불변 모멘트를 이용한 DSTW 기반의 동적 손동작 인식 방법

*지재영 *장경현 **박기태 *문영식

*한양대학교 컴퓨터공학과 , **한양대학교 BK21 사업단

{ *jyji , *khjang }@visionlab.or.kr , { **parkkt , *ysmoon }@cse.hanyang.ac.kr

Recognition of Dynamic Hand Gestures based on DSTW using Invariant Moments

*Ji, Jae Young *Jang, Kyung Hyun **Park, Ki Tae *Moon, Young Shik

*Dept. of Computer Science and Engineering , Hanyang University

**Institute of Hanyang BK21 , Hanyang University

요약

본 논문에서는 Dynamic Space Time Warping(DSTW) 알고리즘을 이용하여 손동작을 다양한 배경에서도 정확하게 인식할 수 있는 방법을 제안한다. DSTW 알고리즘을 이용한 기존의 손동작 인식 방법은 질의영상의 매 프레임 마다 검출된 다수의 손 후보영역을 사용하여 모델영상과 시간 축 상으로 비교하는 방법이다. 그러나 기존의 DSTW 알고리즘을 이용한 손동작 인식 방법은 손을 포함하지 않은 후보영역들(배경, 팔꿈치 등)에 의해 오인식될 수 있는 경로를 생성하며, 그 결과로 사용자가 의도하지 않은 손동작으로 인식될 수 있다. 이러한 단점을 해결하기 위해서, 본 논문에서는 손 후보영역의 불변 모멘트를 이용하여 질감 정보를 추출한 후 후보영역들 사이의 유사도를 비교하였다. 제안한 방법은 유사도를 모델과 질의의 매칭비용에 가중치로 적용하였고, 다양한 실험 결과 제안한 방법이 기존의 방법에 비해 사용자의 손동작을 정확하게 인식하는 것을 확인하였다.

1. 서론

인간은 일상생활에서 말이나 문자와 같은 언어적 수단 뿐 만 아니라 표정, 제스처와 같은 비 언어적 수단을 이용하여 상대방과 의사소통을 한다. 그러나 서로간의 인터페이스가 다르거나 부자연스러울 경우, 문제가 발생하게 된다. 따라서 사람과 컴퓨터간의 보다 효과적인 상호작용을 하기 위해서는 두 개체 간의 의사를 잘 이해할 수 있는 편리하고 자연스러운 인터페이스가 요구된다. 자연스러운 상호작용에 있어서 제스처 기반 사용자 인터페이스는 다른 인터페이스(음성, 촉감, 시점 등)에 비해 비교적 직관적이고 간단하다. 그러나 부가적인 장치(예, 데이터 글로브, 움직임 추적 장치)를 사용해야 하는 불편함이 있다. 따라서 부가적인 장치의 사용에 따른 불편을 해소하고 사용자의 자연스러운 움직임을 보장하기 위해, 컴퓨터 비전(Computer Vision) 기술을 이용한 제스처기반 사용자 인터페이스에 대해 연구가 활발히 진행되고 있다[1]. 일반적으로 컴퓨터 비전 기술을 이용한 사용자 인터페이스는 초기화(Initialization), 추적(Tracking), 포즈 예측(Pose Estimation), 그리고 인식(Recognition) 과정을 거쳐 획득된 영상으로부터 제스처를 인식한다. 그러나 추적 기반의 제스처 인식 기법은 갑작스런 움직임, 가려짐, 복잡한 배경이 존재하는 경우, 객체를 정확하게 추적하기 어렵다. 특히, 색상 기반 손동작 인식 방법의 대부분은 피부색과 유사한 색상의 객체가 존재하거나 또는 배경이 피부색과 유사한 경우 손동작을 정확하게 인식하기 어렵다는 단점이 있다.

