Method For Segmentation Of Articulated Structures Using Depth Images for Public Displays

November 11, 2013

Robin Watson

rjw170@uclive.ac.nz

Department of Computer Science and Software Engineering

University of Canterbury, Christchurch, New Zealand

Supervisor: Dr. Richard Green

richard.green@canterbury.ac.nz

Abstract

A novel method is presented to analyse articulated structures in depth data and is used in an attempt to implement gesture-motion control. The method first uses region growing with a depth threshold to obtain an initial segmentation of the scene into different bodies. Region growing is carried out again on these bodies to produce subregions. A head tracking method and hand tracking method were implemented using the depth analysis. The head tracking had an average of 22 pixel error. The hand tracking was unsuccessful.

Acknowledgments

Thank you Richard, for all your patience, support and insight.

Contents

1 Introduction

Interactive public displays have long been constrained to physical interaction with button or touch screen interfaces. With recent advancements in depth imaging technology, motion controlled interfaces are now possible. Devices such as the Microsoft Kinect camera allow motion control without requiring expensive stages, sensors or markers to be placed on users or the environment. Current depth image software libraries [9, 8] are able to track human body pose very effectively in a reasonably constrained setting, where users are situated close to and in full view of the camera. However a general public display setup will not guarantee such constraints [2]. Such a display is likely to have crowds of people where users may obscure one another and it may be unfeasible to place the camera close to the interaction space. The public setting rules out complex gestures since the majority of users will be first time users. Gestures such as broad hand motions will be well suited as complete motion interaction schemes are achievable and resemble the very accessible and widespread touch interfaces [2, 11]. Because of their shallower associated learning curve, interfaces with simpler interaction are also well suited to brief interaction experiences such as what would be found at a public display.

1.1 Research Contribution

The research presents a novel method to track body pose in simple, broad hand gestures for a crowded scene where people may occlude one another. This paper describes a vision-based system to track motion gestures by segmenting depth data for bodies and then tracking face and hand locations on those bodies. The method does not use a body model and is designed to work in situations where the entire body is not visible to the camera. The lack of a body model means the implementation requires no dataset training. The gestures description obtained from the system by the system would be enough to implement a reasonably complete version of multitouch interaction for the motion display.

1.2 Report Outline

Th remainder of this paper is organised as follows: Section 2 describes the motivations for this research and provides background on similar public interface applications and pose tracking methods. Section 3 describes the design and implementation of the vision system and the methods employed to track users' hands and faces. The evaluation of the effectiveness if the system is given in section 4 and the results are discussed in section 5. The potential abilities of gestures output from the system are also discussed.

2 Background and Related Work

2.1 Motivation

The public setting brings many constraints that make many computer vision approaches unsuitable for motion interaction. To achieve body pose tracking, traditional vision methods have relied on colour have placed constraints on user clothing, background environment and lighting in order to use color information such as skin tone [3]. Public displays may be situated outdoors and methods that rely on strictly controlled illumination will not be applicable. Constraints on user clothing, such as long sleeves and no gloves, will detract from the experience and reduce the overall accessibility of the application. Similarly, fitting specialised tracking equipment or coloured markers to the user will also affect accessibility and may be difficult to manage. Ideally the display can be left unattended. Solutions making use of depth data may overcome all such constraints on lighting and user clothing. A depth image captures how far from the camera the object is at each pixel and from this information it is possible to get a much more informed view of the captured scene (figure 2.1).

2.2 Related Work

Kelly et al. [7, 6] have tracked pedestrians in an unconstrained environment with a disparity depth camera. They used a region growing approach to combine regions in order to segment the pedestrians in the frame. They also incorporated anthropometric constraints during their region combining to constrain the results of their segmentation to realistic human dimensions. They produced very robust tracking of pedestrians however the overall body shape was not preserved which would limit pose recognition. Their system was also slow being designed for surveillance applications rather than real time interaction.

Chu [2] carried out studies on gesture systems for public display interaction. He inves-

Figure 2.1: A depth image. The image instensity corresponds to how close the pixels are to the camera

tigated display interaction and proposed a complete set of interactions for large motion controlled displays. These were partially inspired by multitouch direct manipulation gestures [11, 12, 4] used in touch interfaces. He also assessed a semaphore flag style interface.

Much work on body pose tracking uses classifiers, decision forests and other machine learning approaches. Microsoft's Human Pose Estimation for Kinect project very robustly and efficiently tracks human body pose using decision forests[9, 8]. Other works use an articulated 'parts' body model and statistical means to match the image to model pose[13].

