퍼지 및 지능시스템학회 논문지 2007, Vol. 17, No. 4, pp. 519-525

칼만 필터를 이용한 시청각 음원 정위 및 추적

Audio-Visual Localization and Tracking of Sound Sources Using Kalman Filter

송민규* · 김진영*, 나승유*
Min Gyu Song, Jin Young Kim and Seung You Na

* 전남대학교 전자컴퓨터공학부

요 약

최근 로봇 기술 및 응용에 대한 관심이 고조됨에 따라, 로봇의 청각기술에 대한 연구가 활발하다. 본 기술에서는 로봇 탑재용으로 인간 청각기능중 하나인 음원정위 및 추적기술에 대하여 논한다. 음원 정위 및 추적을 위하여 시청각 정보를 이용하였는데, 시각정보로는 얼굴색 기반 얼굴 탐지 정보를 이용하였으며, 양이(binaural) 기반의 음원 추정 정보가 청각 정보로서 활용되었다. 시각과 청각 정보는 Kalman 필터를 이용하여 통합하였다. 실험결과 시청각 음원 추적 기술은 일부 정보의 유실이 있을 때, 효과적으로 활용될 수 있음을 보였다.

Abstract

With the high interest on robot technology and application, the research on artificial auditory systems for robot is very active. In this paper we discuss sound source localization and tracing based on audio-visual information. For video signals we use face detection based on skin color model. Also, binaural-based DOA is used as audio information. We integrate both informations using Kalman filter. The experimental results show that audio-visual person tracking is useful, specially in the case that some informations are not observed.

Key Words : 개인추적, 얼굴색. 상관함수. 칼만필터, 청각모델

1. 서 론

최근 로봇에 대한 연구가 단순한 산업용 로봇이 아닌 인간 친화적인 로봇에 대한 연구로 많이 바뀌고 있는 추세이다. 대표적인 예로 일본과 미국의 AIBO, 아이로보 등을 들 수 있다. 인간 친화형 로봇이란 사람의 존재를 감지하고 사람과 의사소통이 가능하여야 한다. 이들 로봇은 화자의 위치를 추정하여 마주보고 음성을 인식하여 인간과 더 친화적인 느낌을 유발시킨다. 휴먼 로봇이 이런 일들을 수행하기 위해서는 우선적으로 사람의 위치를 찾는 것이 필수적이다. 사람의 위치를 찾는 방법으로는 시각 정보나 청각 정보, 또는 시각과 청각을 통합한 정보를 이용하는 방법들이 있다[1][2][3].

보통 시각이나 청각 각각 하나의 정보만을 이용하여 사람의 위치를 추적할 때 음성의 정보의 경우 주위 잡음이나 반사음 등, 영상 정보의 경우 빛의 영향이나 배경의 영향 등을 받아 추적 오차가 많이 발생된다. 본 논문에서는 이를 해결하기 위해 시각 정보에서 얻어지는 사람의 위치와 청각 정보에서 얻어지는 사람의 위치를 바탕으로 Kalman 필터를 통해 움직이는 사람의 위치를 추적하였다. 이를 위해서 2 채널(좌/우) 마이크로 입력되는 청각 신호를 바탕으로 두 마이크로 들어오는 신호의 시간차를 이용하여 음원 추적을 구현하였고[4], 카메라를 통해 얻어지는 시각정보를 바탕으로 컬러 영상에서 사람 피부색 기반의 히스토그램을 바탕으로 사람의 위치를 찾는 방법을 사용하였다[5][6]. 논문의 구성은 2장은 청각, 3장은 시각 정보를 이용하여 개인의 위치를 추적하는 방법, 4장은 시청각 정보를 이용하여 얻어진 위치를 통합하는 방법을 논하였으며, 5장은 실험결과, 그리고 마지막 6장은 결론 및 향후 과제 등으로 되어있다.

접수일자 : 2006년 5월 1일
완료일자 : 2007년 5월 29일
※ 본 연구는 학술진흥재단 2004년도 선도연구지원 사업의 지원으로 수행되었습니다.

2. 청각기반 음원 정위

인간은 두 개의 귀를 가지고 다양한 소리 분석을 한다. 즉, 소리의 방향을 삼차원적으로 파악할 수 있으며, 강한 잡음이 존재하는 경우에도 원하는 소리만 분리하여 청취할 수 있다. 본 절에서는 인간이 두 귀를 이용하여 소리의 방향을 정위하는 원리와 음원 정위에 이용한 방법에 대하여 간략하게 설명한다.

2.1 양이(binaural) 기반 음원 정위

음원의 방위판별을 위해 사용되는 가장 중요한 단서는 두 귀사이의 시간 지연(interaural time difference, ITD)이다. 음파는 상온에서 약 340m/s의 속도로 전달된다. 따라서 음원이 사람으로부터 멀리 있다고 가정할 때, 그림 1과 같이 인간의 두 귀에 음이 도달하는 데에는 음원의 방향에 따라 시간차가 발생하게 된다.

