# Motion Capture by Kinect

Karina Hadad de Souza Rosilane Ribeiro da Mota

Pontifícia Universidade Católica de Minas Gerais

## Abstract

This paper describes motion capture using Kinects. Kinect has built-in algorithm for recognition of skeleton of human body, but there are gaps when the joints are competing for the same area of sensor view. Each Kinect contributes to capture for positions that are unreachable by other, making graphical representation of body and movements done closer than other methods. The results showed that it is possible to use more than one Kinect to capture motion in short distance (less than 45º) and for large distance (about 135º) with complementary data, but it is not enough to have a good performance.

**Keywords**: Motion Capture, Kinect.

**Authors' contact**:
kahaddads@gmail.com
rosilane@pucminas.br

## 1. Introduction

One of principal valuable factors to produce animation movies and videogames, according Geroch (2004), is ability of a person to present or to describe what is registered/categorized on your mind, thus creating a 2D or 3D environment. One of ways that creators use to construct these environments is capturing real objects to reconstruct them virtually. They don't consider character's environment aspects and its movements.

Kinect according Shotton et al. (2011) is a dispositive with sensor to enable the interaction between people and computer systems like digital games. This devices can capture specify joint points from human body, but this capture has failed in collected data causing a wrong posture to represent the real one. Maybe, this is one of the reasons applications (specifically games) that using Kinect was simple and didn't require much ability, because Kinect infers the context to complex movements.

By default, Kinect tries to locate form of silhouette from human body when someone is standing up with raised arms in many games (Figure 1). Each time that process stops unexpectedly or something is getting wrong, it is necessary to stay in "T" pose again.

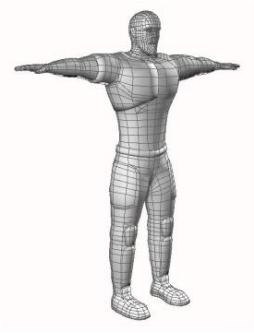

Figure 1: "T" pose

The main idea is minimize errors on data and/or body human position in the screen captured by Kinects. This can be used in more elaborated games if motion capture were done with more than one Kinect. So, it would be more accessible to have some motion capture equipment and more games could use this resource to do interaction more realistic between player and scenario elements.

The general objective from this paper is to propose a method that will use data captured by two Kinects to provide a projection more precision of position where the human body is.

## 2. Literature Review

The work of Flam et al. (2009) was one of the first considered motion capture work. It was done based on movement from a horse to find a standard, it starts with a image in determinate position and it keep capture in sequence until the image coincide with first frame to verify if there is a time that the horse don't touch the floor, Figure 2.

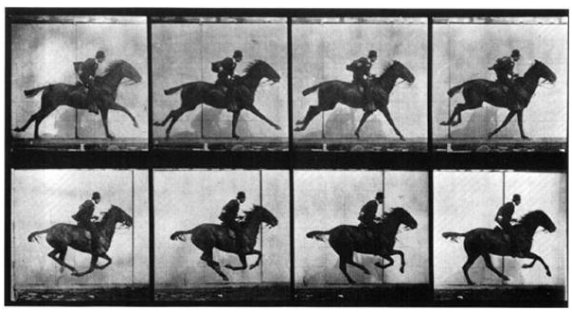

Figure 2: Motion Capture of a horse

The way of capture the environment information varies according objective that it wants to reach. Magnetics system according Geroch (2004) are sensor distributed in many part of body, that have coordinates (x, y, z)

mapped in a tridimensional camp. This information according Geroch (2004) are passed to a central responsible to capture and process. Optic systems are system observers, according Geroch (2004), subject from them is to suffer more extern interference as from objects that are captured as environment than magnetic system, but the capture is more comprehensive.

According Pullen and Bregler (2002), when someone moves with body, another parts can suffer movements – rotation, drivers – once they are connect by articulation. According Agarwal and Triggs (2004), different regions correspond approximately like fragmented movements, it is necessary recognize the actions individually and to translate the sequence to recognize the movement, as a whole.

In Pullen and Bregler (2012)'s work, it was done experiments with a kangaroo to do a reconstruction of their movement by markers identifying their actions and facility the process of data. The magnetic track system related in the work of Pullen and Bregler (2002) showed the inefficient of markers (sensors), because they were on over muscle. The calculations were done to extend the points to intercept and reconstruct the object. It generate errors when articulates were recognized. In a second attempt the capture in Pullen and Bregler (2002), the sensors were places over articulate and the results founds were better, the performance was more satisfied.

