An Autonomous Three-Dimensional Vision Sensor with Ears

Shigeru Ando
Faculty of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113 Japan
Phone: 03-3812-2111 ex.6936
Fax: 03-3816-7805
Email: ando@alab.t.u-tokyo.ac.jp

Abstract — **This paper describes our newly developed intelligent sensor system which comprises two eyes and four ears on a movable head. It can acquire its dynamical visual and auditory image of its surrounding 3-D environment while showing humanlike behavior naturally and autonomously. The most important feature of the sensor system is in a unified sensory architecture throughout low-level and intermediate level visual and auditory functions. This enables us to achieve 1) rapid (5ms) and accurate (±2deg) auditory localization, 2) rapid (0.5s/65536pixel) extraction of motion and texture features, 3) rapid (0.1s/4096pixel) reconstruction of 3-D object profile, 4) rapid (several TV frame times) eye movement and binocular fixation which is activated by auditory localization and motion detection. We describe in this paper the several key items for realizing this sensor.**

1 INTRODUCTION

We are developing a robot head sensor "SmartHead" as shown in Fig.1, in which a binocular vision sensor and a quad-aural auditory sensor are mounted on an autonomously movable head. Both the vision and auditory sensors are ready to detect unusual events, i.e., visual motion/accretion and auditory localization of sound sources, in its surrounding environment as an early warning system. If such an event is found, then head motors and a high-speed saccadic eye movement system are activated and they quickly move the field of view (FOV) of the vision sensor toward a direction of the object. A tracking and fixation system catches the most salient point in that direction and fixes the FOV on the object. Then a high-speed differential stereo vision sensor with microvibrative eye movements is activated, and it reconstructs a solid profile and extracts various image features of the object.

This kind of machine vision systems have been extensively studied in the field of computer vision and image processing. But the most serious problems for applications of the intelligent vision sensor is how to find a target object to pay the biggest attention in the numerous surrounding ones. An autonomous realtime sensing system in an unrestricted environment must decide itself what to concentrate its finite sensing power on. For this purpose, the visual information alone is insufficient because of its limited processing speed and narrow attentive area.

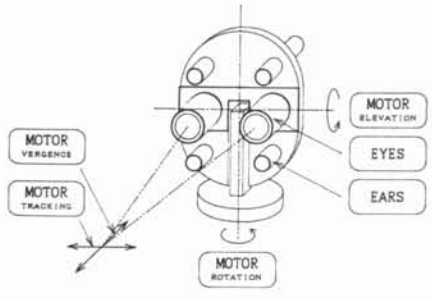

Figure 1. Robot head sensor "SmartHead" with a binocular vision sensor and a quad-aural auditory sensor on an autonomously movable head.

One of the most distinctive feature of our 3-D vision sensor is that it is integrated with an auditory sensor and several motor functions. From it, it gains wider area of attention and automatism. And the second most distinctive feature of our sensor is that both the visual and auditory sensors, throughout low-level feature extraction and intermediate-level integration of them, are based on our unified sensing principle. This enables the superior sensitivity, selectivity and recognition performance of the sensor with the minimal complexity of hardware and software.

2 GRADIENT CORRELATION SENSING PRINCIPLE

The unified sensing principle for our sensor "SmartHead" is shown schematically in Fig.2(a). In the first stage, the difference and/or differential of the incoming signals are constructed on the sensing probes themselves or in the early stage of the circuits. Like any other differential measurement systems, the selectivity for the difference (skew-symmetric) component is emphasized by the symmetric structure. The latter stage of this architecture produces sequences of cross correlation between the output of the gradient stage. The output of this stage forms a temporal and/or spatial sequence of correlation matrix.

The information carried by the differential correlation quantities in the vision section are summarized in Table 1. In these tables, the differential quantities are indicated in the marginals of the rows and columns, and the items in the tables are the cross-correlations between the quantities in the corresponding row and column. The upper-right, off-diagonal items indicate qualitative (discriminative) in-

