PAPER

Man-Machine Interaction Using A Vision System with Dual ViewingAngles

Ying-Jieh HUANG†, ***Member,*** **Hiroshi DOHI**††, ***Nonmember and*** **Mitsuru ISHIZUKA**††, ***Member***

SUMMARY This paper describes a vision system with dual viewing angles, i.e., wide and narrow viewing angles, and a scheme of user-friendly speech dialogue environment based on the vision system. The wide viewing angle provides a wide viewing field for wide range motion tracking, and the narrow viewing angle is capable of following a target in wide viewing field to take the image of the target with sufficient resolution. For a fast and robust motion tracking, modified motion energy (*MME*) and existence energy (*EE*) are defined to detect the motion of the target and extract the motion region at the same time. Instead of using a physical device such as a foot switch commonly used in speech dialogue systems, the begin/end of an utterance is detected from the movement of user's mouth in our system. Without recognizing the movement of lips directly, the shape variation of the region between lips is tracked for more stable recognition of the span of a dialogue. The tracking speed is about 10 frames/sec when no recognition is performed and about 5 frames/sec when both tracking and recognition are performed without using any special hardware.

key words: vision system, dual viewing angles, speech dialogue system, motion tracking, mouth pattern recognition

1 Introduction

During the last thirty years, a major research goal in computer system field has been to make computers intelligent, to work with us, and to be our helpers. An average of 48% of the code in today's application is devoted to the user interface portion according to the results of a survey on human computer interface programming [1].

Despite so much effort, however, computers today still remain difficult to use in common human life. Users have to sit in front of an output device or wear some troublesome device like a goggle, typing on a keyboard, moving a mouse, clicking buttons to express his/her intention. The limitations of interface between the user and the computer restrict the integration of computing power into various human tasks and various daily life styles.

Computer vision makes it possible for a user to use any convenient objects as input signal. These objects include orientation of head [2][3], gaze direction of eyes [4][5][6], finger tips [7], hand gestures [8], mouth movement [9] [10] and even facial expression [11]. The use of computer vision is a key component to realize more free and friendly human interfaces.

Since human eye is one of the most developed visual system and is well studied, many vision systems are modeled on the base of it. The vision systems developed so far are summarized in Table 1 according to the number of cameras used. Only the abilities are listed in Table 1, no matter how well they done. More detailed information about vision systems can be found, for example, in [12].

Table 1 The summary of vision system

	Tracking range	Resolution	Vergence	3D information acquisition	Comment
Monocular vision system	Wide	single	no	no	backgroun compensation is need
		uniform/variable			
Binocular vision system	Narrow	single	yes	yes	no gaze selection
		uniform			
Trinocular vision system	Wide	two	yes	yes	the use of third camera not yet reported
		unifrom			
Two cameras vision system	Wide	two	no	no	only one resolution is used
		uniform			
Dual viewing angles vision system	Wide	two	no	yes	
		unifrom/variable			

When computer vision is used for human computer interaction, the attentive visual search is one of the important factors. A complete human computer interaction should be started automatically when a user enters its viewing field and be ended when the user away from its viewing field. This means that the computer vision for human computer interaction should be able to aware of the existence of user automatically. The required image resolution for recognizing the action of the user is clearly not the same as the one for tracking the motion of the user. This implies that using only one resolution in human action recognition is insufficient.

In this paper, we describe a vision system with dual viewing angles for human computer interaction. The motion tracking and feature recognition of a user in front of it will be done under different image resolution, and a spontaneous dialogue environment constructed with this vision system

Manuscript received May 16, 1996.
Manuscript revised May 14, 1997.

† The author is with the Information and Communication R&D Center of Ricoh Co. Ltd., Yokohama-shi, 222 Japan.
†† The authors are with the Dept. of Information and Communication Engineering, the University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan.

will also be showed.

2 Vision System with Dual Viewing Angles

2.1 Human action recognition with dual viewing fields

The goal of vision system here is to achieve a feature recognition on a movable target with a reasonable movement. For example, it recognizes the action of a person who is allowed to move unrestrained in front of vision system. The problems must to be solved in this application are wide range motion tracking and stable feature recognition. For wide range motion tracking, the camera with a wide viewing angle is preferable. A stable feature recognition could not be expected with this camera, since only low resolution image is available in the area of a target object. Another method to achieve wide range motion tracking is to use a movable camera. This method suffers from the necessary complicated compensation for changing background images [13] and/or the frequently occurred mechanical adjusting of zoom and focus. These make the realization of real-time motion tracking difficult.

For recognition algorithm to work robustly in recognizing the action of human, including gesture, facial expression, gaze direction, mouth shape and so forth from an image, a proper resolution is required. This means that those features must be confined within a narrow viewing field. This is why in many computer vision systems, the input image taken under a controlled environment is required and the target is forced to keep at a proper pose.

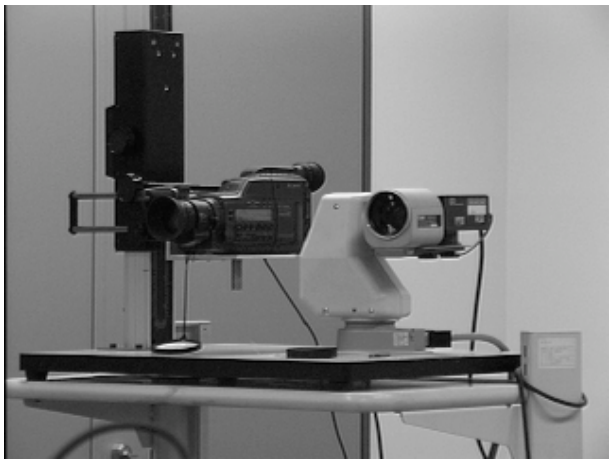


Fig. 1 The vision system with dual viewing angles

2.2 System configuration

We propose a new configuration of using two cameras, as shown in Fig. 1, to meet the both needs of wide-range motion tracking and high-resolution image acquisition. The wide viewing field is provided by the fixed camera with a wide viewing angle. The other camera with a narrow viewing angle is mounted on a rotatable platform which is capable of rotating about two axes, pan and tilt. With this configuration, the vision system provides a foveal-peripheral vision acuity analogues to that of human vision system.

