

THREE-DIMENSIONAL MOTION TRACKING USING STEREO VISION

Sasakthi S. Abeysinghe

Department of Computing
Informatics Institute of Computer Studies, Sri Lanka

Loganathan Krishanthan

Department of Computing
Informatics Institute of Computer Studies, Sri Lanka

ABSTRACT

Today, three-dimensional motion tracking is implemented using magnetic, fibre optic and mechanical techniques that share a common setback: the need for physical contact with the target. Although vision-based techniques provide a contact-free motion tracking solution, they have not been commercially used due to their high resource requirements and code complexity. This paper describes a generic platform that would hide the complexity of vision-based techniques and provide location information via a simple and open protocol. The first step of the solution involves capturing information using multiple image sources, which can be low-cost web-cams or even specialised wide-angle cameras. These image streams are thereafter sent to the Server component that separates the targets from the background using image differencing and a threshold function. Thereafter, a noise reduction algorithm is used to eliminate salt and pepper noise. The flood-fill algorithm is used on the result to identify the borders of each target within each image stream. Finally, the three-dimensional locations of the targets are calculated within the server component using Epipolar geometry. The location information is thereafter sent to software and hardware clients using an open protocol based on the Extensible Markup Language. The RSA encryption algorithm is used in this protocol to ensure the confidentiality of the information being transmitted. Analysis of the developed prototype has demonstrated its practical applicability thereby making vision-based three-dimensional motion tracking more accessible to the commercial and academic worlds.

1. INTRODUCTION

Day-to-day activities of every person, whether it is walking down the street or playing a game of football, involve some type of three-dimensional motion tracking. Objects are seen in three-dimensional space and interacted within a varying degree of complexity.

For humans this is a simple task and is performed sub-consciously, due to the astonishing nature of the brain.

Reproducing this behaviour in computers has been a constant challenge to the academic community, with mixed successes and failures. Notable achievements include creating animated movies using the motion patterns of human actors, tracking the head movement of pilots in military helicopters and even creating robots capable of catching balls thrown at them. Most of these solutions are based on markers, magnetic techniques and mechanical body kits, and not on the more natural vision based techniques.

Reasons for avoiding optical techniques can be found by looking at the complexity of image processing tasks. Similar to information overload in humans, to a computer, a single image is a vast pool of data of which only a minuscule percentage is required for motion tracking and depth perception. In these pools of data, immense processing power and optimised algorithms are required to identify those few key characteristics.

The attention given to alternate technologies have resulted in motion tracking solutions displaying low latency, low processing requirements, high accuracy and low response times making them very attractive for real-time applications [8] [10]. Nevertheless limiting their usage are factors like installation complexity, need for customised hardware and the need of markers or similar devices to be physically attached to the target.

However as technology is evolving at a rapid rate [9], and low-cost high-performance computers are becoming accessible to home users and small businesses, vision-based techniques are now becoming practical and tempting, as they do not display many of the others' shortcomings. This has resulted in a rejuvenated interest in this area [5] [11]. However, the unavailability of a generic platform capable of catering to any requirement has now become one of the strongest limiting factors of this technology, and is the focus of this paper.

The remaining sections of the paper are organised as follows: Section 2 provides more background into three-dimensional motion tracking and gives details on vision-based motion tracking techniques. Section 3 provides an overview of the generic platform suggested by the authors, and goes into detail describing the design of the server component. An extension of this section is presented in Appendix A where the design of two sample client devices is explained in detail. Section 4 describes the implementation of the prototype and the features that it offers. Section 5 makes suggestions for future research while Section 6 provides concluding remarks.

2. MOTION TRACKING TECHNIQUES

Many methods of motion capture are used, and are classified into three fields (inside-in, inside-out and outside-in) based on how the sensors and sources are placed on and around the targets [1].

Mechanical motion tracking systems mainly categorised as inside-in solutions have been widely used as mechanised body kits. These body kits offer high accuracy and low response times, but restrict the movements of the target. Recent advancements in technology have resulted in larger capture areas and six degrees of freedom [8]. Optical fibre based motion-tracking systems, being very similar to the mechanical methods, are blessed with the same strengths and plagued with the same weaknesses. This technology is best used not to monitor the movement of a target, but rather, to monitor the movement of its joints [1].