Dynamic Time Warping(DTW) 알고리즘은 시간 축 상에서 비선형 신장과 축소를 허용함으로서 서로 다른 길이의 두 시계열 패턴을 매칭할 수 있는 알고리즘이다[2]. Alon[3] 등은 DTW를 공간 영역으로 확장한 Dynamic Space Time Warping(DSTW) 알고리즘을 이용하여 손동작 인식 방법을 제안하였다. Alon은 추적 기반 제스처 인식 기법의 단점을 보완하기 위해 질의영상의 매 프레임마다 다수의 손 후보영역을 검출하였다. 그리고 DSTW 알고리즘을 사용하여 모델동작과 질의동작 사이의 매칭 경로를 계산하였다. 그러나 이 방법은 손을 포함하지 않은 후보영역들(배경, 팔꿈치 등)에 의해 오인식될 수 있는 경로를 생성하며, 그 결과로 다른 모델로 인식될 수 있는 단점을 가지고 있다. 이를 해결하기 위해 본 논문에서는 손 후보영역의 불변 모멘트를 이용하여 질감 정보를 추출한 후 후보영역들 사이의 유사도를 비교하였다. 비교된 유사도는 모델과 질의의 매칭비용에 가중치로 적용하였다.

본 논문의 구성은 2장에서 제안한 방법, 후보영역 검출, 특징 추출, 유사도 비교, 그리고 제안한 DSTW 알고리즘에 대해 설명하고, 3장에서 기존의 방법과 제안한 방법에 대하여 각각 성능을 평가하였다. 마지막으로 4장에서 결론 및 향후 연구 방향을 제시하였다.

2. 제안한 방법

본 논문에서 제안한 손동작 인식 방법은 크게 손 후보영역 검출과정, 특징추출 과정, 유사도 비교 과정, 그리고 모델·질의 매칭 과정으로 구성된다. 손 후보영역 검출 과정은 색상(Color)과 움직임(Motion) 정보를 이용하여 K개의 손 후보영역을 검출하는 과정이다. 특징 추출 과정은 검출된 후보영역의 2가지 특징(위치, 속도)을 추출하는 과정이다. 유사도 비교 과정은 첫 프레임의 각 손 후보영역을 기준으로 다른 프레임의 손 후보영역들과 유사도를 비교하는 과정이다. 마지막으로 모델·질의 매칭 과정은 DSTW 알고리즘을 사용하여 최적의 매칭 경로와 비용을 계산하는 과정이다. 그림 1은 제안한 방법의 전체 흐름도를 보여준다.

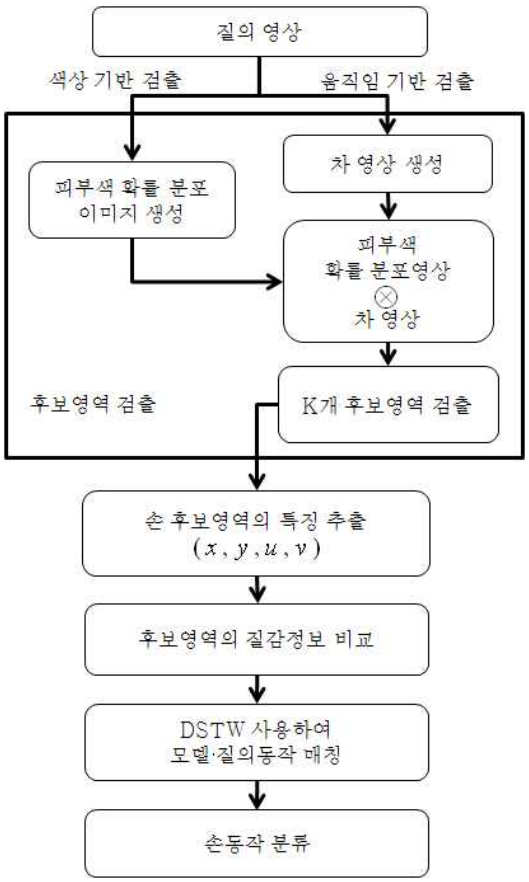
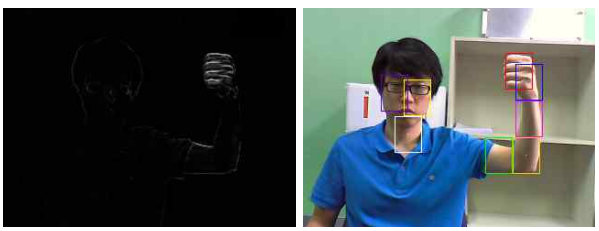


