In Proc. of Int. Workshop on Robot and Human Comm. (RoMAN'96), pp. 158-163, Tsukuba, Japan, Nov. 1996.

Image-based Personal Communication Using an Innovative Space-variant CMOS Sensor

G. Sandini[2], J. Nielsen[1], A. Argyros[7], E. Auffret[4], F. Ciciani[1], P. Dario[6], B. Dierickx[3], F. Ferrari[1], H. Frowein[5], C. Guerin[4], P. Haigron[4], L. Hermans[3], L. Knudsen[2], A. Manganas[7], A. Mannucci[1], P. Questa[2], A. Sassi[6], D. Scheffer[3], T. Tapie[4], W. Woelder[5]

[1] UNITEK Consortium - Italy (ferrari@aitek.it)
[2] DIST, University of Genova, Italy (giulio@dist.unige.it)
[3] IMEC - Belgium (hermans@imec.be)
[4] Thomson - France (auffrete@tcetbs1.thomson.fr)
[5] Instituut voor Doven - The Netherlands
[6] Scuola Superiore S. Anna - Italy (dario@arts.sssup.it)
[7] Knossos Technologies - Greece (manganas@knossos.knossos.gr)

August 28, 1997

Abstract

This paper reports the results of **IBIDEM**[1], a collaborative project supported by the European Union under the *Technology Initiative for Elderly and Disabled People*-TIDE initiative. The goal of the project has been to build a prototype of a videophone, connected to standard PSTN lines, that can be used by hearing impaired persons for tele-communication. The most innovative content of the project has been the design, fabrication and use of a new generation of space-variant visual sensors characterized by a spatial resolution decreasing linearly with distance from the geometric center of the sensor chip. This sampling strategy allows, with a limited number of pixels and consequently a high frame rate, to transmit high resolution information for "speech-reading" and a wide field of view for facial expressions and gestures.

1 Introduction

Image-based communication is, at least potentially, one of the most useful communication technologies not only because it gives a more personal feeling to voice-based communication through facial expressions and gestures, but also because it may allow to extend the use of traditional devices, such as telephones, to persons with some form of hearing disabilities. Currently the major bottleneck that still prevents the spread of image-based communication devices is transmission bandwidth which, if sufficient to transmit reasonable "live images" (such as through ISDN connections) it is still too expensive and, therefore, justifiable at present only for teleconferencing facilities.

On the other hand, in currently available videophones the information content is still largely contained in the audio data and the visual information only gives the "impression" of a face-to-

[1] For information regarding IBIDEM contact: Fabrizio Ferrari - Unitek Consortium - Via Pisa 12/1, 16146 Genova Italy. Ph:+39 10 3620102, Fax: +39 10 314873 - e-mail: ferrari@aitek.it

face communication without allowing the "use" of facial and/or body motions in a meaningful way and do not provide the kind of information required by deaf and hard-of-hearing. Until recently it was felt that these kind of problems could be solved through the spread of ISDN connections. Unfortunately this has not happened at the rate it was expected and, moreover, does not solve the problem of mobile communication.

In order to overcome some of these limitations the project IBIDEM is addressing the bandwidth problem from a somewhat different perspective by trying to match the visual information transmitted to the perceptual capabilities of the human visual system [?].

2 Motivations

Statistics for disabilities are not simple to estimate[2]. Differences between definitions of what constitutes a disability lead to large variations in the estimates from country to country in Europe. The COST219 action has tried to cut through this jungle and provide estimates for the main groups of disabilities [?]. In this study it is found that hearing impairment is a problem for up to 10-15% of the whole population in Europe. Information collected in the Netherlands by the IvD (one of the partners of IBIDEM) tells that about 6.4 % of the Dutch population have problems with their hearing. Information from the AFA Centro REUL in Italy indicates that approximately 7.8 % of the Italian population have a hearing loss that is considered handicap-inducing. The relation between age and hearing-loss is equally clear. The numbers from the italian study (AFA-REUL), the dutch study (IvD), and an american study[3] indicates the same relation: the occurrence of hearing-impairment increases with age (table 1 shows the numbers from the dutch study).

