모션 그래디언트 히스토그램 기반의 시공간 크기 변화에 강인한 동작 인식

(Spatial-Temporal Scale-Invariant Human Action Recognition using Motion Gradient Histogram)

김 광 수 † 김 태 형 †† 곽 수 영 ††† 변 혜 란 ††††

(Kwangsoo Kim) (Taehyoung Kim) (Sooyeong Kwak) (Hyeran Byun)

요 약 본 논문은 동영상에 등장하는 다수 사람의 동작을 검출하여 검출된 동작을 개별적으로 인식하는 방법을 제안한다. 동작이 수행되는 속도 또는 크기 변화에 강인한 인식 성능을 갖기 위해 시공간축 피라미드(Spatial-Temporal Pyramid)방식을 적용한다. 동작 표현 방식을 통계적 특성 기반의 모션 그래디언트 히스토그램(MGH:Motion Gradient Histogram)으로 선택하여 인식 과정에서 발생하는 복잡도를 최소화 하였다. 다수의 동작을 검출하기 위하여 이진 차영상을 축적한 모션 에너지 이미지(MEI: Motion Energy Image) 방법을 적용하여 효율적으로 개별적 동작 영역을 획득한다. 각 영역은 동작 표현 방법인 MGH로 나타내어지고, 크기 변화에 강인하도록 피라미드 방식을 적용하여 학습된 템플릿 MGH와 유사도를 상호 비교하여 최종 인식 결과를 얻는다. 인식 성능의 평가를 위해 10개의 동영상을 활용하여 단일 객체, 다수 객체, 속도 및 크기 변화, 기존 방식과의 비교, 기타 추가 실험 등을 실시하여 다양한 조건의 영상에서 양호한 인식 결과를 확인 할 수 있었다.

키워드 : 동작인식, 다수 동작 검출, 모션 그래디언트 히스토그램, 모션 에너지 이미지

Abstract In this paper, we propose the method of multiple human action recognition on video clip. For being invariant to the change of speed or size of actions, Spatial-Temporal Pyramid method is applied. Proposed method can minimize the complexity of the procedures owing to select Motion Gradient Histogram (MGH) based on statistical approach for action representation feature. For multiple action detection, Motion Energy Image (MEI) of binary frame difference accumulations is adapted and then we detect each action of which area is represented by MGH. The action MGH should be compared with pre-learning MGH having pyramid method. As a result, recognition can be done by the analyze between action MGH and pre-learning MGH. Ten video clips are used for evaluating the proposed method. We have various experiments such as mono action, multiple action, speed and size scale-changes, comparison with previous method. As a result, we can see that proposed method is simple and efficient to recognize multiple human action with scale variations.

Key words : Action recognition, Multiple action detection, Motion gradient histogram, Motion energy image

· 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음(IITA-2007-(C1090-0701-0046)). 본 연구는 한국과학재단 특정기초연구 지원으로 수행되었음(R01-2005-000- 10898-0)

† 정 회 원 : 현대자동차 CL사업부
kwangsoo.kim.kk@gmail.com
†† 정 회 원 : LG전자 MC사업부
erkth@naver.com
††† 학생회원 : 연세대학교 컴퓨터과학과
ksy2177@cs.yonsei.ac.kr
†††† 종신회원 : 연세대학교 컴퓨터과학과 교수
hrbyun@cs.yonsei.ac.kr

논문접수 : 2007년 1월 22일
심사완료 : 2007년 10월 15일

정보과학회논문지: 소프트웨어 및 응용 제34권 제12호(2007.12)

1. 서 론

본 논문은 동영상에서 나타나는 다수 사람의 동작을 각각 효과적으로 인식하는 방법을 제안한다. 동작 인식은 컴퓨터 비전 분야에서 최근 활발히 진행되고 있는 연구 분야 중 하나이다. 동영상에 포함된 각종 움직임들의 효과적인 분석을 통한 비전 기반 인식 시스템은 대용량 비디오 데이타 관리 및 검색, 동영상 분류, 가상현실시스템, HCI(Human Computer Interface), 컴퓨터 애니메이션/게임 등의 기술 응용에 필수 요소라 할 수 있다[2,3]. 특히 사람의 움직임을 검출하고 인식하는 기술은 이동 로봇, 비디오 감시 시스템, 인간과 로봇과의 상호 작용 등 여러 응용에서 연구 되고 있는 중요한 분야 중 하나이다[1].

사람의 동작을 인식하기 위하여 기존의 많은 방법론이 제안되었다. 다양한 방법론을 방법적 특성에 따라 정리해 보면 모델링 특수성, 외형적 의존성, 계산적 복잡성, 환경적 변화성에 따라 방법론들이 나누어진다. 모델링 특수성을 가지는 방법론은 특정 객체의 인식을 위해 모델링을 사용하는 방법으로써 해당 객체마다의 특수한 파라메타에 대한 분석이 사전 작업으로 필요하며 해당 작업의 완성도에 따라 인식 결과에 중요한 영향을 미치게 된다[5,6]. 외형적 요소에 따른 인식은 객체의 크기나 색상 및 텍스쳐 정보 등 객체의 외형적 상황에 그 결과가 민감하고, 객체의 정확한 외형을 항상 추출해 내기 쉽지 않다는 어려움이 있다[4,8]. 계산적 복잡성을 가지는 방법인 광류(Optical Flow) 혹은 상관(Correlation) 함수 등을 사용하여 사람을 동작을 인식하는 방법들[10,11]은 계산이 복잡하기 때문에 처리 속도 및 구현의 복잡도가 높아지므로 낮은 효율을 갖게 된다는 단점이 있다. 외형적 요소를 기반으로 하는 경우 주로 배경 추출 방식을 적용하게 되는데 이 경우 조명, 질감 등의 환경적 외부 요소에 의한 영향을 받게 되기 쉽다는 단점이 있다[4,8].

