Proceedings of the 2007 IEEE/RSJ International
Conference on Intelligent Robots and Systems
San Diego, CA, USA, Oct 29 - Nov 2, 2007

Auditory and Visual Integration based Localization and Tracking of Humans in Daily-life Environments

Hyun-Don Kim, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno

Abstract—**The purpose of this research is to develop techniques that enable robots to choose and track a desired person for interaction in daily-life environments. Therefore, localizing multiple moving sounds and human faces is necessary so that robots can locate a desired person. For sound source localization, we used a cross-power spectrum phase analysis (CSP) method and showed that CSP can localize sound sources only using two microphones and does not need impulse response data. An expectation-maximization (EM) algorithm was shown to enable a robot to cope with multiple moving sound sources. For face localization, we developed a method that can reliably detect several faces using the skin color classification obtained by using the EM algorithm. To deal with a change in color state according to illumination condition and various skin colors, the robot can obtain new skin color features of faces detected by OpenCV, an open vision library, for detecting human faces. Finally, we developed a probability based method to integrate auditory and visual information and to produce a reliable tracking path in real time. Furthermore, the developed system chose and tracked people while dealing with various background noises that are considered loud, even in the daily-life environments.**

I. Introduction

Techniques that allow humans and robots to interact are essential so that robots can detect and understand human intentions and emotions. Therefore, for robots to interact effectively with people, they must be able to identify people in different social and domestic environments, pay attention to their voices, look at speakers to identify them visually, and track them while integrating auditory and visual information [1-5]. The system we developed has some principal techniques that enable humans to be tracked for effective human-robot interactions.

First, robots require the ability to localize sound sources to find the location of the talkers. However, to localize sound sources in environments where several sounds are present, conventional methods require a microphone array, which usually consists of eight microphones, and/or additional information such as impulse response data [3]. Therefore, a method is needed that uses just two microphones and does not need impulse response data. Nakadai et al. have already developed a system to localize multiple sound sources only using two microphones [1, 2]. However, this system can even estimate the impulse response when the shape of a robot's head is a sphere. Therefore, if the shape of the robot's head is changed, the system's capability will be affected. To avoid this, we developed a method to integrate a cross-power spectrum phase analysis (CSP) method [6] and an EM algorithm [7]. This method can localize several moving sound sources by only using two microphones, and it does not need impulse response data. A robot's appearance is variable as long as it knows the delay of arrival (DOA) between two microphones. Moreover, we have confirmed that our method performs better than the conventional methods.

Second, localizing human faces is also necessary for accurately selecting a desired person. OpenCV, a open vision library, has recently become popular and is used to detect faces [5]. However, OpenCV has difficulty detecting faces that are turned or tilted. Although the face detection system included a method to incorporate skin color, detection relied heavily on the illumination and color status. Accordingly, robots have difficulty coping with various races and operational situations. Therefore, we developed a method that can perform face localizations by using skin color features [8] extracted from faces detected by using OpenCV. Also, the robot can recreate the color feature whenever the illumination or background condition changes. To cope with several faces and accurately estimate face areas, we use an EM algorithm to classify skin color distribution extracted by using the skin color feature.

Finally, robots must focus on a specific person to detect and understand their intentions and emotions. Therefore, human tracking is an important part of a human and robot interaction. The human robot interaction systems by using integrating audio and video information have been developed in various forms. For example, Nakadai et al. developed real-time multiple-talker tracking system based on auditory and visual integration [1, 2], HRP-2 of AIST [3] can track a human according to an azimuth, and SIG2 of Tasaki et al. [4] can perform various actions according to a distance by the fusion of audio-visual information. Unfortunately, since these systems did not have abilities to cope with a loud noise and/or a dynamically changed environment. For this reason, we developed a probability based method to integrate audio and video information. Since the developed method is a simple and compact, robots can execute a real time auditory and visual integration. Also, it is easy to manage and modify the program. Moreover, our system was able to track humans

Hyund-Don Kim, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno are with Speech Media Processing Group, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan.
{hyundon, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp.

1-4244-0912-8/07/$25.00 ©2007 IEEE.

while dealing with various background noises, such as music played from audio components or voice signals generated by a TV or radio, typically found in everyday environments.

II. Sound Source Localization

Many methods for sound source localization have been developed and their performance has been steadily improved. Three typical methods for sound source localization are used: HRTF [1, 2], MUSIC [3], and CSP [6]. However, the features of each method differ. For example, HRTF and MUSIC need impulse response data, and their performance degrades when the robot's shape is transformed or local conditions change. Moreover, those methods need to interpolate to manipulate a moving talker because it is available only for discrete azimuth and elevation. However, these methods perform well in a fixed environment and can detect multiple localizations of mixed sounds entering from different directions. On the other hand, CSP can even seldom find multiple localizations simultaneously, does not need impulse response data, and can accurately find the direction of a sound. Therefore, we used a CSP method for sound source localization. The CSP method can usually estimate one delay of arrival (DOA) at a frame. However, for multiple sound localizations, we can estimate multiple directions of sound after we have gathered the CSP results for three frames (one frame consists of 1024 samples) so that the EM algorithm can estimate the distribution of the CSP results.

