Optical Stereoscopy Motion Capture System 2014

Trahern D. Griffiths*

University of New South Wales at the Australian Defence Force Academy.

Thesis Summary Report

The aim of this project is to represent a human like figure in 3D space in order to conduct motion analysis on a human subject. This document outlines the design of a motion capture system that will be used on a kayak in order to analyse the motion techniques of a kayaker. The proposed design of the system was to use GoPro cameras that were to be mounted to the kayak, in order to capture the technique of the subject. The GoPro system was successful at capturing the motion of a subject and representing the data in 3D space. Seen in the results obtained from testing against the Vicon commercial motion capture system, which was used as the control. The response from the GoPro system followed the same response that the Vicon system produced, concluding that the motion capture from the GoPro system was accurate. This was achieved by successfully: segmenting markers, 3D representation, marker tracking and depth estimation.

Contents

*ZEIT4501 Electrical Engineering Project Thesis and Practical Experience

Nomenclature

Symbol	Unit	Description
z	Pixels	Depth of the marker from the camera baseline
d_x	Pixels	Disparity between left and right image planes
m	No units	Gradient of interpreted disparity vs distance relationship
t_x	Metres	Horizontal baseline shift
f_x	Metres	Focal length of camera
$dist$	Metres	Distance of disparity vs distance relationship

I. Introduction

This paper outlines the design of a motion capture system that will be used on a kayak in order to analyse the motion techniques of a kayaker. Motion capture systems are expensive and require sophisticated hardware and software; the design of this system is to be relatively inexpensive; with the ability to be used on the water. The hardware for this system will consist of GoPro cameras and pink polystyrene balls, with the post-processing language of MATLAB. The aim of this project is to represent a human like figure in 3D space in order to be able to conduct motion analysis on a human subject that is on a kayak.

II. Background

II.A. Motion Capture

Motion capture involves measuring an object's position and orientation in physical space, then recording that information in a computer-usable form. Objects of interest include human and non-human bodies, facial expressions, camera or light positions, and other elements in a scene.[2]

Motion capture has many uses in today's science, medicine, sports, and computer graphics. The use of motion capture allows for the data that is record into a usable form to be manipulated and controlled by the user. [1]

Over the years motion capture has developed different forms, that have their own strengths and weaknesses: Optical, Magnetic, and Mechanical; these will be discussed in detail.

Another important question is why use motion capture? Motion capture allows for the motion data of certain points in time and space of the interested subject to be acquired in order to gather information on some of the parameters of the subject, such as speed, direction, size etc.[3] These parameters can be calculated or even control a device. The application of these parameters could be used for: motion analysis, sports analysis, biomechanics, or mimic the motion, tele-surgery, motion feedback control and virtual training. [1]

The first use of motion capture was achieved with cameras taking a sequence of images of the subject of interest. The images were then analysed to determine a 2D motion of certain points of interest on the subject. This is still being used today, however the use of multiple cameras taking images simultaneously advanced the technology into 3D and with the use of computer software, the images could be analysed by computer instead of a person. This then lead to the difficulty of recognising points of interest on the subject.[2] This is why markers or sensors were used to achieve the accuracy in recognising points of interest. Magnetic and mechanical motion capture makes use of sensors, while optical makes use of markers either passive (don't produce their own light, reflective markers) or active (produce their own light, LEDs).[4]

II.B. Types of Motion Capture

II.B.1. Magnetic

These systems use sensors to accurately measure the magnetic field created from the source. Magnetic systems are real time and can provide between 15-120 samples per second (this is dependent on the model and how many sensors). A typical magnetic motion capture system has one or more electronic control units into which the source(s) and sensors are cabled.[2]

Advantages[3]

- It allows for fewer sampling locations and less interred information in order to find the position and orientation information.
- There is minimal device calibration; these systems measure distance and rotation in relation to a single object, the source.
- This system is capable of real time processing, therefore the method allows for an interactive display and verification of captured data.

Disadvantages[3]

- There is a sensitivity to metal, which means that care must be taken no metal is in the vicinity of the subject and the devices.
- The range is limited, maximum effective range is much less than that of its optical counterpart, even though at greater ranges the optical systems lose accuracy.

