Design Considerations for a Marker-free Visual-Based Interfacing Device for Telco Operation (April 2007)

Willem Visser, Yuko Roodt and Willem A. Clarke *Member, IEEE*

Abstract—The parts of the system in the telecommunication environment that is used by technicians are sometimes completely menu driven. The interfaces to these parts can be made much simpler. Visual-based interfacing is a relatively new field of interest with advancements being made toward marker free human input tracking. This paper investigates the issues regarding the design of one such a system. Specifically, it looks at the factors concerning the telecommunication environment as well as factors concerning the setup of the camera being used to capture the user's input. It also investigates how shadows being cast by the user's hand against the background, could affect the detection of user input.

Index Terms—camera setup, image processing, shadow effects, telecommunication, VBI, visual-based interfacing.

I. INTRODUCTION

DURING the two weeks that one of the authors spent in the telecommunication environment, working with Telkom technicians, it was observed that many of the interfaces technicians used were more complex than was necessary. Some of the technicians use parts of the system that are completely menu driven, making standard input devices like a keyboard and mouse, cumbersome. By designing an input device with a front-end system that is easy to use and which incorporates added security, can result in the following benefits:

- User friendliness - A technician's profile travels with him/her
- Easier access to technical and real-time information
- It will be a good, on-site training and simulation tool
- By tracking the information that is required on-site, the knowledge gained can drive better training programmes

Visual based interfacing (VBI), also known as Vision Based Interaction, has gained a lot of interest in the past few years, as it does not require any hardware for the user to give input to the computer. The input is obtained through cameras or sensors and the input is in the form of human motion. In the past, markers were used to track the user's input. This made the tracking of movement relatively easy, but the marker-system was interfering as the user had to attach the markers to his/her body every time before the system could be used.

The EyeToy from Logitech that was specifically developed for the Sony Playstation 2 is an example of a VBI device that does not make use of markers. The commercially successful EyeToy is a colour USB camera which is placed on top of or directly below the television so that the user can be seen on the television screen. The EyeToy requires the user to use his/her body as the input device. The elements of the game are augmented over the recorded view. In the EyeToy: Lemmings game, for instance, the user moves his/her arms to form bridges for the lemmings to walk over. A screenshot of this is shown in Fig. 1.

Fig. 1. Screenshot of EyeToy: Lemmings being played.

This paper will describe some of the design considerations for a VBI system. This will consist of the factors that must be considered when being in the telecommunication environment, the setup of the camera as well as the effect that shadows have on the input.

II. A MARKER-FREE VISUAL BASED INTERFACING DEVICE

The proposed VBI system that is being discussed in this paper will specifically be used with menu driven programs. The system consists of six parts namely a video camera, a projector, an RFID tag-reader, a mat, a computational unit and a wall mounted screen. A model of the system is shown in Fig. 2.

Manuscript received April 15, 2007. W. Visser is a student at the University of Johannesburg, Auckland Park, 2006 South Africa (phone: 082-557-2116; e-mail: glasoog@gmail.com).

Yuko Roodt works as software engineer at the Highquest, Auckland Park, 2006 South Africa (phone: 082-452-8442; e-mail: yuko@highquest.co.za).

W. A. Clarke is with the Electrical Engineering Department, University of Johannesburg, Auckland Park, 2006 South Africa (phone: 011-489-2156; e-mail: willemc@uj.ac.za).

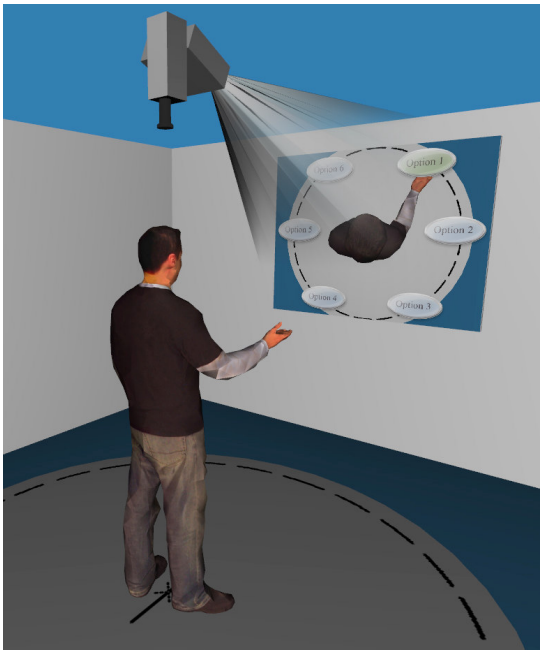


Fig. 2. Design of system.

