Augmenting Communication in a Shared Mixed-Reality Space — Attention Sharing and Eye Contact Recovery

Yuichi Nakamura†, Itaru Kitahara†, Jun'ichi Hoshino†, Katsuhiko Sakaue‡, and Yuichi Ohta†
† IEMS, University of Tsukuba, Tsukuba, 305-8573 JAPAN
‡ National Institute of Advanced Industrial Science and Technology
({yuichi, ohta}@image.esys.tsukuba.ac.jp)

Abstract

A shared mixed-reality space is a space where multiple people share the perception of mixed reality. The people communicate one another in looking and feeling the events in real world augmented by virtual objects. To realize natural and effective communication in this space, we are investigating several techniques which overcome the drawbacks of mixed reality space and of real space. In this paper, we describe the basic idea, fundamental techniques, and their integration.

1 Introduction

Mixed Reality (MR) is integrated technology that merges real world and virtual world. Many applications are proposed based on this framework, some of which realize reality augmentation with virtual objects or views, others realize on-demand information service at anytime and anyplace.

As a natural extension, there are considerable demands for a Shared Mixed Reality Space (hereafter, abbreviated by SMRS) where two or more people share the sense of mixed reality. By allowing two or more people joining the same MR space and by supporting their mutual communication, we can consider a large number of applications, and those will potentially be communication styles of next generation.

Regardless of its importance, we have certain difficulties in supporting convenient communication in SMRS. The devices required for MR, *e.g.* HMD (Head Mount Display), often interfere communication among people, and narrow communication channels usually make sharing awareness difficult. The aim of our research is to improve and augment such communications by restoring the lost information and by managing and editing available information.

In this paper, we introduce our approaches to this problem: restoration of eye-contact by face image synthesis, pointing enhancement by visualizing gazing direction (line of sight) and pointing direction, intelligent video capturing and switching for sharing attention, and annotation capturing for sharing essential information.

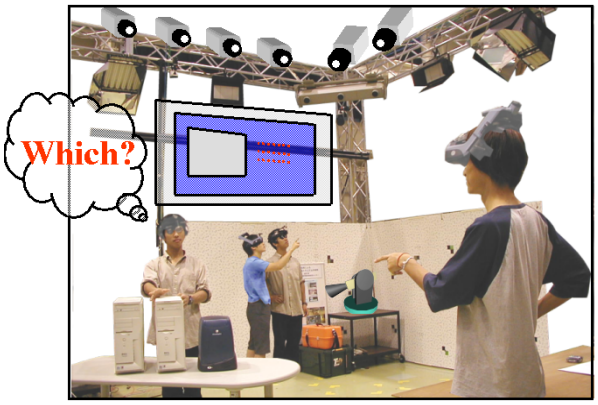


Figure 1. Typical situation (scene)

Table 1. Typical situation (scenario)

(1)Taro: "Hi, Jiro."
(2)Jiro: "Who is it?"
(3)Taro: "It's me."
[Conversation for a while]
(4)Taro: "That DVD drive doesn't work well. Why don't you replace it?"
(5)Jiro: "Which drive?"
(6)Taro: "On the left hand side. Can you see it?"
(7)Jiro: "OK. I got it. How can I replace it?"
(8)Taro: "First, pull that lever... Next, screw the..."

Table 2. Typical situation (scenario cont'd.)

[After Taro went out. Jiro is trying to replace the drive]
(9)Jiro: "Hmm... I think I'm lost... What should I do?"
[When Jiro pointed the DVD drive, the systems playbacks Taro's explanation]
(10)Jiro: "OK. I got it! It's easy ..."

2 Communication in Shared Mixed Reality Space

We have various communication channels in the real world. Non-verbal behaviors, as well as verbal communication, have important roles in our ordinary communications. In SMRS environment, however, such communications are often blocked due to the special devices or narrow communication channels. It causes difficulties in sharing awareness or in directing attention to the right portion.

For better understanding, let us consider an example in Table 1. Suppose that two persons Taro and Jiro meet in a SMRS, and Taro begins to explain an important operation of a machine. In the beginning portion (1)–(3), Taro was not sure if Jiro noticed his presence. When Taro mentioned the DVD drive (4)–(8), he had difficulties in checking if Jiro was also looking at it. He may also have felt inconvenience in specifying the drive.

Thus, in SMRS, we need communication support for sharing attentions. For this purpose, we developed the following mechanism as shown in Figure 2.

Eye-contact restoration: Lost eye-contact is recovered by overlaying the eye images onto the user's facial view. Restored eye-contact works as a good index for the focus of attention.

Pointing augmentation: Pointing line visualization and pointed object emphasis help the users mutually pay attention to the right portion.

Intelligent video capturing and view selection: Simultaneous video capturing for different targets and giving appropriate views are good supports for paying attention to the most relevant portion.

Annotation recording and on-demand playback: By recording the scene and playing back on-demand, communication can be enhanced beyond limitations of time. We can easily think of the situation (9)-(10) in Table 2.

Thus our research augments communication in SMRS by restoring views, overlaying images, switching views, and playing back the recorded movies. We will briefly introduce the above functions in the following sections.

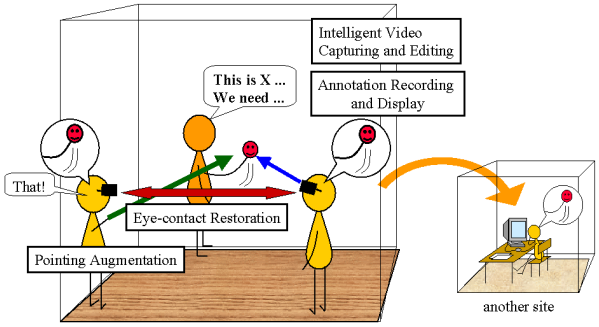


Figure 2. Communication augmentation in Shared Mixed Reality Space

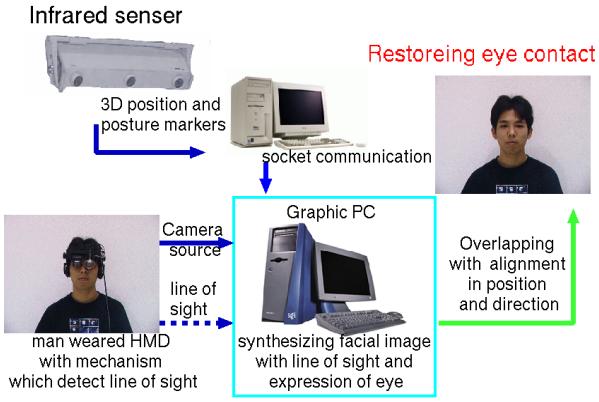


Figure 3. Overview of eye-contract restoration mechanism

3 Support for Nonverbal Communication and Sharing Attention

3.1 Eye Contact Support

As shown in Figure 1, each user's face is partially hidden by an HMD, which blocks eye-contact. Although HMD is getting smaller, it is almost impossible to make an HMD completely transparent.

