Motion Capture in 3D Environment Using Time of Flight Camera: An Implementation Summary of Simple but Robust Algorithm

Muhammd Fahied
Room 18,Villagatan 6
96231 Skelleftea, Sweden
fahied@gmail.com

Abdul Waheed
Room 3, Villagatan 6
96231 Skelleftea, Sweden
Waheed11@gmail.com

David Eriksson
1st line of address
2nd line of address
david@northkingdom.com

ABSTRACT
This article describes inferences derived from practical and theoretical explorations regarding three-dimensional motion capture to develop numerous software applications using Time-of-Flight (TOF) range sensor camera. It illustrates a simple but complete solution towards the development of straightforward 3D motion controlled applications for different public environments. A simple and robust algorithm to capture biggest, brightest and nearest contours (BNBC) of hand motion (by getting nearest pixels with the help of TOF range scanner camera and comprising a set of previous algorithms) is explained along with the set of image processing techniques. This article contains analysis and findings during successful implementation of BNBC to develop a couple of simple 3D motion controlled applications specifically for public places. The BNBC being an optimal hand motion tracking method, focuses on contours while considering the nearest, brightest pixels. A set of practicable programming tools and technologies followed by a detailed research report includes the summary of findings.

Keywords
3D TOF Camera, Laser Range Scanners (LSR), Public Interactive Installations, Motion Controlled Interaction, Stereo Vision (SV) Systems

1. INTRODUCTION
In this paper an efficient and straightforward approach to detect and track three-dimensional hand motion with the help of Time-of-Flight (TOF) camera is explained. A simple but efficient algorithm to track hand motion along with a couple of successfully implemented applications of 3D hand motion tracking is elaborated in this article. It has also been described that which software tools and environments are suitable to develop simple motion controlled applications to be installed at concerned public places by capturing data in three-dimensional (3D) region. A set of research results accumulated during the gradual explorations of numerous computer vision tools and techniques to capture 3D motion by TOF are described. While determining the inferences a simple algorithm named BNBC to capture three dimensional motions is elaborated here. The algorithm is based upon Brightest, Nearest and Biggest Contours drawn over the different objects. In contrast with our previous paper [2], which was written to figure out user behavioral aspects and possibilities of TOF technology to develop public interactive applications, this article addresses how to devise some enduring technical (programming) solution to develop motion controlled multimedia applications especially for public environment by using Time-of-Flight (TOF) camera. It addresses the technological and software related issues to track an object by using bare hands keeping in mind a packed public environment. The proposal of bringing multimedia interactive interfaces into public places is quite new. Moreover the development of 3D public interactive applications using the TOF camera is also a novel and innovative paradigm. To make research and development (R&D) upon these unique and fresh ideas constitutes the whole research work.

Since the authors of Canesta Inc. [7] in article [1] state about Time-of-Flight camera system "Although there are various problems involved in using such a system, we have shown that they can be successfully solved, resulting in a robust and efficient system with a depth resolution of only a few millimeters. We envision a wide variety of applications where our technology can be used to enable everyday devices to perceive and interact with their surroundings in three dimensions." [1] Hence in this article we successfully tried to solve the problems and issues related to usage of TOF camera developed a couple of very interesting applications to be installed at public environments. Moreover our previous article [2] which was written about the possibilities and user behavioral aspects of the PIAs, in which it has been convincingly argued that 3D TOF cameras can be an optimal devices to capture 3D motions in public areas e.g. airports, city centers etc. Moreover we envisaged with the help of contextual enquiry at numerous public environments like airports, and university campuses and a variety of applications were suggested to be designed and to be developed, which led us in developing this 3D hand tracking algorithm effective for public interactive interfaces.

Readers of this article are expected to have knowledge of motion detection and tracking, image processing along with proficiency in C++/C# .NET programming platform.

