Behavior Recognition Based on Head Pose and Gaze Direction Measurement

Yoshio Matsumoto[†], **Tsukasa Ogasawara**[†], **Alexander Zelinsky**[‡]
[†] Nara Institute of Science and Technology
8916-5 Takayamacho, Ikoma-city, Nara, Japan
yoshio@is.aist-nara.ac.jp
[‡] The Australian National University

Abstract

To build smart human interfaces, it is necessary for a system to know a user's intention and point of attention. Since the motion of a person's head pose and gaze direction are deeply related with his/her intention and attention, detection of such information can be utilized to build natural and intuitive interfaces.

In this paper, we describe a behavior recognition system based on the real-time stereo face tracking and gaze detection system to measure head pose and gaze direction simultaneously. The key aspect of our system is the use of real-time stereo vision together with a simple algorithm which is suitable for real-time processing. Since the 3D coordinates of the features on a face can be directly measured in our system, we can significantly simplify the algorithm for 3D model fitting to obtain the full 3D pose of the head compared with conventional systems that use monocular camera. Consequently we achieved a non-contact, passive, real-time, robust, accurate and compact measurement system for head pose and gaze direction. The recognition of attentions and gestures of a person is demonstrated in the experiments.

1 Face and Gaze Detection for Visual Human Interfaces

Smart human interfaces need to know a user's intention and attention. For example, the direction of the gaze can be used for controlling the cursor on a monitor, and the motion of the head can be interpreted as a gesture such as "yes" or "no".

Several kinds of commercial products exist to detect a person's head position and orientation, such as magnetic sensors and link mechanisms. There are also several companies supporting products that perform eye gaze tracking. These products are generally highly accurate and reliable, however all requires either expensive hardware or artificial environments (cameras mounted on a helmet, infrared lighting, marking on the face etc). The discomfort and the restriction of the motion affects the person's behavior, which therefore makes it difficult to measure his/her natural behavior.

To solve this problem, many research results have been reported to visually detect the pose of a head [1, 2, 3, 4, 5, 6, 7]. Recent advances in hardware have allowed vision researchers to develop real-time face tracking systems. However most of these systems use a monocular vision. Recovering the 3D pose from a monocular image stream is known to be a difficult problem, and high accuracy as well as robustness are hard to be achieved. Therefore, some approaches can not compute the full 3D, 6DOF pose of the head, while other methods are not sufficiently accurate as a measurement system. Some researchers have also developed vision systems to passively detect gaze point [8, 9, 10, 11], however, none of which can measure the 3D vector of the gaze line.

In order to construct a system which observes a person without giving him/her any discomfort, it should satisfy the following requirements:

- non-contact
- passive
- real-time
- robust to occlusions, deformations and lighting fluctuations
- compact
- accurate
- able to detect head pose and a gaze direction simultaneously

Our system[12] satisfies all these requirements by utilizing the following techniques:

- real-time stereo vision hardware using a field multiplexing device,
- image processing board with normalized correlation capability,
- 3D facial feature model and model fitting based on virtual springs,
- 3D eye model which assumes the eyeball to be a sphere.

In this paper, the details of the measurement system is described in Section 2. Then the method and experimental results of behavior recognition are described in section 3, and the conclusions and a discussion of the future work are given in Section 4.

2 Head Pose and Gaze Direction Measurement System

2.1 System Overview

The hardware setup of our real-time stereo face tracking system[12] is shown in **Figure 1** . We use a NTSC camera pair (SONY EVI-370DG × 2) to capture images of a person's face. The output video signals from the cameras are multiplexed into one video signal by the "field multiplexing technique"[13]. The multiplexed video stream is then fed into a vision processing board (Hitach IP5000), where the pose of the head and the direction of the gaze are calculated.

The outline of the software configuration for face tracking and gaze detection is shown in **Figure 2** . It consists of three major parts, 1) Initialization, 2) Face Tracking and 3) Gaze Detection.

2.2 Face Tracking

In the initialization stage, the system searches the face in the whole image using a 2D template of the whole face. After a face is found, the system starts face tracking where a 3D facial feature model is used to determine the 3D pose of the head. If the tracking of the face is not successful, the system regards the face to be lost and it jumps back to the initialization stage to find the face again. If the tracking of the face is successful, the system then calculates the direction of the gaze in the gaze detection stage. The 3D head pose and the 3D eye model are used to determine the 3D gaze vector. Finally, the system jumps back to the face tracking stage in the next frame.

