# 비디오기반 사람의 모션 검출

이창수。, 박연출, 박세준, 오해석
숭실대학교 컴퓨터공학과 멀티미디어 연구실
e-mail:hacker@multi.soongsil.ac.kr

# Video Based Human Motion Detection

Chang-Soo Lee, yeon-chool Park, Sae-Joon Park, Hae-Seok Oh
Dept. of Computer Science, Soongsil University

요약

비디오 기반 사람의 모션 캡쳐에 관한 연구는 최근 몇 년 동안 컴퓨터 비전분야에서 폭넓은 연구가 진행되어지고 있다. 본 논문은 비디오 기반으로 사람의 모션을 전체 프레임이 진행되는 동안 프레임 별로 디스플레이 한다. 첫 프레임에서 비디오 세그멘테이션 과정에서 샷을 검출하고 이를 이용하여 객체를 분류한다. 분류된 객체에서 사람의 영역을 추출한다. 추출된 영역은 다음 프레임의 위치를 예측하게 된다.

## 1. 서 론

최근 비디오의 보급과 오디오등 많은 발전을 이룩하면서 실생활에 있어서 많은 부분에 활용이 되고 있다. 이러한 요구에 따라 비디오 분야에 대한 연구가 활발히 이루어지고 있다.

비디오를 기반으로 움직이는 물체를 추적하거나 비디오에서 특정 인물을 찾아내는 것 등 많은 연구가 진행되어 왔다.

기존에 연구된 대표적인 비디오 기반 사람의 모션 연구는 다음과 같다. Gravrila와 Davis는 사람의 몸 전체의 움직임을 측정하였고[1], Chen과 Lee는 17개의 라인 세그멘테이션과 사람의 뼈대 모델을 구성하기 위한 14개의 관절을 사용하여 사람의 움직임을 표현하였다[2].

기존의 연구들의 접근방법에는 다음과 같은 3가지 과정을 통해 모션을 캡쳐 하게 된다.

첫째, 비디오 프레임에서 정확한 객체를 찾아내고 둘째는, 프레임사이에서 일치하는 객체를 찾아낸다. 마지막으로 일치하는 객체로부터 3D 모션을 다시 구성하게 된다. 또는 이미지에 체적측정 모델을 삽입하여 일치하는 부분에 대해 사람의 움직임을 측정하게 된다.

비디오는 크게 프레임(Frame), 샷(Shot), 장면(Scene), 시이퀀스(Sequence)의 4가지의 계층적이며 순차적인 데이터의 모임이다. 여기서 프레임이란 비디오를 구성하고 있는 계층 중 최하위 계층이며 하나의 정지영상이다. 프레임의 상위계층인 샷은 하나의 카메라로 기록된 연속적인 프레임의 모임이며, 장면은 주제가 같은 내용을 가진 인접한 샷의 모임이며, 시이퀀스는 비디오의 최상위계층이며 연관된 장면의 모임이다. 그림 1은 이러한 비디오의 구성을 보여준다.

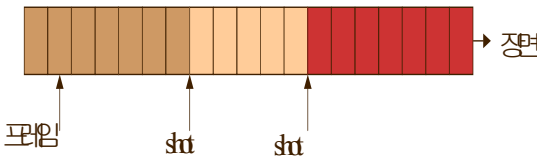


그림 1 비디오 데이터의 구성

본 논문에서는 두 가지 방법으로 접근을 하게 된다. 첫 번째 단계인 각각의 프레임을 분석해 샷을 검출하는 비디오 세그먼테이션(Segmentation)과정과 두 번째 단계로 샷의 특징을 이용하여 사람의 객체

를 얻어낸다.

본 논문에서는 샷을 검출하여 얻어지는 값을 이용하여 팔의 객체를 찾아내고 비디오 프레임 진행동안의 팔의 모션을 디스플레이 한다.

샷을 검출하기 위한 기존의 제안된 연구방법으로는 공간영역의 특징인 각 픽셀의 차를 이용한 방법[4], 픽셀값의 히스토그램의 차이법[5], 분산과 표준편차 등의 통계적인 차를 이용한 방법[6] 등에 관한 연구들이 있다.

위의 제안된 방법을 사용하여 사람의 객체를 찾아내고 프레임마다 사람의 움직이는 위치, 모션을 디스플레이 한다.

본 논문에서는 좀더 효율적인 사람의 모션을 검출하기 위해 간단한 공간영역의 특징인 칼라 히스토그램방법을 각 프레임마다 적용시키고, 픽셀 영역의 확률가정을 통해 정확한 사람의 위치를 찾아낸다.