A. CSP (Cross-power Spectrum Phase analysis)

The direction of a sound source can be obtained by estimating the time delay of arrival (TDOA) between two microphones [5]. When there is a single sound source, the TDOA can be estimated by finding the maximum value of the cross-power spectrum phase (CSP) coefficients [6], as derived from

$$csp_{ij}(k) = IFFT\left[\frac{FFT\left[s_i(n)\right]FFT\left[s_j(n)\right]^*}{\left|FFT\left[s_i(n)\right]\right|\left|FFT\left[s_j(n)\right]\right|}\right] \quad (1)$$

$$\tau = \arg\max\left(CSP_{ij}(k)\right) \quad (2)$$

where k and n are time delays, FFT (or IFFT) is the fast Fourier transform (or inverse FFT), * is the complex conjugate, and τ is the estimated TDOA. The sound source direction is derived from

$$\theta = \cos^{-1}\left(\frac{v\cdot\tau}{d_{\max}\cdot F_s}\right) \quad (3)$$

where θ is the sound direction, v is the sound propagation speed, F_s is the sampling frequency, and d_{max} is the distance with the maximum time delay between two microphones. The sampling frequency of our system is 16 kHz.

B. Localization of multiple moving sounds by EM

Using CSP with two microphones can locate a specific sound source in a frame even if several sound sources are present. Because CSP is unreliable in noisy environments, we developed a new method to estimate the number and localization of sound sources based on probability. To use this method, we first need to gather the CSP results for three frames (shifting every half a frame). Then, the EM algorithm is used to estimate the distribution of the gathered data [7].

Figure 1 (A) shows the sound source localization events extracted by CSP according to time or frame lapses. As shown in this figure, events that lasted 192 ms are used to train the EM algorithm to estimate the number and localization of sound sources. We experimentally decided that the appropriate interval for the EM algorithm was 192 ms. If the interval for the EM algorithm was increased, dealing with sounds that are moving fast would be difficult. Figure 1 (B) shows the training process of the EM algorithm. In detail, events for 192 ms are first converted into histograms. Next, Gaussian components defined by using equation (8) for training the EM algorithm are uniformly arranged on whole angles. Then, the histogram data is applied to the arranged Gaussian components. After training the EM algorithm, the arranged Gaussian components are relocated based on the density and distribution of the histogram data. Finally, if components overlap, the mean and variance of Gaussian components will be one, and each weight value will be added. In addition, components with scant weights are regarded as noise and are removed. Figure 1 (C) shows the results of localizing sound sources by iterating processes (A) and (B) in the same way. The interval for EM training is shifted every 32 ms.

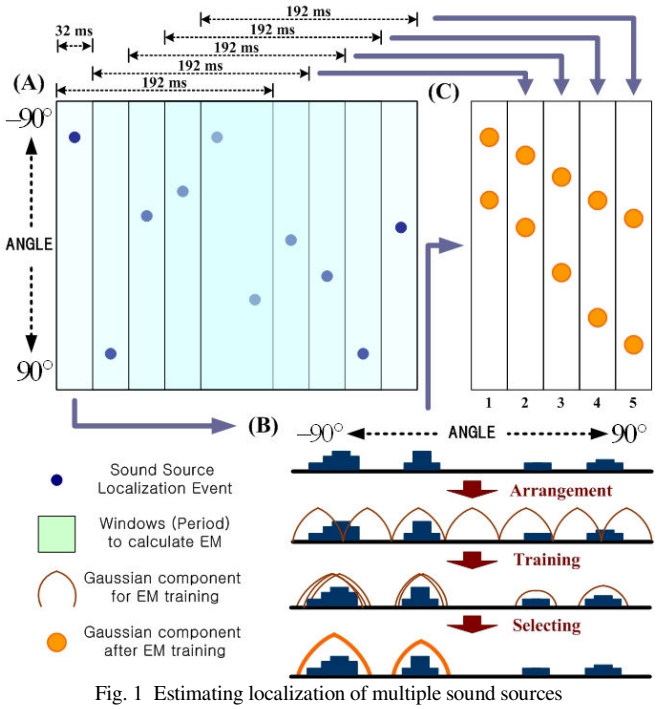

Fig. 1 Estimating localization of multiple sound sources

C. Experiments and Results

The following conditions were used in the experiment to

evaluate localization: the sound sources were 1.5 m from the head of a robot, and the sounds emitted from speakers in the magnitude of 85 dB were recorded female and male speech. We obtained the sound source localization results while the robot's head was rotating from 90° to -90°. By rotating the head, we created the effect of moving sound. The rotation speeds used in these experiments was 1.1 m/s, and this is faster than the average walking speed, 1.0 m/s, of healthy adults. Figure 2 (A) and (C) show the results achieved with the developed method. To compare performance, we also experimented using a conventional method that used HRTF with scattering theory [1, 2] under the same conditions. Those results are shown in Figure 2 (B) and (D).

