韓國컴퓨터情報學會 論文誌
第16卷 第4號, 2011. 04.

2011-16-4-1-1

MHI의 형태 정보를 이용한 동작 인식

김 상 균*

Gesture Recognition using MHI Shape Information

Sang-Kyoon Kim*

요 약

본 논문에서는 MHI(Motion History Image)의 형태학적 정보를 이용하여 동작을 인식하는 제스처 인식(Gesture Recognition) 시스템을 제안한다. 입력되는 영상으로부터 동작에 관한 정보를 제공하는 MHI를 획득하고, 이 MHI로부터 x, y 각각의 좌표에 대한 기울기(gradient) 영상을 추출한다. 각각의 기울기 영상에 형태 문맥 기법(shape context method)을 적용하여 형태 정보를 추출하고, 추출된 형태 정보 값들을 특징 값으로 사용한다. 이렇게 획득한 특징값들을 최종적으로 SVM(Support Vector Machine) 분류기로 학습 및 분류하여 동작을 인식한다. 제안하는 시스템은 MHI의 형태학적인 정보들을 사용함으로써 동작의 방향성을 인식할수 있고 다수 사람의 동작 인식이 가능하다. 뿐만 아니라 간단한 특징 추출 방법으로 높은 인식률의 시스템을 구현하였다.

▸ 키워드 : 동작인식, 모션 히스토리 이미지, 형태 문맥, 스포트 벡터 머신

Abstract

In this paper, we propose a gesture recognition system to recognize motions using the shape information of MHI (Motion History Image). The system acquires MHI to provide information on motions from images with input and extracts the gradient images from such MHI for each X and Y coordinate. It extracts the shape information by applying the shape context to each gradient image and uses the extracted pattern information values as the feature values. It recognizes motions by learning and classifying the obtained feature values with a SVM (Support Vector Machine) classifier. The suggested system is able to recognize the motions for multiple people as well as to recognize the direction of movements by using the shape information of MHI. In addition, it shows a high ratio of recognition with a simple method to extract features.

▸ Keyword : Gesture Recognition), MHI(Motion History Image), Shape Context, SVM(Support Vector Machine)

• 제1저자 : 김상균
• 투고일 : 2010-09-08, 심사일 : 2010-10-26, 게재확정일 : 2011-01-10
* 인제대학교 컴퓨터공학부(School of Computer Engg., Inje University)
※ 본 논문은 2006년도 인제대학교 학술연구조성비 보조에 의한 것임

Ⅰ. 서 론

최근 컴퓨터 기술이 발달함에 따라 인간들은 컴퓨터로 수많은 정보를 교환하고, 컴퓨터 정보 시스템에 대한 의존도가 높아지고 있다. 따라서 컴퓨터와 사용자간의 상호작용을 위한 다양한 기술들이 연구되고 있다. 그 중에서도 제스처 인식(Gesture Recognition)은 먼 거리나 잡음 환경에서 인간과 컴퓨터간의 정보 전달의 수단이 될 수 있어 컴퓨터 비전(Computer Vision) 분야에서는 핵심 기술 분야로 연구되고 있다. 또한 로봇 산업이 크게 성장함으로써 인간과 로봇과의 자연스러운 상호작용을 위하여 비전 기반 제스처 인식에 대한 관심이 높아졌고, 그에 관한 연구들도 활발히 진행되고 있다[1-3].

기존의 제스처 인식 연구는 인간의 정확한 움직임을 인식하기 위해서 인체의 각 관절 부위에 센서와 같은 기계장치를 부착하여 움직임 정보(Motion Information)를 인식하였다[4-7]. 이 방법들은 인체의 각 관절의 각도와 공간적 위치를 직접 측정하기 때문에 복잡한 계산 없이 정확한 데이터를 얻을 수 있어 가상현실 시스템과 같은 분야에 많이 활용되고 있다. 그러나 이러한 센싱에 기반을 둔 시스템들은 장치를 착용하는 과정이 복잡하고 사용자에게 거부감을 유발할 수 있다.

최근에 들어, 특별한 장치의 부착 없이 자유롭게 동작을 하면서 모션 정보를 획득하는 비전 시스템 기반의 제스처 인식 기술들이 많이 연구되어지고 있다. 이러한 비전 기반 시스템은 센싱 기반 기술에 비해 상대적으로 정밀도가 떨어짐에도 불구하고 동작자에게 장치부착의 불편함과 거부감을 주지 않는다는 이유에서 꾸준히 연구되고 있다.

비전 시스템에서 제스처를 인식하는 연구들은 제스처의 표현 방법에 따라 모델링 기반(Modeling based)의 연구와 외형 기반(Appearance based)의 연구로 분류할 수 있다.

모델링 기반의 연구[8][9]는 여러 개의 카메라를 이용하여 사람의 신체 구조를 2차원이나 3차원으로 모델링을 한다. 그리고 이 모델링으로부터 움직임의 정보를 추출하고 분석해서 동작을 인식하는 방법이다. 이 방법은 정밀한 인식 결과를 기대할 수 있으나, 많은 양의 데이터 처리와 높은 구현 난이도, 제한된 환경에서의 실험 및 많은 비용이 요구되어 실세계 적용이 쉽지 않다.

외형 기반의 연구에는 모션 히스토리(Motion History) 정보를 이용하는 방법[10][11]과 실루엣 템플릿(Silhouette template)을 이용하는 방법[12][13]이 있다. 모션 히스토리는 움직임이 일어나는 흔적을 축적하여 하나의 영상으로 나타내어 정보를 이용하는 방법이고, 실루엣 템플릿은 특정 동작의 외형을 템플릿으로 만들어 입력영상에서 이와 같은 실루엣을 찾는 방법이다. 이 방법들은 객체를 모델링을 하지 않고도 좋은 인식 결과를 얻을 수 있지만, 실루엣의 모양과 조명, 복잡한 배경 등과 같은 주변 환경과 조건에 민감하여 결과에 많은 영향을 받을 수 있다는 단점이 있다.

본 논문에서는 별도의 장치 없이 일상생활에서 사용하고 있는 캠 카메라를 이용하여 사람의 제스처를 인식하는 것에 중점을 두고 연구를 진행하였으며, 복잡한 모델링 과정을 거치지 않고 보다 쉽게 움직임의 정보를 얻을 수 있는 외형 기반의 연구의 방법인 MHI를 특징으로 사용하였다. 제안하는 제스처 인식 시스템에 사용된 특징값은 입력되는 영상으로부터 MHI를 추출하고, 이 MHI에 x와 y좌표의 각각의 변화를 이용하여 기울기 영상[14]을 구한다. 이 각각의 기울기 영상에 형태 문맥(shape context)[15]을 적용하여 형태 정보를 추출함으로써 특징값을 획득할 수 있다. 분류 및 인식을 위하여 SVM을 이용한다.

본 논문에서 제안하는 제스처 인식 시스템은 기존의 제스처 인식 연구들에서 해결해야 할 과제중의 하나인 다수 사람의 동작 인식이 가능하고, 사람의 행동에 있어서 중요한 방향성을 정확하게 인식하는 시스템을 구현하였다. 뿐만 아니라 특징을 간단하게 추출함에도 불구하고 인식률이 상당히 높은 시스템을 구현할 수 있었다.

