Real-Time Active Human Tracking by Hierarchical Integration of Audition and Vision

Kazuhiro Nakadai[†], Ken-ichi Hidai[†], Hiroshi G. Okuno[†§], and Hiroaki Kitano [†¶]
† Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Corp.
Mansion 31 Suite 6A, 6-31-15 Jingumae, Shibuya-ku, Tokyo 150-0001, Japan
§ Department of Intelligence Science and Technology, Kyoto University, Kyoto 606-8501, Japan
¶ Sony Computer Science Laboratories, Inc., Tokyo 141-0022, Japan
{nakadai, hidai, okuno, kitano}@symbio.jst.go.jp

Abstract

We present a real-time multiple human tracking system for a humanoid with a pair of microphones and a pair of cameras. The key idea for robust recognition and scalability of multi-modality is the hierarchical integration of multiple sensory information. First, sound direction *by auditory localization module,* speaker ID *by spear identification module,* face location, face ID, *and* object location *by stereo vision, and* motor direction *are extracted as separate events. Next, each stream is formed as a temporal sequence of events. Then, different streams are associated according to their proximity. This association process consists of two stages. At the first stage, an association in* location *or* name *is performed. At the second stage, the higher level association between* location- *and* name-*associated streams may be performed. The hierarchical association enables integration of various kinds of streams and resolve ambiguities of lower level perceptions. The focus-of-attention is mainly controled by hierarchical association streams. As a result, the humanoid demonstrates robust human tracking even when several persons speak simultaneously.*

1 Introduction

Recently, robots such as HONDA ASIMO, Sony AIBO and SDR-3X receive much attention. They have capabilities of biped-walking and friendly simple interaction. These kinds of robots will be deployed for welfare work, housekeeping or as a pet in the future. Because such a robot is expected to behave as a partner of humans, it should have capabilities of perception and recognition. However, from the viewpoint of perception and recognition, a lot of research remains to be done. For example, robots should understand the surrounding environment autonomously by using various sensor information to select a proper action on demand. Such a capability enriches the social interaction between a robot and a human. But, currently, a robot cannot understand the surrounding very robustly, so the social interaction is not rich enough to behave as a partner of humans.

To realize rich social interaction with humans, a robot should at least have similar sensors to those humans have, and deal with information obtained by such sensors properly. In robots, vision is often used, while audition, which is one of five senses for humans, is not popular except for speech recognition. The reasons are as follows: 1. Auditory processing is less accurate. 2. It is difficult to capture sound from a single source in real world, because a microphone captures a mixture of sounds even when a directional microphone is used. 3. Auditory processing is sensitive to reverberation and acoustic change of a room. These affect sound source localization and speech recognition badly.

Robots with auditory processing such as *Kismet* [1] of MIT AI Lab and *ROBITA* [2] of Waseda Univ. have been developed. They have a function of speech recognition by using a microphone worn near the mouth. But they cannot use their own microphones for speech recognition because of a poor signal-noise ratio. They does not have capability of sound source separation. *ROBITA* can localize sound by its own stereo microphones besides speech recognition. But it does not take motor noise into account, so it has difficulty in auditory processing with motion.

To solve these problems and to make a robot with the capability of social interaction, we developed a real-time multiple object tracking system by using face recognition and active audition[3]. Active audition achieves auditory processing with motion by using sound source localization by auditory epipolar geom-