본 연구에서는 이러한 다양한 방법론의 단점을 극복하기 위한 방안을 제시하는데 연구의 주안점을 두었으며, 기존 연구들이 주로 독립된 하나의 객체에 대한 인식을 위주로 연구하는 반면, 본 논문에서는 동영상 내에서 동시에 나타나는 다수 객체에 대한 인식을 수행하는 방안을 제시하였다. 또한, 동일한 동작에 대한 객체의 크기나 움직임의 속도 차이에도 강인한 인식이 가능하도록 하는 방안을 제시하였다.

2. 시스템 개요

본 논문은 동영상 내에서 나타나는 사람의 동작을 인식하는 방법을 제안한다. 동작의 인식을 위해 필요한 사전 학습을 최소화하여 복잡한 모델링 과정 없이 구분되는 동작 자체의 특징만으로 간단하게 인식이 가능하도록 모션 그래디언트 히스토그램(MGH : Motion Gradient Histogram)을 사용하였다. 이는 시공간축 상에서 효과적으로 동작을 표현하는 방법 중 하나이다[9]. 이에 대한 장점은 배경 추출 과정과 객체에 대한 별도의 모델링 과정이 필요하지 않으며 형판(Template)기반의 간단한 학습만으로 사전 학습이 가능하여 계산 복잡도를 크게 낮출 수 있다. 또한 조명, 질감 등의 환경적 외부 요소의 영향에 민감하지 않으며 실루엣 추출이 불필요하다는 등의 외형적 요소의 의존도가 낮다.

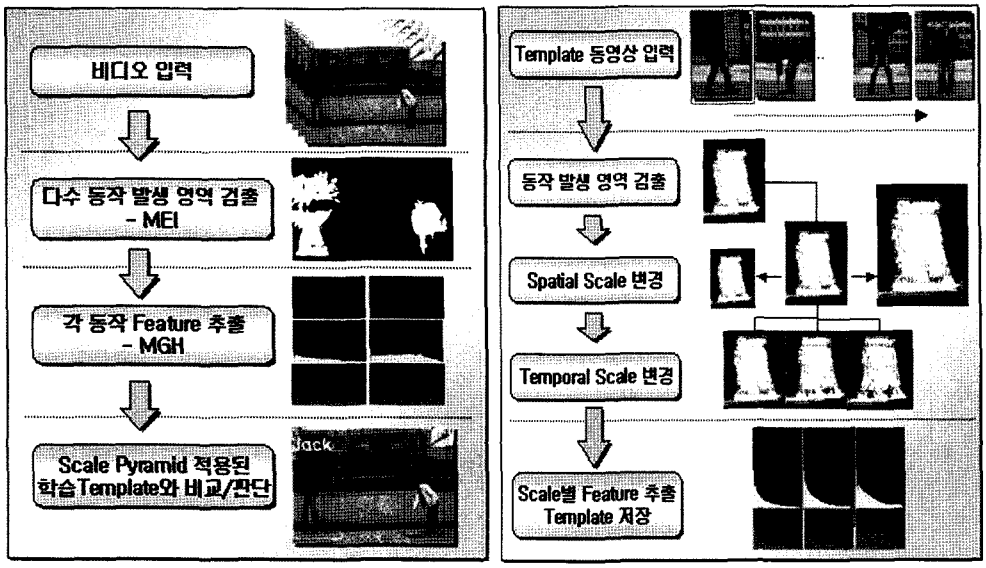
동시에 나타나는 다수 사람의 동작을 각각 인식하기 위해서 동작이 발생하는 영역을 모션 에너지 이미지(MEI : Motion Energy Image)[7]를 이용하여 분리하고 각각의 영역에 대해 MGH를 얻어냄으로써 다수 동작을 개별적으로 인식할 수 있도록 한다.

동일한 동작이라고 하더라도 동작을 수행하는 객체의 크기 차이 혹은 동작 속도 차이가 현저한 경우 인식에 실패하는 경우가 있는데, 이를 보완하기 위해 공간적 크기 변화에 시간적 속도 변화에 따른 학습을 추가적으로 적용한다. 공간적 크기 변화에 따른 학습의 경우 기준이 되는 템플릿 크기의 일정한 비율로 축소 확대하여 템플릿을 생성하게 되고, 마찬가지로 시간적 속도 변화에 따른 학습의 경우 가급적 느린 속도를 기준으로 하여 2배, 4배 빠른 영상을 생성하게 된다. 이 과정은 추가적인 템플릿을 필요로 하는 것이 아니라 기준 템플릿을 바탕으로 자동적으로 생성하는 것이므로 부가적인 데이타 수집 절차가 불필요한 장점을 갖는다.