The outline of this article is as, in Section 2 the core principle of TOF depth sensor camera is described briefly, Section 3 contains a short description of other technologies for 3D motion capture, then in Section 4 a summary of previous relevant work is described, Section 5 illustrates in detail the 3D Public Interactive Applications (3D PIAs), Section 6 elaborates what is BNBC algorithm and its step by step presentation and functioning to track 3D hand motion, finally section 7 tells how the BNBC way of 3D motion tracking is an appropriate choice for development of PIAs. Furthermore section 8 contains a brief but concrete account of tools and technologies chosen (and used) to develop the under consideration complete programming solution. Lastly section 9 provides the conclusive summary of overall inferences and findings.

2. Three Dimensional Time-of-Flight Camera

The core principle of TOF camera is that it uses light of certain frequency; the CMOS chip converts the light signals into electrical signals and watches the phase shift of a modulation envelope of the light source as its property (and measures the amplitude of signals). How far objects are in certain scene can be calculated using the properties of light and the phase shift. The depth sensor is implemented in a single chip using an ordinary CMOS process [1][4]. This camera provides the intensity images and range data with the pixel depth at the same time. TOF camera can provide 2D images of good quality and in addition can provide good vision on scenes in 3D sense along with motion tracking information which is the major challenge of this project.

3. Technologies for Capturing 3D Vision: Brief Comparison

Besides the TOF range sensor camera majorly there are two other technologies are being used to capture 3D motions. Those are Stereo Vision (SV) Systems and laser range scanners (LRS) [4].

The SV systems besides undergoing correspondence problem are also not able to confine an ample range of field of view (FOV) [as mentioned in 5 quoted in 4]. Similarly the LRS always require some mechanical work while installation before use and are not capable of providing intensity images and range data at the same time [4][5].

4. Previous Related Work

Besides Natal XBOX 360 project, majorly companies are also working on numerous (their own) application based on 3D motion capture with the help of their own devices.

a) First one is Canesta [7] which is manufacturer of camera as well, is making 3DTV interface by using TOF camera. Canesta uses the cheap single CMOS chip in TOF camera to transform light signals into electric pulses.
b) Primesense [8] is another vendor who is working on 3DTV interface by using TOF camera. This is the vendor who is providing the 3D sensing technology to Microsoft for project XBOX 360 Natal.
c) Hitachi Full Parallax 3D TV [15]

Above mentioned 3D sensing products are being developed for indoor safe environment and none of them deals with the collaborative public entertainment, whereas our work directly and algorithm deals with motion tracking in packed public environments.

5. 3D Public Interactive Applications

3D motion controlled applications to be installed at public places for information as well as entertainment purposes are called 3D public interactive applications (3D PIAs) in this article. These applications can create new fashions of social interaction amongst people at places like, airport waiting halls, shopping malls and university campuses. A couple of examples of successfully developed 3D PIAs are “Island Explorer Demo” and “Hand Fishing”.

In the “Island Explorer Demo” an UFO (a kind of alien ship) was flown over the animated island. The motion of flying UFO was controlled by the users/players by waving their hands in air in three dimensions, while standing in-front of big screen. Hand motion of User/players was captured in 3D region stipulated in-front of big screen. The 3D coordinates captured from each frame were reflected to virtual world of Island to guide UFO to make it move in three dimensions. This application was exhibited in public hall during the Creative Summit 2010, Sweden [12]. The demo video of project is available in the link [13].

Likewise the idea of “Hand Fishing” is to put a screen in public area. Screen shows animated three-dimensional water aquarium with fishes. User moves her/his hand (in air) in a specific region in front of screen (to catch the fish in animated screen). When the depth pixels (co-ordinates) will match with the scaled pixels of fishes’ location in water, system will show that user has caught the fish. This application is titled as Hand Fishing.

To get the complete idea about feasible and possible installations of 3D PIAs in accordance with users’ perspective readers are recommended to read [2]. PIAs would be interacted by capturing motion of bare hands. To make the motion capture system in crowded public environments robust and efficient a method is devised to capture motion which is elaborated in next section.

6. BNBC Algorithm

BNBC stands for brightest, nearest, and biggest contours which can be drawn from an object being placed in-front of a camera. According to this method a logical amount of nearest, brightest pixels are extracted from individual frames’ bitmap data. Then contours over those selected pixels are drawn which can provide x, y, z data of brightness object along with the distance from camera (in the form of third dimension by TOF depth sensor) can be provided.