Age	
$5-44$	02.8 %
$45-54$	07.1 %
$55-64$	09.6 %
$65-74$	15.4 %
$75-84$	29.1 %
>84	53.0 %

Table 1: Relation between hearing-loss and age

The lesson that also can be learned from the demographic studies, however, is that the elderly make up a large part of this population segment. The results reported in the COST-219 project [?] indicates that attitudes towards technology are not particularly related to age, but that the ease of use of the equipment is paramount for the actual acceptance, especially in this marketplace. Moreover, it is clear from the demographic study that a large segment of the deaf and hard-of-hearing population have residual hearing that should be exploited to weight the individual communication modalities (audio/video): for part of the group, images will serve a supportive function while for others images will be the primary communication modality.

The overall conclusions are that useful videophones are currently not generally available on the market, and that the deaf and hard-of-hearing are primarily offered various form of text-telephones or faxes along with aids for improving the usefulness of normal telephones by people with residual hearing (like amplifier systems, inductive coupling to hearing aids, etc.).

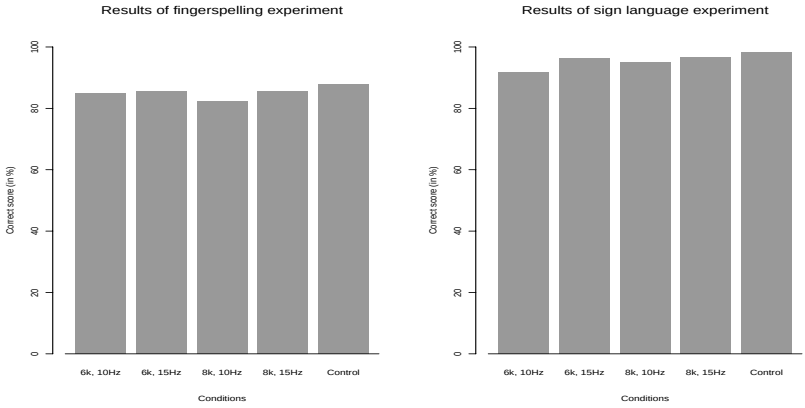
[2]The data reported here are derived from a number of different sources without the possibility for checking the consistency of the individual statistic this study can only serve as an indicator for the area.

[3]a demographic study from the Center for Assessment and Demographic studies, Gallaudet University

3 Human Factors

At the beginning of the project one of our main concern (and a specific request of the funding agency) was to carefully define the user's requirements in terms of both spatial and temporal resolution. This activity was coordinated by the Instituut voor Doven in The Netherlands, who specializes in teaching and training deaf and hard-of-hearing. An experimental paradigm was designed to investigate a number of communication modalities (speech-reading, finger spelling and signing).

For each modality a set of sequences was recorded and processed off-line to simulate different sensor layouts and frame rates [4]. The processed sequences were presented to people with hearing disabilities and their capability to correctly perceive the sentences were measured (for a detailed description of the experimental paradigm and results, see [?]). Results of the experiments are presented in figure 1.



,

Figure 1: Results of the experiments. The left column of the bar-graph is the control score, that is the measure of the subjects capability in understanding the different communication language through the original videotape. The subject scores measured during the experiments with the IBIDEM layouts are then compared with the control score. The experiments performed with finger-spelling and sign language gave excellent results. Mean scores for finger-spelling varied from 84.90% to 85%. Mean control score was 87.80%. Mean score for sign language varied from 91.91% to 96.56%, while the mean control score was 98.22%. In the case of lip-reading (results not shown here) the performance is not as good but it should be pointed out that, due to the use of standard VHS video the resolution of the processed tapes was lower than what can be obtained with the IBIDEM camera. For a perfect simulation a resolution of at least 800x800 pixels would have been necessary to display the full field as well as the maximum resolution in the center.