3 Design and Implementation

3.1 Overview

The goal of the system is to distinguish the different people in the frame and track where their face and hands are. The system uses data from a Microsoft Kinect although other depth sensitive cameras can provide similar data. The system is organised as shown in figure 3.1.

The first segmentation stage separates the image according to the bodies visible in the image. The second stage analyses the depth data for each person to acquire their structure. The next two stages make use of the structural information to determine the position of the hands and face. The data the system operates on is assumed to contain only people. The people in the data performing interactive gestures are all facing toward the camera. This allows for a left-right division of the person directly in the image plane.

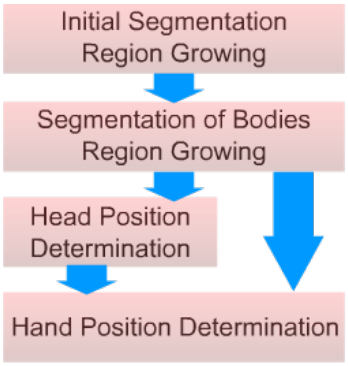


Figure 3.1: system overview

3.2 Segmentation of Bodies

The goal of the first process of the system is to distinguish regions of the image to corresponding to different people in the image. This process results in the image being separated according to the bodies identified in it (figure 3.2). A region growing method operating on the depth data was used to carry out the segmentation. A region growing process on an image starts with a pixel and incrementally checks its neighbours to determine if it belongs to the same region, according to some condition [1]. The condition chosen here is if the pixels are within a fixed depth threshold of each other. The threshold was chosen as approximately 9.5cm. . The borders of the two bodies must have at least 9.5cm between them in their distance to the camera in order to be segmented correctly as different bodies (figure 3.3).

In order to correctly segment the image, the threshold must be tuned precisely. If it is set too high, the segmentation will fail to distinguish between close bodies. If it is set too low, bodies with sharp depth discontinuities such as a forearm held forward will be segmented as separate bodies (figure 3.4). Additionally, segmentation may fail to distinguish two bodies if they are positioned sufficiently close in depth and close side by side such that no intermediate pixels create a boundary between them (figure 3.5). The quality limitations of the Kinect image data may also cause discontinuities in the smaller, finely detailed parts of the bodies such as hands, and cause them to be segmented separately. These generally occur