## 2. 비디오 세그먼테이션 (Video Segmentation)

### 2.1. 픽셀 간의 차를 이용한 방법

인접한 프레임에 대응하는 픽셀간 차를 계산함으로써 두 프레임간의 샷 전환을 감지하는 방법으로 임계치(Threshold)를 계산해 계산된 임계치를 초과하는 프레임을 컷(샷 경계)으로 판단하는 방법이다

$$DP_i(x,y) = |F_i(x,y) - F_{i+1}(x,y)| \qquad (1)$$

$$\frac{\sum_{x,y=1}^{X,Y} DP_i(x,y)}{X \times Y} > T \qquad (2)$$

식(1)에서 $F_i(x,y)$는 $i$번째 프레임 $(x,y)$의 밝기 값을 나타내며, $DP_i(x,y)$는 인접하는 프레임의 픽셀차를 의미한다. 식(2)의 X×Y는 프레임의 크기를 나타내며, T는 컷을 판별하기 위한 임계치이다. 이 방법은 구현이 아주 간단한 반면에 카메라 움직임에 민감하며, 영상 각각의 픽셀에 대해 비교를 해야하기 때문에 시간이 많이 걸린다는 단점이 있다.

### 2.2 히스토그램의 차이를 이용한 방법

이 방법은 가장 일반적인 방법으로 연속되고 인접한 프레임간의 히스토그램 차이를 계산해 주어진 임계치와 비교해 임계치를 초과할 경우 샷으로 판단하는 방법이다. 히스토그램의 차를 이용한 방법은 카메라의 움직임에 강인하며, 연산속도가 빠르다는 장점이 있다[5].

$$\sum_{j=1}^{G} |H_i(j) - H_{i+1}(j)| > T \qquad (3)$$

식(3)의 G는 칼라 레벨의 수이며, $H_i$은 $i$번째 프레임의 히스토그램을 의미한다.

### 2.3 통계적인 차를 이용한 방법

인접하는 프레임간의 통계적인 차를 이용한다. 즉 두 프레임간의 평균, 분산, 표준편차 등을 구한 후 임계치를 설정해 샷을 검출하는 방법이다[6].

$$\frac{[(\frac{\sigma_i + \sigma_{i+1}}{2}) + (\frac{\mu_i - \mu_{i+1}}{2})^2]^2}{\sigma_i \times \sigma_{i+1}} > T \qquad (4)$$

식(4)의 $\sigma_i$는 $i$번째 프레임의 분산이며, $\mu_i$는 프레임의 평균을 의미한다.

`

### 2.4 변환계수의 차를 이용하는 방법

FFT, Wavelet Transform 그리고 DCT와 같은 변환영역의 계수들을 이용하여 샷 경계를 검출하는 방법들이다. 특히, 최근 들어 MPEG 동영상과 같은 압축된 영상에서 압축을 해제할 필요 없이 비디오 데이터로부터 DC계수를 사용하여 샷 경계를 검출하는 방법이 연구 중에 있다. 이 방법은 프레임을 다수의 서버블록으로 나눠 각각의 DC계수를 추출해 추출된 DC계수들을 인접한 프레임과 비교함으로써 샷 경계를 검출하는 방법이다.

## 3. 샷의 특징을 이용한 팔의 객체 분류

샷의 경계가 되는 장면 전환 기법은 크게 두 가지로 나눌 수 있다. 그 중 하나는 급진적인 전환인데 이런 급진적인 전환의 컷 검출은 컷으로 판별되는 프레임과 인접하는 프레임과의 차이가 비교적 크기 때문에 샷을 나누기가 쉽다. 그러나 점진적인 전환은 영상의 시각적인 효과를 위한 카메라 조작으로 장면전환이 여러 프레임에 걸쳐 이루어지므로 컷 검출이 매우 어렵다. 이런 점진적인 전환은 카메라의

조작에 따라서 페이드(Fade), 디졸브(Dissolve), 와이프(Wipe)등이 있다.

일반적으로 카메라의 특수효과로 인해 컷 검출시 잘못된 컷을 감지하게 되는데 이러한 잘못된 컷의 감지를 보완하고, 급진적인 장면전환에 대해서도 좀 더 정확한 컷을 검출하기 위해 본 논문에서는 기존에 연구되어진 칼라 히스토그램과 픽셀영역의 확률 가정을 이용한다[5,8]

### 3.1 칼라 히스토그램의 차이를 이용한 방법

공간 영역에서 서로 인접하는 프레임간에 칼라 채널별로 히스토그램을 구한 후 각각 채널별로 합을 구한다. RGB전체 채널의 평균값으로 연속되는 프레임간에 평균값 차를 이용함으로써 컷을 검출한다. 이 특징값은 컷이 있는 프레임과의 차는 커지고 그렇지 않는 곳에서는 차가 적다[5,7].

$$\frac{\sum_{j=1}^{G}(|R_i(j)-R_{i+1}(j)|+|G_i(j)-G_{i+1}(j)|+|B_i(j)-B_{i+1}(j)|)}{X\times Y} > T \quad (5)$$

식(5)에서의 G는 칼라 레벨의 수이며, X×Y는 프레임의 전체크기, $R_i$, $G_i$, $B_i$는 $i$번째 프레임 각각의 RGB 채널별 픽셀값이다.