On the work Shakernia Vidal and Sastry (2002), an algorithm was developed to estimate the robot movement that carried with camera, it was one-dimensional vision. The robot was equipped with GPS sensors and the movements were made without forecast. We can identify the scenario and the objects according Shakenria Vida and Sastry (2002), by segmentation of corresponding pixels to differentiate aspects of the environment and objects. Features through of differentiation of color and intensity are identified by segmentation.

In (HORPRASET et al. 1998), the systems use the triangulation of image captured to recognize movements represented from draw, like cartoon. It is a combination of analyze of form with track techniques and detection of body, in other words, identification of part from body. In each instance from system, according Horprasert et al. (1998), were send data to a central of control, not only the parts of body identified but the value correspond which indicates a "more certain" of localization of each member.

According Sigal Bala and Black (2012), it is important to note the track as to estimate the pose can have difference performance in 2D, 2.5D and 3D camp, because they correspond in different form to capture and model the human body. According Sigal Balan and Black (2010), the 2D camp refers to model of parts of body that are defined directly in a plan image, while 2.5D keep the model that have depth relativity. Finally, Sigal Balan and Black (2010) define the 3D camp typically as model of human body in three dimensions with spherical parts, rounded and cylindrical.

Kinect capture by a laser grid of pixels that are emitted on environment and the calculation of articulations points is done by interpolations of intercepted points on grid. Kinect according Shotton et al. (2011) have five main characteristics, Figure 3:

- RGB camera that permits facial recognized
- Depth sensor
- Build-in microphone (Audio process - no record)
- Software and API owner
- To detect 48 points of articulation form human body (skeleton reconstruction)

By grid, Kinect also have the characteristic of capture the depth of observed object. Figure 4, the Kienct has a minimum distance to range the object and limits.

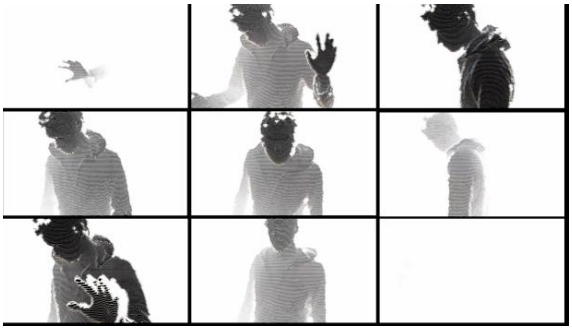

Figure 4: Depth capture

Capture by Kinect how Figure 5, do it in three dimensions (x, y, z) and it software identify specific points of human body, this mean, identify silhouette.

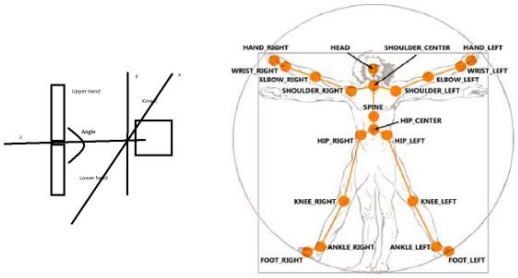


Figure 5: Dimensions and points captured by Kinect.

## 3. Methodology

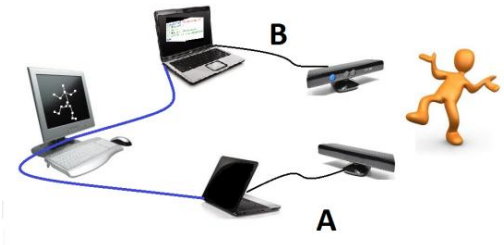

Figure 6: Communication by Kinect

To do this work, it was used two Kinect. In Figure 6, are represented the Kinects and the computer used in a structure constructed to capture.

The work was done in C#, Object Oriented paradigm, due compatibly with software used by Kinect, *Software Development Kit* (SDK), developed by Microsoft, who distributed the sensor. Recourse to aid in this study to communicate between two sensors were developed a Socket application, using *request/response* technique Figure 8.