Acoustic motion tracking systems monitor motion similar to how a bat navigates using echolocation. The outside-in nature of these systems gives them a wide freedom of movement and can therefore be used in large-scale solutions. However, when considering the practical implementation of such a system, the presence of acoustic noise results in erroneous readings [3]. Magnetic motion tracking systems categorised as inside-out or outside-in, are the most successful and widely used solutions in the industry. The reason being their ability to provide information with the accuracy and responsiveness of mechanical systems, but with much less restriction of movement [1]. Optical motion tracking systems, also an outside-in technology, display a similarity to acoustic systems. Although capable of providing reasonably accurate information covering large areas, these systems are considered unattractive due to their resource-intensive nature and sensitivity to illumination levels [5].

Optical motion tracking systems are based on analysing image streams and using a mathematical framework to triangulate the three-dimensional location of the targets. To better analyse images and better understand their information many image-processing techniques have been suggested and are in use.

Discussing the image processing routines used: First, the image is captured and stored in a digital form. This is done by means of sampling which is described in popular theories such as the Shannon sampling theorem [7] [15]. This sampled image is thereafter stored in greyscale or in colour using Red, Green and Blue elements. The next step is separating the background from the foreground. This can be done by image differencing together with a threshold function. Noise patterns that are left over in the image can then be removed using noise removal heuristics [6] [13]. Next, the two-dimensional locations of the targets can be identified by using an edge-detection algorithm [6] [13].

Henceforth, three-dimensional positioning of the targets can take place, the means of which is described by a mathematical machinery named Epipolar Geometry [4] [13]. This model describes the calibration of cameras and thereafter uses that information together with the information in the images to triangulate the three-dimensional location of the targets.

Vision based motion tracking solutions are not without their share of challenges:

- A. *Dependency on light:* All vision-based solutions are dependant on adequate lighting conditions. Although this can be controlled within enclosed areas such as factory environments, controlling lighting levels in large open areas is not a practical task. This problem can be avoided to an extent by using contrast-stretching techniques [6] or by using infrared image capturing devices such as night-vision cameras.
- B. *Need for Line of Sight:* Line of sight is another essential requirement for vision-based techniques, dictating that the targets are within the line-of-sight of the image capturing devices. Nevertheless, these targets tend to be obstructed by objects in the background and even themselves. Although difficult to be solved completely, this problem can be avoided to a certain level by strategically placing many image-capturing devices so that they can compensate for each other.
- C. *High processing power requirement:* Vision based motion tracking techniques demand a large amount of processing power to perform their operations. Most personal computers today still do not have adequate power to perform these tasks at high frame rates and high resolutions. Therefore, optimised algorithms together with low-level vector instructions should be utilised to obtain the maximum performance for the available processing power.

3. DESIGNING A GENERIC PLATFORM

As mentioned, vision-based motion tracking is restricted by its high processing power requirements and connections to image capturing devices. Nevertheless, most practical situations demand small and mobile devices that can react to the three-dimensional movements of the targets. Therefore, a means should be made available for these small and relatively simple devices to be able to track the motions of targets in three-dimensional space.

A client-server based model is suggested where the server equipped with adequate hardware resources performs the image capturing, image processing and target positioning tasks, and makes available that information to client devices via a simple and open protocol. The choice of the communications medium should be made considering the effort required building client devices to communicate in that medium. The medium of choice would be any TCP/IP based network due to its wide reach through the internet, and because WiFi (IEEE 802.11) and similar technologies extend TCP/IP based networks into the wireless domain [12], making it ideal for small portable devices.

Figure 3.1 graphically describes how the image streams are sent from the capturing devices to the Server, which identifies the location information of the target objects and conveys that information to the registered client devices via a TCP/IP based network.

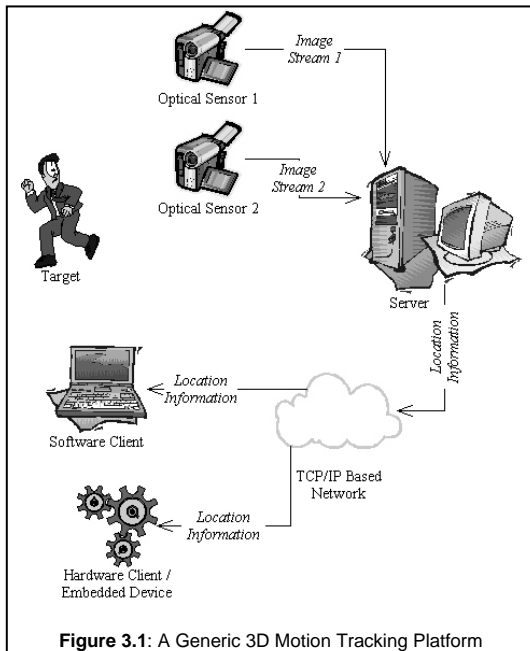


Figure 3.1: A Generic 3D Motion Tracking Platform

A modular approach was preferred, as it would provide extensibility allowing the solution to grow in both its technical capabilities as well as its functional capabilities. The initial proposal contains four main modules, of which two modules are part of the server, and the other two modules are sample client devices.