Figure 2 shows the results of the sound source localization experiment when two sound sources were used with a gap of 60° and 30° between the sound sources. The red dotted lines inside Figure 2 indicate the reference line for the correct track of moving sounds. Based on the results, the developed method (A) and (C) was more accurate than the conventional method (B) and (D) because it could hardly distinguish the two sound sources. The results indicate that the developed method can accurately locate two sound sources moving at 1.1 m/s and with a gap wider than 30° between the two sources. In other words, our method enables robots to locate the voices of people who are walking.

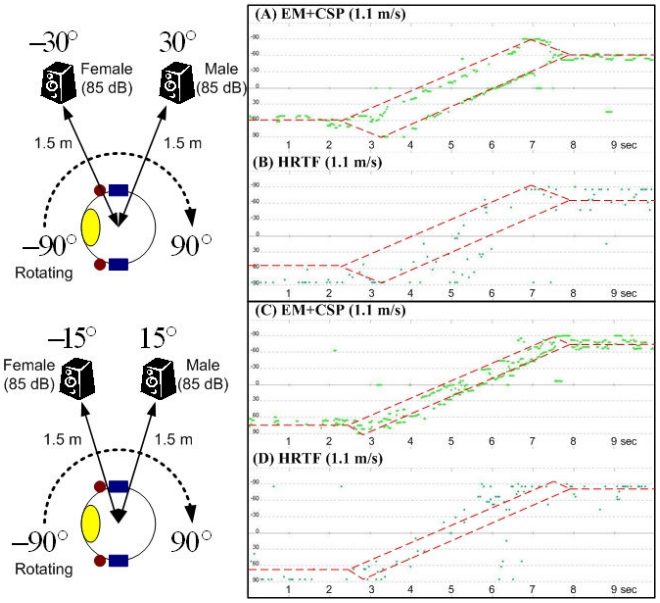

Fig. 2 Sound source localization results

III. Face Localization

Face localization is also necessary for selecting a desired person. The simplest method to detect faces is to use a skin color [8]. However, this method is limited because a performance relies heavily on illumination condition and background color status. Also, dealing with various colored races is difficult. Therefore, robots cannot easily detect faces based on skin colors in environments that are constantly changing. An open vision library called OpenCV has become popular for face detection [5]. However, OpenCV is limited in that, for example, it cannot detect faces that are turned or tilted and has difficulty detecting over 2 m away in the case of 320 x 240 images. To overcome these limitations, we developed a method that after determining face skin color features using faces detected by OpenCV, the constructed color feature extracts the face skin color in images. Robots can then automatically recreate the face skin color feature whenever illumination or background condition changes. Moreover, to accurately localize several different faces, we used an EM algorithm to classify skin colors distribution extracted using the skin color feature. Therefore, the developed method can reliably localize faces regardless of changing environments and various human races.

A. Skin Color Clustering

For color clustering, each pixel in the image is described by a feature vector, $\mathbf{f}=[\mathbf{c}\ \mathbf{p}]^T$, where, $\mathbf{c}=[Y\ U\ V]^T$ describes the color and $\mathbf{p}=[x\ y]^T$ describes the row and column of the pixel. Target pixels in a selected target area are presented as $\mathbf{f}_T(i)=[\mathbf{c}_T(i)\ \mathbf{p}_T(i)]^T$, where $i=1\sim k$, k is the number of the target color, background pixels in the selected background areas are presented as $\mathbf{f}_N(j)=[\mathbf{c}_N(j)\ \mathbf{p}_N(j)]^T$, where $j=1\sim m$, m is the number of the background color, and an unknown pixel is described by $\mathbf{f}_U=[\mathbf{c}_U\ \mathbf{p}_U]^T$.

The $\mathbf{f}_U$ can be classified by comparing the distance d_T.

$$d_T = \min_{i=1\sim k} \left\| \mathrm{f}_T(i) - \mathrm{f}_U \right\|^2 \tag{4}$$

and the minimum distance d_N.

$$d_N = \min_{j=1\sim m} \left\| \mathrm{f}_N(j) - \mathrm{f}_U \right\|^2 \tag{5}$$

If $d_T < d_N$, $\mathbf{f}_U$ is classified as a target pixel, otherwise a background pixel.

