*MVA '96* IAPR Workshop on Machine Vision Applications, November. 12-14, 1996, Tokyo, Japan

# A Vision System with Dual Viewing Angles for Spontaneous Speech Dialogue Environment*

Ying-Jieh HUANG †, Hiroshi DOHI and Mitsuru ISHIZUKA
‡ Dept. of Information and Communication Engineering, The Univ. of Tokyo

## Abstract

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 this 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 a sufficient resolution. For a fast and robust motion tracking, we define a modified motion energy (*MME*) and a existence energy (*EE*) for detecting the motion of the target and extracting the motion region at the same time. Instead of using a physical device such as a foot switch, the begin/end of user's utterance are determined by detecting the movement of mouth.

## 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].

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

When computer vision is used in 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 will also be showed.

## 2 Vision System with Dual Viewing Angles

The goal of our vision system here is to achieve a feature recognition on a movable target with a reasonable movement. 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. 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.

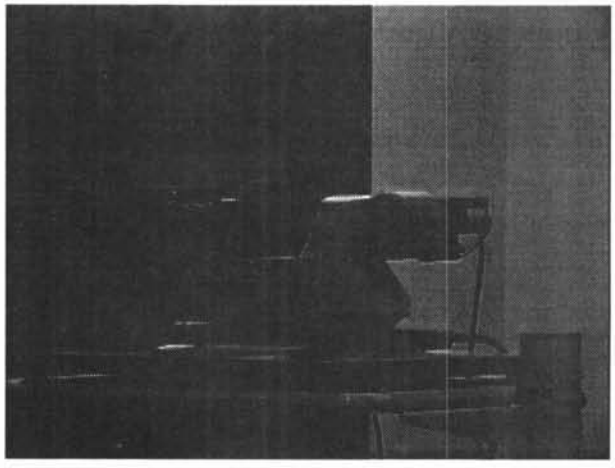

Fig.1 The vision system with dual viewing angles

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

† Presently with Ricoh Information and Communication R&D Center
‡ Address : 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan.
E-mail : Huang@ic.rdc.ricoh.co.jp
E-mail : dohi@miv.t.u-tokyo.ac.jp
E-mail : ishizuka@miv.t.u-tokyo.ac.jp

*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.

and tilt. With this configuration, the vision system provides a foveal-peripheral vision acuity analogues to that of human vision system.

This configuration allows the vision system to track quickly a moving target within the wide viewing field and to get an stable image 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 be also dissolved.

Not only the rotation of pan and tilt, but also the zoom and the focus of the pointable camera can be controlled via RS-232 interface. The hardware specifications of the vision system are shown in Table 1.

Table 1 The hardware specifications of the vision system with dual viewing angles

| 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 |

The camera model used in our work is based on the approximation of pinhole camera model, and the perspective projection transform is used to map the coordinates of points in the 3D world space into 2D image coordinates.

To identify the position of a point in 3D, a fixed coordinate frame of reference, called world frame, is required. The origin of world frame is chosen at the lens center of the fixed camera, and is oriented so that the Z-axis coincides with the optical axis of the fixed camera and parallels to the optical axis of the pointable camera when it 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. 2.

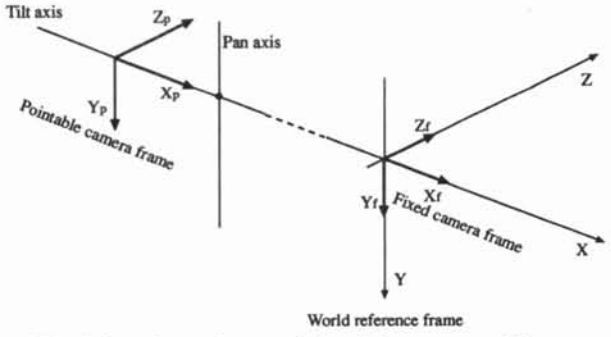


Fig. 2 Coordinate frames of the vision system with two cameras

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**:

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

where $\theta$ and $\phi$ 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:

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

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:

$$X = L\tan\theta - l - r\sec\theta \quad (6)$$

$$Y = -L_p\sin\varphi \quad (7)$$

$$L_p = L\sec\theta - r\tan\theta \quad (8)$$

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 $Q$.

## 3 Motion Tracking and Gaze Initialization in Wide Viewing Field

The image size of a target, e.g. a person, in wide viewing field is too small that recognition-based motion tracking is 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 is detected through a spatiotemporal filter which is implemented simply by image subtraction. Then the motion region can be extracted by thresholding the output of the image subtraction. Since the threshold value is empirically turned, 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:

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

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. According to the values of *MME*, the movement of the object can be described qualitatively such as not moved, move slowly or move fast. Fig. 6 shows the variation of *MME* when a person is in the wide viewing field (gray zone in Fig. 3).

