1999 年度 大韓電子工學會 夏季綜合學術大會 論文集 第 22 卷 第 1 號 99/6

다중 신경망을 이용한 사용자의 응시 위치 추출

박강령, 이정준, 이동재, 김재희
연세대학교 전기·컴퓨터 공학과 지능형 vision 연구실

Gaze Detection Using Two Neural Networks

Kang Ryoung Park, Jeong Jun Lee, Dong Jae Lee and Jaihie Kim
Intelligent Vision Lab, Dept. of Electrical and Computer Eng.,
Yonsei University
E-mail : parkgr@seraph.yonsei.ac.kr

Abstract

Gaze detection is to locate the position on a monitor screen where a user is looking at. We implement it by a computer vision system setting a camera above a monitor, and a user move (rotates and/or translates) her face to gaze at a different position on the monitor. Up to now, we have tried several different approaches and among them the Two Neural Network approach shows the best result which is described in this paper.(1.7 inch error for test data including facial rotation, 3.1 inch error for test data including facial rotation and translation).

1. 서론 (Introduction)

시선의 추출을 통해 사용자의 관심 방향을 알고자 하는 연구는 여러 분야에 응용될 수 있는데, 대표적인 것이 장애인의 컴퓨터 이용이나, 다중 윈도우에서 마우스의 기능 대용 및, 공정 제어 환경에서의 process control 그리고 원격 회의 시스템에서의 view controlling 등이다. 기존의 대부분의 연구들에서는 얼굴의 3차원 움직임량(3D rotation, translation)을 구하는데 중점을 두고 있으며[1][2][3], 모니터, 카메라, 얼굴 좌표계간의 복잡한 변환 과정 때문에 이를 바탕으로 사용자의 응시 위치를 파악하고자하는 연구는 거의 이루어지지 않고 있다[4]. Riker[4]등의 연구에서는 사용자와 모니터 사이의 거리를 고정시키고 모니터상의 한 지점을 쳐다보기 위해서는 얼굴의 회전만 허용하도록 하였다. 그러나 이러한 방법은 실제 환경에서 사용자가 사용하기에 어려울 정도의 제약요소로 작용한다. 이러한 기존의 연구들이 가지고 있는 문제점들을 해결하기 위하여 이 논문에서는 일반 사무실 환경에서 입력된 얼굴 동영상으로부터 얼굴 영역 및 얼굴내의 눈, 코, 입 영역 등을 추출함으로써 모니터의 일정 영역을 응시하는 순간 변화된 특징점들의 위치 및 특징점들이 형성하는 기하학적 모양의 변화를 바탕으로 응시 위치를 계산하였다. 이때 앞에서 기술한 세 좌표계(모니터, 카메라, 얼굴 좌표계)간의 복잡한 변환 관계를 해결하기 위하여, 신경망 구조(다층 퍼셉트론)을 이용하였다. 신경망의 학습 과정을 위해서는 모니터 화면에 22개의 위치를 정하여 각 위치를 응시할 때 추출된 특징값들을 사용하였다. 이때 학습된 22개의 응시 위치이외에 또 다른 응시 영역에 대한 출력값을 얻기 위해, 출력 함수로 연속적이고 미분가능한 함수들을 사용하였다. 실험 결과 linear output function을 사용하였을 경우 가장 우수한 응시 위치 파악 결과를 나타냈으며, 얼굴의 회전에 관한 단일 신경망을 사용하였을 때보다 얼굴의 회전 및 이동에 관한 다중 신경망을 사용하였을 때 보다 정확한 응시 위치 결과를 나타냄을 알 수 있었다.

2. 얼굴 영역 및 얼굴내 특징점 추출

사용자의 응시 위치를 파악하기 위하여, 이 논문에서는 얼굴내의 특징점(양눈, 코, 입)의 위치 변화 및 특징점들이 형성하는 기하학적인 모양의 변화도를 이용한다. 이를 위해 이 논문에서는 먼저 얼굴 영역을 검출한 후 추출된 얼굴 영역내의 제한된 범위 내에서 양 눈과 코 및 입의 양 끝점을 추출한다.