그림 1을 살펴보면 음원과 양쪽 귀와의 거리가 δ 만큼의 차이가 나므로 정면을 기준으로 하는 음원의 방위 θ는 다음과 같이 구할 수 있다.

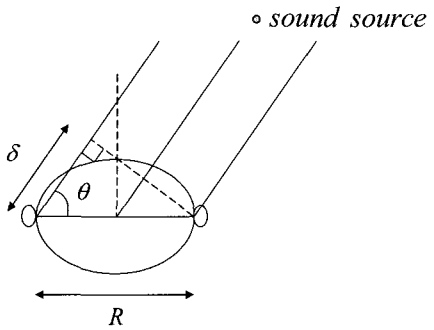

그림 1. 기하학적 관계의 위치 추정
Figure 1. Principle of sound localization

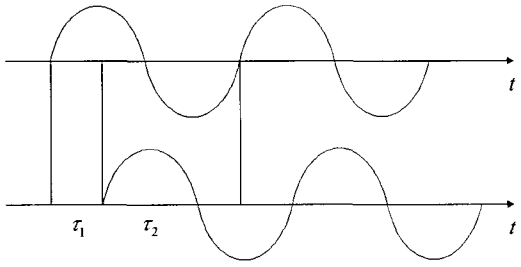

그림 2. 두 귀간의 위상 지연
Figure 2. ITD between two ears.

$$\cos\theta \approx \frac{\delta}{R} \tag{1}$$

양쪽 귀에 들리는 음원이 왜곡이 없고 시간 지연만 있다고 가정하면 그림 2에서 보는 바와 같이 시간만큼의 위상차이가 발생한다. 여기서 음파의 속도를 C라고 하면 $\delta = \tau / C$ 이므로 θ를 구하는 식은 식 (2)와 같이 된다.

$$\cos\theta \approx \frac{\tau}{R \times C} \tag{2}$$

이를 로봇에 적용할 경우에 음파는 일정한 시간간격으로 표본화한 이산 신호로 입력 받게 되며 한쪽신호를 시간 축에 대해 이동시키며 상호 상관을 조사하여 상관도가 높은 점을 찾으면 원래의 파형에서 최대 상관도를 보이는 지점과의 차이가 위상차를 나타내는 것이다. 식 (3)은 위상차 τ를 구하는 식이다.

$$\tau = \arg\max \sum_{n} x_l[k] \cdot x_r[k+\tau_i] \tag{3}$$

일반적으로 ITD는 1.5kHz 이상의 주파수에서는 위상편이에 따른 시간지연 측정이 힘들어지며 따라서 ITD에 의한 음상 정위는 1.5kHz 이하의 저주파 대역의 음에 대한 판별에 사용하게 된다. 그림 2는 시간 축에서 두 귀에 도달하는 음파의 위상 지연을 나타내는 그림이다.

2.2 시간 지연 측정 알고리듬

두 마이크로폰 신호 사이의 시간 지연을 측정하는 대표적인 방법은 상관함수(correlation)를 계산하는 것이다. 두 신호 r1(t)와 r2(t)의 상관함수는 다음과 같이 정의 된다.

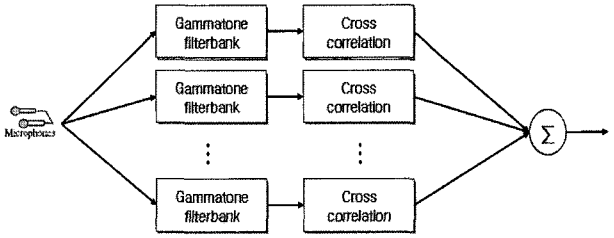

그림 3. 청각모델 기반 상관도
Figure 3. Correlation function based on ear model

$$R_{r_1 r_2}(\tau) = E[r_1(t) r_2(t-\tau)] = \int_{-\infty}^{\infty} r_1(t) r_2(t-\tau) dt \tag{6}$$

여기서E[ ]는 기대값 (시간평균)을 나타낸다. 마이크로폰 신호 r_1에 대한 마이크로폰 신호r_2의 상대적 지연시간은 $\delta_{12} = \tau_2 - \tau_1$ 이므로, $D \equiv \delta_{12}$라 하면 D는 식(6)을 최대로 하는 변수 τ 임을 알 수 있다. 본 논문에서는 두 귀 사이의 시간 지연을 측정하기 위한 방법으로서 청각 모델에 기반한 필터뱅크 (filter bank)를 이용하여 시간지연을 측정하였다.

왜냐하면 인간의 청각이 현재까지 제안된 그 어떠한 방법보다도 잡음에 강인한 성능을 보이기 때문이다. 본 논문에서는 청각 모델로서 'HUTear Matlab ToolBox version2.0'을 이용하였다[7]. 이 귀모델은 크게 6단계로 이루어졌다. 입력신호의 전처리 단계를 거친 다음, 외이와 중이를 모델링 단계 와우각 모델링한 단계(gammatone or gammachirp filterbank), Inner Hair Cell 모델링 단계, 신경 적응 단계, 후처리 단계로 구성되어 있다. 하지만 본 논문에서 사용된 귀모델은 전처리 단계, 외이와 중이를 모델링한 단계 (MAF), 와우각 모델링한 단계(gammatone or gammachirp filterbank)로 구성되어졌다.