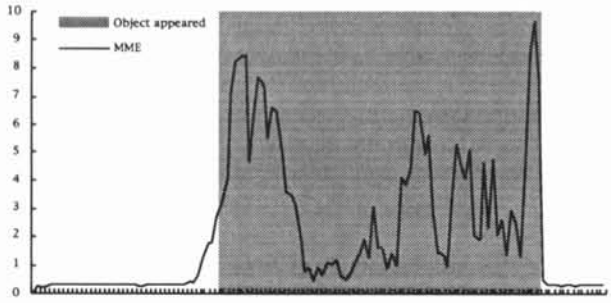


Fig. 3 The MME variation when a person exists in the wide viewing field

When *MME* is calculated from a subtraction image between the image with a 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. 4, the *EE* changes extremely when an object is appeared and disappeared.

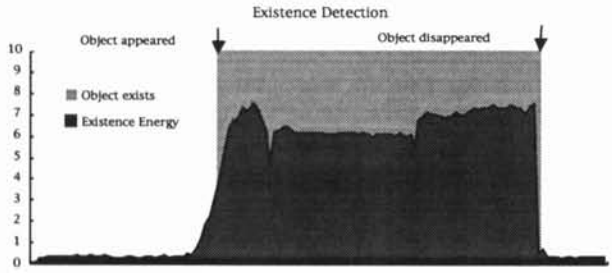


Fig. 4 Object appearance detection with EE

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 be also used to determine the threshold value for extracting it from background.

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.

## 4 Spontaneous Speech Dialogue System

After the high resolution image of the user's head is fetched by the pointable camera, many applications can be implemented on it. We show a implementation of spontaneous speech dialogue system based on our proposed vision system.

In a common dialogue environment, the user must move to a predetermined position, then uses 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.

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 [11], the Q-component is well responded to the lips regions, and the (facial) skin area in I-component exhibits clear peak values.

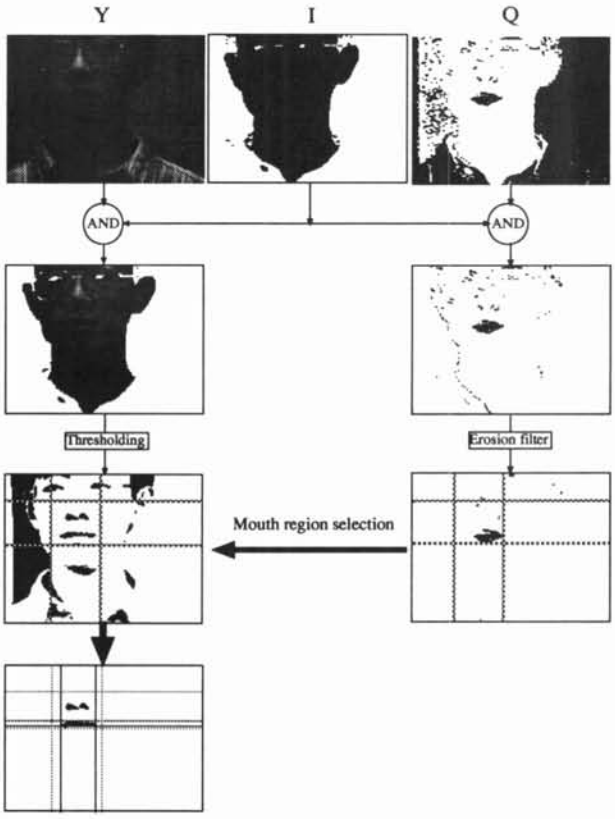


Fig. 5 Flow diagram for extracting the region between lips from YIQ images

The intensity of the lips 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 fact of the intensity of the region between lips is obviously lower than the region around it, and the shape of the region be-

tween 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. 5.

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/close his mouth when utters/not utters, the begin/end of the utterance can be determined by the variations of those parameters. As shown in Fig. 6, 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 no uttering. However, during an utterance, the mouth will not always open. The temporarily closing of mouth when speaking must be detected when the user is speaking, the temporary closing of mouth must be detected, if the open/close of mouth is used to indicate the begin/end of an utterance.

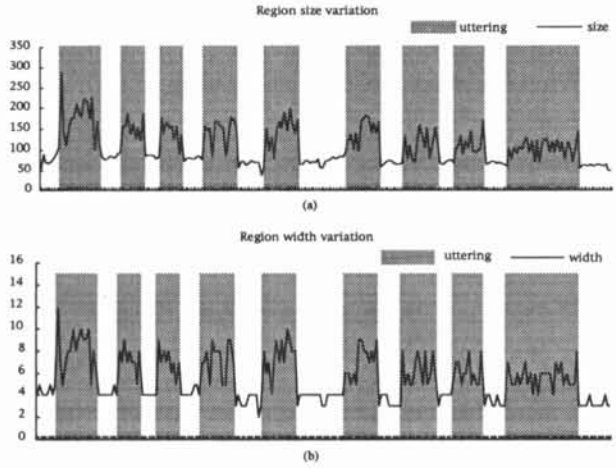


Fig. 6 The variation of size and width of the region between lips during a typical dialogue

From the observation on Fig. 6, 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.

We are developing an anthropomorphous interface agent system called VSA with a realistic facial image and a speech dialogue function [9,10]. At present, the speech recognition system in our VSA uses a foot switch. Incorporating the vision system with dual viewing angles 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.

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

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.

## References


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

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

[3] 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.

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

[5] 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.

[6] 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.

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

[8] 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.

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

[10] 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.

[11] 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.