Unlike the human vision system, the optical axis of the foveal region and the peripheral region of the vision system here are not at the same direction. The optical axis of the camera with wide viewing angle is fixed, and the optical axes of the camera with narrow viewing angle is able to point at any place.

This configuration allows the vision system to track a moving target within wide viewing field fast and to get an image of the target with a sufficient resolution for recognition at the same time. It is notable here that no complicated background compensation is needed and the inconsistency between resolution and field range can also be dissolved.

Besides the rotation of pan and tilt, the zoom and the focus of the camera mounted on platform can also be controlled via RS-232 interface. The hardware specifications of the vision system are shown in Table 2.

Table 2 The hardware specifications of the vision system

Mechanism		
	PAN	TILT
Range of rotation	±170 deg	±60 deg
Velocity of rotation	1～58 deg/s	1～51 deg/s
Resolution of rotation	0.094 deg/step	0.033 deg/step
Optics		
When object is 3 m in front of cameras	Wide viewing angle camera (fixed)	Narrow viewing angle camera (pointable)
Viewing field (Horizontal) (cm)	400	52.5
Resolution (mm/pixel)	25	3.28

2.3 Calibration between two cameras

(a) Model and notation

The camera model used in our work is based on the approximation of pinhole camera model, and weak perspective projection transform is used to map the coordinates of points in the 3D world space into 2D image coordinates as shown in Fig. 2. The image plane in Fig. 2 is perpendicular to the *Z*-axis, the optical axis of the camera, and intersects it at (*0,0,f*), where *f* is the effective focal length in pinhole camera model. The relationships between a point *P*(*X,Y,Z*) in 3D world coordinates and its image *p*(*x,y*) in 2D camera image frame can be expressed as follows:

$$x = f\frac{X}{Z}, y = f\frac{Y}{Z} \quad (1)$$

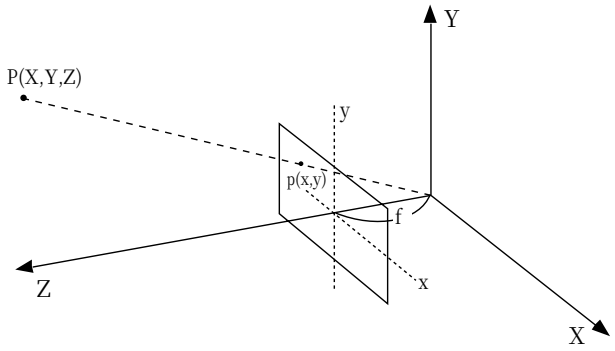


Fig. 2 Perspective projection in pinhole camera model

The coordinate value of a point P between different coordinate systems , (*U,V,W,1*) and (*X,Y,Z,1*), can be related by a 4×4 matrix **T** which describes the transformation of rotation and translation:

$$\begin{bmatrix} U & V & W & 1 \end{bmatrix} = \begin{bmatrix} X & Y & Z & 1 \end{bmatrix} \mathbf{T} \tag{2}$$

where

$$\mathbf{T} = \begin{bmatrix} r_1 & r_2 & r_3 & 0 \\ r_4 & r_5 & r_6 & 0 \\ r_7 & r_8 & r_9 & 0 \\ l & m & n & 1 \end{bmatrix} \tag{3}$$

In (3), $r_1,r_2,...,r_9$ are the rotation parameters and *l,m,n* are the translation parameters.

(b) Extrinsic parameters calibration

Since the amount of pan-tilt of the pointable camera with a narrow viewing angle is determined according to the position information of the target within the image of fixed camera with a wide viewing angle, the extrinsic parameters between the two cameras are also needed to be calibrated.

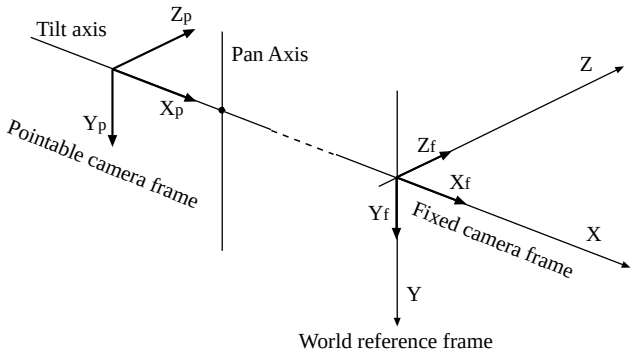


Fig. 3 Coordinate frames of the vision system with two cameras

To identify the position of a point, a fixed coordinate frame of reference, called world frame, is required. The origin of world frame is chosen at the lens center of fixed camera, and be oriented so that the *Z*-axis coincides with the optical axis of fixed camera and parallel to the optical axis of pointable camera which is at its home position. The rotating center of the pointable camera is set on the *X*-axis of world frame with a separation of l from the fixed camera. The coordinates system described above is shown in Fig. 3.

For a point *Q* in the world reference frame (*X,Y,Z*) can be related to the frame of the pointable camera (X_P,Y_P,Z_P) by a series of transformation **T**:

$$\mathbf{T} = \begin{bmatrix} \cos\theta & \sin\theta\sin\phi & \sin\theta\cos\phi & 0 \\ 0 & \cos\phi & -\sin\phi & 0 \\ -\sin\theta & \cos\theta\sin\phi & \cos\theta\cos\phi & 0 \\ l\cos\theta + r & l\sin\theta\sin\phi & l\sin\theta\cos\phi & 1 \end{bmatrix} \tag{4}$$

where θ and ϕ are angles of pan and tilt respectively, *l* is the distance between the origin of the fixed camera and the rotation center of the platform, *r* is the radius of rotation about *Y*-axis. From (4) we can get the relationship between (*X,Y,Z*) and (X_P,Y_P,Z_P) as follows:

$$\begin{aligned} X_p &= X\cos\theta - Z\sin\theta + l\cos\theta + r \\ Y_p &= X\sin\theta\sin\phi + Y\cos\phi + Z\cos\theta\sin\phi + l\sin\theta\sin\phi \\ Z_p &= X\sin\theta\cos\phi - Y\sin\phi + Z\cos\theta\cos\phi + l\sin\theta\cos\phi \end{aligned} \tag{5}$$

The goal of visual tracking is to maintain a fixation on a moving target and to keep the image of the visual target in the center of the viewing field of the pointable camera, i.e. (X_P,Y_P)=(0,0). For this purpose, since the kinematic equations can be derived from (5), we can obtain:

$$X = L\tan\theta - l - r\sec\theta \tag{6}$$

$$Y = -L_p\sin\varphi \tag{7}$$

$$L_p = L\sec\theta - r\tan\theta \tag{8}$$

where *L* is the length of the foot of the perpendicular from *Q* to the *X*-axis. L_p is the distance between the lens center of the pointable camera and *Q*. The visual kinematics described above is depicted in Fig. 4.

