<학술논문> DOI:10.3795/KSME-A.2011.35.7.737 ISSN 1226-4873

음성-영상 융합 음원 방향 추정 및 사람 찾기 기술§

이병기* · 최종석*† · 윤상석** · 최문택** · 김문상** · 김대진***
* 한국과학기술연구원 인지로봇연구단, ** 한국과학기술연구원 지능로봇사업단, *** 포항공과대학교 컴퓨터공학과

Audio-Visual Fusion for Sound Source Localization and Improved Attention

Byoung-gi Lee*, JongSuk Choi*†, SangSuk Yoon**, Mun-Taek Choi**, Munsang Kim**, and Daijin Kim***
* Center for Cognitive Robotics Research, Korea Institute of Science and Technology,
** Center for Intelligent Robotics, Korea Institute of Science and Technology,
*** Dept. Computer Science and Engineering, Postech

(Received December 10, 2010 ; Revised March 29, 2011 ; Accepted April 13, 2011)

Key Words : Audio-Vision Fusion(음성영상융합), Sound Source Localization(음원 방향 추정), Human Attention(사람 찾기), Robot Tracking(로봇 추적)

초록: 서비스 로봇은 비전 카메라, 초음파 센서, 레이저 스캐너, 마이크로폰 등과 같은 다양한 센서를 장착하고 있다. 이들 센서들은 이들 각각의 고유한 기능을 가지고 있기도 하지만, 몇몇을 조합하여 사용함으로써 더욱 복잡한 기능을 수행할 수 있다. 음성영상 융합은 서로가 서로를 상호보완 해주는 대표적이면서도 강력한 조합이다. 사람의 경우에 있어서도, 일상생활에 있어 주로 시각과 청각 정보에 의존한다. 본 발표에서는, 음성영상 융합에 관한 두 가지 연구를 소개한다. 하나는 음원 방향 검지 성능의 향상에 관한 것이고, 나머지 하나는 음원 방향 검지와 얼굴 검출을 이용한 로봇 어텐션에 관한 것이다.

Abstract: Service robots are equipped with various sensors such as vision camera, sonar sensor, laser scanner, and microphones. Although these sensors have their own functions, some of them can be made to work together and perform more complicated functions. Audiovisual fusion is a typical and powerful combination of audio and video sensors, because audio information is complementary to visual information and vice versa. Human beings also mainly depend on visual and auditory information in their daily life. In this paper, we conduct two studies using audiovision fusion: one is on enhancing the performance of sound localization, and the other is on improving robot attention through sound localization and face detection.

1. 서 론

음성-영상 융합 기술은 음성 처리 기술과 영상 처리 기술을 동시에 활용하여 보다 유용한 기능을 개발하고, 보다 신뢰도 높은 성능을 내기 위한 기술이다. 인간의 경우, 오감이라고 불리는 시각, 촉각, 청각, 후각 그리고 미각을 통하여 현실 세계를 이해하고, 감지한다. 그 중 시각과 청각은 단연 중요한 비중을 차지하고 있으며, 이는 이 두 감각의 조합이 상호보완적 특성을 갖는데 기인한다. 시각은 정교하고 정확한 특성을 갖지만, 감지 범위가 근거리의 일정 시야각 이내로 제한되고, 밤과 낮에 따른 조명 환경에 민감한 특성을 보인다. 반면, 청각은 비교적 원거리의 사건까지 감지가 가능하고, 밤과 낮 모두 제약이 없다는 장점이 있으나, 위치 추정의 정교함이나 사건 분석의 정확성은 다소 떨어지는 단점이 있다. 이와 같은 이유로, 로봇에서도 음성 센서와 영상 센서를 융합하여, 보다 신뢰도가 높고, 환경 제약에서 자유로운 기술을 개발하려는 노력이 많이 있어 왔다. 일본의 혼다연구소의 휴머노이드 SIG 에서는 얼굴 검출과 얼굴 인식 등의 영상 기술과 음조 추출, 음원 방향 추정, 음원 분리 등의 음성 기술을 결합하여, 효과적으로 다화자를 추적하는 연구를 수행하였다.[1] 이와 유사한 연구로, Y. Lim 등은 2009 년도 연구에서 얼굴 검출 기술과 음원