The server modules will be described in detail in the following sections. (Details on two sample client modules can be found in Appendix A)

3.1. Server – 3D Positioning module:

The 3D positioning module is responsible for locating targets in each image stream, tracking their movement, and calculating their location in three-dimensional space.

As Figure 3.2 describes, the cameras are first initialised, where their intrinsic and extrinsic parameters are identified. Thereafter, a series of snapshots of the background are taken which are merged to form an averaged background (Formulae 3.1). The multiple images are taken to identify non-stationary objects in the background (Ex: a pendulum) and to merge their movement.

$$A_p = \frac{\sum_{i=1}^n (B_i)_p}{n}$$

$\forall p,$
 B – Background Snapshot
 A – Averaged Background
 n – The number of snapshots
 p – Pixel

Formulae 3.1: Averaging Backgrounds

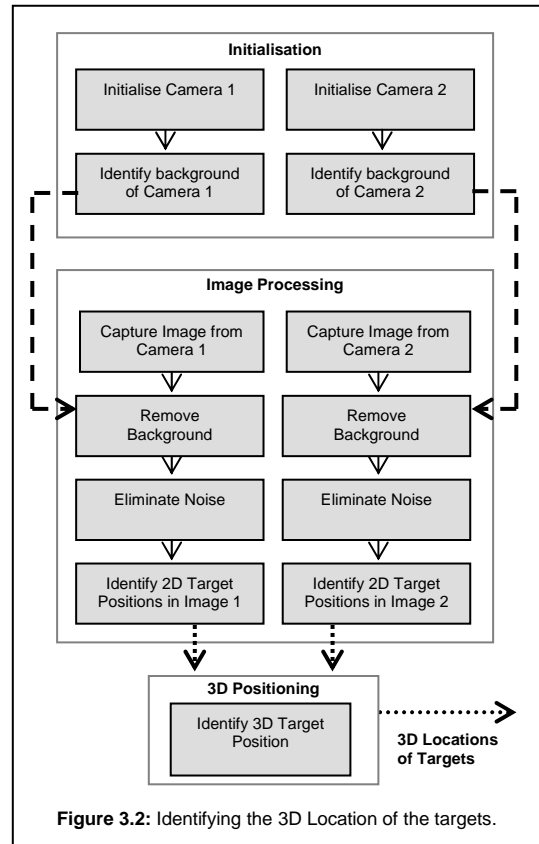


Figure 3.2: Identifying the 3D Location of the targets.

In the Image processing sub-module, when each snapshot is taken, it is compared against the composite background using a threshold function (Formulae 3.2). This is used to identify which pixels belong to the

background and which pixels belong to potential targets.

$$I_p = \begin{cases} S_p, & |S_p - A_p| > T \\ 0, & |S_p - A_p| \leq T \end{cases}$$

I – Output image
 S – Snapshot
 A – Averaged Background
 T – Threshold value
 p – Pixel

Formulae 3.2: Identifying potential target pixels

Although the aforesaid process is reasonably accurate in differentiating between target and background pixels, it is nevertheless hampered by the presence of noise. Any standard noise removal filter can be used to identify the noise pixels and to classify them either as background pixels or as target pixels.