To cope with this problem, we propose a new technique for restoring eye-contact and gaze awareness by synthesizing and overlaying the expected facial image. Figure 3 shows the overview of our system. The 3D appearance of user's face is recovered using the computer vision technique, and is synchronized with the user's current head motion. The eye-appearance is also synthesized based on the eye movements measured by our eye-tracking HMD.

3.1.1 Offline Processing (Model Acquisition)

The model acquisition process has three steps: 3D geometric model acquisition, facial texture acquisition, and acquisition of detailed texture around eyes.

3D geometric model

We reconstruct the user's 3D facial model by using factorization method, that is a computer vision technique. First, we reconstruct a facial 3D shape from three input facial images for which the correspondences of the reference points are known.

In this method, rotation matrix, a geometric 3D model and translation components are calculated by using positions of reference points in three images. Let the matrix which has 2D positions of reference points be W. The position in W is normalized by eliminating translation component. Facial 3D model can be estimated using the following singular value decomposition technique[2].

$$W = RS \tag{1}$$

where the matrix R is the relative position of the camera, and S is the facial 3D model.

Expanded texture

We use the expanded texture, which is a generic texture in which textures from three images are blended. For any orientations of the face, the same generic texture is used for the same portion. This contributes to reduce the computational time, since we can beforehand calculate the blending, which is the most time consuming portion.

Textures are registered based on triangular patches of which vertices are the above reference points. The textures are expanded according to their actual size in the 3D model. For this purpose, we measure the distance of each reference points on 3D model from the central reference point on the nose. Then, we use the 3D distance for placing the vertices of a patch.

Figure 4 shows the outline. The left side of Figure 4 shows the geometric model. The right side of Figure 4 shows the cross section of this plane. We measure the distance from the central reference point to the other reference points on this cross section. The expanded texture are generated by blending multiple input images with these reference points.

Figure 5 shows an example of expanded textures. The local deformation of a texture becomes large as the distance from the central reference points becomes large. However, we do not lose texture information because this deformation usually makes the texture patch larger than the original one.

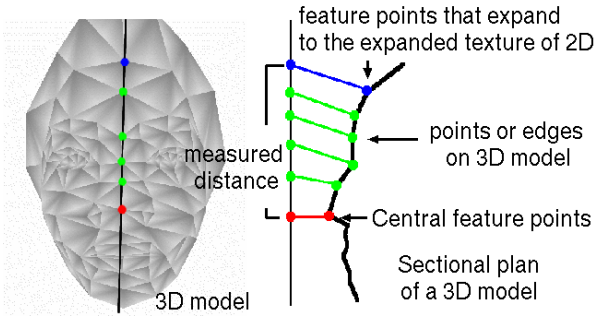


Figure 4. Expanded texture generation

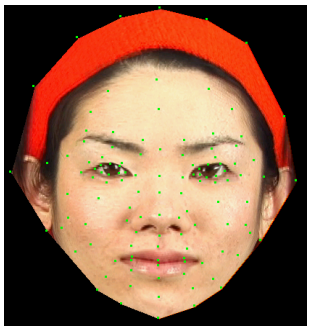


Figure 5. Acquired expanded texture

Texture around eyes

Texture changes on an eyelid are complex, though the eyelid motion is comparatively simple. To simulate eye movements and blinking, we need a different approach for the textures around eyes.

We get eyelid texture at arbitrary state by using 3D models generated from two sets of facial images with open/close eyes.

For textures inside eyes, we approximate 3D rotation of an eyeball by 2D texture changes. To obtain textures, we cut out eyeball texture from facial images as shown in Figure 6. By registering the partial eyeball texture for various gazing directions, we obtain the whole eyeball texture. An example of generated eyeball texture is shown on the right side of Figure 6.

Finally, as shown at the bottom of Figure 7, we obtain the texture for arbitrary facial orientation and gazing directions.

3.1.2 Online Processing (Realtime Rendering)

Face orientation is measured by the above mentioned infrared sensor. A corresponding facial image is rendered through 3D geometric model rotation and facial texture mapping. In addition, gazing direction is measured by the sensing device attached to an HMD. Using

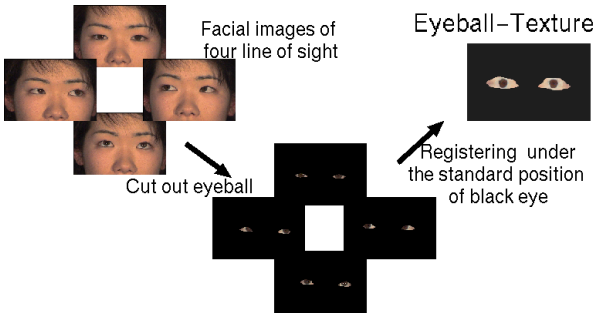
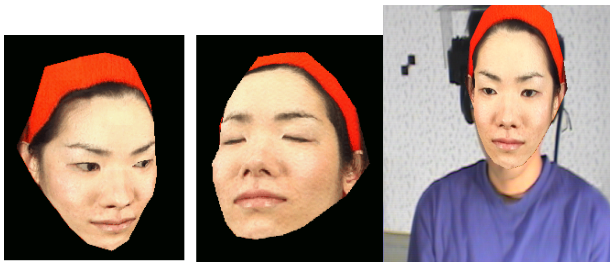