Proceedings of the 2001 IEEE–RAS International Conference on Humanoid Robots

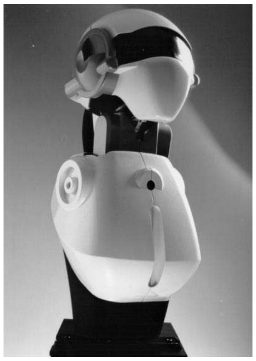

Figure 1: Humanoid *SIG*

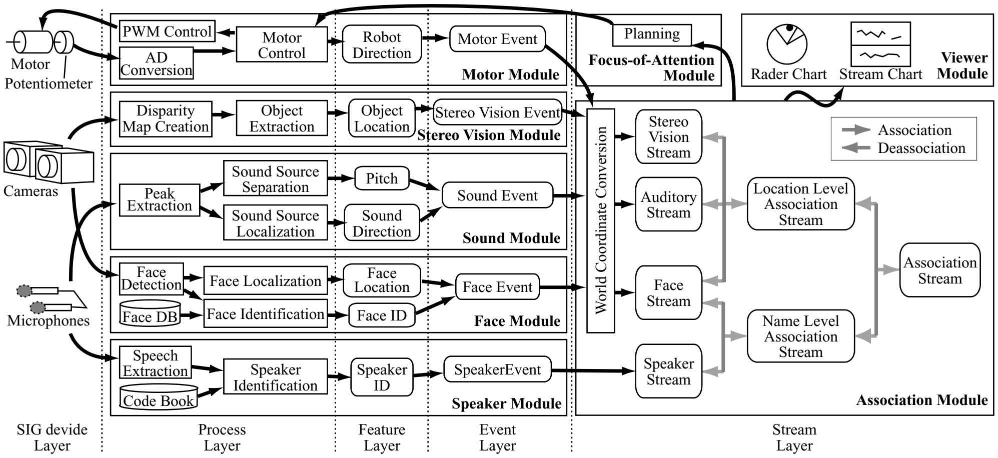

Figure 2: Hierarchical Architecture of Real-Time Tracking System

etry and motor noise cancelation in motion by cover. In addition, auditory, visual and motor processing are integrated to resolve ambiguity which every sensor information has innately. As a result, we achieve robust scene analysis by human tracking even when the person is visually occluded and simultaneous speeches occur. However, the system tracks a person only when he faces the robot face or speaks, since sound and face modalities are simply used to estimate human location. In this paper, we add new modalities of speaker ID and accurate object location by stereo vision to the system. This improves accuracy of human detection because of hierarchical integration – higher level (ID) and lower level (location) integration. Then, the system can estimate location and name of a person even when face detection failed because he looks away.

The paper is organized as follows: Section 2 presents our humanoid *SIG*. Section 3 describes a new real-time human tracking system improved by speaker identification and stereo vision. Section 4 shows evaluation of the system. The last section gives discussion and conclusion.

2 The Humanoid *SIG*

As a testbed of real-time multiple-object tracking, we use an upper-torso humanoid called *SIG*[4] shown in Fig. 1. The cover of the mechanics is made of FRP (Fiberglass Reinforced Plastic) and discriminates the internal and the external world acoustically. *SIG* has two microphones at the left and right ear positions to capture external sounds from outside of the body, and other two microphones within the body to capture internal sounds mainly caused by motor movements. All the microphones are omni-directional microphones of Sony ECM-77S. *SIG*'s body has four DOFs (degree of freedom), each of which is a DC motor controlled by a potentiometer. *SIG* is equipped with a pair of CCD cameras of Sony EVI-G20, but the current vision module uses only one camera.

3 System Design

Fig. 2 depicts hierarchical modules and data flow of the human tracking system.

From the viewpoint of functionality, the whole system can be decomposed into five layers — *SIG Device Layer*, *Process Layer*, *Feature Layer*, *Event Layer* and *Stream Layer*. The *SIG Device Layer* includes sensor equipment such as cameras, microphones and motor system. They send images from cameras and acoustic signals from microphones to the *Process Layer*. In the *Process Layer*, various features are extracted from raw data such as images and signals to send to the *Feature Layer*. Features are transformed to events with observed time for communication, then they are sent from the *Event Layer* to the *Stream Layer*. In the *Stream Layer*, event coordinates are converted into world coordinates. They are connected by taking their time series into account to make a stream. When two streams are close enough to be regarded as originating from a single source, they are associated into an association stream. Three different types of association — location association, name association and association between location and name — can happen according to the characteristics of streams. The status of stream and association influence *SIG* attention and active motion.

From the viewpoint of implementation, the system consists of eight modules, Sound, Face, Speaker, Stereo Vision, Association, Focus-of-Attention, Motor and Viewer. Each module sends and receives data with various abstraction levels over Gigabit Ether network from other modules asynchronously.

3.1 Sound Module

Sound allows a mixture of sounds originating from different directions as input. It is sampled with a sampling frequency of 48 KHz and 16-bit quantization, and its spectrogram is calculated by FFT. Then pitches (fundamental frequency, $F0$) are extracted, and sound sources are separated and localized.

Peak Extraction and Sound Source Separation: First a peak is extracted by a band-pass filter, which lets a frequency between 90 Hz and 3 KHz pass if its power is a local maximum and more than the threshold. This threshold is automatically determined by the stable auditory conditions of the room. Then, extracted peaks are clustered according to *harmonicity*. A frequency of Fn is grouped as an overtone (integer multiple) of $F0$ if the relation $|\frac{Fn}{F0} - \lfloor\frac{Fn}{F0}\rfloor| \leq 0.06$ holds. The constant, 0.06, is determined by trial and error. By applying an Inverse FFT to a set of peaks in harmonicity, a harmonic sound is separated from a mixture of sounds.