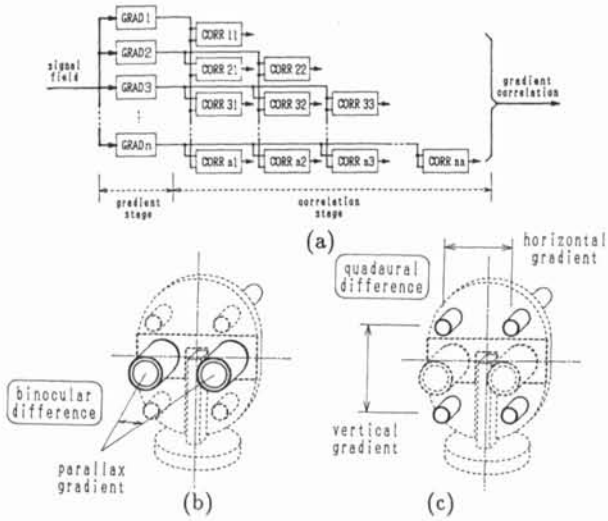

Figure 2. A schematic diagram of the gradient correlation based sensory architecture (a), and two gradient sensing probes (b),(c) used in the vision sensor and auditory sensor. The difference and/or differential of the incoming signals are constructed by the sensing element itself or in the early stage of the circuits. Then short time and/or small-area correlations among them are generated so that they form a correlation matrix.

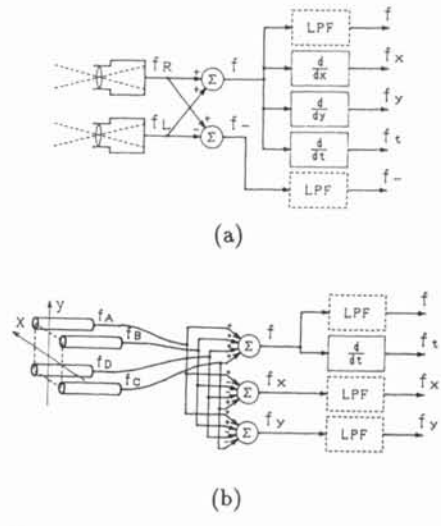

Figure 3. Gradient stage of the visual (a) and auditory (b) gradient correlation architecture. In (a), the left and right cameras followed by a subtraction circuit constitute a parallax differentiator, and in (b), the quad-aural microphones followed by subtraction circuits constitute a sound field differentiator.

formation and the corresponding quantitative information obtained when we find strong correlations between those differential quantities. The lower-left, off-diagonal items indicate the qualitative (discriminative) information obtained when the correlations vanish.

From this table, we know that these quantities are almost universal for low-level image feature extraction. Because description by the correlation matrix is complete, those informative events, even if they occur simultaneously, can easily be perceived while decomposing the co-occurrences and mixtures.

	x gradient f_x	y gradient f_y	time gradient f_t	binocular dif f_d	binocular sum f
f_x	S_{xx}	S_{xy} edge or line (orientation)	S_{xt} motion (x velocity)	S_{xd} binocular fusion (depth)	S_x x slope (x increase rate)
f_y	S_{xy} blob or corner	S_{yy}	S_{yt} motion (y velocity)	S_{yd} binocular fusion (y offset)	S_y y slope (y increase rate)
f_t	S_{xt} x accretion	S_{yt} y accretion	S_{tt}	S_{td} binocular luster (dynamic IID)	S_t t slope (t increase rate)
f_d	S_{xd} depth edge	S_{yd} ?	S_{td} ?	S_{dd}	S_d binocular luster (static IID)
f	S_x x extremum	S_y y extremum	S_t t extremum	S_d ?	S

(IID: Inter-ocular Intensity Disparity)

Table 1 Contents and meanings of the gradient correlation features for binocular vision sensor.

3 SENSING ALGORITHMS

In this section, we describe the visual and auditory sensing algorithms which are based solely on the gradient correlation quantities. The following subsections are placed in order of growing concentration of attention of our sensor. We use the following notation for the correlation quantities:

$$\int_{-\Delta L}^{\Delta L}\int_{-\Delta L}^{\Delta L}\int_{-\Delta T}^{0} f_i f_j \mathrm{d}x\mathrm{d}y\mathrm{d}t \equiv < f_i, f_j > \equiv S_{ij} \quad (1)$$

for the vision section, and

$$\int_{-\infty}^{0} f_i f_j e^{t/\Delta T}\mathrm{d}t \equiv < f_i, f_j > \equiv S_{ij} \quad (2)$$

for the auditory section. The suffix i or j indicates any one of the suffixes $x, y, t, -$ or none of the gradient quantities. For simplicity, the origin of x, y, t coordinates is assumed to be at a measuring point and time. $2\Delta L$ means the spatial extent of the correlation window (typical values are from 2 to 5 pixels in length), and ΔT is a time duration of the correlation window (from 1 to 2 frames of time for vision section and 5 ms for the auditory section).