그림 1. 전체 흐름도

(a) 입력 프레임 (b) 피부색 확률 분포 영상



(c) 차 영상 (d) 손 후보 영역

그림 2. 손 후보영역 검출 ($K=8$)

(a) 손 (b) 부분 손 (c) 얼굴

(d) 팔뚝 (e) 배경

그림 3. 후보영역 예

2.1 손 후보영역 검출 및 특징추출

가. 손 후보영역 검출

본 절에서는 질의영상에서 K개의 손 후보영역을 검출하는 방법에 대해 설명한다. 본 논문에서는 비교적 간단하면서도 효과적인 결과를 보여주는 색상과 움직임 정보를 이용하여 손 후보영역을 검출한다. 우선 피부색 히스토그램[4]을 사용하여 질의영상의 매 프레임마다 피부색 확률 분포 영상을 생성한다. 그리고 전·후 프레임 간의 차 영상을 생성하여 피부색 확률 분포 영상에 적용한다. 그 후 일정한 크기의 블록(30×40)을 사용하여 영상 내에서 피부색 확률 분포 합이 가장 큰 블록 영역을 추출한다. 이와 같은 방법으로 피부색 확률 분포 합이 가장 큰 상위 K개의 손 후보영역을 추출한다. 그림 2는 손 후보영역 검출 과정을 보여주고 있다.

나. 특징추출

본 논문에서는 손 후보영역의 특징으로 위치(Position)와 속도(Velocity)를 사용하였다. 질의영상의 j 번째 프레임에서 k 번째 후보영역의 특징벡터(Q_{jk})는 식(1)과 같다.

$$Q_{jk} = (x_{jk}, y_{jk}, u_{jk}, v_{jk}) \tag{1}$$

x, y는 후보영역의 중심위치이고, u, v는 Optical flow[5]를 사용하여 계산한 속도이다.

2.2. 유사도 비교

본 논문에서는 후보영역들 사이의 유사도를 비교하기 위해 2차·3차 중심 모멘트로 구성된 불변 모멘트(Invariant Moment)[6]를 사용하였다. 불변 모멘트는 이동, 회전, 크기 변화에 불변한 성질을 가지기 때문에 다양한 후보영역들 사이의 유사도를 비교하는데 적합하다. 7개의 불변 모멘트는 식(2)와 같다.

$$\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 \\
&\quad - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \\
&\quad \times [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
\phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
&\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 \\
&\quad - 3(\eta_{21} + \eta_{03})^2] - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03}) \\
&\quad \times [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned} \tag{2}$$

유사도를 비교하기 위해 후보영역을 크게 5개의 그룹(손, 부분 손, 팔뚝, 얼굴, 배경)으로 구분하였다. 5개의 그룹은 실험을 통해 검출 비율이 높은 후보영역들을 선별하였고, 각 후보영역은 200개의 샘플영상을 사용하였다. 그림 3은 후보영역의 예를 보여주며, 그림 4는 각 모멘트마다 정규화를 위한 상수를 곱하여 나타낸 그래프이다. 그림 4를 살펴보면 5차 불변 모멘트는 우수한 분별력을 가진다. 그러나 실험에 따

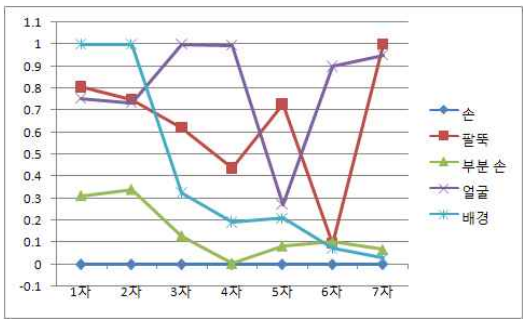


그림 4. 각 후보영역의 정규화된 평균 불변 모멘트

르면 분산 값이 크기 때문에 후보영역들을 정확하게 구분할 수 없다. 6, 7차 모멘트는 손 영역(손, 부분 손)과 나머지 후보영역들의 값이 차이가 거의 없기 때문에 후보영역들을 구분할 수 없다. 7개의 불변 모멘트 중 1～4차 불변 모멘트만 각 후보영역들을 구분할 수 있다. 따라서 본 논문에서는 후보영역간의 유사도를 비교하기 위해 1～4차 불변 모멘트를 사용한다.