Figure 6. Eyeball texture acquisition



(a) generated facial images (b) overlaid on the user's face

Figure 8. Overlaying facial image

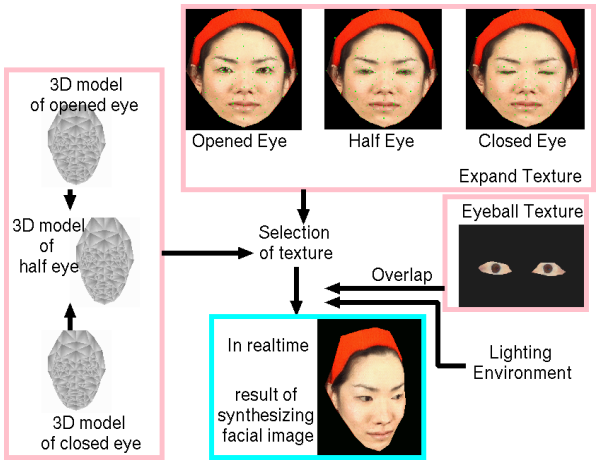


Figure 7. Realtime facial image rendering

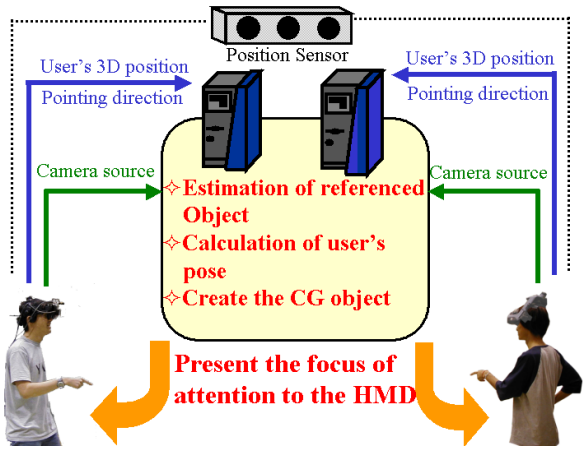


Figure 9. Overview of pointing augmentation

the direction, pre-stored eyeball and eyelid textures, eye expressions are generated in real time. Then, the synthesized facial image is overlaid at the current position of the user's head.

We have developed a prototype system on an ordinary PC to demonstrate the effectiveness of the proposed method. As in Figure 3, 3D position, which is measured in the sensor coordinate system, is sent to the PC via Ethernet. Also, the user's gazing direction is measured in real-time, and sent to the PC. Then, we get the results as shown in Figure 7 and 8.

3.2 Pointing Augmentation

Figure 9 shows the overview of our pointing augmentation system. The location of a referenced object is estimated by the intersection of the user's gazing line and the pointing line. The 3D position of the user and the arm orientation are obtained by using a position sensor. The gazing direction is approximated by the optical axis of the camera attached to the HMD. Then, we estimate the intersecting point by assuming a virtual plane that includes the user's gazing line and by calculating the intersection with the pointing line as shown in Figure 10.

To realize this mechanism, we need to use several positions and orientations measured by different sensors. To deal with their conversion, we use the following four coordinate systems.

World coordinate system W (Xw, Yw, Zw): The 3D coordinate system of the SMRS.

Sensor coordinate system S (Xs, Ys, Zs): The 3D coordinate system of the infrared position sensor.

User coordinate system U (Xu, Yu, Zu): The 3D coordinate system defined on the HMD.

Camera coordinate system C (Xc, Yc, Zc): The 3D coordinate system of the camera attached to the HMD.

Figure 11 illustrates the relationship among these four coordinate systems. We need to consider the following transformation between the coordinate systems.

1. The transformation between the sensor and world coordinate systems (M_{s2w}). This transformation is used to represent the positions of markers attached

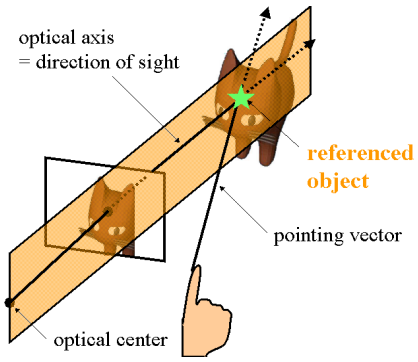


Figure 10. Intersecting point

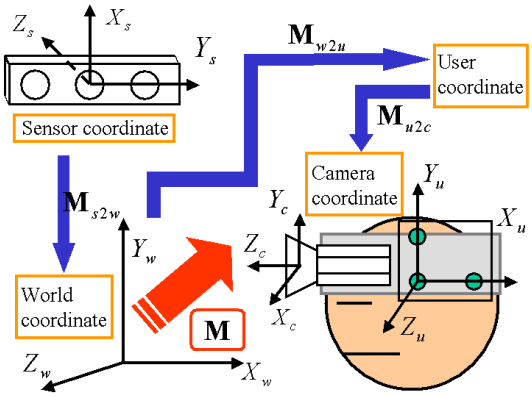
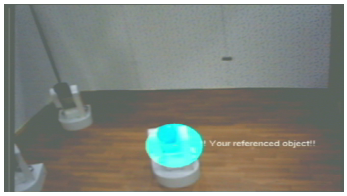


Figure 11. Coordinate systems

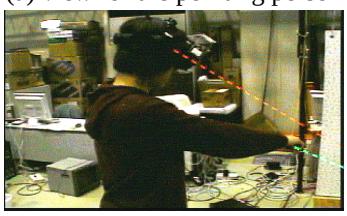
to the user's arm and HMD in the world coordinate system.

2. The transformation between the world and user coordinate systems (M_{w2u}). This transformation represents the position and orientation of the HMD in the world coordinate system. This can be calculated using three marker positions attached to the HMD[7].
3. The transformation between the user and camera coordinate systems (M_{u2c}). This transformation represents the pose of camera attached to the HMD. This can be calculated by multiplying inverse transformation of M_{w2u} and the initial M_{w2c}. The initial M_{w2c} represents the initial position and orientation of HMD. This can be calculated through a camera calibration process.
4. Finally the position and pose of an HMD, M in Figure 11, is obtained by multiplying these transformations.