그림 1(a)와 그림 1(b)는 각각 제안하는 알고리즘에 대한 수행 과정 및 학습 시에 템플릿이 되는 동작 특징(MGH)을 획득하는 과정을 개략적으로 나타낸 것이다.

3. 동작 발생 영역 검출

동작 발생 영역을 얻기 위해서 식 (1)의 MEI(Motion Energy Image)[15]를 구하여 해당 영역을 검출한다. 그림 2의 MEI는 동영상에서 일정한 수의 프레임(FWS:: Frame Window Size) 동안 프레임간 이진 차영상을 축적하여 얻은 영상이다. 동작 발생 영역을 검출할 수 있다는 것은 다수의 동작이 발생하는 경우 개별적으로 동작 영역을 구분하여 계산 할 수 있으므로 다수 개체의 동작 인식이 가능할 수 있다는 것을 의미하므로 매우 중요한 처리 과정이다. 프레임간의 이진 차영상을 이용하므로 부가적인 배경 모델링의 처리 과정을 필요로 하지 않는 장점이 있는 반면, 카메라의 움직임에 민감하게 반응하기 때문에 반드시 카메라가 고정된 상태에서만 정확한 결과를 얻을 수 있는 제약이 있다.



(a) (b)

그림 1 (a) 제안하는 알고리즘 수행 과정 (b) 학습 템플릿 동작 특징(MGH)획득 과정

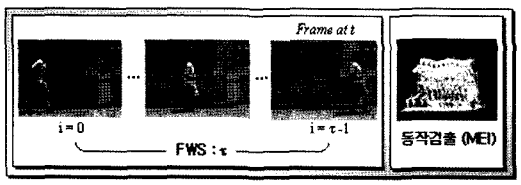


그림 2 MEI를 이용한 동작 발생 검출의 예

$$MEI_\tau(x,y,t)=\bigcup_{i=0}^{\tau-1} D(x,y,t-i) \quad (1)$$

D(x,y,t) : 프레임간 이진 차영상

Y : 프레임 윈도우 크기 (Frame Window Size)

일반적으로 아무런 처리를 하지 않은 MEI는 동작의 속도 및 형태, 주변 환경에 따라 잡음이 발생하며 동일 객체에서 나타나는 하나의 동작임에도 불구하고 차영상의 발생이 연결되어지지 못한 상태로 나타나기도 한다. 이러한 경우를 최소화하기 위해 잡음 제거 과정을 거쳐 잡음을 최소화 하고 모폴로지의 닫힘연산을 수행하여 좁게 끊겨져 있는 부분들과 가늘고 길게 떨어진 부분들을 합치고 작은 구멍들을 제거하여 하나의 동작 영역으로 만든다. 이 경우 잡음으로 제거되지 못한 일부분 혹은 미세한 조명의 변화나 주변 물체의 작은 움직임에 의해서 발생 가능한 차영상 영역이 다소 확실해져 버리는 현상이 있으므로 이를 동작 영역으로 추정되지 않게 하기 위해 동작이 발생된 윈도우의 크기를 일정 크기 이상인 것만 의미 있는 동작 영역으로 판단하도록 하는 과정을 수행함으로써 보다 정확한 동작 발생 영역 검출이 가능하다. 이러한 동작 발생 영역 검출을 통해 다수 동작을 구분하여 처리하게 되는 것이다.

4. 동작 인식

동영상에 나타나는 동작을 인식하기 위해서는 시간축을 고려한 영상의 집합체로써 인식하고자 하는 대상을 다루어야 한다. 즉, 동영상 내의 일정 프레임 간격(FWS)을 기준으로 동작이 나타나는 부분에 대한 다양한 방식의 표현 방법이 하며, 이 중 본 연구에서는 Zelnik[9]이 제안한 모션 그래디언트 히스토그램을 사용한다.

4.1 동작의 표현

MGH은 시공간적 그래디언트(Space-Time Gradient)의 통계적 분포에 기반 한 방법으로써 실루엣 등과 같은 직관적인 외형적 요소를 배제하고, 프레임 간 변화 정도를 중시하므로 배경 모델링과 같은 제약된 상황을 벗어날 수 있다. MGH를 얻는 알고리즘은 다음과 같다. 동영상 내의 일정 프레임 간격에서 발생하는 동작을 표현하기 위해 시공간 고려한 영상 축적체(Space-Time Gradient Volume)를 고려한다. 이러한 S내의 모든 위치에서는 식 (2)에 의해 Local Space-Time Gradient (LSTG)를 얻을 수 있다(실제 계산은 식 (3) 참고).

$$(S_x,S_y,S_t)=\begin{cases}\left(\frac{dS}{dx},\frac{dS}{dy},\frac{dS}{dt}\right), & \frac{dS}{dt}\geq\delta\\ (0,0,0), & \frac{dS}{dt}<\delta\end{cases} \quad (2)$$

$$S_x = (S(x+1,y,t) - S(x-1,y,t)) * 0.5 \quad (3)$$
$$S_y = (S(x,y+1,t) - S(x,y-1,t)) * 0.5$$
$$S_t = (S(x,y,t+1) - S(x,y,t-1)) * 0.5$$

$$(N_x, N_y, N_t) = \frac{(|S_x|, |S_y|, |S_t|)}{\sqrt{S_x^2 + S_y^2 + S_t^2}} \quad (4)$$

