ＡＴＲテクニカルレポート表紙

〔公　開〕

TR－C－0141

Evaluation of Emotion Enhanced Face to Face Meetings Which Uses the Concept of Virtual Space Teleconferencing

リヤナゲ　デ　シルバ　　宮里　勉　　岸野　文郎
Liyanage C. DE SILVA　Tsutomu MIYASATO　Fumio KISHINO

１９９６　３．１５

ＡＴＲ通信システム研究所

Evaluation of Emotion Enhanced Face to Face Meetings Which Uses the Concept of Virtual Space Teleconferencing

Liyanage C. DE SILVA and Tsutomu MIYASATO and Fumio KISHINO

ATR Communication Systems Research Laboratories, Kyoto-fu, 619-02 Japan.

Contents

List of Figures

List of Tables

Abstract

本稿では、臨場感通信システムの革新性について検討する。特に感情認識と伝達について述べ、その観点での本システムの優位点を評価実験で確認する。すなわち、送信側と受信側の間に設置された、知識による感情認識と感情強調によって、より効率的な会話が実現でき、これによりface-to-face会話の欠点になっている、異文化間での感情認識の相違を乗り越えたテレビ会議を実現することが可能であることを示す。さらに、従来テレビ放送で使われているモザイクを用いたプライバシー保護を行なった場合でも話者の感情伝達を可能にする、仮想人物についても述べる。

Abstract

Here we investigate the unique advantages of our proposed Virtual Space Teleconferencing System (VST) in the area of multimedia teleconferencing, with emphasis to facial emotion transmission and recognition. Specially we show that, using this concept, emotions of a local participant can be transmitted to the remote party with higher recognition rate by enhancing the emotions using some intelligence processing in between the local and the remote participants. This leads to a kind of emotion enhanced teleconferencing system which can supersede face to face meetings, by effectively alleviating the barriers in recognizing emotions between different nations. Also in this paper we state about a concept known as a *virtual person*, which is a better alternative to blurred or mosaiced facial images that one can find in some television interviews with people who are not willing to be exposed in public. Finally we compare the amount of data rate required for the proposed method with two other available methods, and confirm that our approach needs a very low data rate compared to those methods.

1 Introduction

We are heading to an era in which the multimedia communications dominated by the Internet. Although mediums capable of heavy traffic handling are developed, Internet users are to be satisfied by slow traffic mediums at some state of their communication due to some bottlenecks of communication. This is due to inherent difficulties in switching from an existing communication medium to a new medium. In contrary to this, most Internet users are finding their ways to get access to high power machines, with the advent of new hardware devices. Hence methods that require low data rate are still very important, although they require high power end equipment.

Also, in the field of inter-personal communication, nonverbal communication plays a very important role compared to verbal communication. According to past research it has been shown that 55% of the information content of a communication is being occupied by non-verbal information as opposed to less than 7% of information occupied by verbal information [1]. Wickens et. al. [2] point out the importance of nonverbal cues such as finger pointing, gesturing and facial expressions. In one study, Chapanis et. al. [3] compared problem-solving performance between verbal and nonverbal modes and concluded that a face-to-face configuration led to a 14% reduction in time required for a pair of subjects to solve geographical location and equipment assembly problems.

The concept of Virtual Space Teleconferencing [4, 5] is a solid answer to both of the above problems, namely, low data rate and enhanced nonverbal commuincation between local and remote participants. In the paper by Slater et. al [6], the importance of virtual bodies in shared virtual environments is discussed. They stress the point of having well defined actors, whose actions resembles the human beings. In the paper by Thalmann et. al [7], it is stated that communicating with virtual humans should be equipped with the ability to "recognize" other virtual humans and "perceive" their facial expressions, gestures and postures. However, there is no need for real recognition or perception since information from the data structures that define the behavior of virtual humans can be passed directly to a second structure. Contrary to above systems, in our virtual space teleconferencing system [4, 5] (see Fig. 1), at each end

we have real humans, but in the virtual space we have their computer generated images. So to make it a live communication, we have to correctly transfer the information from the real world to the virtual world. In this transferring process, intelligent recognizers can enhance the effectiveness by correctly judging the facial expressions and reproducing a widely accepted facial expression either by texture mapping of a 3D facial image or by using Character Animation. We conducted experiments to identify these concepts. Some people do not specifically show their emotions in real life, but an intelligent recognizer which is trained to identify his or her expressions would be used in between to transfer expressions. In our experiment we asked a person to read different sentences with several different emotions such as anger, happiness, sadness etc.,. The evaluation was then done in a manner the recognition was totally free from the context of the sentence. We finally show that the concept of VST is an efficient and faster way of emotion transmission compared to presently available communication methods.

2 The concept of The VST

The Virtual Space Teleconferencing System is a system in which all the participants are graphically represented in a virtual space. The space is shared by the participants who are located at remote locations. The participants can communicate with each other as if they were in a single location since the faces of other participants are projected in front of them. The emotions displayed on the computer generated faces at the local end are the emotions displayed by the remote participants. The emotions are reconstructed in real time. In a given conversation the emotional and gestural inputs are used in the development and handling of virtual objects. The virtual object can be viewed by each participant in their own perspective and they can manipulate it using gloves or by free hand [8]. In the process of transmitting emotions, this system transmits motion of only selected control points of the face to remote ends.

Here, we propose the advantages of virtual person concept, the people who are displayed on the Virtual Space Teleconferencing system. As stated in the paper by Parke [9], a few applications for facial animation, such as visualizing the results of planned surgery, require realistic physically correct facial models. However, many animation applications only require that the faces be believable in passing the correct emotional behavior of the characters they represent. The concept that we are developing can easily be adopted to that kind of situations, as we measure the necessary parameters of facial expression change prior to sending it to the other remote end, which is not the case in conventional teleconferencing systems.

2.1 What is Virtual Person?

In virtual space teleconferencing system, computer generated images are used, instead of using actual human images. During the conference eye gaze, facial expressions, body movements etc., are directly re-produced on the CG image. This kind of systems would be in very high demand in the near future for interviews with people who do not wish to be identified in public, for example, when interviewing a criminal. In a TV program, a CG image with all expressions of the person, who does not like to be exposed in public, could be telecasted, instead of telecasting a mosaic facial image.