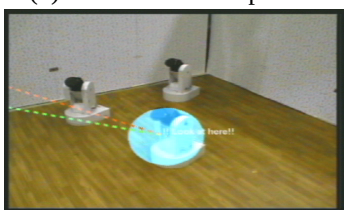
An example is shown in Figure 12. The estimated position of the target is highlighted. Figure 12(a) is the view displayed for the pointing person, (b) and (c) shows the views displayed for another person. Thus the



(a) view for the pointing person



(b) view for another person



(c) view for another person

Figure 12. Example of pointing augmentation

attention of two or more people are directed to the right portion.

4 Communication Enhancement by Intelligent Video Capturing and Annotation

4.1 Intelligent Video Capturing

Suppose a situation that a person is demonstrating the usage of a complicated machine, as shown in Figure 13. When the speaker explains an important device by holding out toward us, we need to carefully look at it. When we want to telecommunicate this scene, we usually prefer a close-up shot (Figure 13(a)) or an extreme close-up shot (Figure 13(b)).

To automatically obtain these shots and present them, we need to computerize the following functions:

camera control: Shooting and tracking of important portions with appropriate cameraworks.

focus recognition: Recognition of the events occurring in a SMRS, and detecting the focus of attention.

selection and emphasis: Selection of the best views and emphasizing important portions.

We are developing a system which realizes the above functions. Figure 14 shows an overview of our system. Multiple pan-tilt cameras shoot at important por-

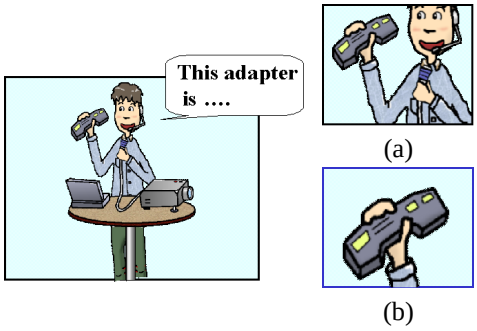


Figure 13. Typical explanation behavior

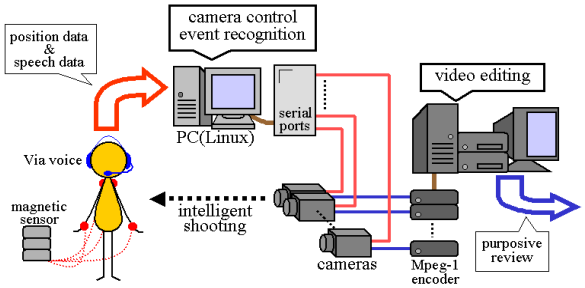


Figure 14. Overview of the intelligent video capturing system

tions. Each camera is assigned a unique target and an objective for shooting. Videos taken by those cameras are transmitted, switched, or recorded in MPEG format. This framework is essential for effective communication, since important portions are often scattered in a scene, and some of them such as hands or important objects often moves arbitrary. For camera control and behavior recognition, we have magnetic sensors for detecting the speaker's position.

This framework also includes event recognition whose output is used for video switching and editing. For this purpose, the system has a speech recognition module and the above mentioned positional sensors. Integration of speech and movements recognition is the key technique to realize automated switching or editing for giving comprehensible videos. According to the recognition results, the system emphasizes the focus of attention by switching the views or choosing the relevant portions. Thus, the system gives views that a speaker wants to show or that viewers want to see.

Figure 15 shows an example, in which an object is held out by a speaker. One camera always tracked the right hand of the speaker, and when he held out an object as shown in Figure 15(a), the system switched the video to the close-up views as shown in Figure 15(b). Thus, the focused object drew the audience's attention.

Here we will briefly describe two important points of

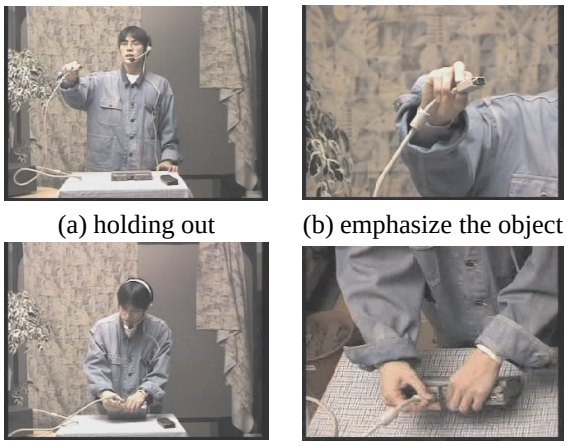


Figure 15. Example of video capturing and view selection

this research.

Camerawork

First, we proposed adaptive camera control based on what *target (subject)* we want to shoot, and what *aspect-of-target* we want to capture. Basically, a target is the object to be tracked by a camera, and the above aspect-of-target determines how to track it.

We currently consider four kinds of targets.

<speaker> a speaker, a lecturer, or an instructor.

<workspace> a *dynamic* space where a manipulation such as assembling or cooking is going on.

<object> an important object to be paid attention.

<place> an important *static* place to be paid attention.

Next, we categorized aspect-of-target considering the focus of attention:

<circumstance> Target's circumstance that includes position, trajectory, or spatial relationship to other objects. This is suitable for giving the overview of a presentation or manipulation with a wide-angled view.

<movement> Target movements that may include frequent small motions such as hand motions in manipulations.

<appearance> Target's appearance, such as shape, or color.

