동영상에서 MGH을 이용한 실시간 다수 동작 인식

김태형° 변혜란
연세대학교
{erkth°, hrbyun}@cs.yonsei.ac.kr

Real-Time Multiple Action Recognition on Video using Motion Gradient Histogram

Taehyoung Kim°, Hyeran Byun
Department of Computer Science, Yonsei University

요 약

본 논문은 모션 그래디언트 히스토그램(Motion Gradient Histogram : 이하 'MGH')을 적용하여 동영상에서 나타나는 다수 객체들의 동작 검출 및 인식을 실시간으로 구현하는 방법을 제안한다. 인식하고자 하는 대상에 대한 기본적인 템플릿 동영상들의 MGH와 일정 프레임 간격마다 동영상의 MGH를 비교하여 검출 및 인식이 이루어진다. 동시에 다수의 동작이 있는 경우 동작이 발생하는 영역을 모션 에너지 영상(Motion Energy Image : MEI)기법으로 추출하여 해당 영역별 MGH를 구함으로써 다수 동작을 인식할 수 있도록 한다.

Keywords: 동작 인식, 모션 그래디언트 히스토그램, 모션 에너지 영상, 비디오 인덱싱

1. 서 론

동영상에 포함된 각종 움직임(Action)들의 효과적인 분석을 통한 비전 기반 인식 시스템은 비디오 인덱싱, 브라우징, 영상 분류(Clustering) 등의 기술 응용에 필수 요소이다. 특히 사람의 동작을 검출하고 인식하는 기술은 이동로봇, 감시 시스템, 인간과 로봇과의 상호작용 등 여러 응용에서 연구되고 있는 중요한 분야 중 하나이다.

현재의 비전 기술을 이용하여 실시간 사람의 동작을 인식하는 일은 쉽지 않음에도 불구하고, 컴퓨터 성능의 발달과 영상처리 기법의 다양한 시도들과 더불어 이에 관한 연구가 활발히 진행되고 있다.

사람의 동작 인식을 위한 주된 연구는 다음과 같이 크게 3가지로 구분해 볼 수 있다. Yacoob[1], Bregler[2]와 같이 2D,3D 인체 모델링을 통해 인체 각 부분의 상호 연관성을 분석함으로써 동작을 인식하는 방법이 있다.(Modeling-based) 이러한 모델링 기반 방식은 인식 정밀도가 높다는 장점이 있는 반면 구현 난이도가 높고, 실험 환경의 제약 조건 및 실시간 처리가 어려워 특수한 응용에 제한되는 단점이 있다. Bobick[3]이 제안한 축적된 움직임 정보(Motion History)를 이용한 방법과 Blank[4]처럼 실루엣을 이용한 외형 기반의 동작 인식 방법이 있다.(Appearance-based) 이는 시간축상 움직임의 정보를 하나의 이미지(Motion History Image, Space-time shape)로 만들어 효과적으로 표현함으로써 복잡한 모델링 과정 없이 좋은 인식 결과를 얻을 수 있는 장점이 있으나, 인식 기준이 되는 템플릿과 인식 대상 동영상 간의 높은 시공간적 제약성을 가지며 실루엣 등의 외형적 의존도가 높으므로 주변 환경에 따라 영향을 받는 한계가 있다. 마지막으로는 광류(Optical Flow)를 적용한 Efros[5]와 같이 움직임 자체에 대한 영상 분석으로 동작을 인식하는 방법이 있다.(Motion-based) 이는 광류의 정밀도에 의존하게 되어 인식률이 변화할 가능성이 높고, 계산량이 많다는 단점이 있다.

본 논문에서는 위 방법들에 대한 단점을 보완하고자 Zelnik[6]이 제안한 모션 그래디언트 히스토그램을 이용하여 동작 인식을 구현하였으며, 특히 기존 연구들이 주로 하나의 객체 동작 인식에 중점을 두었으나 본 논문에서는 다수 객체의 동작을 분할 검출하여 각각의 인식이 가능한 방법을 제안한다.