청각모델에 기반한 상관도 계산 방법을 그림 3에 보였다. 그림에 보인 바와 같이 각 밴드패스 필터를 거친 신호에 대하여 정규화된 상관함수를 구한 후 평균하여 최종적인 상관함수를 구하게 된다.

3. 시각 기반 개인 추적

로봇에게 중요한 소리는 결국 인간의 음성이다. 따라서 카메라의 영상정보에서 얼굴을 탐지하면, 정확한 음원의 방향을 찾을 수 있을 것이다. 본 논문에서는 얼굴색에 기반하여 개인을 찾아 추적하는 방법을 사용하였다. 특히 얼굴색 필터와 x, y 축 사상(projection)을 이용하여 빠른 계산으로 얼굴을 탐지하는 방법을 제안하였으며, 얼굴탐지의 애매성(ambiguity)에 기반한 적응적 잡음 모델을 제안하여 정확한 얼굴 탐지 및 추적이 가능하도록 하였다. 얼굴의 가부를 판단하고 애매성을 측정하기 위하여 신경회로망을 이용하였다.

3.1 피부색 기반 얼굴 영역 추출

피부색 모델에 기반하여 얼굴 영역 후보를 필터링하여 얼굴 영역을 찾는 방법을 구현하였다. 얼굴색 모델은 Kwato의 (a,b) 파라미터를 이용하였다. 컬러영상의 (R,G,B) 정보는 다음의 식을 통하여 (a,b) 공간으로 변환된다.

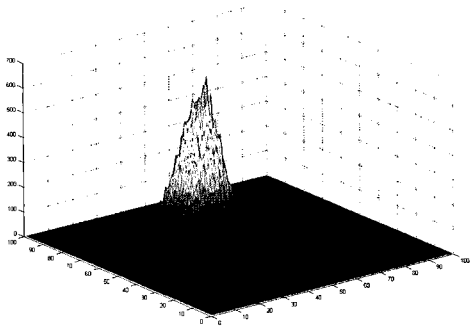

그림 4. 피부색 영역에 대한 누적 히스토그램
Figure 4. Histogram of skin color

그림 5. 얼굴일 확률이 높은 부분의 추출
Figure 5. Skin Color filtering

$$r = R/(R+G+B),\ g = G/(R+G+B) \tag{12}$$

$$a = r + g/2,\ b = \sqrt{3}/2g \tag{13}$$

피부색의 확률적 모델은 히스토그램 기반의 방법을 이용하였는데, 얼굴에 대한 히스토그램을 얻기 위해 총 100장의 얼굴 이미지를 배경과 눈, 입 등 피부색과 다른 부분은 손으로 제거한 후 (a,b)좌표로 변환하여 누적 히스토그램을 얻었다. 얻어진 히스토그램은 전체 화소 개수로 나누어 확률모델로 변경되게 된다. 다음의 그림 4는 피부색 영역 DB를 이용하여 구한 (a,b) 영역에서의 히스토그램을 보여준다. 그림 5는 주어진 영상에서 피부색 확률값이 높은 부분만을 임계치를 주어 필터링한 결과를 보인 것이다.

3.2 수직, 수평 투영을 이용한 얼굴영역 추출

본 논문에서는 얼굴영역을 찾기 위해 얼굴후보영상 각 픽셀의 확률 값을 수직·수평 투영하여 평탄화 작업 후 얼굴영역의 폭과 길이를 구하였다. 이 방법을 이용하여 빠르고 효과적으로 얼굴 영역만을 추출하였다. 피부색 필터링 결과를 $F(i,j)$라고 할 때, 수식, 수평 투영은 다음의 식과 같이 표현된다.

$$X(i) = \sum_{j=0}^{M-1} F(i,j) \tag{14}$$

$$Y(j) = \sum_{i=0}^{N-1} F'(i,j) \tag{15}$$

그림6은 입력 영상에서 수직 투영을 통해 얼굴의 폭을 검출하는 예를 보여주고 있다. 수직투영을 통해 얻어진 얼굴의 폭을 이용해 그 부분만의 수평 투영을 구한 후 그림 7과 같이 얼굴 영역을 검출 하였다. 본 논문에서 제안한 방법은 간단한 수직 및 수평 투영을 이용하기 때문에 오류를 발생하기