Sound Source Localization: Sound localization for a robot or an embedded system is usually solved by using interaural phase difference (IPD) and interaural intensity difference (IID). These values are calculated by using a Head-Related Transfer Function (HRTF). However, the HRTF depends on the shape of head and it also changes as environments change. For real-world applications, sound localization without HRTF is preferable. We proposed a method based on the auditory epipolar geometry, an extension of epipolar geometry in stereo vision to audition [4]. However, we failed in doing the job in real-time, because they stuck to pure-tone processing. Then we improved the robustness of the job to process in real-time 1. by exploiting the harmonic structure to extract peaks precisely, 2. by solving the uncertainty in sound source localization by Dempster-Shafer theory, and 3. by introducing *active audition* for sensorimotor task with canceling motor and mechanical noises.

From the extracted harmonic structure of left and right channels, a pair of harmonic structures is obtained. Then the IPD, P_s, is calculated. Auditory Epipolar Geometry generates a hypothesis of IPD P_h for each 5° candidate, θ [3]. Since the IPD is ambiguous for frequencies of more than 1.5 KHz, the distance, $d(\theta)$, in IPD between the data and a hypothesis is defined by

$$d(\theta) = \frac{1}{n_{f<1.5\mathrm{KHz}}} \sum_{k=0}^{n_{f<1.5KHz}-1} \frac{\left(P_h(\theta, f_k) - P_s(f_k)\right)^2}{f_k} \tag{1}$$

where $n_{f<1.5KHz}$ is the number of overtones of which frequency is less than 1.5KHz.

From the distance obtained by Eq. (1). The belief factor of IPD, B_{IPD}, is calculated by using probability density function defined by

$$B_{\mathrm{IPD}}(\theta) \quad = \quad \int_{-\infty}^{\frac{d(\theta)-m}{\sqrt{\frac{s}{n}}}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \tag{2}$$

where m and s are the average and variance of $d(\theta)$, respectively. n is the number of d.

A similar relation may hold for IID, but our experience with IID proves that it can discriminate at most the side, that is, center, left or right. Suppose that $I_s(f)$ is the IID for peak frequency f. If the value of $I = \sum_{f=1.5\,\mathrm{KHz}}^{3\,\mathrm{KHz}} I_s(f)$ is non-negative, the direction is decided as left, otherwise as right. According to the value of I, the belief factor of IID, $B_{\mathrm{IID}}(\theta)$ is defined by Table 1. Then, belief factors of B_{IPD} and B_{IID}, are integrated using Dempster-Shafer theory defined by

$$\begin{aligned} B_{\mathrm{IPD+IID}}(\theta) &= B_{\mathrm{IPD}}(\theta)B_{\mathrm{IID}}(\theta) + \\ &\left(1 - B_{\mathrm{IPD}}(\theta)\right)B_{\mathrm{IID}}(\theta) + B_{\mathrm{IPD}}(\theta)\left(1 - B_{\mathrm{IID}}(\theta)\right) \end{aligned} \tag{3}$$

where θ for the maximum $B_{\mathrm{IPD+IID}}$ is treated as the sound source direction of the harmonics.

Finally, Sound sends an auditory event consisting of pitch ($F0$) and a list of 20-best directions (θ) with reliability factor and observation time for each harmonics.

Table 1: Belief Factor of IID, $B_{\mathrm{IDD}}(\theta)$

	θ	90° ∼ 35°	30° ∼ −30°	−35° ∼ −90°
I	+	0.35	0.5	0.65
	-	0.65	0.5	0.35

3.2 Face Module

Face detects and recognizes multiple faces, and send face events. To implement on a robot and apply to the real world, this module employs fast and robust processing for frequent changes in the size, direction and brightness of a face in real-time.

Face Detection and localization:

The face detection submodule detects face robustly by combining skin-color extraction, correlation based matching, and multiple scale image generation [5].

The face localization submodule converts a face position in the 2-D image plane into 3-D world coordinates. Suppose that a face is $w \times w$ pixels located in (x, y) in the image plane, whose width and height are X and Y, respectively (see screen shots shown in Fig. 5). Then the face position in world coordinates is obtained in terms of distance r, azimuth θ and elevation ϕ by

$$r = \frac{C_1}{w}, \; \theta = \sin^{-1}\left(\frac{x - \frac{X}{2}}{C_2\ r}\right), \; \phi = \sin^{-1}\left(\frac{\frac{Y}{2} - y}{C_2\ r}\right) \tag{4}$$

where C_1 and C_2 are constants defined by the size of the image plane and the image angle of the camera.