After target pixels are identified, the two-dimensional perimeters of the targets need to be found. This can be done using any standard edge-detection algorithm. Nevertheless, we suggest the use of the flood fill algorithm, as it provides an easy way of identifying multiple targets that are located within the same bounding rectangle. The Flood-Fill algorithm also provides much better performance than the standard matrix based edge detection algorithms, Nevertheless the flood-fill instructions cannot be converted into vector instructions, and therefore in the presence of a vector processor, using a matrix-based algorithm might offer better performance.

$$x = \frac{x_1(2u_{2h} - W_h) - x_2(2u_{1h} - W_h)}{2(u_{2h} - u_{1h})}$$

$$z = z_c + \frac{2W_h(x - x_2)}{2(u_{2h} - W_h) \tan\left(\frac{\alpha_h}{2}\right)}$$

$$y = y_c + \frac{(W_v - 2u_{2v})(z - z_c) \tan\left(\frac{\alpha_h}{2}\right)}{W_h}$$

(x, y, z) – Centre position of the target
 (x₁, y_c, z_c) – Focus of the 1st camera
 (x₂, y_c, z_c) – Focus of the 2nd camera
 W – Width of the cameras in pixels
 u – Pixel position of the target
 α – Angle of vision
 h/v – Horizontal / Vertical
 1/2 – 1st Camera / 2nd Camera

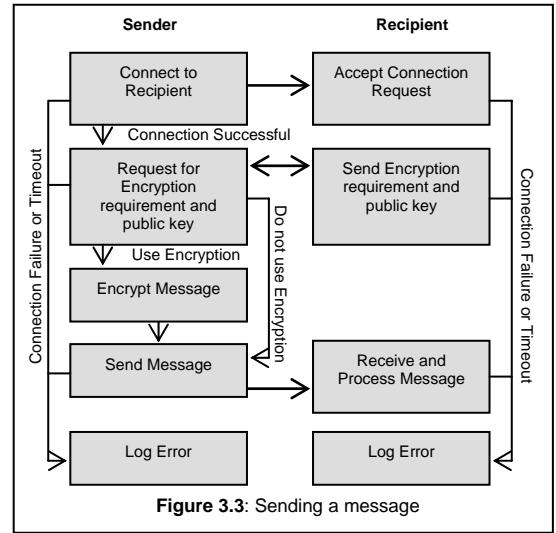
Formulae 3.3: Calculating the 3D Location using the Canonical Configuration

After applying the aforesaid routines to the snapshot the relative positions of the targets are available for both cameras. Thereafter this information is applied to the mathematical machinery known as Epipolar geometry [4] [13] to obtain the three-dimensional coordinates of the targets. Formulae 3.3 describes the

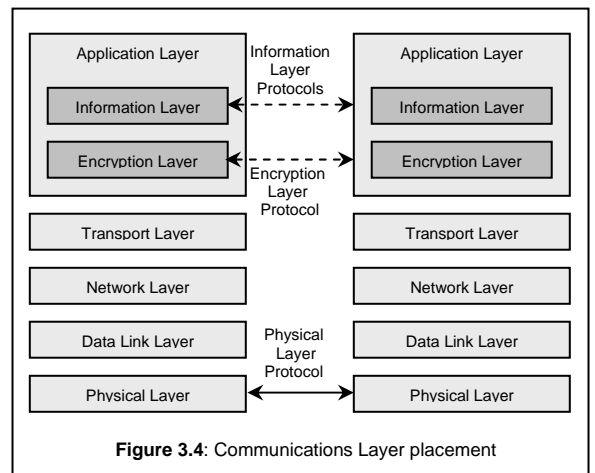
three-dimensional Cartesian coordinates of a target using the canonical configuration of Epipolar geometry, where the cameras are placed on the same z and y axes, and are parallel to each other.