To use the system, the user must walk onto the middle of the mat. The user is then recorded from above using a video camera that is mounted against the roof. The live video feed is then projected together with overlaid menu options against a white wall or screen. What the user views on the screen can be compared to the face of a grandfather clock. The user is the middle of the clock with his/her arms being the arms of the clock. The menu options are displayed around the user like the numbers of the clock. However, instead of there being twelve options, like that of a clock, there can be any number of options being displayed, up to a certain limit. The user can then choose a menu option by pointing one of his/her arms in the direction of the desired option. The option will then be chosen and the next level of options in the hierarchy will be displayed. The user does not need any computer skills to operate the system.

III. Related Work

Research indicates that different authors have found different ways to obtain user requests. Hansen et al. focused their research on tracking the human eye [1]. They provided an improved likelihood model to cope with major local and global lighting changes and made use of an infrared camera to obtain their input. Manresa-Yee et al. uses a standard web-cam to obtain their input, which is in the form of eye winking detection and nose tracking [2]. By tracking the nose they developed a system that replaces the use of the mouse. What makes it different from the Camera Mouse is that it uses the detection of eye winks to replace mouse button clicks.

Another popular tracking option is the human hand as it is a natural interface device for human beings. Kölsch et al. presented a fast hand tracking method that is robust against indoor and outdoor lighting and dynamic backgrounds and that can be used by different people [3]. Their flock of features method uses KLT (Kanade, Lucas and Tomasi) features, which detects a feature as an area with a steep brightness gradient along at least two directions. They use this together with foreground-background separation where the hand is the foreground. The separation is done by looking at the normalised RGB histogram of the hand area together with a horseshoe-shaped area around the hand and by doing skin colour detection of these two areas. The skin colour of the specific user is learnt as the user uses the system and is not done a priori. The learning algorithm is a variation of that developed by Störring et al. [4].

Skin colour detection is very important in the field of VBI. Research presented by Störring et al. in 2001 offered a robust skin detection algorithm which provided a huge breakthrough in human-computer interaction. They discovered that if an r-g plot is made of all the pixels in the image, where r = R/(R+G+B) and g = G/(R+G+B), then the skin coloured pixels are grouped together in a Skin Area in the plot. They also discovered that the Skin Area is dependent on the Correlated Colour Temperature (CCT) of the light source or combination of light sources. The Skin Area is slightly different for different people and particularly for different races of people, as can be expected. Fig. 3 illustrates these phenomena.

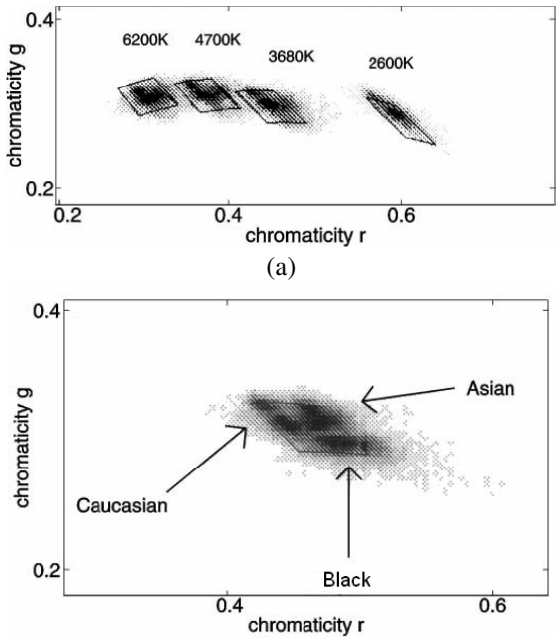
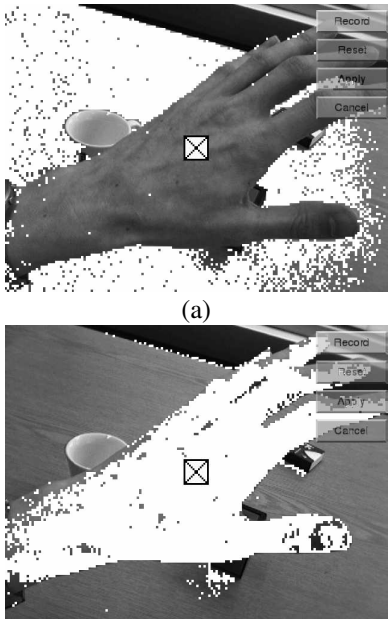


Fig. 3. (a) r-g Plot of Caucasian skin under four different illuminations (the solid line shows the skin colour area, modelled for each candidate); (b) r-g plot of three different races' skin under an illumination with CCT = 3680 K.