By implementing this algorithm in motion controlled applications at public places it has been accumulated that the hand motion is captured quite effectively to get the X, Y, Z coordinates of the pixels in 3D expanse using 3D TOF camera. By getting and optimizing 3D data as output of TOF camera this tracking method can be implemented which is proved to yield evenhanded results. According to this scheme following steps are followed to track the specific point of object to be tracked. The tracked points from real world are reflected to the virtual object in game to navigate directed movements. Contours are drawn upon brightest and nearest pixels captured by detecting hand and the biggest contour amongst all the contours in one frame is taken into account.

BNBC algorithm encompasses following sequence of steps to capture hand motion with the help of suitable programming and image processing tools.

1. Defining 3D ROI in front of camera
2. Extracting bitmap data from frames
3. Compiling bitmap image
4. Applying smoothing filters and converting into grayscale image
5. Extracting the pixels which are brightest in image data
6. Getting the nearest pixels potentially inclusive to those chosen brightest pixels using the range coordinate provided by depth sensor
7. Removing the outliers from chosen pixels
8. Drawing the contours upon each frames finally filtered and selected pixels.

9. Getting the biggest contours and ignoring the rest
10. By calculating the convexity defect extracting the tips of fingers of hand
11. Getting center point of that (biggest brightest nearest contour).
12. Detecting and capturing

6.1 Defining ROI

A specific region of interest is defined in front of camera. This three dimensional region is defined at certain distance from camera and visual screen of system to be interacted. To define this region is essential to avoid the unnecessary movements to be tracked. Users are allowed to move their hand in this ROI. Then 3D positions of some specific pixels of hand are captured with the help of BNBC algorithm. This algorithm is implemented by using C++ Mathematical functions supported by rich libraries of OpenCV [3]. To make the tracking robust crowded public places the objects outside this region are neglected.

6.2 Extracting and compiling bitmap data from frame

In this preliminary step 2D bitmap data in the form of raw pixels is acquired from each frame. Data is comprised of each pixel's brightness, its x, y position along with third dimension (distance from lens) is received. All this data extracted and saved in IPLImage[3] type object which is an useful data structure introduced by OpenCV [3]. This whole data is saved in the form of grayscale image. OpenCV smoothing (cvSmooth) operations are applied over complied IPLImage object.

6.3 Filtering and extracting the brightest pixels

This is very imperative segment of algorithm. At this stage the pixies are filtered according to their brightness level. The more significant part of the object to be tracked is taken into account.

A number of brighter pixels which exist in close proximity of each other are extracted from compiled grayscale IPLImage type object. Brightness of each pixel is set by scaling the values of brightness from 0 to 255.

$0 \leq \text{Brightness} \leq 255$

By experimenting and tracking different objects a minimum threshold value for brighter pixels is stipulated which can provide better results. Rest of the comparatively less bright pixels is ignored. This threshold value of minimum brightness can be set in accordance with the overall brightness level of environment. If stipulated threshold value of brightness is β then the concerning brighter pixels can be shown as

$$Xb=i=\beta 255Xi$$

6.4 Extracting adequate amount of nearest pixels

At this vital step a stipulated (but sufficient) amount of hand's pixels which are nearest in each frame are taken out. These stipulated minimum pixels are chosen from each frame by measuring and comparing the range co-ordinate (third dimension) of pixels. This theory is established by considering, observing and watching the people how would they like to interact with the system. Considerably bare hands would be used to interact and play with the system. And Hands always remain closer to system as compare to rest of the body. That's why users are allowed to move their hands in 3D ROI in front of system screen. Only the hand motion is being captured while ignoring rest of the pixels belonging to other parts of body, as those are considered as scraps of background.

If the pixel X which is figured out as the nearest pixel to the system lens then X is deemed as the j^{th} pixel after that at least total number of k pixels next to X (or can lying at same distance) are accumulated as X_j, X_{j+1}, X_{j+2} ..... X_{j+k}, where k is minimum stipulated amount of pixels for further operations. Following total number of nearest pixels from brighter cloud are selected.

$$\square\square = \square=\square\square=\square+\square\square\square$$

Here *Xn* are nearest pixels chosen from already selected *Xb* Brighter pixels.