2.3. 제안한 Dynamic Space Time Warping

DSTW는 DTW를 공간영역으로 확장한 알고리즘으로써 질의영상에서 매 프레임마다 검출된 K개 손 후보영역의 특징을 이용한다.

전체 m개의 프레임으로 구성된 모델영상의 시퀀스는 $M=(M_1,...,M_m)$이며, M_i는 i번째 프레임에서 추출된 특징벡터를 나타낸다. 전체 n개의 프레임으로 구성된 질의영상의 시퀀스는 $Q=(Q_1,...,Q_n)$이다. j번째 프레임 Q_j는 K개의 손 후보영역을 포함하며 $Q_j=\{Q_{j1},...,Q_{jK}\}$로 나타낼 수 있다. 이때 Q_{jk}는 j번째 프레임의 k번째 손 후보영역에서 추출된 특징벡터다.

손 후보영역의 중심위치 (x,y)는 간단하면서 다루기 쉬운 특징벡터다. 하지만 중심위치는 절대좌표이기 때문에 모델영상의 손동작 위치가 질의영상의 손동작 위치에 불변(Translation Invariance)하기 위해 중심위치를 상대좌표로 변환할 필요가 있다. 즉, 첫 번째 프레임의 k번째 손 후보영역을 기준으로 나머지 프레임에 속한 모든 손 후보영역의 중심위치를 상대좌표로 변환한다. 이렇게 첫 번째 프레임의 각 손 후보영역을 기준으로 총 K번의 DSTW 알고리즘을 실행한 후 구한 매칭 비용 중 가장 작은 값이 M과 Q 사이의 최적의 매칭 비용이다.

기존의 DSTW 알고리즘은 비교적 단순한 특징인 중심위치와 속도만 사용하여 모델영상과 질의영상 사이의 매칭 경로를 계산한다. 만일 질의영상 내에서 손 후보영역들이 손동작과 관련이 없는 경로를 생성하여 다른 모델로 매칭 된다면 사용자가 질의한 손동작은 정확히 인식될 수 없다. 따라서 본 논문에서는 이를 해결하기 위해 후보영역의 질감정보를 비교하여 잘못된 매칭경로의 비용을 최대화하는 방법을 제안한다. 제안한 방법은 불변 모멘트를 사용해 첫 번째 프레임에서 기준이 되는 손 후보영역과 나머지 프레임에 속한 모든 손 후보영역들 사이의 유사도를 계산한다. 유사도는 식 (3)과 같이 1～4차 불변 모멘트의 차의 누적 합으로 계산하였다.

$$l(A,B)=\sum_{i=1}^{4}\left|\phi_i^A-\phi_i^B\right| \qquad (3)$$

$l(A,B)$는 기준이 되는 손 후보영역과 Q_{jk}의 유사도를 나타내며, M_i와 Q_{jk}의 거리차를 계산할 때 가중치로 적용된다. 최적의 경로(Warping Path) W는 M과 Q 사이의 최적의 매칭경로이며 $W=(w_1,...w_T)$로 나타낼 수 있다. W의 한 요소는 $w_t=(i,j,k)$이며 모델영상의 특징벡터(M_i)가 질의영상의 특징벡터(Q_{jk})와 매칭됨을 의미한다. 유사도를 가중치로 적용하여 M과 Q 사이의 최적의 경로 W를 계산하는 알고리즘은 그림 5와 같다. $d(i,j,k)$는 M_i와 Q_{jk}의 거리차이며, L_{rjk}는 기준이 되는 손 후보영역(S_r)과 다른 프레임의 손 후보영역(Q_{jk}) 사이의 정규화된 유사도 가중치이다. DSTW 알고리즘은 최종적으로 최적의 매칭경로 W^*와 매칭비용 D^*를 산출한다.