The whole tracking process takes approximately 30[ms] which is well within the NTSC video frame rate. The accuracy of the tracking is approximately ±1[mm] in translation and ±1[deg] in rotation.

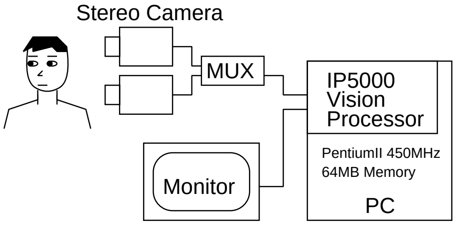


Fig. 1 : Hardware Configuration of Measurement System.

Even when deformations of the facial features and partial occlusions occurs, our tracking system works quite robustly due to the model fitting method. By utilizing the normalized correlation function on the IP5000, the tracking system is also tolerant to significant fluctuations in lighting.

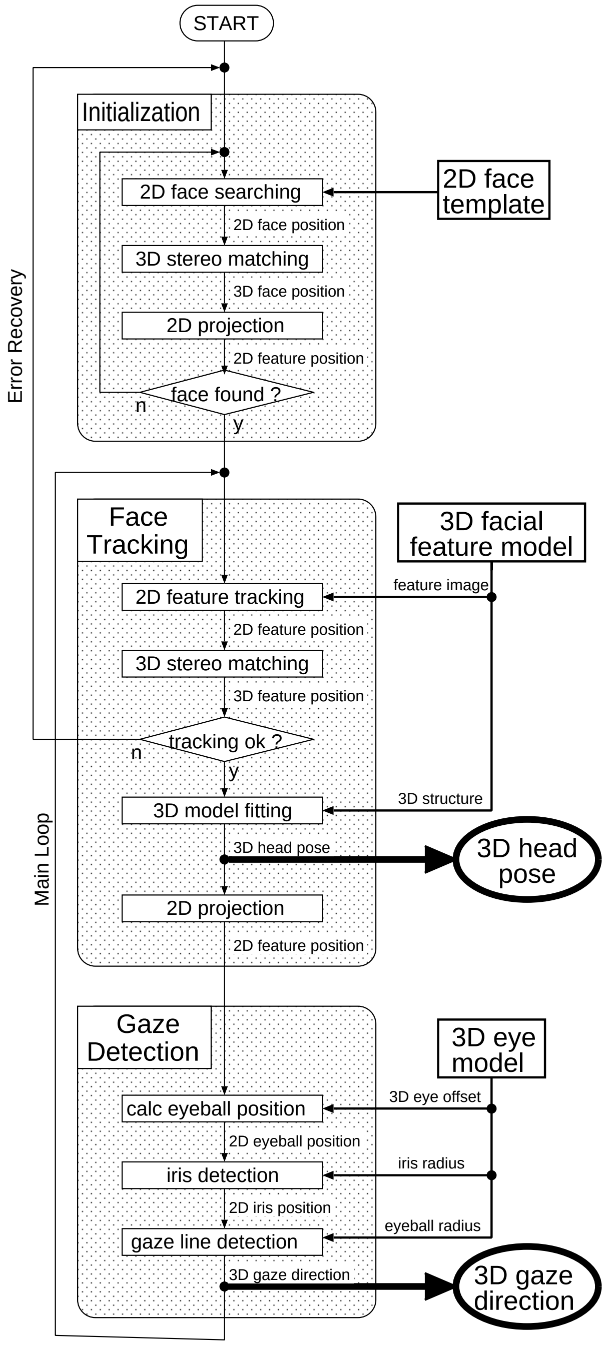


Fig. 2 : Software configuration.

2.3 Gaze Detection

In the modeling of the gaze line, the eyeballs are regarded as spheres. Gaze direction is determined based on both the pose of the head and the position of the irises of the eyes. The 3D eye model consists of following parameters:

- the relative position of the center of the eyeball respect to the head pose,
- radius of the eyeball,
- radius of the iris.

The relative position of the center of the eyeball is defined as a 3D vector from the mid-point of the corners of an eye to the center of the eyeball, and termed as an "offset vector." The radius of the eyeball is a value around 13[mm], and the radius of the iris is a value around 7[mm]. These parameters are currently determined by the manual adjustment through a training sequence where the gaze point of a person is known.