Ⅱ. 제스처 인식 시스템

본 논문은 입력되는 사용자의 동작을 인식하고 분석하는 제스처 인식 시스템을 제안한다.

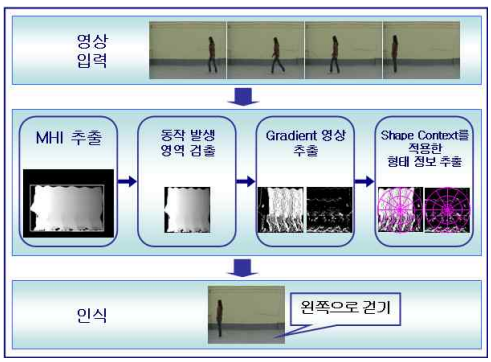


그림 1. 동작 인식 시스템
Fig. 1. Main structure for gesture recognition

본 논문에서 제안하는 제스처 인식 시스템의 전체 구성 절차는 그림 1과 같다. 실시간 입력되는 320×240 크기의 영상으로부터 동작의 히스토리 모션 정보를 나타내는 MHI를 추출한다. MHI를 추출함으로써 동작에 대한 시간과 공간적인 특징들을 획득할 수 있고, 동작이 발생한 영역을 검출해 낼 수 있다. 동작 발생 영역의 검출은 동영상에 여러 사람이 등장하여 동작을 수행하여도 각각의 동작 영역을 검출해 낼 수 있어서 다수 사람의 동작 인식이 가능하다.

III. 특징 추출(Feature Extraction)

1. MHI(Motion History Image) 추출

1.1 MHI

먼저 입력되는 영상으로부터 MHI를 추출한다. MHI는 공간상과 시간상에서의 동작 변화 정도를 표현한 명암도 영상(gray image)으로, 점점 밝아지는 영상의 명암은 동작 발생 시간에 비례하는 특징을 가지고 있다. MHI를 표현함으로써 동작이 어디서 발생했으며, 어느 방향으로 얼마의 시간만큼 진행하였는지의 시공간적인 정보를 획득할 수 있다.

$$H_\tau(x,y,t)=\begin{cases}\tau & \text{if } D(x,y,t)=1\\ \max(0,H_\tau(x,y,t-1)-1) & otherwise\end{cases} \quad \cdots\cdots (1)$$

- $H_\tau(x,y,t)$: MHI
- $D(x,y,t)$: 프레임 사이의 이진영상 (차영상)
- τ : 움직임 지속 기간

식(1)은 A. F. Bobick과 J. W. Davis[10]에 의해 정의된 MHI이다. 본 논문의 연구에서는 명암에 대한 의미를 부각시키기 위해 식(1)을 다음 식(2)와 같이 수정하여 사용한다.

$$H_\tau(x,y,t) = \begin{cases}\min(255,\tau\times k) & \text{if } D(x,y,t)=1\\ H_\tau(x,y,t-1) & otherwise\end{cases} \quad \cdots\cdots (2)$$

- $H_\tau(x,y,t)$: MHI
- $D(x,y,t)$: 프레임 사이의 이진영상 (차영상)
- τ : 움직임 지속 기간, k : 상수

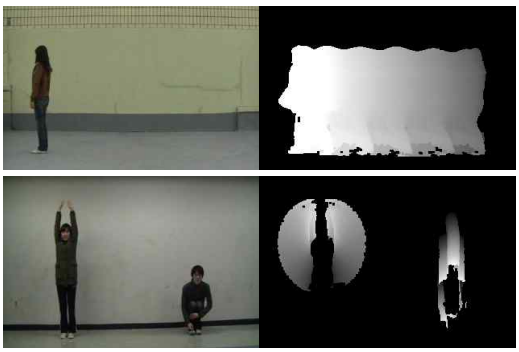


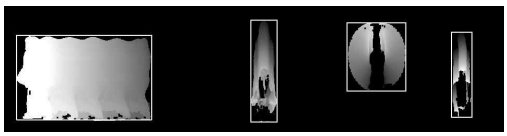
그림 2. MHI 추출
Fig. 2. MHI extraction

그림 2는 식(2)에 의해서 MHI가 만들어지는 과정을 보여준다. MHI는 프레임 사이의 이진 차영상을 이용하므로 구현은 다소 간편하나, 조명이나 복잡한 배경에 민감하여 사람의 동작 영역이 아닌 부분에서도 원하지 않는 잡음이 생길 수 있다. 또한, 카메라의 움직임에 민감하게 반응하기 때문에 정확한 결과를 얻기 위해서는 반드시 고정된 카메라를 사용하여야만 하는 제약이 따른다. 본 논문의 실험에서는 고정된 카메라를 사용하고 모폴로지(morphology)를 적용하여 잡음을 제거한 MHI를 얻었다.

1.2 동작 발생 영역 검출 및 정규화

추출된 MHI를 그림 3(a)와 같이 동작 발생 영역을 따로 검출 한다. 동작 영역을 검출함으로써 한 사람이 동작을 수행했을 때뿐만 아니라 다수의 사람이 동작을 수행하더라도 동작 영역을 각각 검출하여, 검출된 동작 영역을 개별적으로 인식하는 다수 사람의 동작 인식이 가능하다.

본 논문에서는 일정한 실험 조건을 위해서 다양한 크기의 동작 발생 영역들을 그림 3(b)와 같이 160×160으로 정규화를 한다.



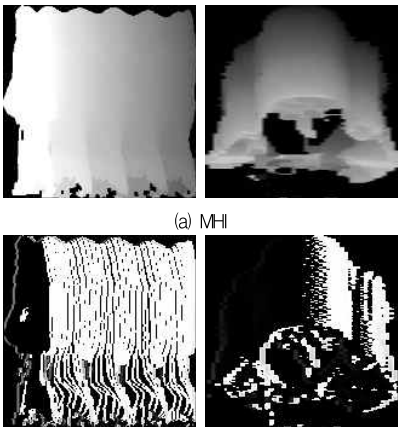
(a) 동작 영역 검출(Gesture region detection)

(b) 정규화(Normalization)
그림 3. 동작 영역 검출 및 정규화
Fig. 3. Detection of a gesture region and normalization

2. 기울기(Gradient) 영상 추출

본 논문에서 기울기 영상은 형태 문맥 기법을 적용하여 특징을 추출하는 과정의 입력 영상으로 사용한다. 선행 단계에서 얻어진 MHI 정규화 영상에서 공간상의 x, y 좌표의 변화 정도를 나타내는 기울기 영상을 추출하기 위해 Zelnik[16]이 제안한 지역 화소값 기울기(local intensity gradient) 측정 알고리즘을 사용한다. 단, 본 논문의 연구에서는 MHI에 대한 공간적인 정보를 측정하기 위해 기울기를 추출하는 것임으로 Zelnik이 제안한 지역 화소값 기울기에서 시간의 정보를 제외한 공간상의 기울기만을 추출한다. 공간상의 기울기는 다음 식(3)과 같이 정의 되고 그림 4와 같이 표현된다.

$$S_x = \frac{S(x+1, y) - S(x-1, y)}{2}, \quad S_y = \frac{S(x, y+1) - S(x, y-1)}{2} \quad \cdots\cdots (3)$$



(a) MHI

(b) x좌표의 변화를 나타낸 S_x기울기 영상
(S_x gradient images presenting a change of x-coordinate)

(c) y좌표의 변화를 나타낸 S_y기울기 영상
(S_y gradient images presenting a change of y-coordinate)

그림 4. 기울기 영상
Fig. 4. gradient image

그림 4와 같이 MHI에서 공간상의 지역 화소값 기울기 정보를 추출한 결과, 아래의 표 1, 그림 5와 같은 방향성의 정보를 획득 할 수 있다.