즉, 모든 화소(x,y)에서 일반적인 정지영상의 그래디언트와 시간축 프레임간 차이를 동시에 고려한 벡터가 존재하게 된다. 특히 프레임간 차이가 특정 임계값 이하 픽셀인 경우 움직임이 없는 것으로 판단하여 해당 위치의 LSTG를 0으로 함으로써 해당 부분을 제외하고 움직임이 명확한 부분에 대해서만 처리해 줄 수 있다. 이 경우 대체로 급격하지 않은 조명 변화나 노이즈에 강인한 결과를 얻게 된다. 이와 같이 구해진 LSTG를 식(4)에 의해 정규화된 STG(N-STG : Normalized Space-Time Gradient)로 나타낼 수 있다. 정규화를 통해서 외형 요소(조명, 질감, 색상 등)의 영향을 최소화 시킬 수 있으며, 절대값을 취함으로써 방향 성분(좌,우)을 제거한다. 방향성분을 제거하는 이유는 예를 들어, 좌측에서 우측으로 걸어가는 사람의 동작과 반대로 우측에서 좌측으로 걸어가는 사람의 동작이 상호 동일한 동작임을 표현하기 위한 방안이다. 이 경우 동작 자체만을 판단하기 위해서는 유용한 방법이긴 하나, 반면에 방향 정보를 알고 싶은 경우는 그것이 불가능하다는 단점을 갖고 있으므로, 절대값 여부는 어떤 것에 주안점을 두느냐에 따른 선택의 문제라 볼 수 있다. 본 연구에서는 방향성을 고려하지 않으므로 절대값을 사용하고 있다.

정해진 FWS 내에서 모든 점들의 N-STG를 얻게 되면, 각 성분별 히스토그램 ({hx, hy, ht})을 계산하여 이를 MGH라 한다. 일반적으로 FWS는 1초 정도의 간격을 갖는 프레임 수만큼 결정한다. (30fps 동영상을 기준으로 32프레임 정도) 그러나 실제로 FWS의 결정이 인식의 결과에 중요한 영향을 미치는 요소이므로 인식하고자 하는 동작 대상이 얼마 동안의 최소 시간으로 인식 가능한지에 따라 FWS를 변경시켜야 할 필요가 있다. 예를 들어 1초 정도에 인식이 가능한 동작이라면 큰 문제가 없겠지만 동작의 특성상 4~5초 정도의 긴 시간을 요하는 인식 동작이라면 1초의 FWS로는 인식 오류가 당연히 높아질 수밖에 없는 것이다. 이러한 문제는 일반적인 동영상을 다루는 인식 연구에서 나타나는 근본적인 문제점으로써 이에 대한 보완적인 해결 연구가 필요할 것으로 보인다.

계산되어 얻어진 3개의 1차원 히스토그램 집합인 MGH은 해당 동영상 구간에서 발생하는 동작을 나타내는 표현 방법으로 사용된다. 즉, 서로 다른 동영상 간의 MGH 유사도를 비교함으로써 동작의 유사성을 판단할 수 있게 되는 것이다. 특히 동작 특징이 대상체의 질감에 따라 민감하지 않게 하기 위해 전처리로써 블러링 과정을 추가해 주는 것이 일반적이다.

4.2 동작 유사도 판단

동작 인식이 이루어지는 과정은 학습된 템플릿 MGH와 동영상 MGH간의 유사도를 비교하여 해당 유사도가 가장 높은 템플릿의 동작을 채택하게 된다. 궁극적으로 히스토그램의 유사도를 비교하는 것이므로 아래 식 (5)의 카이-제곱근 다이버전스(Khai-Square Divergence)를 통해 유사도(SM)를 계산하게 된다. 동영상 S1과 S2간의 MGH유사도 값이 0에 가까울수록 유사한 동작으로 판단할 수 있다. 정해진 임계값 이상의 유사도가 나올 경우 인식하고자 하는 동작들에 속하지 않은 동작으로 판단한다.

$$SM(S_1, S_2) = \sqrt{\sum_{k,i} \frac{[h_{1k}(i) - h_{2k}(i)]^2}{h_{1k}(i) + h_{2k}(i)}} \quad (5)$$

$SM(S_1, S_2)$: 동영상 S_1, S_2 간 유사도

$h(i)$: 히스토그램 I번째 bin에서의 수

$k \in \{x, y, t\}$

5. 시공간 크기 변화에 강인한 알고리즘

5.1 공간축 피라미드(Spatial Pyramid)

