# 스테레오 영상을 이용한 3차원 포즈 추정

양 욱 일, 송 환 종, 이 용 욱, 손 광 훈
연세대학교 전기전자공학과
전화 : 02-2123-2879

# 3D Head Pose Estimation Using The Stereo Image

Ukil Yang, Hwanjong Song, Yonguk Lee, Kwanghoon Sohn
Dept. of Electrical and Electronic Engineering, Yonsei University
E-mail : starb612@diml.yonsei.ac.kr

## Abstract

This paper presents a three-dimensional (3D) head pose estimation algorithm using the stereo image. Given a pair of stereo image, we automatically extract several important facial feature points using the disparity map, the gabor filter and the canny edge detector. To detect the facial feature region , we propose a region dividing method using the disparity map. On the indoor head & shoulder stereo image, a face region has a larger disparity than a background. So we separate a face region from a background by a divergence of disparity. To estimate 3D head pose, we propose a 2D - 3D Error Compensated-SVD (EC-SVD) algorithm. We estimate the 3D coordinates of the facial features using the correspondence of a stereo image. We can estimate the head pose of an input image using Error Compensated-SVD (EC-SVD) method. Experimental results show that the proposed method is capable of estimating pose accurately.

## I. 서 론

본 논문은 스테레오 영상을 이용하여 얼굴의 포즈 변화를 추정하는 기법에 대하여 제안한다. 한 장의 2차원 영상으로부터 얼굴의 3차원 포즈 변화를 추정하는 것은 매우 어려운 기술이며, 추정된 포즈 변화도 많은 오차를 나타내고 있다.[1] 따라서 본 논문에서는 스테레오 영상을 이용하여 보다 정확한 얼굴의 포즈 변화를 추정하는 알고리즘을 제안한다. 먼저 스테레오 영상을 정합하고, 정합된 영상을 이용해서 얼굴 영역과 배경을 분리한다. 분리된 영상에 gabor filter를 적용하여 특징 영역을 추출한다. 추출된 특징 영역에서 canny edge detector를 사용하여 특징점을 추출한다. 추출된 특징점의 변이 정보를 이용하여 얼굴의 3차원 포즈 변화를 추정한다.

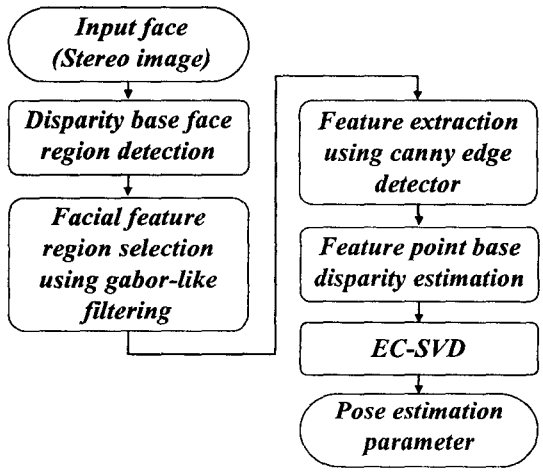


그림 1 Algorithm flow

*본 연구는 한국과학재단지정 생체인식연구센터의 지원을 받아 이루어졌습니다.

## Ⅱ. 얼굴 영역 결정

### 2.1 스테레오 정합

입력 영상의 경우 Max Planck Institute (MPI)에서 획득한 3차원 얼굴 data를 이용하여 2차원 스테레오 영상을 획득한다. 이는 2차원 data와 3차원 data의 일치성을 이용하여 포즈 추정을 하기 때문이다.

스테레오 정합은 화소 정합 방식의 밝기 기반 스테레오 정합 방법을 사용한다. Epipolar constraint를 사용하여 수평 방향으로만 검색한다. 검색 영역은 좌우 각각 최대 128 화소로 제한하며, 화소 정합을 위해서 5×5 창 크기를 사용한다. 정합 기준은 *minimize* absolute error (MAE)를 사용한다.

$$d_r = arg\ min \left| \sum_m \sum_n \{L(x+m, y+n) - R(x+m+d, y+n)\} \right| \quad (1)$$

$$d_l = arg\ min \left| \sum_m \sum_n \{L(x+m+d, y+n) - R(x+m, y+n)\} \right| \quad (2)$$

좌영상을 기준으로 우영상의 변이 $d_r$, 우영상을 기준으로 좌영상의 변이 $d_l$를 각각 계산한다.

그림 2 Left image

그림 3 Right image

그림 4 Disparity map

그림 5 Disparity map

### 2.2 Disparity 정보를 이용한 얼굴 영역 결정

스테레오 정합을 통해서 얻은 변이 정보는 카메라의 *intrinsic, extrinsic* 상수들을 사용하여 깊이 정보로 변화된다. 이러한 깊이 정보는 카메라를 기준으로 물체들의 앞뒤 위치에 대한 정보를 가지고 있다. 즉 변이는 깊이 값에 비례한다.