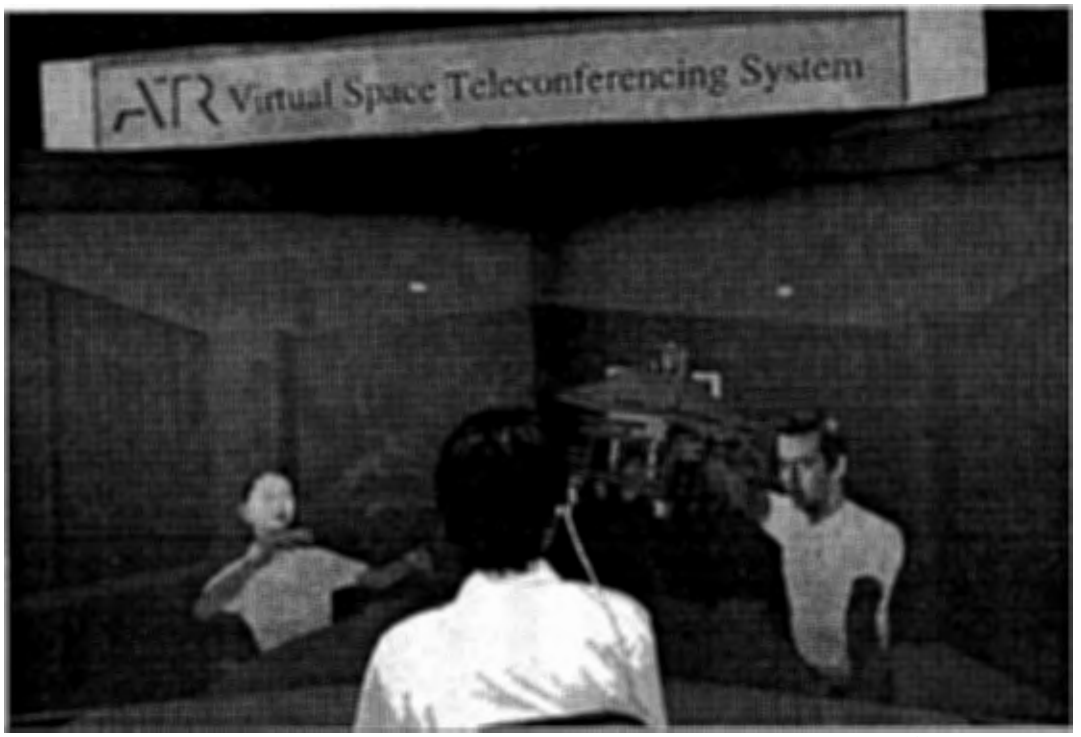


Figure 1: The Virtual Space Teleconferencing (VST) system

On the other hand we are in an era of multimedia. We are trying to make use of higher image quality as much as possible. Higher quality is a demanding request. But nowadays most of the time we see a lot of mosaic images during News Conferences. Why not try the idea proposed in the Virtual Space Teleconferencing be used in such situations? (see Fig. 2) Again if you are in a bad temper, instead of speaking with your real face, you may prefer to speak virtually. Also you may agree that most people do not like video phones. Why? They do not like to be exposed in person. They do not like to show what they are really doing. But, compared to a voice only communication if you can communicate visually the efficiency would be so much higher. Help of a *virtual person call* would make quite a big difference in these kind of situations.

2.2 Effect of Emotion Exaggeration on Context Understanding

It is a widely understood fact that human emotional changes are not clearly visible. But to understand a context more clearly, emotions are very important. For example news readers or nursery teachers usually change their emotions according to the context. But some people do not make much effort to change their emotions even though they really need to do it, specially during teleconferencing sessions, although they do it in day to day conversations. In teleconferencing people do not get an instant feed back, due to inherent problems in present day teleconferencing systems, such as loss of eye contact from the remote participant (In De Silva et. al. [10] a system which is free from loss of eye contact is proposed). Hence, they fail to show much emotional changes, although actually they like to show them. In such a situation, if the emotion can be detected by an external device and make necessary exaggerations before

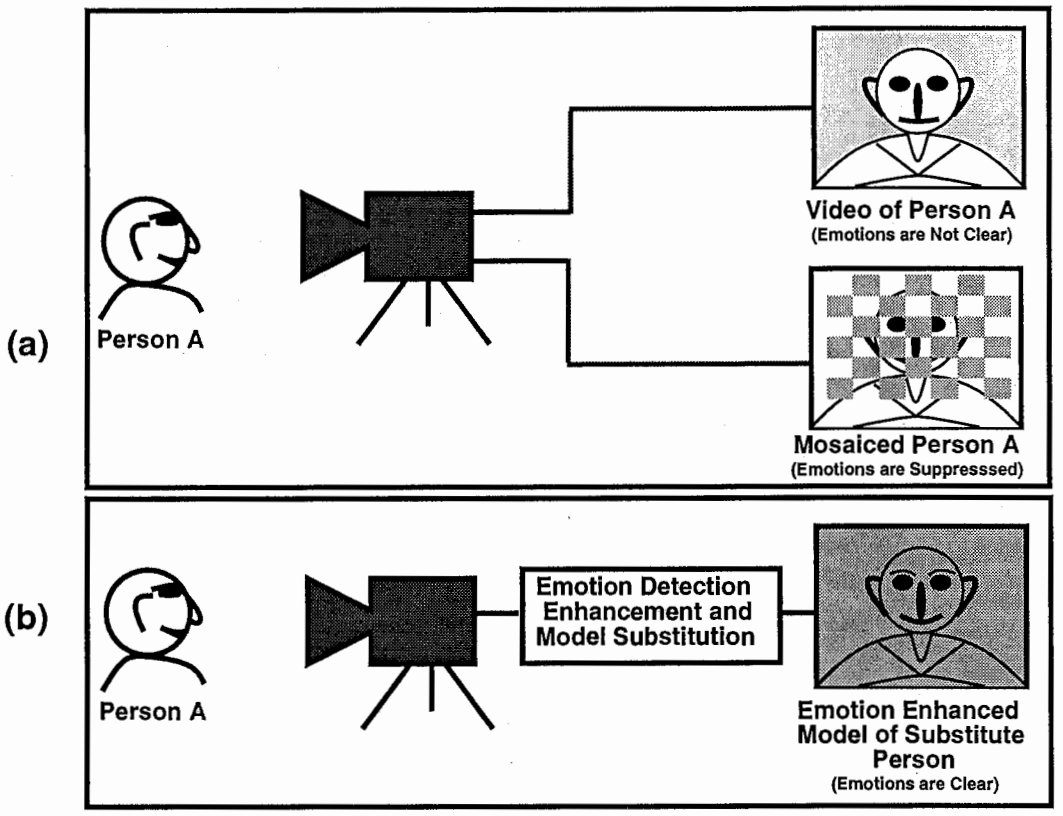


Figure 2: Substitution of a Virtual Person to Improve Visual Sensation (a) Conventional methods (b) Using the proposed concept

presenting it to the remote party then the impact would be quite impressive. For example, if one wants to scold one of his colleagues for a wrong thing, then he can use a kind of exaggerated emotion and input a stream of words recorded without much emotional changes.

There are quite a large number of on going researches in the field of emotion detection such as the one proposed by Ebihara et. al. [11] and the one proposed by Sakaguchi et. al. [12]. Once an emotion is detected, it can be transmitted and can be reproduced in several ways. As discussed earlier it can be directly mapped on to a real human like facial image, or it can be mapped on to a cartoon like character. As human beings are well trained to identify emotions of cartoon like faces, the speed of recognition of the emotions at the remote end would be very high.

3 Subjective Evaluation of The VST concept from the point of view of Emotion Detection

3.1 Evaluation of the Virtual Person Concept in VST

Here we evaluate the VP concept by generating some human emotions such as: anger, happiness, sadness, surprise and dislike. Here we deliberately omitted the fear emotion as it is rarely occur in face to face meetings. The evaluation was focused on identifying the advantages of

The Virtual Person concept used in the VST, subjected to some restricted conditions such as: audio only, mosaiced image, slow frame rate. (see Fig. 3).

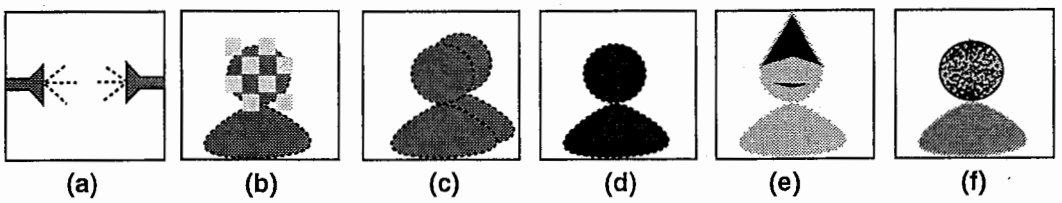


Figure 3: Evaluation of the necessity of visual clues in emotion recognition (a) audio only - AU (b) mosaic image - MO (c) slow frame rate - SL (d) video - VI (e) animation - AN (f) texture mapping - TM & 3D

Each of the above emotions were observed by several subjects and subjective evaluations were being carried out. Subjects were asked to recognize the emotion under the above restricted conditions and the percentage recognition was evaluated.

3.2 Procedure of the Experiment

In this experiment, first, a person (the remote person) was asked to instruct another person (the local person) to build a certain object in the virtual environment. Here we consider building of a portable shrine. Partly completed object was displayed on the screen. Then the following sentences were recorded. (Note that although the sentences given below are in English, the actual sentences read by the speaker were in Japanese.) During the instructions the emotional status of the face and the voice were asked to be changed according to the context of the sentence.

1. *(Anger) Mr. Tanaka, that roof is awful.*
2. *(Happiness) Mr. Tanaka, that roof is very good.*
3. *(Sadness) Mr. Tanaka, I have a big problem.*
4. *(Surprise) Mr. Tanaka, my goodness, what nice work.*
5. *(disgust, dislike) Mr. Tanaka, I don't like the roof of that shrine.*

3.3 Equipment and Conditions of the Experiment

The emotions were recorded in NTSC format. During the act the performer was asked to directly look at the camera. No complex scenes were used as the background. During recording, always the camera framing was done so that at least 60% of the screen area is covered by the performing person's head. In the evaluation two different voice tracks were used, Japanese and non-Japanese. The voice generated during the above act of making emotions were used in the narrations for the Japanese voice track. Voice needed for the non-Japanese voice track was recorded by a native Sinhala Language Speaker (see Section 4.2 for details).