Although the Skin Area is different for different candidates and under different light sources, an adaptation algorithm can be used together with the above mentioned theory so that it can be used in a VBI application, as was done by Heidemann et al. in their research for the VAMPIRE project [5]. By recording only four or five images of the user's hand, they can model the skin colour successfully. To show the pixels that fall within the Skin Area, they change the pixel's colour to white. They use a Head Mount Display (HMD) as their output device and two stills from the HMD of the skin detection are shown in Fig. 4.



(a)

(b)

Fig. 4. (a) Untrained skin detector detects wood as skin; (b) Trained skin detector.

Detecting skin colour is very useful when only the hand or face needs to be tracked, but it is sometimes necessary to track the whole body. The final part to be discussed in this section is Human Motion Tracking (HMT).

The direction in which HMT started to move is that of marker-free tracking. Using markers to help with tracking simplifies the job, but is invasive and time-consuming as the user has to apply the markers each time before using the system. Marker-free tracking on the other hand gives the user total freedom of movement. Marker-free tracking is usually model-based and input can be retrieved from one or sometimes multiple cameras [6].

The easiest way to do marker-free tracking is to use a stationary camera and to extract the moving human silhouette from a background image. The background is removed using background subtraction. Devlaeminck did research on human motion tracking which presented a system that is based on Zimmermann and Svoboda's probabilistic estimation of human motion parameters [7]. The system made use of 12 cameras in a room. These cameras were positioned in such a way that the user can be recorded by at least three of the cameras at any given time. The input from the cameras was then used to position a model of the user in a 3D virtual space in such a way that the model mimics the user.

Han et al. used both colour and infrared cameras to better the technique [8]. By incorporating an infrared camera, which operates in the long wave band of 8-12 μm, they recorded a thermal image with pixel values representing temperatures. The thermal camera has the advantage that lighting conditions as well as the colour of the human's clothes and skin are disregarded. Also, the temperature of the subject is usually significantly different from that of the background, making this a robust type of detection.

IV. Telco Environmental Considerations

For the system to work properly in the telecommunication environment, the following factors must be considered:

- Training requirements, i.e. illiteracy
- Access control when sharing the infrastructure with other companies (this will be more important once local loop unbundling has been implemented)
- The actual environment – heat, size, space, dust
- Acceptance of the system by the people that have to use it.

The system must also cater for different users. Here is a list of some differences that should be taken into account:

- Users speaking different languages
- Users having different skin colours
- Users with different job types
- Users operating at different speeds

By considering the above-mentioned issues, a system can be designed that is robust as well as user friendly.

V. Camera Design Considerations

In order for the system that was described in II to work properly with a specific user or group of users, it is important that the setup of the system is done correctly. There are a number of factors that come into play when setting up the system. In Fig. 5 the setup is shown together with the factors that need to be set for the system to work correctly.

In order for the system to work correctly it is important that the user is in the middle of the screen, occupying an area with diameter X_{human}, and that the space around the user X_{frame} is enough so that the option buttons can be displayed. X_{human} and X_{frame} are directly proportional to the angles γ and φ, respectively, which accumulates to the viewing angle of the camera, θ. As the height H_{sh} and width W_{sh} of the user's shoulders are fixed for a specific user, the desired input image can be obtained by either adjusting the height of the camera H_{cam} and keeping the viewing angle of the camera θ fixed, or by changing the viewing angle θ and keeping the height of the camera fixed. This means that the system can be set up in a place with a low or high ceiling, with the ceiling height depending on the viewing angle θ of the camera and visa versa.

A person using the system rarely knows the height and width of his/her shoulders and measuring it can become cumbersome and the accuracy of the measurements questionable. He/she will most probably know their body height H_{human}.