3.1 *Invoking Visual Attention by Auditory Localization*

In the auditory sensor, a set of four closely located microphones measure the spatiotemporal gradients f, f_t, f_x, f_y of the sound pressure field. Let a Cartesian coordinate with its origin at a measuring point be x, y, z. Indicating the clockwise microphones outputs as f_A, f_B, f_C, f_D, we obtain the gradients as

$$f \simeq \frac{1}{4}(f_A + f_B + f_C + f_D) \quad (3)$$

$$f_x \simeq \frac{1}{2W}(f_A - f_B - f_C + f_D) \quad (4)$$

$$f_y \simeq \frac{1}{2W}(f_A + f_B - f_C - f_D) \quad (5)$$

$$f_t = \frac{\partial}{\partial t} f \simeq \frac{1}{4}\frac{\partial (f_A + f_B + f_C + f_D)}{\partial t}, \quad (6)$$

where W is the separation of the microphones. The approximation is valid if W is much smaller than the wavelength. In our system, the signals are low-pass-filtered below 1.4 kHz (wavelength~24.3 cm) similarly to the human auditory system so that $W = 10.6$ cm is valid.

In the front half-space $z > 0$, there are multiple uncorrelated sound sources S_i $(i = 1, 2, \cdots)$ of which positions are (x_i, y_i, z_i) (unknown). Let the sound velocity be C, the distance from a sound source S_i to the sensor at an origin be R_i (unknown), source sound be $g^i(t)$ (unknown), and sound pressure generated by each source be $f^i(t)$ (unknown). Then, the total pressure f (observable) is expressed by a sum of spherical waves emitted from the sources as

$$f = \sum_i f^i(t) = \sum_i \frac{1}{R_i} g^i(t - \frac{R_i}{C}). \quad (7)$$

Therefore, x, y and t gradients (observable) of the sound field at a measuring point are written as

$$f_x = -\sum_i \{\xi_x^i f^i(t) + \tau_x^i f_t^i(t)\} \quad (8)$$

$$f_y = -\sum_i \{\xi_y^i f^i(t) + \tau_y^i f_t^i(t)\} \quad (9)$$

$$f_t = \sum_i f_t^i(t) \quad (10)$$

by using x and y directional phase gradients τ_x^i and τ_y^i, and amplitude gradients ξ_x^i and ξ_y^i of the wavefront as

$$\xi_x^i = \frac{x_i}{R_i^2}, \quad \xi_y^i = \frac{y_i}{R_i^2} \quad (11)$$

$$\tau_x^i = \frac{x_i}{CR_i}, \quad \tau_y^i = \frac{y_i}{CR_i}, \quad (12)$$

where $f_t^i(t)$ is a time derivative of $f^i(t)$.

We rewrite this equation as

$$\begin{bmatrix} f_x \\ f_y \\ f_t \\ f \end{bmatrix} = \begin{bmatrix} \xi_x^1 & \tau_x^1 & \xi_x^2 & \tau_x^2 & \cdots \\ \xi_y^1 & \tau_y^1 & \xi_y^2 & \tau_y^2 & \cdots \\ 0 & 1 & 0 & 1 & \cdots \\ 1 & 0 & 1 & 0 & \cdots \end{bmatrix} \begin{bmatrix} f^1 \\ f_t^1 \\ f^2 \\ f_t^2 \\ \vdots \end{bmatrix} \quad (13)$$

so that the source position variables and source waveform variables are separated. Hereafter, the rank of distribution of $(f_x, f_y, f_t, f)^T$ is determined by the number of sources and spatial arrangement of the sources. In particular, if the number of sources is one, the rank is always 2. If it is two, the rank is usually 3 for distant sources and 4 for very near sources[16]. This means that, from degeneration of the distribution $(f_x, f_y, f_t, f)^T$ in a short interval, we can obtain instantaneous sound source arrangement information.