II.B.2. Optical

Optical motion capture systems are based on high contrast video imaging of retro-reflective markers which are attached to the object whose motion is being recorded.[3]

The markers are imaged by high-speed digital cameras. The number of cameras used depends on the type of motion capture. Facial motion capture usually uses one camera, sometimes two. Full body motion capture may use four to six (or more) cameras to provide full coverage of the active area.[5] To enhance contrast, each camera is equipped with infrared- (IR) emitting LEDs and IR (pass) filters are placed over the camera lens. The cameras are attached to controller cards, typically in a PC chassis.[3]

After a motion capture session, the recorded motion data must be post-processed or tracked. The centroids of the marker images (either computed then, or recalled from disk) are matched in images from pairs of cameras, using a triangulation approach to compute the marker positions in 3D space. Each marker's position from frame to frame is then identified. Several problems can occur in the tracking process, including marker swapping, missing or noisy data, and false reflections.[5]

Advantages[3]

- Depending on the system used and the precision required, the motion capture area can be arbitrarily large.
- The subject is not physically attached to the motion capture system. This allows for the long in-run (for the subject to get up to speed) and out-run (for the subject to slow down) areas required for full-speed running motion.
- Since the markers are active elements of the motion capture system, additional markers cost very little. Theoretically, hundreds of markers could be included in a given scene. However, given the problems of occlusion and the limitations of the tracking software, the practical maximum is probably less.

Disadvantages[3]

- Since current optical systems are contrast based, backgrounds, clothing, and ambient illumination may all be issues.

- Wet or shiny surfaces (mirrors, floors, jewellery, and so on) can cause false marker readings.
- Since a marker must be seen by at least two cameras (for 3D data), the total or partial occlusion caused by the subject, props, floor mats, or other markers, can result in lost, noisy, displaced, or swapped markers. Common occlusions are hand versus hip (standing), elbow versus hip (crouched) or hand versus prop in hand or opposite hand.

II.B.3. Mechanical

Subject wears a human-shaped set of straight metal pieces (like a very basic skeleton) that is hooked onto the subject's back. As the performer moves, this exoskeleton is forced to move as well and sensors in each joint detect the rotations. [2]

Advantages[3]

- No interference from light or magnetic fields.

Disadvantages[3]

- The technology has no awareness of ground, so there can be no jumping, plus feet data tends to slide.
- Equipment must be calibrated often.
- Unless there is some other type of sensor in place, it does not know which way the performer's body is pointing.
- Absolute positions are not known but are calculated from the rotations.

II.C. Stereoscopy

Stereoscopy is a technique used for creating or enhancing the illusion of depth in an image. Most stereoscopic methods present two offset images separately to the viewer's left and right eye. These different two dimensional (2D) images are then combined in the brain to give the perception of depth, creating three dimensions. Human vision uses the following cues in order to perceive depth;[10]

- Stereopsis
- Accommodation of the eye
- Overlapping of one object by another
- Change in size of textured pattern detail
- Haze, desaturation and a shift of bluishness
- Vertical position (objects higher in the scene are perceived as further away)
- Linear perspective
- Subtended visual angles of an object of known size

[8]

All of these cues except stereopsis and accommodation of the eye are present in 2D images.[7] By presenting a slightly different image of the scene to the left and the right eye, the brain combines both these 2D offset images and creates the 3D of depth.There are many different methods that can be used to create a 3D illusion for the viewer:[10]

- **Side-by-side.** Enhancing the depth perception by providing the brain with two slightly different images of the same object.
- **3D viewers.** Active and passive viewers use filters to provide constant streams of binocular inputs to the appropriate eye.

- **Autostereoscopy.** Optical components in the display, that splits the image up directionally to the viewer's eyes.

[7]

II.C.1. Stereo photography techniques

In order to create the stereoscopic image two photos must be taken from different horizontal positions. This can be achieved by having two separate cameras side-by-side, creating the left and right eye image planes. Both these cameras must be in the same plane so that the correct stereoscopic image can be captured.[9]

In general purposes the goal stereo photography is to duplicate the natural human vision, and in this case the correct baseline (the distance between where the left and right images were taken) should be the same as the distance between the left and right eye. Most stereo photography makes use of 50-80mm baseline, and is referred to as "normal".[10] However there may be situations when it would be desirable to use either longer or shorter baselines.

To produce a dimension of depth from two camera views there are three common approaches; depth synthesis from stereo matching, active depth sensing, and 2D-3D conversion from a single view.[10]

II.C.2. Estimation of Depth

Estimating the depth of an object given the two different views from the cameras can be explained geometrically, seen in Figure 1. Consider the simple case when two cameras horizontal planes are in same plane as the image plane with only a horizontal shift between the two cameras.[6] Here this setup is similar to that of human eyes, left and right in the same plane, looking at the same object from to different views. Therefore when the image from the left camera is directly compared to the image from the right, the point of interest appears in different locations. The relationship to estimate the depth is not quiet simple, the disparity of the position of the point of interest between the cameras d_x, the depth of the point in this scene z, the focal length of the camera f_x and the horizontal baseline shift between the cameras t_x.