### 3.2 픽셀영역의 확률가정

칼라 히스토그램의 차이를 이용하여 연속되는 프레임간의 컷을 검출하고 사람의 모션을 가정, 그리고 반복되는 확률적 방법을 통한 사람의 모션 검출을 통하여 밀집한 모션 영역을 측정한다[8].

순차적인 이미지를 $I^{(1)},I^{(2)},...,I^{(t)}$,라하고 세그먼트 객체를 Θ 라고 가정한다.

$$\Theta = \arg max_{\Theta} P(I^{(1)},I^{(2)},...,I^{(t)}|\Theta) \quad (6)$$

$$P(I^{(1)},I^{(2)},...,I^{(t)}|\Theta) \propto P(\Theta|I^{(1)},I^{(2)},...,I^{(t)}) \cdot P_{jointed}(\Theta) \cdot P_{HMM}(\Theta) \quad (7)$$

Θ와 객체는 $\Theta^{(1)},\Theta^{(2)},...,\Theta^{(t)}$을 기본으로 이미지 프레임은 순차적인 흐름을 갖는다. 각 이미지 프레임은 단일 영역의 고정된 수를 포함하는 $\Theta_1^{(1)},\Theta_2^{(2)},...,\Theta_k^{(t)}$ 와 배경모델인 $\Theta_0^{(t)}$를 기본으로 한다. $\Theta_k^{(t)}$를 기본으로 하는 영역은 사람의 영역이거나 모션의 관절영역이 된다.

사람의 세그멘트나 관절의 가정인 $\Theta_k$는 세 가지 피라미터의 범주를 가진다.

1) Gaussian Blob Model - 크기 $\mu_k$는 2D의 중앙부이고, $\Sigma_k$는 2×2 공분산 메트릭스
2) Motion parameters - $R_k$는 2×2 회전 메트릭스, $d_k$는 2D전송 벡터(관절은 회전 메트릭스를 가지지 않는다), $\sigma_k$ 는 픽셀 변수값
3) a Score coefficient - 가우시언 혼합을 위한 혼합계수

배경모델은 이미지 $[R_0,d_0,\sigma_0]$에서 가장 두두러진 모션을 위한 파라미터를 포함한다. 그림 2는 이러한 가정을 나타내고 있다.

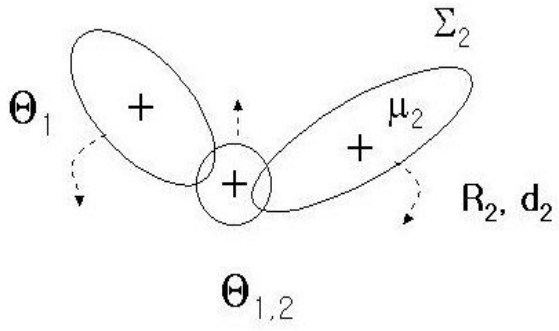


그림 2 Two connected arms segments with the three corresponding blob models

이미지 프레임이 $I^{(t-1)}$이고 예상되는 객체를 $\Theta^{(t)}=[\Theta_1,...,\Theta_k]$ 이다. 다음 이미지 프레임 $I^{(t)}$ 을 위한 이미지 확률 모델을 얻게 된다. 이러한 것은 Gaussian blob model의 픽셀 값을 예측하기 위해서는 전 단계의 프레임과의 일치하면 예측을 하게 된다.

$$P(I^{(t)}|\Theta^{(t)},I^{(t-1)}) = \prod_{X,Y} P(I^{(t)}(x,y),x,y|\Theta^{(t)},I^{(t-1)}) \quad (8)$$

$$P(I^{(t)}(x,y),x,y|\Theta^{(t)},I^{(t-1)}) = \sum_{k=0}^{k}(\omega_k \cdot P(x,y|\mu_k,\Sigma_k) \cdot P(I^{(t)}(x,y)|x,y,R_k,d_k,\sigma_k)) \quad (9)$$

모든 스코어 $\omega_k$는 1,2,3,.....로 순차적인 오름 값을 가지고, $P(x,y|\mu_k,\Sigma_k)$ 는 $k'th$ 번째 Blob 모델의 Gaussian 분산이다.