입력 : 모델의 특징 벡터 $M_i,\ 0\le i\le m$
질의의 특징 벡터 $Q_j=\{Q_{j1},...,Q_{jK}\}$
$1\le j\le n$
기준 후보 영역 $S_r,\ 1\le r\le K$
정규화 상수 C
S_r과 Q_{jk}의 정규화된 유사도 $L_{rjk}=C\times l$
출력 : 매칭비용 D^*, 최적의 경로(Warping Path) W^*

```
for r = 1 : K do
  j = 0
  for i = 0 : m do
    for k = 1 : K do
      D(i,j,k) = ∞
    end
  end
  D(0,0,1) = 0

  for j = 1 : n do
    for i = 0 : m do
      for k = 1 : K do
        if i = 0 then
          D(i,j,k) = ∞
        else
          w = (i,j,k)
          L_rjk = C × ( Σ_{i=1}^{4} |φ_i^{S_r} − φ_i^{Q_jk}| )
          for w' ∈ N(w) do
            C(w',w) = D(w') + (L_rjk × d(w))
          end
          D(w) = min_{w'∈N(w)} C(w',w)
          b(w) = argmin_{w'∈N(W)} C(w',w)
        end
      end
    end
  end

  k* = argmin_k {D(m,n,k)}
  D* = D(m,n,k*)
  w*_T = (m,n,k*)

  w*_{t-1} = b(w*_t)

end
```

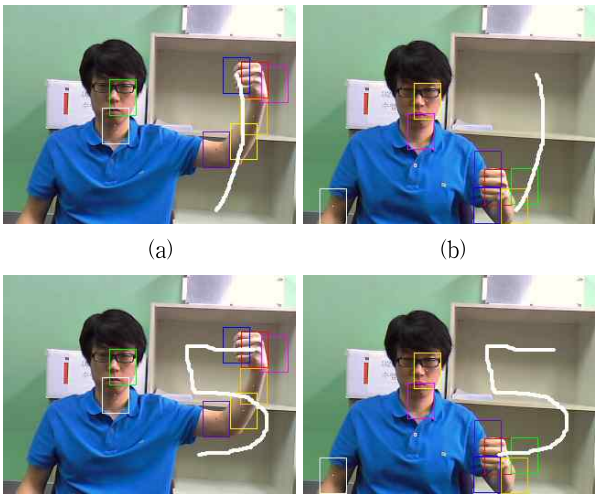
그림 5. 제안한 DSTW 알고리즘

3. 실험 결과

본 논문에서 제안하는 방법의 성능을 평가하기 위해서 질의영상에 대한 손동작 인식실험을 수행하였다. Visual Studio C++ 6.0과 OpenCV 1.0을 사용하여 구현하였고, 삼성 SPC-A130M 웹캠을 사용하여 초당 30프레임으로 획득된 영상(320×240)을 사용하였다. 인식에 사용된 제스처는 필기체 입력을 위해 Palm사에서 고안한 Graffiti 숫자[7]를 사용하였다. 사용된 영상은 각 숫자를 3회씩 생성하였으며 모델영상은 30개, 그리고 질의영상은 긴팔, 반팔 각각 30개이다.

그림 7은 사용자가 숫자 5를 질의한 결과를 나타낸다. 그림 7의 (a), (b)는 피부색 배경에서 기존의 방법을 사용했을 때 모델영상(숫자 1)과 매칭된 경로의 시작과 끝을 보여주고 있다. 그림 7의 (c), (d)는 제안한 방법을 사용하였을 때 모델영상(숫자 5)과 매칭된 경로의 시작과 끝을 나타낸다. 기존의 방법은 숫자 1로 잘못 인식하는 반면 제안한 방법은 숫자 5로 정확하게 인식하는 것을 확인할 수 있다.