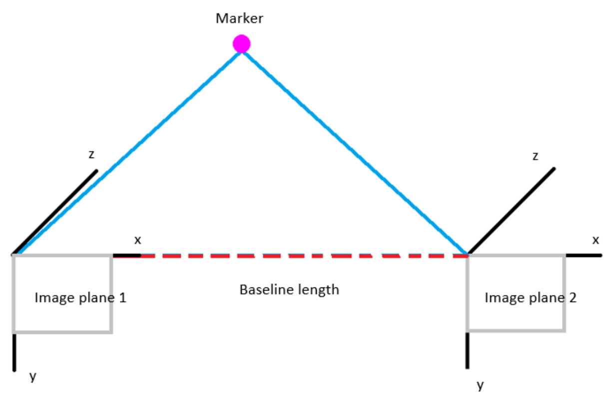


Figure 1: Camera Depth Setup

$$z = \frac{t_x f_x}{d_x} \tag{1}$$

[6]

III. Design and Experimentation

III.A. Proposed Design

Using stereoscopy, optical motion capture was determined to be the best solution for the project design. The proposed design of the system was to use four GoPro cameras that were to be mounted to the kayak, in order to capture the technique of the subject. At the current stage eleven markers are going to be used on the subject; one on the head, each shoulder, elbow, and wrist, and back as seen in the Figure 2. Two Go Pros are to be placed at the rear and two at the front, in order to be able to determine where the markers are if they are lost behind a body

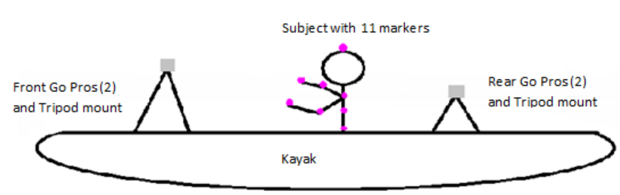


Figure 2: Proposed Design

part from the front or the rear. The subject will approximately be 1.7m away from the front camera plane; therefore most tests were conducted at 1.7m.

III.B. Choice of Baseline Length

The distance of the baseline between the two cameras is a significant factor with stereoscopic imaging. This is because it ultimately determines how much overlap there is between the two images. If the cameras are 10mm apart, approximately the "normal" distance, there is no real difference between the two images. Therefore the disparity between where the markers appear in the two images will be very minimal. This then leads to difficulty when creating a significant depth scale. With testing different distances for the baseline the optimum distance for pixel disparity could be achieved.

Another important aspect to be considered when the distance for the baseline was to be determined is at what point the physical support for the cameras become unstable. If the camera's are too far apart then the support will be more prone to shaking. This means when the video is being captured on the kayak the marker locations between the left and right planes could be out of calibration, leading to incorrect representation of marker location and therefore the technique of the subject.

The differences in disparity were compared over a range of distances from the camera baseline to determine the ideal length of the camera baseline. In this experiment the baseline length was swept through from 0.1m to 1m. Figure 3 shows the disparity response.

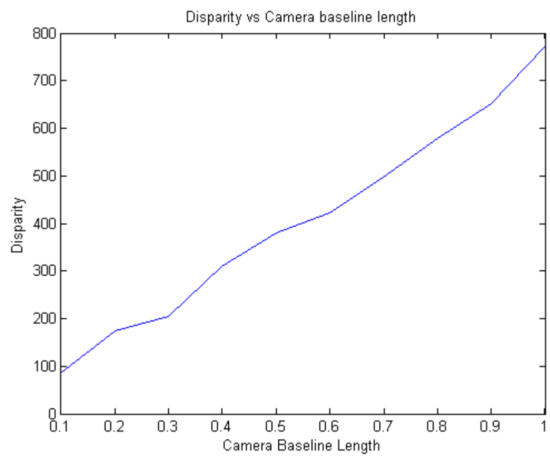


Figure 3: Disparity vs Camera Baseline Length

From this plot it can be seen that the best length of the camera baseline is 1m. As the baseline length keeps increasing the disparity between the two markers in the images will become greater, however as the marker becomes closer to the baseline it may not be seen by either of the cameras as there may not be enough overlap between the images. Therefore at a certain distance from the baseline the increase in baseline length will become negative due to missing the markers. Furthermore as mention above if the baseline length becomes to large the stability becomes lower, this is why no lengths were tested above 1m. For further experiments the baseline length of 1m was used.

III.C. Marker segmentation

In order to create an optical motion capture system it is critical to segment the areas that are to be analysed by the system. In this case the areas to be segmented are the wrists, elbows, shoulders and the head. The most effective means of segmenting these areas were determined to be by using coloured markers. Motion capture systems such as the Vicon sytem use reflective markers, LED light, and infra-red technology. However due to the subject being on water and an outdoor environment the idea of having reflective markers and infarred technology was determined to not be effective for marker identification, due to the reflective nature of water and the extra manipulation to the GoPro cameras. This lead to the decision of using coloured markers. A colour could be used that was not standard in the environment (similar to why reflective markers are used) as it can be separated from other objects in the image. Using fluorescent pink markers, a colour filter could be applied to the image after it was taken and then from there it can be isolated. As seen in the Hue-scale in Figure 4. Firstly, the HUE colour space was used to segment the pink markers from the