객체의 다양한 크기에 대해 강인성(Spatial Scale Invariant)을 갖기 위해 그림 3(b)와 같이 그림 3(a)의 학습 템플릿 동영상을 축소, 확대 시켜 얻은 MGH를 사전 학습 과정 단계에서 생성, 저장한다. MGH의 특성상 크기 변화가 심하지 않은 경우는 상호 유사도가 높게 나타나므로 실험을 통해 유사도의 차이가 크게 나타나는 크기 비율을 고려하여 공간축 피라미드를 생성한다. 그림 3(b)에서는 원본의 50%, 150%를 각각 비교한 것으로, 50%인 경우 원본의 MGH와 많은 차이가 나타나는 것을 알 수 있고, 150%인 경우 원본과 유사함을 알 수 있다. 즉, 50%의 축소된 MGH는 인식 오류를 최소화 하는데 의미가 있으며, 150%인 경우는 큰 의미가 없으므로 그 이상의 확대를 통해 인식의 오류를 최소화 할 수 있다고 잠정적으로 판단할 수 있다. 이와 관련해서 본 논문에서는 공간축 피라미드 적용시 최대 50%, 100%, 180%의 세 단계를 설정(SP=3)하여 실험하였다. 200%의 확대를 하지 않은 이유는 일부 동작의 경우 200% 확대 시 전체 영역을 벗어나는 동작이 발생하므로 이 경우 인식 오류의 가능성이 크기 때문에 해당 비율 이상의 확대는 배제하였다. 이론적으론 축소, 확대가 어떠한 비율로도 가능하지만 실제 정상적인 결과를 얻기 위해서는 너무 작은 크기의 동작은 동작 발생 영역으로 추출될 수 없으므로 의미를 지니지 못하게 되며, 반대로 동작이 나타나는 전신(full body)이 한 장면에

포함되어야 하는 조건이 있으므로 확대 시에도 크기의 제약이 있을 수밖에 없다.

5.2 시간축 피라미드(Temporal Pyramid)

객체 움직임의 속도 차이에 의한 인식의 오류를 최소화(Temporal Scale Invariant) 하기 위해 그림 3(c)와 같이 학습 템플릿 동영상의 시간축 피라미드를 구성하여 얻은 MGH를 저장한다. 즉, 사람이 동작을 행할 때 동일한 동작에 대해서도 개인별 혹은 상황에 따라 동작이 발생하는 속도가 차이가 나게 되는데 이러한 차이를 극복하기 위해 사전에 속도 변화에 따른 다수의 MGH를 얻을 필요가 있는 것이다. 그림 3(c)는 템플릿 동영상의 기준(100%) 크기에서의 시간축 피라미드를 3단계로 구성하였을 경우(TP=3)의 예이다. 좌측에서부터 정상 속도(100%), 2배 빠른 속도(200%), 4배 빠른 속도(400%)의 MEI와 MGH를 각각 나타내고 있다. 공간축 피라미드의 경우는 축소와 확대가 가능하지만 시간축 피라미드의 경우는 기준 템플릿에서 빠른 속도의 동영상만 생성 가능하므로 이를 고려하여 다소 느린 동작의 동영상을 학습 템플릿으로 하여야 효율적인 시간축 피라미드가 생성됨을 알 수 있다. 이 경우 '느리다', '빠르다'의 기준이 수치적으로 애매하므로 일반적인 관점에서 특정 동작을 해당 동작이라 인정할 정도의 최소한의 속도로 동작을 수행할 경우를 학습 템플릿 동작의 속도로 여겨야 할 것이다.

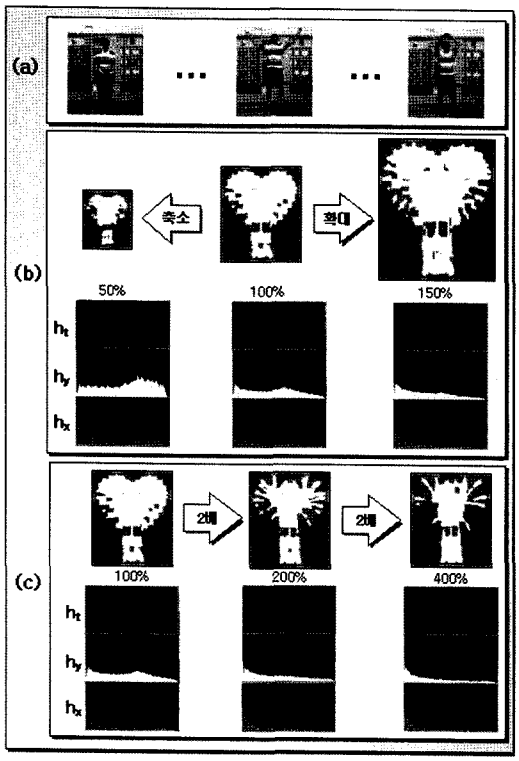


그림 3 시공간축 피라미드 적용 MEI/MGH의 예 : (a) 템플릿 동영상 샘플(팔벌려뛰기) (b) 공간축 피라미드 생성 (c) 시간축 피라미드 생성

6. 실험 및 결과

본 논문에서 제안한 시스템은 Windows 2000에서 Visual C++ 6.0을 이용하여 구현하였으며, 펜티엄-IV의 CPU 2.4 GHz와 1GB RAM의 하드웨어에서 실험하였다. 실험 데이타는 실내외 다양한 환경에서 촬영된 10개의 데이타(A-tc1, B-t1, C-t2, D-t2, E-t1, F-ntb1, G-t1, H-nc1, I-t1, J-tbc1)를 이용하였다(t:실외, n:실내, b:순간적인 장면전환 포함, c:조명변화 심한 중저화질, 숫자:동시 등장인물의 수). 실험 데이타 A~G는 등장인물이 기본 6동작(Walk, Bend, Jack, Jump, Wave, Run)중 일부를 수행하며, H는 다이어트체조 6동작, I는 집총회전 2동작, J는 축구공 다루는 3동작을 수행하는 시나리오로 되어 있다. 제안한 방법은 1명 동작 인식시 평균 8.5fps의 속도를 보였으며, 2명 동작 인식시 평균 6.5fps의 속도를 보였다.