Basically, every camera is always controlled to shoot at its target. Rapid and frequent motion, however, causes shaky and irritating view changes, which we need to avoid as much as possible. For this purpose, we propose (a) camera motion smoothing by the Kalman filter and

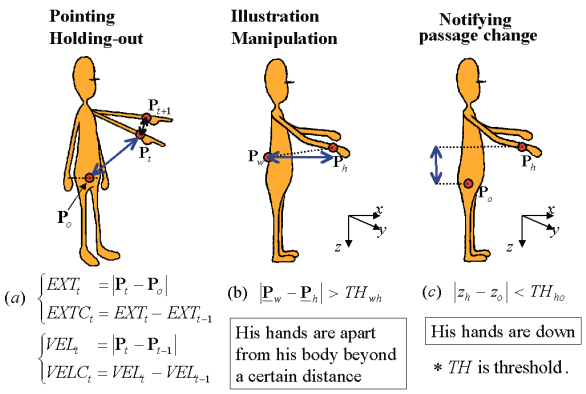


Figure 16. Motion and posture detection

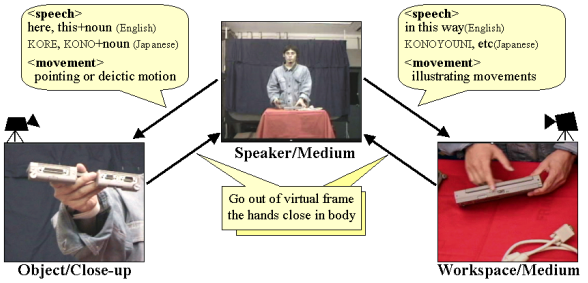


Figure 17. Condition of view switching

(b) camera motion suppression by the *virtual-frame control*. We can adapt a camerawork for various purposes by tuning the parameters of those methods. Details are described in [17].

Behavior recognition

The second important point is behavior recognition, which detects the focus of attention. We are currently dealing with five kinds of behaviors. To detect those behaviors, we need multimodal processing.

Table 3 shows the speech clues that we are currently using, and shows the focus suggested by the clues. The first column gives the words, the second gives the behaviors, and the third gives the focus of attention. In each pair of rows, the upper row shows Japanese expressions, and the lower row shows English expressions. Since this system is designed for Japanese, the words in the upper row are actually the targets of speech recognition.

Speaker's motions or postures, *e.g.* hand position or velocity, are also the most useful clues. The following features are simple and do not need any sophisticated analysis, and efficient realtime processing is possible.

- local maxima of arm stretch
- local minima of pseudo velocity change of a hand

If the arm-stretch-change at a local maxima is larger than the threshold value, the first condition is satisfied. For detecting that both hands are on/above the desk, the distance between body and a hand is calculated.

The condition for detecting each behavior is as follows:

pointing/holding-out: Table 3(a), (b) and Figure 16(a)

manipulation/illustration: Table 3(c) and Figure 16(b)

notifying passage changes: Table 3(d) and Figure 16(c)

If the system detects both speech clues and motion clues within a certain period, the system accepts the corresponding behavior. We previously investigated the occurrence time difference between speech and motion[18]. The statistics showed that, in around 90% cases, speech clues and motion clues occurs within 2 seconds from each other. This duration is enough for the condition on offline processing. For online processing, however, the speech recognition sometimes has a delay longer than 2 seconds. We set the duration to 3 seconds for online processing.

Video switching

By selecting the most relevant view according to the events, we can obtain a comprehensible video as shown in Figure 18. This selection is fully automated by using an electronic switcher controlled by a host computer.

The switching condition is briefly shown in Figure 17. When holding-out or pointing behaviors are detected, the system selects the view through camera2 (shooting at <object>). Similarly, in the case of manipulation or illustration behavior, the system selects the view through camera3 (shooting at <workspace>). Comparing the raw video in Figure 19 with the edited video in Figure 18, we can easily understand that the system selects appropriate views and the result is quite satisfactory.

4.2 Annotation by Video Clips

Videos captured by the above system can be useful information sources for the people in a SMRS. If we can see the recorded explanation as in the example in Table 2, our communication is augmented beyond limitations of time and place.

Automatic object recognition and tracking is necessary for realizing this mechanism. Objects are usually the keys in manipulation or instruction. For example, parts are most important when we assemble a machine, and they are often the focus of our attention.

Table 3. Typical examples of speech, behaviors, and focus

	Typical Words (upper row: Japanese, lower row: English)	Behavior	Focus	Example
(a)	KORE, KONO (+ object name), etc.	pointing, holding-out	object	KONO NEJIMAWASHI WO TSUKAIMASU
	(this + object name), this, etc.			use this screw driver
(b)	KOKO, KONO (+ place name), etc.	pointing	place, location	KOKO NI OKIMASU
	(this + location name), here, etc.			put it here
(c)	KONOYOUNI, KOUSHITE, etc.	manipulation, illustration	manipulation, hand motion / locus	KONOYOUNI NEJI WO MAWASHIMASU
	in this way, like this, etc.			drive the screw in this way
(d)	KOKODE, SOREDEHA, TSUGINI, etc.	notifying passage changes	speaker	SOREDEHA TSUGI NO STEP WO HAJIMEMASU
	So, OK, Next, etc.			OK, go on to the next step

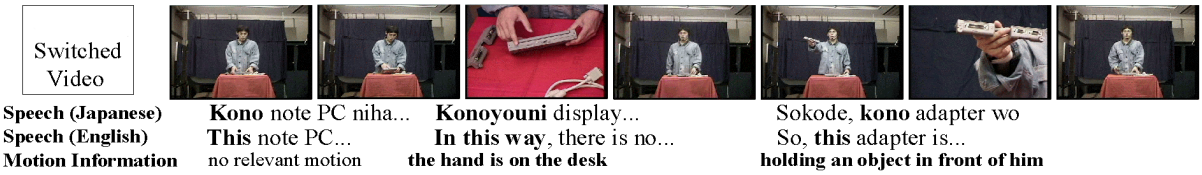


Figure 18. Result of view switching

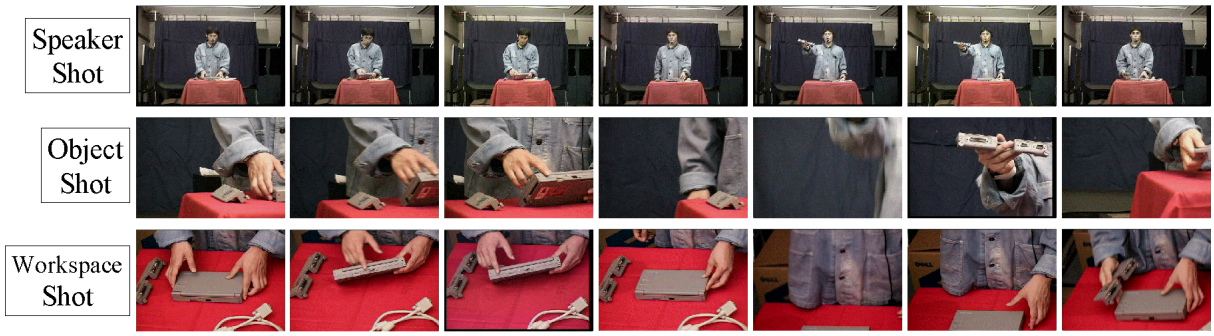


Figure 19. Videos from three cameras

For this purpose, we consider object tracking on the following conditions:

- There is no prior knowledge about object's size, color, texture, and so on.
- The background may change at any time during manipulation.