The analyses of the emotion detection accuracy was carried out using various different kinds of communication methods. Altogether 7 different types of experiments were carried out per each emotion such as:

1. audio only (AU) - Here the video signal was totally suppressed.
2. mosaiced actual facial image sequence (MO) - Here each 16x16 pixel block was approximated by a single gray shade.
3. slow frame rate (4 frames/s) (SL) - Here same frame was repeatedly recorded and some frames were deleted in between still frames.
4. actual video sequence (VI) - Real time video.
5. 3D Cartoon character animation with the selected emotion (AN) - As shown in the Fig. 4 the mouth was moved according to the actual persons mouth shape, while a model that fits the emotion was selected manually.
6. Texture mapped 3D facial image - displayed as a 2D image (TM) - Data was processed by using a DCT based facial expression detection [11] method as shown in the Fig. 4.
7. Texture mapped 3D facial image - displayed as a 3D image (3D) - This was done in real time, where the person who makes the facial expressions was at a remote station while the observers make their judgments. The data preparation process was same as the above item 6.

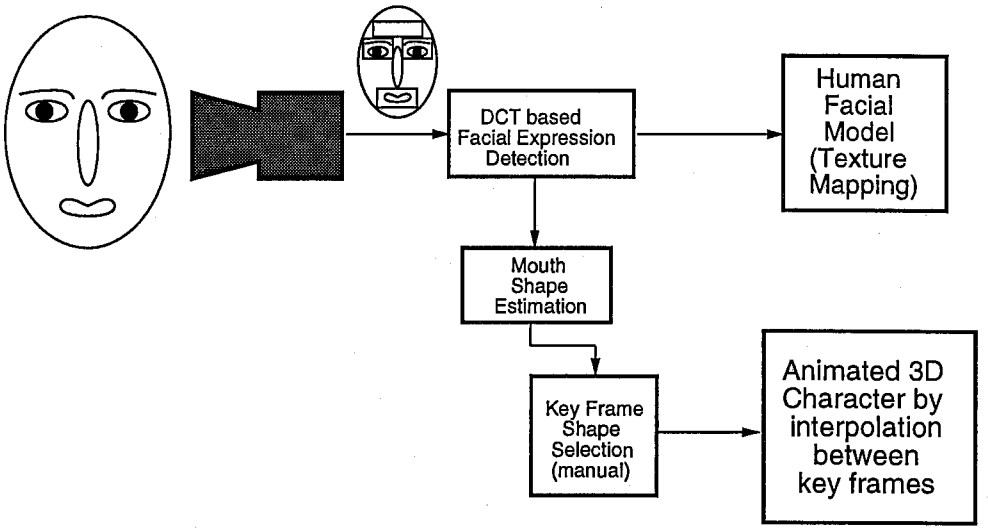


Figure 4: Data preparation for the Animation and Texture Mapping experiments

An NTSC video camera, a video editor and an NTSC VTR were the equipment that have been used in the 1-4 of the above experiments. Mosaic and slow frame rate image sequences were generated by using the video editor. The equipment used in the rest of the three types of the experiments were those explained in the paper [4]. Basically it consisted of an image processing board which estimates the facial muscle movements and which then converts those movements into a computer generated animating character or facial model.

During the evaluation the images were displayed on a large screen TV display of which the diagonal dimension was $34in$ ($86cm$). In average, participants were asked to sit approximately $240cm$ away from the screen so that the solid angle of the total head, looking from the subject's position, lie in between 7.5° to 10.5° in all cases. In a book written by Kato [13], he states that, according to psychological aspects the effective distance between two participants in a face to face meeting is 160cm (business distance). The solid angle of the participant's head seen by the counterpart at this distance is approximately 7.0°. So in our evaluation experiments, in order to get better clarity of the facial expressions, an angle larger than 7.0° was used .

3.4 Image Sequences used in the Evaluation

Figs 5, 6, 7 and 8 show a frame from each image sequence used in the evaluation process. As stated in the section 3.1, here we considered only the five emotions: anger, happiness, sadness, surprise and dislike. Although in some literature [14] reference to a very large number of different emotion types are made, we have selected only these five emotions for our evaluation, as we believed these are the most common type of emotions in face to face meetings.

(a) (b) (c) (d) (e)

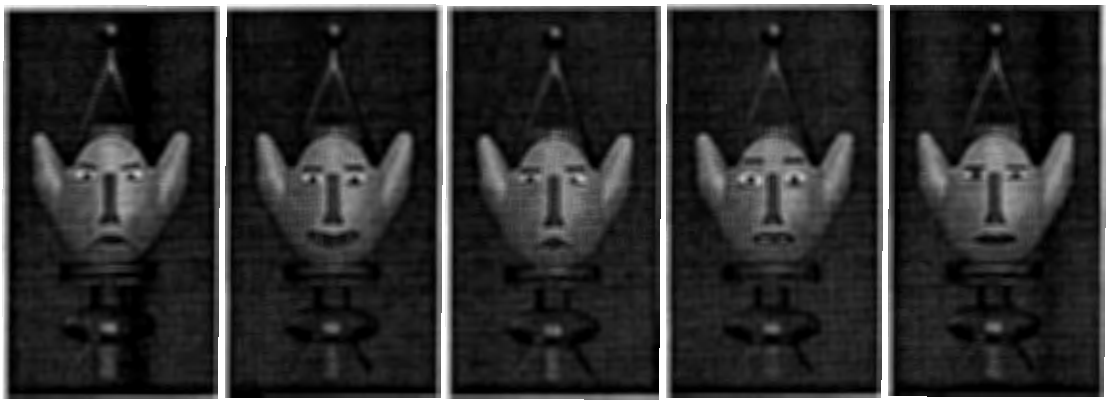
Figure 5: Images used for video and low frame rate emotion recognition (a) angry (b) happy (c) sad dislike

(a) (b) (c) (d) (e)

Figure 6: Images used for mosaic-ed emotion recognition (a) angry (b) happy (c) sad (d) surprise (d) dislike

4 Preliminary Evaluation using Whole Sentences with different Emotions

In reality effectiveness of face-to-face meetings are heavily dependent on facial expressions as well as verbal and non-verbal information. Verbal informations are the content that understood by people who can understand the language spoken by the speaker. In this evaluation study



(a) (b) (c) (d) (e)

Figure 7: Images used for Cartoon Character (Animation) emotion recognition (a) angry (b) happy (c) sad (d) surprise (d) dislike

(a) (b) (c) (d) (e)

Figure 8: Images used for Texture Mapped emotion recognition (a) angry (b) happy (c) sad (d) surprise (d) dislike

we try to find out the effect of non verbal clues, in communication, by suppressing the verbal content of the conversation.

Although, in the previous section we have shown five English sentences, the actual sentences read by the speaker were in Japanese, but context and the emotions were in the same order. We first carried out a preliminary evaluation to see the effect of emotion detection after introducing various constraints on verbal information. The preliminary evaluation consisted of the following three types of tests:

- Test 1: long version - Japanese voice track (key words deleted - Japanese audience)
- Test 2: long version - Non-Japanese voice track (Japanese audience)
- Test 3: long version - Japanese voice track (Non-Japanese audience)

We call these 3 tests *long version*, since the sentence patterns used here were longer than the one used in the final evaluation. In Test 1 of the long version although the participants could understand most of the context, they were not provided with a clue to understand it verbally. In the second and third cases, they could not understand the context totally, since the voice track selected were completely in-comprehensible to the participants.

4.1 Preliminary Test 1: long version - Japanese voice track (keywords deleted - Japanese audience)

Here, the words such as those which give the meaning *awful* etc., were deleted in the Japanese voice track and the total video clip was presented to the audience. Here, all the 5 emotions (randomly arranged) were presented with 6 different types of communication methods (above AU to TM).

