HMM을 이용한 알파벳 제스처 인식

윤 호섭, 소 정, 민 병우
컴퓨터소프트웨어연구소 영상처리연구부
e-mail:hsyoon@yoon seri.re.kr

Alphabetical Gesture Recognition using HMM

Ho-Sub Yoon, Jung Soh, Byung-Woo Min
Image Processing Division, Systems Engineering Research Institute
e-mail:hsyoon@yoon.seri.re.kr

Abstract

The use of hand gesture provides an attractive alternative to cumbersome interface devices for human-computer interaction(HCI). Many methods for hand gesture recognition using visual analysis have been proposed such as syntactical analysis, neural network(NN), Hidden Markov Model(HMM) and so on. In our research, a HMMs is proposed for alphabetical hand gesture recognition. In the preprocessing stage, the proposed approach consists of three different procedures for hand localization, hand tracking and gesture spotting. The hand location procedure detects the candidated regions on the basis of skin-color and motion in an image by using a color histogram matching and time-varying edge difference techniques. The hand tracking algorithm finds the centroid of a moving hand region, connect those centroids, and thus, produces a trajectory. The spotting algorithm divides the trajectory into real and meaningless gestures In constructing a feature database, the proposed approach use the mesh feature code for codebook of HMM In our experiments, 1300 alphabetical and 1300 untrained gestures are used for training and testing, respectively. Those experimental results demonstrate that the proposed approach yields a higher and satisfying recognition rate for the images with different sizes, shapes and skew angles.

I. 서론

컴퓨터 기술의 빠른 발전은 유연한 사용자 접속의 요구를 증대시키게 되었다. 사용자 접속기술은 영상과 음성을 주로 하는 멀티미디어 기술 및 마우스 등과 같은 사용도구의 발전에 의해 매우 편리하게 되었으나, 궁극적으로는 사람과 사람 사이의 의사소통 수준에 이르러야 한다. 이를 위해 음성과 시각정보에 기반한 사용자 접속이 되어야 하며, 제스처는 사용자 접속에서 중요한 분야가 되었다. 제스처란 사람이 지니고 있는 사고(concept)를 손을 포함한 몸의 움직임으로 나타낸 물리적 표현으로, 사람이 인식하는 제스처는 물리적인 사고기 신체의 움직임이라는 물리적인 운동을 거쳐, 우리의 눈에 들어오는 신체의 운동, 즉 사고의 시각정보로서 정의된다. 우리는 대화를 하는 중에 무의식적 혹은 의식적으로 어떤 동작을 취함으로써 자신의 의사를 상대편에게 보다 잘 전달이 되도록 하며 또한 언어로서 나타나지 않는 자신의 감정까지도 표출하게 된다. 이와 같이 제스처는 우리의 일상생활에서 자신의 의사를 표현하는 데 중요한 보조 수단으로 활용되고 있다. 즉, 제스처는 2 차원 혹은 3 차원 상에서 적합한 입력도구로 사용될 수도 있으며 사람간의 의사 전달에 언어 및 눈의 움직임 등과 함께 사용되는 정보의 주된 이동 경로라 할 수 있다[1]. 이에 따라 제스처를 인식하고자 하는 연구가 국내외에서 활발하게 연구되고 있으며 과거에는 주로 데이터 글러브나 Marker등의 특정 하드웨어를 이용한 방법이 많이 연구되었으나 현재는 주로 카메라로부터 획득된 영상을 처리하는 방법이 주로 연구되고 있다

본 논문에서도 카메라를 이용해 실시간으로 제스처를 인식하는 시스템을 기술한다. 본 논문의 구성은 2 장에서 영상처리를 이용한 전처리에 대해 설명하고 3 장에서는 HMM(Hidden Markov Model)을 이용한 인식기를 그리고 4 장에서는 실험 결과를, 마지막으로 5장에서 결론을 논의하고자 한다