§ 이 논문은 대한기계학회 2010 년도 추계학술대회(2010. 11. 3.-5., ICC 제주) 발표논문임
† Corresponding Author, cjs@kist.re.kr

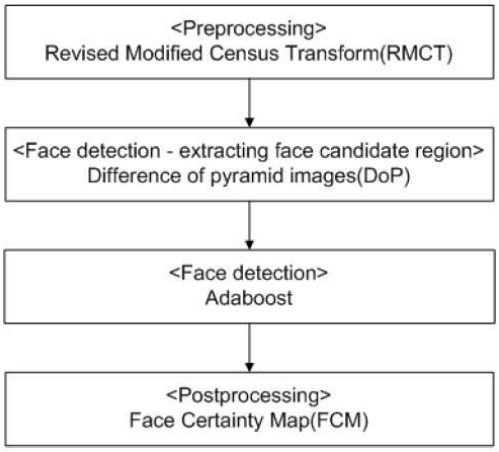


Fig. 1 Overall procedure of the face detection

방향 검지 기술을 파티클 필터 기법으로 융합하여, 효과적으로 다화자를 추적하고, 현재 발화자를 추정하는 로봇을 개발하였다.[2]

한편, 영상 기술로 부터의 정보를 활용하여 자가 학습형 음원 방향 추정 기술이 개발되기도 하였다. 소리 신호를 분석하여, 음원 방향을 추정하기 위해서는 사전에 미리 플랫폼에 따른 방위별 음성 신호의 특징 정보를 미리 구축 해야 하는데, 영상으로부터 음원의 위치를 파악하고, 동시에 소리 신호의 방향 특징을 수집하여, 학습적 방법으로 음원 방향을 추정하는 것이다. 연구 사례로는 유럽의 i-CUP 로봇에서 이루어진 오디오-모터 맵 빌딩 연구[3]가 있고, 호주 시드니 대학의 오디오-비주얼 퓨전 실험[4] 사례가 있다.

본 발표에서는 앞에 소개된 사례들과 마찬가지로 음성 신호 처리 기술과 영상 신호 처리 기술을 융합한 연구들인, 반향에 대응 가능한 음원 방향 검지 기술과 사람을 찾고 주의를 집중하는 로봇 어텐션 기술을 소개하고자 한다. 이제, 2 장과 3 장에서는 기본이 되는 기술인 얼굴 검출 기술과 음원 방향 검지 기술을 각각 간단히 설명하고, 4 장에서는 반향 환경 학습 기술을, 5 장에서는 로봇 어텐션 기술을 소개한다.

2. 얼굴 검출 기술

2.1 전체적 수행 단계

일반적으로, 얼굴 검출은 전처리 단계, 얼굴 검출 단계, 후처리 단계의 3 단계로 나뉘어 수행된다. 첫번째 단계인 전처리 단계에서는 조도의 변화에 강인한 처리를 위해 본래의 이미지를 RMCT (Revised Modified Census Transform)으로 변환하게 된다. 그 다음 단계인 얼굴 검출 단계에서는 DoP (Difference of Pyramid)를 이용하여 빠르게 window searching 을 하며, cascade 형태의 다중 단계 Adaboost 로 얼굴인지 아닌지 판단하게 된다. 마지막으로, 후처리 단계에서는 FAR (False Acceptance Rate)를 낮추기 위해, FCM (Face Certainty Map)을 제안하였고, 이를 통해, 보다 안정적인 얼굴 검출 결과를 얻을 수 있다.

2.2 전처리 단계

Zabin 과 Woodfill 은 CT (Census Transform)이라는 조도에 강인한 국지적 변환 방법을 제안하였다.[5] 이후 Froba 와 Ernst 는 이를 개선하여 MCT (Modified Census Transform)을 발표하였다.[6] 우리는 이 MCT 를 보완하여, RMCT (Revised Modified Census Transform)을 제안한다.

$N(x)$ 가 x 를 포함하는 x 지점 근방 화소들의 집합이라고 하자. $I(x)$ 는 x 지점의 화소값 (Intensity)이고, $\bar{I}(x)$ 는 $N(x)$ 집합에 대한 화소값의 평균이다. 그러면, x 지점에서의 RMCT 변환은 다음과 같다.

$$\Upsilon(x) = \underset{y \in N}{\otimes} C\left(\bar{I}(x)+r, I(y)\right),\ r = 2 \text{ or } 3 \qquad (1)$$

여기서, $\otimes$ 는 연결 연산자 (concatenation operation)이고, $C(\cdot,\cdot)$ 는 앞의 인자와 뒤의 인자를 비교하여 앞의 인자가 작으면 1, 아니면 0 인 비교 함수 (comparison function)이다.