Face Identification: The face identification submodule projects each extracted face into the discrimination space, and calculates its distance d to each registered face. Since this distance depends on the degree (L, the number of registered faces) of discrimination space, it is converted to a parameter-independent probability P_v as follows.

$$P_v = \Gamma\left(\frac{1}{2}, \frac{d^2}{2}\right) = \int_{\frac{d^2}{2}}^{\infty} e^{-t}\, t^{\frac{L}{2}-1} dt \qquad (5)$$

Linear Discriminant Analysis (LDA) can create an optimal subspace to distinguish classes. Therefore, we use Online LDA [6]. In addition, this method continuously updates a subspace on demand with a small amount of computation.

Finally, Vision sends a visual event consisting of a list of 5-best Face ID (Name) with reliabilities, observation time and position (distance r, azimuth θ and elevation ϕ) for each face.

3.3 Speaker Identification Module

As an engine of speaker identification, Speaker uses "Juno" which is a software developed by Akita *et al.* of Kyoto University. It can identify a single speech by two methods, that is, vector quantization (VQ) and Gaussian mixture model (GMM). Speaker use GMM for identification because it is robust against environmental noises. It can be seen as a single state Hidden Malcov Model (HMM).

Let S be the number of speakers. In GMM, a speaker model is represented as a mixture of M Gaussians. When $G(i,j)$ is a jth Gaussian of ith speaker, and acoustic observation v divided into T frames is given by

$$v = \{v(k) | k = 1, 2, \cdots, T\}, \qquad (6)$$

$L(i)$, a likelihood of ith speaker, is defined by

$$L(i) = \prod_{k=1}^{T} \sum_{j=1}^{M} (W(i,j) \cdot N(i,j,k)) \qquad (7)$$

where $N(i,j,k)$ is a component density of $G(i,j)$ corresponding to a kth frame $v(k)$, and $W(i,j)$ is a weight value for $G(i,j)$.

Then Speaker assumes that a speaker with the maximum $L(i)$ is talking. Finally, Speaker sends an speaker event a list of 5-best speakers with reliability factor and observation time.

3.4 Stereo Vision Module

Stereo Vision extracts lengthwise objects such as persons from a disparity map to localize them. First a disparity map is generated by an intensity based area-correlation technique. This is processed in real-time on a PC by a recursive correlation technique and optimization peculiar to Intel architecture [7].

In addition, left and right images are calibrated by affine transformation in advance. An object is extracted from a 2-D disparity map by assuming that a human body is lengthwise. A 2-D disparity map is defined by

$$DM_{2D} = \{D(i,j) | i = 1, 2, \cdots W, j = 1, 2, \cdots H\} \quad (8)$$

where W and H are width and height, respectively and D is a disparity value.

As a first step to extract lengthwise objects, the median of DM_{2D} along the direction of height shown as Eq. (9) is extracted.

$$D_l(i) = Median(D(i,j)). \qquad (9)$$

A 1-D disparity map DM_{1D} as a sequence of $D_l(i)$ is created.

$$DM_{1D} = \{D_l(i) | i = 1, 2, \cdots W\} \qquad (10)$$

Next, a lengthwise object such as a human body is extracted by segmentation of a region with similar disparity in DM_{1D}. This achieves robust body extraction so that only the torso can be extracted when the human extends his arm. Then, for object localization, epipolar geometry is applied to the center of gravity of the extracted region. Finally, Stereo Vision creates stereo vision events which consist of distance, azimuth and observation time.

3.5 Association Module

Association forms a stream by connecting events to a time course. In addition, it associates streams to create a higher level stream, which is called an *association stream* (see Fig. 2). Three kinds of association can create three types of association streams. The flow of processing in stream formation and association is summarized in Figs. 4(a)-(d).

Table 2: Event Generation Cycle and Latency

Face Event	150 ms
Sound Event	30 ms
Speaker Event	200 ms
Stereo Vision Event	50 ms
Motor Event	100 ms
Network Latency	10 – 200 ms

World Coordinate Conversion: Every event except a speaker event includes location information. But, since location information is observed in a *SIG* coordinate system, event coordinates should be converted into world coordinates for integration shown in Fig. 3. For conversion, the system compares an event with a motor event observed at the same time. However, each type of events has a different delay time and generation cycle (see Table 2). So the system stores events in two second short-term memory to synchronize them. If a corresponding motor event is unavailable, synchronization interpolates a motor event by linear regression. Then, events in world coordinates are obtained with 200 ms delay, since it is longer than the maximum delay shown in Fig. 4(a).