If the range of pan angle is small enough, the relationship between the amount of rotation of the pointable camera and the position of the target in viewing field of the fixed camera can be approximated by linear equations. We expand the equations (6) and (7) to first order linear equations by using Taylor's series as follows:

$$\theta = 57.3\left(\frac{x}{f_x} + \frac{l}{W_x}\frac{w_x}{f_x} + \frac{r}{W_x}\frac{w_x}{f_x}\right) \tag{9}$$

$$\phi = 57.3\left(\frac{W_y y}{0.01745 r w_y \theta - W_y f_y}\right) \tag{10}$$

where (x,y) are the projection of (X,Y) in a fixed camera image frame, f_x, f_y are the focal length in x, y, and w_x,w_y are the measured length in an image frame of the target of size W_x, W_y.

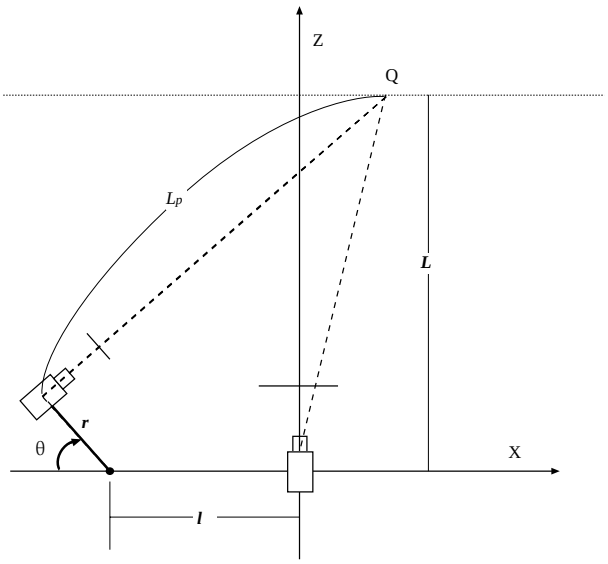


Fig. 4 The visual kinematics of vision system with two camera

3 Motion Tracking and Gaze Initialization in Wide Viewing Field

3.1 Modified motion energy and exist energy

The image size of the target, e.g. a person, in wide viewing field is so small that recognition-based motion tracking will not suitable for this work. On the other hand, for a motion tracking to be as general as possible, it should be able to follow a moving target whose identity is not known, i.e., not require an object recognition. Motion energy detection is one of the methods suited for this purpose. The motion energy detection is implemented through a spatiotemporal filter, and the simplest implementation of this filter is image subtraction since each image has a previous image in the image sequence subtracted from it. Then the motion region can be extracted by thresholding the output of the image subtraction. Since the threshold value is empirically tuned, this will cause the motion energy to be not robust enough in motion detection. To determine the threshold value dynamically, we define a modified motion energy (*MME*) based on the output of image subtraction as follows:

$$MME \triangleq \sqrt{\frac{\sum_k (x-m)^2}{k-1}} \tag{11}$$

where x , m and k are the pixel value, mean value and number of pixel in the subtraction image. The *MME* in (11) shows the variation of the pixel values in motion region compared with the region out of it. The threshold value for extracting the motion region can be determined from:

$$\begin{aligned} Th_{low} &= m - MME \\ Th_{high} &= m + MME \end{aligned} \tag{12}$$

where Th_{low}, Th_{high} are lower bound and higher bound of the threshold value, m is a mean value of the subtraction image as in (11).

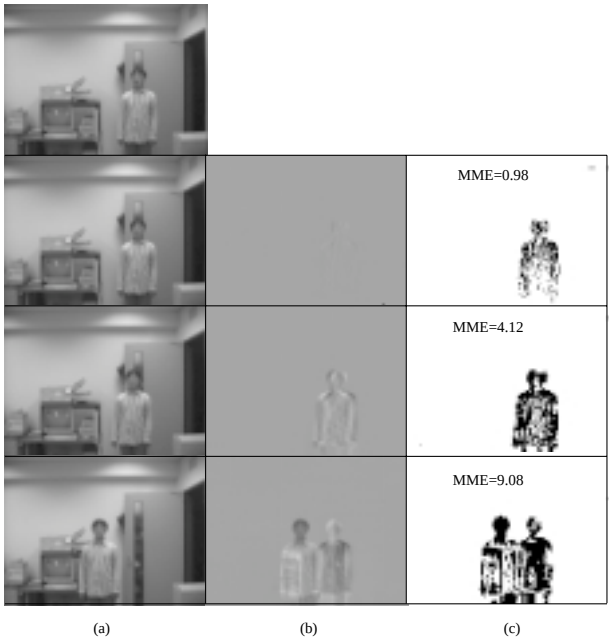


Fig. 5 Motion region detection and extraction with *MME*

In Fig. 5, (a) shows four continuous cuts of an image sequence, (b) shows the output of subtracted images between two successive images, and the extracted motion regions by using *MME* are shown in (c) together with the values of *MME* . It should be noted that the value of *MME* is in positive proportion with the amount of displacement. According to the values of *MME*, the movement of the object can be described qualitatively such as not moved, move slowly, move fast. The variation of *MME* when a person is in the wide viewing field (of gray zone) is shown in Fig. 6. It should be noted that the *MME* of a moving object, especially a people, will not keep staying at very low value stably for a few frames since it is impossible to make a

completely static. The situation that the *MME* keep staying at a very low value is either the moving object is out of the wide viewing field or the background is changed, for example, a people leaves his baggage and then go out. The timing for updating background image will be described later.