RMCT 는 근방 화소 집합이 3x3 크기일 때, 각 화소값을 0~510 으로 변환시켜 주며, 이는 근방 화소 집합의 평균값에 근거하기 때문에, 조도 변화에 강인한 특성을 보여 준다. 또한, RMCT 는 국지적 화소의 변화 패턴을 잘 반영하기 때문에, 얼굴 검출에 적합하다.

2.3 얼굴 검출 단계

얼굴 검출은 RMCT 변환 이미지와 Adaboost classifier 을 이용하여 이루어 진다. 학습 이미지들을 RMCT 이미지로 변환한 뒤, 이를 이용하여 weak classifier 를 학습 시키고, 이들 weak classifier 를 선형조합 (linear combination)으로 묶어 strong classifier 를 구성한다. 각각의 weak classifier 는 이미지에서 특정위치를 담당하며, 입력된 RMCT 값에 대해서 신뢰값 (confidence value)을 출력한다. Weak classifier 의 신뢰값들을 선형조합으로 종합한

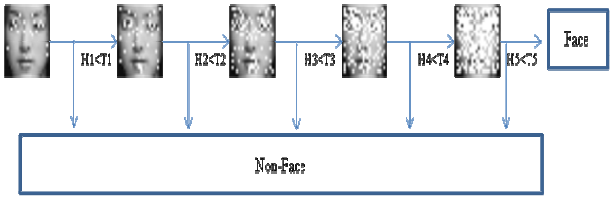


Fig. 2 Multi-stage classifier cascade

값이 strong classifier 의 신뢰값이 되며, 기준 문턱값 보다 신뢰값이 높은 경우, 얼굴 영역으로 인식하게 된다.

이 뿐 아니라, 얼굴 검출의 성능은 높이고 수행시간은 낮추기 위해, 다중 단계 classifier 를 구성하게 된다. 나중 단계의 classifier 는 앞단계의 classifier 보다 더 많은 weak classifier 로 구성된다.

만약 주어진 이미지가 얼굴이 아닐 경우, 첫단계에서 즉시 비얼굴 영역으로 판별하여 속도를 높이게 된다. 그리고, 나중 단계의 classifier 는 앞 단계에서 판별에 실패한 이미지들로 학습시켜 검출 성능을 높이게 된다. 한편, 주어진 이미지에 대한 얼굴 검출은 입력 사이즈에 맞는 윈도우를 이미지 전 영역에서 움직이며 이미지 조각을 얻고, 이를 얼굴인지 아닌지 판별하며 검출하게 된다. 이때, 다양한 크기의 얼굴들을 검출할 수 있게, 이미지를 다양한 크기로 줄여가며 윈도우 searching 을 실시하게 된다. (이렇게 차츰 크기가 줄어드는 이미지들을 이미지 피라미드라고 부른다.) 이러한 형태의 full-search 방식은 제일 좋은 성능을 내지만, 수행시간이 길어지게 된다. 이러한 문제를 해결하기 위해, DoP 를 통한 윈도우 searching 방식을 제안한다. Fig.3 은 (i-1) 번째 피라미드 이미지와 i 번째 피라미드 이미지의 차를 보여주고 있다. 이 사례에서 확인할 수 있는 것처럼, 배경 이미지는 DoP 가 크지 않는 특성을 보인다. 이를 이용하면, 기준 문턱값 보다 높은 DoP 를 갖는 영역에 대해서만 선택적으로 얼굴 검출을 수행하여 수행시간을 단축 할 수 있다.

2.4 후처리 단계

얼굴 검출 단계를 거치면, 얼굴 영역으로 판별된 수많은 윈도우들을 얻게 된다. 이때, 실제 얼굴 영역도 검출이 되지만, 얼굴이 아닌 영역이 얼굴 영역으로 검출된 경우도 발생하게 된다. 이를 관찰해 보면, 실제 얼굴 영역인 경우, 많은 윈도우들이 동시에 얼굴로 판별되고, 얼굴이 아닌 영역은 소수의 윈도우가 얼굴로 판별되는 것을 볼 수 있다.