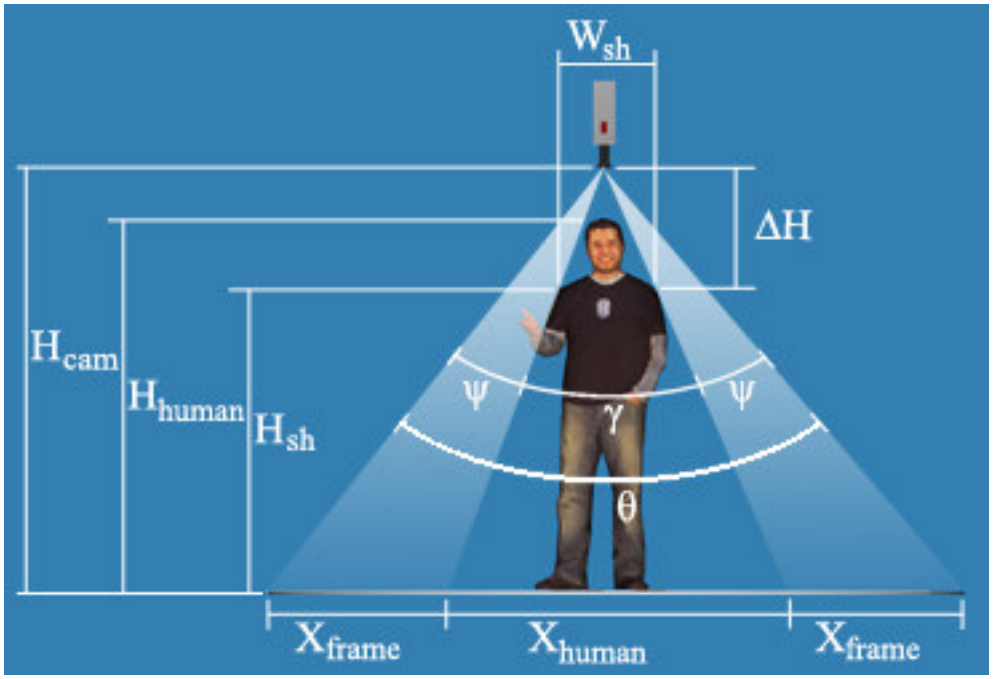


Fig. 5. Camera setup

By making use of the golden ratio φ

$$\varphi = \frac{1+\sqrt{5}}{2} = 1.618 \tag{1}$$

the shoulder width and height can be calculated by only knowing the user's height H_{human} [9]. From Fig. 6 it can be seen that the shoulder width W_{sh} is determined as

$$W_{sh} = \frac{H_{human}}{\varphi^3} \tag{2}$$

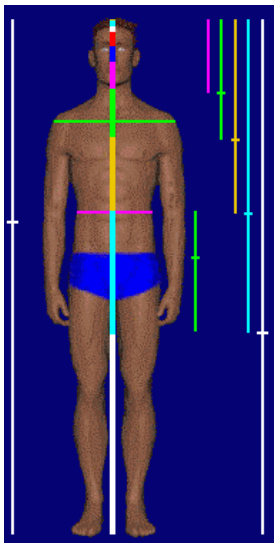


Fig. 6. The golden ratio in human proportions

By using the following important property of the golden ratio

$$\varphi^2 = \varphi + 1 \tag{3}$$

(2) can be simplified to

$$W_{sh} = \frac{H_{human}}{2\varphi+1} \tag{4}$$

The height of the user's shoulder is determined as

$$H_{sh} = H_{human} - \left[H_{head} + \left(\frac{W_{sh} - H_{head}}{\varphi} \right) \right] \tag{5}$$

The user's head height is determined as

$$H_{head} = \frac{W_{sh}}{\varphi} \tag{6}$$

By substituting (6) into (5) and taking (3) into consideration, the height of the user's shoulders is determined by knowing only his/her height as (5) is simplified to

$$H_{sh} = H_{human} \left(\frac{39\varphi+17}{46\varphi+19} \right) \tag{7}$$

To return to the setup of the camera, the user has one of two choices, either he/she can give the specifications of the camera that is to be used, in which case the lowest height H_{cam} the camera must be placed from the floor is determined, or the camera height H_{cam} is given in which case the minimum specifications of the camera is determined.