An overall evaluation of the experiments has shown that from a perception point of view there is no difference between the results using the two proposed spatial resolutions, while a substantial difference exists between the two tested images frame rates; this is in accord with the studies by Frowein et al. [?] where they have investigated the effect of frame rate in speech recognition, and have observed that there is an appreciable improvement in speech reception score increasing the frame rate up to 15 Hz, but the performance does not improve above this value.

[4] Two different spatial resolution, namely 96 receptive fields on each of the 64 concentric circles and 128 on 64 circles, and two temporal refresh rates (10 Hz and 15 Hz) were investigated.

4 The IBIDEM System

The most innovative part of the system is the use of a space-variant sampling strategy adopted for the sensor and derived from the distribution of photoreceptors in the human retina. Broadly speaking the overall system shown in fig. 2 includes: *i)* the CMOS sensor; *ii)* the camera; *iii)* the control and interface module *iv)* the pan/tilt unit and pointing device.

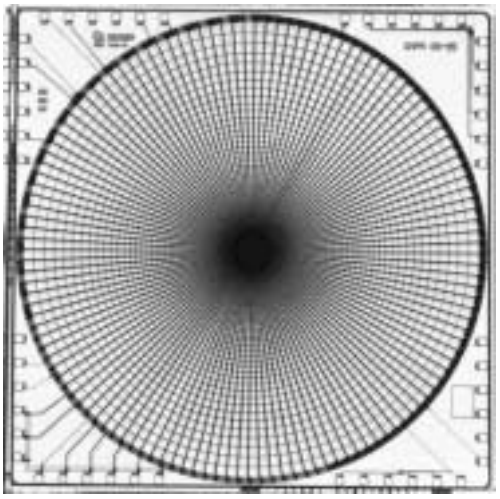
Figure 2: The IBIDEM system

4.1 The Sensor

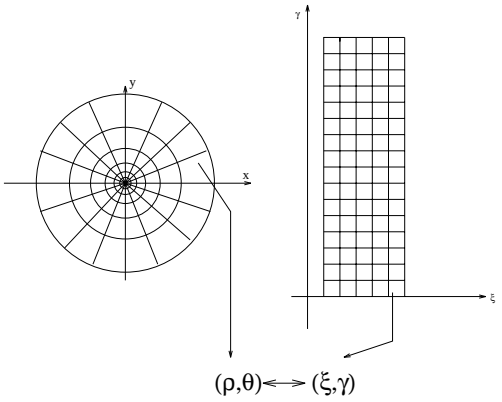
The main feature of the sensor is shown in figure 3. Its advantage with respect to standard TV cameras is that, with a relatively small number of pixels (about 8,000 at most), it will allow high resolution in the areas of interest (e.g. lips and/or fingers) while still maintaining a wide field of view in order to perceive the facial expression of the speakers and the overall pattern of arms gesture.

The images acquired by such a space-variant sensor are Cartesian images in polar coordinates and, due to the linear increase of inter-receptor distance a logarithmic compression of the topology with eccentricity is also performed. This mapping, which is called *log-polar mapping* (see figure 3), is conformal and, therefore, allows any local compression technique developed for standard images, to be adapted to these particular images.

The idea of mimicking the distribution and size of the human receptive fields in the human retina for sampling non uniformly a scene has been first introduced by Chaikin and Weiman [**?**] and further developed by Sandini and Tagliasco [**?**]. The non uniform sampling grid consists of receptive fields displaced over concentric circles; the size of the receptive fields increases linearly from the inner circle towards the outermost.