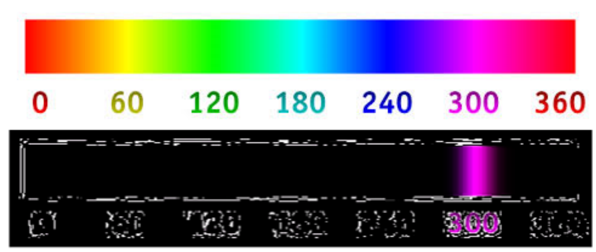


Figure 4: Hue-scale & Filtered Hue-scale

background of the frames. This lead to difficulties in later experiments where some foreign objects would be recognised leading to incorrect data points.

Alterations to the image after capture were made to try and correct these misrepresentations, such as the intensity of the markers colour was adjusted so that the pink was more prominent in the image so it could be successful identified. From here further isolation was applied to the markers. The objects that were on the border of the image could easily be removed using the in built MATLAB function *imclearborder* which allowed for a better isolated image. Once this was done the circular objects were found using *imfindcircles* which identified circles within the specified radius range. This allowed for the marker to be identified within the image with an x and y pixel location.

However this was deemed to be inefficient, therefore the conversion to the LAB colour space was tested. The frames were then post processed; converting it into the LAB colour space. This colour space was deemed more effective for isolating the desired colour over the HUE colour space as previously used, due to the LAB involving just a lightness and two dimensions a and b as can be seen in Figure 5.

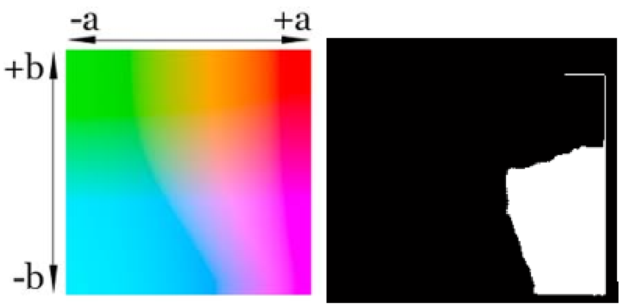


Figure 5: LAB & Filtered LAB

Where as the HUE involves both the dimensions, lightness and angle of rotation in addition. Once the image was converted into the LAB colour space a threshold could be used dependent on the dimension of a and b to convert the image into a binary image. Furthermore if any unwanted objects did fall between these threshold values, an upper and lower boundary dependent on area was used to remove these objects. Obtaining the binary images in both image planes then meant that the markers were completely segmented from the surrounding areas in the original image/frame. Having the isolated markers now allowed for marker analysis, and in turn 3D representation.

III.D. Third Dimension Calculations

Depth is the third dimension, which needs to be calculated from the two different image planes. To test the depth the baseline; disparity and marker distance from the camera plane were adjusted. With the x location, the disparity between the pixel location for the left and right images could be calculated. The disparity between the pixels is important as to achieve a significant depth, there would have to be approximately 1 pixel/mm to be able to have a significant difference between depths. The known distance of the marker to the camera plane is useful. Due to having a known distance of 100mm this can be used to test if the disparity is large enough to achieve the 1 pixel/mm.

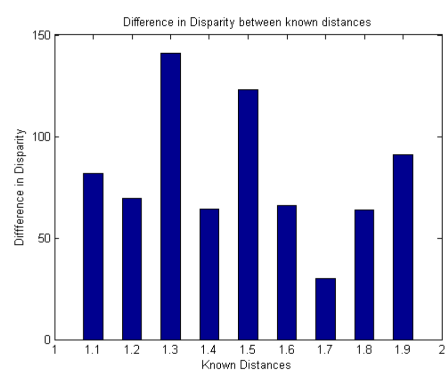


Figure 6: Difference in Disparity between known distances

Taking images of the marker from 1m and then 1.9m from the camera baseline the difference in the disparity in pixels at these distances could be calculated. This is seen in Figure 6.

This showed that as there is approximately 1 pixel/mm. To achieve 1 pixel/mm the difference in disparity had to be approximately 100 pixels/10 cm. From here the average disparity per 10cm was 81.15 pixels, therefore 0.8115 pixels/mm. This response resulted in a relatively linear line over the test area, therefore this relationship could be used later for interpolation. As seen in Figure 7.

Using equation (1), the depth value for the known distance could be calculated in order to have a different value at known distances. The results indicate that there is an increase in depth as the distance from the marker to the camera baseline is increased. However this increase between steps was too small in order to distinguish different known distances, therefore was not used to estimate the depth of the markers. Instead the relationship between the gradient and the known distances in Figure 7 were used, and will be discussed in Depth Estimation(III.F).