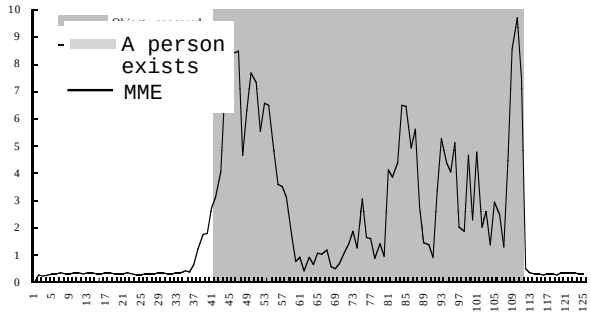


Fig. 6 The *MME* variation when a person exists in the wide viewing field

When *MME* is calculated from a subtraction image between the image with an target in it and the background image, the *MME* can be used to detect the appearance of an object. We define the *MME* as an existence energy (*EE*) when the image subtraction is carried with a background image. As shown in Fig. 7, the *EE* changes extremely when an object appeared and disappeared.

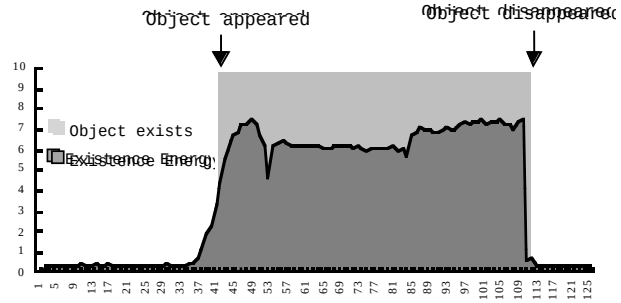


Fig. 7 Object appearance detection with *EE*

As can be seen, the values of existence energy (*EE*) are always kept large enough to distinguish the differences between appearance and disappearance of an target. If a target exists, the values of *EE* can also be used to determine the threshold value for extracting it from background.

Fig. 8 shows the segmentation results of using existence energy *EE*. The first image in Fig. 8(a) is the background image, images subtracted from background image are shown in (b), and segmentation results are shown in (c).

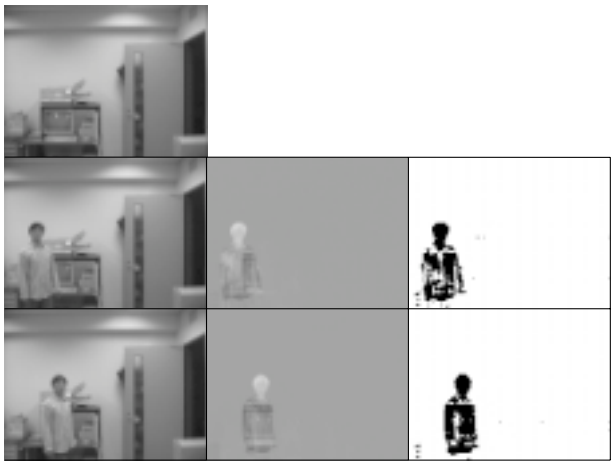


Fig. 8 Segmentation result with *EE*

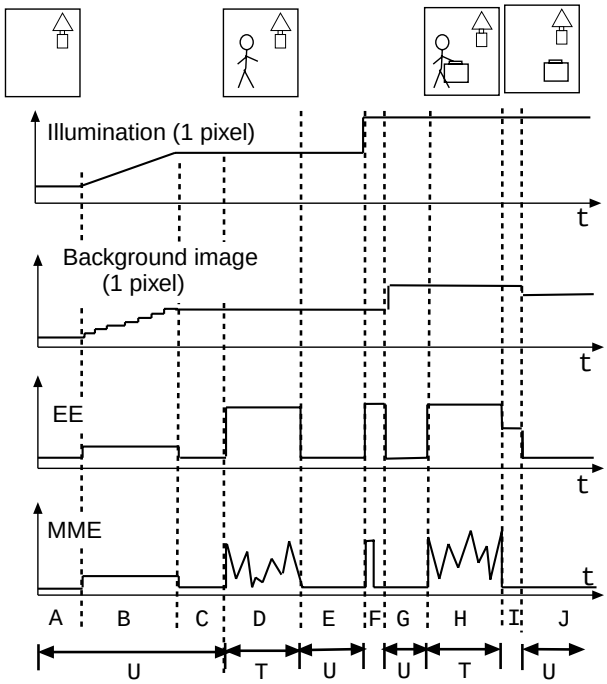


Fig. 9 Background image updating timing.

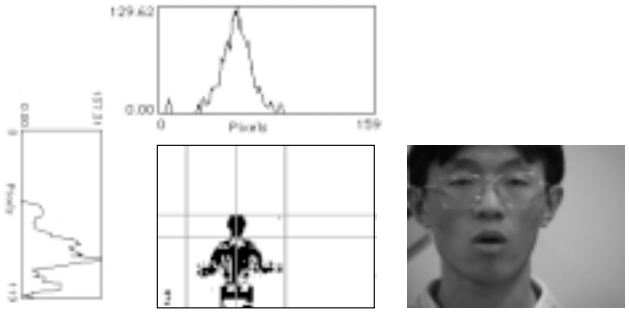
It is obvious that to keep a brand-new background image is important for motion region extraction. The background image is updated only when there is no moving object in wide viewing field, i.e., when *EE* is small enough. If the value of *MME* in current frame is small enough, the background image will be replaced by the current frame from time to time as shown in Fig. 9 the duration of A, C, E, G and J. Under an environment of slow variation of illumination as in duration B of Fig. 9, the value of *MME* is still small enough, and the background image can be

updated. However when a rapid illumination change has occurred as in duration F of Fig. 9, both values of *MME* and *EE* are increased like in the duration of D, E. Anyway, after the value of illumination is stable, the *MME* will back to stable low value, then the background image can be updated to make *EE* work properly. Beside the variation of illumination, another situation need to force background image be updated is real background changed. For example, in Fig. 9 duration H, a people walks through the wide viewing field, and left a baggage within the viewing field. Although, the value of *EE* will not be small enough, the background image can be updated when *MME* remains low for few frames. Under the environment of our office, the motion region can be extracted correctly from day to night, thus this method is robust against the daily illumination variation.