(a) Retina plane



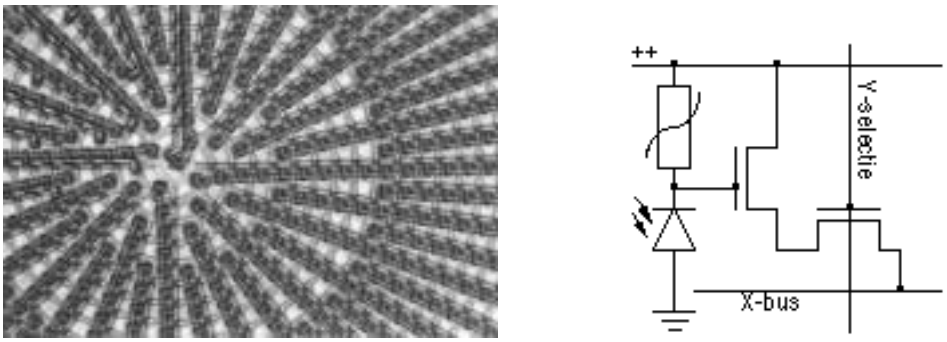
(b) Log-polar plane

(c) Log-polar image

(d) Retina Image

Figure 3: (a) Layout of the CMOS sensor; (b) Any position P in the retinal plane can be expressed in terms of (ρ, θ) coordinates (left). In the log-polar plane (right) the same position is identified by (ξ_i, γ_i). (c) Example of logpolar image acquired by the sensor; This image is compressed and transmitted through the telephone lines. (d) Log-polar image remapped onto Cartesian plane. This image is shown to the user.

The image sensor (which has been fabricated at IMEC) is an active pixels CMOS image sensor [?] (see fig. 3a) The sensitive area has a 8 mm diameter. The array consists of 76 circles. The 56 outer circles (the "retina") have 128 pixels per circle. As the minimum pixels size is 14 μm, the 20 inner circles (the "fovea" shown in fig. 4a) have less pixels: 1 circle with 1 pixel, 1 circle with 4, 1 circle with 8, 2 circles with 16 pixels and 5 circles with 32, and 10 circles with 64 pixels.



(a) Retina plane **(b)** Log-polar plane

Figure 4: a) The "fovea" of the sensor; b) Pixel architecture

The pixels in the array are random addressable by using a circle number/spoke number address. The pixel architecture is shown in fig. 4b. The sensitive element of the pixels is a np diode. These pixels have a logarithmic intensity to voltage conversion. In order to solve the problem of the highly different pixels dimensions, yielding highly different levels of photo current, a scaling of the amplifying transistor is performed along with the scaling of the pixel. However, due to different gain factors of the individual pixels a pixel-by-pixel offset correction is still necessary.

The sensor, which operates on a 5V supply voltage, has 8013 pixels in total that can be readout at a frame rate of more than 100 Hz. It has full digital inputs, an analog output, and an on-chip illumination control facility.

Figure 5: Image acquired with the IBIDEM system re-mapped as they appear to the videophone's user

4.2 The Camera

In order to allow a full testing of the IBIDEM system, a special purpose camera has been designed by Thomson incorporating some analog and digital preprocessing. Mechanically, the camera is divided in two parts: the camera head and the driving unit.

The head is composed of the sensor boards and the chassis intended to be clamped on the pan/tilt device (see figure 2 on top of the monitor).

The hardware part of the camera's driving unit includes several different items:

- a programmable DSP based on a MOTOROLA 56K component, which in conjunction with the surrounding electronics (PROM, E2PROM, RAM) gives the possibility of doing all the real-time processing in software.
- a programmable logic device (EPLD) in charge of all synchronous aspects of the camera.
- an analog front end part in charge of processing the analog output from the sensor (including amplifiers to fit the range of the 12-bit Analog to Digital Converter (ADC), and the analog part of the feedback loop for black level).
- a power supply.

The software part is composed of:

- the real time algorithm in charge of processing the video signal from the ADC.
- The EPLD software (this kind of component looks like a software element as it has to be programmed and simulated).

The last item to be pointed out is the interface between the camera and the compression system. This interface is completely synchronous, and uses a protocol allowing very simple exchange from the camera to the compressor.

4.3 Control and Interface Board

The control and interface subsystem (designed and implemented by Unitek) serves a number of different purposes (see fig. 6).

An important factor underlying the design of both the hardware and the software is the use of the H.324 suite of standards for low bitrate multimedia communication terminals. The choice of this approach is based on analyses of the technical feasibility as well as the exploitation potential. The choice of the H.324 suite of standards has wide implications for the architecture of the terminal as will be seen in the following.

4.3.1 Software Architecture

The overall software architecture of the IBIDEM terminal is shown in figure 7 reflecting the architecture of a standard H.324 low bitrate multimedia communication terminal.