B. Accurate Face Localization by using EM

For the purpose of face localization, a robot firstly uses OpenCV. If OpenCV fails to detect a face, the system will try to detect faces using skin color. To detect a face this way, we need to first create a Lookup table that contains skin and background color information. This creation requires that a robot should initially detect the face by using OpenCV. The system can then extract a target (face skin color) feature, $\mathbf{f}_T$, and a background feature, $\mathbf{f}_N$, from areas inside and outside the detected face. The top image in Figure 3 is a detected face by using OpenCV and the designated area used to create feature vectors. Step (A) indicates that the created $\mathbf{f}_T$ and $\mathbf{f}_N$ features are saved in the Lookup table and are used to detect faces by using skin color. Since creating the Lookup table needs about 0.5 second (Celeron 2.4 GHz, 512 M ram, and 320 x 240 image), the robot only updates the Lookup table when the update conditions are satisfied: First, not many skin colors are present in the detected face when OpenCV succeeded in a face detection; and second, many skin colors are present in the background areas. Our system uses 320 x

240 images and can calculate about five images per second without updating the Lookup table.

Figure 3 shows the process used to extract human faces by using skin colors and the EM algorithm. Step (B) is where an image is captured from a camera and is converted to an image that contains only pixel related to skin color determined by equations (4) and (5) with the constructed Lookup table. The circles in step (C) indicate the arranged Gaussian components defined by equation (9) and are used to find the number and area of faces in the image. Step (D) represents the relocated Gaussian components based on the density and distribution of the skin color data after executing the EM algorithm. Step (E) is where the number and size of faces are estimated using the means and variances obtained by executing the EM algorithm. If the variances calculated by EM algorithm in this step are broad and many skin colors exist in the background area, the robot will create new color feature vectors and create a new Lookup table as shown in step (F) based on satisfying the second update condition.

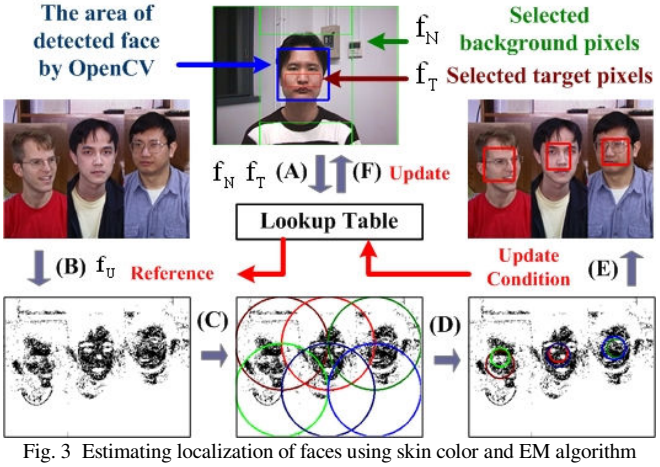

Fig. 3 Estimating localization of faces using skin color and EM algorithm

C. Experiments and Results

We experimented using the Georgia Tech face data base (we downloaded a zip file at www.anefian.com/face_reco. htm), which contains images of 50 people. First, for single face detection, we chose 60 images of 17 people, including faces that are turned or tilted. For several faces detection, we created 20 images including three people. The face detection results are listed in Table I.

TABLE I
FACE DETECTION RESULTS

Face Detection	Single Face		Several Faces	
	OpenCV	Skin Model	OpenCV	Skin Model
Success rate	52%	80%	58%	72%

Success rate in Table I indicates the percentage of when the areas of faces are accurately detected. The success rate of a skin model for single faces increased 28% compared to that using only OpenCV. The Lookup table to classify skin colors was automatically updated seven times based on satisfying the update conditions as shown in Figure 3. That is to say, our system will update the Lookup table when the skin color is not or few inside the detected face by OpenCV and/or when a lot of skin colors exist at the background area. For several faces, we got the performance increasing 14% compared to using only OpenCV. At that time, the updates of six times was performed and we used additional 20 images including a single face with 20 images including three faces because our system can do update when there is a single face each image (See Figure 3). Since updating the Lookup table requires much execution time, the robot should just update when the update condition is satisfied. Figure 4 shows some of the result images detected by a skin model even if OpenCV failed to detect those because most faces were turned or tilted. The red boxes indicate the faces detected by using the skin model and EM algorithm.

