of forest node

다리와 발의 관절만을 선택하는 것은 적절하지 않다. 혐오 동작을 수행하는 동안 허벅지에 연결되지 않은 관절이 오히려 동작을 판별하는 정보를 제공할 수도 있기 때문이다.

마지막으로 추출된 동작에서 발생한 오류를 시뮬레이션하기 위해 가우시안 잡음이 있는 테스트 관절 데이터로 바꾸어 자세 기반 특징의 강건성을 테스트했다. 분류 정확성 대비 잡음 결과는 Fig. 7과 같다. 잡음이 많아지면 가속도 특징의 분류 성능은 빠르게 저하되는 반면 관절 거리와 평면 특징의 분류 성능은 점차적으로 낮아지면서 각 관절에 추가된 잡음 범위가 75mm에 이를 때까지 비슷한 분류 성능을 유지함을 알 수 있다. 잡음 추가 범위가 100mm에 이르면 자세 기반 동작 인식 성능은 외형기반 동작 인식 방법과 거의 비슷해진다. 측면의 스켈레톤은 추가된 잡음량을 의미한다.

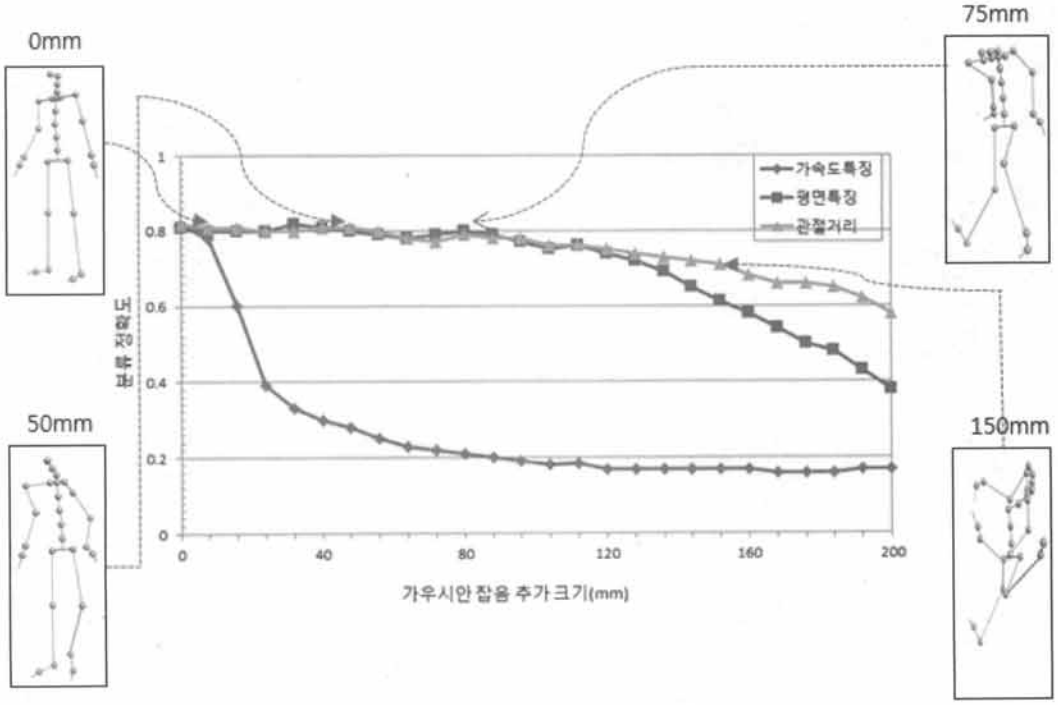
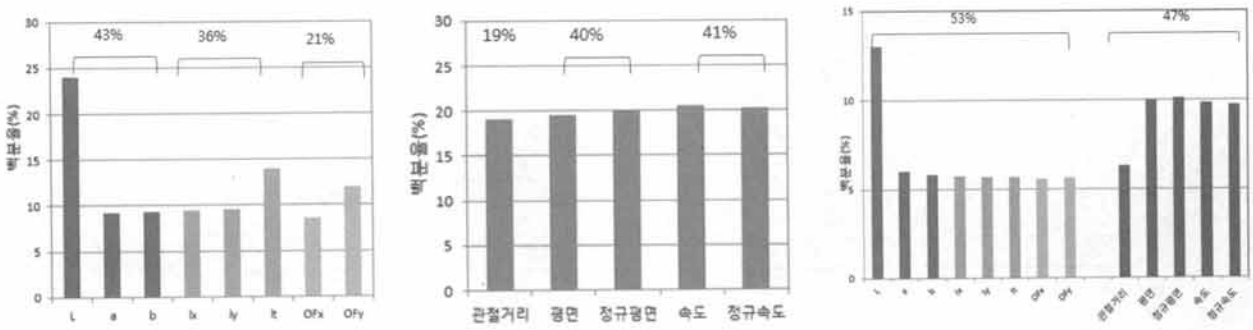