실내에서 촬영된 head&shoulder stereo 영상의 스테레오 정합 결과를 관찰하면 disparity 값에 대한 히스토그램이 크게 두 부분으로 분할됨을 알 수 있다. 즉, 얼굴 영역에 해당하는 부분과 배경 영역에 해당하는 부분으로 분할된다. 이렇게 깊이 정보의 히스토그램을 이용해서 얼굴 영역을 결정한다.

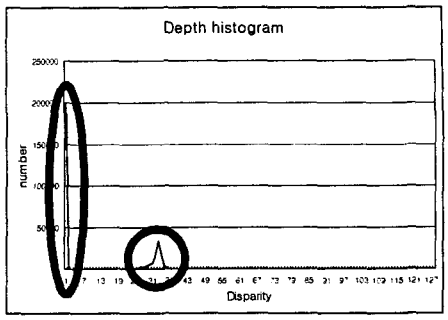


그림 6 Disparity map의 히스토그램

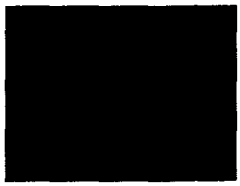

그림 7 영역 추출 전

그림 8 영역 추출 후

## Ⅲ. 특징 영역 선택

### 3.1 2D gabor-like filter

Gabor filter는 자동으로 특징 추출이 가능하고, 인간의 시각적 특징을 모델링한 필터이다.[2] 이 gabor 필터의 even symmetric부분과 odd symmetric 부분을 분리하여 사용하는 gabor-like filter는 다음의 형태를 가진다.

$$CosG(x,y) = cos[(\frac{2\pi}{\lambda})(x\cos\theta + y\sin\theta)] exp[-\frac{(x^2+y^2)}{2\sigma^2}] \quad (3)$$

$$SinG(x,y) = sin[(\frac{2\pi}{\lambda})(x\cos\theta + y\sin\theta)] exp[-\frac{(x^2+y^2)}{2\sigma^2}] \quad (4)$$

$\sigma$는 gaussian aspect ratio, $\theta$는 orientation, $\lambda$는 wave length of the harmonic modulation function이다. 시뮬레이션에서는 $\sigma = 1$, $\lambda = 1$의 값을 사용한다.

Gabor-like 필터는 주어진 $\theta$의 방향에 따라서 그 방향의 edge 성분만을 추출하는 band pass filter의 역할을 한다. 얼굴 영역 내에서 타원 형태의 특징 영역인 눈, 코, 입을 추출하기 위해서 0° 에서 180° 까지 9° 간격으로 생성된 gabor-like 필터를 입력 영상에 적용시키고, 결과 값의 평균과 분산을 구한다. 계산된 평균

과 분산을 기준으로 threshold를 정하고, 정해진 기준에 따라서 특징 영역에 대한 binary 영상을 만든다. 이런 방법을 통해서 폐쇄된 특징 영역을 나타내는 영상을 획득한다.

그림 9 얼굴 영상

그림 10 특징 영역

### 3.2 특징 영역 선택

2.2의 과정을 통해서 획득한 얼굴 영역 정보를 이용해서 얼굴 윤곽선 부분의 특징 영역을 제거한다. 나머지 특징 영역의 깊이 정보를 검색한다. 얼굴 내 특징 영역인 눈, 코, 입을 비교할 경우 눈이 가장 함몰된 영역이며, 코가 가장 도출된 영역이다. 이런 깊이 특징을 이용해서 눈과 코의 후보 영역을 정한다. 이와 동시에 영역들 사이의 깊이 변화률을 확인하여 noise 영역이라 생각되어 지는 영역을 제거한다. 그리고 후보 영역간의 위치 관계와 눈, 코, 입의 위치 관계를 고려해서 눈, 코, 입의 특징 영역을 결정한다.

그림 10 영역 결정 전

그림 14 영역 결정 후

## Ⅳ. 특징점 추출

### 4.1 Canny edge detector를 이용한 특징점 추출

스테레오 영상을 이용한 3차원 포즈 추정 알고리즘의 경우 좌우 영상간의 정확한 correspondence를 바탕으로 추정된 깊이 정보를 이용한다. correspondence가 만족되지 않을 경우 깊이 정보가 잘못 추정되어, 포즈 추정에 오차를 발생시킨다. 이와 같은 오차를 최소화하기 위해서 특징 영역을 추출한 후, 특징점을 찾는 방식을 사용한다.