4.2 Preliminary Test 2: long version - Non-Japanese voice track (Japanese audience)

Here, the total audio track was re-recorded with a language (in Sinhala - a Language which is used in Sri Lanka) which was totally incomprehensible by the participants. However the necessary intonations were added to present the emotion. Then, all the 5 emotions (randomly arranged) were presented with 6 different types of communication methods (above AU to TM).

4.3 Preliminary Test 3: long version - Japanese voice track (Non-Japanese audience)

Here, Japanese sentences with the emotions were shown to a group of non Japanese people (those who do not know even a greeting in Japanese). Here also, all the 5 emotions (randomly arranged) were presented with 4 different types of communication methods (above AU, VI, AN and TM).

4.4 Data Analysis

At each experiment, each person in the audience was given a questionnaire which consists of a table (see Table 1). On the table a sample answer given by a participant (let's say that the shown emotion was designated as *angry*) was marked.

Table 1: Judgment Results

	Expression 1					
	20%	40%	60%	80%	100%	Can not Judge
Anger			√			
Happiness						
Sadness	√					
Surprise						
Dislike						

Then we calculated the coefficients of a score matrix from the results (Table 2).

Table 2: Score Matrix

		Emotion Designated				
		Ang.	Hap.	Sad.	Sur.	Dis.
	Ang.	aa	ha	sa	pa	da
Emo-	Hap.	ah	hh	sh	ph	dh
tion	Sad	as	hs	ss	ps	ds
Det-	Sur.	ap	hp	sp	pp	dp
ected	Dis.	ad	hd	sd	pd	dd

The score for the judgment shown in the Table 1 is calculated as follows. Out of 100 points the viewer had already granted 80 to two emotions. The remaining 20 points was being distributed evenly, since that is the amount of uncertainty, which is considered to be equi-probable. Then the estimated coefficients (before normalization) are: $aa = 64, ah = 4, as = 24, ap = 4, ad = 4$. Similarly we can calculate the remaining coefficients of the score matrix.

The results obtained for the long audio version with Japanese voice track (for 8 people P1 to P8) is shown in Table 3. Here expression designated was *Angry*.

Table 3: Sample results obtained for long audio only experiment with Angry Designated Emotion

Expression 1 (Angry)										
	P1	P2	P3	P4	P5	P6	P7	P8	Tot.	Norm.
Ang	64	36	12	20	16	4	16	16	184	0.230
Hap	4	16	12	20	16	24	16	16	124	0.155
Sad	24	16	52	20	36	4	16	16	184	0.230
Sur	4	16	12	20	16	24	16	16	124	0.155
Dis	4	16	12	20	16	44	36	36	184	0.230

Note that although the designated emotion was *Angry*, the scores for Sadness and Dislike are also high. These results are for the audio only test, the results of which are far from correct. Similar process was carried out for the rest of the emotions and experiment types.

The participants of the Japanese audience tests were: 3 male and 5 female. All the participants were in the age group 20-35. The participants for the non-Japanese audience were: 4 male and 4 female in the age group of 18-40. None of them know a single word of Japanese, and also they never visited Japan. The experiment was carried out outside Japan.

Here the final score is calculated by adding all the 5 diagonal elements of the normalized score matrix. If all the subjects judged a designated emotion as it was designated then the normalized score for that particular emotion should be 1.0. In total we considered 5 emotions. Therefore in the results of subjective evaluation always we get a score less than 5.0 as the final answer, which represents the results of all the 5 emotions.

4.5 Results of the Preliminary Experiments

Results of the preliminary experiments are shown in Fig. 9 and Fig. 10. The results of animation leads all the other types of communication methods. This is because, human beings are well trained to identify symbolized character emotions than natural emotions. If we can improve such emotional clues in all the other methods we will be able to get much better results.

Compare the scores shown in Fig. 9 and Fig. 10. The scores for Test 3 is generally lower than both scores of Tests 1 and 2. Lower score means higher mis-judgments. In all these experiments we showed the same video of a Japanese person making emotions to a group of Japanese (in Tests 1 and 2) and to a group of non-Japanese (in Test 3). Although the Japanese group identified the emotions, the non-Japanese group were unable to identify some emotions made by the Japanese person, hence the final score became low in Test 3. Also in Test 2 most of the Japanese were able to identify the emotions with higher rate of accuracy, after adding auditory emotions made by a non-Japanese (results of Test 1 are by showing auditory emotions made by a Japanese plus visual emotions those are same as Test 2), although the content is totally incomprehensible, compared to that of Test 1. This means the auditory emotions made my the non-Japanese person were much prominent than those made by the Japanese person. These are fine examples of problems in emotion generation and recognition between different nations. So exaggerated or commonly recognizable set of emotions are vitally important in Teleconferencing between different nations.

Although in these experiments we used only a 2D type of texture mapped facial images, the results show that some improvements of emotion recognition can be obtained by using a kind of 3D (stereo) type representation. So we extended the testing procedure to 3D type displays at the same time reducing the length of the expressions. We believe that quick emotion recognition helps the local person to act in time in face to face meetings. So we select the following experiment.

5 Final Evaluation with a short emotion clip

As we have seen in the preliminary evaluation, the emotion recognition rate differs from each kind of experiment type. In all of the above tests we checked the results for a whole sentence. But, here we carried out the evaluation test as follows.

- Test 1: short version without voice
- Test 2: short version with voice

The idea behind this test was to find out how fast people can judge an emotion embedded in a sentence, only observing the first word of the sentence. The faster a person can judge an emotion the better the communication method is.

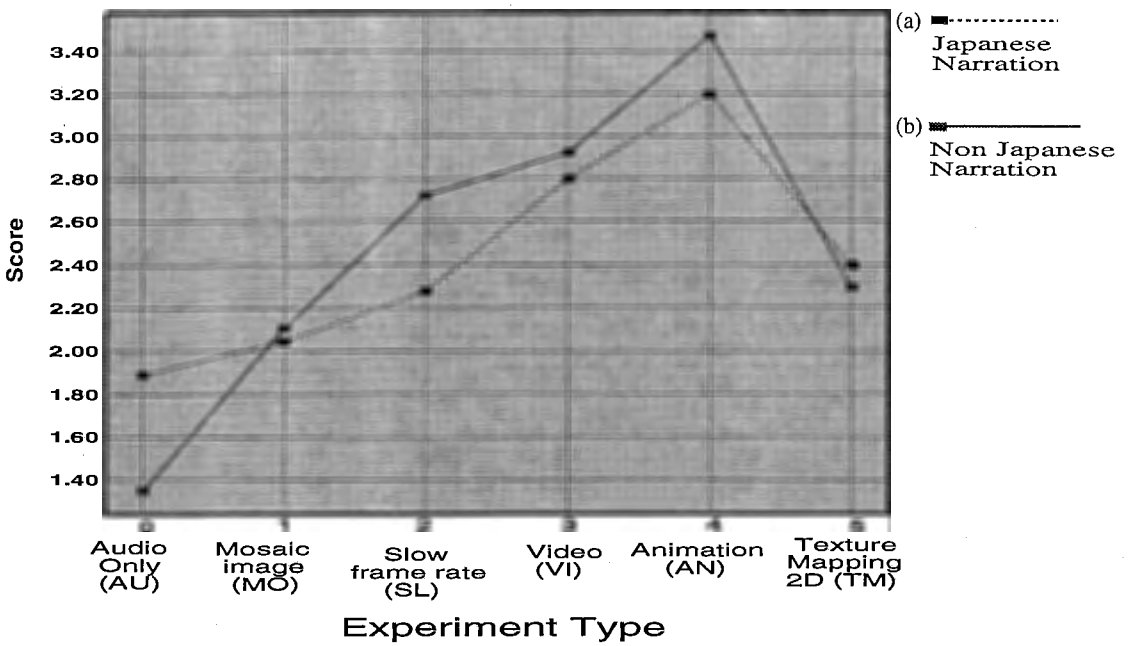


Figure 9: Scores of Showing (a) Whole Japanese Sentences after deleting the key words and (b) Replacing the sound track with Non-Japanese Narration, to a Japanese Audience (Results of Preliminary Tests 1 and 2)

5.1 Final Test 1: short version - without voice

In this test, only the video (without audio) of the first word, when the speaker is pronouncing *"Mr. Tanaka"*, was presented to the audience, for possible emotion detection. All the 5 emotions (randomly arranged) were presented with 6 different types of communication methods (above MO to 3D).