As these relatively nearest pixels are extracted from already accumulated comparatively brighter pixels so following mathematical expression would express those nearest pixels which are potentially inclusive to chosen brightest pixels.

$$\square\square=Xb-X\square$$

Here Xn are aforementioned nearer pixels accumulated by subtracting comparatively distant pixels Xd from Xb (Brighter pixels). And Xd can be represented by this statement

$$\square\square = \square=\square+1\square=\square+\square\square\square$$

Here Xd the distant pixels are derived by starting accumulating the pixles after the mth Pixel (which was the last nearest pixel chosen) till the farthest pixels.

Detailed expression may look like this

$$Xn=i=\beta 255Xi-\square=\square+1\square=\square+\square\square\square$$

6.5 Removing the outliers inside ROI

To achieve more robustness, for each frame the set of proximal brightest pixels is spotted at an exclusive three dimensional space in ROI. Considering vicinity of points the x,y,z co-ordinates) the pixels tracked at marginally distant point but having optimal level of brightness are also ignored. The average of depth/range values of those spotted pixels is computed. All the pixels' depths are compared individually with the average depth. And the pixels having depth values differentially beyond the average depth value are considered as outliers.

6.6 Drawing Contours around selected pixels

Lastly chosen pixels belonging to hand are again converted into an IPLImage. Then contours upon those pixels are found and drawn. Besides that fingers of hands are also recognized by using convexity defect. Then against each frame the biggest contour amongst all the drawn contours is found. At one of the previous stages maximum brightness was considered to capture maximum image gradient in a grayscale image to find exact place of biggest contours.

Lastly three co-ordinates of center point or starting point (or whatever required) of biggest drawn contour (over selected pixels) is computed. These accumulated tracked points vary from frame to frame. This sequence of points is sent to system's virtual world to be mapped over the movement of different virtual objects.

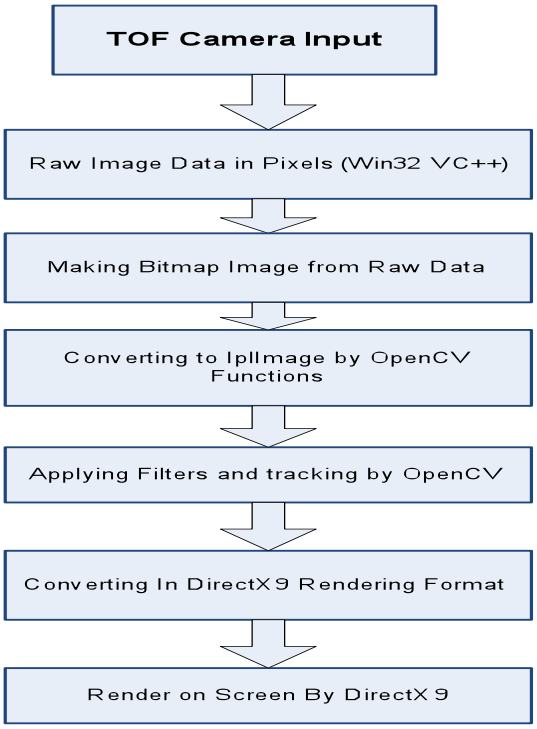

Fig. 6.1: Step by step processing of 3D data

7. Suitability of BNBC for PIAs

As described earlier public applications are to be installed mostly at crowded and packed places. BNBC assures robust and efficient 3D object tracking in defined ROI in those concerned places. With the support of BNBC algorithm 3D TOF camera for PIAs provides remarkably functional results in combination of range data from a sequence of frames. That can be implemented to track any object or (specific significant part of objects) in given ROI. Considerably less space is required to provide manage effective interaction with any multimedia application in some public place. By figuring out the distance from camera to object the BNBC continuously calculates the nearest pixels (a few hundred) that are why it is easy for it to ignore rest of the surrounding objects. Further it gauges and concentrates on the biggest of the contours being drawn around selected pixels. Then by identifying the finger tips and accumulating the x, y and z position of selected pixels of finger tips average position in ROI is computed to remove the outliers which make the tracked points more accurate and exact. This helps in tracking focused points in places which are swarming with people. One or maximum two can play at the same time and attract the attention of others around them.