2.1 차영상 정보와 칼라 정보를 이용한 얼굴 영역의 추출

이 논문에서는 시간적으로 연속된 두 영상간의 차영

상 정보와 얼굴의 살색 정보를 이용하여 얼굴 영역을 검출한다. 이때 차영상 정보이외에 살색 정보를 같이 이용한 이유는 사용자의 뒷 배경에서 움직임이 있는 물체를 얼굴 영역으로 오인식하는 경우를 막기 위해서이다[5]. 얼굴의 살색 영상 처리부에서는 입력된 얼굴의 살색 칼라 정보에 대한 RGB신호를 YIQ model로 변환함으로써 얼굴의 살색 정보에 민감한 I성분 구간(110~150)을 바탕으로 얼굴 영역을 검출한다[5]. 이때, 차영상내에서 얼굴로 검출된 부분과 color model에서 얼굴로 검출된 부분에 대한 공통 부분(intersection)을 택함으로써 그림 1처럼 입력되는 얼굴 영상으로부터 빠르고 정확하게 얼굴 영역을 검출 할 수 있다.

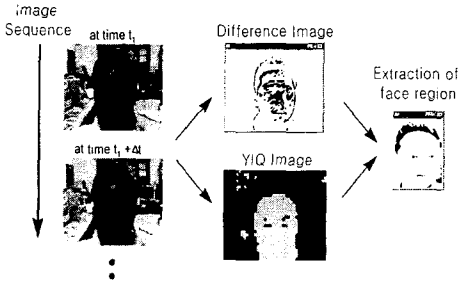


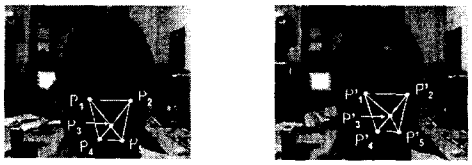
Fig. 1 얼굴 영역의 검출

2.2 수평·수직 히스토그램 분석법을 이용한 눈동자, 코 및 입의 양 끝점 추출 및 움직임 추적

추출된 얼굴 영상은 히스토그램 평활화 및 이진화과정을 통해 이진 영상으로 변환한다. 이때, 얼굴내의 눈, 코/입의 위치에 대한 사전정보와 이진 영상에 대한 제한된 범위 내에서 수직, 수평히스토그램의 최고치를 계산함으로써 눈/코/입의 위치를 정확하게 추출할 수 있다. 또한 초기 영상에서의 특징점 추출 방법과는 달리 이후 연속 영상에서는 매번 얼굴 영역을 다시 추출하지 않고, 이전에 추출된 특징점의 위치로부터 현재 특징점의 위치를 예측하는 알고리듬을 사용함으로써 특징점의 움직임을 추적한다[6].

3. 신경망 입력을 위한 특징값 및 이의 정규화 과정

사용자의 응시 위치를 파악하기 위해 이 논문에서는 양눈과 코, 입의 위치를 특징점으로 사용하였다. 아래의 그림 2.(a)와 그림 2.(b)는 각각 모니터의 정중앙과 일정 영역을 응시하는 순간에 추출된 특징점들의 위치를 나타낸 것이다. 이때, 추출된 특징점들로부터 응시 위치를 파악하기 위해 이 논문에서는 다음과 같은 20개의 특징값들을 신경망의 입력 노드로 사용하였다.



(a)모니터 정 중앙 응시 (b) 모니터 일정 영역 응시

Fig. 2 모니터 정중앙과 일정 영역을 응시하는 순간의 특징점의 위치 변화

▷ 모니터의 정중앙을 응시 할 때
P_1 (왼쪽눈 : X_1, Y_1), P_2 (오른쪽눈 : X_2, Y_2),
P_3 (코 : X_3, Y_3), P_4 (입의 왼쪽 끝 : X_4, Y_4)
P_5 (입의 오른쪽 끝 : X_5, Y_5)
◁ 모니터의 일정 영역을 응시 할 때
P'_1 (왼쪽눈:X'_1, Y'_1), P'_2 (오른쪽눈:X'_2, Y'_2),
P'_3 (코:X'_3, Y'_3), P'_4 (입의 왼쪽 끝:X'_4, Y'_4)
P'_5 (입의 오른쪽 끝:X'_5, Y'_5)