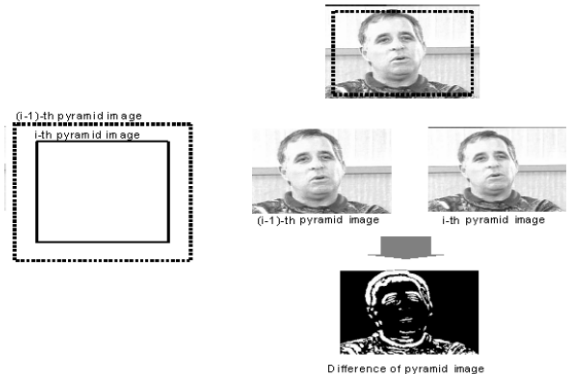


Fig. 3 An example of DoP image

이를 바탕으로 다음과 같은 FCM (Face Certainty Map)에 기반한 후처리 방법을 제안한다. 이 방법을 사용하면, FAR (False Acceptance Rate)을 현저히 줄일 수 있다.

1. (x,y) 를 중심으로 하는 윈도우에 대해, 다중 단계의 i 번째 cascade classifier 의 신뢰값은 다음과 같이 구한다.
$$H_i(\Upsilon) = \sum_{p \in S_i} h_p(\Upsilon(p)) \tag{2}$$
여기서, p 는 p 번째 특징 추출 위치, S_i 는 특징 추출 위치의 집합을 나타낸다.
2. 다중단계 classifier 전체의 신뢰값은 다음과 같이 구한다.
$$S(x,y) = \begin{cases} \sum_{i=1}^{n} H_i(\Upsilon), \text{ if } H_i(\Upsilon) > \text{threshold for all } i \\ 0, \text{ otherwise} \end{cases} \tag{3}$$
3. 모든 피라미드 이미지에 대해 (3)을 계산할 수 있으며, p 번째 피라미드 이미지에 대해 신뢰값 $S_p(x,y)$ 을 얻는다.
4. 이제 FCM 은 다음 4 가지 값으로 정의할 수 있다.
$$\begin{cases} S_{\max}(x,y) = \max_p S_p(x,y) \\ W_{\max}(x,y) = (\text{width of window having } S_{\max}(x,y)) \\ H_{\max}(x,y) = (\text{height of window having } S_{\max}(x,y)) \\ C(x,y) = \sum_{p=1} S_p(x,y) \end{cases} \tag{4}$$
5. FCM 의 $S_{\max}(x,y)$ 와 $C(x,y)$ 가 동시에 기준 문턱값 보다 높을 때, 최종 얼굴 영역으로 검출한다.

본 알고리즘을 적용한 얼굴 검출기를 CMU + MIT 얼굴 검출 테스트 셋에 테스트하여 결과는 Table 1 과 같다.[7]

3. 음원 방향 추정 기술

3.1 마이크로폰 어레이

음원 방향 추정 기술은 다수의 마이크로폰을 사

Table 1 Results of face detection

Detector	False Detection
RMCT, adaboost and FCM	3
MCT and adaboost	93
Viola-Jones	78
Rowley-Baluja-Kanade	167
Bernhard Froba	27

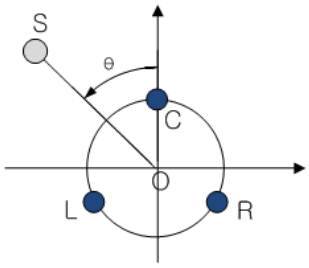


Fig. 4 3-channel Microphone array

용하여 음원 신호를 동시에 측정하고, 마이크로폰 사이의 측정 신호차를 이용하여 음원의 방향을 추정하는 기술이다. 따라서, 마이크로폰 어레이의 형태나 위치는 음원 방향 추정의 성능을 크게 좌우하는 요소이다. 본 연구에서는, 마이크로폰을 3 개 사용하며, 7.5cm 반경의 원 위에 120 도 간격으로 정삼각형 모양을 이루게 배치한다. 이 구조의 마이크로폰 어레이는 0~360 도 범위의 음원 방위각을 측정하는데 적합하다.