Stream Formation: The converted events are connected to a stream with some error corrections described in Fig. 4(b). The algorithm of stream formation is as follows:

- **Sound Event:** A sound event is connected to a sound stream when it satisfies two conditions: 1. they have harmonic relationship, and 2. their direction difference is within ±10°. The value of ±10° is defined by the accuracy of auditory epipolar geometry.
- **Face Event:** A face event is connected to a face stream when they have the same face ID and their distance is within 40 cm. The value of 40 cm is defined by assuming that human motion speed is less than 4 m/sec.
- **Speaker Event:** At most, a single speaker event is created at the same time, because Speaker does not support simultaneous speeches. Therefore, a speaker event is always connected to a speaker stream if any.
- **Stereo Vision Event:** A stereo vision event is connected to a stereo vision stream when their distance is within 40 cm.

After checking, an event which did not find any stream to be connected to is transformed to a new stream. A stream is terminated if there is no event to be connected for more than 500 ms. The advantages of creating streams by taking time series into account are as follows:

- Sound : Acquisition failure of a fundamental tone can be corrected.
- Face : Error of face identification can be corrected because face IDs are checked over a whole stream.
- Speaker : In a speaker stream, an later connected event has higher belief of identification. So, even if speaker ID is wrong on creation of a stream. The wrong ID can be corrected with progress of time.

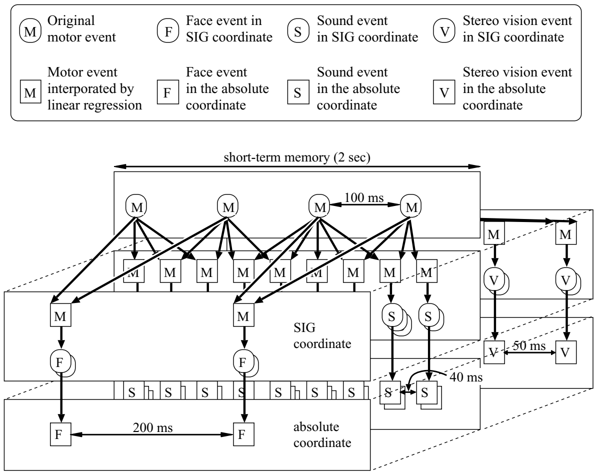
M Original motor event
F Face event in SIG coordinate
S Sound event in SIG coordinate
V Stereo vision event in SIG coordinate
M Motor event interporated by linear regression
F Face event in the absolute coordinate
S Sound event in the absolute coordinate
V Stereo vision event in the absolute coordinate
short-term memory (2 sec)
100 ms
SIG coordinate
absolute coordinate
200 ms
50 ms
40 ms