특징값 1 ~ 5 : $X'_i - X_i$ (i = 1, 2, ... 5),
특징값 6 ~ 10 : $Y'_i - Y_i$ (i = 1, 2, ... 5)
특징값 11 : $S(\angle P'_1P'_2P'_3) - S(\angle P_1P_2P_3)$,
특징값 12 : $S(\angle P'_1P'_3P'_4) - S(\angle P_1P_3P_4)$
특징값 13 : $S(\angle P'_2P'_3P'_5) - S(\angle P_2P_3P_5)$,
특징값 14 : $S(\angle P'_3P'_4P'_5) - S(\angle P_3P_4P_5)$
특징값 15 : $S(\triangle P'_1P'_3P'_4)/S(\triangle P'_2P'_3P'_5) - S(\triangle P_1P_3P_4)/S(\triangle P_2P_3P_5)$
특징값 16 : $S(\triangle P'_3P'_4P'_5)/S(\triangle P'_1P'_2P'_3) - S(\triangle P_3P_4P_5)/S(\triangle P_1P_2P_3)$
특징값 17 : $\{(X'_1+X'_4)/2 - X'_3\} - \{(X_1 + X_4)/2 - X_3\}$
특징값 18 : $\{X'_3-(X'_2+X'_5)/2\} - \{X_3 - (X_2 + X_5)/2\}$
특징값 19 : $\{(Y'_1+Y'_5)/2-Y'_3\} - \{(Y_1 + Y_5)/2 - Y_3\}$
특징값 20 : $\{Y'_3-(Y'_1+Y'_2)/2\} - \{Y_3 - (Y_1 + Y_2)/2\}$

그러나 사용자의 앉은 키와 모니터와의 거리에 따라 입력 특징값의 차이가 크다면 정확한 응시 위치를 나타낼 수 없을 것이다. 그러므로 이 논문에서는 정규화 과정을 통해 입력 특징값의 변화도를 수용하고자 한다. 그러나 실제의 경우 카메라가 모니터의 위에 설치되어 있는 관계로 사용자의 앉은 키에 따른 변화도는 크지 않은 결과를 나타냈으므로, 이 논문에서는 모니터와 사용자간의 거리에 따른 변화도만 정규화한다. 즉, 아래식 (1)과 같이 초기에 사용자로 하여금 모니터의 22영역중 최우측 상단과 최좌측 하단을 응시하게 함으로써 얻어진 각 특징값들의 최대 최소치의 변위 차이로 입력 특징값들을 나눔으로써 정규화하였다.