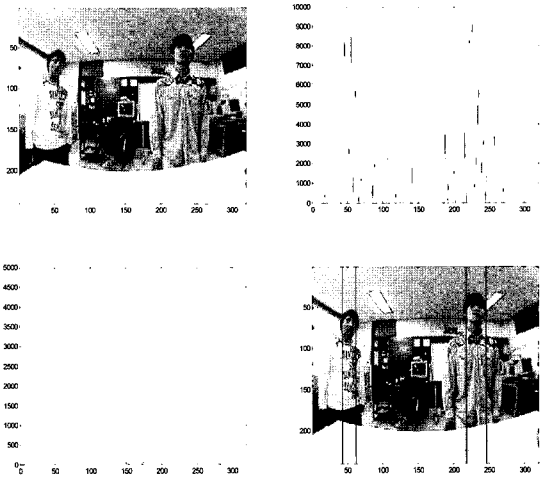

그림 6. 수평투영을 이용한 얼굴영역 추정
Figure 6. Face region detection based on x-projection
(원본 영상(왼쪽 위), 수직 투영 히스토그램(오른쪽 위) 수직 투영 히스토그램 평탄화(왼쪽 아래), 얼굴의 폭 추정(오른쪽 아래))

도 하는데, 오류를 분석결과 두 가지로 나눌 수 있었다.

첫째 수직, 수평 투영 시 높은 피크를 가지고 있는 영역이라 할지라도 피부와 비슷한 영역의 색이 서로 겹쳐서 높은 피크를 가질 수도 있어 얼굴이 아닌 영역을 얼굴로 인식되어 결과를 보일 수도 있다.

둘째, 탐지된 후보 영역에 얼굴의 일부만 보이거나 또는 후보 영역이 얼굴보다 넓게 탐지될 수 있다.

3.2 신경망을 이용한 신뢰도 계산

위절에서 피부색에 기반 한 얼굴 탐지 방법에 대하여 기술하였으며, 제안한 방법에서 발생할 수 있는 오류에 대하여 설명하였다.

본 논문에서는 위 두 문제를 해결하기 위하여 얼굴영역 후보의 검증을 위하여 은닉층을 갖은 다층신경회로망을 사용하였다. 신경회로망은 두 개의 출력 노드를 갖도록 하였는데, 첫 번째 출력은 탐지된 얼굴의 진위를 측정하기 위한 노드이며, 두 번째 노드는 탐지된 영역에서 얼굴의 편이를 측정하기 위하여 사용되었다. 즉, 첫 번째 노드의 출력은 (-1,1)사이의 출력을 갖는 노드로서, 이 값이 적당한 임계값 이상이면 얼굴영상으로 수락하고, 작으면 거절하게 된다. 두 번째 노드의 출력도 출력값의 범위는 첫 번째 노드와 동일하나, 임계값이 0.5이상이면 오른쪽 편이, -0.5이하이면 왼쪽 편이 그리고 중간값이면 편이가 없다고 해석되게 된다. 그림 9는 본 연구에서 사용된 신경회로망의 의미를 간략하게 보인 것이다. 신경망의 학습에 사용된 데이터는 표1에 정리하였다. 여기서 신경망의 입력으로 사용된 이미지는 32×32의 크기로 샘플링 된 이미지이다. 그림 9에 보인 구조의 신경망을 이용하게 됨에 따라 부가적인 이점이 존재한다.

즉, 두 번째 출력노드 (output2)의 결과를 이용함으로서 추출된 영역에서 얼굴의 편이값을 추정할 수 있어, 추정값의 보정이 가능하다는 점이다. 또한 첫 번째 출력노드 (output1)을 측정의 애매함(ambiguity)에 대한 측정이라고 할 때, 이 값에 따라 얼굴 탐색의 잡음을 탄력적으로 조절할 수 있다는 점이다. 다음 그림 10은 출력노드 output1의 값을 신뢰도라고 해석하여 그린 얼굴 탐색의 오차를 그린 그림이다. 이 잡

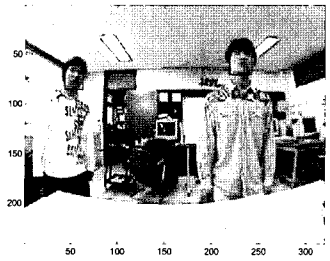

그림 7. 검출된 얼굴 영역
Figure 7. Detected face regions

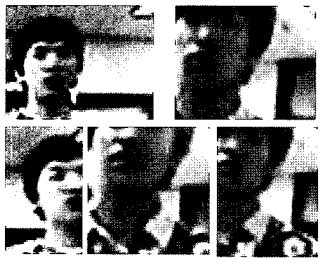

그림 8. 얼굴영역 탐지의 오류
Figure 8. Mis-detections of face regions

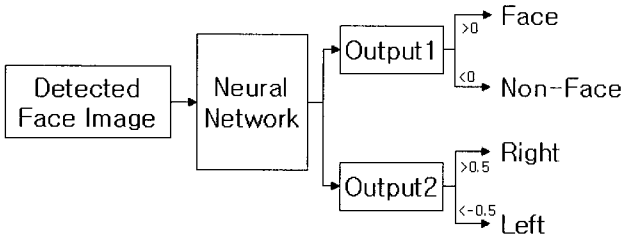

그림 9. 신경망의 흐름도
Figure 9. Blockdiagram of neural network

표 1. 신경망 학습에 사용된 데이터
Table 1. Database for neural network training

사용이미지	개수	수렴값	사용이미지	개수	수렴값
얼굴 이미지	1000개	(1,0)	비 얼굴 이미지	2000개	(-1,0)
얼굴 일부(우)	500개	(0.6,1)	얼굴 일부(좌)	500개	(0.6,-1)
얼굴+배경(우)	500개	(0.2,1)	얼굴+배경(좌)	500개	(0.2,-1)