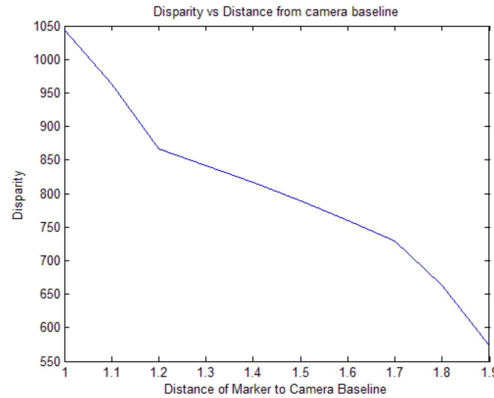
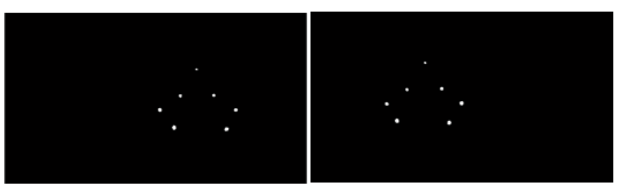


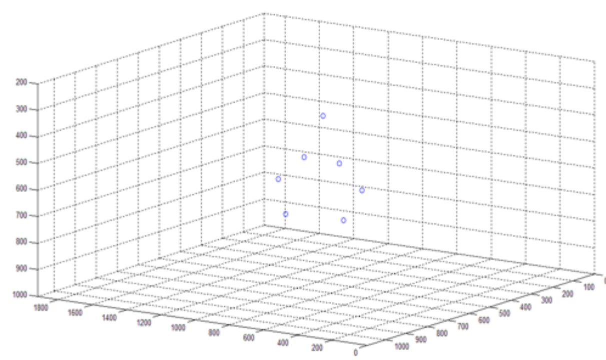
Figure 7: Disparity vs Difference from Camera Baseline

III.E. 3D Representation

Now that the markers have be segmented from both the image planes, the centroid of these markers could be found. This was achieved by analysing each marker as a solid area, using the x and y pixel value of the markers centroid, they could be compared to the other image planes respective markers, which can be seen in Figure 6. By taking the difference between the centroids x and y pixel values over the two image planes, the x and y coordinates for that marker in 3D space could be represented. In order to achieve a third dimension, the disparity between the pixels was used as another dimension. By testing the disparity change over a range of distances it was deemed an effective means of representing the third dimension. It was seen that the relationship was relatively linear as the markers moved further away from the camera baseline. Using the disparity as the third dimension the markers could successfully be represented in 3D space as seen Figure 8.



(a) Left & Right Binary Images



(b) 3D Representation

Figure 8: Binary Images & 3D Representation

Each frame of the video was analysed individually and then represented into 3D space. From here each markers coordinates in each frame were saved into an array. This allowed for each marker, over the duration of the video to be represented in a sequence of 3D plots. Meaning that the model of the human figure could be represented in 3D space for analysis.

III.F. Depth Estimation

With the markers in 3D space, the coordinates/scale must be transformed into metres in order to compare the systems accuracy against a system that already exists, in this case the Vicon System. Now since the testing has been conducted at approximately 1.5 - 2 metres (which is also the distance where the system will be operating), the x and y coordinates could be compared to the known distance between two markers, and interpret the number of pixels in a metre at 1.5 - 2 metres. This was found to be 159.061 pixels per 0.45 metres or 353.468 pixels per metre. To find the distance the marker was away from the camera baseline, the disparity was tested over a number of known distance (1 - 1.9 metres) as previously mentioned, and this lead to an interpretation of disparity 1550 pixels being where distance equalled 0 metres and disparity of 0

pixels where distance equalled 3.2 metres. From this test (Figure 7), as mentioned earlier, it was determined relativity linear. Therefore the gradient of this linear line could be used to find the depth of the markers.

$$z = (\frac{d_x}{m}) \times dist \tag{2}$$

Where:

- z=Depth;
- d_x=Disparity;
- m=Gradient;
- $dist$=Total distance interpreted (3.2 metres)

Using this equation the depth and coordinates could be estimated in metres instead of pixels, which is much more beneficial when analysing the motion capture.

III.G. Vicon System Test

In order to test the effectiveness of the system developed, the GoPro System was tested against a known commercial optical motion capture system; Vicon motion capture system. The Vicon system consisted of ten cameras that use LED lights are reflective markers, the system required five reflective markers. The Vicon system was setup as seen in Figure 9. The origin of this camera system was in the centre. Where as the GoPro System origin was at the centre of the camera baseline, which was positioned 1 metre above the ground. The means that the scales on the following results are not the same, however the movements made in the test were the same. Therefore all distance should be the same, yet respective to each systems origin.

The subject that was used in these tests was a homemade paddle which had two pink GoPro markers and five reflective Vicon markers.