Figure 3: World Coordinate Conversion

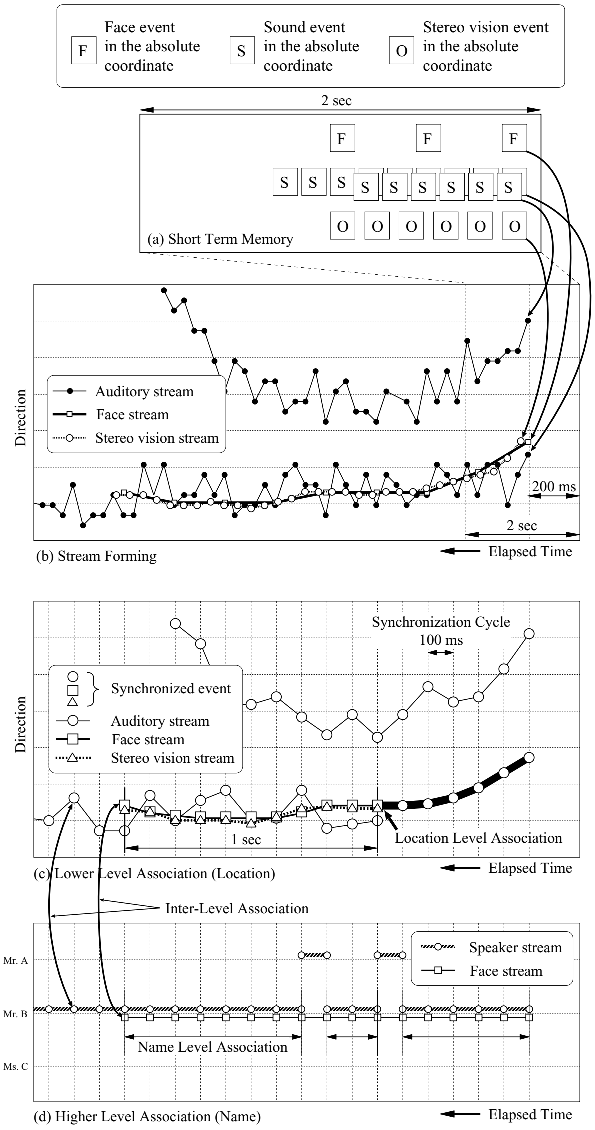

Figure 4: Stream Formation in Association Module

- Stereo Vision : Movement of objects can be grasped (effective in an association).

Association: When the system judges that multiple streams originate from the same person, they are associated into an association stream. When one of the streams forming an association stream is terminated, the terminated stream is removed from the association stream. If a condition for association is not satisfied by this removal, the association stream is deassociated to some separated streams.

A stream is classified into lower and/or higher level streams according to the abstraction level of information included in it to make hierarchical stream structure. Higher level streams include speaker and face streams because they have a name ID. Lower level streams include sound, stereo vision and face streams because of their location information. A face stream belongs to both level streams because it has both information.

In association, first, streams are synchronized every 100 msec to calculate the distance between streams (see Fig 4(c) and (d)). Then, stream association is checked in lower level streams and in higher level association. Association between lower and higher level streams is also checked. Thus, three kinds of association is checked on the hierarchical stream structure.

Lower Level Association: Location information is common in lower level streams. When the distance between streams keeps close for more than a constant time, they are regarded as streams originating from the same object and integrated into an location level associated stream shown in Fig 4(c). Location information consists of three parameters – distance(r), azimuth(θ) and elevation(ϕ). However, the number of parameters in location information depends on the stream type as shown in Table 3. Accuracy and available range of location information also depends on the stream type because of different sensors and extraction methods. For example, azimuth is available for all lower level streams. But the stream with the most accurate direction is stereo vision, the second is face, and the last is sound. On the other hand, the available range of sensors is getting wider in the reverse order. Then, the definition of distance between streams depends on the stream types to be calculated as follows:

Sound – Face : Sound and face streams are associated if their azimuth difference is within the range of ±10° and this situation continues for more than 50% of a 1 sec period.

Face – Stereo Vision : Face and stereo vision streams are associated if their distance is within the range of 10 cm and this situation continues for more than 50% of a 1 sec period.

Stereo Vision – Sound : Stereo vision and sound streams are associated if their azimuth difference is within the range of ±10° and this situation continues for more than 50% of a 1 sec period.

Thus, distance functions should be customized because of using location information with different accuracy. But the most accurate information is not always used in the system because less accurate information can compensate for missing information to improve robustness of the system. For example, stereo vision and face modules cannot localize occluded or out-of-sight objects. Sound can localize such objects if they make sounds. Face may misextract a face in a picture on the wall, and it cannot extract a face of a person looking away. To correct such errors, the system can use speaker ID and sound direction. Although a sound stream has only azimuth information, lower level association can give it distance and elevation information.

Table 3: Location Information in Streams

	Sound	Face	Streo Vision
r		✔	✔
θ	✔	✔	✔
ϕ		✔	

Higher Level Association: Speaker and face streams belong to higher level streams including name information. Higher level association shown in Fig. 4(d) can happen when speaker and face IDs are the same. The advantage of higher level association is that the system keeps a stream continuously even when speaker or face ID is missing.

Inter-Level Association: Association between higher and lower level association streams can occur, which is represented as arrows between Figs. 4(c) and (d). Associations of face ID and face location, and of speaker ID and sound direction are allowed for this case. But the former is done in creation of face stream because a face event includes location and name information in advance. The latter is based on two rules as follows:

1. When a single sound stream exists, a speaker stream is associated with it.
2. When multiple sound streams exist, a speaker stream is associated with a sound stream which has the nearest start time.

In Association, the following rules are used to avoid contradiction in association.

- Streams which are not associated yet can be associated when their stream types are different.

- Streams which are associated can be associated when they do not have any common stream type.

3.6 Focus-of-Attention Control

Focus-of-Attention selects *SIG* action according to the status of streams and sends motor events to Motor. We call it audio-visual servo, because this focus-of-attention is controlled by both auditory and visual information. Robust and complex focus-of-attention control can be done in comparison with either visual or audio servo. The principle of focus-of-attention control hereby is as follows:

1. An auditory stream has the highest priority,
2. an associated stream has the second priority, and
3. a visual stream has the third priority.