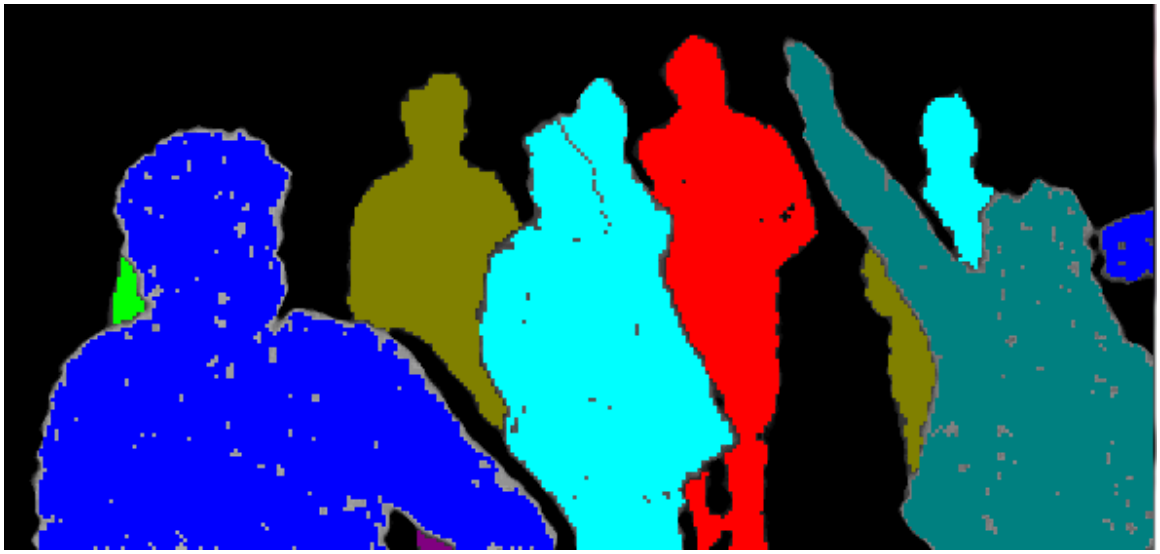


Figure 3.2: A segmented crowd scene

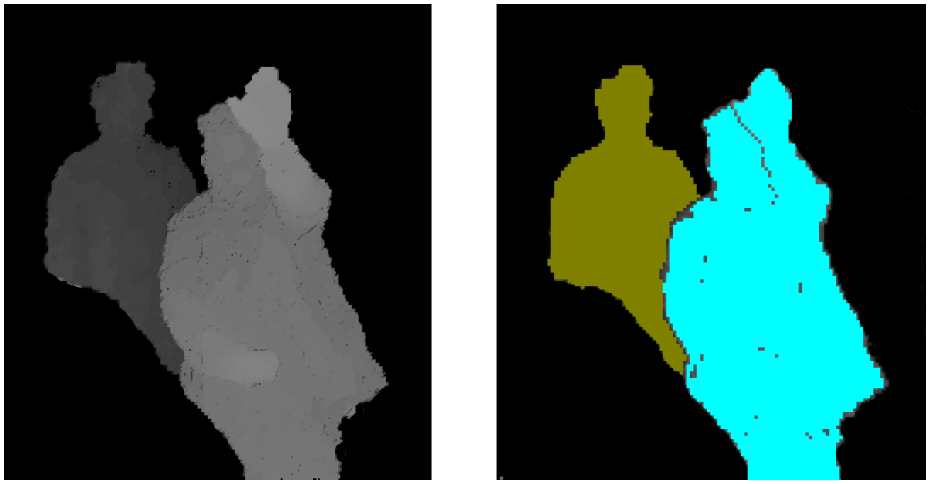


Figure 3.3: Left: the two bodies are at different depths. Right: The two are segmented as different bodies, even though they overlap in the image

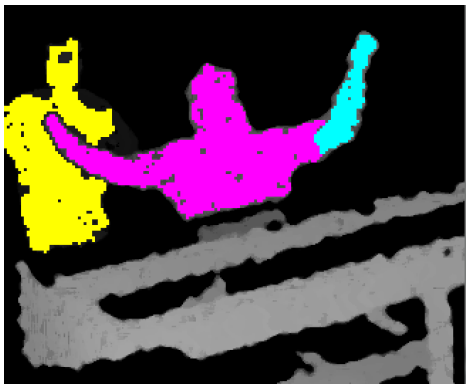


Figure 3.4: The person's arm was held forward enough to exceed the connectivity threshold resulting in failed segmentation

only at the tips and extremities of a body and don't affect the overall structure significantly.

The depth data allows the region growing process to be initialised to filter out data from a close range or a far range. This defines a volume of interest, oriented with respect to the camera. This allows for isolation of a smaller operating area within the camera's view.

After segmentation, the segments are then filtered for a minimum size in order to produce only bodies of a similar size to people. The minimum size is given as a simple count of pixels(mass) in these experiments. This filtering accounts for noise and removes stray body parts that were occluded so could not connect to their whole body. The minimum size could also be a function pixel mass and per-pixel depth. This would better match the size of object in real world dimensions rather than in the camera's view. However in the test

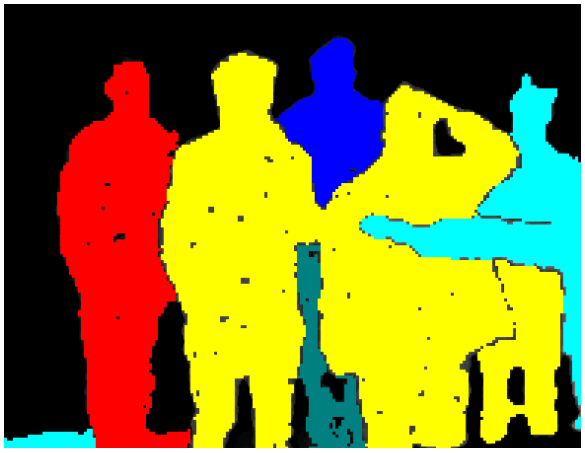


Figure 3.5: The two people in front stood close enough at similar depths that they were segmented as one body

data used, the range of depths of the bodies did not vary significantly enough to necessitate depth-dependent size filtering.

3.3 Further Segmentation of Bodies Into Parts

The second stage segments bodies further, distinguishing different regions of the body's articulated structure (figure 3.6.).

Figure 3.6: The person has been segmented into body parts, revealing their underlying structure.