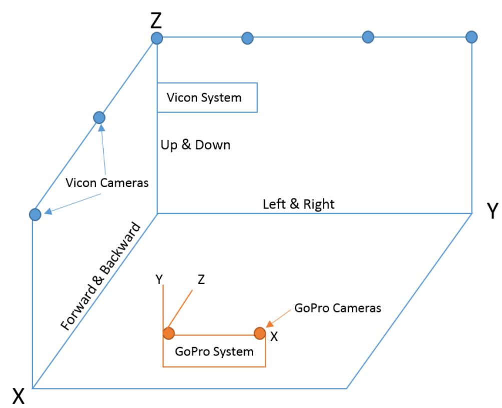


Figure 9: Vicon Test Camera Setup

The first test to be undertaken was a paddling motion, where the subject was at the x, y origin of the Vicon system and 3 metres away from the GoPro system origin. The expected result of this motion is to have sinusoidal waveform in each direction (x, y, z), which are slightly out of phase from one another. Both systems were tested at the same time in order to keep results obtained relatively in synchronisation, however the human error of starting and stopping the two systems at the same time effected the overall synchronisation; as seen in the beginning and end of the plots below. Even though there was this inaccuracy, the major outcome of this test was to see if the GoPro system followed the same pattern as the Vicon system. Evidently seen in Figure 10, both systems produced sinusoidal waveforms in each direction as expected.

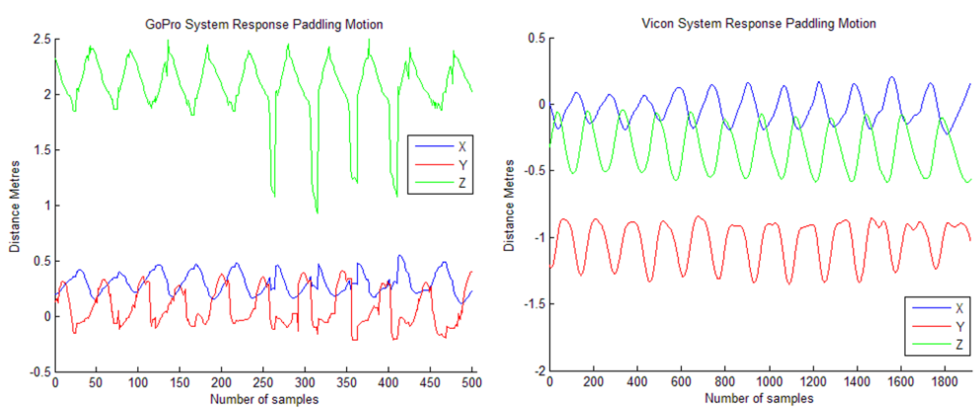


Figure 10: GoPro system vs Vicon system

It is important to note that the each systems coordinates have been represented in GoPros coordinate system. From here it can be seen that the Vicon system does produce a much smoother response, however this is due to that the Vicon system samples at approximately four times the rate than the GoPro system. Another reason for the GoPro system not being as effective at tracking the motion is that it is was tested at its maximum effective range, as the system was designed to operate with in 0.5 - 2 metres.

The next test that was undertaken involved the rigid paddle moving around in test area. The motions that were tested were up and down, forward and back, and left and right; as seen in Figure 9. The motions from both systems were compared to one another with the Vicon system used as the control (known/correct) response for the motion captured. All motions were done individually (i.e. left and right, then up and down). The first motion to be test was up and down, therefore in the y direction for the GoPro system and z for the Vicon system. What is evident from the plots in Figure 11 is that the motion captured from the GoPro system follows the fit of the Vicon system. The drop in the data in plot for the GoPro system at approximately the 800 sample mark is due to when the subject moved away from the camera baseline. This is an inaccuracy that is of no significance, due to the subject moving outside the effective area of the system. The design of the system only requires to capture the motion data to a maximum of 2.5 metres.