Fig. 4 Faces detected by using skin color and EM algorithm

D. Face Distance Estimation

To estimate distances between detected faces and a pair of CCD cameras, we used a correlation method as derived from

$$R(x_k, y) = \sum_{i=0}^{W} \sum_{j=0}^{H} \left[T(x_i', y_j') \cdot I(x_k + x_i', y + y_j') \right] \quad (6)$$

where T is the template image captured from a right CCD camera and template images are detected faces, x' and y' describes the row and column of the pixel, W and H are the width and height of detected faces respectively, I is the image captured from a left CCD camera, R is a correlation result, the range of x_k is from $-W/2$ to $W/2$, and y indicates y'_0. Then, we can estimate the distances of detected faces through calculating the maximum value of R as derived from

$$\tau_{\text{distance}} = \arg\max_k R(x_k, y) \quad (7)$$

where τ_{distance} is the estimated disparity between the right and left image.

IV. EM (Expectation-Maximization) Algorithm

We used an EM algorithm to process sound and face localization. This algorithm allows us to effectively classify the number and area of the distributed data. To use EM, we need to first properly arrange Gaussian components before the system runs the E-step and M-step. We used a one-dimensional Gaussian model, denoted as equation (8), for sound localization and the Bivariate Gaussian model, denoted as equation (9), for face localization.

$$P\left(X_m|\theta_k\right)=\frac{1}{\sqrt{2\pi\sigma_k{}^2}}e^{-\frac{(X_m-\mu_k)^2}{2\sigma_k^2}} \tag{8}$$

$$P\left(X_{m,xy}|\theta_{k,xy}\right)=\frac{1}{2\pi\sigma_{k,x}\sigma_{k,y}}e^{-\frac{1}{2}\left[\left(\frac{X_{m,x}-\mu_{k,x}}{2\sigma_{k,x}}\right)^2+\left(\frac{X_{m,y}-\mu_{k,y}}{2\sigma_{k,y}}\right)^2\right]} \tag{9}$$

where μ_k is the mean, σ^2_k is the variance, θ_k is a parameter vector, m is the number of data, x and y are the values of m-th data on the basis of a x axis and a y axis in two dimensions, and k is the number of mixture components. The objective is to find the parameter vector, θ_k, describing each component density, $P(X_m|\theta_k)$, through iterations of the E and M step. This EM step is described as follows:

1) E-step: The expectation step essentially computes the expected values of the indicators, $P(\theta_k|X_m)$, that each data point X_m is generated by component k, given N is the number of mixture component, the current parameter estimates θ_k and weight w_k, using Bayes' Rule derived as

$$P\left(\theta_k|X_m\right)=\frac{P\left(X_m|\theta_k\right)\cdot w_k}{\sum_{k=1}^{N}P\left(X_m|\theta_k\right)\cdot w_k} \tag{10}$$

2) M-step: At the maximization step, we can compute the cluster parameters that maximize the likelihood of the data assuming that the current data distribution is correct. As a result, we can obtain the recomputed mean using equation (11), the recomputed variance using equation (12), and the recomputed mixture proportions (weight) using equation (13). The total number of data is indicated by M.

$$\mu_k=\frac{\sum_{m=1}^{M}P\left(\theta_k|X_m\right)X_m}{\sum_{m=1}^{M}P\left(\theta_k|X_m\right)} \tag{11}$$

$$\sigma^2_{k,xy}=\frac{\sum_{m=1}^{M}\left(X_{m,x}-\mu_{k,x}\right)\left(X_{m,y}-\mu_{k,y}\right)P\left(\theta_k|X_m\right)}{\sum_{m=1}^{M}P\left(\theta_k|X_m\right)} \tag{12}$$

$$w_k=\frac{1}{N}\sum_{m=1}^{M}P\left(\theta_k|X_m\right) \tag{13}$$

After the E and M steps are iterated an adequate number of times, the estimated mean, variance, and weight based on the current data distribution can be obtained.

V. Human Tracking System

A. System Overview

Figure 5 shows an overview of the structure of our system used in human tracking and a humanoid robot called SIG2. The robot has two omni-directional microphones inside humanoid ears at the left and right ear position. Its head has three degree of freedom (DOF) and the body has one DOF, each of which is enabled by a DC motor controlled by an encoder sensor. SIG2 is equipped with a pair of CCD cameras. Our system consists of five modules (audition, vision, motor, viewer, and tracking). The audition module generates sound events using sound source localization. Specifically, after it judges that sound signals exist by using the first value of mel-frequency cepstral coefficients (MFCC) [9], it locates the sound source with the CSP method. The vision module generates face events using face localization and each face event includes the distance information of a detected face. The tracking module can track humans by using events extracted by the subsystems. The position of a desired path is sent to the motor module to turn the robot's head.