8. Suitable Software Tools

To yield a real time inclusive programming solution implementing the BNBC algorithm, a complete cycle of R&D has been passed through to find out appropriate choices for development environment, language to code, image processing tool, front end presentation layer etc.

After experimenting a wide range of contrivances the following set of environments, programming languages and libraries were finally chosen to make the scheme of object tracking practicable in packed public environment.

1. Visual C++.NET along with Microsoft .NET 2008 is used to write code
2. By operating the hardware of TOF camera using C++ language 3D data is extracted in the form of frames (upto 60 frames per second).
3. Each frame contains raw bitmap data which is stored in one of the data structures of DirectX 9 e.g. PointList, TriangleList etc.
4. Using OpenCV[3] function get2D() and set2D() frame's Bitmap raw pixels data are saved in OpenCV's efficient data structure IPLImage object. [3]
5. Third dimension (Range information) belonging to each pixel is accumulated and saved in different C++ float variable. Besides that range data X and Y position of each pixels is also saved in different variables
6. Data is smoothed using numerous OpenCV functions i.e. cvErode(), cvDilate(), cvThreshold()and saved again in an IPLImage object. [3]
7. By performing various mathematical functions of C++ pixel data is filtered and sorted with respect to their distance from camera lens and brightness. Some steps of BNBC are performed. [3]
8. Then OpenCV API's such as cvFindContour() and cvDrawContour for contour detection play their roll to position and capture the motion of moving object. [3]
9. Finally using OpenCV functions cvCheckContourConvexity, cvConvexityDefect for convexity recognition along with depth of pixels the hand tracking is performed and exact 3D points's set against each frame is computed and sent to front-end platform. [3]
10. Unity3D is used to fulfill the roll of front-end applications and presentation layers.
11. 3D points in the form of direction vectors notifying the position of captured moving object in front of camera are directly sent to Unity's presentation layer. At this phase received 3D points are processed using C# language and provided to application layer of animated world which is interacted by user's real world.
12. At application layer, animated character can be navigated with help of received 3D points guided by captured hand motion from user's 3D environment.

    This whole cycle is run for each frame. And at least 30 times per second the animated character at application layer gets its new position, eventually making the mapping between user's 3D environment and virtual animated environment more natural. And user gets pleasure of new interaction by moving the animated character merely waving her hand without touching the screen as well as without holding any controller.

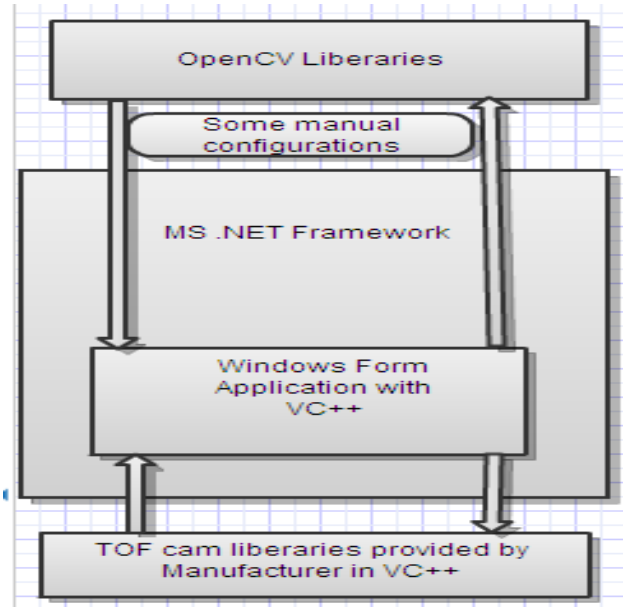

Fig. 8.1: Environmental configuration of programming tools

9. Future Work

Further future work may encompasses the augmentation of idea and development of progressively interesting applications providing pleasure of this novel interaction at public places. Following technical points can appeal to developers

- Gesture recognition
- Developing library of gestures dynamically
- Tracking and detecting multiple objects interacting in public environment.
- Envisioning the further social applications for public places entertaining a number of people at the moment with unique sound feedback to individual users while interacting with the system.
- By using the results of BNBC in CCD (contracting curve density) algorithm [6][16] of contour matching, shape recognition can be achieved robustly.