5.2 Test 2: short version - with voice

In this test, both video and audio of the first word, when the speaker is pronouncing *"Mr. Tanaka"*, was presented to the audience for possible emotion detection. All the 5 emotions (randomly arranged) were presented with all the 7 different types of communication methods (above AU to 3D).

5.3 Evaluation Results

Now we present the results of the final evaluation experiment, in which we used one word of the sentence for the evaluation.

Score = 1.45

Similar score matrices can be obtained for the rest of the experiment types with and without

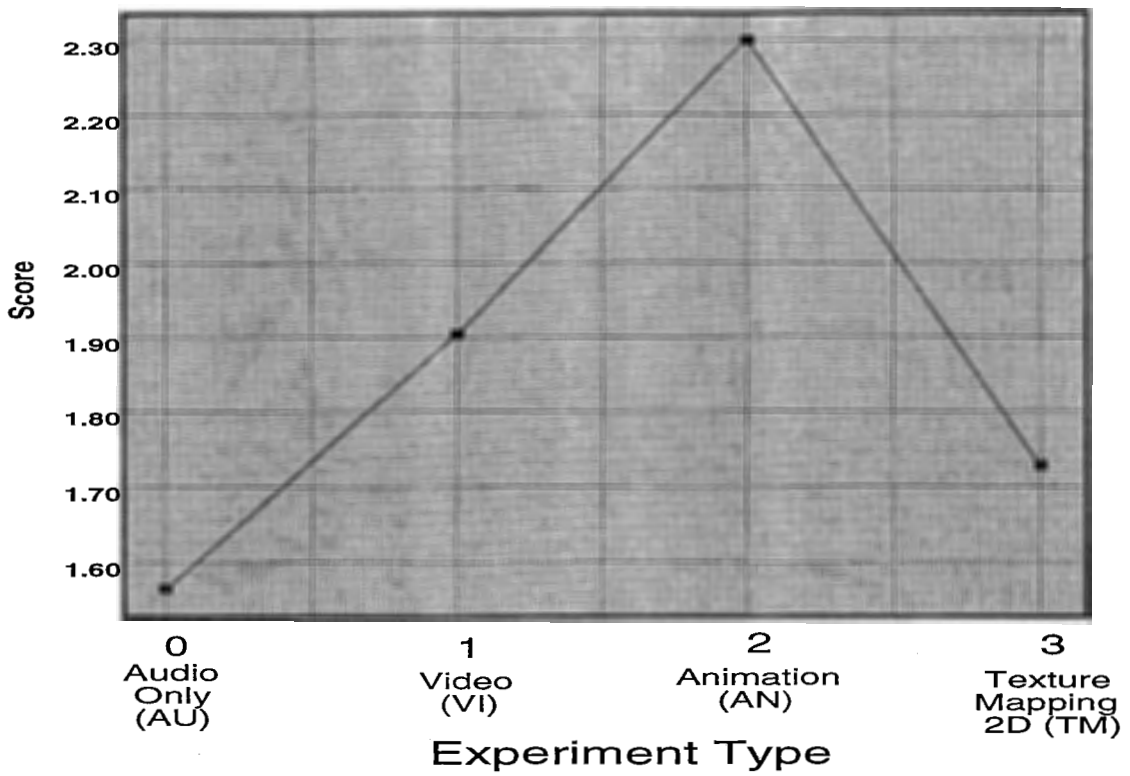


Figure 10: Scores of Showing Whole Japanese Sentences to a Non Japanese Audience (Results of Preliminary Test 3)

audio.

The combined results based on emotions, for each experiment with audio are shown in Fig. 11.

Total Normalized Scores for each experiment type with and without voice are shown in Fig. 12.

Table 4: Normalized score matrix for the short audio only (AU) version

		Emotion Designated				
		Ang.	Hap.	Sad.	Sur.	Dis.
	Ang.	0.230	0.260	0.205	0.055	0.175
	Hap.	0.155	0.185	0.180	0.180	0.200
Emo.	Sad.	0.230	0.160	0.155	0.055	0.225
Det.	Sur.	0.155	0.185	0.155	0.655	0.175
	Dis.	0.230	0.210	0.305	0.055	0.225

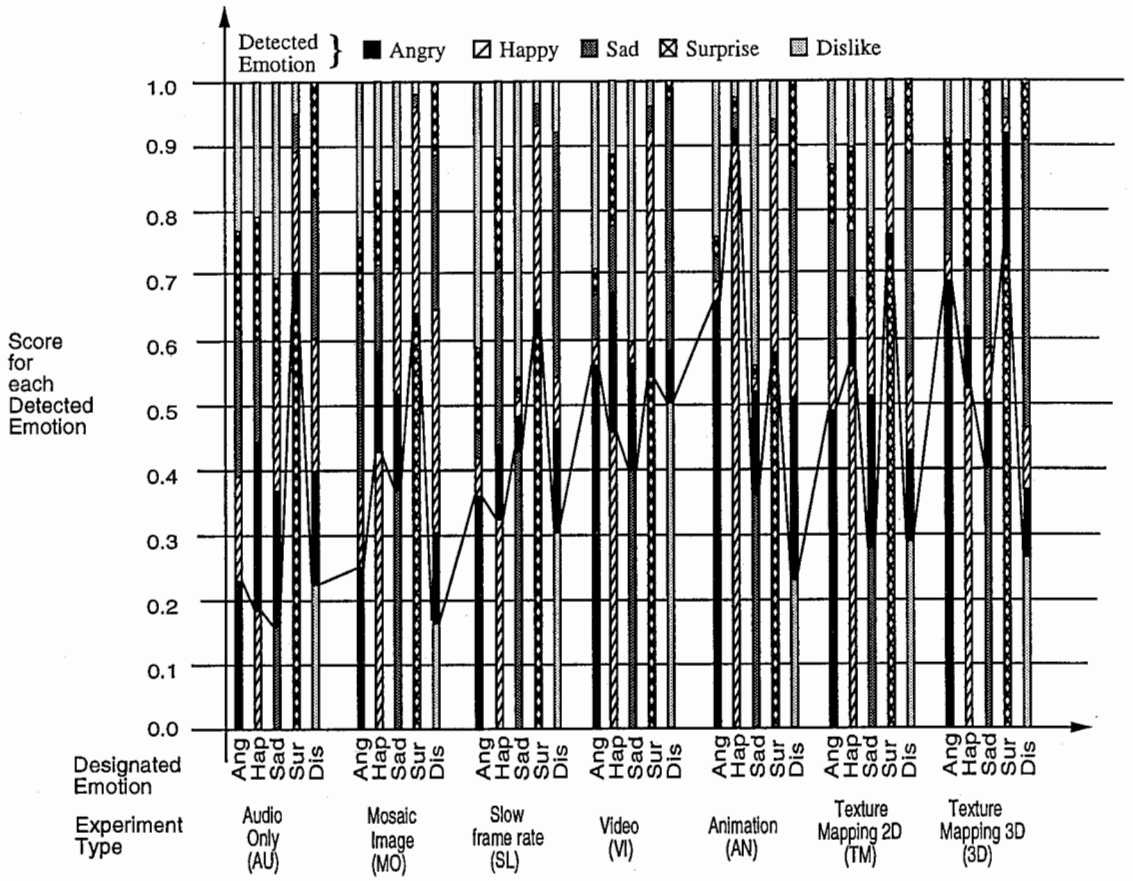


Figure 11: Scores of each experiment based on Emotions. Note that the detected emotions which lie above the zig-zag line shown in the graph are mis-judgments

5.4 Discussion of Results

In Fig. 11 scores obtained for all the 7 types of experiments based on emotions are shown as stacked bars. The bottom most bar of each stacked bar is the correct judgment (designated emotion detected as designated). The taller the bottom most bar is, the more accurate the judgment is. Note that the total height of the bottom most bars in the audio only experiment type is very much less than that of Animation type. Also we can see that angry, happy and surprise emotions in the animation were easy to detect than the rest of the two, as those two are too ambiguous. In the Texture Mapping 3D experiment except the dislike emotion the rest of the emotions were easy to distinguish, hence showed a higher recognition rate.