3.2. Server – Communications Module:

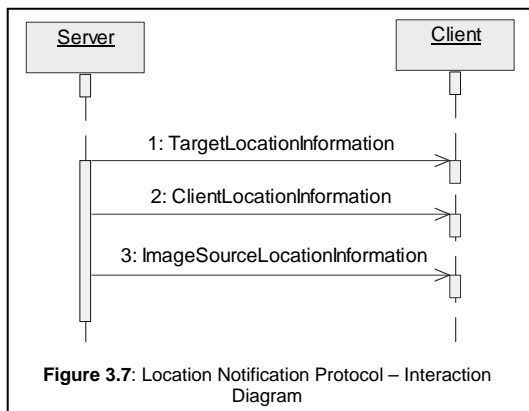
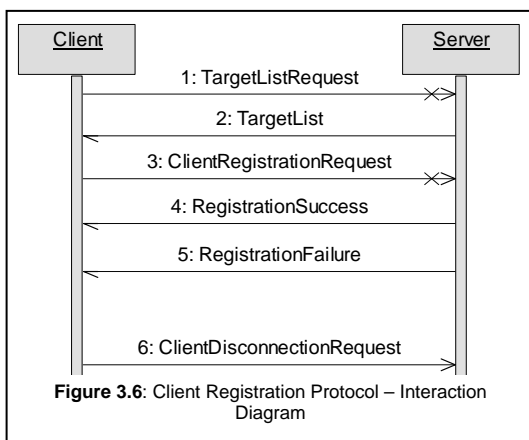
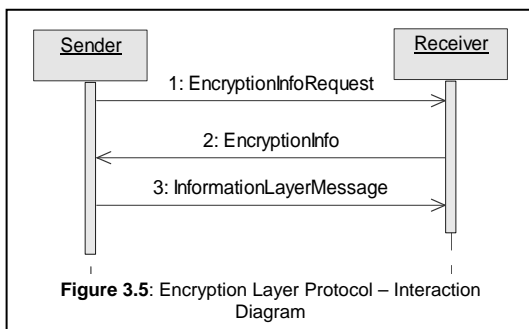
The Communications Module is responsible for the distribution of the location information to the client devices. Figure 3.3 describes the events that take place when sending a message between two devices. Further enhancements into this communications module can include a cross-protocol communications sub-module, and a data-compression sub-module. These can be used to increase the availability of the information provided by the server, and will result in better utilisation of the communications medium.



The communications protocol as mentioned above need to be simple, open and extensible, making it easy for customised client devices to tap into the information provided by the server. To achieve that end it is suggested using the following XML based protocol.



As shown in Figure 3.4 the communications protocols operate in the Applications layer in a TCP/IP layer structure [14]. The applications layer is divided into two layers, namely the Encryption Layer and the Information Layer. The Encryption Layer Protocol is graphically described in Figure 3.5, and the two Information layer protocols, the Client Registration protocol and the Location Notification protocol are graphically described using Figure 3.6 and Figure 3.7 respectively. Further information including a W3 Schema Definition can be found in [2].



4. IMPLEMENTATION OF A PROTOTYPE

The prototype was developed using Borland Delphi 7.0 due to its high performance, ease of use, and wide array of open source units that together, provide a good

foundation for Rapid Application Development while generating optimised, high-performance code.

The prototype displayed the following features, and was used to demonstrate the practical applicability of such a generic platform. It was also used to display the ease of creating customised clients capable of utilising the services provided by the server.

- A. Ability to track the movement of multiple targets simultaneously in three-dimensional space in real-time.
- B. Reliance on low-cost optical sensor based inputs (Web cams).
- C. Provide information to external client devices so that they can react to the tracked movement.
 - a. The prototype hardware device was a motor controller capable of rotating an object (i.e.: spot light) to always point at a target.
 - b. The prototype software application was a motion-monitoring tool that monitored the movements of multiple targets within an area, and displayed it in a virtual 3D world.
- D. Communicate with client devices using secure communication techniques so that unauthorised third parties cannot read the information being transmitted. The RSA encryption algorithm was used to encrypt the messages due to its inherent solution to the key exchange problem.
- E. Able to act as a generic platform that can be used to create motion-tracking solutions for any commercial or academic requirement.
- F. Easily deployable in any environment with suitable lighting and networking infrastructure.

During the implementation, many challenges were faced. One was the large amount of processing power required to perform the image processing tasks. This was overcome by means of optimised algorithms performing image-processing tasks only on the areas required, and by skipping frames to avoid a cumulative delay that can appear when processing all available frames.

Another challenge faced was the occlusion of objects. This was solved to an extent using a heuristic rule, which assumed that objects continued their movement in the same direction even after they were no longer visible. Attempts at solving this can be made by incorporating an object recognition module based on either Artificial Neural Networks or other image processing algorithms.

5. FUTURE DEVELOPMENTS

Evolution is how the humans came into being, and should be the same with any solution that aims at perfecting itself. During the course of the project, we were presented with many ideas, some of which were incorporated, and others left in the “To-do” list due to time constraints. Implementing these would result in a better three-dimensional motion tracking solution capable of adding value to the final product.