On the other hand, we can naturally assume the following restrictions.

- Most of the important objects are moved or manipulated by human hands.
- The space (volume) in which important objects potentially appear is known.

Even with the above two restrictions, the above conditions are still severe. Object rotation or occlusion caused by grasping can easily alter the object's texture, and moving people in the background add serious noise that cannot be easily eliminated.

To cope with this problem, we proposed a novel method for tracking objects which appear in manipulations. For the robust object tracking under few constraints, we use multiple image sensors, that is, an RGB camera, a stereo camera[15], and an IR (infrared) cam-

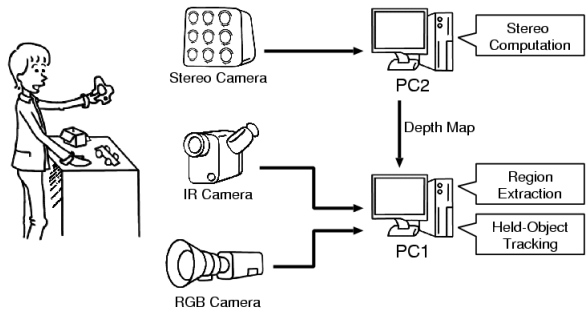


Figure 20. Region detection and integration

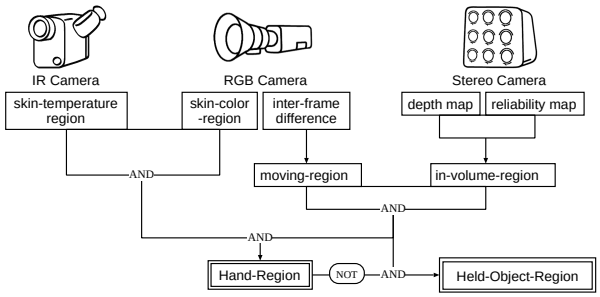


Figure 21. Flow of region detection and integration

era. With this method, our system tracks a hand-held object at video-rate even there are other moving objects or people with skin color regions in the background.

The overview of our system is shown in Figure 20. The system detects the following regions from three sensors.

RGB Camera: *skin-color-region* and *moving-region*.

Infrared Camera: *skin-temperature-region*, that is a region with the intensity corresponding the skin temperature, *i.e.* around 34°C.

Stereo Camera: *in-volume-region*, that is a region in the volume in which hands and related objects appear.

By integrating the above regions, *hand-region* and *held-object-region* are detected based on the following idea.

$$\text{hand-region} = \text{in-volume-region} \wedge \text{moving-region} \wedge \text{skin-temperature-region} \wedge \text{skin-color-region} \quad (2)$$

$$\text{held-object-region} = \text{in-volume-region} \wedge \text{moving-region} \wedge \neg\text{hand-region} \quad (3)$$

Each image sensor

For detecting skin-color-region, we created a skin color model by gathering the statistics of pixels corresponding to skin regions, and determined the parameters of the distribution.

For skin-temperature-region, we examined the pixel values in real hand region and those in typical background, and determined the threshold for extracting skin-temperature-region. Our IR camera captures infrared light with the wavelength between 7 and 14μm, which covers the dominant wavelength that a human body emits.

For in-volume-region, we assumed that the width, height, and depth of the workspace are known. These can be changed according to the spatial arrangement of the workspace and the camera position. Objects in this volume can be detected by using the depth map obtained by the stereo camera.

Integration for multiple image sensors

Prior to the actual region extraction and tracking, we need geometric compensation and synchronization of the images from three cameras. For geometric compensation, we use a quadratic model. Although the IR camera has heavy radial distortion, 25 reference points are enough to calculate the parameters. To compensate the latency of stereo computation and transmission time, each image is attached its captured time. The depth map image captured at the nearest time is used with the other two images.

As shown in Figure 21, hand-region is detected by taking logical AND operation of the four regions as shown in equation 2. The extracted hand region candidates are labeled after region expansion-contraction. Then, at most two regions whose area are larger than the threshold are registered as hand-region.

Through the position smoothing by the Kalman filter, the final estimated position of the object is determined. By repeating the above process at video-rate, the estimated position of a held object is obtained at every frame. The detected regions are shown in Figure 22, and examples of tracking result are shown in Figure 23. As we can see in these figures, the held object is well detected and tracked even when the output of each sensor has much noise.

As shown in Figure 23, we evaluated our system in two situations. Scene A is a simple scene in which one person is holding and moving an object. Scene B is a more complicated scene with multiple objects on the worktable and with another person walking behind.

Regions were correctly detected for 97% and 93% of frames in scene A and scene B, respectively. For scene B, tracking was a little more failed than in scene A, since the box with skin color and the walking person make misleading regions. However, the rate of tracking failure is still less than 6%, which is difficult to achieve by a

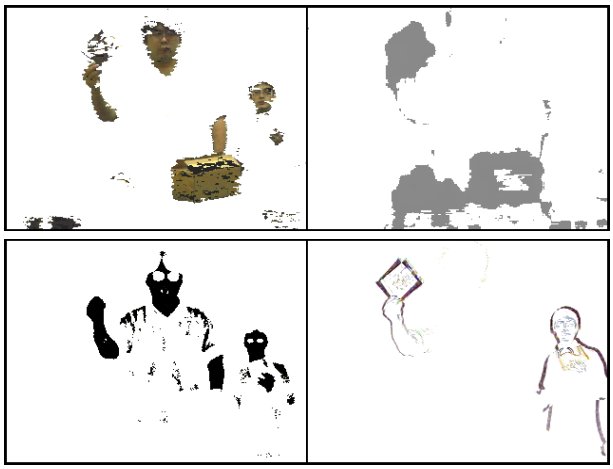


Figure 22. Detected regions (upper-left: skin-color-region, lower-left: skin-temperature-region, upper-right: in-volume-region, lower-right: moving-region)

Table 4. Detection and tracking performance

	#Total	#Detection failure	#Tracking failure
Scene A	1350 frames	30 (2.2%)	4 (0.3%)
Scene B	1350 frames	11 (0.8%)	80 (5.9%)

single image sensor.