$$d_i = \frac{d_i}{|\max(d_i) - \min(d_i)|} \cdots\cdots\cdots(1)$$

단, $d_i\,(i=1,2,\ldots 20)$: 신경망 입력 특징값

4. 응시위치 파악을 위한 다중 신경망

이 논문에서는 사용자의 응시위치 파악을 위해 신경망(다층 퍼셉트론)을 사용하였다. 이때 신경망의 입력 노드로는 앞에서 설명한 20개의 특징값들을 사용하고, 출력노드를 통해 모니터의 X, Y축 응시 위치를 나타내도록 하였다. 학습 데이터로는 19인치 모니터상에서 임의로 정한 22개의 응시 위치를 쳐다보는 순간 측정된 특징값들을 사용하며 출력함수로는 학습된 22영역 이외의 응시 영역에 대한 정확한 출력값을 나타낼 수 있도록 세종류의 미분 가능하며 연속적인 함수들(sigmoid, inverse sigmoid, linear 함수)들의 성능을 비교하여 가장 우수한 성능을 나타내는 함수(linear 함수)를 사용하였다. 신경망의 학습을 위해서는 generalized delta rule을 사용하였으며 총 10000~50000번의 반복 횟수등에 대하여 실험하였다. 그런데 일반적으로 사용자가 모니터상의 여러 지점을 응시하는 경우에는 얼굴의 회전(rotation) 뿐아니라 얼굴축의 이동(translation)이 같이 발생하게 된다. 그러므로 이 논문에서는 다음 그림 3과 같이 입력 영상으로부터 사용자의 얼굴 윤곽선을 추출하여 이 윤곽선의 중심점이 이동된 량을 얼굴축의 이동량으로 간주하고 앞에서 추출한 20개의 특징값들을 얼굴의 회전량으로사용하는 다중 신경망을 이용하여 사용자의 응시 위치 판단 결과를 실험하였다. 얼굴의 윤곽선을 추출하기 위하여 입력 영상을 8방향 sobel edge operator를 이용하여 edge 영상으로 변환한 후, 미리 저장된 타원 모양의 edge template을 이용하여 template matching함으로써 얼굴의 윤곽 위치를 찾는 방식을 사용하였다. 이때 타원 모양의 에지정보만을 template에 사용한 이유는 얼굴내의 눈, 코, 입등은 사람마다 차이를 나타내기 때문이다. 앞의 20개의 특징값들과 얼굴의 윤곽선 추출을 통해 구한 얼굴 중심점의 이동량을 나타내는 2개의 특징값을 신경망의 입력으로 하여 사용자의 응시 위치를 추출한다.

5. 실험 결과

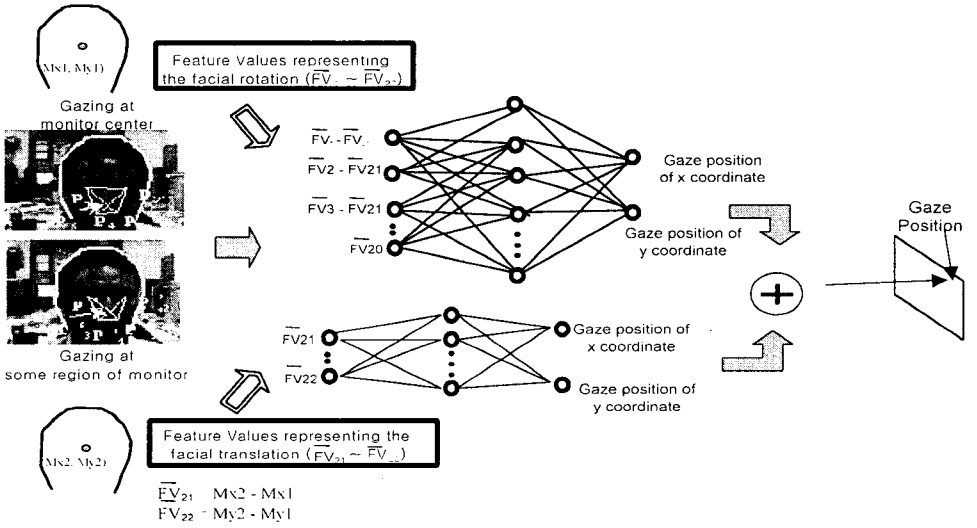
신경망의 학습 과정에는 19" 모니터를 기준으로 다양한 앉은 키와 자세를 갖는 총 220개의 학습 sample(10명분×22응시 영역)을 사용하였으며, 이때 모니터에서 사용자사이의 거리는 50~70cm정도의 거리를 유지하였다. 모니터의 각 영역에 대한 응시 위치의 정확도는 학습에 사용하지 않은 10명분의 test 데이터를 이용하여 실험하였다.

(단위 : inch)

방법	선형 보간법	단일 신경망	다중 신경망
RMS error	1.81	1.61	1.7

표 1. 얼굴의 회전에 의한 응시 위치 정확도
(5명 × 22응시 위치 = 110 test data)

표 1에 의하면 얼굴의 회전만 있는 경우에는 단일 신경망을 사용하여을 때에 가장 우수한 성능을 나타내지만 다음 표 2와 같이 얼굴의 회전과 이동이 동시에



발생하는 경우에는 다중 신경망이 가장 우수한 성능을 나타냄을 알수 있었다.