3.2에서 결정된 눈, 코, 입 각각의 특징 영역에 대해서, 특징 영역을 포함하는 직사각형의 창을 생성한다. 생성된 창 내부 영역에 대해서 canny edge detector[3]를 적용한다. 검출된 edge들을 contouring 방식을 통해서 검색하여 conner point를 선택한다. 좌우 영상 각각에서 검출된 conner point들의 correspondence를 확인한다.

## Ⅴ. 3차원 포즈 추정

### 5.1 3차원 좌표 결정

추출된 특징점들의 정확한 깊이 정보를 얻기 위해 특징기반 변이 추정 방법을 사용한다.

$$d_i = \sqrt{(l_{x_i} - r_{x_i})^2 + (l_{y_i} - r_{y_i})^2} \qquad (5)$$

특징 기반 변이 추정 방법에 의해서 추정된 변이는 스테레오 카메라의 intrinsic, extrinsic 상수에 의해서 실제 깊이 정보로 변화되며, $l_{x_i}$와 $r_{x_i}$와 함께 3차원 좌표를 이루게 된다.

### 5.2 EC-SVD

5.1에서 결정된 특징점들의 좌표를 이용해서 먼저 특이치 분해(SVD) [4] 기반으로 초기 얼굴 포즈 추정한다. 이렇게 추정된 회전각을 이용해서 3차원 정규화 얼굴 공간에서의 오류 보상 과정을 수행하여 보다 정확한 얼굴 회전 각도를 추정한다.

특이치 분해 과정을 통한 회전각은 다음과 같이 수행된다. 먼저 입력영상의 6개의 얼굴 특징점 $p_i = \{x_i, y_i, z_i\}, i = 1, 2, \cdots, n$의 3차원 좌표를 얻은 후, 데이터 베이스의 3차원 mean head의 얼굴 특징점을 $q_i = \{x_i, y_i, z_i\}$라고 하면, 다음과 같은 관계를 얻을 수 있다.

$$\mathrm{p_i} = \mathrm{R_{SVD}q_i} + \mathrm{t},\ i = 1, 2, \cdots, n \qquad (6)$$

여기서 $R_{SVD}$는 SVD를 통하여 얻은 회전 행렬이고 $t$는 전이 벡터이다. $R_{SVD}$와 $t$를 얻기 위해서는 SVD 기법을 통하여 다음의 최소 자승 문제를 해결함으로써 가능하다.

$$\mathrm{minimize} \sum_{i=1}^{n} \| \mathrm{p_i} - \mathrm{R_{SVD}q_i} - \mathrm{t} \| \qquad (7)$$

$R_{SVD}$는 $3 \times 3$ 회전 행렬로서, $R_{SVD}^T = R_{SVD}^{-1}$의 성질을 갖는다. 또한, 얼굴 중심을 원점으로 옮김으로써 전이 벡터는 추후 보상될 수 있다.

$$\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i,\ \bar{q} = \frac{1}{n}\sum_{i=1}^{n} q_i \qquad (8)$$

여기서 $\bar{p}$와 $\bar{q}$는 얼굴 특징점 집합 $\{p_i\}$와 $\{q_i\}$의 중심점이다. 따라서 입력 특징점과 데이터 베이스의 특

징점과의 공분산 행렬 M을 구하기 위해 다음과 같은 과정을 수행한다.

$$\text{Feature}_{\text{input}} = p_i - \bar{p}_i,\ \text{Feature}_{DB_i} = q_i - \bar{q}_i \quad (9)$$

$$M = \sum_{i=1}^{n} Feature_{input_i} * Feature_{DB_i}^{T} \quad (10)$$

식 (10)로부터 얻어진 공분산 행렬 M으로부터 회전행렬 $R_{SVD}$는 다음과 같은 식을 만족해야 한다.

$$R_{SVD} = MQ^{-1/2},\ Q = M^{T}M \quad (11)$$

여기서, 행렬 Q에 대하여 SVD 과정을 수행하면 다음과 같은 식을 얻는다.

$$Q = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \lambda_3 v_3 v_3^T \quad (12)$$

$$Q^{-1/2} = \frac{1}{\sqrt{\lambda_1}} v_1 v_1^T + \frac{1}{\sqrt{\lambda_2}} v_2 v_2^T + \frac{1}{\sqrt{\lambda_3}} v_3 v_3^T \quad (13)$$

여기서 $\lambda_i$와 $v_i$는 각각 고유값과 고유벡터를 나타낸다.