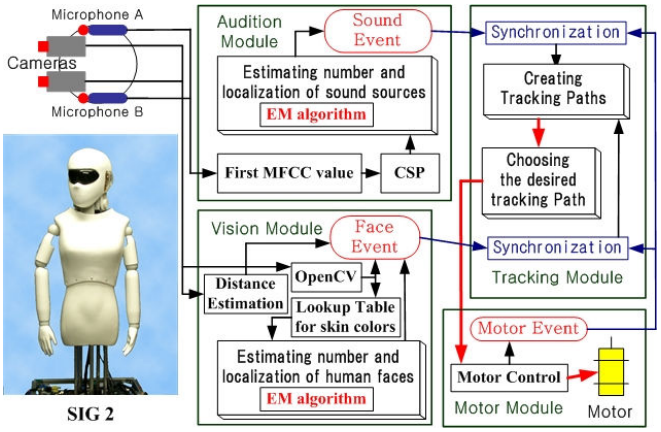

Fig. 5 System overview

B. Auditory and Visual Integration

How to integrate auditory and visual information is an important problem for improving the information processing ability. Since the aim of this research is to choose and track a desired person, we estimated the location of targets using a probability based method to integrate sound and face localization.

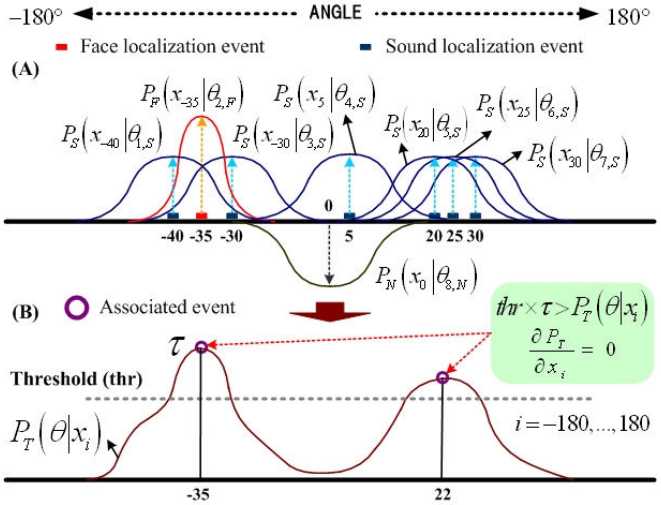

Fig. 6 Integrating auditory and visual information

Figure 6 represents auditory and visual integration. A tracking module initially receives events generated from the audition and vision modules every 0.1 second. At this time, each event includes the angle information. Next, the received events are applied to Gaussian models, as indicated by equation (8). Step (A) in Figure 6 is where parameters are

determined in which the number of received events is k, the angle of each events is μ_k, the variance of events is σ^2_k, θ_k is each parameter vector, x_i is an angle, and $i=-180,..,180$. In other words, Gaussian models are created for every angle in each generated event. The variance of Gaussian models for face events is narrow at that time because the accuracy range of face localization is normally narrower than that of sound source localization. Second, corresponding probability is calculated using

$$P\left(x_i\middle|\theta_{k,FS}\right)=P_F\left(x_i\middle|\theta_{k,F}\right)+P_S\left(x_i\middle|\theta_{k,S}\right)-P_N\left(x_i\middle|\theta_N\right) \quad (14)$$

where P_F is the probability of face localization, P_S is the probability of sound localization, and P_N is the probability to cope with sound noise. Therefore, if sound noise did not happen at angle i, $P_N(x_i|\theta_{k,N})$ will be 0. The total probability density can be calculated using

$$P_T\left(\theta\middle|x_i\right)=\frac{\sum_{k=1}^{n}P\left(x_i\middle|\theta_k\right)\cdot w_k}{\sum_{i=-180}^{180}\sum_{k=1}^{n}P\left(x_i\middle|\theta_k\right)\cdot w_k} \quad (15)$$

where n is the total number of received event, w_k is the weight concerning each Gaussian component. Here, w_k is fixed in $1/n$ when sound events are only present, but when face events are present, face events generated at near distance have the higher w_k than face events generated at far distance. Figure 6 (B) shows the result calculated when using equation (15) with whole angles. The maximum value, τ, can be calculated using

$$\tau=\arg\max_i P\left(\theta\middle|x_i\right),\ i=-180,...,180 \quad (16)$$

Finally, we obtained two positions, -35 and 22 degree, which are peak values in intervals satisfying the condition where $thr\cdot\tau > P_T(\theta|x_i)$. These positions were used to choose and track a desired person, and we call those associated events.

C. Creating and Choosing a tracking path

Robots must create tracking paths, including the status of received events, based on the real time so that robots can deal with various situations. Then, after the robot choose appropriate tracking path, it should track a designated path in order to interact with humans. In our system, the left part of Figure 7 shows that the robot made associated events every 0.1 second by using equation (15) and (16).