음 분포는 Kalman 필터를 이용한 추적시 관측 오차로서 이용될 수 있다. 즉, 관측오차는 신뢰도 값에 따라 가우시안 분포(평균과 표준편차)로서 모델링 되며, 평균값의 부호는 신경망 출력노드 output2에 의하여 결정된다.

4. Kalman 필터를 이용한 시청각 정보 통합

위에서 설명한 청각 정보와 시각 정보는 Kalman 필터를 통하여 통합되었으며, 음원의 추적이 이루어졌다.

Kalman 필터는 분석하고자 하는 시스템의 미래현상을 현재의 상태와 미래의 입력 값을 이용해 설명할 수 있음을 모델링 하는 것으로 추적에 좋은 성능을 보이는 필터이다. Kalman 필터를 개인 또는 음원추적에 이용하기 위해서는

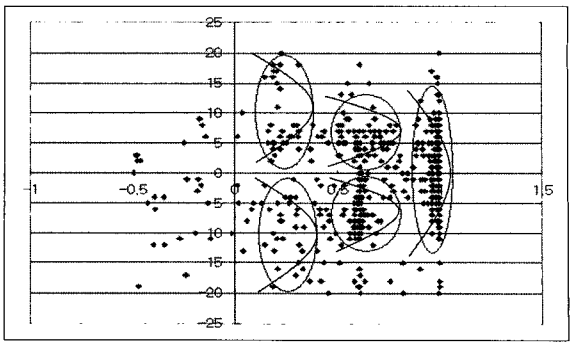

그림 10. 신뢰도에 따른 오차 분포도
Figure 10. Errors depending on confidence

관측값에 대한 상태표현이 필요하다. 지금까지 일상생활에서 시간에 따른 사람의 위치 변화를 표현하는 활동 모델들은 여러 가지가 존재한다[9][10]. 활동 모델들 중 간단하면서도 실제 환경에서 잘 동작하는 것으로 알려진 Langevin 모델[10]을 있다.

이 모델에서 Cartesian 좌표의 각각의 source 움직임들은 서로 독립적이다. x좌표에서 이 움직임은 다음과 같이 표현된다.

$$\dot{x}(k) = a_x \dot{x}(k-1) + b_x F_x \tag{16a}$$

$$x(k) = x(k-1) + \Delta T \dot{x}(k) \tag{16b}$$

$$a_x = e^{-\beta_x \Delta T} \tag{16c}$$

$$b_x = v_x \sqrt{1-a_x^2} \tag{16d}$$

$$X(k) = \begin{bmatrix} \dot{x}(k-1) \\ x(k-1) \end{bmatrix} = \begin{bmatrix} a_x & 0 \\ \Delta t & 1 \end{bmatrix} \begin{bmatrix} \dot{x}(k-1) \\ x(k-1) \end{bmatrix} + b_x \begin{bmatrix} Fx \\ 0 \end{bmatrix} \tag{17}$$

여기서 F_x는 정규화 된 랜덤 변수이고, $\Delta T = L/f_s$ (L은 샘플들의 프레임 길이, f_s는 샘플링 주파수)는 추정치들 사이의 시간 간격을 말하며, v_x는 source의 속도를 뜻한다. 이 모델의 변수들은 실험에 의해 $\beta_x = 10s^{-1}$, $v_x = 1ms^{-1}$로 정해졌다. 이 활동 모델과 변수들은 다른 Cartesian 차원에서도 동일하다. 우리는 이 활동 모델을 사용하여 Kalman 필터를 수행하였다. Kalman 필터를 수행하기 위한 기본 식은 다음과 같다.

시스템 방정식 :

$$X(k) = \Phi_{k-1} X(k-1) + w_{k-1}, \quad w_k \sim N(0, Q_k) \tag{18}$$

측정방정식 :

$$Z(k) = \begin{bmatrix} x_V(k) \\ x_A(k) \end{bmatrix} = H(k)X(k) + v(k), \quad v(k) \sim N\begin{bmatrix} R_I(k) \\ R_A(k) \end{bmatrix} \tag{19}$$