표 1. 지역 화소값 기울기 결과에 따른 방향성 정보
Table 1. Direction information according to local intensity gradient result

지역 화소값 기울기의 결과	방향	기울기 영상에서의 표현
$S_x > 0$	왼쪽 → 오른쪽	회색 (gray)
$S_x < 0$	오른쪽 → 왼쪽	흰색 (white)
$S_y > 0$	위 → 아래	회색 (gray)
$S_y < 0$	아래 → 위	흰색 (white)

표 1을 보면, 왼쪽에서 오른쪽 방향으로 움직이는 동작은 S_x의 결과가 양수로 나타났고 S_x 기울기 영상에서 회색(gray)으로 표현이 된다. 반대로 오른쪽에서 왼쪽 방향으로 움직이는 동작은 S_x의 결과가 음수로 나타났고 S_x 기울기 영상에서 흰색(white)으로 표현이 된다. 또한, 위에서 아래 방향으로 움직이는 동작은 S_y의 결과가 양수로 S_y 기울기 영상에서는 회색으로 표현되고, 아래에서 위 방향으로 움직이는 동작은 S_y의 결과가 음수로 S_y 기울기 영상에서는 흰색으로 표현이 된다. 이 4가지의 방향성을 조합하면 그림 5에서와 같이 총 8개의 방향 정보를 획득할 수 있다.

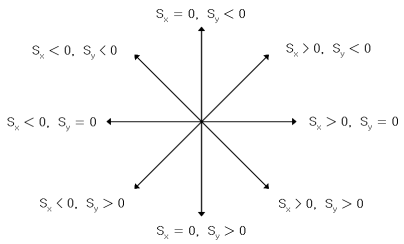


그림 5. 지역 화소값 기울기(local intensity gradient) 결과에 따른 방향성 정보
Fig. 5. Information of direction according to a result of local intensity gradient

3. 형태 정보를 이용한 특징 추출

3.1 형태 문맥(Shape context)를 이용한 형태 정보 추출

형태 문맥은 S. Belongie와 J. Malik[15]가 제안한 특징 서술자(feature descriptor)이다. 형태 문맥은 형태적 유사성의 척도로서 객체 인식에 유용하게 활용될 수 있다. 기본적인 아이디어는 객체의 윤곽선상에서 특정 위치에 해당하는 포인트들의 개수를 계산하여 형태 정보로 활용하는 것이다. 이 방법으로 추출된 형태 정보는 관련된 위치 분포에 강인하며, 사용이 간단하면서도 높은 성능을 보여준다.

그림 6은 형태 문맥을 계산하는 과정을 순서대로 보여준

다. 그림 6의 (a)와 (b)는 형태 정보로 추출된 에지 포인트들의 예이며, (c)는 형태 문맥 계산에 사용된 로그폴라(log-polar) 다이어그램이다. 계산된 결과는 (d), (e), (f)와 같이 각도(θ)로 나누어진 조각(bin)들과 반지름($\log r$)으로 나누어진 조각(bin)들로 조합된 패턴인 형태 문맥이 된다.

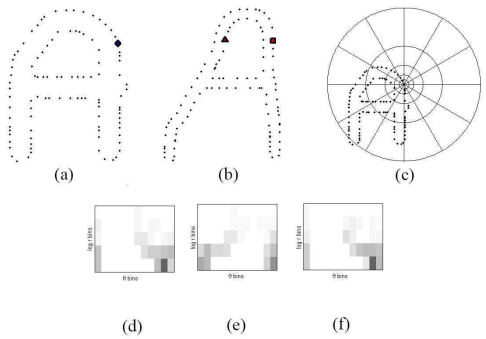
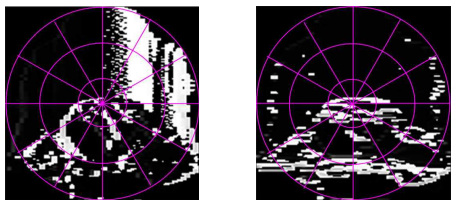


그림 6. 형태 문맥 계산
Fig. 6. Operation of shape context

각도와 반지름을 조각내는 개수는 많으면 많을수록 세밀한 특징을 추출할 수 있으나 계산 양이 증가하여 속도가 느려지고, 개수가 작으면 속도는 빠르나 특징 추출이 둔해짐을 고려하여 실험에 맞는 적절한 개수를 선택하도록 한다.

3.2 특징값 패턴 생성

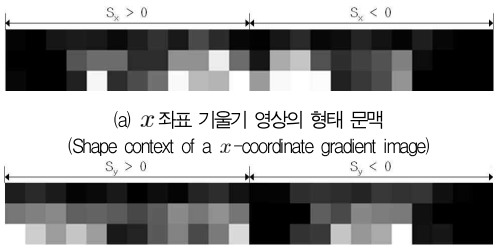
본 논문에서는 이전 단계에서 계산된 x, y 각각의 기울기 영상에 그림 6의 (c)와 같이 각도 θ는 12개의 조각으로, 반지름 r은 3개의 조각으로 나누어 총 36개의 조각들을 가지는 로그폴라 다이어그램을 생성하여 적용한다. 그림 7은 각 기울기 영상에 내접하는 로그폴라 다이어그램을 적용한 예이다.



(a) x좌표(x-coordinate) (b) y좌표(y-coordinate)
그림 7. 각 기울기 영상에 적용한 로그폴라 다이어그램
Fig. 7. Log-polar diagram adapted to each gradient image

각각의 bin에서 흰색 포인트의 개수($S_x < 0,\ S_y < 0$)와 회색 포인트의 개수($S_x > 0,\ S_y > 0$)는 동작의 특징값으로 사용된다. 즉, 36개의 조각(bin)들을 가지는 형태 문맥이 x좌표 기울기 영상에서 2개, y좌표 기울기 영상에서 2개가 생성되어 총 144개의 특징 값들을 학습패턴으로 구성한다.

그림 8은 그림 7의 각각의 로그폴라 다이어그램에서 형태 문맥을 계산하여 나타낸 것이다. 밝기가 밝을수록 많은 포인트들이 존재함을 나타낸다.