(단위 : inch)

방법	선형 보간법	단일 신경망	다중 신경망
RMS error	4.54	4.40	3.4

표 2. 얼굴의 회전및 이동에 의한 응시 위치 정확도 (5명 × 22응시 위치 = 110 test data)

또한, 이 연구에서는 실험 환경을 3차원 그래픽 워크스테이션(Silicon graphic ws : Indigo-II) 으로 확장하여 그림 4와 같이 사용자의 응시 위치 결과에 따라 3차원 그래픽으로 구현된 가상 비행체를 격추시키는 시스템을 구현하였다. 3차원 그래픽은 SGI전용 open inventor로 작성하였으며, 실험 결과 기존의 mouse를 사용하였을 경우보다 얼굴의 움직임으로 가상의 비행체를 격추했을 때 보다 현실감나는 가상 현실 환경을 제공할 수 있었다

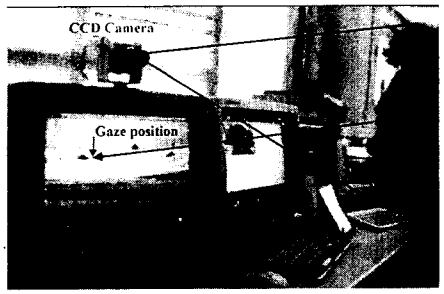


Fig. 9 가상 현실 환경에서 얼굴의 응시 위치 결과를 이용한 비행체 targetting

6. 결론

응시 위치 추적은 모니터위에 설치된 카메라로부터 입력된 사용자의 얼굴 영상으로부터 현재 모니터상에 사용자가 쳐다보고 있는 지점을 파악해 내는 기술이다. 이 연구에서는 사용자가 모니터상의 임의의 지점을 쳐다보기 위하여 눈동자보다는 주로 얼굴을 회전(rotation)하거나 이동(translation)하는 경우를 대상으로 하였다. 현재까지 사용자의 응시 위치를 파악하기 위해 많은 알고리듬(선형 보간법, 단일 신경망, 다중 신경망)을 개발하였으며, 이중 다중 신경망(Two Neural Network)에 의한 응시 위치 파악이 가장 우수한 성능을 나타났다(얼굴의 회전만 있는 경우에 1.7 인치 응시 위치 에러, 얼굴의 회전과 이동이 같이 있는 경우에 3.1 인치 응시 위치 에러).

7. 참고문헌

[1] A. Azarbayejani, "Visually Controlled Graphics", in Proc. IEEE PAMI, Vol. 15, No. 6, pp. 602-605, June 1993

[2] Andrew Kiruluta, "Predictive Head Movement Tracking Using Kalman Filter", in Proc. IEEE Trans. on SMC, Vol.27, No.2, April 1997

[3] T. Fukuhara, T. Murakami, "3-D motion estimation of human head for model-based image coding", in IEE Proc. , Vol. 140, No.1, Feb., 1993

[4] T.Rikert, M. Jones, 1998. Gaze Estimation using Morphable Models, Proceedings of the 3th International Conference on Automatic Face and Gesture Recognition, Japan, pp. 436-441.[5] Ramesh Jain, Machine Vision, McGraw-Hill International Edition, 1995

[5] 남시욱, 박강령, 정진영, 김재희, "얼굴의 칼라정보와 움직임 정보를 이용한 얼굴 영역 추출" 1997년 대한전자공학회 하계 종합학술대회 논문집 pp.905-908, 1997년 6월

[6] 남시욱, 박강령, 한승철, 김재희, "다중 모드 인터페이스 환경에서 동가속도 예측 알고리듬을 이용한 얼굴 특징점 추적" 1998년 제 10회 영상처리 및 이해에 관한 워크샵 발표 논문집, pp. 209-214, 1998년 1월

이 연구는 1999년도 한국학술진흥재단의 대학부설연구소 연구비에 의하여 일부 지원되었음