3.2 도달 지연 시간차 (TDOA)

음원으로부터 각 마이크로폰 L, C, R 에 이르는 거리는 조금씩 차이가 난다. 이 거리차에 의해, 음원으로 부터 각 마이크로폰까지 도달되는 소리들 사이에는 시간 지연이 발생하고, 이것을 TDOA (Time Delay of Arrival) 라고 부른다. TDOA 는 마이크로폰 어레이의 중심 O 로부터 음원까지의 거리에 따라서는 많이 변화하지 않고, 주로 음원의 방위각에 따라 변화하는 것으로 알려져 있다. 따라서, TDOA 를 분석하면, 음원의 방향을 추정할 수 있게 된다. TDOA 는 다음과 같이 계산될 수 있다.

$$\begin{cases} \tau_{LC}(\theta) = \dfrac{\overline{LS_\theta} - \overline{CS_\theta}}{v_{sound}} \\ \tau_{CR}(\theta) = \dfrac{\overline{CS_\theta} - \overline{RS_\theta}}{v_{sound}} \\ \tau_{RL}(\theta) = \dfrac{\overline{RS_\theta} - \overline{LS_\theta}}{v_{sound}} \end{cases} \quad (5)$$

3.3 교차 방위 상관도 (CAC)

두 신호간 닮은 정도는 상관도(correlation)을 계산하여 수치화 할 수 있다. 한 신호에 대해 다른 한 신호를 시간 이동시켜가며 상관도를 조사하고, 가장 상관도가 커지는 시간을 찾으면 두 신호간의 시간차를 구할 수 있다. 이것을 교차 상관도 (cross-correlation)이라고 한다. 마이크로폰 L 에서 얻은 신호를 $mic_L(t)$ 라고 하고, 마이크로폰 C 에서 얻은 신호를 $mic_C(t)$ 라고 하면, $mic_L(t)$ 과 $mic_C(t)$ 사이의 교차 상관도는 다음과 같이 정의 된다.

$$r_{LC}(\tau) = \frac{\langle mic_L(t), mic_C(t-\tau) \rangle}{\|mic_L(t)\| \|mic_C(t)\|} \quad (6)$$

그렇지만, 우리가 원하는 형태의 함수는 음원의 각도에 따른 상관도를 구할 수 있어서, 상관도가 가장 높게 나타나는 음원의 각도를 쉽게 알 수 있는 함수이다. 따라서, 식 (5)와 (6)을 결합하여 다음과 같은 새로운 교차 상관도 함수를 정의 한다.

$$r_{LC}(\theta) = \frac{\langle mic_L(t), mic_C(t-\tau_{LC}(\theta)) \rangle}{\|mic_L(t)\| \|mic_C(t)\|} \quad (7)$$

한편, 마이크로폰이 3 개이므로, r_{LC} , r_{CR} , r_{RL} 의 교차 상관도를 얻게 되고, 이들을 결합하여 다음과 같이 교차 방위 상관도 (cross-angle-correlation, CAC)를 정의한다.

$$R(\theta) = \frac{r_{LC}^+(\theta) r_{RL}^+(\theta) + r_{CR}^+(\theta) r_{LC}^+(\theta) + r_{RL}^+(\theta) r_{CR}^+(\theta)}{3} \quad (8)$$

, where $r_{XY}^+(\theta) = \max(0, r_{XY}(\theta))$

CAC 가 구해지면, 최대값을 갖는 각도에 음원이 있다고 추정하게 된다.

$$\hat{\theta} = \arg\max_\theta R(\theta) \quad (9)$$

4. 음성-영상 융합 반향환경 학습

4.1 반향환경에서의 음원 방향 추정

밀폐된 실내 환경에서 발화를 할 경우, 벽면에서 반사되어 오는 반사파가 생기게 되기 마련이다. 어느 정도 수준의 반사파는 소리를 풍부하게 해주어 사람이 소리를 듣는데 있어 도움을 준다. 그렇지만, 심한 반향이 있는 경우는 우리가 소리를 알

아든거나, 음원의 방향을 추정하는데 어려움을 느끼게 된다. 특히나, 음원 방향 추정 기술에 있어서, 반사파는 또 다른 음원으로 검지되어 올바른 음원 방향 추정을 하는데 큰 혼란을 야기하게 된다. 사람의 경우, 반향이 심한 장소에서도 어느 정도 효과적으로 방향을 알아낼 수 있는데, 이것은 선행효과 (precedence effect)라고 알려진 원리에 기반한다.[8] 선행 효과란, 아주 짧은 시간 간격을 두고 (~40ms) 두 가지 다른 방향에서 소리가 들려올 때, 우리 뇌에서는 먼저 도착한 소리의 공간정보만 인식하고, 나중에 도착한 소리의 공간정보는 무시한다는 것이다.