(a) x 좌표 기울기 영상의 형태 문맥
(Shape context of a x-coordinate gradient image)

(b) y좌표 기울기 영상의 형태 문맥
(Shape context of a y-coordinate gradient image)
그림 8. 형태 문맥 영상
Fig. 8. Shape context image

4. 분류 및 인식

4.1 SVM 분류기

SVM은 1995년에 Vapnik에 의해 제안되었고 VC(Vapnik-Chervonenkis)이론에 근간을 두고 있으며, 뛰어난 일반화 성능을 보여준다. SVM은 원래 이원 패턴인식 문제를 해결하기 위한 방법으로써 구조적 에러를 최소화하는 기법이다. 기존의 경험적 에러 최소화 기법인 다층신경망과 비교하여 학습에 필요한 파라미터 일부가 자동적으로 결정된다. 기본적인 아이디어는 두 클래스에 속한 데이터의 집합이 있을 경우, SVM은 두 클래스 간의 거리를 최대화하여 두 클래스를 분류하는 최적의 초평면을 찾는 것이다.

선형분리 가능한 이진 클래스의 경우, 입력 값들을 다른 클래스간에 데이터 사이를 최대거리로 분리할 수 있는 초평면(Hyperplane)을 찾기 위해서 고차원의 특징 공간(Feature space)으로 변환시킨다. 두 클래스 군집을 선형 분리하는 초평면과 가장 가까운 점을 'Support Vector(SV)'라고 한다. 그리고 SV와 결정 평면간의 거리를 '마진(Margin)'이라고 한다.

식(4)는 특징값들을 모아놓은 벡터로, SVM 분류기의 입력 벡터가 된다.

$$\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{iM}]^T, \quad 1 \le i \le l \quad \cdots\cdots (4)$$

총 차원 수 M은 특징값의 개수인데, 본 논문에서의 특징

값은 72개의 조각으로 이루어진 형태 문맥이 x좌표 기울기 영상, y좌표 기울기 영상에서 각각 생성되어 총 144차원으로 구성된다. l은 인식하고자 하는 동작들에 대해서 수집한 샘플들의 총 개수이다.

특징값 $\mathbf{x}$를 사용자의 특정한 동작으로 결정하기 위해 특징공간에서 초평면과 특징값의 구조적인 관계를 생각할 수 있다. 최적의 결정함수를 식(5)라 할 때, $\mathbf{u}$는 결정 경계인 초평면의 단위 방향 벡터이고, 원점에서 초평면을 수직으로 가리키는 방향 벡터는 $\mathbf{u}^T\mathbf{p}$이다. 원점과 초평면 사이의 거리를 D_u라고 할 때 초평면과 특징값 $\mathbf{x}$와의 거리 $D_{\mathbf{x}}$는 식(9)와 같다.

$$f_{\mathbf{w},b} = \mathrm{sgn}\left(\mathbf{w}^T\mathbf{x} + b\right)\} \quad \cdots\cdots (5)$$

$$\mathbf{u} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \quad \cdots\cdots (6)$$

$$\mathbf{u}^T\mathbf{p} = -\frac{b}{\|\mathbf{w}\|} \quad \cdots\cdots (7)$$

$$D_u = \frac{|b|}{\|\mathbf{w}\|} \quad \cdots\cdots (8)$$

$$D_{\mathbf{x}} = |\mathbf{u}^T\mathbf{x} - \mathbf{u}^T\mathbf{p}| \quad \cdots\cdots (9)$$

$$= \left| \frac{|\mathbf{w}^T\mathbf{x}|}{\|\mathbf{w}\|} + \frac{b}{\|\mathbf{w}\|} \right|$$

$$= \left| \frac{\mathbf{w}^T\mathbf{x} + b}{\|\mathbf{w}\|} \right|$$

최대의 마진을 가지는 초평면을 구하기 위해 $D_{\mathbf{x}}$의 최대값을 구해야 한다. 따라서 분모 $\|\mathbf{w}\|$이 최소가 될 때 $D_{\mathbf{x}}$의 최대값을 구할 수 있다. 최대의 마진을 구하기 위한 식 $\tau(\mathbf{w})$는 식(10)과 같다. 뿐만 아니라 데이터를 정분류하기 위한 조건식(11)을 만족하여야 한다. y_i는 목표 값으로 1과 -1을 나타낸다.

$$\tau(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \quad \cdots\cdots (10)$$

$$y_i \cdot \left((\mathbf{w} \cdot \mathbf{x}_i) + b\right) \geq 1, \quad i = 1, \cdots, l \quad \cdots\cdots (11)$$

식(11)의 제약조건을 만족하면서 동시에 식(10)의 최소값을 구하는 문제는 라그랑즈 승수(Lagrange Multiplier) α를 사용한 식(12)로 표현된다.

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i \left(y_i \cdot ((\mathbf{x}_i \cdot \mathbf{w}) + b) - 1\right) \quad \cdots\cdots (12)$$

$\tau(\mathbf{w})$ 및 $y_i \cdot \left((\mathbf{w} \cdot \mathbf{x}_i) + b\right) \geq 1$가 미분 가능하다고 할 때 조건식 $y_i \cdot \left((\mathbf{w} \cdot \mathbf{x}_i) + b\right) - 1 = 0$을 만족하는 점들에 대한 함수 $\tau(\mathbf{w})$의 극소값을 구하기 위해 식(13)과 같이 편미분 하여 극값을 갖는 조건식(14)를 구한다.

$$\frac{\partial}{\partial b^*} L(\mathbf{w}^*, b^*, \alpha) = 0, \quad \frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}^*, b^*, \alpha) = 0 \quad \cdots\cdots (13)$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \qquad \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i = \mathbf{w}^* \quad \cdots\cdots (14)$$

식(14)의 관계를 만족 할 때 최대의 마진을 가지는 초평면을 구할 수 있다. 이것을 식(5)에 대입하여 다음 식(15)와 같은 결정함수를 얻는다.

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i (\mathbf{x}_i \cdot \mathbf{x}) + b^* \right) \quad \cdots\cdots (15)$$

본 논문에서는 마진을 최대로 하는 결정함수 $f(\mathbf{x})$를 이용하여 사용자의 행동을 분류한다.

IV. 실험 결과

1. 실험 환경 및 대상

1.1 실험환경

본 논문에서 제안하는 제스처 인식 시스템의 실험은 표 2와 같은 PC 환경에서 구현 및 수행하였다.

표 2. 실험 PC 환경
Table 2. Experiment PC environment

CPU	2.4GHz Pentium4-PC
운영체제	Microsoft Windows XP
프로그래밍 도구	Microsoft Visual C++ 6.0 Intel OpenCV 라이브러리

1.2 인식 대상 동작의 종류

본 논문의 실험에서 인식하고자 하는 동작은 그림 9와 같이 왼쪽으로 걷기(Walk to Left), 오른쪽으로 걷기(Walk to Right), 앉기(Sit Down), 일어서기(Stand Up), 팔 올리기(Hands Up), 팔 내리기(Hands Down), 왼쪽으로 달리기(Run to Left), 오른쪽으로 달리기(Run to Right), 바닥의 물건을 줍는 동작과 비슷한 구부리기(Bend)로 총 9개의 동작이다.