Figure 3 illustrates the process to determine the 3D gaze direction. As shown in **Figure 3** (1), the 3D position of the eyeball can be determined from the pose of the head using the offset vector, although the eyeball center cannot be seen. By projecting the eyeball with known position and size back onto the image plane, the 2D appearance of the eyeball can be determined (**Figure 3** (2)). Next, the center of the iris is detected by using the circular Hough Transform as shown in **Figure 3** (3). Since the corner of the eyes are already known, the iris detection is executed on a small region between them, which typically takes about 10[ms].

The relationship between the iris center and eyeball center in the image plane defines the orientation of the gaze vector. **Figure 3** (4) indicates how to compute the horizontal angle of the gaze vector θ, and the vertical angle can be computed in the same manner. The modeling of the gaze direction in our system is quite straightforward, however, since it requires accurate head pose to determine the eyeball center, no other research has successfully adopted such simple modeling so far.

There are four eyes in total in stereo image pair, therefore four gaze direction are detected independently. However each measurement is not sufficiently accurate, mainly due to the resolution of the image. The field of view of the camera is set to capture the whole face in the image, then the width of an eye is only about 30[pixel] and the radius of the iris is only about 5[pixel] in a typical situation. Therefore it is hard to determine the "gaze point" in a 3D scene by calculating the intersection of the detected gaze lines. Therefore the those four vectors are currently averaged to generate a single gaze vector in order to reduce the effect of noises.

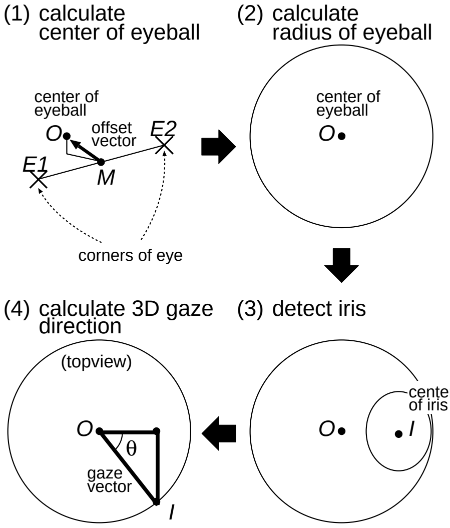


Fig. 3 : Modeling of gaze direction.

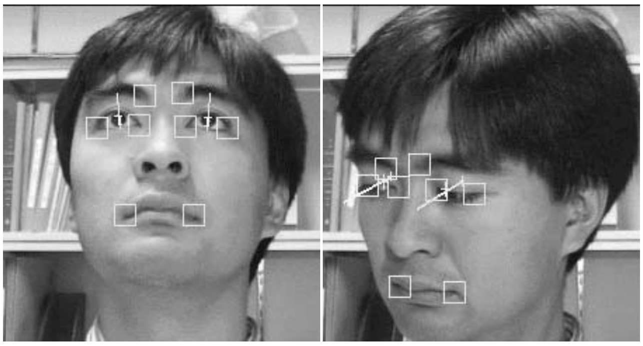


Fig. 4 : Result of detection of gaze direction.

2.4 Experimental Results

Figure 4 shows some snapshots obtained in a real-time gaze detection experiment. The 3D gaze vectors are superimposed on the tracking result. The whole process including face tracking and gaze detection takes about 45[ms], thus the 3D gaze vector can be determined at 15[Hz].

The accuracy of the gaze direction are evaluated through a experiment using a board with markers, which is shown in **Figure 5** . The person sits 0.8[m] away from the camera pair. The distance between the markers which is 10[cm], which corresponds to 5.7[deg] in terms of the gaze direction. The results shown in **Figure 5** indicates the accuracy of the gaze vector is about 3[deg] in the worst case.