More concretely, we obtain the correlation matrix

$$\begin{bmatrix} S_{xx} & S_{xy} & S_{xt} & S_x \\ S_{xy} & S_{yy} & S_{yt} & S_y \\ S_{xt} & S_{yt} & S_{tt} & S_t \\ S_x & S_y & S_t & S \end{bmatrix}. \quad (14)$$

Usually, S_{tt} is small and the rank of this matrix is full because of circuit noises or orderless environmental sounds. If S_{tt} becomes large and the matrix shows some degeneration, we know some distinctive sounds are present. According to the rank and the corresponding eigenvectors, we can obtain the location information as follows: 1) the azimuth and distance of one sound source (rank=2), 2) the azimuth and distance of a curve which passes through two sound sources (rank=3).

3.2 Invoking Visual Attention by Motion Detection

As shown in Table 1, the existent or nonexistent correlations between f_t and f_x and f_y supplies us the motion information [6]. This is because the moving image field (optical flow) satisfies Euler's relation[5]

$$v_x f_x + v_y f_y + f_t = 0 \quad (v_x, v_y : x, y \text{ velocity}), \quad (15)$$

which is equivalent to saying that the distribution of f_x, f_y, f_t shows perfect degeneration if motion exists there[7]. If degeneration does not occur although the image changes, we know that some events other than motion, e.g., dynamic occlusion/accretion or changing illumination, occur[12].

More concretely, we observe the correlation submatrix

$$\begin{bmatrix} S_{xx} & S_{xy} & S_{xt} \\ S_{xy} & S_{yy} & S_{yt} \\ S_{xt} & S_{yt} & S_{tt} \end{bmatrix} \quad (16)$$

for the left and right added image anywhere in FOV. Usually, the energy of brightness change S_{tt} is small because the sensor and its surrounding environment are stationary. If some strong/large values appear, we obtain their position information (direction) from the image coordinates, and, according to the rank and the corresponding eigenvectors, we know whether 1) there is a moving object with a velocity v_x, v_y (rank=1), or 2) brightness changes possibly because an object appears or disappears suddenly (rank$\geq$ 2).

3.3 FOV Movement and Binocular Correspondence on the Most Salient Point

The visual attention process starts with the activation of the FOV movement system (two vergence motors and two head motors) so that the center of FOV is directed toward the object. In this stage, however, stable fixation and binocular fusion are not established. To achieve them, the extraction of feature points begins simultaneously from both right and left images.

Again, let the brightness be f (the following argument on f applies to both a right brightness f_R and a left brightness f_L), and its spatiotemporal gradients be f_x, f_y and f_t. Correlation values among these quantities are elements of the correlation matrix shown in Table 1, as follows:

$$\begin{bmatrix} S & S_t & S_x & S_y \\ S_t & S_{tt} & S_{xt} & S_{yt} \\ S_x & S_{xt} & S_{xx} & S_{xy} \\ S_y & S_{yt} & S_{xy} & S_{yy} \end{bmatrix} \tag{17}$$

As shown in Table 1, the existent or nonexistent correlations among f and f_x and f_y supplies the information on local image patterns. This is because the linear dependence (degeneration)

$$af_x + bf_y + f = 0 \quad (a^{-1}, b^{-1} : x, y \text{ increasing rate}) \tag{18}$$

implies that the image pattern there is exponential, and the dependence

$$af_x + bf_y = 0 \quad (a, b : \text{orientation of pattern}) \tag{19}$$

indicates that it varies only in one direction. Both indicate that a brightness slope or edge exists there. If no correlation is found even though the image is varying, we know there is a brightness extremum and can classify it according to its homogeneity and orientation[13]. The degree of correlation is determined by the canonical correlation coefficient[14] as

$$Q_2 \equiv \frac{SS_{xx}S_{yy} + 2S_xS_yS_{xy} - SS_{xy}^2 - S_{xx}S_y^2 - S_{yy}S_x^2}{S(S_{xx}S_{yy} - S_{xy}^2)} \quad \text{(heterogeneous peak)}. \tag{20}$$

See [11, 13] for the detailed theory and complete set of operators.

Using the categorization and ranking given by these operators, we extract the three most important feature points as follows: 1) edge feature and its strength and orientation. 2) line-element feature and its sign (dark or bright), strength and orientation, and 3) blob (point) feature and its sign, strength and orientation.

The correspondence is established using the above-mentioned pattern 1) of a right and left image as a primary candidate. Evaluating and ranking the quality of match between attributes and the horizontal disparity, an optimum correspondence is determined. If there are no good correspondences using 1), the search is repeated using 2) and then 3) as secondary candidates.