This stage uses region growing again, this time operating on individual bodies identified in the last step, finding subregions of the bodies (figure 3.6). These subregions(body parts) are segmented with a much lower connectivity threshold than used in the full-body segmentation. This lower connectivity threshold divides the body parts on a small scale,

rather than the large scale of the initial body segmentation described in section 3.2. The threshold ranges from 0.2cm to 3.4cm and is an independent variable for the experiments described in section 4. An important property of the structure of the body parts is the 'bordering' of parts. Body parts are considered to be bordering if they share perimeter pixels that are within a compatible distance in both depth and the image plane. Compatible depth is defined as if the parts have some overlap in their depth ranges. This means that either part has a depth extreme between the other part's extremes. Because adjacent parts may have gaps between them (figure 3.6), parts considered bordering may exist with an allowable number of pixels between them, here called the 'bleed' distance.

3.4 Determination of Head Position

In order to help define the extents of the hands' control range, the head position is determined for each body so that hand positions may be taken as relative to the head. The approach taken assumes the largest body part identified will be the torso. The bottom of the head will be positioned at the top of the torso. The y coordinate of the head is taken to be the top boundary of the torso. The x coordinate is taken as the x coordinate of the centre of mass of the torso. This approach requires that the head be effectively distinguished from the body during segmentation. The presence of the jaw generally provides enough of a discontinuity between head and chest, especially if the user tilts their head up in order to look at a display. The experiments described in section 4 assess the accuracy of this approach.

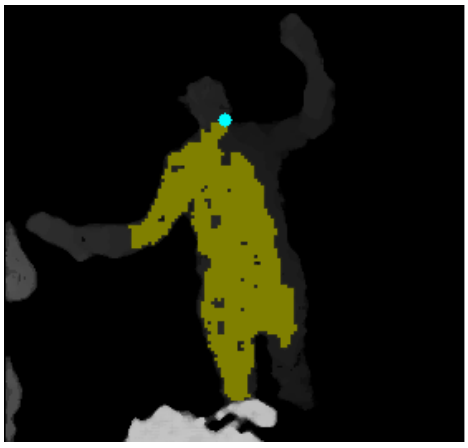


Figure 3.7: The largest body part is shown in yellow and is deemed the torso. The head is marked in blue at the top of the torso

3.5 Determination of Hand Position

A function is evaluated per body part to determine if it is a hand or part of a hand. The approach considers hand body parts to be extremities in the body structure and expects that their bordering parts will all be off to one side. This means that there will be a large angle between two angularly adjacent bordering parts3.9. This is termed here the 'clearance angle'. An extremity is assumed to have a clearance angle of over 330°. The hand will only be an extremity when the user moves it sufficiently far from their body. For example, if the user had their hand in their pocket the system would regard it as being connected to their body. This grey area is further discussed in section 4.2

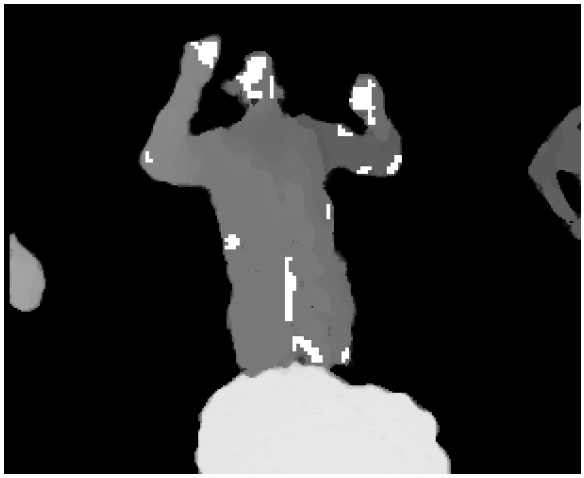


Figure 3.8: The body parts that are identified as hands by the algorithm. Many false positives are present

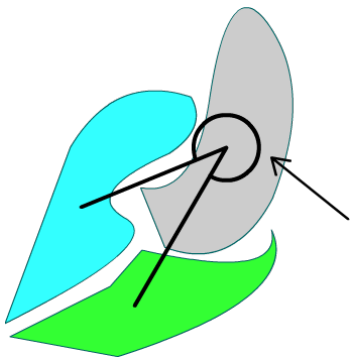


Figure 3.9: The clearance angle of the grey body part, as defined by its neighbouring body parts

4 Evaluation and Results

Evaluation of both the head and hand position determination was carried out using a groundtruth labelled video feed. The system evaluation was carried out on a frame by frame basis to test the methods' instantaneous accuracy, rather than with the aid of temporal information to enhance stability.