In Fig. 12 we can see that the score of the Animation Experiment is the highest compared to all the other methods. This is because Human can identify instantly, if they see a symbolic emotion. This is where we are aiming at. If we can enhance the emotions that are easily recognizable, then it will become a good teleconferencing environment.

We can see that the results obtained for the 2D Texture Mapped experiment is higher than that of Audio only, Mosaic and Slow Frame Rate methods. This means that, by only using a very

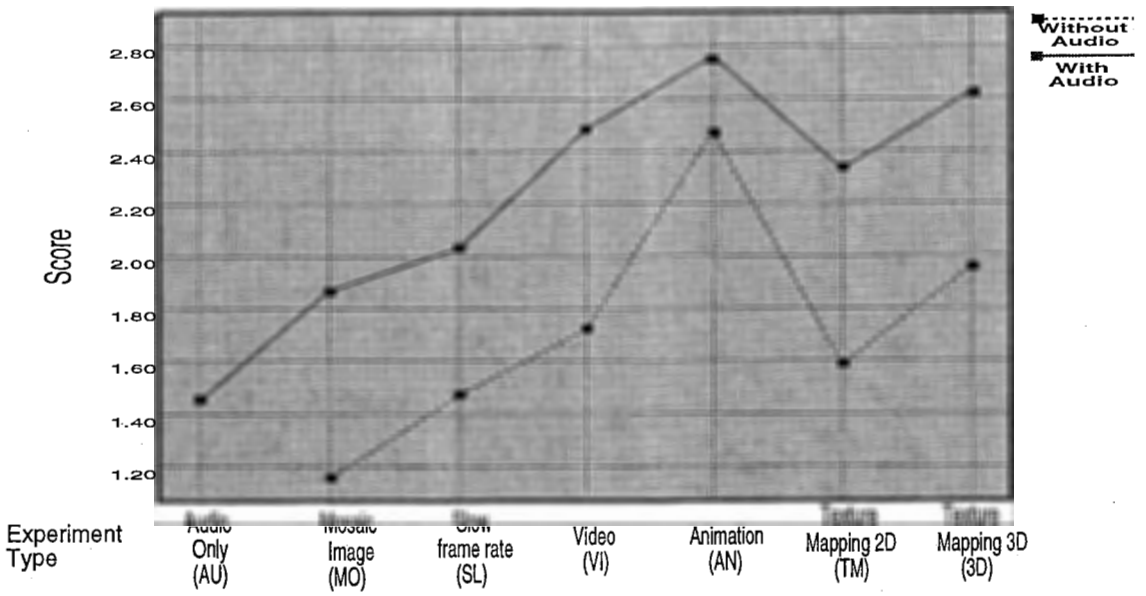


Figure 12: Scores of Showing One Word having 5 Different Emotions with and without Audio

robust Computer Graphic image, one can improve the quality of the communication beyond that is capable of most of the present day communication methods.

By using a stereo type display unit we were able to reproduce the emotions, which gave a higher recognition rate due to some improvements and emotion exaggerations in emotion reproduction process. (see Texture Mapping 3D in Figs. 11 and 12). With some more improvements we hope this recognition rate can be increased as close as to that of Animation type, as the basic idea of both are same. Finally, we can state that this kind of a system can be used to improve the face to face teleconferencing beyond that is capable of direct video type systems.

6 Comparison of Data Rate with Commonly Available Methods

Here we compare the required data rate for the proposed method with those of two other methods. Slow frame rate (commonly found in low rate video phones) and real time video (commonly found in high speed teleconferencing) are two commonly available modes of communication with visual clues. Here we compare the data rate required for each of those type of communications using presently available data compression techniques, with our approach (see Table 5). In our approach at the beginning of communication, texture data and control point data are to be transmitted to the remote end. Typical size of such initial data are also shown in the table. However non of the other two methods require such initial data. The final column shows that after 100 sec the proposed approach requires a less number of data bits per second compared to any other method.

More specifically the plot of Data Rate verses Time for each different type of communication

	Data Rate	Initial Bytes	Data Rate After 100 sec
Slow frame rate	$\approx$ 64 kb/s (using H.261)	0	64 kb/s
Video	$\approx$ 1.5 Mb/s (using MPEG)	0	1.5 Mb/s
Proposed Method	$\approx$ 7 kb/s† (without compression)	4.096 M	48 kb/s

Table 5: Required data rates for each type of communication method († 4 byte each for 14 control points x 15 frames per second = 840 bytes/sec $\approx$ 7 kb/s)

method in Fig. 13 shows that we get a break even point around 72 secs. Which means that the proposed concept is better than both of the conventional types of communication in the point of view of data rate, if the time of conversation is longer than 72 sec, even without using any data compression techniques.

7 Conclusions

In this paper we showed that the Virtual Person concept is a unique way of communication, which has the capability of emotion enhancement between the local and remote participant. Although in the animation experiment we selected the emotions by a manual process the final outcome was far better than using a direct video sequence. Using an automatic emotion recognition process [12, 15, 16] in between the transmitter and the receiver, we can obtain much better results for 3D texture mapping also. This means that we can expect an emotion enhanced teleconferencing system that supersedes the normal face to face meetings, by effectively alleviating the emotional barriers between different nations.

Also we have showed that it is a better alternative to the blurred or mosaic-ed facial images that one can find in some television interviews with people who are not willing to be exposed in public. Finally we compared the data rates required for two different types of commonly available communication methods with our approach. Even without using any data compression techniques the proposed method require very low rate, if the conversation is longer than a specified minimum time.

8 Acknowledgment

The authors wish to thank Dr. N. Terashima, President of ATR Communication Systems Research Laboratories, Dr. K. Habara, Executive Vice President of ATR International (Chairman of the Board of ATR Communication Systems Research Laboratories) for the thoughtful advises and encouragement on this research. Authors also wish to thank Dr. J. Ohya, Mr. K. Ebihara, Mr. S. Ura, Mr. M. Hirose, Mr. T. Ochi, Mr. S. Imura, and A. A. Pasqual for helping the setting of the experiment and also to all the subjects who took part in the subjective evaluation.

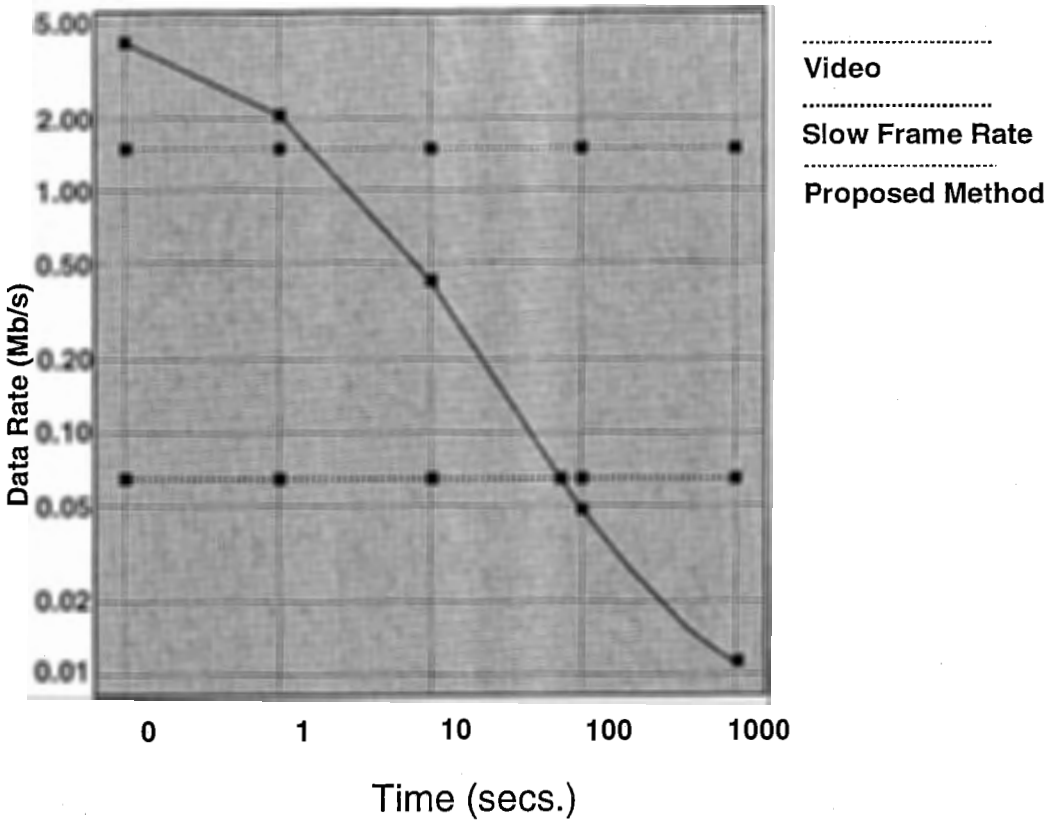


Figure 13: Data rate versus time for different communication methods

References

[1] Y. Matsushita et. al. *Groupware for Multimedia Era (in Japanese).* Ohmu, 1994.