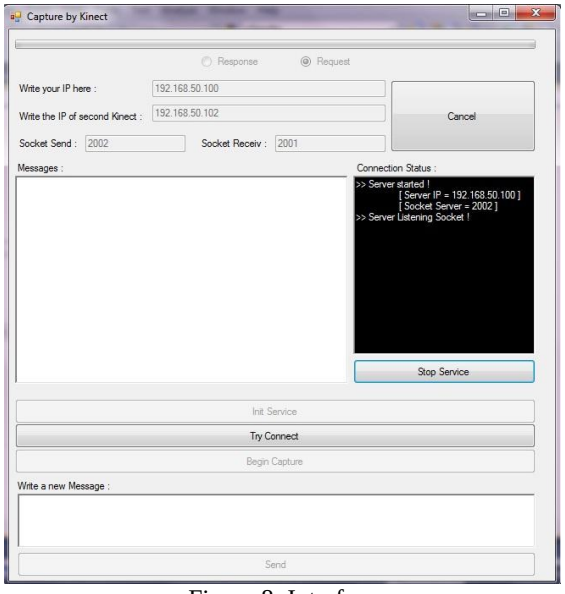


Figure 8: Interface

The "A" set receive this information and capture movements of player, too. For each frame captured, it verifies the difference between coordinate (x, y) from two points is less than three, five, seven and ten, because it is possible that they are in the same direction. The "A" set request and consult the vector field of points captured by "B". Case, the vision camp of Kinect of "B" set, the both points identified have no next abscises, less than three, and the application concludes that points are in the same direction by vision of "A".

Who requires stars to draw on screen the corrected points in a step before, more precise, ensuring the integrity of real data. This transmits more ratability on the skeleton reconstruction in vision of player.

## 4. Test and experiments and analysis of results

The platform used was Windows 7. Although the documentation about SDK to const the possibilities to use more than one Kinect in only one PC, the driver used to control the sensor didn't support two Kinects connected in distinct *Universal Serial Bus* (USB) ports in the same computer. The solution was to use one computer with one Kinect and other computer with a second dispositive, changing the data.

After done this analysis, the next step was developed a method to change the data. The application was done using Socket using request/response. This technique was necessary due the quantity captured by one Kinect in a period of time – middle 30 frames per second – has poor performance and stopped the application, invalidating a future analysis of data. So it was use a control to require this package, it is, the data are not sending every time anymore, only when a set, Figure 6, require the package.

When the captures is done, the both Kinects was positioned in distinct angle and to analysis how much efficient is the method developed, it was tested four distance different between dispositive. The first it was 45º, later 90º, 135º and 180º.

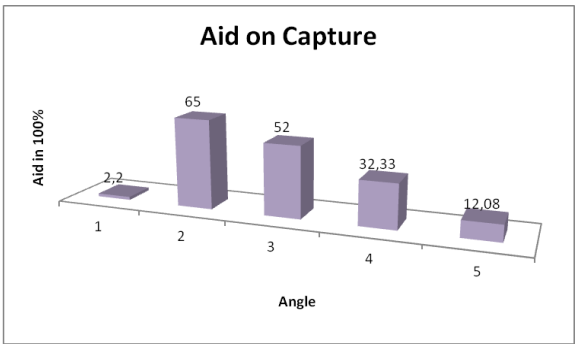


Figure 10: Aid on Capture

In the graphics of Figure 10 it can analysis how much aid in the reconstruction when change the positions of Kinects. On Figure 11 the green points do refer to elbow, knee and foot. There are letters next of each sphere R – red, Y – yellow and G - green.

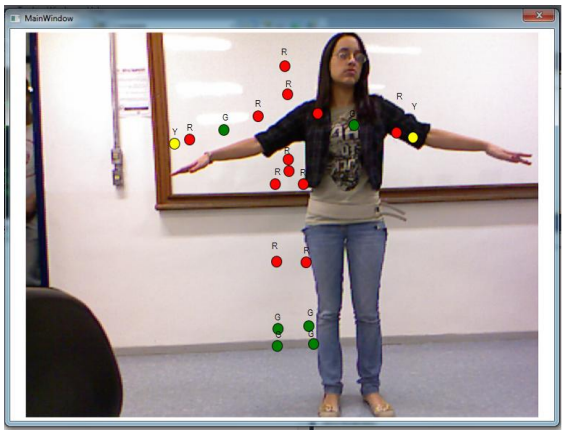


Figure 11: Identify some points of articulation

In other moment, withdraw some spheres that were doing refer the body but hands, the yellows spheres were misrepresenting the elbow. When it was return with original code, the yellow spheres were doing refer

to hand. With this, we can conclude the inconsistence of data into software from application that weren't confinable to selections less points to be sample data.