(a) 왼쪽으로 걷기(Walk to Left)

(b) 오른쪽으로 걷기(Walk to Right)

(c) 앉기(Sit Down)

(d) 일어서기(Stand Up)

(e) 팔 올리기(Hands Up)

(f) 팔 내리기(Hands Down)

(g) 왼쪽으로 달리기(Run to Left)

(h) 오른쪽으로 달리기(Run to Right)

(i) 구부리기(Bend)

그림 9. 인식대상 동작들

Fig. 9. Recognized gesture of objects

방향성 인식을 위해 걷기 동작과 뛰기 동작은 오른쪽과 왼쪽으로 나누어서, 앉기와 일어서기, 팔 올리기와 팔 내리는 동작 모두 쌍으로 인식 대상 동작에 포함하였다. 또한 걷기와 달리기, 그리고 앉기와 구부리기 같은 유사한 동작도 인식이 가능함을 보여주기 위해 인식 대상 동작에 포함하였다. 또한 배경 색과 비슷한 옷을 입은 실험(a),(e)와, 복잡한 배경에서의 실험(i), 동작이 영상의 가운데에서 일어나지 않는 실험(c)등 다양한 실험을 진행하였다.

실험 및 테스트에 사용한 영상은 일상생활에서 흔히 사용하는 웹캠을 이용하여 획득하였다.

2. 학습(Training)

2.1 학습 패턴 구성

본 논문에서는 SVM을 이용하여 학습을 하였다. 그림 10은 학습 패턴을 구성하는 과정의 예를 보여준다. 먼저, 학습을 위해 그림 9에서 제안한 총 9개의 인식 동작들에 대해서, 각각 10개씩의 영상(샘플)들을 수집하였다. 총 90개의 영상들마다 3.3.2장의 그림 8과 같이 144개의 특징값을 추출하여 학습 데이터를 생성한다. 이 학습 데이터는 144차원의 특징값을 가지는 90개의 샘플로 구성되어 SVM의 입력 벡터로 사용된다.

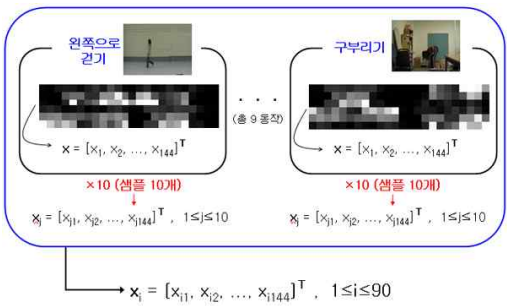
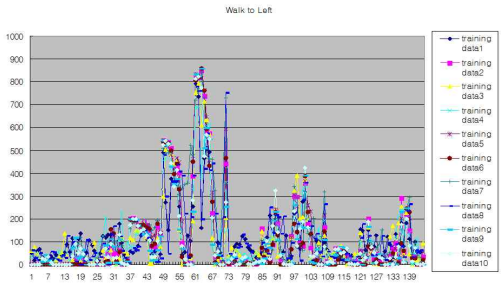


그림 10. 학습 패턴 구성 과정

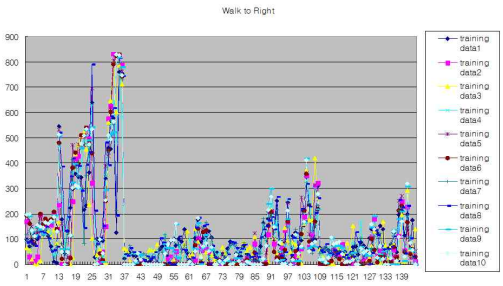
Fig. 10. A constructing process of a training pattern

2.2 학습 데이터 분석

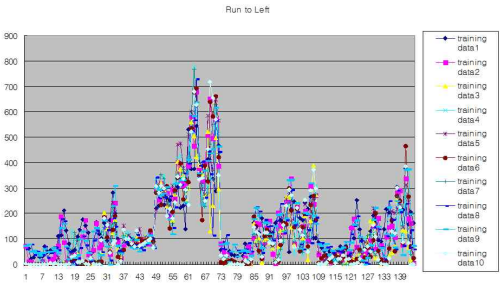
본 논문의 실험은 인식하고자 하는 9개의 각 동작마다 10개의 샘플을 수집하여 SVM을 이용하여 학습을 하였다. 다중 인식을 위하여 one-others model을 사용하였다. 하나의 행동인식을 위한 동작과 나머지 8가지의 동작데이터를 2가지 클래스의 학습데이터로 각각 구성하여 학습하였고 9개의 동작인식을 위해 9개의 SVM분류기를 구성하였다. 그림 11은 본 논문에서 사용하는 일부 동작 학습 데이터의 특징들의 분포를 나타낸 그래프들이다.



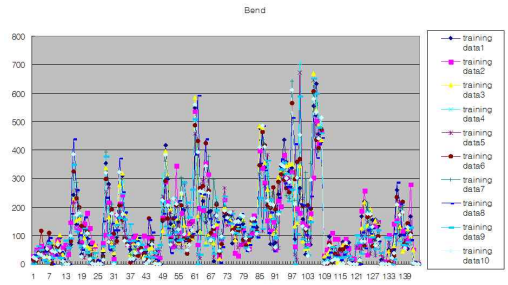
(a) 왼쪽으로 걷기(Walk to Left)



(b) 오른쪽으로 걷기(Walk to Right)



(c) 왼쪽으로 뛰기(Run to Left)



(d) 구부리기(Bend)

그림 11. 학습 데이터의 특징 분포 그래프
Fig. 11. Feature graph of training data

그림 11의 그래프들을 보면, 각 동작마다 특정한 특징 패턴이 있음을 확인 할 수 있다. 이를 이용해서 방향과 속도에 따른 동작들을 분류할 수 있다.

3. 실험 결과 및 분석

3.1 실험 결과 및 분석

본 논문의 실험은 학습 샘플 10개와 한 사람이 한 개의 동작을 수행하는 영상 10개 이상, 한 사람이 여러 개의 동작을 수행하는 영상 11개와 두 사람이 동작을 수행하는 영상 12개를 테스트 하였다.

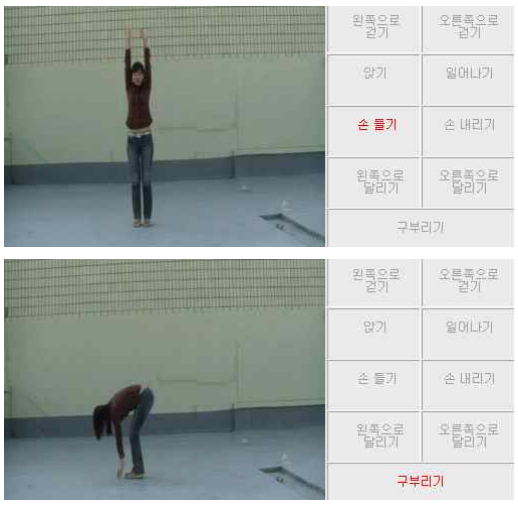


그림 12. 동작 인식 결과 화면
Fig. 12. Result image of gesture recognition

그림 12는 한 사람이 여러 개의 동작을 수행하였을 때의 결과 화면이다.

그림 13. 다수 사람의 동작 인식 결과 화면
Fig. 13. Result image of gesture recognition of multi persons.