II. 영상처리를 이용한 전처리

본 절에서는 제스처 추적(tracking)에 대하여 기술한다. 제스처 추적의 목적은 사용자의 손 움직임을 최대한 정확하게 추적하여 그 결과 생성되는 손의 이동 궤적(trajectory)을 제스처 인식 모듈에 입력하는 데 있다. 제스처 추적은 실제로는 인속되는 각 프레임에서 손 영역을 추출하고 그 위치들을 시간 흐름에 따라 연결함으로써 이루어진다. 따라서 제스처 추적의 가장 핵심적인 부분은 정지된 하나의 영상에서 손 영역을 추출하는 알고리즘이라 할 수 있다. 본 절에서는 먼저 제스처 추적의 입출력에 관해 기술하고, 손 영역 추출 알고리즘을 상세하게 설명한 후, 끝으로 제스처 추적을 기술한다.

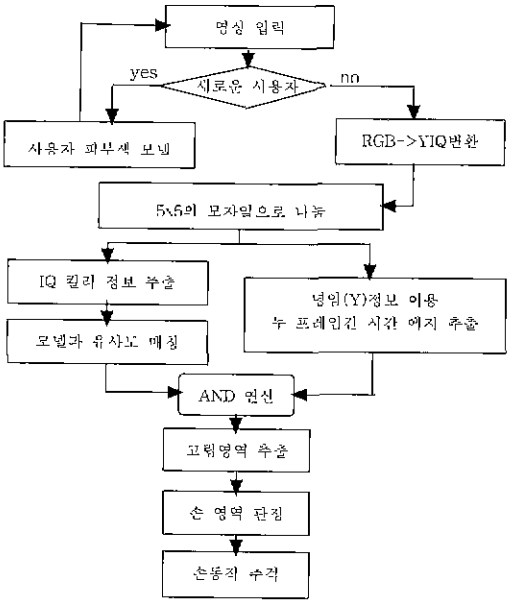

그림 1. 전처리 흐름도

2.1 컬러 좌표계 변환

RGB 컬러 좌표계는 이용하면 RGB 성분 모두에 색상 정보가 포함되어 있으므로 16,777,216가지 색상 모두에 대한 고려가 필요하다는 것을 의미한다. 따라서 대부분의 컬러 영상처리 응용 프로그램들에서는 색상과 명도 정보가 혼합된 RGB 컬러 좌표계 상의 영상 정보를 색상과 명도 정보가 분리되는 다른 컬러 좌표계로 변환하여 처리한다. 색상과 명도 정보가 분리되는 컬러 좌표계로 본 연구에서는 YIQ 공간을 이용하였다.[2]

2.2 사용자 모델 데이터 생성

일반적인 사람의 눈으로는 사람의 피부 색상이 쉽게 구분되지만, 카메라를 통해 입력되는 디지털 영상에서 사람의 피부 색상은 배경, 조명 상태 및 개인의 특성에 따라 그 차이가 매우 크다. 따라서 구현된 시스템에서는 조명 및 사용자의 변화에 따라 새로운 피부 색상 모델을 설정할 수 있도록 하였다.

2.3 사용자 모델과 입력 영상간의 컬러 정보 매칭

모델의 IQ 히스토그램이 구해지면 다음은 새로운 입력 영상을 5x5 영역을 갖는 각각의 모자익(Mosaic)으로 분리한 후, 부속 모자익의 IQ 히스트그램과 모델의 IQ 히스토그램과의 유사성 매칭을 통해 피부색인지를 결정한다. 본 연구에서는 히스토그램과의 유사도 결정을 위해 David Saxe가 사용한 식 (1)을 사용하였다[3].

$$S_{M,C}=\frac{\sum_{i,q}\min(H^{M}(i,q),\ H^{C}(i,q))}{\sum_{i,q}H^{M}(i,q)} \quad \cdots\cdots \quad \text{식 (1)}$$

식 (1)에서 H는 히스토그램을, M은 Model을 C는 현재 영역(Current Area)을 (i,q)는 IQ 테이블을 의미한다. 즉, $S_{M,C}$는 모델과 현재 영역 사이의 유사성 값으로서 두 히스토그램을 겹쳐 최소 값을 취함으로 얻어진다. 이 값은 0.0 ~ 1.0 사이의 다양한 분포를 가지므로 이 값에서 피부 컬러를 결정하기 위해 본 연구에서는 Otsu 알고리듬[4]을 사용한 임계치 결정 알고리듬을 사용하였다.