3 Attention Recognition

3.1 Modeling of Environments

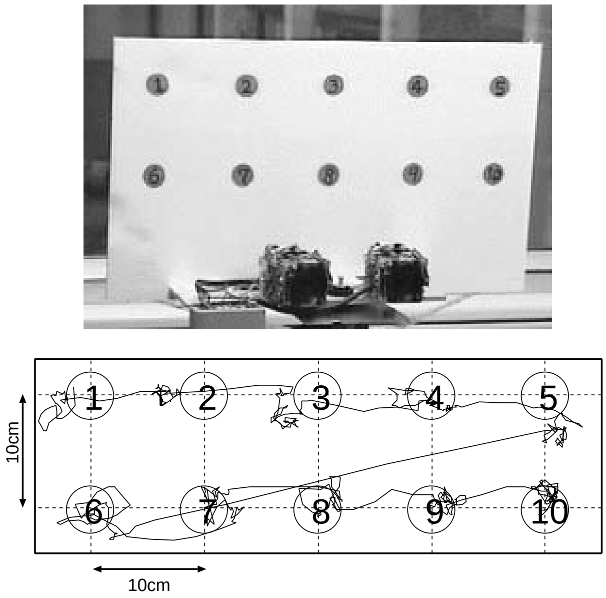


Fig. 5 : Result of accuracy assessment of gaze direction.

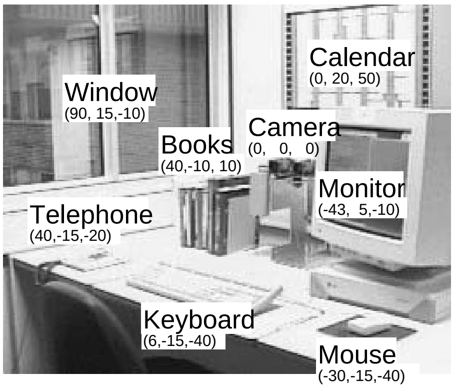


Fig. 6 : Experimental environment in office.

Figure 6 shows an desktop environment for the attention recognition experiment. There are objects such as a stereo camera pair, a telephone, books, a keyboard, a mouse, and monitor on the desk. A calendar on the wall and a window also exist in the environment. The approximate 3D coordinates of those objects are manually measured, and are registered in the system beforehand.

3.2 Modeling of Attention

Our system can only detect one direction as the gaze. The vergence angle which is related with the distance to the object cannot be obtained. This is because the gaze directions obtained independently are averaged to reduce the noise. Thus the object which a person is paying attention to is recognized based on the angle between the gaze vector and the vector from the head of the person to the object. More concretely, $object_i$ which has smallest angle θ_i in **Figure 7** is selected as the attentional object.

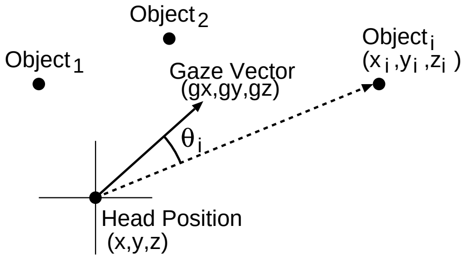


Fig. 7 : Detection of object in attention.

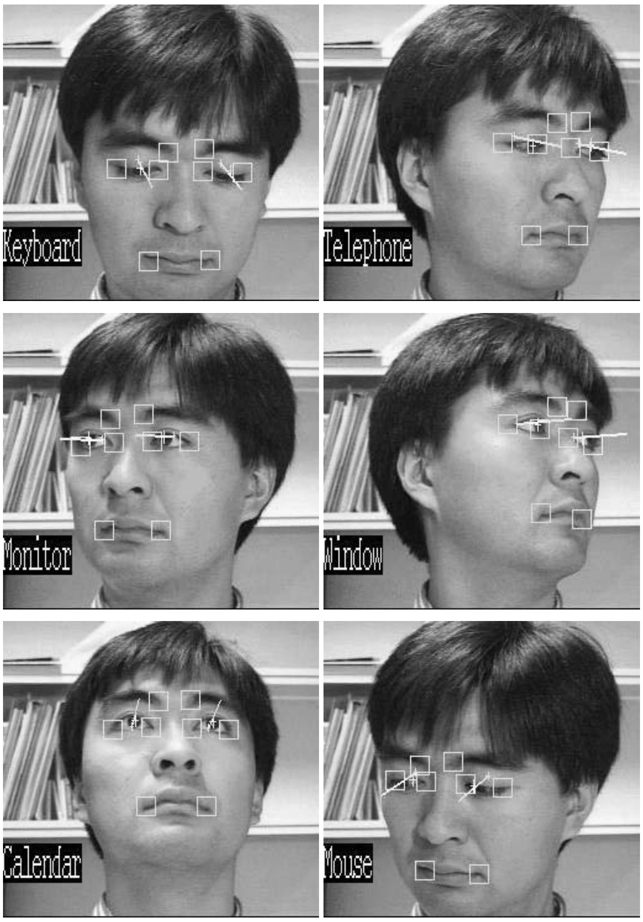


Fig. 8 : Experimental results of attention recognition in an office environment.