6.1 인식 성능 평가 기준

성능 평가 결과는 정해진 FWS 간격 내에서 발생한 동작에 대해 관찰자가 직접 확인하여 학습된 동작으로 판명되고, 이때 실험 결과로써 모니터 상에 표시되는 동작이 해당 동작과 일치하는 경우, 해당 FWS에서 성공으로 판단한다. 인식 성공율(SR)은 아래 식 (6)과 같이 계산한다.

$$SR(\%) = \frac{ST}{Floor\left(\frac{TFN \times N}{FWS}\right)} \times 100 \qquad (6)$$

TFN : 동영상의 총 프레임 수
ST : 개별 등장 인물의 성공 횟수 총합
N : 동시에 등장하는 인물의 수
FWS : 인식의 단위 프레임 수 (8~60사이의 정수값)

6.2 평가 결과 및 분석

실험 데이타에 따른 인식 결과를 표 2에 나타내고 그 중 일부 수행 장면을 그림 4에서 보여준다. 데이타에 따라 프레임의 수와 적용된 FWS, 시공간축 피라미드의 정도(TP/SP), 등장하는 객체의 수, 인식하는 동작의 수, 영상이 촬영된 환경 등이 각각 상이하여 처리속도의 편차가 나타남을 알 수 있다. 그러나 순간적인 장면전환 및 조명변화가 심한 중저화질의 J데이타의 결과를 제외하고는 전반적으로 양호한 인식결과(평균 90.35%)를 나타내었으며 C,D데이타에서 다수 등장시의 인식도 가능함을 보여주었다. 특히 E,F데이타의 경우, 각각 시간축 피라미드와 공간축 피라미드의 효과를 알아보고자 목적

표 2 실험 데이타별 인식율 결과

	A	B	C	D	E	F	G	H	I	J
총프레임수	2371	552	500	792	985	448	792	2208	407	420
동시객체수	1	1	2	2	1	1	1	1	1	1
TP/SP	2/1	2/1	2/1	2/1	3/1	2/3	2/2	2/1	3/1	2/2
FWS	24	24	24	24	32	14	24	32	12	16
인식동작수	5	6	6	6	3	3	6	6	2	3
속도(fps)	8.68	8.24	6.83	6.37	6.45	4.53	5.30	7.92	6.23	5.14
인식율(%)	91.90	86.96	90.24	89.39	93.33	90.62	90.90	94.20	96.00	80.00

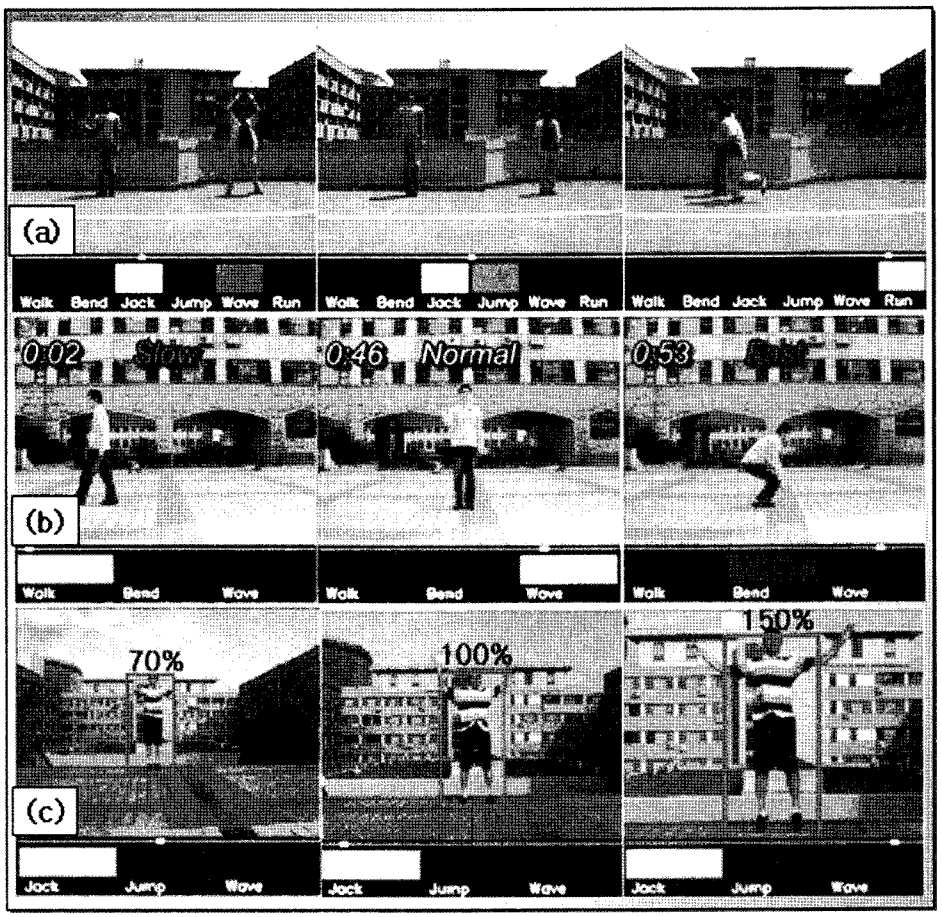


그림 4 제안한 시스템에 의한 수행 결과
(a) C데이타 : 다수 등장, (b) E데이타 : 동작 속도 변화, (c) F데이타 : 객체 크기 변화