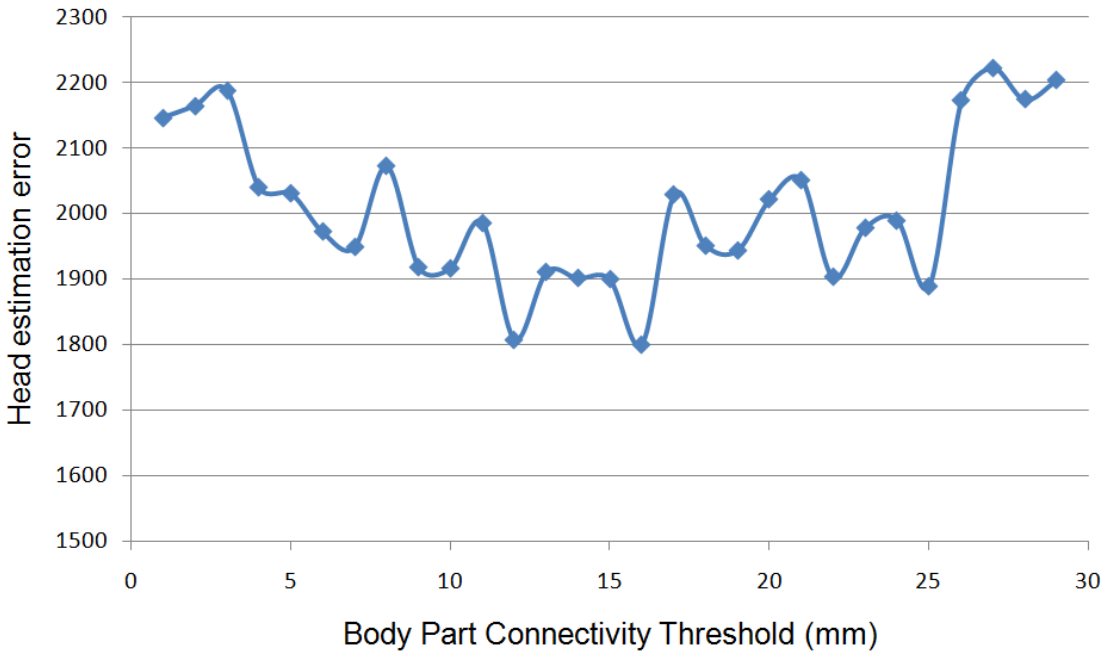
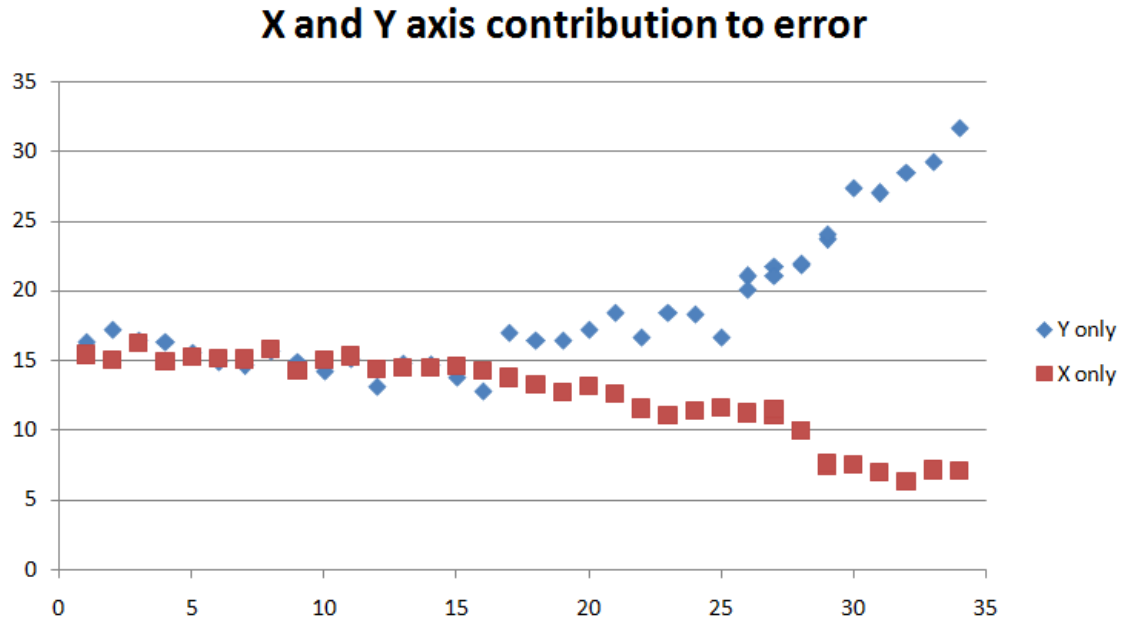