2.4 시간 에지를 이용한 차 영상 추출

동 영상 처리에서 가장 일반적으로 접근하는 방법은 연속되는 두 영상간의 차를 이용하는 방법이다. 본 연구에서도 차 영역 추출을 위해 Takahashi가 사용한 시간 방향 에지를 구한다[5]. Takahashi의 차 영상은 시간 에지 마스크가 저주파 통과 필터의 성격을 가지므로 동작시의 미리, 어깨 등에서 발생하는 미세한 움직임을 어느 정도 제거한 차 영상을 얻을 수 있다.

2.5 손 영역 추출

추출된 피부색 영역에는 손 피부와 동일한 얼굴 영역 및 잡음 영역이 포함된다. 이러한 얼굴 영역 및 잡음 영역은 시간 에지와의 간단한 AND 연산을 통해 쉽게 제거된다. 즉, 피부색 영상과 시간에지 영상에서 동시에 존재하는 부분만을 구하면 원하는 그림 2의 손 영역 후보 영역이 찾아진다. 이 가정은 영상에서 손만이 움직인다는 가정 하에서 성립되며, 만일 손이 움직이지 않고 정지해 있다면 손동작 추적에 의해 얻어진 이전 프레임에서의 손의 위치를 손 영역으로 결정하면 된다.

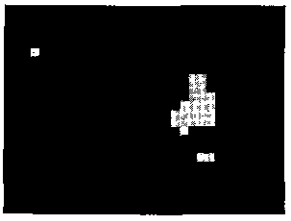

그림 2. 추출된 손 영역 후보 영역

2.6 손동작 추적

제스처 인식을 위한 손동작의 추적은 시간 t 동안 구해진 손 영역의 중심점을 연결하면 얻어진다. 즉 각각의 프레임에서 중심점을 구한 후, 시간 t 동안 연결하면 손동작의 궤적을 구할 수 있다.

III. HMM(Hidden Markov Model)을 이용한 인식기

3.1 특징 추출

HMM을 이용한 인식을 수행하기 위해서는 먼저 제스처 궤적으로부터 특징을 추출해야 한다. 가장 쉽게 특징을 추출할 수 있는 방법으로 제스처 궤적에서 MBR(Minimum Bounding Rectangle)을 구한 후 Mesh 특징을 이용할 수 있다. 이 특징은 전체 영역을 공간적으로 분할해서 위치 정보를 이용해 특징 코드를 생성하는 대표적인 방법이다.

3.2 HMM

Hidden Markov Model (HMM)[6]은 위의 유한 상태 오토마타에 확률 개념을 도입한 것이라고 볼 수 있으며, 음성 인식에 가장 많이 사용되는 인식 모델이다. HMM은 상태(state)라 불리는 노드와 이들간의 전이를 나타내는 선분으로 구성된 그래프로 표현될 수 있다. 그래프의 각 노드에는 공간적인 특성을 모델링하는 관측 심볼 확률 분포와 초기 상태 확률 분포가 저장되어 있으며, 각 선분에는 관측열의 시간적인 특성을 모델링하는 상태 전이 확률 분포가 저장되어 있다. HMM은 아래와 같은 요소로 구성된다.

1. N: 상태의 수, $S=\{S_1,S_2,\ldots,S_N\}$ 상태 상태의 집합, q_t 시간 t의 상태,
2. M: 관측 심볼의 수, $V=\{v_1,v_2,\ldots,v_M\}$ 관측 심볼의 집합
3. $A=\{a_{ij}\}$:상태 전이 확률분포
   $a_{ij}=P(q_{t+1}=S_j|q_t=S_i),\ 1\le i,j\le N$; 상태 i에서 상태 j로 전이할 확률 상태 전이 확률 분포,
4. $B=\{b_j(k)\}$,관측 심볼 확률 분포, $b_j(k)=P(v_k \text{ at } t|q_t=S_j)$, $1\le j\le N,\ 1\le k\le M$, j에서 심볼 v_k를 관측할 확률
5. $\pi=\{\pi_i\}$:초기 상태 확률 분포, $\pi_i=P(q_1=S_i),\ 1\le i\le N$, 초기 상태가 i일 확률