10. Conclusion

Research results of our previous paper "Designing Future Multimedia Applications for Public Environment using 3D TOF" [2] have led us to develop 3D applications for public places. Here it has been argued convincingly that the usage of TOF camera to implement BBNC in development of public interactive applications can bring a new and easy way of interaction between human and multimedia applications for entertainment and information purposes.

By this whole study and accomplishment it can be concluded that motion controlled applications for public places are very good and novel notions in multimedia world and by using BNBC their implementation is not an intricate goal anymore. Step by step illustration of BNBC algorithm to capture motion of objects in form of TOF camera a number of simple but pretty interesting and attractive applications can be developed just like aforementioned Island Explorer and Hand fishing. The usage of TOF depth sensor camera has proved to be an trouble-free choice for capturing motion in 3d public environments. With the help of these technical inferences Implementation of PIAs can open a gateway towards a modern social interaction not only in between human and multimedia rather it can provide social interaction amongst humans themselves. Process 3D point could delivered by BNBC can be used for other detection and recognition Algorithms e.g. CCD for enhanced shape recognition and gesture recognition.

11. REFERENCES (have not updated yet)

[1] S. Burak Gokturk, H. Yalcin, C. Bamji, "A Time-Of-Flight Depth Sensor – System Description, Issues and Solutions" Canesta Inc 2010.

[2] A. Waheed, M. Fahied, D. Eriksson, E. Borglund, "Designing Future Multimedia Applications", IEEE International Conference on Information and Emerging Technologies 2010.

[3] Bradski, G.R., Kaehler, A., (2008) "Learning OpenCV: Computer Vision with the OpenCV Library," ISBN-10: 0596516134

[4] Houssmanne, S and Edeler, T. 2009 "*Performance improvement of 3D-TOF PMD camera using a pseudo 4-phase shift algorithm,*" I2MTC2009, International Instrumentation and Measurement Technology Conference.

[5] Houssmanne, S., Ringbeck, T. and Hagebeuker, B., "*A performance review of 3D TOF vision systems in comparison to stereo vision systems*," in *stereo visio* , I-Tech Education and Publishing, Vienna, Austria.

[6] R. Hanek, T. Schmitt, S. Buck, M. Beetz, "Fast Image-based Object Localization in Natural Scenes" TU Munchen, Institute for Informatic, Boltzmannstr. 3, 85748 Garching b. Munchen, Germany. http://www9.in.tum.de/agilo/

[7] Canesta Inc. TOF camera manufacturer using the CMOS chip, http://www.canesta.com/ 20:00, 4 April 2010.

[8] Primesense Natural Interaction. Provided 3-D sensing technology to Microsoft Project Natal [14]. http://www.primesense.com/ 08:00, 4 April 2010.

[9] Wii Technology, http://www.wiitechnology.com/ .14:00, January 20, 2010.

[10] Microsoft (Kinect) Project Natal http://www.popsci.com/gadgets/article/2010-01/exclusive-inside-microsofts-project-natal. 11:30, January 25, 2010.

[11] Faugeras, O. *Three Dimensional Computer Vision: AGeometric Viewpoint. MIT press,Cambrigde Massechusets 1993*.

[12] Creative Summit June, 2010. Skellefteå Sweden. http://www.creativesummit.se/

[13] Unity, three dimensional game development tool http://unity3d.com/

[14] Three Dimensional Island Explorer Demo. http://www.youtube.com/watch?v=OFN9zVBQviQ&feature=channel

[15] Hitachi Full Parallax 3D TV http://3dguy.tv/hitachi-full-parallax-3d-tv/

[16] G. Panin, A. Ladikos, A. Knoll, "An Efficient and Robust Real-Time Contour Tracking System", IEEE International Conference Computer Vision Systems, 2006 ICVS '06.