Figure 4.1: Head estimation error for range of connectivity thresholds

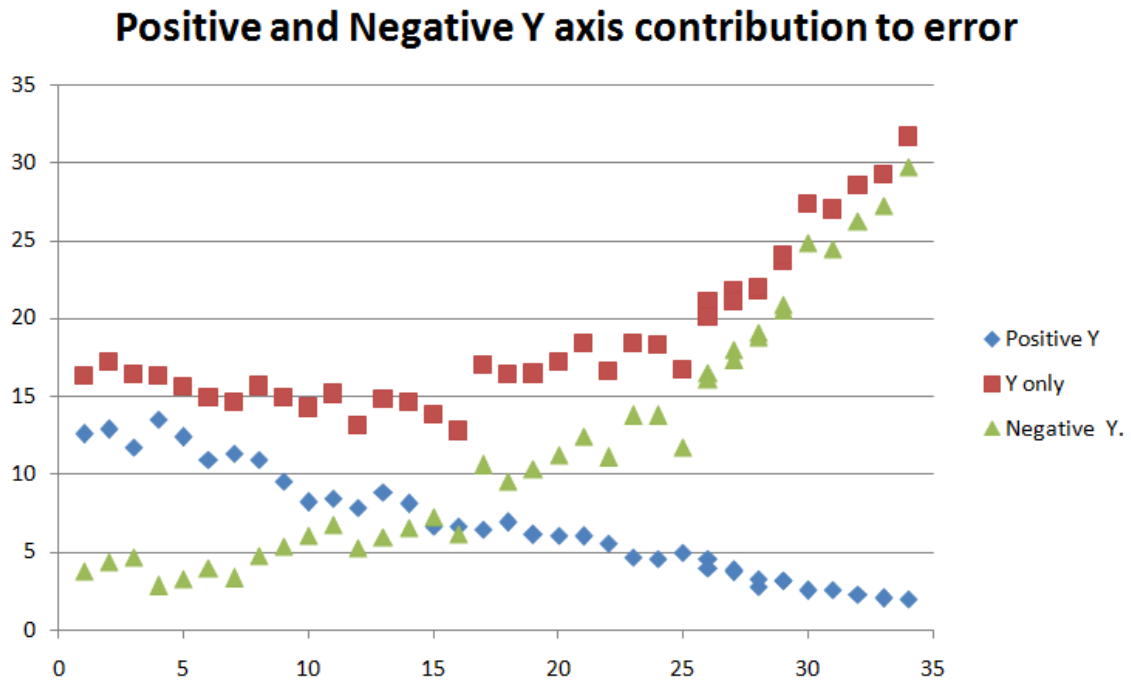
4.1 Evaluation of Head Position Determination

For the head position evaluation, the groundtruth marked the position of the bottom of the head of each person in frame. The error in head position as determined by the system is measured as its deviation from the marked groundtruth head position for a given body. The units of measured error are screen pixels, that is, the distance in the screen plane. The average error summed over all the video frames was measured under a range of connectivity thresholds for the body part segmentation stage (see section 3.3)



4.2 Evaluation of Hand Position Determination

The groundtruth for the hand position evaluation had approximately all pixels marked if they contained the hands of the users in view. This was done for each frame of the video. To account for the grey area where a hand transitions from a body connected position to an extremity, only hand positions above the head line were considered. This is because in general a person raising their hand will have moved it far enough from their body by the time they raise it above their head. Hands that were raised up to the head but were not raised out were not marked as hands in the groundtruth. The groundtruth also recorded per frame how many hands were raised for each person in frame.



The evaluation considered, for every frame, the body parts above the head identified as hands. If any of these body parts had pixels that overlapped with the groundtruth marked hands, it counted as a positive identification for that hand. Since many, small hand segments may overlap the marked hand, the maximum score per person was the number of hands marked for them in the frame. The total number of hands positively identified divided by the total available in the video was the score for that experiment. The experiment was run for combinations of body part connectivity threshold and bleed distance.