3.3 Experimental Results

The minimum distance between two objects in this environment was about 10[cm], and the system was able to recognize human's attention robustly. **Figure 8** shows experimental results. When the person looked down on the desk, irises are mostly hidden by eyelids and hard to be seen. However, the system was still able to select an object which the person is paying attention to. This process runs at 15[Hz].

4 Gesture Recognition

4.1 Gesture Recognition based on CDP

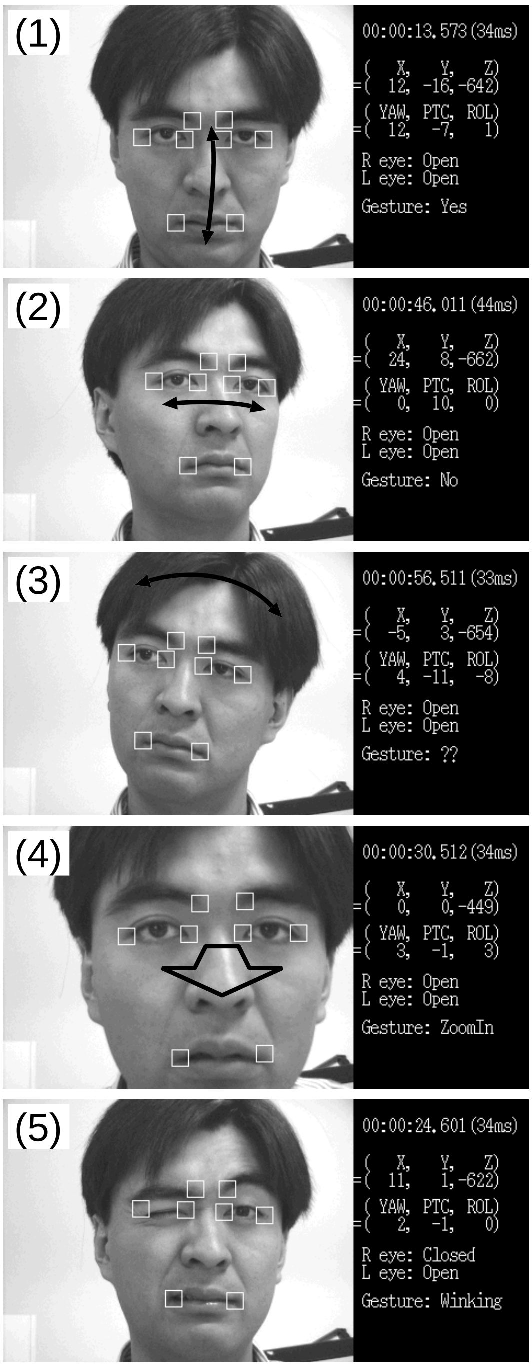
Since the system can measure the head motion of a person accurately, gestures represented by head motions can also be recognized. The high accuracy in head tracking enables the system to recognize not only explicit and over-acted gestures but also gestures with small head motions.

In our experiments of gesture recognition, Continuous DP matching[14] is used with angular velocities of three degrees as input data. Thus a gesture with low CDP value (matching error) is regarded as the current gesture. In this experiment, several gestures such as "Yes", "No", "Maybe", "Look Left", "Look Right", "Zoom-In" and "Zoom-Out" are registered.

4.2 Experimental Results

Figure 9 and **Figure 10** shows experimental results of gesture recognition. **Figure 9** shows five snapshots from recognized gestures. They correspond to "Yes", "No", "Maybe", "Zoom-In" and "Winking" gestures respectively. The "Winking" gesture is recognized based on the horizontal edge detection, while other gestures are recognized based on the angular velocities.