4 Experiment and Evaluation

A scenario shown in Fig. 5 is used as a benchmark. In the scenario, two speakers, Mr. A and B, express behaviour for about 20 seconds, and *SIG* selects action according to their behaviour. Their behaviour in t_n in Fig. 5 is as follows:

t_1: Mr. A starts talking out of *SIG*'s visual field. *SIG* turns to the direction of the sound stream on him.

t_2: A speaker stream on Mr. A is created, and speaker and sound streams on him are associated.

t_3: Mr. A starts moving in front of *SIG*. When *SIG* catches sight of him, a face and a stereo vision stream are created.

t_4: They are associated.

t_5: All streams on Mr. A are associated. *SIG* has much interest in the association stream, and it goes on tracking him.

t_6: During tracking, Mr. B starts talking out of *SIG*'s visual field. A sound and a speaker stream are created and associated. *SIG* turns to the direction of a new sound source to obtain further information on it.

t_7: Because *SIG* cannot see Mr. A by turning to Mr. B, the association stream on Mr. A is deassociated.

t_8: Mr. B is in sight, and a face and a stereo vision stream are created. He pauses to speak for a while.

t_9: A face and a stereo vision stream are associated.

t_{10}: All streams on Mr. B are associated. *SIG* goes on tracking him.

The performance of integrated auditory and visual tracking shown in Fig. 5 proves that robust tracking is achieved by strong association on all types of streams from t_5 to t_7, and after t_{10}. From t_6 to t_8, accurate stream separation is shown under simultaneous speeches. When a sound stream out of the visual field