2. 동작 특징 추출

동영상에서 객체의 동작을 인식하고자 할 경우 시공간적 조건이 고려된 특징으로 해당 동작을 효과적으로 나타낼 수 있어야 한다. 모션 그래디언트 히스토그램은 움직임이 포함된 화소값의 변화량을 통계적 접근을 통해 객체의 공간적 이동 변화, 외형의 변화, 주변 환경의 변화 등에 강인한 동작의 특징을 나타낼 수 있는 방법이다. 또한 동일 동작의 속도 차이를 극복하기 위해 시간축 피라미드 동영상을 생성하여 문제를 해결한다.

2.1 모션 그래디언트 히스토그램

일반적인 시공간상 그래디언트(Space-Time Gradient)는 각 시공간 점(Space-Time Point) (x,y,t)에서 동영상 S에 대하여 아래 식(1)과 같이 추정할 수 있다.

$$(S_x,S_y,S_t) = \begin{cases} \left(\frac{dS}{dx}, \frac{dS}{dy}, \frac{dS}{dt}\right), & \frac{dS}{dt} \geq \delta \\ (0,0,0), & \frac{dS}{dt} < \delta \end{cases} \quad (1)$$

$\frac{dS}{dt}$의 값이 특정 임계값(δ)보다 큰 경우에만 주된 움직임이 발생하는 것으로 판단한다. 각 값들의 크기는 움직이는 객체의 외관상 특성(조명, 옷의 질감, 색깔 등)에

따라 크게 영향 받을 수 있으므로 이러한 영향을 제거하기 위해 아래 식(2)에 의해 크기가 1로 정규화된 그래디언트를 구성하는데 이를 모션 그래디언트로 정의한다. 또한 각 값의 절대값을 이용함으로써 움직임의 방향 변화에 무관한 특성을 가질 수 있다.

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

식(2)에 의해 구해진 (N_x,N_y,N_t)은 일정 프레임 윈도우 내에서 각각 $\{h_x, h_y, h_t\}$의 히스토그램으로 표현할 수 있고 이를 모션 그래디언트 히스토그램(MGH)이라 한다. 즉, $\{h_x, h_y, h_t\}$ 이러한 3개의 1차원 히스토그램 정보로 동영상내에서 일정한 프레임 윈도우의 동작 특징을 추출할 수 있는 것이다. 그림 1은 걷기(W1,W2), 팔벌려 뛰기(Jack), 점프(Jump) 동영상에 대한 각각의 h_t를 계산한 결과를 나타내고 있다.

(a) W1 (b) W2 (c) Jack (d) Jump

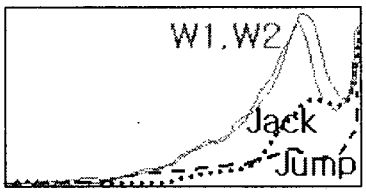


그림 1. 일부 동영상의 MGH(h_t) 결과

보는 바와 같이 서로 다른 동작에 대해 각 MGH가 차이가 나는 것을 알 수 있으며, 특히 (a),(b)는 서로 다른 배경에서 다른 사람이 반대 방향으로 걷고 있더라도 동일 동작으로 판단 할 수준의 유사한 MGH가 나타남을 알 수 있다.

2.2 시간축 피라미드(Temporal Pyramid)

동일한 걷는 동작이라고 하더라도 사람마다 그 속도가 조금씩 차이가 나게 된다. 시간축상으로 발생하는 이러한 속도의 차이점을 보완하기 위해 아래 식(3)과 같이 시간축 피라미드 형태의 동영상을 추가로 구성한다. 일정 프레임 윈도우 동영상 S에 대해, 시간축 L-피라미드 집합은

$$TP^L(S) = \{S^1(=S), S^2, S^3, ..S^L\} \quad (3)$$

(S^{l-1}프레임수 = 2xS^l프레임수)

와 같이 정의된다. 즉, 프레임 윈도우 32(FW=32)이고 3-피라미드(TP=3)인 경우, 동영상 S에 대해 각 프레임수가 32,16,8인 $\{S^1,S^2,S^3\}$의 $TP^3(S)$를 얻게 된다.

3. 동작 인식

동영상내에 발생하는 다수 객체의 동작을 인식하기 위해서는 움직임이 발생하는 각각의 영역을 검출해 내고, 각 영역의 움직임에 대해 사전 준비된 인식하고자 하는 템플릿 동영상과의 유사도를 판단하여 어떤 동작인지를 결정하게 된다.