4.2 반향 구별 classifier

우리는 인간의 선행효과를 모방하여, 현재 프레임의 음원 방향 추정 결과가 반향에 의한 것인지 아닌지를 구별하는 classifier 를 제안하고자 한다.

먼저, △-power 필터를 다음과 같이 정의한다.

$$f_{\gamma,\delta}(n,\theta)=\gamma\cdot f_{\gamma,\delta}(n-1,\theta)+\mu_{\delta}(\Delta p)\cdot R(n,\theta)$$
$$\text{, where } \mu_{\delta}(\Delta p)=\frac{1}{(1+\exp(-2(\Delta p-\delta)))} \quad (10)$$

여기서, n 은 현재 프레임 번호, Δp 는 이전 프레임과 현재 프레임간 음성신호의 파워 변화량, $R(n,\theta)$ 는 현재 프레임의 교차 방위 상관도이다. △-power 필터는 γ 와 δ 의 두 가지 파라미터를 가지게 되며, γ 는 필터의 시간 특성을, δ 는 필터의 △-power 특성을 규정하게 된다. 다양한 γ 와 δ 에 대해 D 개의 △-power 필터를 필터 뱅크로 구성하고, 이로부터, D-차원 음원 특징 벡터를 구성하게 된다.

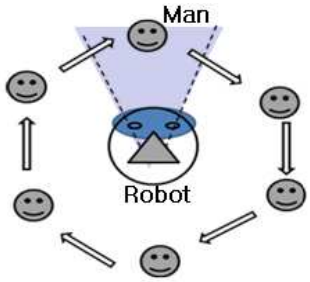
$$(\zeta_{\gamma_1,\delta_1}(n),\zeta_{\gamma_2,\delta_2}(n),...,\zeta_{\gamma_D,\delta_D}(n))\in\mathbf{R}^D$$
$$\text{,where } \zeta_{\gamma,\delta}(n)=\sum_{\theta}f_{\gamma,\delta}(n,\theta)\cdot R(n,\theta) \quad (11)$$

이렇게 구성된 특징 벡터를 이용하여, 학습 데이터를 만들어 classifier 를 학습시키면, 각 프레임에서 계산된 CAC 에 대해 신뢰값 (confidence value)를 classifier 가 산출하게 되고, 기준 문턱값을 이용하여 반향인지 아닌지 구별하게 된다. 이때 사용되는 classifier 는 어느 종류를 사용하나 무리가 없으나, 본 연구에서는 잘 알려져 있는 인공신경망 (artificial neural network)를 사용하였다.

4.3 음성-영상 융합 반향 학습

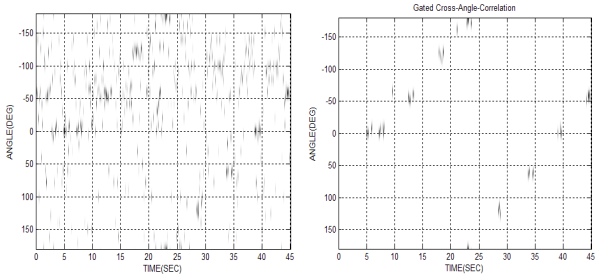
반향의 정도나 패턴은 장소에 따라 달라질 수

Fig. 5 SIL-BOT platform



* RT_{60}=0.6sec

Fig. 6 Experiment in a large hall



(a) Original CAC (b) Echo-removed CAC

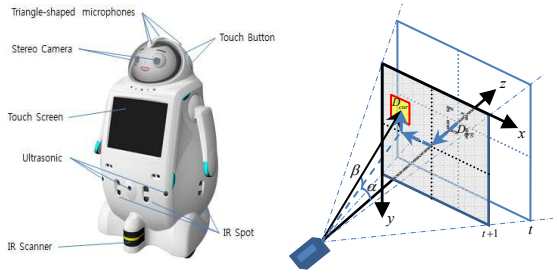
Fig. 7 Experiment result

밖에 없으며, 같은 방에서도 벽과 가까운 위치와 방 한가운데서의 반향 패턴은 다를 수 밖에 없다. 따라서, 서비스 로봇과 같이 이동하는 플랫폼에서는 실시간으로 반향 구별 classifier 를 학습 시키는 것이 중요하다. 그렇지만, 음성 센서만으로 실시간으로 학습 데이터를 생성하는 것은 많은 어려움이 따른다. 이때, 영상 신호 처리의 얼굴 검출 기술을 이용하면, 실시간으로 음원(발화자)의 위치를 파악하는 것이 가능하고, 반향 구별 classifier 의 학습을 자연스럽게 수행할 수 있다.