The upper graph in **Figure 10** shows the measured head motions in terms of angular velocities (roll, pitch and yaw), while the lower graph shows the CDP values for"Yes" and "No" gestures. As described above, low CDP values in the lower graph in **Figure 10** indicate that the measurement matches the registered actions. Therefore the meshed parts, Ⓐ,Ⓑ and Ⓒ in the graph correspond to gestures of "Yes", "No" and "Yes" respectively.



```
00:00:13.573(34ms)
(   X,    Y,    Z)
=(  12,  -16, -642)
( YAW,  PTC,  ROL)
=(  12,   -7,    1)
R eye: Open
L eye: Open
Gesture: Yes
```

```
00:00:46.011(44ms)
(   X,    Y,    Z)
=(  24,    8, -662)
( YAW,  PTC,  ROL)
=(   0,   10,    0)
R eye: Open
L eye: Open
Gesture: No
```

```
00:00:56.511(33ms)
(   X,    Y,    Z)
=(  -5,    3, -654)
( YAW,  PTC,  ROL)
=(   4,  -11,   -8)
R eye: Open
L eye: Open
Gesture: ??
```

```
00:00:30.512(34ms)
(   X,    Y,    Z)
=(   0,    0, -449)
( YAW,  PTC,  ROL)
=(   3,   -1,    3)
R eye: Open
L eye: Open
Gesture: ZoomIn
```

```
00:00:24.601(34ms)
(   X,    Y,    Z)
=(  11,    1, -622)
( YAW,  PTC,  ROL)
=(   2,   -1,    0)
R eye: Closed
L eye: Open
Gesture: Winking
```

Fig. 9 : Recognized gestures: (1) "Yes", (2) "No", (3) "Maybe", (4) "Zoom-In" and (5) Winking gestures respectively.

5 Conclusion

In this paper, a behavior recognition system based on a head pose and gaze direction measurement was presented. The system consists of a stereo camera pair and a standard PC equipped with an image processing board. The measurement system is (1) non-contact, (2) passive, (3) real-time and (4) accurate, all of which have not been able to be achieved by previous research results. The recognition of attentions and gestures of a person is demonstrated in the experiments using the measurement system.

This system can be applied to various targets, such as psychological experiments, ergonomic designing, products for the disabled and the amusement industry. In our future work, we will evaluate the accuracy of the head pose and the gaze direction. We also aim to improve the accuracy and processing speed of the gaze detection.