[2] Christopher D. Wickens. *Engineering Psychology and Human Performance*, chapter 5, pages 167–210. Harper Collins, 2nd edition, 1992.

[3] A. Chapanis, R. B. Ochsman, R. N. Parrish, and G. D. Weeks. Studies in interactive communication: I. the effect of four communication modes on the behaviour of teams during cooperative problem-solving. *Human Factors*, 14(6):487–509, 1972.

[4] Haruo Noma, Yasuichi Kitamura, Tsutomu Miyasato, and Fumio Kishino. Multi-point virtual space teleconferencing system. *IEICE Trans. Communication*, E78-B(7):970–979, July 1995.

[5] Fumio Kishino, Tsutomu Miyasato, and Nobuyoshi Terashima. Virtual space teleconferencing - "commuincation with realistic sensations". In *Procs. of 4th Int. Workshop on Robot and Human Communication (ROMAN'95)*, pages 205–210, July 1995.

[6] Mel Slater and Martin Usoh. Body centered interaction in immersive virtual environments. In Nadia Magnenat Thalmann and Daniel Thalmann, editors, *Artificial Life and Virtual Reality*, chapter 9, pages 123–147. John Wiley and Sons Ltd, 1994.

[7] Nadia Magnenat Thalmann and Daniel Thalmann. Creating artificial life in virtual reality. In Nadia Magnenat Thalmann and Daniel Thalmann, editors, *Artificial Life and Virtual Reality*, chapter I, pages 1–10. John Wiley and Sons Ltd, 1994.

[8] Akira Utsumi, Fumio Kishino, and Tsutomu Miyasato. Multi-camera hand pose recognition system using skeleton image. In *Procs. of 4th Int. Workshop on Robot and Human Communication (ROMAN'95)*, pages 219–224, July 1995.

[9] Frederic I. Parke. Techniques for facial animation. In Nadia Magnenat Thalmann and Daniel Thalmann, editors, *New Trends in Animation and Visualization*, chapter 16, pages 229–241. John Wiley and Sons Ltd, 1991.

[10] Liyanage C. De Silva, Mitsuho Tahara, Kiyoharu Aizawa, and Mitsutoshi Hatori. A teleconferencing system capable of multiple person eye contact (mpec) using half mirrors and cameras placed at common points of extended lines of gaze. *IEEE Trans. on Circuits and Systems for Video Technology*, 5(4):268–277, August 1995.

[11] Kazuyuki Ebihara, Jun Ohya, and Fumio Kishino. A study of real time facial expression detection for virtual space teleconferencing. In *Procs. of 4th Int. Workshop on Robot and Human Communication (ROMAN'95)*, pages 247–252, July 1995.

[12] Tatsumi Sakaguchi, Jun Ohya, and Fumio Kishino. Facial expression recognition from image sequence using hidden markov model (in japanese). *Journal of Institute of Television Engineers*, 49(8):1060–1067, August 1995.

[13] Takayoshi Kato. *Awareness and Image of the Space (in Japanese)*, pages 139–141. Shinyousha, 1986.

[14] Paul Ekman and Wallace V. Friesen. *Unmasking the Face.* Prentice-Hall Inc., 1975.

[15] Yaser Yacoob and Larry Davis. Recognizing faces showing expressions. In *Proc. of Int. Workshop on Automatic Face-and Gesture-Recognition, Zurich*, 1995.

[16] Mark Rosenblum, Yaser Yacoob, and Larry Davis. Human emotion recognition from motion using a radial basis funtion network architecture. In *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects, Austin, TX*, November 1994.

Liyanage C. De Silva was born in Sri Lanka, on January 12, 1962. He received B.Sc.Eng.(Hons.) degree from the University of Moratuwa Sri Lanka in 1985, M.Phil. degree from The Open University of Sri Lanka in 1989, M.Eng. and Ph.D degrees from the Univ. of Tokyo, Japan in 1992 and 1995 respectively. He was with the University of Tokyo, Japan, from 1989 to 1995. Currently, he is pursuing his post doctoral research as a research with ATR Communication Systems Research Laboratories, Kyoto, Japan, as a Research Associate. His current research interests are Image Processing, Computer Vision and Visual Communication. Dr. De Silva received the 1995 Best Student Paper Award from SPIE - The International Society for Optical Engineering.

Tsutomu Miyasato was born in Okinawa, Japan, in 1953. He received the B.E. degree in electronic engineering from the University of Electro-Communications, Tokyo, Japan, in 1976, and the M.E. degree in electronic systems from Tokyo Institute of Technology, Tokyo, Japan, in 1978. He received the Ph. D. degree from Tokyo Institute of Technology in 1991. Since 1978, he was with the Research and Development Laboratories of the Kokusai Denshin Denwa (KDD) Co., Ltd., Tokyo, Japan, and worked in the field of high efficiency coding of handwritten

signals, image processing in videotex. Since 1993, he is with ATR Communication Systems Research Laboratories. He is currently engaged in a teleconference system based on virtual reality techniques. Dr. Miyasato is a member of the Institute of Electronics, Information and Communication Engineers of Japan, the Institute of Television engineers of Japan, and the Information Processing society of Japan.

Fumio Kishino received the B.E., M.E. and D.E. degrees from Nagoya Institute of Technology, Nagoya, Japan, in 1969, 1971 and 1995, respectively. In 1971, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation, where he was involved in work on research and development of image processing and visual communication systems. In mid-1989, he joined ATR Communication Systems Research Laboratories as a head of the Artificial Intelligence Department. His research interests include image processing, artificial intelligence, and communication with realistic sensations. Dr. Kishino is a member of IEEE, the Institute of Electronics, Information and Communication Engineers, and the Institute of Television Engineers of Japan.

A Appendix: Sentence Patterns Used in the Evaluation

Japanese Sentences Used in the Evaluation Experiment with non-Japanese Audience

1. (怒り) 田中さん、あの屋根の形はひどい。
2. (幸福) 田中さん、あなたのモデルはとてもいいですね。
3. (悲しみ) 田中さん、大問題があります。
4. (驚く) 田中さん、すばらしい作品ですね。
5. (いや気) 田中さん、あの形 (屋根の形) 、私は大きらいです。

Japanese Sentences Used in the Evaluation Experiment with Japanese Audience - key words deleted

1. (怒り) 田中さん、あの屋根の形は＊＊＊。
2. (幸福) 田中さん、あなたのモデルは＊＊＊＊＊ですね。
3. (悲しみ) 田中さん、＊＊＊があります。
4. (驚く) 田中さん、＊＊＊＊＊作品ですね。
5. (いや気) 田中さん、あの形 (屋根の形) 、私は＊＊＊＊です。

Words Used in the Final Evaluation Experiment

1. (怒り) 田中さん、...。
2. (幸福) 田中さん、...。
3. (悲しみ) 田中さん、...。
4. (驚く) 田中さん、...。
5. (いや気) 田中さん、...。

B Appendix: Making of the Evaluation Video

Table 6: No of video clips used in each experiment

Experiment type	long version with Japanese audience (key words deleted)	long version with Japanese audience (non Jap. voice)	long version with non Jap. audience (Jap. voice)	short version with audio	short version without audio
Audio Only (AU)	5	5	5	5	-
Mosaic Image (MO)	5	5	-	5	5
Slow Frame Rate (SL)	5	5	-	5	5
Video (VI)	5	5	5	5	5
Animation (AN)	5	5	5	5	5
Texture Map. 2D (TM)	5	5	5	5	5
Texture Map. 3D (3D) (This was done in real time)	-	-	-	5	5
Total Number of Video Clips Used in the Experiment					145