에 맞게 촬영한 영상(속도를 구분하여 동작을 실시하고, 줌인/줌아웃으로 객체의 크기를 의도적으로 변화시킴)으로써, 두 데이타 모두 TP/SP=1/1로 하였을 때 각각 66.67%, 68.75%의 저조한 인식율을 보였다. 즉, 시공간 축 피라미드의 적용으로 성능의 개선이 있음을 확인할 수 있었다.

6.3 기존연구와의 비교 실험

제안하는 인식방법과 기존 연구와의 성능 비교를 위해 동작 인식 방법 중 일반적으로 잘 알려진 Bobick과 Davis가 2001년도에 제안한 MHI(Motion History Image) 방법[7]과 비교 실험을 하였다. 기본 6동작을 임의로 실시하는 단일객체에 대한 인식결과와 MHI 방법의 실험결과를 그림 5에 나타내었다. 위쪽 프레임이 본 논문에서 제안하는 방법이고 아래쪽 프레임이 MHI 방법의 결과이다. (a)와 같이 전박적으로 양호한 결과를 볼 수 있었는데, 동작 전이 부분 혹은 동작의 초기부분(b), 그리

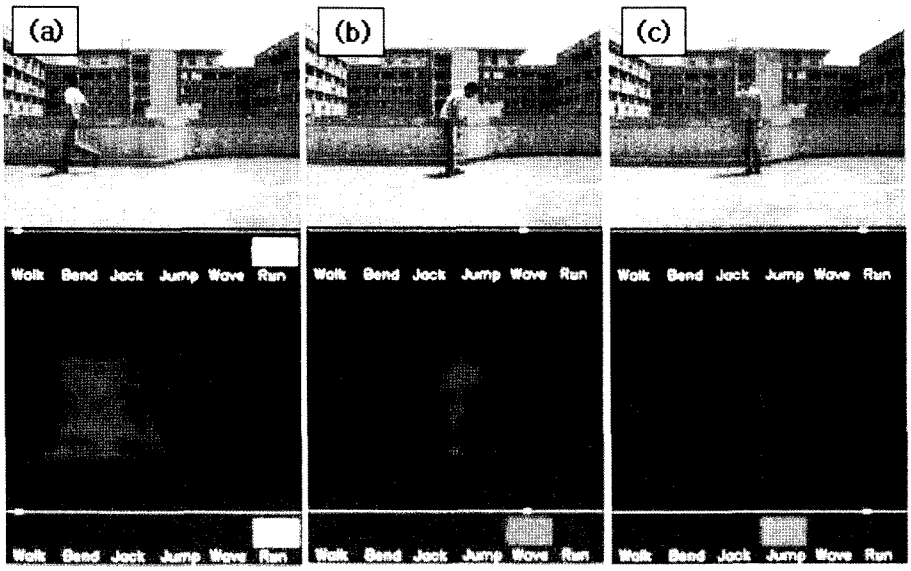


그림 5 제안하는 방법과 MHI 방법의 인식 결과(<H>동영상)

고 무의미한 동작이 발생하는 부분(c)에서 MHI 방식 경우 일부 오류가 발생하는 것을 볼 수 있었다. <H>동영상에서의 인식율은 제안하는 방법 경우 90.90%(FW=24, TP=2, SP=2), MHI방법 경우 81.82%로 제안하는 방법이 다소 양호함을 알 수 있었다. 특히 MHI 방법의 경우 동시에 다수의 동작이 발생하는 경우에 대한 해결 방안이 없으므로 다수 동작에 대한 비교가 이론상 불가능하여 본 논문에서 제안하는 방법과의 비교 실험 결과는 의미가 없는 것으로 판단된다.

단, 이와같은 실험은 FWS 간격 내에서 발생하는 상이한 동작의 중첩으로 나타나는 오류가 발생할 가능성이 있으므로, 그러한 오류를 배제하고 순수하게 동작의 인식율을 비교하기 위해 동작이 시작되는 시점을 수동으로 직접 지정하여 동작의 인식율을 상호 비교하는 추가 실험을 실시하였다. 그 결과 제안하는 방법의 경우 93.94%, MHI방법 경우 87.88%의 성능을 얻었다. MHI 방법에서 인식율은 개선이 되었으나, 제안하는 방법의 인식율이 여전히 더 높다는 것을 알 수 있다.

7. 결론 및 향후연구방향

본 논문에서는 동영상에서 다수 사람의 동작을 인식하고, 동작의 속도나 크기 변화에 강인하게 인식하는 방법을 제안하였다. 제안하는 방법론을 통해 기존의 동작 인식 방법들이 갖는 몇 가지 문제점을 해결하고 성능면에서 비교 가능할 정도의 결과를 얻었다.