Fig. 5 는 우리가 음성-영상 융합 반향 학습 기술을 구현한 로봇 플랫폼을 보여준다. 로봇의 정수

Table 2 Classifier Performance[9]

Room	Hit[frames]	Miss[frames]	
		Pass invalid	Block valid
Large Hall	2197 (87.77%)	195 (7.79%)	111 (4.43%)



(a)SIL-BOT platform (b) Face position on the image

Fig. 8 Human tracking system

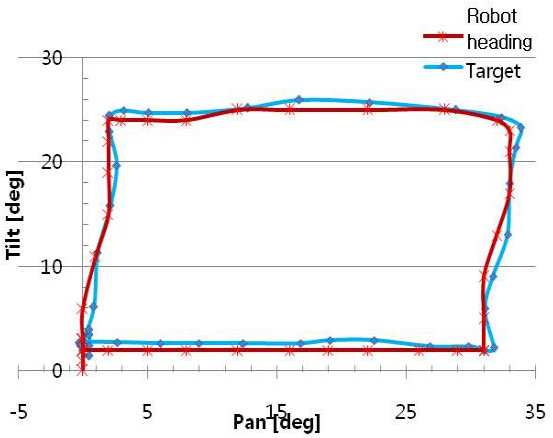


Fig. 9 Human tracking for moving target

리 부분에 3 개의 마이크로폰이 위치하고 있으며, 두 개의 카메라를 장착하고 있다. (얼굴 검출에는 하나의 카메라만 사용된다.) Fig. 6 과 Fig. 7 은 RT60 가 약 0.6sec 인, 어느 정도 반향이 있는 공간에서의 음원 방향 추정 실험과 실험 결과를 각각 보여준다. 로봇은 고정되어 있으며, 실험자가 로봇 정면에서 시작하여 로봇 주위로 60 도씩 이동하며 발화를 한다. 실험자가 로봇의 정면 시야각 내에 있을 때에만 로봇은 반향 학습을 하게 되고, 이후, 반향 구별 classifier 에 의해 반향음으로 판별된 CAC 결과는 제거된다. Fig. 7 의 (a)는 반향음 CAC 가 제거되기 전의 다소 혼란스런 음원 방향 추정 결과이며, (b)는 반향음 CAC 가 제거된 음원 방향 추정 결과이다. Table 2 는 전체 2500 여 frame 에 대하여 반향음 CAC 가 제거된 경우의 정량적인 결과를 나타낸다.

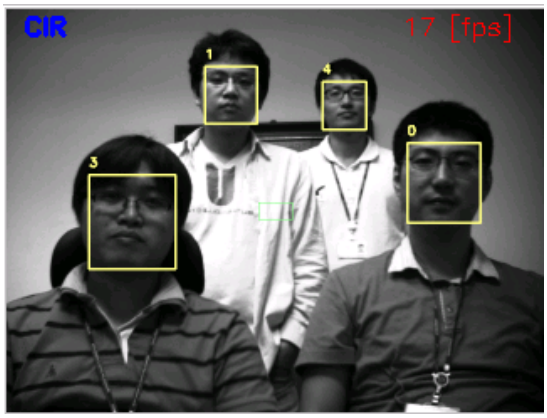


Fig. 10 Human attention on the image

5. 음성-영상 융합 사람 찾기 기술

5.1 얼굴 추적

발화자 및 검출된 사람을 찾아내고 해당 사용자에 대하여 지속적인 추적이 가능하도록 하기 위하여 Fig. 8(a)와 같이 3 개의 마이크로 폰 및 스테레오 카메라를 장착한 로봇 플랫폼을 구성 하였다. 또한, 강인한 얼굴 추적을 수행하기 위하여 카메라 이미지 상에서 얼굴 검출기를 통해 얻어진 얼굴 정보를 Fig. 8(b)에서 보는 바와 같이 이미지의 중심과의 각도 오차로 환산하여 항상 이미지의 중심에 맞추어지도록 설계하였다.