- A. Add an object recognition module implemented capable of recognising objects that come into the field of vision. This can be used to solve the occlusion problem as well as provide automated target acquisition.
- B. Investigate the feasibility of performing edge detection or contrast stretching on the snapshots, thereby making the solution more resilient to lighting level changes in the background.
- C. Create a Cross-Protocol communication module enabling the proposed solution to be used with existing motion tracking infrastructure.
- D. Introduce data compression into the communications protocol, thereby reducing the network bandwidth utilisation.
- E. Use the general configuration of Epipolar geometry for three-dimensional depth perception, and thereby allow the use of more than two cameras to detect three-dimensional motion. This will improve the accuracy of the solution, and can be used to solve the occlusion problem. It will also enable the movement of the image sources giving a larger field of vision.
- F. Integrate other motion tracking techniques (magnetic, acoustic, mechanical) to the proposed solution thus obtaining better accuracy without being penalised with movement restrictions.
- G. Enable multiple Servers to communicate with each other when observing the same targets from different perspectives. This information can be used to increase accuracy and can be extended to incorporate dynamic load balancing.

6. CONCLUSION

Most three-dimensional motion tracking solutions available today are based on magnetic and mechanical techniques that need to be in physical contact with the target being tracked. This can restrict the movements of the target as well as alter their natural behavioural patterns. Vision based motion tracking provides a solution for these restrictions, but is underutilised due to its high performance requirements and dependency to light.

To address the issues mentioned above, a generic platform was suggested which discussed a means of isolating the processor intensive image processing tasks within a server, and distributing the location information using a simple, open and extensible protocol to any client device capable of tapping into the communications medium. Furthermore, a design was proposed for the server that supports a modular architecture enabling new functionality to be integrated with only a few modifications to the existing solution. These features together form a powerful platform capable of making three-dimensional motion tracking using vision based techniques more accessible to the commercial and academic worlds.

REFERENCES

- [1] 4th Wave Inc. (2001, Aug. 08). Motion Tracking Tutorial. [Online] 4th Wave Inc. <<http://www.wave-report.com/tutorials/MoTrak.htm>> [2003, Sept. 15]
- [2] Abeysinghe S.S. (2004), MoTrack 3D – Three Dimensional Motion Tracking using Stereo Vision, *Final Year Project Thesis*, Informatics Institute of Computer Studies, Sri Lanka.
- [3] APR Inc. (2002). [Online]. Acoustic Positioning Research Inc. <<http://www.positioning-research.com>> [2003, Oct, 24]
- [4] Fisher, B. (1997), Epipolar Geometry. [Online]. University of Edinburgh. <http://www.dai.ed.ac.uk/CVonline/LOCAL_COPIES/EPSRC_SSAZ/node18.html> [2003, Oct 24]
- [5] Goncalves, L. (1996). Monocular tracking of the Human Arm in 3D [Online]. California Institute of Technology. <http://www.vision.caltech.edu/luis/luis_old/luis.html> [2003, Oct 19]
- [6] Gonzalez R.C. and Woods R.E. (2003). *Digital Image Processing 2nd ed*: Pearson Education.
- [7] Hearn D. and Baker P. (2002). *Computer Graphics 2nd ed*. India: Prentice Hall.
- [8] Meta Motion (2000). Mechanical motion capture – Gypsy 3 Motion capture system [Online] Meta Motion, Inc. <<http://www.metamotion.com/gypsy/gypsy-motion-capture-system.htm>> [2003, Oct 18]
- [9] Moore, G.E. (1965). ‘Cramming more components onto integrated circuits’, *Electronics*, Vol 38, No 8
- [10] Polhemus, Inc. three-dimensional scanning, position/orientation tracking systems, eye and head. (2003). [Online] Polhemus, Inc. <<http://www.polhemus.com>> [2003, Sept 18].
- [11] Sanders-Reed, J.N. (1996) Vehicle Real-time Attitude Estimation System [Online], SVS R&D Systems Inc, <<http://www.swcp.com/~spsvs/VisualFusion/VraesSpie.pdf>> [2003, Sept 10]
- [12] Smith W. (2003). How WiFi Works: all the basics on 802.11a, 802.11b and 802.11g. [Online] Wireless-Network-Guide.com <<http://www.wireless-network-guide.com/wi-fi-basics.php>> [2003, Nov 18]

- [13] Sonka, M., Hlavac, V., and Boyle R. (1999), *Image Processing, Analysis and Machine Vision*, 2nd ed., India: Vikas Publishing House.
- [14] Stallings, W. (2000). *Data and Computer Communications*. 6th ed. India: Pearson Education
- [15] Wikipedia Encyclopaedia (2003) Nyquist – Shannon sampling theorem. [Online]. Wikipedia Encyclopaedia. <http://en.wikipedia.org/wiki/Nyquist-Shannon_sampling_theorem> [2003, Oct 26]

APPENDIX A

A.1. Sample Software Client

The software client module is responsible for acquiring location information of targets, clients and image sources from the Server module and displaying that information to a user in the form of a virtual three-dimensional world.

