PTZ Control with Head Tracking for Video Chat

Kota Yamaguchi
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, 1138656 Japan
Kota_Yamaguchi@ipc.i.u-tokyo.ac.jp

Masatoshi Ishikawa
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, 1138656 Japan
Masatoshi_Ishikawa@ipc.i.u-tokyo.ac.jp

Takashi Komuro
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, 1138656 Japan
Takashi_Komuro@ipc.i.u-tokyo.ac.jp

CHI 2009, April 4 – 9, 2009, Boston, MA, USA
ACM 978-1-60558-246-7/09/04.

Abstract
This paper describes a user interface for video chat that is capable of panning, tilting, and zooming (PTZ) operation using head tracking. The approach is to map a captured 3D position from head tracker to PTZ parameters of a remote camera so that a user can intuitively change the view just as people change their sight by moving their head. The preliminary user study gave encouraging results and clarified the point for further improvement.

Keywords
HCI, vision-based interaction, head tracking, panning, tilting, zooming, PTZ control, wide field of view video chat

ACM Classification Keywords
H5.2. User Interfaces: Graphical user interfaces (GUI), H5.2. User Interfaces: Interaction styles (e.g., commands, menus, forms, direct manipulation), I4.9. Applications

Introduction
With the advancement of computer vision technology, vision-based interaction has become one of the popular ways to enhance user experience in human computer interaction. Successful applications that utilize vision-

based interaction include fancy video effects in video chat or automatic shutter by smile recognition in a digital camera. Vision-based interaction has several different characteristics over traditional input devices such as mouse or keyboard, and is gathering more interests as a way to provide a different experience. The trend of the lowering price of CMOS cameras would further increase the use of computer vision techniques for interaction design.

As one of attempts to use computer vision technique for interaction, we have worked on a user interface to control panning, tilting, and zooming (PTZ) parameters of a camera using head tracking. Our approach is to first capture 3D head position from a camera in front of the viewer, and then map it to PTZ parameters of a remote camera. Unlike mice, an absolute position is used as an input in our approach. All the processing works in real time.

We pick up the PTZ control in this work because we see head tracking as one of the ideal interaction methods for a task to control views. Since people unconsciously expect the change of their sight when they move their head, it is probable that changing a view along with the position of the head is acceptable and fits into the usual sense of people. People would feel stronger integrity of senses between their head position and sight than between other body parts like finger and sight. Our attempt is to incorporate this characteristic of head tracking for an interaction design and create more intuitive and comfortable interface.

Also, head tracking is suitable for controlling continuous parameters like those of PTZ. Parameter control typically requires users to find the best position with trial-and-error adjustment. For this kind of task, it is important that users can go back to a "home" position of the control, or users may lose the reference direction. To keep users aware the reference position, one good strategy is to use an absolute signal for an input channel like head position. With a head tracker configured to capture absolute 3D position of viewer's head, viewers just need to get back their pose to revert PTZ parameters.

Our head tracking interface isn't meant to replace well-established interface devices such as mouse or keyboard. Rather, our attempt is to find yet another supportive input that hasn't been utilized so far, or to improve user experience for a device that doesn't have comfortable input methods like the case of a mobile phone.

The target application of our PTZ control is video chat. By extending the remote view with configurable PTZ parameters, users could have a better feeling of sharing two sites in video chat. However, we do not limit our approach to video chat. The approach itself is generic enough to be used in other applications such as video conferencing or surveillance which requires PTZ control.

Related Works

Vision-based interaction is relatively new and still in the process of active research. However, several important discussions have been already made so far. Kjeldsen outlines in [1] various aspects of issues in vision-based interaction along with required tasks, input signals, and mapping between them. In [2], Jaimes summarizes types and techniques used in vision-based interaction along with modalities in human. As discussed in [2], we

believe that the integrity of senses is the key to design successful interaction using vision techniques.

Several studies have been made for head or face tracking for human computer interaction [3-6]. Wang et al. makes systematic research the effect of head tracking to game experience in [3], and reports positive results in its use. Attempts to improve GUI environment using head tracking are also presented in [4-6] with positive results and feedbacks. Like games, an immersive application may be directly benefited by the use of head tracking because of the similarity between actions and input methods. However, it is noticeable that the use of head tracking still contributes the improvement of usability for an artificial workspace such as desktop in GUI environment where users are not leveraged their spatial sense in the usual 3D space. As long as the use of head tracking incorporates the style of target action, users could feel comfortable with head tracking as a way for interaction in various applications.

The focus of interaction design for PTZ control has been mainly on the improvement of experience in video conferencing with new input methods such as sensor-based automation [7] or gesture input [8].

Our Approach

On designing vision-based interaction, it is a good practice to first clarify control action, control signal, and mapping between them [1]. In our approach, these are adjustment of PTZ parameters, captured position by head tracking, and a mapping function between two. Figure 1 illustrates the overview.