Fig. 7. Classification rate with Gaussian noise



(a) Appearance-based features (b) Pose-based feature (c) Combined features

Fig. 8. Features Selection

5.3 외형 기반 특징과 자세 기반 특징 조합

외형기반 특징과 자세 기반 특징을 조합했을 때 인식 성능은 별로 개선되지 않았으며 평균 0.806의 분류 성능을 보였다. 특징을 조합한 분류기의 인식률은 Table 3(c)와 같다. Fig. 5(c)의 정규화된 분류기 성능을 보면 조합한 특징 인식 성능이 자세 기반의 분류기와 거의 유사하기 때문에 감정 동작 인식 시 자세 기반의 특징에 강하게 반응함을 알 수 있다. 하지만 Fig. 8(c)의 트리 노드에서 선택한 특징을 보면 자세와 외형 특징이 높은 중복성을 보이면서 외형기반의 특징이 53% 정도로 더 선택되었다.

Fig. 1에서 설정한 스켈레톤 이미지 외형이 실제 동작보다 더 정확하게 동작을 설명하기 때문에 외형기반과 자세기반 특징을 조합한 방법을 통해 인식율이 대폭 개선되기를 기대했지만 Table 3은 자세기반 특징으로도 감정 동작을 이미 정확하게 예측함을 보여주고 있다.

대상 동작은 잉여 특징의 채널 수가 증가할 때 랜덤포레스트를 통해 관측하였다. 모든 특징을 같이 적용했을 때 노드에서 특징 선택의 결과는 Fig. 8(c)와 같이 거의 비슷하였다. Fig. 8(a)는 외형기반 특징을 이용한 분류기에서 노드의 43%에서 색상특징(L,a,b)을 선택한 결과이다. 기울기 특징 (I_x, I_y, I_t)는 36%, 광류(OF_x, OF_y)는 노드의 21%에서 선택되었다. (b)는 자세 기반의 분류기에서는 노드의 41%에서 관절거리 19%, 평면특징 40%, 가속 특징을 선택한 결과이다. (c)는 외형기반과 자세기반 특징을 조합하여 이용한 분류기에서는 선택한 외형, 자세기반 특징으로 노드의 53%에서 외형 기반 특징을 선택하였고 노드의 47%에서 자세기반 특징을 선택하였다.

5. 결 론

본 연구에서는 감정을 드러낸 동작을 대상으로 자세 예측을 이용한 자세기반 동작 인식 기술을 제안하였다. 제안 방법의 타당성을 설명하기 위해 현재 가장 일반적인 동작 인식 기술로 이용되고 있는 외형특징 기반 방식과의 비교를 수행하였다. 이를 위해 비디오 데이터에서 추출한 저수준의 외형 특징만으로 높은 동작 인식 성능을 보이는지를 확인하는 방법으로 외형특징 기반 방식과 자세 예측 기술을 이용한 자세특징 기반 방식을 비교하였다. 실험결과는 동일한 데이터집합에 대해 동일한 분류기를 적용했을 때 자세기반의 특징방식이 외형 기반의 특징보다 높은 분류 성능을 보였다.

자세 기반의 동작 인식 방법은 신체부위의 정확한 세그멘테이션과 다리 움직임추적에 별도의 전처리과정을 요구하기 때문에 동작 인식 연구자들이 선호하지 않았으나 본 연구를 통해 상황에 따라 자세 예측을 하면 자세 기반 방식에 별도의 전처리 과정이 필요하지 않음을 알 수 있다. 또한 비디오 데이터에 잡음이 많아지면 완벽한 자세 예측을 할 수 없기 때문에 자세 기반의 특징방식이 외형기반 방식과 동작 인식 성능이 같거나 더 나은 분류 성능을 보임을 확인하였다.

물론 외형기반의 접근 방법은 자세기반 방법보다 더 다용도로 사용가능하며 자세를 추출하지 못하는 여러 상황에서 쉽게 적용할 수 있는 장점이 있다. 또한 외형기반의 특징은 자세만으로 포착하지 못하는 상황 정보를 인코딩할 수 있다는 점에서 계속 동작 인식에 다양하게 적용될 것이지만 실험결과, 자세가 복잡한 동작 클래스인 경우 저수준 특징에서는 직접 동작을 학습하기 아주 어려우므로 고수준 정보를 통해서 주로 동작 인식이 이루어지기 때문에 잡음이 많은 동작데이터와 복합한 외형상의 문제로 인해 외형기반 접근 방식의 적용이 어려운 경우 자세 예측을 이용한 자세기반 특징 기반의 동작 인식 방식을 적용할 수 있음을 확인하였다.

이 외, 동작을 자세만으로 분류하지 못할 때 외형과 자세 특징을 조합한 방법이 가장 이상적인 인식 수단이 될 것으로 예상했으나 단일 특징을 이용한 인식 방법보다 조합한 방식의 인식 성능이 높지 않아 단일 특징을 이용한 인식방법의 장점을 보였다.

향후 연구에서는 데이터로부터 추출한 저수준 정보와 저수준 정보 중 어느 쪽을 이용해야 상황정보를 더 잘 학습할 수 있는지에 대한 문제를 실험하고 신체 동작을 통한 감정 표현 외 다양한 상황에서 나타나는 대상의 주요 움직임을 직접 이해할 수 있도록 동작 시나리오를 다양화하여 제안 방식을 계속 테스트할 계획이다.