Actual application example is presented in the next section.

5 Experimental System at University of Tsukuba

We are developing an integrated system at University of Tsukuba. The above four functions are implemented in two rooms. One room mainly has video capturing and annotation recording systems, and the other has eye-contact restoration and pointing augmentation systems. These two sites are connected by Gigabit Ethernet, which can transmit several MPEG2 streams in both directions. The overview of our system is shown in Figure 24.

In the followings, we briefly introduce possible scenario on our system. It is composed of two parts: Scene 1 is prepared for demonstrating intelligent video capturing and annotation recording, and Scene 2 is prepared to demonstrate pointing augmentation in SMRS. For these scenario, our experimental results are shown in the followings.

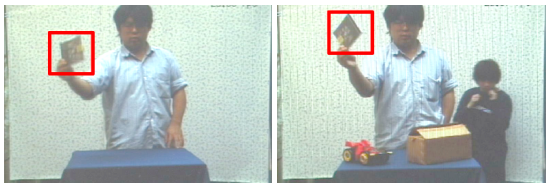


Figure 23. Scene A (left) and Scene B (right)

Scene 1: Intelligent video annotation capturing

By combining object tracking with the intelligent video capturing system, we can get movie clips that are directly linked to the objects in a SMRS. This scenario is on the conversation between a guest and a waiter in a restaurant.

Figure 25 shows the several images of the scene. A waiter is giving an explanation of the dish to the guest. First, when he held the dish, the system detected it, and the rectangle with the dotted lines shows the location. When he gave the explanation of the dish by speaking "This dish is ....", the system recognized the behavior, and registered his annotation as the information on dishes. This step is noticed by the red thick lines overlaid on the dotted lines. When he put the object on the table, the texture and the position of the object were registered, and the captured annotation was linked to the object region.

Scene 2: Pointing augmentation

By pointing augmentation, the system supports conversation among two or more people. Figure 26 illustrates the system configuration used in this demonstration.

This scenario is on two users' talking about exhibited food samples. Two people came to a restaurant, and they began to talk about the exhibited food samples as shown in Figure 27. When one user pointed out a food, it was highlighted, and at the same time an annotation movie tagged to this food was presented to the partner's HMD (Figure 28). The annotation movie was stored as the movie database created by the above system. Thus the users can communicate with fully understanding his partner's attention. Two users can also share the attention by presenting the augmented pointing even if the users are temporally and/or spatially apart.

6 Conclusion

In this paper, we introduced the basic idea of SMRS. To support convenient communications in SMRS, we

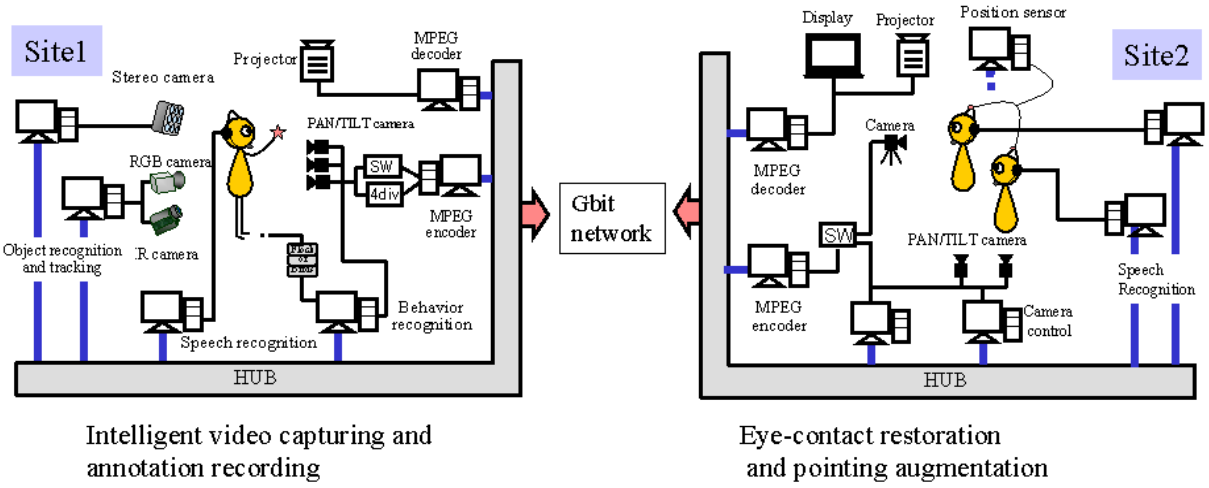


Figure 24. System overview

proposed eye-contact restoration, pointing augmentation, intelligent video capturing, and annotation recording for sharing essential information. Through some experiments as shown above, we verified that our method effectively supports for keeping awareness and sharing attentions.

Although elementary functions are working, we are still integrating them into a system. In the near future, it will be a really human-oriented cyberspace that supports and augments our ordinary and natural communications.

Acknowledgements

This work is supported in part by the Ministry of Education, Culture, Sports, Science and Technology under the Grant-in-Aid for Scientific Research (11230202).

The authors thank M. Ozeki, Y. Sato, M. Itoh, M. Takemura, R. Ogata, and T. Takemasa for their contribution. Our experimental system is developed based on their programming and system integration.

References

[1] J. Tang, Findings from observational studies of collaborative work, Int. J. Man-Machine Studies 34(2), 143-160.