If correspondence is established, the FOV movement system is again activated to catch the most salient object point at both FOV centers of the binocular camera. At this stage, we obtain precise position information (direction and distance) of the object by means of the stereo ranging principle.

3.4 Perceiving Shape from Binocular Stereo

Suppose we use a binocular camera with an interocular separation D (known) and a distance H ($H \gg D$) (known from the stereo ranging) to its base plane. Let an image observed at the center of the cameras be $\tilde{f}(x, y)$ (unknown). By using $\tilde{f}(x, y)$, left and right images $f_L(x, y)$ and $f_R(x, y)$ (observable) are described as a small symmetric intensity change $1 \pm \xi$ and a small symmetric positional shift $\pm\Delta$ with respect to $\tilde{f}(x, y)$ as

$$f_R(x, y) \simeq (1 + \xi)\tilde{f}(x + \Delta, y) \tag{21}$$

$$f_L(x, y) \simeq (1 - \xi)\tilde{f}(x - \Delta, y), \tag{22}$$

respectively where Δ is expressed as

$$\Delta \equiv \frac{Dh(x, y)}{2(H + h(x, y))} \simeq \frac{D}{2H}h(x, y) \tag{23}$$

using the relative height $h(x, y)$ with respect to the base plane. The base plane (known) coincides with a plane parallel to an image plane and intersecting the object point at the center of FOV. Let addition and subtraction of left and right images be $f(x, y)$ and $f_-(x, y)$, respectively. Applying first-order approximation of Δ and ξ to these two images $f(x, y)$ and $f_-(x, y)$, and adding and subtracting them so that the unknown $\tilde{f}(x, y)$ is eliminated, we obtain the relation

$$f_-(x, y) \simeq \xi f(x, y) + \Delta f_x(x, y), \tag{24}$$

where $f_x(x, y)$ is an x-directional derivative of the addition image $f(x, y)$.

Solving eq.(24) by the least squares method, we obtain a position disparity Δ and an intensity disparity ξ as

$$\xi = \frac{S_{xx}S_{+-} - S_{x+}S_{x-}}{S_{xx}S_{++} - S_{x+}^2} \tag{25}$$

$$\Delta = \frac{S_{++}S_{x-} - S_{x+}S_{+-}}{S_{xx}S_{++} - S_{x+}^2}, \tag{26}$$

where each correlation value is obtained from a submatrix

$$\begin{bmatrix} S & S_x & S_- \\ S_x & S_{xx} & S_{x-} \\ S_- & S_{x-} & S_{--} \end{bmatrix} \tag{27}$$

of the correlation matrix in Table 1. From Δ, we know the relative height distribution $h(x, y)$ (shape) from eq.(23). From ξ we can detect small interocular intensity disparity owing to the viewing angle difference on the surface[4], and relate it to the luster feeling (glossness of surface)[8].

Another important quantity obtained from the matrix (27) is a measure for the goodness of binocular fusion[9]. It is defined as

$$J_{\mathrm{ERR}} \equiv \frac{(S_{xx} + S_{++})(S_{--} - \xi S_{+-} - \Delta S_{x-})}{W^2(S_{xx}S_{++} - S_{x+}^2)} \tag{28}$$

which is a sum of expected error correlations of Δ and ξ. By thresholding J_{ERR} as

$$J_{\mathrm{ERR}} < J_{\mathrm{MIN}} \tag{29}$$

using an appropriate threshold J_{MIN} (e.g., (0.5pixel length)^2), we identify

where the binocular fusion is established. These decisions are transfered to the depth accumulation system so that a complete object shape is reconstructed by connecting the partial depth slices within each fusion zone during the microvibrative eye movements.

4 GRADIENT CORRELATION HARDWARE

Although the gradient correlation principle enables us to use a unified and systematic architecture, it requires tremendous computational power to obtain all sensory information listed in Tables 1 and 2 simultaneously, unless it is supported by specialized hardware. Therefore we developed a digital circuit for computing the gradient correlation quantities for the vision section shown in Table 1. The circuit is hard-wired (not DSP-based) according to the architecture shown in Fig.2. For the gradient correlation hardware for the auditory section, we used an analog circuit developed previously[15]. Fig.4 shows a photograph of this circuit (two slot-wide VME board). Although the newly developed circuit is very simple, it can generate whole elements of the gradient correlation matrix within a four pixel time of a TV camera. All succeeding processes are performed by conventional UNIX workstations (Solbourne 5/600 with 1 CPU, and a desktop SUN 4).