During the tests, it was possible to percept that method was more effective to little angles between Kinect Figure 12. When the experiments was done angle 45º, the performance was better than angle 135º. Because the points were too close.

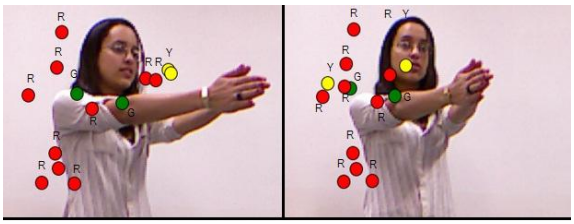

Figure 12: With improve and improvement

When in developed method the condition was done by distance between the points the performance was good. But when the difference between points were calculated in pixels the performance was late, because the information were data from an image which difficult the identification, while the other the information are data from software Kinect.

It also can percept it was insignificant the time spent to change the data. Before to implement the control of requisition of points, the quantity of capture were so much that stop application.

## 5. Conclusion and future works

The results shown that it is possible to improve the precision of Kinect capture when it is used more than one dispositive. But the efficient is identified when other Kinect is closer of one. In a future work it can be mentioned the use more Kinects to aid capture, beyond more computer to change the package of data. In the same time, it will be necessary to careful with late of communication among computer, once the answer of application should work in real time.

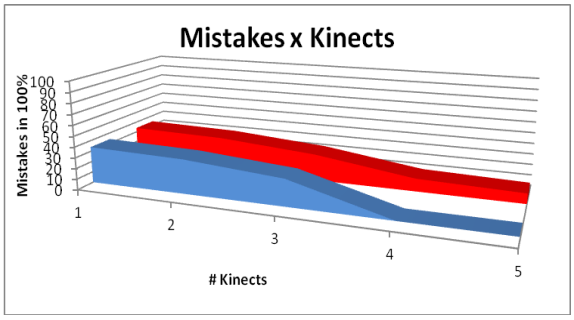


Figure 13: Mistakes x Kinects

On the graphics form Figure 13 it can see the blue about three Kinects, on the definitions the process of data in function of distance of points, this mistake tends to decrease. The red late of use three Kinect, on the graph it is shown the probability of mistakes when use four or five Kinects, and late of this the level of mistakes tends to be next.

Beyond to take in consider the quantity of Kinects, other kind of test that can be done it is the position of Kinects in different distance from player in the same direction. The capture of depth can take some positive use, like in cases of approach. After change of message and capture the data a next step will apply depth fundamentals, its mean, the aspects that represent idea of tridimensional capture. Thus, a 3D modeling of object it being mapping.

## References

FLAM, D. et al. Openmocap: An open source software for optical motion capture. In: IEEE. *Games and Digital Entertainment (SBGAMES), 2009 VIII Brazilian Symposium on*. [S.l.], 2009. p. 151–161.

GEROCH, M. Motion capture for the rest of us. *Journal of Computing Sciences in Colleges*, Consortium for Computing Sciences in Colleges, v. 19, n. 3, p. 157–164, 2004.

HORPRASERT, T. et al. Real-time 3d motion capture. In: *Second workshop on perceptual interfaces*. [S.l.: s.n.], 1998. v. 2.

PULLEN, K.; BREGLER, C. Motion capture assisted animation: texturing and synthesis. ACMTrans. Graph., ACM, New York, NY, USA, v. 21, p. 501–508, July 2002. ISSN 0730-0301. Dispon´ıvel em: <http://doi.acm.org/10.1145/566654.566608>.

SHAKERNIA, O.; VIDAL, R.; SASTRY, S. Infinitesimal motion estimation from multiple central panoramic views. In: IEEE. *Motion and Video Computing, 2002. Proceedings. Workshop on*. [S.l.], 2002. p. 229–234.

SHOTTON, J. et al. Real-time human pose recognition in parts from single depth images. In: *CVPR*. [S.l.: s.n.], 2011. v. 2, p. 7.

SIGAL, L.; BALAN, A.; BLACK, M. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, Springer, v. 87, n. 1, p. 4–27, 2010.