3.1 움직임 검출

아래 식(4)로 구해지는 이진 축적 영상인 모션 에너지 영상(Bobick[2])을 통해 일정 프레임 윈도우에서 발생하는 움직임의 영역을 구할 수 있다.

$$MEI_\tau(x,y,t) = \bigcup_{i=0}^{\tau-1} D(x,y,t-i) \quad (4)$$

$D(x,y,t)$: 프레임간 이진 차영상

τ : 프레임 윈도우 크기

3.2 동영상간 유사도 판단

두 동영상S_1, S_2간의 움직임에 대한 유사도(SM)를 판단하기 위해서, 식(5)의 카이제곱 계산법으로 각각의 MGH의 유사도를 비교하게 된다.

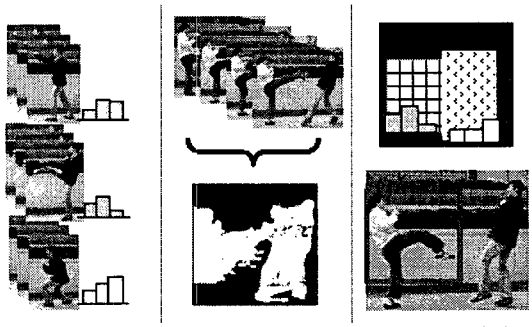
$$SM(S_1,S_2) = \sqrt{\sum_{k,i} \frac{[h_{1k}(i) - h_{2k}(i)]^2}{h_{1k}(i) + h_{2k}(i)}} \quad (5)$$

$k \in \{x,y,t\}$, $i = 1, \ldots,$ 히스토그램 bin 수

구해진 값이 0에 가까울수록 두 동영상내 동작이 유사하다고 판단 가능하다.

4. 실험 및 결과

전반적인 시스템 구성은 아래 그림 2와 같이 도식적으로 표현할 수 있다.



(a)템플릿 동영상 (b)움직임 검출 (c)동작 인식

그림 2. 시스템 구성도

(a) 인식하고자 하는 각 동작에 대한 템플릿 동영상의 MGH를 각각 구한다.

(b) 실제 동영상에서 일정한 프레임 윈도우마다 MEI를 이용하여 움직임을 검출한다.

(c) 검출된 영역 각각에 해당하는 MGH를 계산하여 템플릿내의 가장 유사한 MGH를 찾아 어떤 동작인지 판단한다.

4.1 실험 환경

주간 실외 환경에서 일반 디지털 카메라(후지 파인픽

스 F-11)로 촬영한 동영상과(320x240 24fps으로 변환하여 사용), Blank[3]에서 사용한 템플릿 동영상(180x140 25fps)을 사용하여 P4-1.8GHz Windows2000 환경에서 VC++6.0로 구현하여 실험하였다.

4.2 단일 동작 인식 결과

표 1은 Blank[3]에서 사용한 템플릿 동영상을 편집하여 만든 걷기(A), 달리기(B), 팔벌려 뛰기(C), 물건줍기(D), 제자리 뛰기(E) 등 5가지 동작의 동영상을 이용한 실험 결과를 나타낸다. 각 테스트 동영상은 7명의 서로 다른 사람들의 동일한 동작을 연속적으로 편집하여 만든 데이터이며 템플릿 동영상(T(X))은 한 사람씩 5가지 동작의 템플릿을 이용하여 7번 반복 실험하였다. (TP=2, FW=32 기준)

표 1. 단일 동작 인식 결과

		템플릿 동영상 인식(%)					NG (%)	처리속도 (fps)	총프레임수
		T(A)	T(B)	T(C)	T(D)	T(E)			
테스트 동영상	A	79.3	16.1	-	-	-	4.6	12.10	710
	B	4.3	91.8	-	-	-	3.9	11.23	435
	C	-	-	93.0	2.5	2.8	1.7	13.91	520
	D	-	-	10.3	73.5	8.8	7.4	12.72	639
	E	-	-	9.0	4.8	80.7	5.5	14.28	756

달리기(B)와 팔벌려 뛰기(C) 동작은 확연한 동작의 특성상 인식율이 높았고, 걷기(A) 경우 시간축 피라미드의 적용으로 인해 달리기와 혼동된 결과를 가져오는 경우가 있었다. 실험상에서 실제 동작이 일어나고 있음에도 불구하고 유사 동작을 판단하지 못하는 경우(NG)가 발생되는 원인은 서로 다른 사람의 동작 간 서로 전이가 발생 할 때 해당 MGH의 값이 템플릿의 MGH와 너무 낮은 유사도를 갖기 때문에 발생하는 것으로 판단된다. 전반적인 처리속도는 초당 10프레임 이상으로 실시간 응용에 양호한 수준임을 알 수 있다.