The main difference between the typical H.324 terminal and the IBIDEM terminal is the need for the image remapping block. This block is used to transform the log-polar image to the image that is displayed on the screen. It can also be seen that the image remapping function is really not an integrated part of the *standardized* components of an H.324 terminal.

The components of the software architecture are (following the H.324 standard):

H.263 Video coding for low bitrate communication

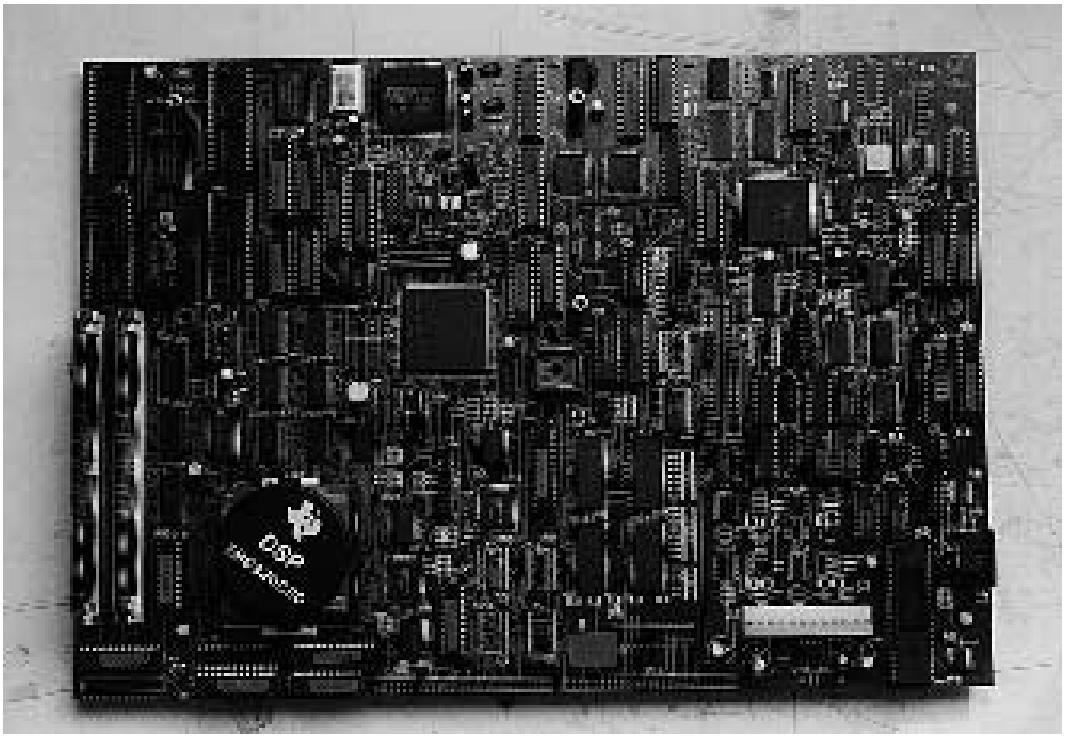


Figure 6: The control and interface board

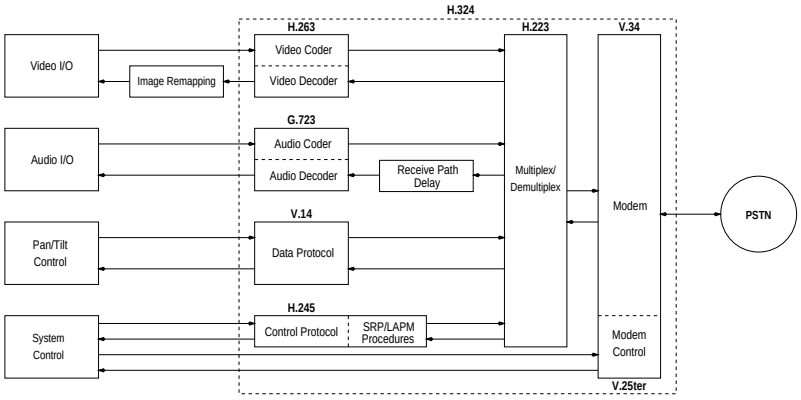


Figure 7: Block diagram of the Control and Interface Module Software Architecture