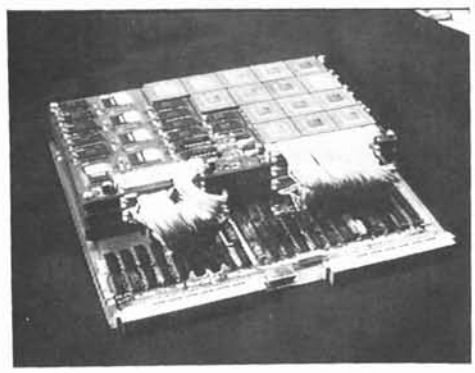

Figure 4. A photograph of the gradient correlation hardware for the vision section of "SmartHead". Input of this hardware is binocular TVC signals (8bit/pixel each), and the output is the 5 x 5 gradient correlation matrix (32 bit/element each within four pixel time).

5 AUTONOMOUS SENSING ALGORITHM

One of the most important objectives of "SmartHead" sensor is acquiring a world image, i.e., a dynamical 3-D map of the surrounding environment, while performing humanlike actions naturally and autonomously. We realized such actions through the coordination of several processes. The tasks and algorithms of each process installed in "SmartHead" are as follows.

Binocular Correspondence Task — Establish the most reliable correspondence and trigger saccadic eye movement.

Tracking and Fixation Task — Compute binocular disparity and motion in the FOV center and pursue the object continuously.

3-D Feature Extraction Task — Fuse binocular images and extract 3-D shape and image features.

Object Segmentation Task — Identify an object figure from a background by using the shape and features.

Visual Early Warning Task — Examine marginal views and find salient or moving objects as candidates of a new FOV position.

Auditory Early Warning Task — Localize sound sources and find salient ones as candidates of a new FOV position.

Fig.5 shows a block diagram of the FOV control system realized by the binocular correspondence process and the tracking/fixation process. Fig.6 shows a photograph of when the sensor localized on a sound source and then caught by eyes a calling woman. Fig.7 shows examples of outputs of this sensor. The primary output is not a CRT monitor image such as is shown but arrays of digital data in which the shape and features in the processing region are stored.

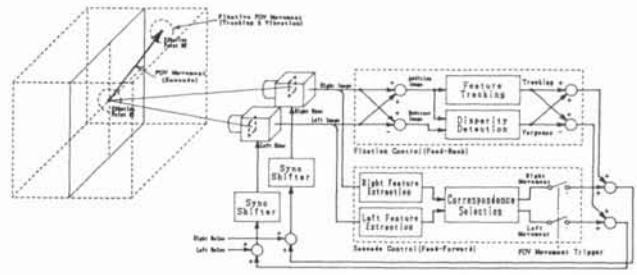

Figure 5. A simplified block diagram of a saccadic and smooth pursuit FOV control system of "SmartHead". The smooth-pursuit eye motion is performed by dual feedback control: one for azimuthal motion using an image coordinate and 2-D velocity of a featured point, and one for range motion using binocular disparity.

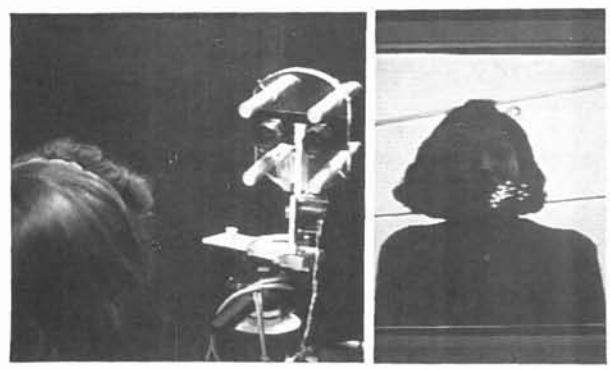

Figure 6. An addition image of both eyes immediately after completing the FOV movement and fixation. The vision of "SmartHead" correctly catches the mouth of the calling woman.

Figure 7. An example of the real-time shape reconstruction (a) and the image feature extraction (b). The central square of the monitor shows the processing region (64 x 64 measuring points in 128 x 128 area). In the region, a relative height map (a) with respect to the base plane and a strength map (b) of "ridgeness" feature is expressed by the degree of brightness.

6 PERFORMANCE

The performance of this sensor is as follows.