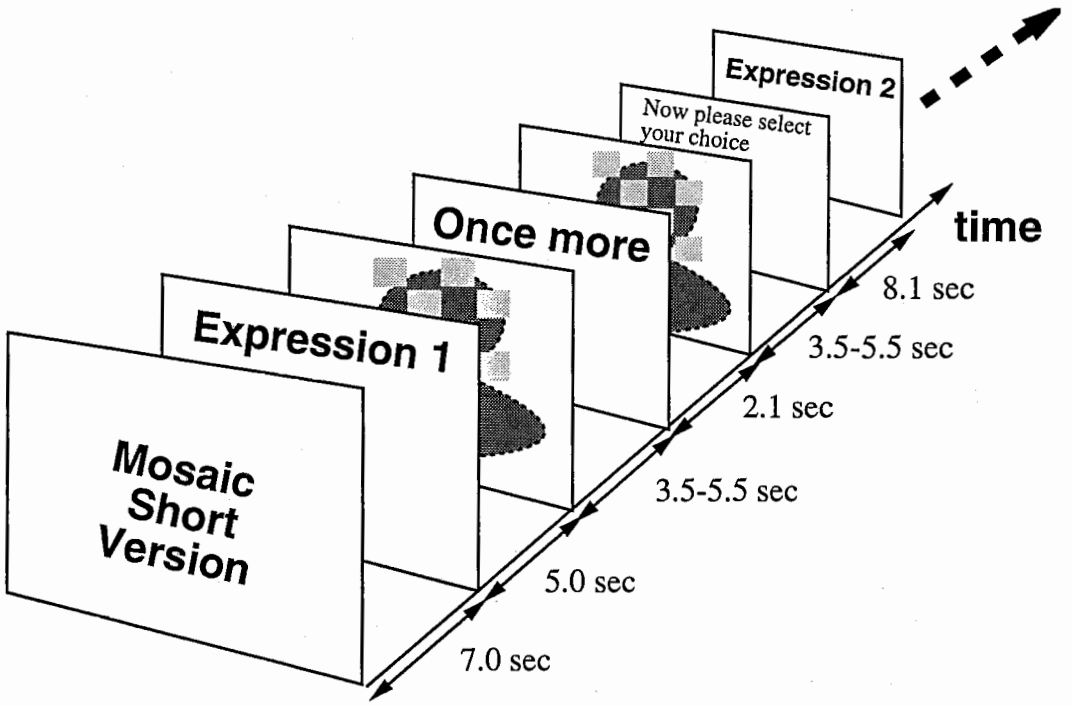


Figure 14: An example video clip shown to the subjects for the emotion detection

C Appendix: An Example Subject's Response Sheet

Mosaic and Audio - short form(MS)

Expression 1						
	20%	40%	60%	80%	100%	Can't Judge (判定できない)
Anger (怒り)						
Happiness (幸福)						
Sadness (悲しみ)		√				
Surprise (驚き)						
Disgust/Dislike (嫌悪)						
Expression 2						
	20%	40%	60%	80%	100%	Can't Judge (判定できない)
Anger (怒り)						
Happiness (幸福)						
Sadness (悲しみ)			√			
Surprise (驚き)						
Disgust/Dislike (嫌悪)						
Expression 3						
	20%	40%	60%	80%	100%	Can't Judge (判定できない)
Anger (怒り)						
Happiness (幸福)		√				
Sadness (悲しみ)						
Surprise (驚き)						
Disgust/Dislike (嫌悪)						
Expression 4						
	20%	40%	60%	80%	100%	Can't Judge (判定できない)
Anger (怒り)						
Happiness (幸福)	√					
Sadness (悲しみ)						
Surprise (驚き)				√		
Disgust/Dislike (嫌悪)						
Expression 5						
	20%	40%	60%	80%	100%	Can't Judge (判定できない)
Anger (怒り)						
Happiness (幸福)						
Sadness (悲しみ)	√					
Surprise (驚き)						
Disgust/Dislike (嫌悪)		√				

D Appendix: An Example Evaluation Sheet

Mosaic and Audio - short form(MS)

Expression 1										
	P 1	P 2	P 3	P 4	P 5	P 6	P 7	P 8	Tot.	Normalized
Anger (怒り)	12	16	12	8	12	28	8	20	116	0.145
Happiness (幸福)	12	16	12	8	12	8	68	20	156	0.195
Sadness (悲しみ)	52	36	52	68	52	8	8	20	296	0.370
Surprise (驚き)	12	16	12	8	12	8	8	20	96	0.120
Disgust/Dislike (嫌悪)	12	16	12	8	12	48	8	20	136	0.170
Expression 2										
	P 1	P 2	P 3	P 4	P 5	P 6	P 7	P 8	Tot.	Normalized
Anger (怒り)	8	16	16	8	12	32	4	16	112	0.140
Happiness (幸福)	8	16	36	68	52	12	64	16	272	0.340
Sadness (悲しみ)	68	36	16	8	12	12	24	16	192	0.240
Surprise (驚き)	8	16	16	8	12	12	4	16	92	0.115
Disgust/Dislike (嫌悪)	8	16	16	8	12	32	4	36	132	0.165
Expression 3										
	P 1	P 2	P 3	P 4	P 5	P 6	P 7	P 8	Tot.	Normalized
Anger (怒り)	12	12	20	12	12	32	4	20	124	0.155
Happiness (幸福)	52	52	20	52	52	12	84	20	344	0.430
Sadness (悲しみ)	12	12	20	12	12	12	4	20	104	0.130
Surprise (驚き)	12	12	20	12	12	12	4	20	104	0.130
Disgust/Dislike (嫌悪)	12	12	20	12	12	32	4	20	124	0.155
Expression 4										
	P 1	P 2	P 3	P 4	P 5	P 6	P 7	P 8	Tot.	Normalized
Anger (怒り)	0	0	0	4	4	12	4	0	24	0.030
Happiness (幸福)	20	80	0	4	4	12	4	60	184	0.230
Sadness (悲しみ)	0	0	0	4	4	12	4	0	24	0.030
Surprise (驚き)	80	20	100	84	84	52	84	40	544	0.680
Disgust/Dislike (嫌悪)	0	0	0	4	4	12	4	0	24	0.030
Expression 5										
	P 1	P 2	P 3	P 4	P 5	P 6	P 7	P 8	Tot.	Normalized
Anger (怒り)	8	16	16	44	52	32	8	20	196	0.245
Happiness (幸福)	8	16	16	4	12	12	8	20	96	0.120
Sadness (悲しみ)	28	36	36	44	12	32	8	20	216	0.270
Surprise (驚き)	8	16	16	4	12	12	8	20	96	0.120
Disgust/Dislike (嫌悪)	48	16	16	4	12	12	68	20	196	0.245

Here the evaluation results of 8 subjects (P1 to P8) are shown. The column P1 is calculated from the results shown in **Appendix C**. *For example the values in the P1 column for* **Expression 1** *is calculated as follows. Out of 100 points the subject has already granted 40 points to Sadness. The remaining 60 points are being distributed evenly, since that is the amount of*

uncertainty, which is considered to be equi-probable. Therefore the score for this expression is $40 + \frac{60}{5} = 52$. *This procedure is repeated with all the 8 subjects, hence the total score is divided by 800 to get the normalized score for each expression. If all the subjects judged a designated emotion as it is designated then the normalized score for that particular emotion should be 1.0. Then the final score is calculated by adding all the 5 score values obtained for each expression. Therefore in the results of subjective evaluation always we get a score less than 5.0 as the final answer, which represents the results of all the 5 emotions.*

	Expression 1	*Expression 2*	*Expression 3*	*Expression 4*	*Expression 5*
Designated as	*Sadness*	*Disgust*	*Happiness*	*Surprise*	*Anger*
Score	*0.370*	*0.165*	*0.430*	*0.68*	*0.245*
Total score					1.89

The above result is shown in Fig. 12 as the data point MO on the cure marked as "with audio". The rest of the data points are calculated similarly .