위 식 (18)의 시스템 방정식은 식 (17)로 주어진 source 다이나믹스 식을 이용하게 된다. 한편, 식 (19)로 주어진 관측 방정식에서 관측값은 청각측정값 및 시각측정값이 된다. 또한 $H(k)$는 diagonal 행렬이다. 여기서 관측 오차 $v(k) = \begin{bmatrix} R_V(k) \\ R_A(k) \end{bmatrix}$ 중 $R_V(k)$는 영상을 통해 얻어지는 사람의 위치 중 배경이나 조명 등에 의해 발생되는 오차를 모델링한 것이다. 본 논문에서는 $R_V(k)$의 제어를 위하여 신경회로망의 결과를 이용하였다. $R_A(k)$는 음성을 통해 얻어지는 위치

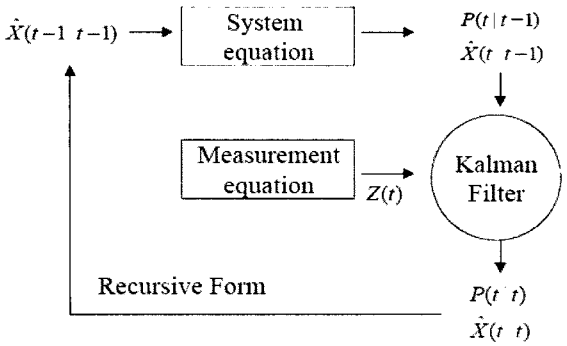

그림 11. Kalman 필터링
Figure 11. Kalman filtering

중 주위 잡음이나 반향파, 또는 SNR이 떨어짐에 따라 발생되는 오차를 모델링한 것이다. 실험을 통해 각 각도별로 어느 정도의 오차가 있는가를 계산한 후 각 각도에 따른 오차 모델을 만들어 사용하였다. 다음 그림 11은 Kalman 필터 과정을 보여준다. 한편 신경망에 의한 시각 신호의 추정된 편이값은 관측값에 더해져서 필터링을 수행하게 된다. 왜냐하면 칼만필터는 잡음의 평균을 0이라고 가정하기 때문이다.

5. 실험결과

실험을 위해 우리는 일반 실험실 환경에서 100°의 시야각을 가지는 카메라와 두개의 마이크로폰을 사용하여 2개의 DB를 제작하였다. 첫 번째 DB는 총 7명의 사람을 0~180° 사이에 위치시킨 후 각각의 위치에서 일정 시간동안 특정 문장을 읽는 DB이며, 다른 하나는 한명의 사람이 0~180° 사이를 애국가를 부르면서 이동하는 DB를 제작하였다. 이렇게 제작된 DB를 음성과 영상으로 분리한 후 각각의 데이터를 바탕으로 사람의 위치를 분석하였다.

5.1 청각신호 기반 음원 추적

청각신호에 의한 음원 정위는 정지된 사람의 위치 정위 실험과 동적인 움직임을 갖는 사람의 위치 정위 실험을 수행하였다. 표 2는 정지 음원에 대하여 음원추적 윈도우의 크기에 따른 추적 결과를 보여준다. 표 2에 의하면, 중앙에서 좌우 0° 및 180°에 음원이 위치함에 따라 추정 편이가 커진다는 사실을 알 수 있다. 또한, 일반 상관함수를 사용한 경우와 청각모델을 이용한 경우를 비교하여 보면 청각 모듈을 사용하는 것이 일반 상관함수를 이용하는 것 보다 우수한 성능을 보임을 확인 할 수 있다.

한편 그림 12는 이동하는 음원에 대한 정위 결과를 보여준다. 음원은 0° 와 180°사이를 왕복한 경우이다.

그림에서 볼 수 있는 바와 같이, 움직임 경우에 대해서는 음원 추적의 성능이 저하되고 있음을 알 수 있으며, 종종 90도 방향으로 추적되고 있음을 알 수 있다. 이는 입력 신호의 신호대잡음비가 좋지 않은 경우에 발생하였는데, 실험 데이터 녹음시 중앙에 위치하였던 컴퓨터의 잡음이 간섭을 일으킨 것이다. 이 경우 본 실험에서는 신호대잡음비에 대한 임계치를 설정하여, 신호대잡음비가 좋지 않은 경우에는 오디오에 의한 음원정위 정보를 무시하도록 하였다.

5.2 시각신호에 의한 음원추적

표 2. 윈도우 크기에 따른 음원 추적 결과
Table 2. Results of sound localization

실제위치	방법	50ms	100ms	200ms	400ms	1s
0°	CC	46.264	45.442	35.006	48.792	49.321
	Ear	40.621	30.848	22.879	29.773	27.548
50°	CC	57.351	57.280	59.020	57.036	52.183
	Ear	49.899	49.543	53.096	48.498	42.860
70°	CC	60.360	60.967	62.626	66.118	66.521
	Ear	71.729	71.686	71.327	72.442	67.326
90°	CC	87.345	85.594	86.259	84.605	81.037
	Ear	89.134	89.534	88.498	88.343	87.791
110°	CC	104.036	103.375	104.228	103.450	99.617
	Ear	103.650	105.411	108.558	108.963	109.520
130°	CC	125.245	128.254	128.211	127.510	129.987
	Ear	126.978	132.303	129.503	132.475	132.940
180°	CC	139.270	145.036	145.325	146.581	146.025
	Ear	143.602	146.104	147.210	146.519	146.025