[2] Y. Mukaigawa, et al., Face Synthesis with Arbitrary Pose and Expression from Several Image — An integration of Image-based and Model-based Approach—, ACCV98

[3] Y. Sato, et al., Attention Sharing in MR Community Space -Enhanced Visualization System of User's Indication (in Japanese), the 6th VRSJ 2001

[4] T. Ohwa, Y. Ohta, Image-Based Face Synthesis and Video Rewriting (In Japanese), MIRU 2000

[5] D.Piponi, G.Borshukov, Seamless Texture Mapping of Subdivision Surface by Model Pelting and Texture Blending, SIGGRAPH 2000 Conference Proceedings

[6] P.Debevec, et al., Acquiring the Reflectance Field of a Human Face, SIGGRAPH 2000 Conference Proceedings

[7] M.Kanbara, et al., A Stereoscopic Video See-though Augmented Reality System Based on Real-time Vision-Based Registration, IEEE Virtual Reality 2000 International Conference, 2000

[8] M.Billinghurst, et al., Shared Space, collaborative information spaces. HCI International, 1997

[9] Z.Szalavari, et al., Augmented reality enabled collaborative work in studiertube, EURO-VR, 1997

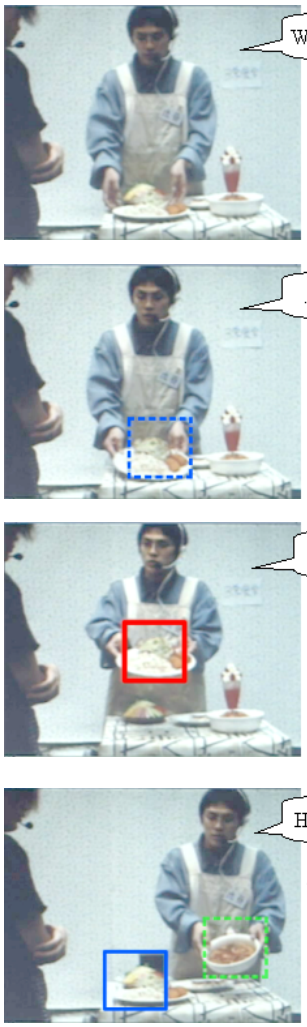
[10] M.Takemura, et al., Restoration of eye-contact in mixed community space by human image processing (in Japanese). the 6th VRSJ 2001

[11] M.Ozeki, Y.Nakamura, Y.Ohta, Sharing Attention in Mixed Community Space - Recording Annotation for Objects by Recognizing Human Behaviors (in Japanese), the 6th VRSJ 2001

[12] Y.Kameda, et al., A live video imaging method for capturing presentation information in distance learning. Proc. IEEE ICME, 2000.

[13] S. Mukhopadhyay and B. Smith, Passive capture and structuring of lectures, Proc.ACM Multimedia, 1999

[14] N.Ohno and K.Ikeda, Video stream selection according to lecture context on remote lecture (In Japanese), Proc. 5th Intelligent Information Me-



(a) Left image shows the result of held-object tracking. The right person is giving an explanation of dish for the guest.

(b) When he held the dish, the system detected it, and the rectangle with the dotted lines shows the location.

(c) When he gave the explanation of the dish, the system recognized the situation and registered his annotation as the dish information.

(d) When he put the object on the table, the texture and the position of the object were registered, and captured annotation was linked to the object region. Then, the system detected new held-object.

Figure 25. Example of combining object tracking with the intelligent video capturing system

dia, 1999

[15] http://www7.airnet.ne.jp/komatsu/stereo/stereoj/page2.htm (In Japanese)

[16] M. Ozeki, Y. Nakamura, and Y. Ohta, An intelligent system for recording presentations (In Japanese), Proc. 6th Intelligent Information Media, 2000

[17] M. Ozeki, Y. Nakamura, and Y. Ohta, Camerawork for intelligent video production, Proc. ICME, 2001

[18] Y.Nakamura, et al., MMID: Multimodal multi-view integrated database for human behavior understanding, Proc. IEEE International Conference on Automatic Face and Gesture Recognition, 1998

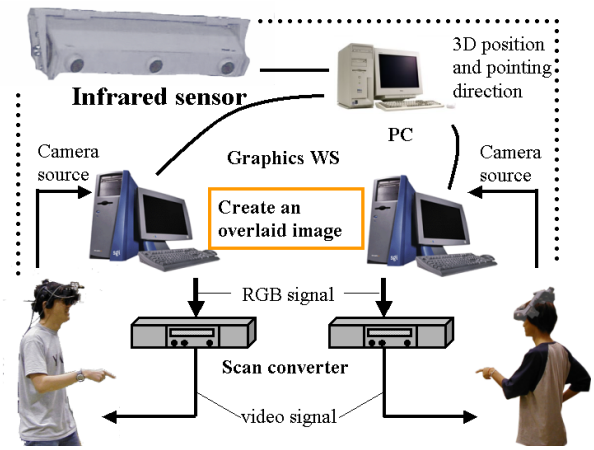


Figure 26. Configuration of the demonstration system

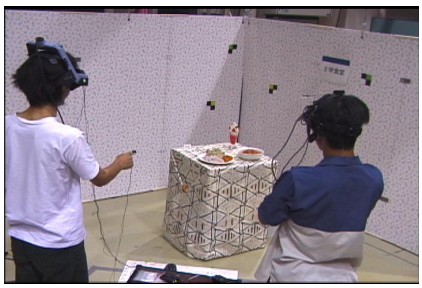
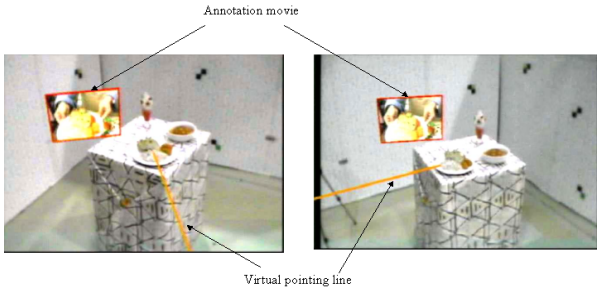


Figure 27. Situation of pointing augmenting



(a) view for the left user (b) view for the right user

Figure 28. Pointing augmentation with the pre-recorded video clips