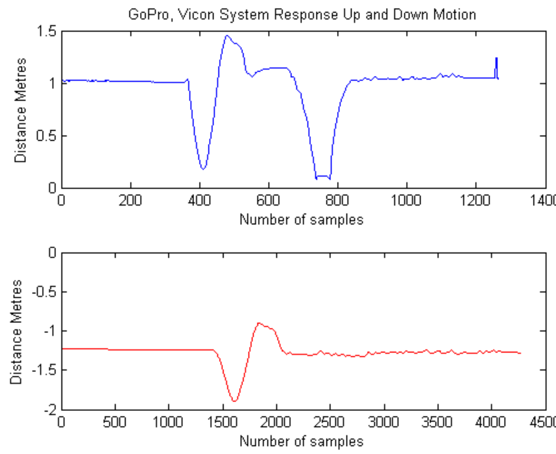


Figure 11: GoPro vs Vicon Test; Up and Down Motion

The next motion that was captured was the subject moving forward and backward from the GoPro system's camera baseline. Again the GoPro system follows the trend seen in the Vicon systems data, as seen in Figure 12. As previously mentioned the Vicon system has a much smoother motion capture response, this is due to the Vicon system sampling at a much higher rate. As can be seen in all the plots involved with this test is that the number of samples from the Vicon system exceed 4160 whereas the GoPro system only reaches 1255 samples, even though the same time frame was taken for both systems. This explains why there are some abnormalities in the GoPro system's data. Again due to the test pushing the capabilities of the GoPro system, some inaccuracies could be recognised. Firstly, the GoPro system calculated the maximum distance that the subject was from the camera baseline as approximately seven metres whereas it should have only been at approximately five metres. However again has no real significance due to the design only requiring to effectively capture the motion at maximum of 2.5 metres.

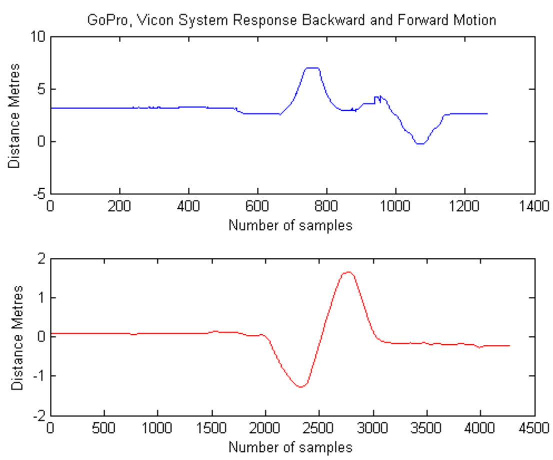


Figure 12: GoPro vs Vicon Test; Up and Down Motion

The last directions to be tested was left and right. Again the GoPro system could be successfully compared to the response from the Vicon system, as seen in Figure 13. With the left and right motions being inverse to the Vicon system. From all of these motion responses, in can be concluded that the GoPro system can effectively capture the motion of the subject being analysed. However due to the design of the system to work within a distance of 2.5 metres there were errors in the systems response to certain motions, yet the errors are not significant enough to question the outcome of the project's aim.

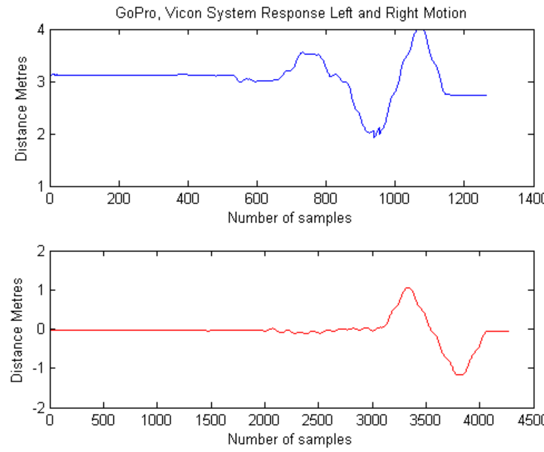
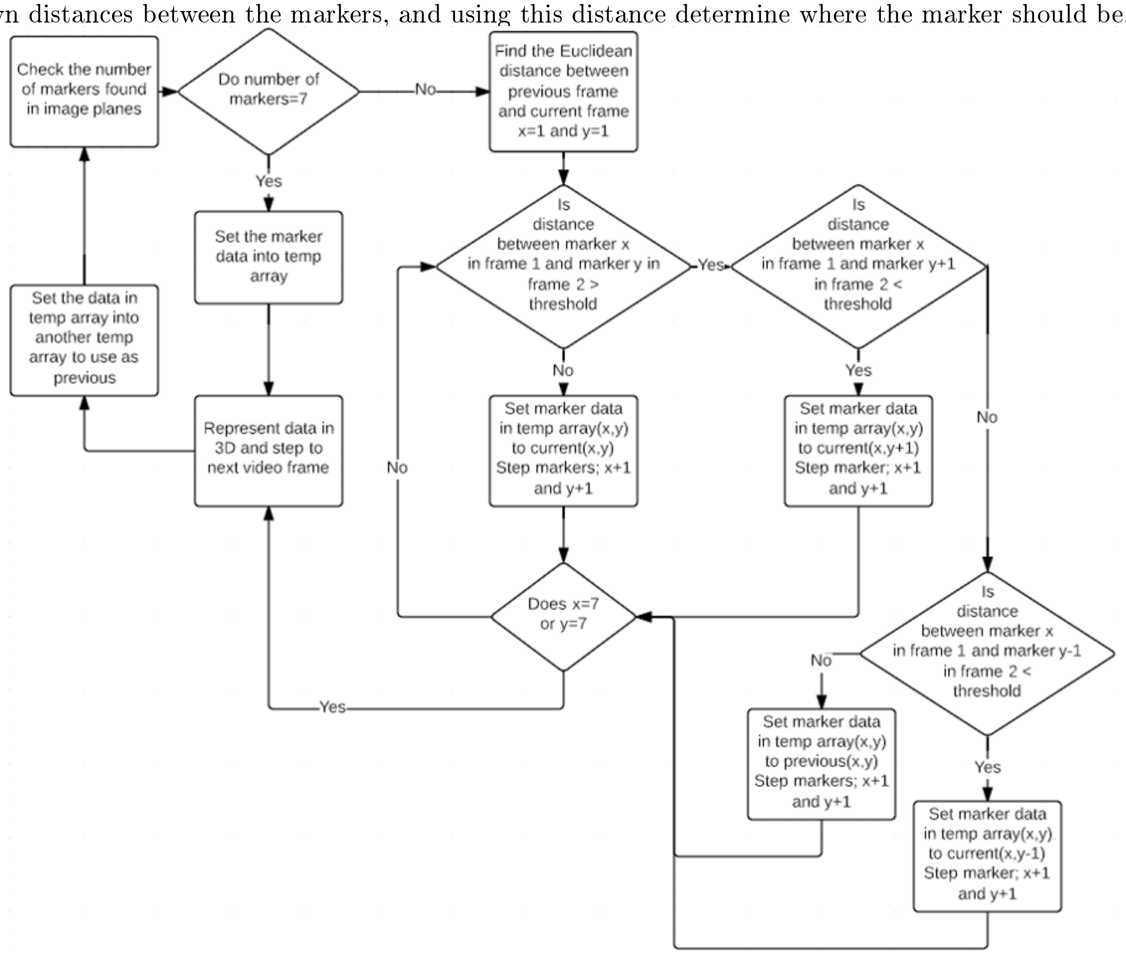