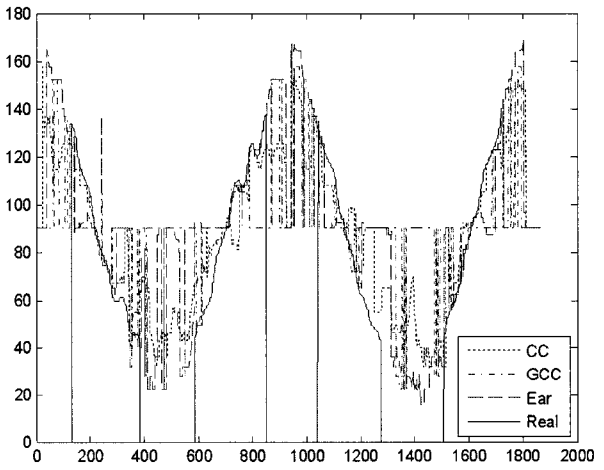

그림 12. 이동하는 음원 정위 결과
Figure 12. Localization results on moving source

얼굴색(피부색) 기반 얼굴 탐지를 이용한 영상 추적은 상당히 좋은 성능을 나타낸다. 즉, 대상 음원의 얼굴이 카메라 시야각 안에 존재하는 경우는 상당히 정확하게 음원을 찾을 수 있다. 다음의 표 3은 영상신호를 기반으로 사람의 위치를 추적한 결과를 보여준다.

표 3은 얼굴탐지 기반 개인 추적인 매우 정확함을 입증하고 있다.

한편 그림 13은 움직임 음원을 대상으로 한 개인 정위의 결과를 보여주고 있다. 그림 13에 보인 바와 같이 얼굴탐지에 의한 개인 추적은 정확하게 수행하고 있음을 알 수 있다. 단, 영상에 의한 개인 정위는 간혹 얼굴후보영상의 검증 오류에 의하여, 개인 정위에 실패하고 있다. 또한 움직이는 사람이 카메라의 시야각 밖에 존재하는 경우에는 시각정보를 통하여서는 움직이는 음원을 추정하는 것이 불가능하다. 그러므로 시각과 청각정보를 동시에 이용하는 방법에 대한 고려가 필요하다고 할 수 있다.

표 3. 영상기반 개인정위 결과
Table 3. Person localizaed results based on video information

실제위치	50°	70°	90°	110°	130°
추정위치	48.98	69.56	90.9	111.54	132.5

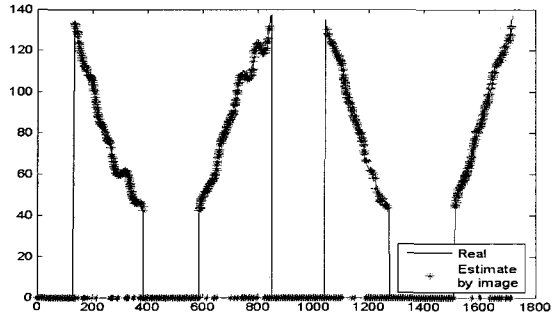

그림 13. 이동하는 개인 정위 결과
Figure 13. Localization results on moving person

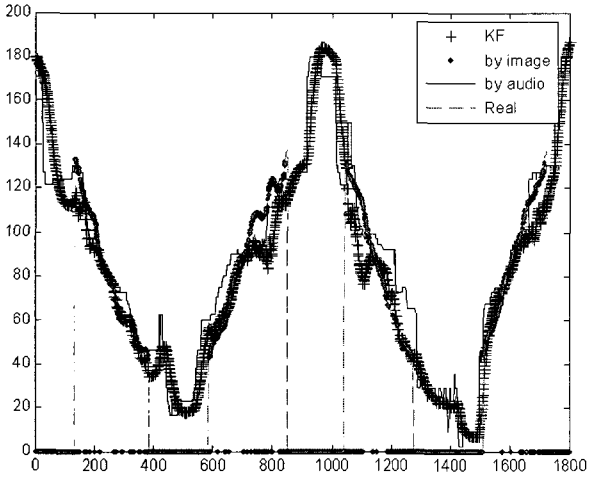

그림 14. 시청각 음원 추적 결과
Figure 14. Audio-visual sound source localization results

5.3 시청각정보 통합에 의한 음원 추적

본 논문에서는 시청각 정보의 통합방법으로서 Kalman 필터를 이용하였다. 이는 오디오 정보의 경우 잡음이 영상정보에 비하여 많다는 단점, 비디오 신호는 시야각 바깥에 존재하는 음원을 정위할 수 없다는 문제를 해결하기 위한 것이다. 그림 14는 Kalman 필터를 이용한 시청각 음원 추적 결과를 보이고 있다. 실험에서 청각정보가 소실된 경우 청각정보 관측 잡음을 무한대로 두었으며,

시각정보가 존재하지 않는 경우에도 동일한 방법으로 Kalman 필터 적용시 무시되도록 하였다. 그림에서 볼 수 있는 바와 같이 통합된 음원 추적이 시각정보 또는 청각정보의 소실에 무관하게 음원을 추적하고 있음을 알 수 있다. 그러나 정확하게 추정된 시각정보가 존재하는 경우에도 청각정보의 오류가 전체 Kalman 필터의 결과에 영향을 미치고 있음을 알 수가 있다. 그러므로 향후 연구로서 시각정보 및 청각정보의 잡음도 또는 측정 정확도에 따라 최적으로 가중하는 Kalman 필터에 대한 연구가 필요하다고 할 수 있다.