$$\alpha = \tan^{-1}\left(\frac{W/2-u}{f}\right), \beta = \tan^{-1}\left(\frac{v-H/2}{f}\right) \quad (12)$$

여기서, f 는 초점 거리, W 와 H 는 이미지의 가로 및 세로 크기, u 와 v 는 이미지 내 검출된 얼굴의 위치, 그리고 α 와 β 는 이미지 중심에서 가로 및 세로 방향에 대한 각도를 나타낸다.

5.2 Human Attention

특정 사람에 대하여 주목하는 기술은 마이크를 통하여 사람의 음원 위치를 추정하고 카메라를 이용하여 얼굴 검출 기술로 사람의 위치를 추정하는 단계, 그리고 Wheel 및 Pan-tilt 를 이용하여 추정 위치로의 이동이 가능하도록 하는 사람 검증 단계, 마지막으로 검증된 사람에 대하여 항상 Eye Contact 의 효과를 보여주기 위한 단계로 구분된다. Fig. 9 는 이미지 내의 사람이 상하좌우로 30 도를 이동하였을 때 Pan-tilt 를 이용한 로봇의 얼굴 지향 방향이 사람 얼굴을 강인하게 추적하는 성능을 보여준다.

Fig. 10 에서 보는 바와 같이 이미지 내에 다수

의 얼굴 검출 정보를 추출하고 각 얼굴 별 id 를 부여 함으로써 각종 상황 및 해당 id 별 얼굴 추적이 가능하도록 구성되어 있다.

6. 결 론

본 발표에서는 음성 기술과 영상 기술을 융합하여 반향 환경을 학습하는 기술과 로봇의 어텐션 기술을 구성하는 것을 소개하였다. 음성과 영상 기술은 조합 되어 사용할 때, 좋은 성능뿐만 아니라, 다양한 기술 개발이 가능하다. 때문에, 앞으로도 많은 음성-영상 융합 기술이 연구 개발될 것으로 기대한다.

후 기

본 연구는 지경부 21 세기 프론티어 인간기능 생활지원 지능로봇 기술개발사업의 일환으로 진행되었음을 밝히며, 관계자 여러분께 감사를 전합니다.

참고문헌

(1) Nakadai, K., Hidai, K., Okuno, H.G. and Kitano, H., 2001, "Real-Time Multiple Speaker Tracking by Multi-Modal Integration for Mobile Robots," in *Proc. Eurospeech 2001*, pp. 1193~1196.

(2) Lim, Y. and Choi, J., 2009, "Speaker Selection and Tracking in a Cluttered Environment with Audio and Visual Information," *IEEE Trans. Consumer Electronics*, Vol. 55(3), pp. 1581~1589.

(3) Hornstein, J., Lopes, M., Santos-Victor, J. and Lacerda, F., 2006, "Sound Localization for Humanoid Robots – Building Audio-Motor Maps based on the HRTF," in *Proc. IEEE/RSJ IROS 2006*, pp. 1170~1176.

(4) Chan, V., 2009, "Audio-Visual Sensor Fusion for Object Localization," *INE NewsLetter,* 8 June.

(5) Zabin, R. and Woodfill, J., 1994, "Non-Parametric Local Transforms for Computing Visual Correspondance," In *Proc. the 3rd European Conference on Computer Vision*, pp.151~158.

(6) Froba, B. and Ernst, A., 2004, "Face Detection with the Modified Census Transform," in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pp.91~96.

(7) Jun, B.-J. and Kim, D., 2007, "Robust Real-time Face Detection Using Face Certainty Map," in *Proc. ICB 2007*, pp.29~38.

(8) Haas, H., 1972,"The Influence of a Single Echo on the Audibility of Speech," *Journal of the Audio Engineering Society*, Vol. 20, pp.146~159.

(9) Lee, B.-G., Choi, J. S., Kim, D. and Kim, M., 2010, "Verification of Sound Source Localization in Reverberation Room and its Real Time Adaptation Using Visual Information," in *Proc. ARSO2010*, pp.176~181.