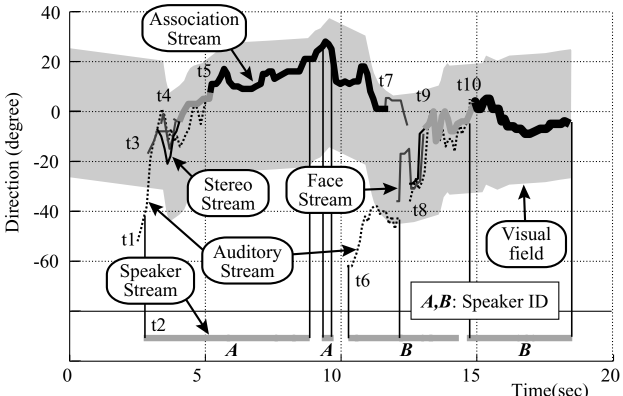

Figure 5: Tracking of Two Speakers

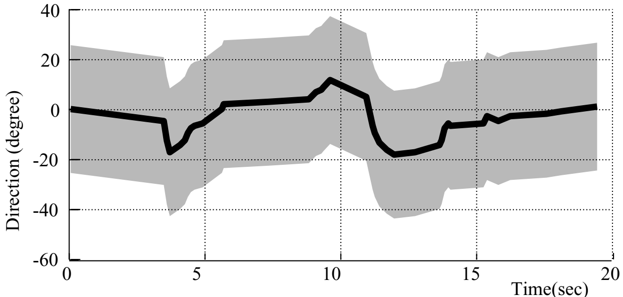

Figure 6: Motor Direction and Visual Field in Fig. 5

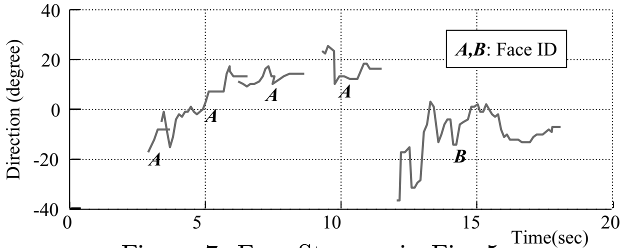

Figure 7: Face Streams in Fig. 5

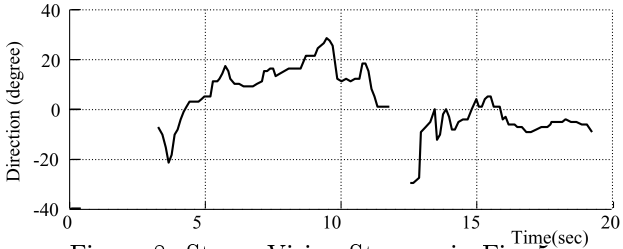

Figure 8: Stereo Vision Streams in Fig. 5

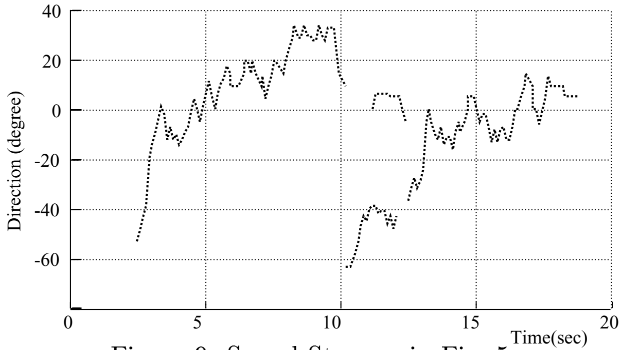

Figure 9: Sound Streams in Fig. 5

is created, *SIG*'s attention is changed to get further information on the new stream under the condition of multiple streams (t_6).

Fig. 6 shows the front direction and visual field of *SIG*. Focus-of-Attention realizes smooth human tracking in spite of fine vibration of stream direction information. When an unseen speaker exists, *SIG* accurately tracks him by auditory information.

Fig. 7 shows face streams. A face stream has an advantage of both location and name information. But face extraction often fails because of bad lighting and a person looking away. This divides a stream into some fragments in $t = 3.5s$, $6s$ and $9s$. In spite of the fragmentation of a stream, *SIG* can go on tracking due to the association shown in Fig. 5.

Fig. 8 shows stereo vision streams. A stereo vision stream has accurate location information. But because stereo cameras are required, the visual field is narrower than a single camera for face streams. Fig. 5 proves that association can compensate for such a narrow visual field.

Fig. 9 shows sound streams. Although microphones for a sound stream are omni-directional, sound streams from $t = 5s$ to $t = 10s$ in Fig. 9 indicate that estimation of sound direction is not very accurate. Association can compensate for the ambiguity of the sound direction by vision information.

We already showed that robust tracking can be done under missing visual information by occlusion [3]. In Fig. 5, the second half of a speaker stream starting at t_6 should be associated with a sound stream starting at t_8. This error was caused by missing stream formation or association. A mechanism of stream reformation and re-association by feedback would be required to resolve it.

5 Conclusion

In this paper, we described the method for robust real-time human tracking by integration of sound direction, speaker ID, face location, face ID, object location by stereo vision. New modalities of speaker ID and object location by stereo vision are added. Stereo vision not only gives the system accurate location information, but also compensates for missing face information when a person looks away. Speaker information gives speaker ID for a stream out of sight unless *SIG* turns to the sound source. When stereo vision cannot localize objects because of occlusion, sound direction can compensate for such missing location information. Errors of speaker ID can be corrected by face ID if *SIG* turns to it. Thus, this proves that it makes a system more robust to add modalities even if each modality has some ambiguities. Robust processing of dynamic changes of the environment and multiple speaker identification are the object of further research.

Acknowledgements

We thank Mr. Yuya Akita and Professor Tatsuya Kawahara of Kyoto University for their discussions and offering Juno.

References

[1] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," in *Proceedints of the Sixteenth International Joint Conference on Atificial Intelligence (IJCAI-99)*, 1999, pp. 1146–1151.

[2] Y. Matsusaka, T. Tojo, S. Kuota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi, "Multi-person conversation via multi-modal interface — a robot who communicates with multi-user," in *Proceedings of 6th European Conference on Speech Communication Technology (EUROSPEECH-99)*. 1999, pp. 1723–1726, ESCA.

[3] K. Nakadai, K. Hidai, H.G. Okuno, and H. Kitano, "Real-time auditory and visual multiple-object tracking for humanoids," in *Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI-2001)*. AAAI.

[4] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proceedings of 17th National Conference on Artificial Intelligence (AAAI-2000)*. 2000, pp. 832–839, AAAI.

[5] K. Hidai, H. Mizoguchi, K. Hiraoka, M. Tanaka, T. Shigehara, and T. Mishima, "Robust face detection against brightness fluctuation and size variation," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2000)*. 2000, pp. 1397–1384, IEEE.

[6] K. Hiraoka, S. Yoshizawa, K. Hidai, M. Hamahira, H. Mizoguchi, and T. Mishima, "Convergence analysis of online linear discriminant analysis," in *Proceedings of IEEE/INNS/ENNS International Joint Conference on Neural Networks*. 2000, pp. III–387–391, IEEE.

[7] Okada K. Inaba M. Inoue H. Kagami, S., "Real-time 3d optical flow generation system," in *Proc. of International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI'99)*, 1999, pp. 237–242.