그림 6. Palm's Graffiti 숫자



(a) (b)

(c) (d)

그림 7. 기존의 방법과 제안한 방법의 매칭경로 (숫자 5)
(a),(b) 기존의 방법 (c),(d) 제안한 방법

표1. 배경 환경에 따른 인식률 ($K=8$)

환경	단순한 배경		복잡한 배경		피부색 배경	
방법	기존 DSTW	제안한 DSTW	기존 DSTW	제안한 DSTW	기존 DSTW	제안한 DSTW
긴팔	93.3	96.6	93.3	93.3	90.0	93.3
반팔	90.0	96.6	83.3	93.3	83.3	93.3

표2. 사용자별 인식률 (피부색 배경 , 반팔)

사용자별	기존 DSTW	제안한 DSTW
동일한 사용자	83.3	93.3
다른 사용자	78.3	90.0

표 1은 다양한 배경환경에 따른 인식률의 실험 결과를 보여주고 있다. 실험 결과로서 기존의 방법은 0, 1, 4, 7과 같이 비교적 간단한 궤적을 가지는 숫자는 정확하게 인식하는 반면 2, 3, 5, 6, 8, 9와 같이 복잡한 궤적을 가지는 숫자는 정확하게 인식하지 못하였다. 특히 배경에 피부색과 유사한 색상이 많이 분포하거나, 사용자의 신체 부위가 많이 노출될수록 오인식 되는 결과가 두드러지게 나타났다. 이는 배경에서 잘못된 손 후보영역을 생성하거나, 손뿐만 아니라 팔뚝의 피부색 영역이 잘못된 경로를 생성하기 때문이다. 표 2는 사용자가 다른 모델영상과 질의영상의 인식률에 대한 실험 결과를 보여 주고 있다. 실험결과로써 제안한 방법이 기존의 방법보다 인식률이 우수함을 확인하였다. 인식률은 제안한 방법이 기존의 방법에 비해 3%~14% 개선된 것을 확인하였다.

4. 결론 및 향후 연구방향

본 논문에서는 각 후보영역의 질감정보를 고려한 DSTW 기반의 손동작 인식 방법을 제안하였다. 제안한 방법은 후보영역 사이의 질감정보를 비교하기 위해 7개의 불변 모멘트 중 분별력이 우수한 4개의 불변 모멘트를 사용하였다. 유사도 비교는 질의영상 내 첫 번째 프레임의 각 후보영역을 기준으로 나머지 후보영역들에 대해 불변 모멘트의 거리 차로 계산되며 모델영상과 질의영상을 매칭할 때 정규화된 가중치로 사용하였다. 모델영상과 질의영상에서 추출한 특징벡터의 거리 차에 정규화된 가중치를 적용함으로써 질감이 서로 다른 후보영역들로 생성된 경로가 다른 모델로 오인식될 수 있는 것을 제한하였다. 실험 결과를 통해 제안한 방법이 기존의 방법보다 다양한 배경에서도 인식률이 우수함을 확인하였다. 향후 과제로는 모델영상과 질의영상의 숫자 크기에 따른 비교 연구가 필요하다.

Acknowledgement

이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2009-0077434)

참고문헌

[1] 홍동표, 우운택, "제스처기반 사용자 인터페이스에 대한 연구 동향", Telecommunications Review, 18, 3, pp 403-413, 2008.

[2] E. Keogh. "Exact indexing of dynamic time warping", *International Conference on Very Large Data Bases*, pp 406-417, 2002.

[3] Jonathan Alon, Vassilis Athitsos, Quan Yuan, Stan Sclaroff, "Simultaneous Localization and Recognition of Dynamic Hand Gestures", wacv-motion, vol. 2, pp.254-260, *IEEE Workshop on Motion and Video Computing (WACV/MOTION'05)* - Volume 2, 2005

[4] M. Jones and J. Rehg. "Statistical color models with application to skin detection", *IJCV*, 46(1):81-96, January 2002.

[5] Q. Yuan, S. Sclaroff, and V. Athistos. "Automatic 2D hand tracking in video sequences. In Proc. *WACV*, 2005.

[6] Hu, M.K. "Visual pattern recognition by moment invariants", *IEEE Transactions on information Theory*, IT-8, pp 179-187, 1962.

[7] Palm. Graffiti alphabet. http://www.palmone.com