As Figure A.1 describes, all three messages of the Location Notification Protocol are used to obtain information on the world objects. Upon the receipt of these messages, the coordinates are converted from the global coordinates provided by the server module to the virtual coordinates used by the software client. This information is modelled in the 3D world, rendered into to a 2D scene and displayed to the users.

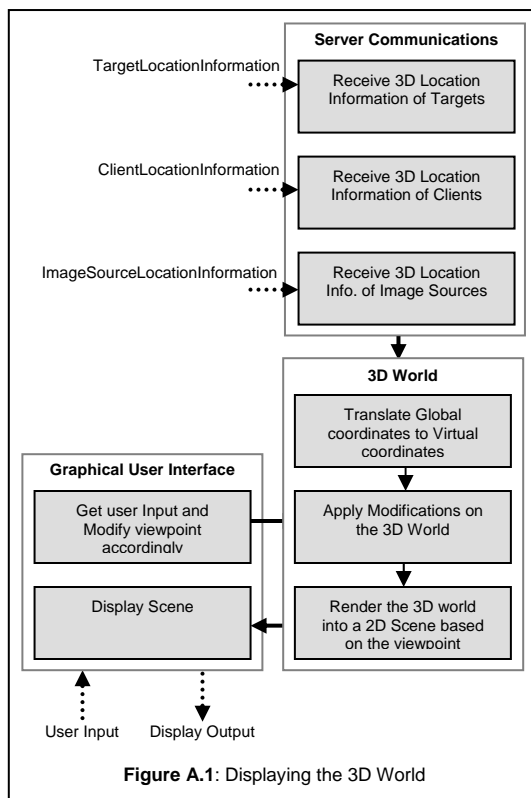


Figure A.1: Displaying the 3D World

The users can also modify the viewpoint of the 3D scene using the Graphical User Interface, thereby

enabling them to navigate in the virtual three-dimensional world.

A.2. Sample Hardware Client

In a commercial environment when using a hardware client to track the movement of a target, it should be directly pluggable to the TCP/IP network. However, as a proof of concept, the hardware client was created using the computer as an intermediary to connect to the communications medium.

As described in Figure A.2, the client computer receives location information from the server using the TargetLocationInformation message of the Location Notification protocol. Thereafter, it calculates motor rotational angles, translates it into motor driver signals, and sends those signals to the motor driver via the parallel port. The motor driver receives the input as two binary coded decimals instructing it to move its two motors to the specified states. The implementation of this motor controller circuit can be simplified by looking at the unipolar stepper motors as state machines, each with four states, able to navigate from one state to the other. The circuit receives information specifying to which state the motor must move next, and thereafter converts that information into pulses that are sent to the motors using four transistors in the switching mode.

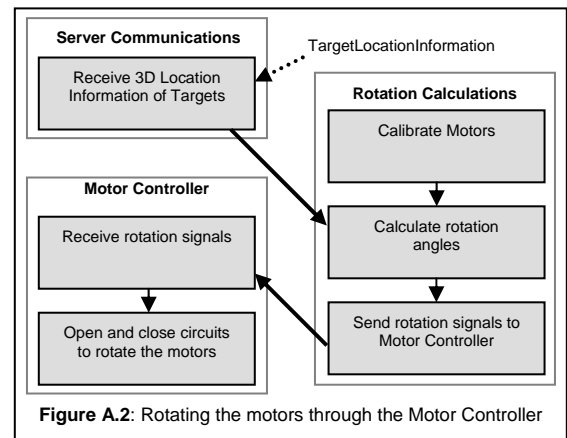


Figure A.2: Rotating the motors through the Motor Controller

Analysing Figure A.3 in detail, the 7805 acts as a voltage regulator, converting the unregulated 15V input into a regulated 5V output. The 470µF, 1000µF and 0.1µF capacitors give further stability to the output voltage by filling in the current if the input supply drops below 8V.

The 7407 buffer IC and the 7404 NOT gate IC both act as protection to the computer by propagating only the voltage across. This results in a minimum current to be drawn from the parallel port, thereby protecting it from the power drain of the motors. The 7404 NOT gate IC and the 7408 AND gate ICs convert the 2-bit binary