일반적으로 하나의 HMM은 $\lambda=(A,B,\pi)$로 표시된다. 주어진 모델 와 관측열 $O=O_1,O_2,\ldots O_T$에 대해 생성 확률은 아래와 같다.

$$P(O|\lambda)=\sum_{\text{for all } Q}\left[\pi_{q_1}b_{q_1}(O_1)\prod_{t=2}^{T}a_{q_{t-1}q_t}b_{q_t}(O_t)\right] \quad \text{..... 식(2)}$$

HMM의 응용에는 세 가지의 해결해야 할 문제가 있다. 즉, 평가, 해석, 그리고 학습에 있으며 이들은 각기 Forward-Backward 알고리즘, Viterbi 알고리즘 그리고 Baum-Welch 알고리즘으로 해결된다. 다음 그림은 본 연구에서 사용된 일반적인 Left-Right HMM의 형태를 보여준다.

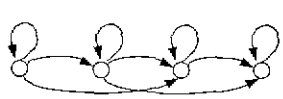

그림 3. Left-Right HMM 모델

IV. 실험 결과

본 연구에서는 동적 제스처 인식을 일반 PC상에서 실 시간으로 구현하였다. 사용된 컴퓨터는 IBM PC Pentium- Pro 200MHz로서 OS로는 Windows95를 이용하였다. 영상 입력을 위한 하드웨어로는 영상 입력 장치로 일반 캠코더나 일반 CCD 카메라를, 영상 캡처 보드는 100만원대의 저가형인 Meteor보드를 사용하였다. 개발 프로그램 언어는 Visual C++ 4.2를 사용하여 구현하였으며, 카메라를 통해 입력된 영상의 해상도는 160 x 120 pixels 및 24-bit true 컬러이다. 실험에 이용된 제스처는 알파벳 제스처 26 자로 실제 알파벳을 거의 변형시키지 않고 같은 모양으로 제스처를 생성하여 인식율을 을 테스트 해 보았다. Table 1에서 실험 결과를 볼 수 있다.

Table 1. 알파벳 제스처 실험 결과

Recognition Results(%)		
Trained Data	New Data	Overall
1232/1300(94.8)	1220/1300(93.9)	2452/2600(94.3)

V. 결론

본 연구에서는 HMM을 이용한 실시간 알파벳 손 제스처인식에 관해서 기술하였다. 제스처 인식 시스템은 실제 응용에 있어 아직 요원한 것이 사실이나 알파벳과 같은 독립된 제스처가 인식 가능하다면 향후 알파벳의 집합인 단어도 인식 가능하리라 여겨진다. 단어가 인식 가능하다면 컴퓨터와 사람 사이에 또 하나의 인터페이스로 실제 사용될 수 있을 것으로 기대된다. 그러므로 향후 연구 계획으로는 연속된 단어를 인식할 수 있는 시스템의 개발이 이루어 져야 할 것이다.

참고 문헌

[1] P. A Harling, "Gesture Input using Neural Networks", B S Thesis, Dept. of C. S , University of York 1993.

[2] R. C Gonzalez, R. E. Woods, Digital Image Processing, Addison-Wesley, 1992

[3] 최 형일, 컴퓨터 비젼 입문, 홍릉과학출판사, 1991.

[4] Dvid Saxe, Richard Foulds, "Toward Robust Skin Identification in Video Images", International Workshop on Automatic Face and Gesture Recognition, Vermont, pp 379-384, 1996.

[5] Tomoichi Takahashi et al, "A Hand Gesture Recognition Method and Its Application", IEICE D-II, Vol. J73-D-II, No. 12, pp 1985-1992, Dec. 1990

[6] L. R. Rabiner A tutorial on hidden Markov models and selected application in speech recognition, *Proc. IEEE 77*, pp. 267-293, 1989.