그림 13은 다수 사람의 동작 인식의 결과 화면이다. 다음 표 5는 두 사람이 한 개의 동작을 수행하였을 때(a)와 여러 개의 동작을 수행하였을 때(b)를 테스트한 결과이다.

표 3은 한 사람이 여러 개의 동작을 수행하는 영상 11개에 대한 테스트를 한 결과이다. 영상 10에서 오른쪽으로 걷기를 오른쪽으로 달리기로 올바르지 못한 인식이 발생하였으나 나머지 모든 동작들에 대해서는 올바른 인식을 하였다.

표 3. 한 사람이 여러 개의 동작을 수행하는 영상 11개의 테스트 결과
Table 3. The test result of 11 videos that one person make lots of gestures

		WL	WR	SD	SU	HU	HD	RL	RR	Bend
영상1	동작			O	O	O	O			
	인식			SD	SU	HU	HD			
영상2	동작	O	O			O	O			
	인식	WL	WR			HU	HD			
영상3	동작	O		O	O		O			
	인식	WL		SD	SU		HD			
영상4	동작				O				O	O
	인식				SU				RR	Bend
영상5	동작			O	O				O	O
	인식			SD	SU				RR	Bend
영상6	동작	O	O							O
	인식	WL	WR							Bend
영상7	동작		O	O	O	O	O	O		
	인식		WR	SD	SU	HU	HD	RL		
영상8	동작				O	O	O			O
	인식				SU	HU	HD			Bend
영상9	동작	O	O	O	O	O	O			
	인식	WL	WR	SD	SU	HU	HD			
영상 10	동작	O	O	O	O	O	O			
	인식	WL	RR	SD	SU	HU	HD			
영상 11	동작			O	O			O		O
	인식			SD	SU			RL		Bend
정인식		5/5	4/5	7/7	9/9	6/6	7/7	2/2	2/2	5/5
오인식			1/5							

표 4. 두 사람이 동작을 수행하는 영상 17개의 테스트 결과
Table 4. Test result of 17 videos that two persons does gesture

(a) 두 사람이 한 개의 동작을 수행한 결과
(The test result that two person makes one gesture)

	동작 및 인식			정인식	오인식
영상 1	동작	SD	RL	2/2	
	인식	SD	RL		
영상 2	동작	HU	WL	2/2	
	인식	HU	WL		
영상 3	동작	WR	Bend	2/2	
	인식	WR	Bend		
영상 4	동작	HU	HU	2/2	
	인식	HU	HU		
영상 5	동작	HU	SD	2/2	
	인식	HU	SD		
영상 6	동작	SD	SD	2/2	
	인식	SD	SD		
영상 7	동작	HU	WR	1/2	1/2
	인식	HU	RR		
영상 8	동작	WR	SD	1/2	1/2
	인식	WR	WR		
영상 9	동작	HU	WR	2/2	
	인식	HU	WR		
영상10	동작	WR	Bend	1/2	1/2
	인식	WR	SD		

(b) 두 사람이 여러 개의 동작을 수행한 결과
(The test result that two person makes lots of gestures)

	동작 및 인식					정인식	오인식
영상11	동작	SD	HU	SU	HD	4/4	
	인식	SD	HU	SU	HD		
영상12	동작	HD	SD	Bend	SU	4/4	
	인식	HD	SD	Bend	SU		
영상13	동작	SU	SU	HU	Bend	4/4	

	인식	SU	SU	HU	Bend		
영상14	동작	HU	WR	HD	HU	2/4	2/4
	인식	RR	RR	HD	HU		
영상15	동작	HD	SD	Bend	WR	3/4	1/4
	인식	HD	SD	Bend	RR		
영상16	동작	SU	SU	WR	Bend	2/4	2/4
	인식	SU	SU	RR	SD		
영상17	동작	HU	SU	HD	HU	4/4	
	인식	HU	SU	HD	HU		

표 4의 (a)의 결과를 살펴보면, 영상 7에서 오른쪽으로 걷기를 오른쪽으로 달리기로 잘못 인식하였다. 이는 빠르게 걷는 경우와 뛰는 경우 판단하기 모호하기 때문이다.

또한, 표 4의 (a)에서, 영상 8은 구부리기 동작을 인식하지 못하였다. 영상 8은 그림 14와 같이 한 사람이 먼저 동작을 수행하고, 다른 사람이 나중에 동작을 수행하는 영상이다. 영상 15의 경우 손올리기와 오른쪽으로 걷기 동작이 두 사람이 서로 겹치면서 인식하지 못하였다.

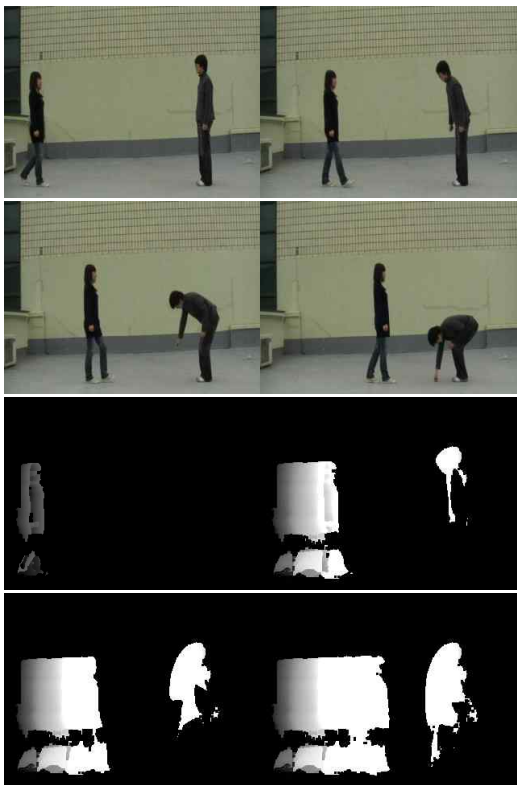


그림 14. 두 사람의 동작 시간이 차이가 날 경우
Fig. 14. Case that gesture time of two persons has a gap

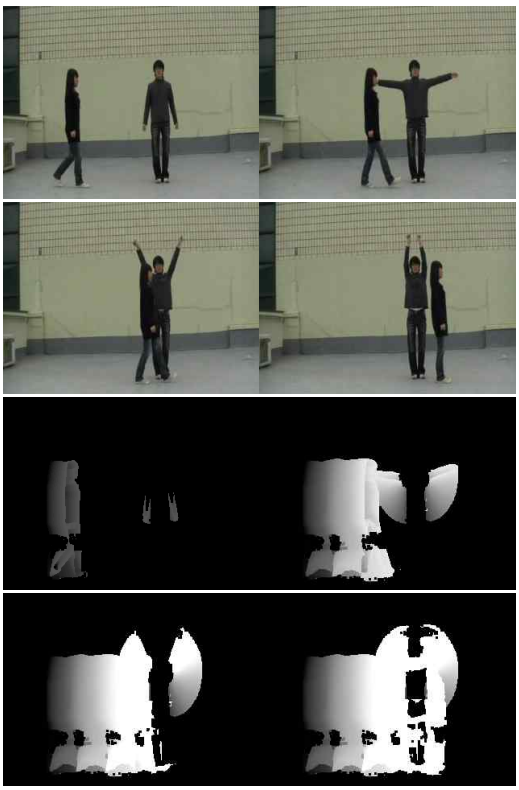


그림 15. 두 사람의 동작 영역이 겹쳐지는 경우
Fig. 15. Case that gesture regions of two persons overlap

한 사람이 동작할 때와 다수의 사람이 동작할 때와의 비교 실험 결과 본 논문에서는 그림 3(a)와 같이 동작 발생 영역을 따로 검출하기 때문에 다수의 사람이 동작을 수행하더라도 동작 영역을 각각 검출하여, 검출된 동작 영역을 개별적으로 인식한다. 그래서 다수 사람이 동작을 수행하더라도 검출된 영역에서 동작 인식이 가능하다. 하지만 그림 14와 같이 여러 사람의 동작 시작시간이 다를 경우와 그림 15와 같이 여러 사람의 동작 시간이 다를 경우 한 사람이 동작을 시행할 때 보다 인식률이 떨어진다.