$P(I^{(t)}(x,y)|x,y,R_k,d_k,\sigma_k)$는 변수 $\sigma_k$ 와 밀접한 이미지 일치값을 가지는 이미지의 위치 $(x,y)$ 에 픽셀 값을 가지는 표준 분산이다.

### 3.3 칼라 히스토그램의 특징값과 픽셀영역의 확률값에 의한 예측

일반적으로 비디오의 샷과 샷사이의 프레임의 개수는 3 프레임을 초과한다. 그러므로 먼저 느슨한 임계치로 추출된 칼라 히스토그램의 특징값과 픽셀영역의 확률값을 적용하여 동일한 샷 경계를 나타낼 때 추출된 샷 경계는 정확한 샷 경계라고 할 수 있다. 즉, 이러한 방법으로 추출된 샷 경계에서 전후 연속되는 3 프레임까지는 또 다른 샷 경계가 없다고 판단한다. 그 이유는 카메라 조작이나 비디오에서 정지영상을  추출시 또는 일반적으로 포함될 수 있는 잡음 때문에 같은 샷이지만 픽셀의 밝기값 등의 심한 변화를 일으킬 수 있으므로 이런 변화에 좀 더 강인성을 갖기 위해서이다.

Gaussian 모델에서 한 프레임에서 가지는 확률값과 프레임에서 적용되는 히스토그램의 특징값을 사용하여 다음 프레임을 예측하고, 프레임에서 객체의 명확성을 위하여 전 프레임의 특징값과 비교를 하게 된다.

## 4. 실험 결과

본 논문에서 사용된 데이터는 AVI형식의 동영상 파일이며 펜티엄 350Mhz PC를 사용하였고, Visual c++ 6.0을 사용하여 프로그램하고 동영상데이터에서 프레임별로 저장하여 디스플레이 하였다. 그림 3에서는 프레임별 세그멘테이션 결과를 나타낸 것이고, 그림 4에서는 본 영상의 프레임에 Gaussian 모델을 기초로 하여 중심점을 표시한 것이다.

그림 3 프레임별 세그멘테이션

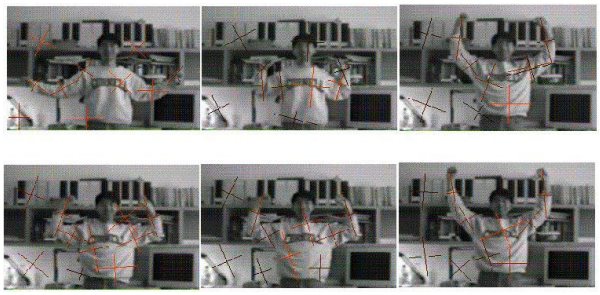

그림 4 프레임에서 센터를 찾은 결과

## 5. 결론 및 향후 연구과제

본 논문에서 제안한 방법으로 칼라 히스토그램의 특징값 적용 방법과 픽셀영역의 확률값에 의한 예측값을 이용하여 비디오 프레임에서 객체의 분류를 향상시키고 있다.

위의 그림 4의 결과는 사람의 객체뿐만 아니라 움직임이 있는 객체를 인식을 하고 있다.

향후 과제로는 비디오 내에서 검출되는 특징값을 이용하여 3D객체를 모델링하고 객체의 움직임을 연결하는 연구하는 것이다.

## [참고 문헌]


[1] D.M Gavrila and L.S Davis, "Towards 3-d model-based tracking and recognition of human movement : a multi-view approach", In Proc, OfInt, Workshop on Automatic Face- and Gesture-Recorgnition, Zurich, 1995.

[2] Z, Chen and H,J.Lee, "Knowlege-guided visual perception of 3D human gait from a single image sequence", IEEE Trans, On Systems, Man, and Cybernetics, 22(2), pp.336-342, 1992.

[3] Michael Philips, Wayne Wolf, "Video Segmentation Techniques for News," Proc. of SPIE Vol. 2916, pp. 243-251, 1996.
[4] J. S. Boreczky and L. A. Rowe, "Comparison of video shot boundary detection techniques," SPIE Storage and Retrieval for Image and Video Databases, Vol. 2670, pp.170-179, 1994.
[5] N. V. Patel and I. K. Sethi, "Video Shot Detection and Characterization for Video Databases," Pattern Recognition: Special issue on multimedia, 1996.
[6] I. K. Sethi and N. Patel, "A Statistical Approach to Scene Change Detection," SPIE Storage and Retrieval for Image and Video Databases, Vol. 2420, 1995.
[7] 나재형, “내용기반 검색을 위한 뉴스비디오에서 자막추출”, 숭실대학교 석사논문, 1997.12.
[8] Christoph Bregler, "Probabilistic of Human Actions", May 28, 1996.