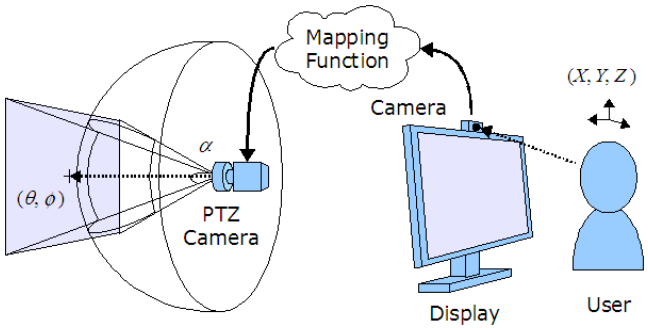

Figure 1. Overview of PTZ control with head tracking. Captured 3D head position (X, Y, Z) is mapped to PTZ parameters (θ, φ, α) using designated function.

Control Action: PTZ Parameters Adjustment
Parameters to be controlled are panning angle θ, tilting angle φ, and angle of view α. All parameters are continuous and independent each other.

Control Signal: Captured Head Position
We use a camera as a sensor to capture 3D position (X, Y, Z) of user's head. There are several approaches to capture head position from camera. Each approach has different characteristics in accuracy, speed, or robustness. In this work, we choose the combination of the face detection method based on Haar-like features [9] and Kalman filter for its speed and robustness.

In principal, a monocular camera can not capture the depth of head. It can measure only the 2D position (X, Y) and the size of face in the view. However, under the assumption that the size of actual face is constant, we can calculate the depth from the size of face in the view.

Mapping Function

It is not trivial to design an appropriate mapping from the control signals to the control actions, because captured position (X, Y, Z) and PTZ parameters (θ, φ, α) are independent measures. We put the following assumptions on designing the function.

- It is natural for a user if the direction of the optical axis of PTZ camera is synchronized to the direction of the display from the user's head.
- Zooming should occur when a user gets close to the display since people usually get close to look at the detail of a target.

As one of the possible mapping, we designed the following function in this work.

$$\theta = K_1 \arctan \frac{\sqrt{X^2 + Y^2}}{Z}$$

$$\varphi = K_2 \arctan \frac{Y}{X}$$

$$\alpha = \arctan \frac{K_3}{Z}$$

K_n is a parameter, and was chosen heuristically in this work. Note that it is possible that we have a better mapping function other than above. The function is chosen just as one possibility.

System Prototype

For a prototype, we implemented a video communication system composed of a client laptop that has a webcam and a fisheye camera connected to a PC in a remote place. The webcam of the laptop is used for head tracking, and the fisheye camera is used as a digital PTZ camera. We chose a fisheye camera instead of a regular PTZ camera to avoid control latency.

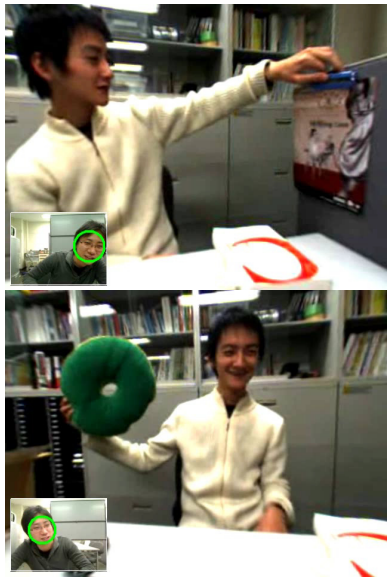

Figure 2. Screen captures from the prototype.

Using a captured head position, the client laptop synthesizes an appropriate perspective view from fisheye video stream in real time. The fisheye camera can capture 185 degree of view, however, only a small portion of range is used for synthesizing a perspective view due to the limitation of the webcam.

The prototype system is implemented using GStreamer, OpenGL, and OpenCV library on Ubuntu Linux. Head tracking routine can achieve 15 fps with a regular laptop under the configuration that the webcam captures color images with 160x120 pixels. The system is currently only capable of one directional communication without audio channel.

Preliminary Evaluation

We conducted a preliminary user study for our prototype to evaluate its usability and get feedbacks. 25 male subjects were introduced to our system. The average age was 24 years old.

In the evaluation, subjects were first given an explanation on how to control PTZ parameters with their head. Then, they sit in front of a client laptop and were told to look at a target indicated by an experimenter in a remote room. The experimenter moved around in the room so that subjects needed to change PTZ parameters to follow the target. One session took about 2 to 3 minutes. After the session, subjects were asked to respond to questions on a 5-point Likert scale.

The results from our user study are shown in Figure 3. Most of subjects reported positive feeling for the ease to learn the usage and showed their preference for our approach in video chat. Also, some commented that they would feel more comfortable with supportive input such as mouse using their spare hand. These would be benefits to use a camera as yet another interaction method.