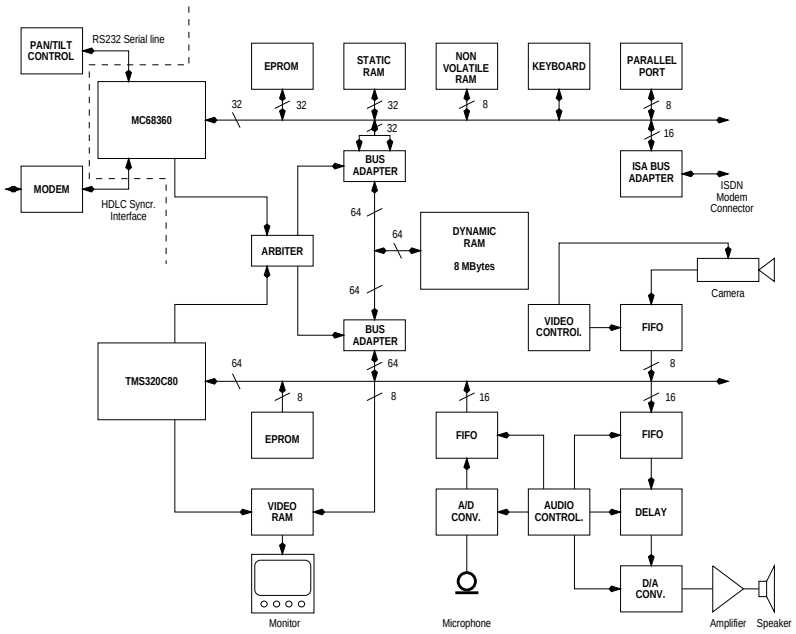


Figure 8: Block Diagram of control and interface module

G.723 Dual rate speech coder for multimedia communicating at 5.3 & 6.3 kbit/s

H.245 Control protocol for multimedia communication

V.14 Transmission of start-stop characters over synchronous bearer channels.

H.223 Multiplexing protocol for low bitrate multimedia communications.

V.34 A modem operating at data signaling rates of up to 28800 bit/s for use on the GSTN and on leased point-to-point 2-wire telephone-type circuits

V.25ter "Modem control" (exact title unknown).

Image Remapper The image remapper is used to convert the incoming decoded "cortical" image to a viewable "remapped" image. This process is taken care of by a relatively simple lookup table mechanism.

4.3.2 Hardware Architecture

The block diagram of the control and interface module can be seen in figure 8. All control operations of the system are taken care of by the Motorola MC68360 microprocessor. In particular the processor takes care of input/output operations (modem, keyboard, and pan/tilt system). The Texas Instruments TMS320C80 parallel DSP takes care of the audio/video compression/decompression and the image re-mapping onto the display.

It has been decided to use the MC68360 as it provides the HDLC protocol embedded in the processor. The HDLC protocol is used in H.324 recommendation for data transfer.

As can be seen from the block diagram, the processor is connected to a 32 bit EPROM/FLASH memory bank, a static 32 bit RAM bank, and a non-volatile memory bank where the system configuration is stored. The video memory is realized with a 1 Mbyte VRAM bank to facilitate

flicker-free operation by employing a double-buffering scheme; the TMS320C80 integrates a video controller that facilitates the interface towards the video RAM generating all the necessary signals and implementing the blanking and synchronizing signals for the monitor.

The pan/tilt unit is controlled by one of the MC68360 serial lines while the keyboard is interfaced using a standard chip for this purpose.

For the audio path the method is basically the same. Output data from the microphone are digitized with an A/D converter with a sampling frequency of 8 KHz and, following the G723 standard, converted into a 16 bit linear PCM data flow. A control circuit stores this data in a 16 bit FIFO memory. The TMS320C80 subsequently copy it into the dynamic RAM for processing and transmission.