In both cases the camera specifications are the focal length f and the size of the sensor I as the angle of view θ can be determined from these values [10] as

$$\theta = 2\arctan\left(\frac{I}{2f}\right) \tag{8}$$

Irrespective of the user's height, the system will work when

$$\psi_{\min} = \delta\theta \tag{9}$$

with δ being a predefined fraction.
The angle γ that the user is occupying is then given by

$$\gamma = (1-2\delta)\theta \tag{10}$$

It can also be shown that the user's occupying angle γ can be given by

$$\gamma = 2\arctan\left(\frac{W_{sh}}{2\Delta H}\right) \tag{11}$$

where

$$\Delta H = H_{cam} - H_{sh} \tag{12}$$

By substituting (4), (7) and (12) into (11) the user's occupying angle γ can be given by

$$\gamma = 2\arctan\left(\frac{H_{human}}{2(2\varphi+1)(H_{cam}-H_{sh})}\right) \tag{13}$$

Finally, by substituting (10) into (13), the tie between the viewing angle θ and the camera's height is given as

$$\theta = \frac{2}{(1-2\delta)}\arctan\left(\frac{H_{human}}{2(2\varphi+1)(H_{cam}-H_{sh})}\right) \tag{14}$$

Thus if the user is limited by the height of the ceiling at which the system is to be used, (14) together with (8) can be used to determine the camera's minimum specifications. If, on the other hand, the user has a relatively high ceiling and he/she already has a camera with fixed specifications, the camera height can be determined by rearranging (14) as

$$H_{cam} = \left(\frac{H_{human}}{2(2\varphi+1)\tan\left(\frac{(1-2\delta)\theta}{2}\right)}\right) + H_{sh} \tag{15}$$

VI. Experimental Study of the Effects of Shadows

An easy method of obtaining user input from the input image, is to see if the background over which the option button has been placed, changes dramatically in light intensity. In order for the user to choose the desired input option without his/her shadow making a conflicting option choice, it is important to investigate what the difference is between skin intensities, background intensities and shadow intensities.

The experiment consisted of a 100W light source with 5-level variable light intensity being shone onto a wall from a distance of 2m, the nearest distance a ceiling would be from the floor. A piece of hardboard was placed at a distance of 1.2m from the wall, casting a shadow against the wall. This shadow mimics the shadow that would be cast on the floor by the user's hand. Photos were taken of the wall with the shadow and a hand placed on the wall. Four different hands were photographed at each light intensity level, with two of the hands being Caucasian, one Asian and one Black. This experiment took place in an area with little ambient light. Each time the light reflected from the hand, the wall and the shadow were measured with a light meter. The experiment was repeated in an area with sufficient ambient light. Fig. 7 below contains a Caucasian hand at the five different intensity levels in an area with little ambient light and then in an area with sufficient ambient light. Fig. 8 below contains all five hands at the highest intensity level of the light source in an area with little ambient light and then in an area with sufficient ambient light.

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 7. Caucasian hand, white wall and shadow in area with little ambient light with the light source set to (a) Level 1; (b) Level 2; (c) Level 3; (d) Level 4; (e) Level 5. Caucasian hand, white wall and shadow in area with sufficient ambient light with the sufficient source set to (f) Level 1; (g) Level 2; (h) Level 3; (i) Level 4; (j) Level 5.

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 8. Light source set to level 5 with little ambient light with the hand being (a) Caucasian; (b) Caucasian; (c) Asian; (d) Black. Light source set to level 5 in area with sufficient ambient light with the hand being (e) Caucasian; (f) Caucasian; (g) Asian; (h) Black.

The light intensity readings were tabulated and are presented below in Fig. 9 and 10 in graph form.

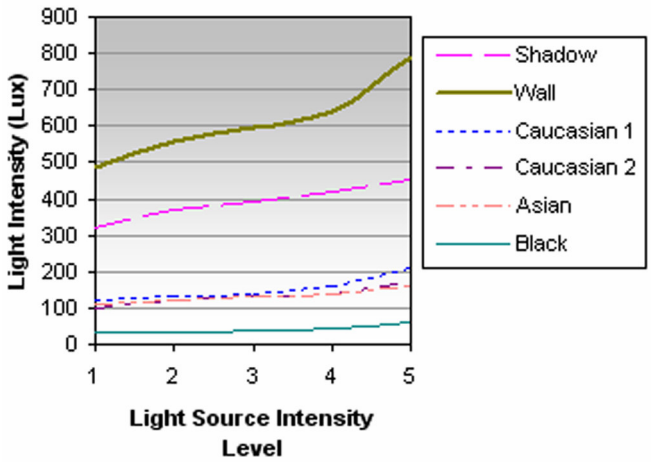


Fig. 9. Light intensities of shadow, wall and skin of different skin types in an area with little ambient light