Bleed (px)	0	1	2	3	4	5	6	7	8	9
Threshold(mm)										
2	0	0.11	0.13	0.17	0.20	0.21	0.22	0.25	0.25	0.26
4	0	0.11	0.19	0.18	0.21	0.19	0.25	0.26	0.22	0.22
6	0	0.17	0.20	0.23	0.23	0.21	0.28	0.25	0.27	0.25
8	0	0.18	0.21	0.25	0.26	0.22	0.24	0.25	0.28	0.24
10	0	0.19	0.21	0.24	0.23	0.23	0.22	0.25	0.25	0.24
12	0	0.25	0.26	0.26	0.28	0.26	0.27	0.26	0.26	0.21
14	0	0.25	0.28	0.25	0.27	0.28	0.27	0.28	0.23	0.23
16	0	0.21	0.24	0.23	0.23	0.26	0.24	0.24	0.21	0.22
18	0	0.21	0.20	0.24	0.22	0.23	0.19	0.21	0.20	0.19

Table 4.1: Sensitivity of hand identification

5 Discussion

5.1 Head Position Determination

The head position determination achieves reasonably steady tracking of the head position per frame. As can be seen in figure 4.1, The variation in error as a result of sensitivity to threshold is approximately 12% (2.71 error for values of 23). The figure shows that there is a threshold value around 15mm that produces a minimum in error of around 22 pixels. The error increases sharply as threshold increases beyond this value. Figure 4.1 shows the x and y axis component contributions to the error. The x-axis contribution steadily declines as the threshold is raised and the body is segmented into larger parts. This appears to be because the x position output by the system converges to the centre of the body as the whole body converges to one segment, as will happen with a high connectivity threshold. The y coordinate error increases as the body converges to one part. Figure 4.1 shows the y-axis contribution to the error from positive values and negative values (which are when the system output lies below and above the actual location, respectively). The error contribution from estimations below (positive y) the actual location decrease as threshold rises and the errors from above (negative) increase. This is again consistent with the body converging to one segment; the head position will be estimated as high, since the whole body will be given as the torso and the head will be placed at the top.

5.2 Hand Position Determination

From the hand position determination results it is clear the method is not effective. Under all variable configurations, not one trial identified more than 30% of the hands presented. It is not a simple case of inverting a 20% success rate into an 80% rate, as the true-negative/false-negative rate would also invert. The system's hand tracking is not a success.

5.3 Gestures

Chu [2] proposed that the set of gestures used for the iPhone would be effective and complete for motion controlled public displays (table 5.3). To work in motion controlled displays these gestures essentially require two hands, a 2 dimensional range of motion in which they are suitably tracked and an action or pair of actions that correspond to touch-press and release. A user moving their hands toward and away from the display can facilitate the press and release actions. Once a more effective method for hand tracking is implemented, this set of gestures will be available to any users present in front of the depth camera that are visible enough for the system to effectively handle.

Gesture	Action
Tap	To press or select a control or link (analogous to a single mouse click event).
Double Tap	To zoom in and centre a block of content or an image. To zoom out (if already zoomed in)
Flick	To scroll or pan quickly
Pinch open	To zoom in
Pinch close	To zoom out
Drag	To move the viewport or pan. (Analogous to a mouse drag event)
Slide	To unlock and confirm turning it off. The technique is also used for deleting files in certain screens such as videos, images and e-mails.
Two finger tap	Zoom out of a map quickly
Touch and hold	To display an information bubble, magnify content under the finger.
Two-finger scroll	To scroll up or down within a text area, an inline frame. (Analogous to a mouse wheel event).
Two-finger rotate	To rotate pictures clockwise and anti-clockwise in the photo album.

Table 5.1: iPod Touch/ iPhone gestures and their actions [2]

6 Conclusion and Future Work

This paper presented a novel method to analyse the articulated structure of bodies in a scene. The method provided an effective means to track users' head positions but failed to achieve accurate hand tracking. Once an effective means of hand tracking however is determined, the system will be capable of a rich motion experience for large numbers of users simultaneously.

6.1 Limitations

6.1.1 Segmentation

The segmentation phase has the limitation that people standing in very close proximity to each other will be segmented as a single body. This can be overcome using anthropometric constraints during region growing as was done by Kelly et al[7]. However their bottom-up joining of regions did not preserve the outline of the user well. A top down, division of the adjoined bodies may be possible and the points at which to divide the bodies may be obtained through convexity defect analysis [10]

6.1.2 Head Position

There were still cases when the head became joined to the torso and head tracking was affected. Temporal stabilisation using a rolling average of head positions will help overcome this.