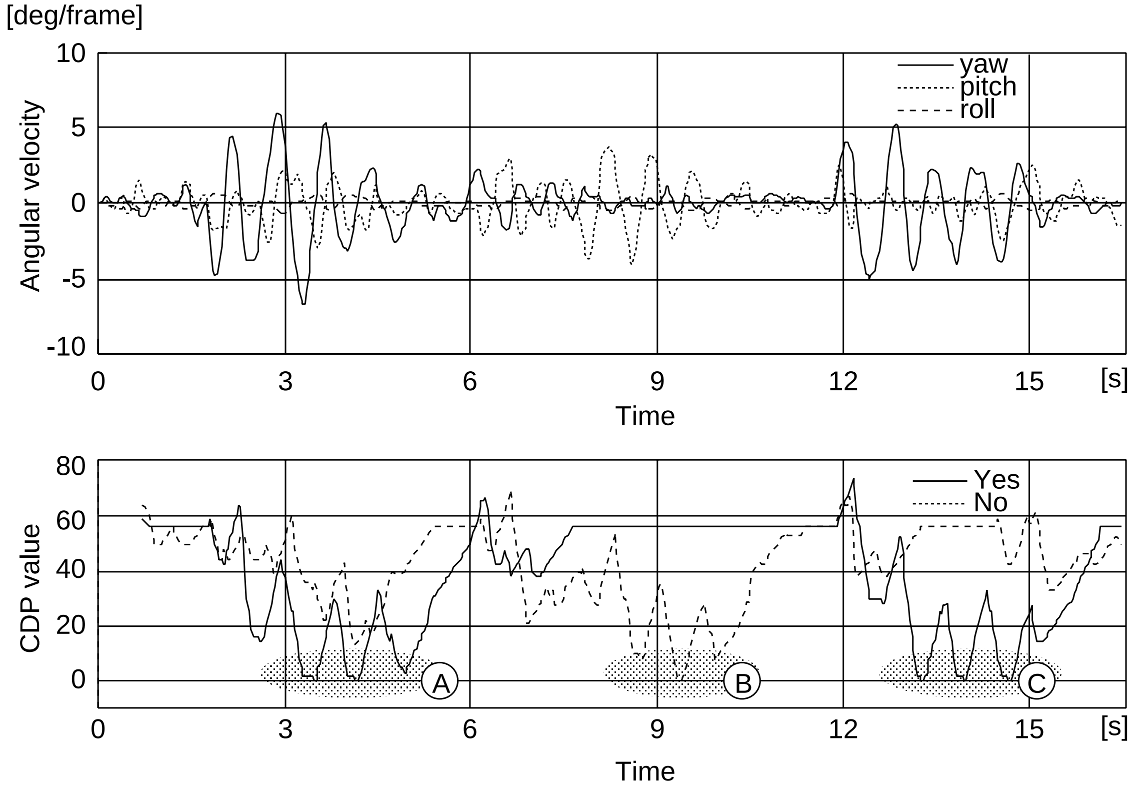


Fig. 10 : Experimental results of gesture recognition: (Upper) Input angular velocities, (lower) Output CDP Values (i.e. matching errors).

References

[1] A.Azarbayejani, T.Starner, B.Horowitz, and A.Pentland. Visually controlled graphics. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(6):602–605, 1993.

[2] A.Zelinsky and J.Heinzmann. Real-time Visual Recognition of Facial Gestures for Human Computer Interaction. In *Proc. of the Int. Conf. on Automatic Face and Gesture Recognition*, pages 351–356, 1996.

[3] P.Ballard and G.C.Stockman. Controlling a Computer via Facial Aspect. *IEEE Trans. Sys. Man and Cybernetics*, 25(4):669–677, 1995.

[4] Black and Yaccob. Tracking and Recognizing Rigid and Non-rigid Facial Motions Using Parametric Models of Image Motion. In *Proc. of Int. Conf. on Computer Vision (ICCV'95)*, pages 374–381, 1995.

[5] S.Birchfield and C.Tomasi. Elliptical Head Tracking Using Intensity Gradients and Color Histograms". In *Proc. of Computer Vision and Pattern Recognition (CVPR'98)*, 1998.

[6] A.Gee and R.Cipolla. Fast Visual Tracking by Temporal Consensus. *Image and Vision Computing*, 14(2):105–114, 1996.

[7] Kentaro Toyama. Look, Ma – No Hands! Hands-Free Corsor Control with Real-time 3D Face Tracking. In *Proc. of Workshop on Perceptual User Interface (PUI'98)*, 1998.

[8] Shumeet Baluja and Dean Pomerleau. Non-Intrusive Gaze Tracking Using Artificial Neural Networks. Technical Report CMU-CS-94-102, CMU, 1994.

[9] C.Colombo, S.Andronico, and P.Dario. Prototype of vision-based gaze-driven man-machine interface. In *Proc. of IEEE/RSJ Int. Workshop on Intelligent Robots and Systems*, pages 188–192, 1995.

[10] J.Heinzmann and A.Zelinsky. 3-D Facial Pose and Gaze Point Estimation using a Robust Real-Time Tracking Paradigm. In *Proc. of the Int. Conf. on Automatic Face and Gesture Recognition*, 1998.

[11] R.Stiefelhagan, J.Yang, and A.Waibel. Tracking Eyes and Monitoring Eye Gaze. In *Proc. of Workshop on Perceptual User Interface (PUI'97)*, 1997.

[12] Y.Matsumoto and A.Zelinsky. An Algorithm for Real-time Stereo Vision Implementation of Head Pose and Gaze Direction Measurement. In *Proc. of the Int. Conf. on Automatic Face and Gesture Recognition (FG'2000)*, pages 499–505, 2000.

[13] Y. Matsutmoto, T. Shibata, K. Sakai, M. Inaba, and H. Inoue. Real-time Color Stereo Vision System for a Mobile Robot based on Field Multiplexing. In *Proc. of IEEE Int. Conf. on Robotics and Automation*, pages 1934–1939, 1997.

[14] T.Nishimura and R.Oka. Spotting recognition of human gestures from time-varying images. In *In Proc. of the Second Intl. Conf. on Automatics Face and Gesture Recognition*, pages 318–322, 1996.