In the other direction audio data is coming from the modem. After being processed by the TMS320C80 it is stored into another FIFO memory from where it is retrieved, sent to the D/A converter and subsequently through an amplifier to the loudspeaker. Before the conversion a data delay circuit has been inserted to allow compensation of the de-phasing between audio end video.

4.4 Pan-tilt and Pointing

Because of the peculiarity of the foveated sensor it may be necessary for the user to control the position of the high resolution part of the sensor. This function can be controlled either from the remote site or locally. The role of Scuola Superiore S. Anna in Pisa in IBIDEM was to design and implement an easy-to-use device and its interface to the overall system (see figure 9). The unit consists of a two DOF actuator system. To obtain a simple mechanical solution the pan and the tilt axes are centered on the camera optical axis. The tilt motor is attached at the shaft of the camera's support and allows to rotate it; the pan motor, placed in a fixed position over the camera, is attached at the tilt's support so it rotates both the camera and the tilt motor.

The total weight of the unit (only mechanical part) is about 390 g. As for most video conferencing systems, the frame has been designed for allowing to place the camera immediately above the monitor case in order to minimize the eye-contact angle;

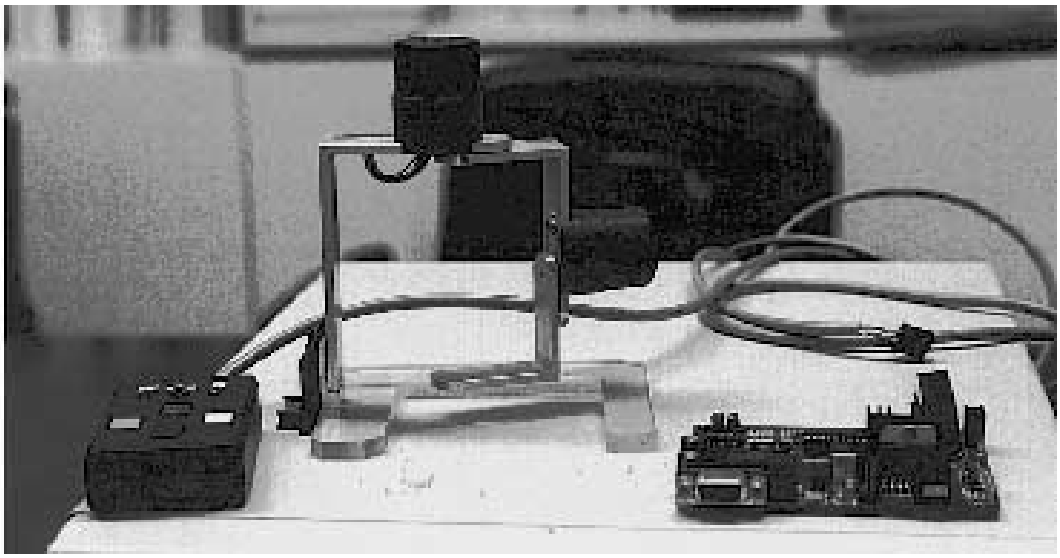


Figure 9: The Pan-tilt unit

5 Prototype Trials

To assess the functional value of the IBIDEM videophone user trials had to be carried out. It was the purpose to get both performance (objective) information and subjective information. In order to get an useful and good insight into the functional value of the IBIDEM videophone it was essential that many people would be trying this videophone and that different communication modes would be evaluated; based on the earlier experiments it was decided to evaluate speechreading and sign language. In the speechreading experiment the number of subjects was 14 and in the sign language experiments 10. Each subject participated in one session and in each session two subjects participated.

Performance (objective) information was obtained by using the Continuous Discourse Tracking (CDT) procedure and the Word Guessing (WG) task. In the Continous Discourse Tracking task the test leader read aloud a short story and the subject had to repeat every word that was said. The Word Guessing task is a question-answer game where a subject should guess at a word that the other subject was given.