6.2 Future Work

An alternative method to allow hand control would be to place virtual targets around the user's tracked head. They could then move their hand into a target to use it. This is similar to a means of control implemented by Joshua Scott [5]. This method doesn't require precise identification of the user's hands, only that the user is given space to both fill and leave the target empty. Another alternative is to count the pixel mass above the user's head. Taking pixel mass counts from both the left and right side would reflect general hand position although would not be a direct pinpoint on the hands' actual locations. The user would not notice this if the visual representation of the tracking is not overlain on the user's image.

Further work on the body part segmentation method could investigate properties of graphs constructed from the 'bordering' relationship of the body parts. These graphs would have bordering body parts represented as adjacent nodes.

Bibliography

[1] R. Adams and L. Bischof. Seeded region growing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(6):641–647, 1994. ISSN 0162-8828. doi: 10.1109/34.295913.

[2] Cheng-Tse Chu. Robust Upper Body Pose Recognition in Unconstrained Environments Using Haar-Disparity. Master's thesis, University of Canterbury. Computer Science and Software Engineering, 20 Kirkwood Ave, Christchurch, Canterbury, New Zealand, 2008.

[3] Luis M. Fuentes and Sergio A. Velastin. People tracking in surveillance applications. *Image and Vision Computing*, 24(11):1165 – 1171, 2006. ISSN 0262-8856. doi: http://dx.doi.org/10.1016/j.imavis.2005.06.006. URL `http://www.sciencedirect.com/science/article/pii/S0262885605000879`. ¡ce:title¿Performance Evaluation of Tracking and Surveillance¡/ce:title¿.

[4] Jefferson Y. Han. Low-cost multi-touch sensing through frustrated total internal reflection. In *Proceedings of the 18th annual ACM symposium on User interface software and technology*, UIST '05, pages 115–118, New York, NY, USA, 2005. ACM. ISBN 1-59593-271-2. doi: 10.1145/1095034.1095054. URL `http://doi.acm.org/10.1145/1095034.1095054`.

[5] Richard Green Joshua Scott. Public interactive displays. *University of Canterbury*, 2012.

[6] Philip Kelly, Noel E. O'Connor, and Alan F. Smeaton. Pedestrian detection in uncontrolled environments using stereo and biometric information. In *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, VSSN '06, pages 161–170, New York, NY, USA, 2006. ACM. ISBN 1-59593-496-0. doi: 10.1145/1178782.1178807. URL `http://doi.acm.org/10.1145/1178782.1178807`.

[7] Philip Kelly, Noel E. OConnor, and Alan F. Smeaton. Robust pedestrian detection and tracking in crowded scenes. *Image and Vision Computing*, 27(10):1445 – 1458, 2009. ISSN 0262-8856. doi: http://dx.doi.org/10.1016/j.imavis.

2008.04.006. URL http://www.sciencedirect.com/science/article/pii/S0262885608000863. ¡ce:title¿Special Section: Computer Vision Methods for Ambient Intelligence¡/ce:title¿.

[8] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, and Andrew Blake. Efficient human pose estimation from single depth images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2821–2840, 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.241.

[9] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Commun. ACM*, 56(1):116–124, January 2013. ISSN 0001-0782. doi: 10.1145/2398356.2398381. URL http://doi.acm.org/10.1145/2398356.2398381.

[10] Richard Green Simon Flowers. Hand tracking and gesture recognition as user input. *University of Canterbury*, 2012.

[11] Andrew D. Wilson. Touchlight: an imaging touch screen and display for gesture-based interaction. In *Proceedings of the 6th international conference on Multimodal interfaces*, ICMI '04, pages 69–76, New York, NY, USA, 2004. ACM. ISBN 1-58113-995-0. doi: 10.1145/1027933.1027946. URL http://doi.acm.org/10.1145/1027933.1027946.

[12] Andrew D. Wilson. Playanywhere: a compact interactive tabletop projection-vision system. In *Proceedings of the 18th annual ACM symposium on User interface software and technology*, UIST '05, pages 83–92, New York, NY, USA, 2005. ACM. ISBN 1-59593-271-2. doi: 10.1145/1095034.1095047. URL http://doi.acm.org/10.1145/1095034.1095047.

[13] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2878–2890, 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.261.