To create a tracking path, the present tracking position can be estimated from some past tracking positions by using this model:

$$x_t=x_{t-1}+\dot{x}_{t-1}\cdot T_s+\ddot{x}_{t-1}\frac{T_s^2}{2} \quad (17)$$

where T_s is a sample period, x_t is a estimated tracking position, x_{t-1} is the previous tracking position, $\dot{x}_{t-1}$ is the differential value between x_{t-1} and x_{t-2}, and the differential value between $\dot{x}_{t-1}$ and $\dot{x}_{t-2}$ is $\ddot{x}_{t-1}$.

The tracking path includes associated events within a range of $\pm 15°$ of the expected position calculated by using equation (17), shown as step (B) in Figure 7. Tracking paths can then be continuously created every 100 ms. As shown in step (B) in Figure 7, the tracking path has not started if some associated events are missing in the intervals of some of the following frames. On the other hand, once the tracking path has started, the tracking path can be continuously estimated even if some associated events are missing in a few frames. Also, the tracking path is terminated if some associated events are missing at the adjacent interval of several frames.

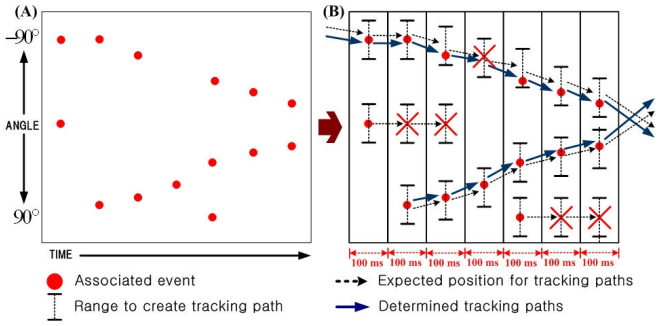

Fig. 7 Estimating tracking paths

To achieve reliable human tracking in a real environment, the developed system was designed with the following points in mind: First, if several tracking paths are created, the system will follow the first tracking path created or the tracking path nearest the current motor position, the angle of the robot' head. Second, various sounds are present in domestic environments, such as music played from audio components or voice signals generated by a TV or radio. Therefore, to identify noise, the robot first turns a camera to the direction of the noise. If a face event is not extracted in 3 seconds and the range of location variations extracted by sound source localization is narrow, in other word, the sound source is not moving and a human face is not detected, the robot will regard the sounds as noise. The robot then applies P_N of equation (14) to the angle of the detected noise. Therefore, the robot can create associated event only using face events where noise is effective. On the other hand, if noise is not generated from a designated noise direction for several seconds, the robot will release sounds entering from the noise direction from noise. Third, the path including sound events has higher priority for tracking than the path including the face events. Finally, face events generated at a near distance has higher priority for tracking than face events generated at the far distance. Also, if several face events occurred in 0.1 second, the robot will adjust the weight proportion, w_k, of equation (15) in order to track the face that is closer to the robot. As a result, the robot can reliably track and choose a designated person while reducing the effect of noise.

D. Experiments and Results

Figures 8 represents events generated from an audition

module and a vision module, the current position of a motor, and the status of human tracking received from the tracking module. The red lines indicate created tracking paths and the red rings indicate the selected tracking path for human tracking. We experimentally evaluated our methods ability to track humans while coping with domestic noises. (A) in Figure 8 shows that the robot was tracking a human face which was close to the robot. (B) shows that the robot responded to a call of a person at -60° and changed the tracking target. Then, after the robot turned its head to the call direction, it continued to track a person using face events. (C) shows that after the robot turned its head to 80° where sounds whose magnitude was 85 dB happened, it checked whether there is a human face or not. If face localization events is not created and the range of location variations of sound events is narrow for 5 seconds, the robot will discontinue tracking the sounds because it will regard the sounds as noise generated by equipment. (D) shows that the robot tracked the person through integrating sound and face events while dealing with noises. (E) shows that the robot used only face events at the noise range so as to track the person. (F) shows that after the robot automatically recognized that noises were extinct, it then continued tracking the person even in the former noise range while integrating of sound and face events.