참 고 문 헌

[1] L. Campbell, A. Bobick. "Recognition of human body motion using phase space constraints". ICCV(International Conference on Computer Vision), 1995, pp.624-630.

[2] D. Gavrila, L. Davis. "Towards 3-d model-based tracking and recognition of human movement: a multi-view approach". Int. Workshop on Face and Gesture Rec., 1995, pp.272-277.

[3] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld. "Learning realistic human actions from movies". CVPR(Computer Vision and Pattern Recognition), 2008, pp.1-8.

[4] M. D. Rodriguez, J. Ahmed, M. Shah. "Action mach: A spatio-temporal maximum average correlation height filter for action recognition". CVPR(Computer Vision and Pattern Recognition), 2008, pp.58-65.

[5] J.G. Liu, J.B. Luo, M. Shah. "Recognizing realistic actions from videos in the wild". CVPR(Computer Vision and Pattern Recognition), pp.1996-2003, 2009.

[6] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie. "Behavior recognition via sparse spatio-temporal features". Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), 2005, pp.65-72.

[7] A. Efros, A. Berg, G. Mori, J. Malik. "Recognizing action at a distance". ICCV(International Conference on Computer Vision), Vol.2, 2003, pp.726-733.

[8] J. Gall, V. Lempitsky. "Class-specific hough forests for object detection". CVPR(Computer Vision and Pattern Recognition), 2009, pp.1022-1029.

[9] J. Sivic. "Efficient visual search of videos cast as text retrieval". IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.31, No.4. pp.591-605, 2009.

[10] K. Schindler, L. Van Gool. "Action snippets: How many frames does human action recognition require?", CVPR(Computer Vision and Pattern Recognition), 2008, pp.1-8,

[11] J. Bandouch, M. Beetz. "Tracking humans interacting with the environment using efficient hierarchical sampling and layered observation models, Int. Workshop on Human-Computer Interaction", 2009, pp.2040-2047,

[12] J. Gall, A. Yao, L. Van Gool. "2d action recognition serves 3d human pose estimation", ECCV(European Conference on Computer Vision), 2010, pp.425-428.

[13] G. Taylor, L. Sigal, D. Fleet, G. "Hinton. Dynamical binary latent variable models for 3d human pose trackin". CVPR(Computer Vision and Pattern Recognition), 2010, pp.631-638.

[14] L. Breiman. "Random Forests. Machine Learning", Vol.45, No.1, pp.5 - 32, 2001.

[15] L. Kovar, M. Gleicher. "Automated extraction and parameterization of motions in large data sets". ACM Trans. Graph., Vol.23, pp.559 - 568, 2004.

[16] C. Rao, A. Yilmaz, M. Shah. "View-invariant representation and recognition of actions". IJCV(International Journal of Computer Vision), Vol.50, No.2, 2002, pp.203 - 226.

[17] D. Weinland, E. Boyer, R. Ronfard. "Action recognition from arbitrary views using 3d exemplars". ICCV(International Conference on Computer Vision), 2007, pp.1-7.

[18] D. Weinland, E. Boyer. "Action recognition using exemplar-based embedding", CVPR(Computer Vision and Pattern Recognition), 2008, pp.1-7.

[19] R. Li, T.P. Tian, S. Sclaroff, M. H. Yang. "3d human motion tracking with a coordinated mixture of factor analyzers". IJCV(International Journal of Computer Vision), Vol.87, 2010, pp.170 - 190.

[20] C. Thurau, V. Hlavac. "Pose primitive based human action recognition in videos or still images". CVPR(Computer Vision and Pattern Recognition), 2008, pp.1-8.

[21] A. Kläser, M. Marszałek, C. Schmid, A. Zisserman. "Human focused action localization in video". Int. Workshop on Sign, Gesture, and Activity (SGA), 2010.

[22] Z. Zeng, M. Pantic, G. Roisman, T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.31, No.1, pp.39-48, 2009.

[23] Kim Jin Ok, "A Study on Visual Perception based Emotion Recogniton using Body-Activity Posture", The KIPS Transactions: Part B, Vol.18, No.5, pp.305-314, 2010.

[24] M. Müller, T. Röder, M. Clausen. "Efficient content-based retrieval of motion capture data". ACM Trans. Graph., Vol.24, pp.677 - 685, 2005.

김 진 옥

e-mail : bit@dhu.ac.kr

1989년 성균관대학교(학사)

1998년 성균관대학교(석사)

2002년 성균관대학교 전기전자및컴퓨터공학과(박사)

1992년~1994년 현대전자산업(주) 정보통신사업부 재직

1994년~1999년 현대정보기술(주) 인터넷사업부 재직

2004년~현 재 대구한의대학교 국제문화정보대학 모바일콘텐츠학부 부교수

관심분야 : 멀티미디어공학, 패턴인식, 영상처리, 유비쿼터스 컴퓨팅, 융합공학 등