6. 결론

본 논문에서는 일반 실험실 환경에서 시청각 정보를 이용하여 사람의 위치를 추적하는 방법에 대하여 논하였다. 실험결과 음성의 경우 중앙, 즉 90° 근처에 위치한 사람의 경우는 비교적 추적결과가 좋으나 양 끝단 쪽으로 즉, 0°나 180° 근처로 갈수록 왜곡이 심하게 나타났다. 영상의 경우도 배경의 피부색과 유사한 색이 많이 존재하거나 빛의 영향을 많이 받을 시 사람의 위치를 추적하지 못하는 결과들을 보였다. 전체적인 구간에서 음성 정보 보다는 영상 정보에 의해 얻어지는 사람의 위치정보가 더욱 정확함을 알 수 있었다. 시청각 정보를 이용하여 사람의 위치를 찾을 시 영상 정보에 의해 사람의 위치 정보가 구해지면 영상 쪽에 가중치를 두고 사람의 위치를 추적하는 방법이 더 좋은 결과를 보였다. 향후 2개 이상의 많은 마이크로폰과 카메라를 이용하면 보다 정확한 사람의 위치를 추적할 수 있을 것이다. 또한 다중 화자가 존재하는 환경에서 사람의 위치를 추적하기 위해 화자 분리 기술에 대한 연구가 필요하다.

참 고 문 헌

[1] J. Segen and S. Pingali, "A camera-based system for tracking people in real time," in International Conference on Pattern Recognition, vol. 3 pp. 63-67, 1996

[2] D. J. Beymer and K. Konolige, "Real-time tracking of multiple people using stereo," In Frame-rate99, 1999

[3] J. Vermaak and A. Blake, "Sequential monte carlo fusion of sound and vision for speaker tracking," In International Conference on computer Vision, 2001

[4] P. S. Chang, "Performance of 3D Speaker Localization Using a Small Array of Microphones," In Proc.. of IEEE International Conference on Thirty-First Asilomar, vol. 1, pp. 2-5, 1997

[5] S. Kawato and J. Ohva, "Automatic Skin-color Distribution Extraction for Face Detection and Tracking," In ICDSP2000 : The 5th Int. Conf. on Signal Processing, vol II, pp 1415-1418, 2000

[6] F. Tomaz, T. Candeias and H. Shahbazkia, "Improved Automatic Skin Detection in color Inages," In Proc. VIIth Digital Image Compution : Techniques and Applications, Sun C. Talbot H. Ourselin S. and Adriaansen T., pp 10-12, 2003

[7] A. Harma. (2000, March 7). THTear Matlab Toolbox(version2.0) Available : http://www.asustic.hut.fi/software/HUTear/HUTear.html

[8] H. A. Rowley, S. Baluja, T. Kanade, "Human face detection in visual scenes," CMU-CS-95-158, Carnegie Mellon University, November, 1995

[9] M. Isard and A. Blake, "Condensation-Conditional density propagation for visual tracking," In Int. J. Computer Vision, vol. 29, no1, pp5-28, 1998

[10] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environment," In Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-01), Salt Lake City, UT, USA, May 2001.

저 자 소 개

송민규(Min Gyu Song)
2004년 전남대학교 졸업.
2006년 동 대학원 전자공학 석사
2006년~현재 동 대학원 전자공학 박사과정

관심분야 : 시청각 개인추적, 임베디드 시스템

Phone : 062-530-0472
Fax : 062-530-1759
E-mail : smg686@lycos.co.kr

김진영(Jin Young Kim)
1986년 서울대학교 졸업.
1988년 동 대학원 전자공학 석사
1994년 동 대학원 전자공학 박사
1995년~현재 전남대학교 교수

관심분야 : 음성처리, 시청각 신호처리

Phone : 062-530-1757
Fax : 062-530-1759
E-mail : beyondi@chonnam.ac.kr

나승유(Seung You Na)
1977년 서울대학교 졸업.
1984년 Univ. of Iowa, Dept. of ECE Master
1986년 Univ. of Iowa, Dept of ECE Ph. D
2005년~현재 전남대학교 교수

관심분야 : 지능제어 및 계측, 신호처리

Phone : 062-530-1757
Fax : 062-530-1759
E-mail : syna@chonnam.ac.kr