4.3 다수 동작 인식 결과

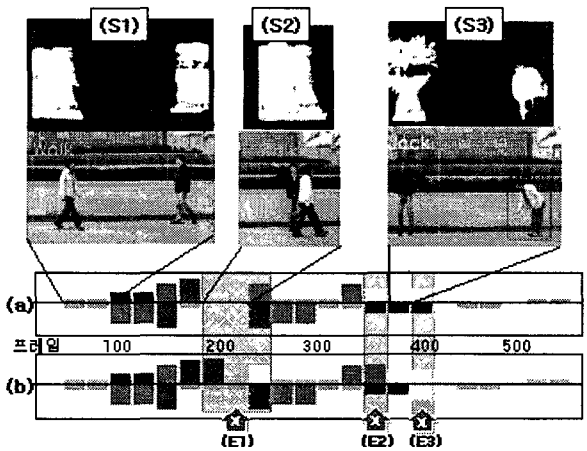


그림 3. 다수 동작 인식 실험 결과

위 그림 3은 두 사람이 동시에 등장하는 동영상에서의 아래 그림 4의 5가지 동작(걷기:Walk, 물건줍기:Bend, 팔벌려 뛰기:Jack, 제자리뛰기:Jump, 팔흔들기:Wave)에 대한 인식 결과를 나타내고 있다.(총570프레임, TP=2, FW=24)

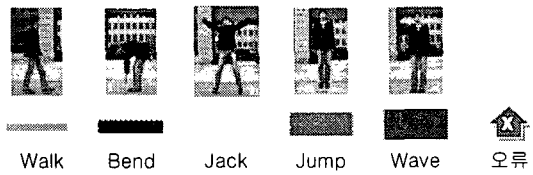


그림 4. 템플릿 동작 영상 및 범례

(a)는 제안한 방법에 의해 처리한 결과이고 (b)는 실제 발생한 동작으로써 두 그래프를 비교해 볼 때 전반적으로 (S1),(S3)등과 같이 양호한 결과를 보였다. 오류 (E1)는 (S2)에서처럼 상호 가려짐으로 MEI상에서 동작의 분할 검출이 어려워 발생하였으며, (E2),(E3)는 전후 동작 간 전환 시 MGH의 이전 정보값의 축적으로 오류가 발생한 것으로 판단된다.

5. 결 론

본 논문에서 제안한 방법을 통해 복잡한 모델링 및 추가적인 학습 과정 없이도 다수 객체의 동작을 실시간적으로 양호하게 인식함을 알 수 있었다. 일부 가려짐에 의한 동작 검출 오류와 동작 전이 시 발생하는 오류에 대해서는 프레임 윈도우 슬라이딩과 같은 보완이 필요할 것으로 여겨지며, 외형상 전혀 다른 동작에 대해 유사함이 나타나는 문제를 해결하기 위해서는 외형적 특성을 고려한 유사도 비교를 추가해야 할 것으로 여겨진다. 향후 이러한 개선이 이루어진다면 보다 나은 성능을 얻을 것으로 기대된다.

감사의 글

본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음, IITA-2005-(C1090-0501-0019).

참고 문헌

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

[2] C.Bregler, "Learning and recognizing human dy-namics in video sequences", CVPR, June 1997

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

[4] M.Blank, L,Gorelick, E,Shechtman, M.Irani and R. Basri,"Actions as Space-Time Shapes", ICCV pp.1395-1402, 2005

[5] A.Efros, A.Berg, G.Mori and J.Malik,"Recognizing action at a distance",ICCV Vol.II pp.726-733, 2003

[6] L.Zelnik Manor ,and M.Irani, "Event-based analy-sis of video", CVPR Vol.II pp.123-130, Hawaii, 2001