3.2 Gaze point initialization

After the position of the target in wide viewing field is detected, the pointable camera will be rotated so that the image of the target will be centered at the viewing field of the pointable camera. Since the viewing angle of the pointable camera is narrow, the gaze point must be further determined when the size of the target is too large. For example, to gaze at the head or the hand of a person, or even to gaze at his/her eye or mouth, the gaze selection can be done in the wide viewing field and/or in the narrow viewing field depending on the models used in motion tracking and in feature recognition.

In many human-interface applications, it is necessary to fetch the image of head or face first. The shape of a human's body in wide viewing field is clear as shown in Fig. 8. The center of the head in *X*-axis can be detected by finding the maximum projection in *X*-axis, and the neck can also be detected under an assumption that the neck is the slenderest along the center axis of the body as shown in Fig. 10(a). The image taken from pointable camera according to the result of Fig. 10(a) is shown in Fig. 10(b).



(a) (b)

Fig. 10 Extraction of the head from the segmentation result in wide viewing field

.

3.3 Tracking strategies

A complete vision system must be able to aware of the appearance and disappearance of a target. In our vision system, this awareness is achieved by calculating the modified motion energy (*MME*) and the existence energy (EE) in wide viewing field. To reduce the affects from illumination variation to as small as possible, the background image must be refreshed frequently since it will be used to segment the object. The background image is updated whenever the existence energy (*EE*) is small enough. When an object exists in field view of wide viewing angle camera, the following four conditions must be considered:

1) Has the object moved ?

The first one is to decide whether the object is moved (with a large displacement) or not. This can be made according to the amount of modified motion energy (*MME*) between two successive images.

2) How large does it move (if moved)?

If the modified motion energy (*MME*) exceeds a predetermined value, i.e., if the object moves, the required rotation angles (pan and tilt) for pointable camera to track it must be decided. The amount of rotation for pan and tilt can be calculated by using the inverse kinematic equations of this vision system as described in Chapter 3 according to the position of the object.

3) Where is it?

The position of the target can be located by calculating its *EE* in current frame and then be extracted.

4) Does the object still exist?

The last condition should be considered is the disappearance of a target. When the target is disappeared, i.e., away from the viewing field of the wide viewing angle camera, the background must be updated. The disappearance of the object can be detected by comparing the modified motion energy (*MME*) and existence energy (*EE*) of the object.

The Fig. 11 shows some tracking results of a people who moving at about 3 meters in front of the vision system. The people moving in front of it will be tracked according to images caught by the camera with wide viewing angle, and then images with higher resolution of his head are caught by the pointable camera. In Fig. 11, the left most column shows the 5th, 10th, ... frames of a series of images taken by the camera with wide viewing angle at about 12 frames/sec, and the extracted motion region from each frame are shown in the middle column, and the right most column shows images caught by the pointable camera. It takes less than 5ms to send out the command of pan and tilt from each frame in wide viewing field if the movement of the object is detected.

When the people walking at a speed below 28 cm/sec, the center of the image of his head will be kept within the viewing field of pointable camera. Note that through the selection of a wider viewing angle of the pointable camera, the people is allowed to walk at a higher speed and the image of his head is also kept within the viewing field of pointable camera. Since the position of the people is tracked by the fixed camera, no matter how fast the people moves, as long as he stop moves within the wide viewing field of the vision system, the pointable camera is still able to point at his head although after few frames.

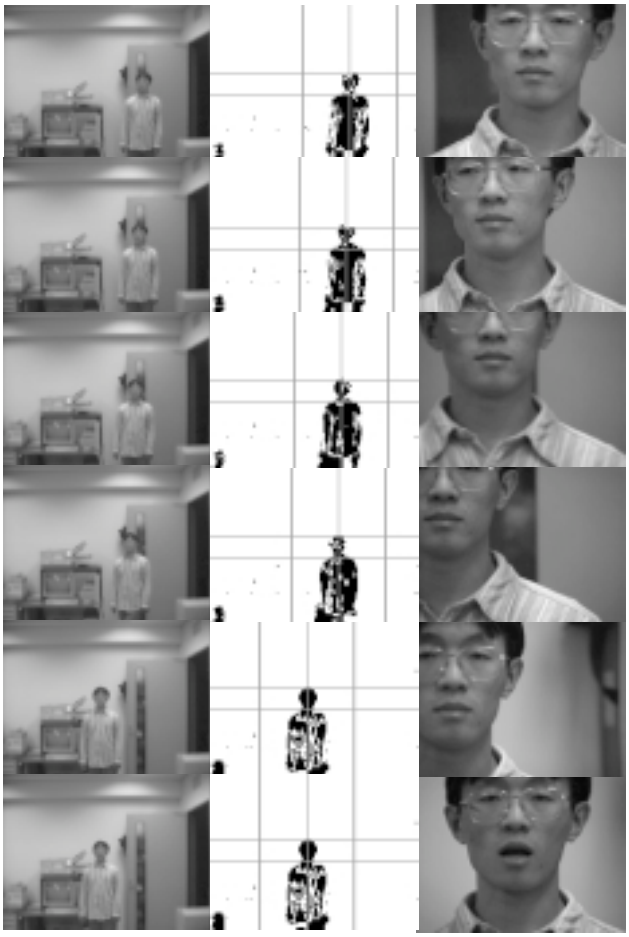


Fig. 11 Tracking result of vision system with dual viewing fields

4 Application: Spontaneous Speech Dialogue System

4.1 Unconstrained speech dialogue environment

After the high resolution image of the user's head is fetched by the pointable camera, many applications can be implemented on it. In this chapter, we describe a user-friendly speech dialogue system using the vision system with dual viewing angles. Instead of using a physical device such as a foot switch which is required in many continuous speech recognizers, the begin/end of an utterance are detected by recognizing the movement of the user's mouth.