Figure 13: GoPro vs Vicon Test; Up and Down Motion

III.H. Marker Tracking

With the motion capture system working successfully, the next obstacle of interpreting the data lost had to be overcome. When the subject is performing kayaking strokes at certain points, some of the markers are lost from the sight of the cameras or cross over changing the order the markers are listed, due to the cameras only being mounted in front of the subject. In order to develop the figure of the subject in 3D space, it is critical that one of the joints is not lost or misrepresented. Therefore an algorithm was created to track the markers, and then once a marker was lost, use it's last know location. Below is the algorithm used. The algorithm effectively represents the lost marker data until the marker has reappeared, however can not determine what that marker is doing once it is lost from sight. This can lead to some inaccuracies within the motion capture. A way that this inaccuracy can be overcome is by measuring the known distances between the markers, and using this distance determine where the marker should be.



IV. Conclusion

In conclusion, the GoPro system was successful at capturing the motion of a subject and representing the data in 3D space. As seen in the testing and results obtained against a Vicon system, that is known working system for motion capture. Although there were some errors in the GoPro system's responses they were deemed not significant enough to question the project's design and aims. Marker segmentation, 3D representation, depth estimation and marker tracking were all crucial in achieving successful motion capture.

The GoPro system created is very adaptable with the system able to be easily manipulated to use extra cameras, with some slight adjustments to the design in camera placement. Having extra cameras in the system would allow for the GoPro system to be used for other motion capture applications, as well as provide more accurate representation in 3D space and marker tracking.

References

[1] Jonathan Deutscher, Andrew Blake, and Ian Reid. Articulated body motion capture by annealed particle filtering. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 126–133. IEEE, 2000.

[2] Scott Dyer, Jeff Martin, and John Zulauf. Motion capture white paper, 1995.

[3] Maureen Furniss. Motion capture. *posted at http://web. mit. edu/mit/articles/index-furniss. html on Dec*, 19, 1999.

[4] Adam G Kirk, James F O'Brien, and David A Forsyth. Skeletal parameter estimation from optical motion capture data. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 782–788. IEEE, 2005.

[5] Kazutaka Kurihara, Shin'ichiro Hoshino, Katsu Yamane, and Yoshihiko Nakamura. Optical motion capture system with pan-tilt camera tracking and realtime data processing. In *ICRA*, pages 1241–1248, 2002.

[6] Pickering. Stereoscopic and multiview video coding. 1, 2014.

[7] A Puri, RV Kollarits, and BG Haskell. Basics of stereoscopic video, new compression results with mpeg-2 and a proposal for mpeg-4. *Signal Processing: Image Communication*, 10(1):201–234, 1997.

[8] John M Rolfe and Ken J Staples. *Flight simulation*. Number 1. Cambridge University Press, 1988.

[9] Ardrea Shetley. Easy 3d: Stereo photography for everyone, 2011.

[10] Various Sources. Stereoscopy. 1, 2013.