1) A 3-D sound localization sensor with ±2deg (in anechoic room) and ±5deg (in usual room) accuracy, real-time processing speed (less than 1ms delay time), and high temporal resolution (5ms observation window time).

2) A high speed saccadic eye movement system which responds quickly to motion and salient features over an entire FOV (several TV frame times from detection to matching).

3) A tracking and fixation system which pursues both the range motion and azimuth motion (feedback delay is less than 1 TV field time).

4) A high speed differential stereo vision sensor with microvibrative eye movements (from 0.1sec to 0.5sec for reconstructing solid on 64 x 64 measuring points).

5) A visual and auditory early warning system using realtime motion detection (0.5sec for entire image search) and 3-D sound localization (1ms/localization).

REFERENCES

[1] Y.Yakimovsky and R.Cunningham: A System for Extracting Three-Dimensional Measurements from a Stereo Pair of TV Cameras, Computer Graphics and Image Processing, vol.7, pp.195-210 (1978)

[2] E.Krotkov, K.Henriksen, and R.Kories: Stereo Ranging with Verging Cameras, IEEE Trans. Pattern Anal. Machine Intell., vol.PAMI-12, no.12, pp.1200-1205 (1990)

[3] S.D.Cochran and G.Medioni: 3-D Surface Description from Binocular Stereo, IEEE Trans Pattern Anal. Machine Intell., vol.PAMI-14, no.10, pp.981-994 (1992)

[4] B.D.Lucas and T.Kanade: An iterative image registration techniques with an application to stereo vision, Proc. 1981 Int. Joint Conf. Artificial Intell., pp.674-679 (1979)

[5] B.K.P.Horn and B.G.Schunck: Determining optical flow, Artificial Intelligence, vol.17, pp.185-203 (1981)

[6] C.Cafforio and F.Rocca: Methods for measuring small displacement of television images, IEEE Trans. Information Theory, vol.**IT-22**, no.5, pp.573-579 (1976)

[7] S.Ando: A Velocity Vector Field Measurement System Based on Spatio-Temporal Image Derivatives, Trans. Soc. Instrumentation and Control Engineers, vol.**22**, no.12, pp.1330-1336 (1986) (in Japanese)

[8] S.Ando: Detection of Intensity Disparity in Differential Stereo Vision Systems with an Application to Binocular Luster Perception, Trans. Soc. Instrumentation and Control Engineers, vol.**23**, no.6, pp.619-624 (1987) (in Japanese)

[9] S.Ando and T.Tabei: Differential Stereo Vision System with Dynamical 3-D Reconstruction Scheme, Trans. Soc. Instrumentation and Control Engineers, vol.**24**, no.6, pp.628-634 (1988) (in Japanese)

[10] S.Ando: Image feature extraction operators based on curvatures of correlation function, Trans. Soc. Instrumentation and Control Engineers, vol.**24**, no.10, pp.1016-1022 (1988) (in Japanese)

[11] S.Ando: Texton finders based on Gaussian curvature of correlation function, Proc.1988 IEEE Int.Conf.Syst.Man Cybern., Beijing/Shenyang, pp.25-28 (1988)

[12] S.Ando: Gradient-Based Feature Extraction Operators for the Classification of Dynamical Images, Trans. Soc. Instrumentation and Control Engineers, vol.**25**, no.4, pp.496-503 (1989) (in Japanese)

[13] S.Ando: Image Grayness Feature Extraction Based on Extended Gradient Correlations, — Theoretical Analysis and Numerical Simulation —, Trans. Soc. Instrumentation and Control Engineers, vol.**27**, no.9, pp.982-989 (1991) (in Japanese)

[14] P.Chen and S.Ando: A Generalized Theory of Waveform Feature Extraction Operators for High-Resolution Detection of Multiple Reflection of Sound, Annual Report of Engineering Research Institute of University of Tokyo, vol.**50**, pp.167-174 (1991) (in Japanese)

[15] S.Ando: An Analog Electronic Binaural Localization Sensor, Technical Digest of 8th Sensor Symposium, Tokyo, pp.131-134 (1989)

[16] S.Ando, H.Shinoda, K.Ogawa, and S.Mitsuyama: A Three-Dimensional Sound Localization System Based on the Spatio-Temporal Gradient Method, Trans. Soc. Instrumentation and Control Engineers, vol.**29**, no.5, pp.520-528 (1993) (in Japanese)