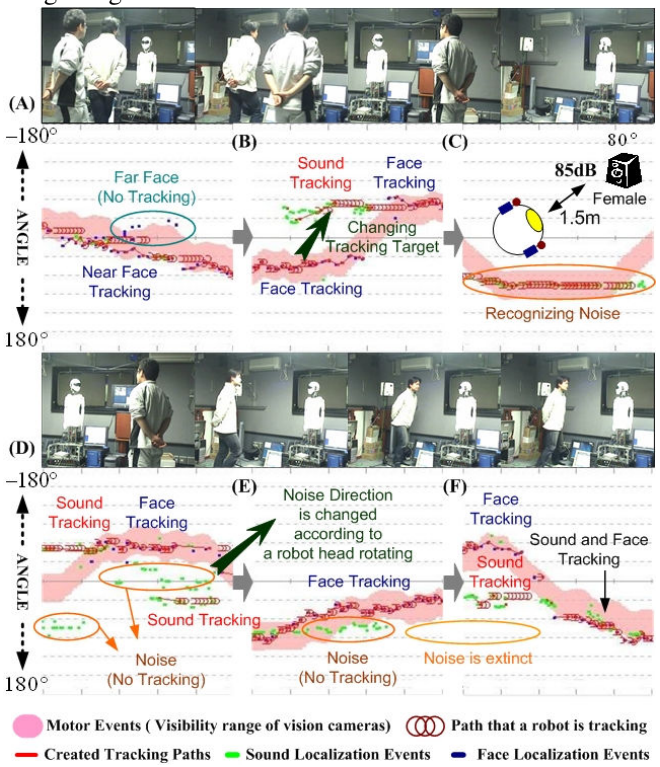

Fig. 8 Human tracking results in noisy environments

VI. Conclusion

We developed a system that enables robots to choose and track a desired person in daily-life environments and confirmed that our system performed well. Our system has some principal capabilities. First, the algorithm we developed can localize multiple sound sources even if it has only two microphones and a normal sound card device because the EM algorithm helps the system cope with multiple sound sources in real time. Results indicated that our system reliably located two sound sources moving less than 1.1 m/s and with a gap wider than 30° between two sources. Therefore, robots equipped with our system can simultaneously localize the voices of two people that are walking. Second, we combined the advantages of OpenCV and a skin color method to detect human faces. Therefore, a robot cannot be significantly affected by different angles of human faces and illumination conditions when performing face localization. An EM algorithm was shown to help face localization system classify human faces in images. Finally, to choose and track a desired person while dealing with the noise from electric home appliances, the robot produced tracking paths which integrated auditory and visual information. Therefore, the robot was able to choose and/or track a human by referring to tracking paths.

Acknowledgment

This research was partially supported by MEXT, Grant-in-Aid for Scientific Research, and Global COE program of MEXT, Japan.

References

[1] Kazuhiro Nakadai, Ken-ichi Hidai, Hiroshi Mizoguchi, Hiroshi G. Okuno, and Hiroaki Kitano, “Real-Time Auditory and Visual Multiple-Object Tracking for Humanoids,” in Proc. of 17th International Joint Conference on Artificial Intelligence (IJCAI-01), Seattle, Aug. (2001) pp. 1425-1432.

[2] Hiroshi G. Okuno, Kazuhiro Nakadai, Ken-ichi Hidai, Hiroshi Mizoguchi, and Hiroaki Kitano, “Human-Robot Interaction Through Real-Time Auditory and Visual Multiple-Talker Tracking,” in Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2001), Oct. (2001) pp. 1402-1409.

[3] I. Hara, F. Asano, Y. Kawai, F. Kanehiro, and K. Yamamoto, “Robust speech interface based on audio and video information fusion for humanoid HRP-2,” IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2004), Oct. (2004) pp. 2404-2410.

[4] T. Tasaki, S. Matsumoto, H. Ohba, M. Toda, K. Komatani, T. Ogata and H. G. Okuno, “Dynamic Communication of Humanoid Robot with multiple People Based on Interaction Distance,” RO-MAN2004 Int. Workshop on Robot and Human Interactive Communication, Okayama, Japan, Sept. 2004, pp.81-86.

[5] H. D. Kim, J. S. Choi, and M. S. Kim, "Speaker localization among multi-faces in noisy environment by audio-visual integration", in Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA2006), May (2006) pp. 1305-1310.

[6] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, “Localization of multiple sound sources based on a CSP analysis with a microphone array,” IEEE/ICASSP Int. Conf. Acoustics, Speech, and Signal Processing, June (2000) pp 1053-1056.

[7] T. K. Moon. “The Expectation-Maximization algorithm,” IEEE Signal Processing Magazine, Nov. (1996) 13(6) pp. 47-60.

[8] Chunsheng Hua, Haiyuan Wu, Qian Chen, Toshikazu Wada, "A Pixel-wise Object Tracking Algorithm with Target and Background Sample," 18th International Conference on Pattern Recognition (ICPR'06), pp. 739-742, 2006.

[9] J. K. Shah, A. N. Iyer, B. Y. Smolenski, and R. E. Yantormo, “Robust Voiced/Unvoiced classification using novel feature and Gaussian Mixture Model,” IEEE/ICASSP Int. Conf. Acoustics, Speech, and Signal Processing, Montreal, Canada, May, 2004.