In a common dialogue environment, the user must move to a predetermined position, as shown in Fig. 12(a), to use a physical switch to inform the system: I am here now, I will start my utterance now, my utterance is end. These make the dialogue system difficult to be integrated into the normal human life. To free the user from these restrictions, a steerable phase-array microphone is used in [14]. In which, several microphones are used, and an available computation power is required.

We will show that an unconstrained dialogue environment can be realized with the vision system with dual viewing angles. In Fig. 12(b), the user is allowed to move around as long as within the wide viewing field. The pointable camera is then to tracks the user's head and gets his mouth image with a sufficient resolution for recognition.

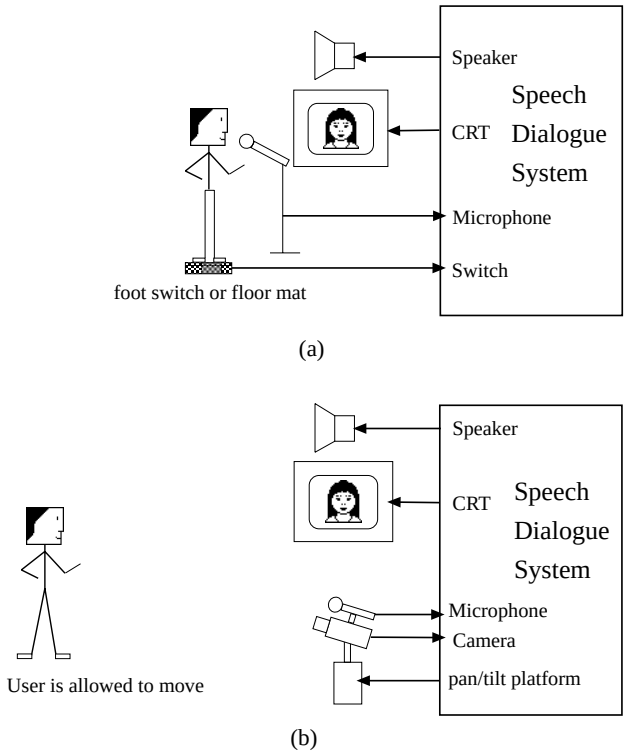


Fig. 12 Speech dialogue system from constrained (a) to unconstrained (b).

Fig. 13 shows the configuration of our spontaneous speech dialogue system with the dual viewing angles vision system. The transform circuit in Fig. 13 sends an on/off signal to the acoustic speech recognition system, which is installed on another workstation, according to the recognition result of the movement of the user's mouth.

4.2 Gaze selection on mouth in narrow viewing field

The image information in non-**RGB** color space has shown less influenced by image acquisition conditions than in normal **RGB** color space [15]. The face segmentation

using color information has been studied [16], in which color spaces of **HSV** (hue, saturation and value) and **YIQ** are used.

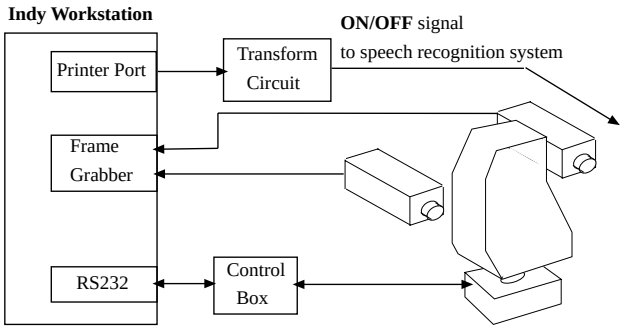


Fig. 13 The configuration of dialogue environment using the vision system with dual viewing angles

We here present a robust method to segment the mouth region from a color face image sequence which is taken from the pointable camera with narrow viewing angle. The images taken from pointable camera are represented with transformed **YIQ** formats instead of the original **RGB** images. With the empirical knowledge [17], the **Q**-component is well responded to the lips regions, and the (facial) skin area in **I**-component exhibits clear peak values.

Fig. 14 shows the 3 input **YIQ** images in (a)-(c), the results of logical AND operation between **I** image and **Q** image in (d), and the results of filtering with 4-neighborhood erosion filter in (e). Based on the result of filtering, we can set a proper region to indicate the position of mouth region. As shown in Fig. 14(e), the box around lips regions will be used to segment the mouth region from **Y** image. The segmentation result is shown in Fig. 14(f).

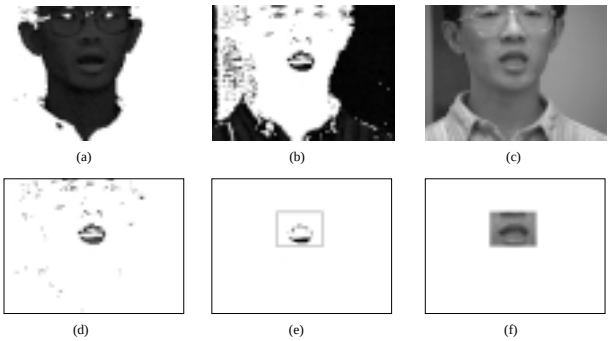


Fig. 14 A segmentation result of mouth region from **YIQ** images

4.3 Dialogue span detection

The intensity of the lips in Fig. 14(f) is so similar to the one of skin around lips that makes it difficult to extract the lips region stably. For this reason, many systems use special lighting or require the user to paint his lips with a special color for lip movement analysis. From the observation that the intensity of the region between lips is obviously lower than the region around it, and the shape of the region between lips meaningfully expresses the movement of lips, we analyze the shape variation of the region between lips to determine the open/close of mouth rather than to analyze the variation of lips movement directly. The flow diagram from the input of **YIQ** images of a user's head to the output of the region between two lips is shown in Fig. 15.