3차원 정규화 얼굴 공간에서의 오류 보상 과정은 다음과 같이 수행된다. 새로운 complete rotation matrix R을 정의하여 SVD에서 얻은 결과를 이용하여 다음과 같은 새로운 식으로 전개될 수 있다.

$$R = R_X R_Y R_Z = R_{SVD_X} R_{\theta_x} R_{SVD_Y} R_{\theta_y} R_{SVD_Z} R_{\theta_z} \quad (14)$$

여기서,

$R_X = R_{SVD_X} R_{\theta_x}$, $R_Y = R_{SVD_Y} R_{\theta_y}$, $R_Z = R_{SVD_Z} R_{\theta_z}$이며, $R_{\theta_x}, R_{\theta_y}$ 그리고 $R_{\theta_z}$는 각 축으로 보상되어야 할 회전 행렬를 나타낸다. 따라서 complete rotation matrix R을 이용하면 식 (6)은 다음과 같은 식으로 전개된다.

$$f_i' = Rf_i + t = R_X R_Y R_Z f_i + t \quad (15)$$

$$= R_{SVD_X} R_{\theta_x} R_{SVD_Y} R_{\theta_y} R_{SVD_Z} R_{\theta_z} f_i + t \quad (16)$$

여기서 $f_i'$ 회전된 입력 영상에서의 특징점이고, $f_i$는 정면 영상일 경우의 얼굴의 특징점을 나타낸다. 결국 전이벡터 $t$도 역시 얼굴 중심을 원점으로 옮김으로써 전이벡터는 추후 보상될 수 있다. 따라서 다음의 수식을 통하여 회전행렬을 구한다.

$$p_i = R^{-1} p_i' = R_Z^{-1} R_Y^{-1} R_X^{-1} p_i' \quad (17)$$

$$= R_{\theta_z}^{-1} R_{SVD_Z}^{-1} R_{\theta_y}^{-1} R_{SVD_Y}^{-1} R_{\theta_x}^{-1} R_{SVD_X}^{-1} p_i' \quad (18)$$

## VI. 모의실험 결과

5명의 3차원 얼굴 data로부터 생성된 여러 가지 포즈를 가지는 스테레오 영상을 이용하여 시뮬레이션 하였다. 포즈 추정에 두 눈의 특징점을 사용하기 때문에 포즈 변화는 두 눈의 특징점을 찾을 수 있는 포즈 변화로 제한한다. 또 회전, 이동, 확대/축소 가운데 회전에 대한 포즈 변화만을 생각하였다.

| Data | 회전 각도 (x축, y축, z축) | 추정된 각도 (x축, y축, z축) | 오차(%) |
|---|---|---|---|
| 1 | (20, 0, 0) | (20, 0, 0) | 0% |
| 2 | (0, 20, 0) | (0, 20, -0.0066) | |
| 3 | (0, 0, 20) | (0, 0, 19.98) | |
| 4 | (15, 15, 0) | (15, 15, -0.005) | |
| 5 | (0, 15, 15) | (0, 15, 14.98) | |
| 6 | (15, 0, 15) | (15, 0, 14.99) | |
| 7 | (30, 23, 30) | (30, 23, 29.96) | |
| 8 | (38, 20, 25) | (38, 20, 24.97) | |

표 1 모의 실험 결과

## VII. 결 론

기존의 2차원 얼굴 인식 방법의 경우 한 장의 2차원 영상을 사용한다. 그렇기 때문에 화소의 밝기와 색상 정보를 이용한 특징 추출 방법이 많이 사용된다. 화소의 밝기와 색상 정보의 경우 화소의 위치 변화 및 빛의 방향과 세기 변화에 매우 민감하다. 그래서 포즈 변화와 illumination에 대해서 인식률이 급속히 떨어지는 현상을 보인다.

본 논문에서는 이러한 2차원 얼굴 인식 방법의 문제점을 해결하는 방안으로 스테레오 영상을 이용할 것을 제안한다. 스테레오 영상을 이용할 경우 화소의 밝기와 색상 정보와 함께 3차원 구조적 정보인 객체의 깊이 정보를 이용할 수 있기 때문이다. 2차원 얼굴 인식에 3차원 정보를 첨가한 보다 정확한 얼굴인식 방법으로 발전할 수 있을 것이라 생각한다.

## References


[1] Haralick, R.M.; Joo, H.; Lee, C.; Zhuang, X.;Vaidya, V.G.; Kim, M.B.; "Pose estimation from corresponding data," *Systems, Man and Cybernetics, IEEE Transactions on,* Volume: 19 Issue: 6, Nov-Dec. 1989

[2] T. S. Lee, "Image representation using 2D Gabor wavelets", *IEEE Trans. Pattern Analysis and Machine Intelligence* 18 959~971(1996).

[3] J. Canny, "A computational approach to edge detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, 1986, 679-698

[4] T.S. Huang, A.N. Netravali, "Motion and structure from feature correspondences: A Review," *Proceedings of the IEEE*, vol.82, no.2, pp. 252-264, Feb. 1994.