첫째, 단일 객체의 동작에 대한 인식뿐 아니라 동작이 발생하는 영역을 검출하는 방식을 적용하여 다수의 객체에 대한 동작 인식을 가능하게 하였다.

둘째, 동작을 수행하는 주체 혹은 상황에 따라 동일한 동작임에도 다른 속도 혹은 다른 크기로 나타날 수 있는데, 이에 대해 시공간축 피라미드를 적용하여 속도, 크기 변화에 강인한 인식 결과를 얻을 수 있었다.

셋째, 동작의 표현을 모션 그래디언트 히스토그램을 적용함으로써 복잡한 인체 모델링이 불필요하고, 외형적/환경적 요소(실루엣, 색상, 텍스쳐, 조명, 배경 추출 등)의 영향에 강인하며, 계산의 복잡도를 최소화 할 수 있었다. 이는 인식을 위한 사전 학습 과정을 단순하게 할 수 있었으며, 원하는 어떤 동작이라도 단 시간 내에 학습시켜 인식 할 수 있다는 장점을 갖는다. 즉, 사람의 동작 뿐 아니라 일정한 동작 패턴을 갖는 객체의 어떠한 동작이라고 이론적으로 인식 가능한 것이다.

그러나 아직도 해결해야 할 문제점이 남아있다. 본 연구는 카메라의 이동 환경(팬, 틸트, 줌 등)을 고려하지 않았고, 동작의 객체가 전신이 모두 나타나는 것을 전제로 하였다는 제약 조건이 있어, 일반적인 동영상에서 모두 적용이 어렵다는 문제가 있으므로, 이동 카메라 환경에 대한 연구가 필요할 것이다. 또한 동작 인식의 단위(Frame Window Size)가 고정되어 있어 해당 단위 내에 두 가지 이상의 동작이 겹치는 경우(동작 전이 상태)로 인한 인식 오류의 가능성이 높으므로 보다 나은 인식율을 위해서는 이전의 일정 간격 마다 반복해서 인식을 수행하는 전향 슬라이딩 윈도우(Forward Sliding Window) 방식을 고려해 볼 필요가 있다고 본다. 그 외에 실시간 응용(사람의 동작을 컴퓨터 게임에 적용하는

등의 인터렉션 기법 응용)에 적용하기 위해서는 짧은 시간내에 동작을 구분할 정도의 특성을 갖도록 동작의 인식 단위(FWS)를 최소로 하는 특정 동작 위주로 인식이 이루어져야 하는데(8~12프레임), 이를 위해서는 인식하고자 하는 동작의 대상에 많은 제약이 따를 것으로 여겨진다. 또한 동작 선정의 문제 외에도 실시간 처리가 가능하기 위해 수행시간 단축을 위한 최적화 과정도 고려해야 할 것으로 판단된다.

참 고 문 헌

[1] I.Haritaoglu, D.Harwood, L.S.Davis, "W4:real-time surveillance of people and their activities," IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(8), 2000, pp. 809-830.

[2] Shearer, Bunke., Venkatesh, "Video indexing and similarity retrieval by largest common subgraph detection using decision trees," Pattern Recognition 34, 2001, pp. 1075-1091.

[3] Alex Pentland, "Looking at people: sensing for ubiquitous and wearable computing," IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(1), 2000, pp. 107-119.

[4] M.Yang, N.Ahuja, and M.Tabb, "Extraction of 2D motion trajectories and its application to hand gesture recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(8):pp. 1061-1074, 2002.

[5] Y.Yacoob and M.J.Black, "Parameterized modeling and recognition of activities," Journal of Computer Vision and Image Understanding 73(2):pp. 232-247, 1999.

[6] S.X.Ju, M.J.Black, and Y.Yacoob, "Cardboard people: A parameterized model of articulated image motion," In 2nd Int. Conf. On Automatic Face and Gesture Recognition, pp. 38-44, Oct. 1996.

[7] A.Bobick and J.Davis, "The recognition of human movement using temporal templates," IEEE Pattern Analysis and Machine Intelligence, 23(3):pp. 257-267, 2001.

[8] M.Blank, L,Gorelick, E,Shechtman, M.Irani and R. Basri, "Actions as Space-Time Shapes," IEEE International Conference on Computer Vision, pp. 1395-1402, 2005.

[9] L.Zelnik Manor and M.Irani, "Event-based analysis of video," IEEE Conference on Computer Vision and Pattern Recognition, Vol.2, pp. 123-130, 2001.

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

[11] E. Shechtman and M. Irani, "Space-Time Behavioral Correlation," IEEE Conference on Computer Vision and Pattern Recognition, Vol.1, pp. 405-412, 2005.

김 광 수

정보과학회논문지 : 소프트웨어 및 응용 제 34 권 제 10 호 참조

김 태 형

1995년 연세대학교 전기공학과 졸업. 1999년 공군 학사장교 중위제대. 2006년 연세대학교 컴퓨터과학과 석사 졸업. 1999년~현재 LG전자 MC연구소 재직 중. 관심분야는 영상처리, 컴퓨터비전 및 패턴인식

곽 수 영

정보과학회논문지 : 소프트웨어 및 응용 제 34 권 제 10 호 참조

변 혜 란

정보과학회논문지 : 소프트웨어 및 응용 제 34 권 제 10 호 참조