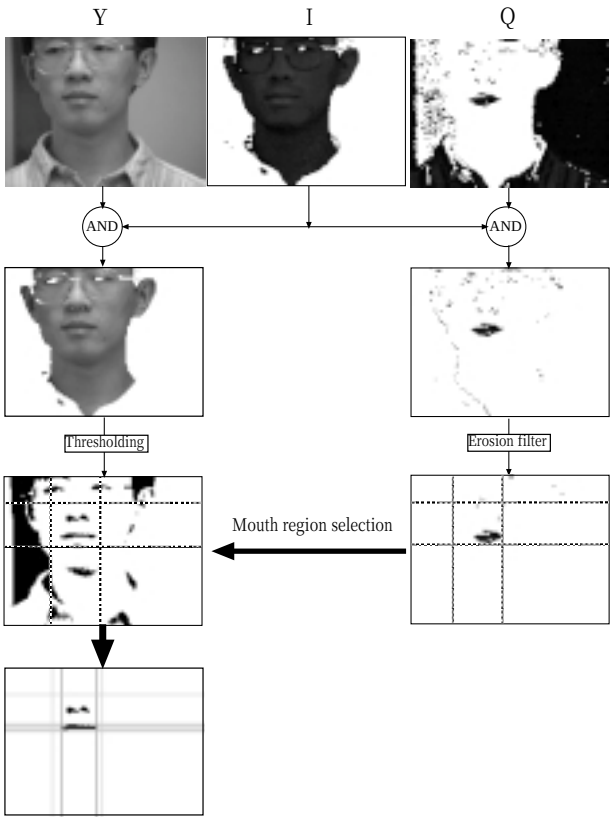


Fig. 15 Flow diagram for extracting the region between lips from **YIQ** images

To determine whether the mouth is open or closed from the shape of the region between lips, we need to describe it quantitatively. The width and the size of the region between lips are used as parameters to determine the open/close of mouth. Assuming that the user will open his mouth when uttering and close his mouth when not uttering, the begin/end of an utterance can be determined by the variations of the parameters. Fig. 16 shows an example of the variation of size and width of the region between lips when a user stand in front of the vision system, and user is allowed to move around as long as the image of his mouth kept within the viewing field of pointable camera. As shown in Fig. 16, both the size (a) and width (b) of the region between lips will be increased when the utterance is begun, and will be decreased when the utterance is ended. The size and the width of the region between lips stably remain small during not uttering.

However, during an utterance, the mouth will not always open. The temporarily closing of mouth when user speaking must be detected if the open/close of mouth is used to indicate the begin/end of an utterance.

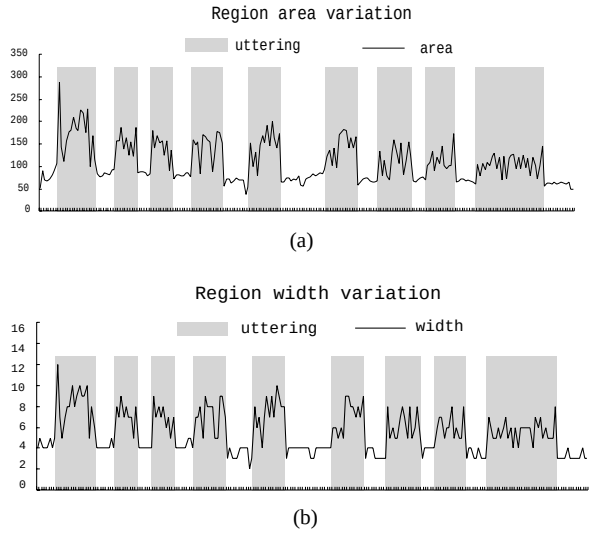


Fig. 16 The variation of size and width of the region between lips during a typical dialogue

From the observation of Fig. 16, when an utterance is really ended, the size of the region between lips will remain small at least two frames. The real end of an utterance can be detected by setting a counter to monitor the closure of mouth. The detection of the span of dialogue can be summarized as follows:

Begin of an utterance:

Both increase on the width and the size of the region between lips.

End of an utterance:

Both decrease on the width and the size of the region and the region size remains small at least two frames.

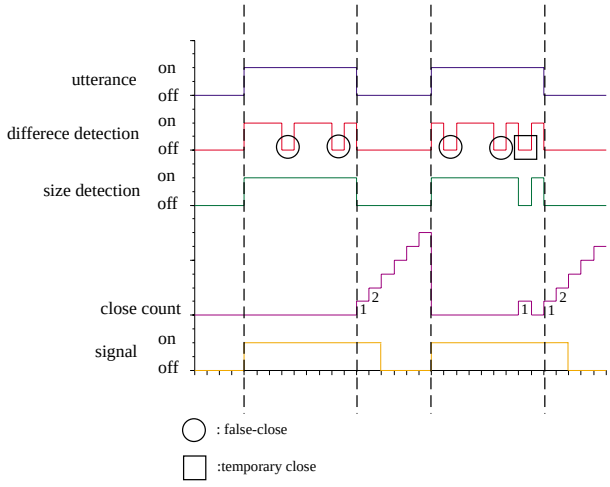


Fig. 17 Signal output according to the detection result of an utterance

The difference detection in Fig. 17 includes the increase/decrease detections of the region width and region size, and the result of size detection is on, if the size of region between lips remains small at least two frames. The on/off signal will be sent to the continuous speech recognition system according to the result of dialogue span detection. We are developing an anthropomorphous interface agent system called VSA with a realistic facial image and a speech dialogue function [18,19]. At present, the speech recognition system in our VSA uses a foot switch. Incorporating the vision system of this paper into the VSA, we are planning to make a more unconstrained and user-friendly environment of the VSA.

5 Conclusion

We have proposed a vision system with dual viewing angles which is capable of simultaneously tracking and recognizing a person in front of it, and constructed a user-friendly speech dialogue environment based on it.

The camera with wide viewing angle is fixed and provides wide viewing field for wide range motion tracking. The other camera with narrow viewing angle is mounted on a rotatable platform to centralize the image of target in its viewing field.

For catch up with the moving object fast, we have proposed a modified motion energy (*MME*) for estimation of movement, and the *MME* can also be used to determine a threshold value dynamically to extract the motion region between two consecutive frames. When background image is used in calculating *MME*, an existence energy (*EE*) is defined to detect the existence of an object and to segment the target from background image if it exists.