The results of the videophone communication CDT task and WG task were compared to the results of the same tasks in face-to-face communication so the face-to-face condition served as base-line. Subjective information was obtained by giving the subjects a Social Conversation (SC) task. After the Social Conversation the subjects had to fill in a questionnaire concerning their opinion of videophone communication, by which the face-to-face communication served as base-line.

The results from the Continuous Discourse Tracking task show that the performances are higher in the face-to-face condition than using the videophone; in the face-to-face condition the mean correct score is about 30% higher (86.87% versus 55.60%) and the number of correctly repeated words per minute is about three times higher (24.39 wpm versus 8.55 wpm).

It also can be concluded that the number of correctly guessed words and the number of turntakings in the word guessing game were approximately the same face-to-face and on the videophone, in both the speechreading and sign language experiment.

The mean percentage misunderstandings was lower in the face-to-face condition than in the videophone condition in the speechreading experiment (12% versus 24%). In the sign language experiment, on the other hand, there was no difference between the mean percentages misunderstandings in the two conditions (3% versus 8%). The performance of the speechreaders was more affected by the videophone than the performance of the signers.

It can be concluded, based on subjective judgement, that there was not much difference between sign language communication in the face-to-face condition versus the videophone condition for social conversation. In the face-to-face condition the signing rate was a little bit slower and the signs were made more distinct and careful, but the communication was rather intact and there were almost no misunderstandings.

The speechreaders and signers were equally positive concerning:

- the frequency of repetitions; all signers and almost all speechreaders said that the frequency was the same or a bit higher;
- the wish to have a videophone; about 40% would like to have a videophone and about 20% perhaps like to have a videophone.

The signers were more positive concerning:

- the overall image quality; 60% of the signers and 37% of the speechreaders judged the quality as 'average';
- the movements of the interlocutor; 60% of the signers and 42% of the speechreaders judged the movements as 'average/good';

- the way, upon which they could follow what the interlocutor signed; 90% of the signers and 68% of the speechreaders judged it as positive ('very easy/easy/average');
- the frequency, in which they have to ask their interlocutor to repeat things; half of all signers and only a quarter of the speechreaders said that the frequency was the same;
- the degree in which the conversation was tiring; 90% of the signers and 79% of the speechreaders judged the conversation as 'not at all tiring/a bit tiring';
- the communication rate; 70% of the signers judged the signing rate as 'normal', while 68% of the speechreaders judged the speaking rate as 'slower';
- the videophone conversation; 60% of the signers said that videophone conversation was the same or a bit different from face-to-face communication, while 68% of the speechreaders said that the two conditions, face-to-face versus videophone, were very different.

6 Conclusions

In general it can be concluded that the IBIDEM videophone prototype is a promising technology; communication by sign language is almost perfect whereas communication by speechreading is averagely functional. It has to be emphasized that in this experiment the speechreading condition was vision-only! It is a fact that many hearing-impaired and deaf people need auditory information; it is very likely that using visual and auditory information would increase the overall performance of the speechreaders!

The results of these user trials correspond with the results of the previous experiments with the simulated IBIDEM images; the performance in the sign language experiment is almost perfect. The difference in performance on the two conditions, namely in the user trial face-to-face versus videophone and in the previous experiments video versus simulated condition, is similar; in the user trials the CDT mean score in the videophone condition was about 64% of score in the face-to-face condition and the mean sentence score in the simulated condition was about 68%. The only difference between the user trial and the experiments is that the mean scores of the former were much higher than those of the latter, namely 86% versus 54%. Probably this is caused by the fact that the test material was different; in the user trials texts were used, while in the previous experiments sentences which had no relation with each other were used.

As a final remark we would like to point out that in its current configuration some of the components of the system are clearly overdimensioned to allow testing of different alternatives. The re-evaluation of the design has already begun and will be completed as soon as possible. Unitek, as the coordinating partner of the consortium, has, in collaboration with the other industries involved, already started actions to commercially exploit the result of IBIDEM. DIST and IMEC are starting a new collaboration to exploit the use of the camera in other fields (such as robotics) and to improve some of the features of the CMOS sensor.

Acknowledgements

The work described here has been supported by the EU under the TIDE initiative Project No. 1038