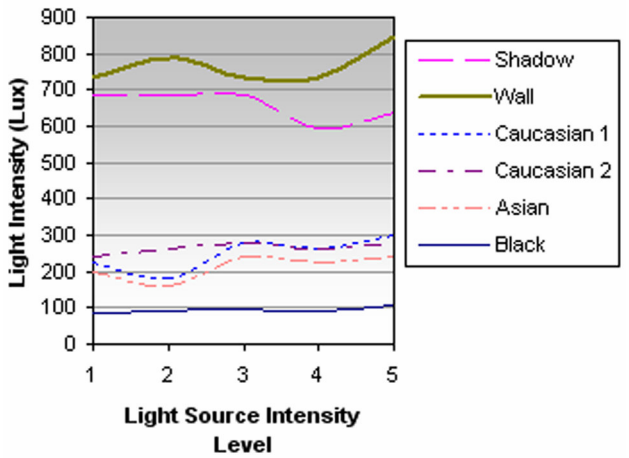


Fig. 10. Light intensities of shadow, wall and skin of different skin types in an area with sufficient ambient light

There are two observations that can be made from Fig. 9 and 10. Firstly that even in an area that is a bit dark, the difference in light intensity between a shadow of a hand and the actual hand is significant. This is due to the fact that the hand casting the shadow on the floor is far away from the floor, relative to the light source intensity. Finally the figures also show that if it happens that shadows do start to interfere, the problem can be resolved by increasing the ambient light.

VII. Conclusion

After investigating the different factors that play a role in the setup of a marker-free VBI device, two design equations were formulated to easily set up the camera that is used for the system.

It has also been shown that the effect of shadows that are cast by the user's hands, is negligible if there is sufficient ambient light where the system is used.

References

[1] D. W. Hansen, R. I. Hammoud, "An improved likelihood model for eye tracking", *Computer Vision and Image Understanding*, 2007.

[2] C. Manresa-Yee, X. Varona, F. J. Perales López, "Towards Hands-Free Interfaces Based on Real-Time Robust Facial Gesture Recognition", *AMDO*, 2006, pp. 504-513.

[3] M. Kölsch, M. Turk, "Fast 2D Hand Tracking with Flocks of Features and Multi-Cue Integration", IEEE Workshop on Real-Time Vision for Human-Computer Interaction (at CVPR), 2004.

[4] M. Störring, H. J. Andersen, E. Granum, "Physics-based Modelling of Human Skin Colour under Mixed Illuminants", *Robotics and Autonomous Systems*, 35(3-4), 2001, pp. 131-142.

[5] G. Heidemann, I. Bax, H. Bekel, "Multimodal Interaction in an Augmented Reality Scenario", *ICMI'04*, 2004, pp. 1-8.

[6] F. Remondino, "Tracking human movements in image space", *Internal technical report at IGP – ETH,* 2001.

[7] R. Devlaeminck, "Human Motion Tracking with Multiple Cameras Using a Probabilistic Framework for Posture Estimation", Masters Degree in Electrical and Computer Engineering, Purdue University, Indiana, 2006.

[8] J. Han, B. Bhanu, "Fusion of Colour and Infrared Video for Moving Human Detection", *Pattern Recognition*, 2006.

[9] G. Meisner, "The Human Body", 2007, http://goldennumber.net/body.htm, Referenced on 13 April 2007.

[10] C. Demant, B. Streicher-Abel, P. Waszkewits, "Industrial Image Processing: Visual Quality Control in Manufacturing". Berlin: Springer-Verlag, 1999, p. 254.

Willem Visser studied B.Eng. in Electrical and Electronic Engineering with endorsement B.Sc. in Information Technology at the University of Johannesburg, Gauteng, South Africa and graduated in 2005. He is currently working towards an M.Eng. with telecommunication as his major field of study.

Yuko Roodt studied B.Sc. in Information Technology at the University of Johannesburg, Gauteng, South Africa and graduated in 2006. He is currently working at Highquest and has an active interest in graphics pipeline programming and advance rendering and simulation techniques.

Willem Clarke heads up the Telkom/SAP Centre of Excellence in Operational Support Systems at the University of Johannesburg. He holds a D.Eng (Electrical and Electronic Engineering).