The gaze selection of the pointable camera can be done in either wide viewing field or narrow viewing field or in both viewing fields, according to the size of the target at which will be gazed. In our work, the gaze point selection on head from the body of a person is done in wide viewing field, and then followed a selection of gaze point on mouth from head in narrow viewing field.

To demonstrate the advantage and the performance of using dual viewing angles, we have constructed a novel human interface in speech dialogue system. By using the dual viewing angles, we have shown that the spatial constraints on common speech dialogue systems can be solved by the use of computer vision to detect the open/close of the user's mouth, and then to indicate the continuous speech recognition system the begin/end of the user's utterance. We have also shown that the detection of the span of dialogue is stable when we use the region between lips to express the open/close of mouth.

Acknowledgments

This research was supported by the proposal-based advanced industrial technology R&D program of NEDO and the part-in-aid for developmental scientific research (No. 06558045) of the Ministry of Education.

References

[1] B. A. Mayer and M. B. Rosson, "Human Factor in Computing Systems," Proc. SIGCHI'92, Monterrey, CA, 1992.

[2] A. Azarbayejani, T. Starner, B. Horowitz and A. Pentland, "Visually Controlled Graphics," IEEE Trans. on PAMI, Vol. 15, No. 6, pp. 602-605, 1993.

[3] A. H. Gee and R. Cipolla, "Non-Intrusive Gaze Tracking for Human-Computer Interaction," Proc. Mechatronics and Machine Vision in Practice, Toowoomba, Australia, 1994.

[4] T. E. Hutchinson, K. P. White, W. N. MArtin, K. C. Reichert, and L. A. Frey, "Human-Computer Interaction Using Eye-Gaze Input," IEEE Trans. on Systems, Man and Cybernetics, Vol. 19, No. 6, pp. 1527-1534, 1989.

[5] S. Baluja and D. Pomerleau, "Non-Intrusive Gaze Tracking Using Artificial Neural Networks," Tech. Report, Carnegie Mellon Univ., CMU-CS-94-102, 1994.

[6] A. Tomono, F. Kishino and Y. Kobayashi, "Pupil Extraction Processing and Gaze Point Detection System Allowing Head Movement," IEICE Trans. Information and System, Vol. J76., No. 3, pp.636-646, 1993.

[7] J. M. Rehg and T. Kanade, "DigitEyes: Vision-Based Human Hand Tracking," CMU-CS-93-220, 1993.

[8] T. Baudel and M. Beaudouin-Lafon, "Charade: Remote control of Objects Using Free-Hand Gestures," Communication of the ACM, Vol. 36, No. 7, pp. 28-35, 1993.

[9] Y. Huang, H. Dohi and M Ishizuka,"A Realtime Visual Tracking System with Two Cameras for Feature Recognition of Moving Human Face," Proc. 4th IEEE Int' Wrokshop on Robot and Human Communication(RO-MAN'95), pp. 170-175, Tokyo, 1995.

[10] K. Mase and A. Pentland, "Automatic Lipreading by Optical-Flow Analysis," IEICE Trans. Information and System, Vol. J73., No. 6, pp.796-803, 1990.

[11] K. Ebihara, J. Ohya, F. Kishino, "A Study of Real Time Facial Expression Detection for Visual Space Teleconferencing," IEEE Int'l. Workshop on Robot and Human Communication, Tokyo, pp. 247-252, 1995.

[12] N. Kita, "Active vision System using Human Vision as Inspiration," Jour. IPS Japan, vol. 36, No. 3, pp. 264-272, 1995.

[13] D. Murray and A. Basu, "Motion Tracking with an Active Camera," IEEE Trans. on PAMI, Vol. 16, No. 5, pp. 449-459, 1994.

[14] A. Pentland, "Machine Understanding of Human Action," Proc. Int' Forum on the Frontier of Telecommunication Tech., Tokyo, 1995.

[15] D.H. Ballard and C. M. Brown, "Principles of Animate Vision," Computer Vision Graphics and Image Processing, Vol. 56, No. 1, pp. 3-21, 1992.

[16] T. Miyawaki, S. Ishibashi and F. Kishino, "A Region Segmentation Method Using Color Information," Proc. IMAGE'COM90, Bordeaux, 1990.

[17] S. Akamatsu, T. Sasaki, H. Fukamachi and Y. Suenaga, "Automatic Extraction of Target Images for Face Identification Using the Sub-Space Classification Method," IEICE Trans. Information and System, Vol. E76-D, No. 10, pp. 1190-1198, 1993.

[18] H. Dohi and M. Ishizuka, "A Visual Software Agent connected with WWW/Mosaic," Proc. Multimedia Japan '96, pp. 392-397, Yokohama, 1996

[19] Y. Hiramoto, H. Dohi and M. Ishizuka, " A Speech Dialogue Management System for Human Interface employing Visual Anthropomorphous Agent," Proc. 3rd IEEE Int'l Workshop on Robot and Human Communication(RO-MAN'94), pp. 277-282, Nagoya, 1994.

Ying-Jieh Huang

was born in Tainan, Taiwan, in 1963. He received the B.S. in electronic engineering from National Taiwan Institute of Technology in 1988, and M.S., Ph. D. in electronic engineering from the University of Tokyo in 1993 and 1996. He now works at Information and Communication R&D Center of Ricoh Co. Ltd.. His research area includes image coding, image processing and computer vision.

Hiroshi Dohi

received his B.S. and M.S. in electrical engineering from Keio University. He is now a research associate at Dept. of Information and Communication Eng., University of Tokyo. Current research interests include Internet-based anthropomorphic interface agent, multimodal interface, and advanced human computer interaction.

Mitsuru Ishizuka

earned his B.S. , M.S., and Ph. D. in electronic engineering from the University of Tokyo. He is now a professor at Dept. of Information and Communication Eng. of the same university. Prior to his current position, he worked at NTT Yokosuka Lab. and Institute of Industrial Science, the University of Tokyo. His current research area includes artifical intelligence, multimodal anthropomorphic interface and software angent for WWW information space.