However, some results indicate a room for improvement. Although most of negative comments for panning, tilting, and zooming usability were simply related to errors in head tracking, some subjects reported that tilting was more difficult than panning or zooming because moving their head upwards and downwards was not so easy on a chair. We put panning and tilting in one question because the control mapping was the same, but we would need to deal with them independently for the future with taking account into ergonomics. The result from subjects' feeling for motion range supports this response.

In addition, there were a few but important comments on PTZ control mapping to head position. Two subjects reported that they felt panning and tilting should occur in the opposite direction to the current setup as in the usual 3D translation. The difficult problem is that a camera does not have the capability to move its viewpoint unless the camera is on a robot. We might be able to pretend the view in that manner, but further research is needed on the preferable control mapping.

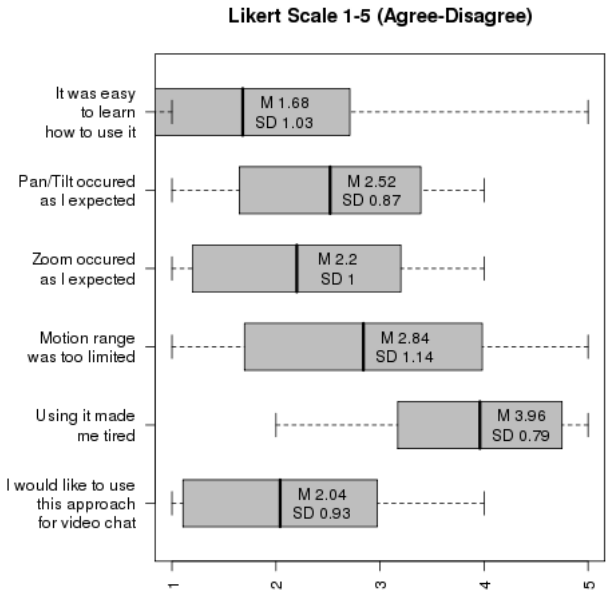

Figure 3. Results from subjective evaluation showing the mean (center), standard deviation (box), and min-max (break line) of the responses.

Another notable comment was that panning and tilting range was too limited when the subject tried to zoom up an object. It was due to the limited angle of view of the head tracker, but solving it would require a light-weight algorithm to detect a face in a wide field of view.

Conclusion and Future Works

We have presented the head tracking interface for PTZ control in video chat, and showed the results from preliminary evaluation using our prototype system. The results were encouraging, but indicated several improvements.

The most important point is ergonomics consideration. Since most of vision-based interaction captures the move or posture of body, its usability is significantly affected by the required range of motion. In our system, we would be able to improve the mapping between captured head position and PTZ parameters by taking into account the range of upper body movement when a viewer is sitting on the chair.

Further development for our prototype is ongoing to implement the capability of audio communication and bidirectional communication. At the same time, we will focus on the better design for the vision input and the control action. We hope that this paper will help research and development of vision-based interaction.

Acknowledgements

Authors would like to thank Yoshihiro Watanabe, Carson Reynolds, and Alvaro Casinelli for several useful advices and insightful discussions.

References

[1] Kjeldsen, R. and Hartman, J. Design issues for vision-based computer interaction systems. In *Proc. PUI 2001*, ACM Press (2001), 1-8.

[2] Jaimes, A. and Sebe, N. Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*. Elsevier Science Inc. (2007), 116-134.

[3] Wang, S., Xiong, X., Xu, Y., Wang, C., Zhang, W., Dai, X., and Zhang, D. Face-tracking as an augmented input in video games: enhancing presence, role-playing and control. In *Proc. CHI 2006*, ACM Press (2006), 1097-1106.

[4] Kitajima, K., Sato, Y., and Koike, H. Vision-Based Face Tracking System for Window Interface: Prototype Application and Empirical Studies. *Ext. Abstracts CHI 2001*, ACM Press (2001), 359-360.

[5] Ashdown, M., Oka, K., and Sato, Y. Combining Head Tracking and Mouse Input for a GUI on Multiple Monitors. *Ext. Abstracts CHI 2005*, ACM Press (2005), 1188 – 1191.

[6] Reynolds, C., Cassinelli, A., and Ishikawa, M. Meta-perception: reflexes and bodies as part of the interface. In *Proc. CHI 2006*, ACM Press (2008), 3669-3674.

[7] Strubbe, H. and Lee, M.S. UI for a Videoconference Camera. *Ext. Abstracts CHI 2001*, ACM Press (2001), 333-334.

[8] Liao, C., Liu, Q., Kimber, D., Chiu, P., Foote, J., and Wilcox, L. Shared interactive video for teleconferencing. In *Proc. MULTIMEDIA 2003*, ACM Press (2003), 546-554.

[9] Viola, P. and Jones, M. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR 2001*, I-511-I-518.