실내에서 단순한 배경과 복잡한 배경의 실험 그리고 실외에서의 실험 결과 실내에서는 배경의 단순함과 복잡함의 경우 실험에 크게 영향을 미치지 않았다. 하지만 그림 16에서 보는 바와 같이 배경(가방)의 명도 값이 움직이는 사람의 명도 값과 유사할 경우 동작영역에 잡음이 발생해서 동작을 제대로 검출하지 못하는 문제점이 발생했다. 그리고 실외 환경의 경우 배경의 복잡도와 함께 그림 17에서 보는바와 같이 바람으로 인해 주변 나뭇가지 등이 흔들렸을 때 흔들리는 나뭇가지를 동작영역으로 잘못 인식해서 인식률이 크게 저하되는 문제가 발생했다.

그림 16. 실내 환경에서의 동작 인식
Fig 16. Gesture recognition in indoor environment

그림 17. 실외 환경에서의 동작 인식
Fig 17. Gesture recognition in outdoor environment

3.2 기존 연구와의 비교 실험

본 논문에서 제안하는 연구 방법은 Zelnik이 제안한 방법, 특징 추출 방법, 다른 분류기를 사용해서 비교 및 실험을 하였다. Zelnik가 제안한 방법은 입력되는 영상에서 각 프레임마다 지역 화소값 기울기(local intensity gradient)를 추출하여 정규화 과정 후, 클러스터링을 하여 동작을 인식하는 방법이다. Zelnik는 추출한 지역 화소값 기울기를 정규화하면서 방향성의 정보를 상실하여 방향을 고려하지 않는다. 즉, Zelnik이 제안하는 방법에서는 왼쪽으로 걸어가는 동작과 오른쪽으로 걸어가는 동작을 같은 동작이라고 인식한다. 본 논문에서는 일상생활에서 사람에 의해 일어나는 동작들은 방향성이 중요하다고 판단하여 방향을 인식하는 연구 방법을 제안하였다. 그림 18은 방향성을 판단하는 영상이다.

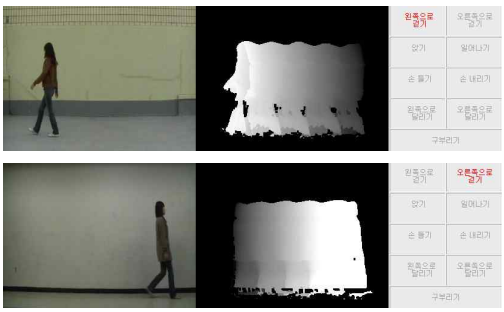


그림 18. 본 논문에서 제안하는 방향성 인식
Fig. 18. Direction recognition proposed in this paper

또한, Zelnik이 제안하는 방법은 영상이 입력되면 각 프레임마다 모든 픽셀을 사용하여 특징을 추출하기 때문에 다수의 사람이 수행하는 동작을 개별적으로 인식하는 것이 불가능하다. 본 논문에서는 MHI를 이용하여 동작 발생 영역을 따로 검출할 수 있기 때문에 다수의 사람의 동작을 개별적으로 인식이 가능하다. 하지만 본 논문의 경우에도 다수 사람의 동작 발생시간이 다를 경우 오인식하는 문제가 발생한다. 그림 19는 다수 사람의 동작을 인식하는 영상이다.

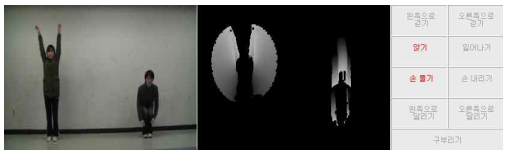
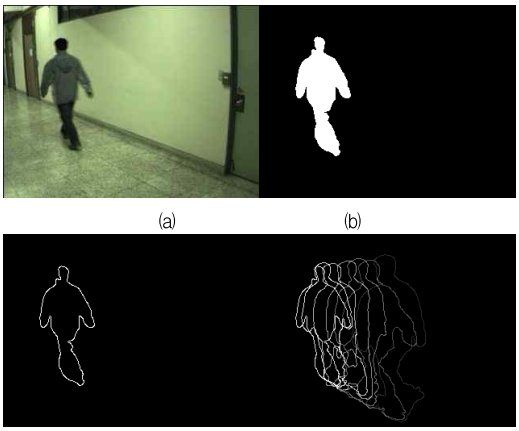


그림 19. 본 논문에서 제안하는 다수 사람의 동작 인식
Fig. 19. Gesture recognition of multi persons proposed in this paper

특징 추출 방법에 따른 비교 실험에서는 실루엣 추출후 동작 인식하는 방법과 비교 실험을 실시하였다. 먼저 실루엣을 만들기 위해서 초기 배경을 만들고 그림 20(b)에서 보는 바와 같이 배경에서 객체를 추출한 다음 canny-edge를 이용해서 그림 20(c)와 같이 최종적인 실루엣을 추출하였다.[17] 다음으로 실루엣에 수식(2)를 적용해서 그림 20(d)와 같은 실루엣 영상을 추출하였다. 이후의 실험내용은 본 논문에서 제안한 방법을 그대로 적용했다.



(a) (b)

(c) (d)

그림 20. 실루엣 검출
Fig. 20. Extraction silhouette

실험결과 표 5에서 보는 바와 같이 제안하는 방법에 비해 인식률이 떨어짐을 알 수 있다. 이는 실루엣 추출 단계에서 색상 값을 이용함으로 인해 배경의 색상 값에 영향을 많이 받아서 실루엣을 제대로 추출할 수 없어서 발생한다.

표 5. 특징 추출법에 따른 인식률 비교
Table 5. Comparison of recognition rate according to a feature extraction method

실험환경	제안방법	실루엣
학습데이터	100%	100%
한 사람이 한개의 동작	100%	95%
한 사람이 여러개의 동작	98%	94%
두 사람이 한개 동작	85%	75%
두 사람이 여러개 동작	82%	71%

다른 분류기와의 비교 실험에서는 신경망 알고리즘을 사용해서 비교실험을 실시하였다. 신경망 알고리즘으로는 BP(back propagation)알고리즘을 사용하였다. 실험을 위해 SVM 입력값과 같이 144개의 특징 값을 분류기의 입력 값으로 사용하였다. 그리고 본 연구에 사용된 신경망의 구조는 144개의 입력노드로 구성된 입력층과 49개의 은닉노드로 구성된 은닉층, 그리고 분류할 동작의 개수인 9개의 출력노드로 구성된 출력층으로 설계하였다. 신경망의 학습 에러 값(minimum error)이 0.1이 될 때까지 연결강도(weight)를 조정하여 학습하였다.

표 6. 분류기에 따른 인식률 비교
Table 6. Comparison of recognition rate according to a classifier

실험환경	SVM	신경망
학습데이터	100%	100%
한 사람이 한개의 동작	100%	100%
한 사람이 여러개의 동작	98%	96%
두 사람이 한개 동작	85%	85%
두 사람이 여러개 동작	82%	79%

실험 결과 표 6에서 보는 바와 같이 인식률에서는 큰 차이를 보이지 않았다. 하지만 신경망 분류기의 경우 학습시간이 많이 걸리고 학습시에 지역 최소 값(local minimum)에 자주 빠지는 문제가 발생했다. 이는 입력 값으로 사용된 144개의 많은 특징값과 49개의 은닉노드, 그리고 9개의 출력 노드로 인해 계산량이 많아져서 발생하는 문제로 생각된다.

V. 결론

논문에서는 MHI의 형태학적 정보를 이용하여 제스처를 인식하는 방법을 제안하였다. 본 논문에서 제안한 방법은 MHI를 이용하여 형태 정보를 추출할 뿐만 아니라 동작 영역을 검출할 수 있어 기존의 제스처 인식 연구 방법들에서 해결해야 할 문제점 중의 하나였던 다수 사람의 동작 인식이 가능하였다. 또한 MHI의 x, y 좌표의 변화를 나타내는 기울기 정보와 형태 문맥을 적용한 형태 정보를 추출함으로써 동작의 인식 성능을 높였을 뿐만 아니라 정확한 방향 인식도 가능하였다. 또한 제안하는 연구 방법은 특별한 모델링 과정을 거지치 않고, 특징 값을 쉽게 추출할 수 있어 간편하고, 신속하면서도 인식률이 우수한 제스처 인식 시스템을 만들 수 있었다.

그러나 고정된 카메라를 사용하여야 한다는 점과, 동작이 서로 겹치면 인식이 불가능 하다는 점, 동작과 동작 사이를 구분하기 위한 장치를 둔다는 점 등의 여러 가지 제약 사항들이 존재한다. 이는 이동 카메라 환경에 대한 연구와 인식 하고자 하는 사람과 같은 객체를 검출하여 라벨링하는 등의 객체 검출 연구가 필요할 것이다. 뿐만 아니라 사람의 동작의 시작과 끝을 찾는 동작 구분에 대한 연구도 반드시 필요한 과제중의 하나이다.

제안하는 시스템은 특별한 동작의 구별 및 인식이 필요한 영화 및 방송 산업과 사용자와의 상황 인식을 위한 제스처인식이 요구되는 스마트 홈 산업 및 로봇 산업 등의 기반 연구로 활용될 수 있다.

참고문헌

[1] A. Corradini, H.-M. Gross, "Camera-based gesture recognition for robot control," Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks IJCNN 2000, Vol.4, pp.133-138, July 2000.

[2] A. Corradini, "Dynamic Time Warping for Off-line Recognition of a Small Gesture Vocabulary," In Proceedings of the International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems, pp. 82-89, July 2001.

[3] Y. Zhu, G. Xu "A Real-Time Approach to the Spotting, Representation, and Recognition of Hand Gestures for Human-Computer Interaction," Computer Vision and Image Understanding, pp.189-208, Mar 2002.

[4] H. Tanie, K. Yamane and Y. Nakamura "High Marker Density Motion Capture by Retroreflective Mesh

Suit," International Conference on Robotics and Automation, pp.2884-2889, Apr 2005.

[5] S. Hashi, Y. Tokunaga, S. Yabukami, M. Toyada, K. Ishiyama, Y. Okazaki, K.I Arai, "Development of real-time and highly accurate wireless motion capture system utilizing soft magnetic core," IEEE Transactions on Magnetics, Vol.41, pp.4191-4193, Oct 2005.

[6] N. Miller, O.C. Jenkins, M. Kallmann, M.J. Mataric, "Motion capture from inertial sensing for untethered humanoid teleoperation," IEEE/RAS International Conference on Humanoid Robots, Vol.2, pp.547-562, Nov 2004.

[7] S. Yabukami, H. Kikuchi, M. Yamaguchi, "Motion Capture System of Magnetic Makers Using Three-Axial Magnetic Field Sensor," IEEE Transactions on magnetics, Vol.36, pp.3646-3648, Sept 2000.

[8] Y.Yacoob and M.J.Black, "Parameterized modeling and recognition of activities," Journal of Computer Vision and Image Understanding, Vol.73, pp.232-247, Feb 1999.

[9] C.H. Esteban, F. Schmitt, "Silhouette and stereo fusion for 3D object modeling," Computer Vision and Image Understanding, Vol.96, pp.367-392, Dec 2004.

[10] A. F. Bobick and J. W. Davis, ''The Recognition of Human Movement Using Temporal Templates,'' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.23, No.3, pp.257-267, Mar 2001.

[11] R. Venkatesh Babu, K.R. Ramakrishnan, "Recognition of human actions using motion history information," Image and Vision Computing, Vol.22, pp.597-607, Aug 2004.

[12] R. Muñoz-Salinas, R. Medina-Carnicer, F.J. Madrid-Cuevas, A. Carmona-Poyato, "Depth silhouettes for gesture recognition", Pattern Recognition Letters, Vol 29, pp.319-329, Feb 2008.

[13] L. Díaz-Más, R. Muñoz-Salinas, F.J. Madrid-Cuevas, R. Medina-Carnicer, "Shape from silhouette using Dempster-Shafer theory", Pattern Recognition, Vol 43, pp.2119-2131, June 2010.

[14] L.Zelnik Manor and M.Irani, "Event-based analysis of video," IEEE Conference on Computer Vision and Pattern Recognition, Vol.2, pp.123-130, Dec 2001.

[15] S. Belongie, J. Malik, J. Puzicha, "Shape matching and object recognition using shape contexts," IEEE Transactions On Pattern Analysis and Machine Intelligence, Vol.24, No.24, pp.509-522, Apr 2002.

[16] V. Vapnik, "The Nature of Statistical Learning Theory," Springer, New York, 1995.

[17] Y.H. Kim, S.J. Kim, "Movement Detection Algorithm Using Virtual Skeleton Model," Journal of Korean Institute of Intelligent Systems, Vol 18, pp. 731-736, Dec 2008.

저 자 소 개

김 상 균

1991 : 경북대학교 통계학과 이학사
1994 : 경북대학교 컴퓨터공학과 공학석사
1996 : 경북대학교 컴퓨터공학과 공학박사
현재 : 인제대학교 컴퓨터공학부 부교수
관심분야 : 패턴인식, 정보검색, 컴퓨터비전
Email : skkim@inje.ac.kr