Nicholas Jones
English 3302
Tom Akbari
Unit 1 Final Draft
January 26, 2015

Analysis of "Real-Time Human Pose Recognition in Parts from Single Depth Images"

OVERVIEW

"Real-Time Human Pose Recognition in Parts from Single Depth Images" presents research into a new algorithm that allows computers to map 3-dimensional human skeletons to images of people in real time. This incredibly powerful ability played a key role in the creation of Microsoft's Xbox Kinect camera, which is used by researchers, engineers, and gamers around the world [1(pg. 116)]. The paper was written by Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Andrew Blake, Mat Cook, and Richard Moore following their research at Microsoft, and it was released in Volume 56, Issue 1 of "Communications of the ACM" in January 2013. Although the subject is highly technical and the research is targeted towards computer vision researchers, the authors' use of simple organization and colorful graphics make the paper clear and understandable.

ORGANIZATION

The paper is organized in a straight-forward manner. First, it summarizes the problem, prior work in the field, the algorithm, and how the algorithm improves in pre-existing techniques. This is important, since it lays the foundation that the rest of the paper can build upon. Once readers can relate this work to other work in the field, they have a sense of the direction and value that this research provides.

Next, the paper discusses the related data—a series of depth images* obtained from motion-capture data [1(pg. 118)]. Additionally, it describes how this initial data was used to generate a much larger, more diverse set of synthetic data, which was a further improvement on prior techniques and allowed them to train their machine learning models with greater accuracy [1(pg. 119)]. By discussing this up front, they give the reader a starting point that they can picture and modify as the algorithm progresses.

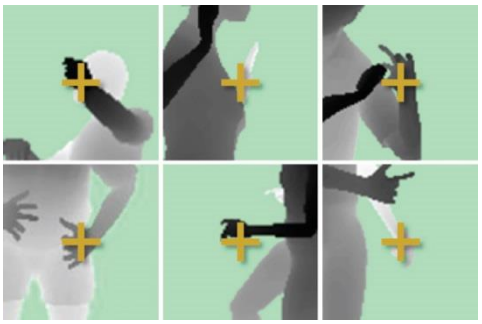

Figure 1: Figure 1 from paper shows synthesized depth images. The image was used to show the ways a single body part can appear in various poses [1 (pg. 118)].

Over the next few sections, the paper progresses through the algorithm explaining the tools, equations, and algorithms used to map human skeletons to depth images in real time. Each section presents a single, distinct step, following it by addressing concerns and issues that the researchers encountered during their implementation. By addressing these concerns here, the authors expose their thought processes and guard against critical analysis from the research community. These sections contain the meat of the paper, and, without them, the paper would provide little valuable information to the field.

After explaining the algorithm, the authors show "both qualitative and quantitative results on several challenging datasets and compare with both nearest-neighbor approaches and the state of the art. [1(pg. 121)]" They clearly explain the test data they used to perform these tests and the error metrics that they applied. This section gives authority to the rest of the paper by proving that the algorithm works better than other techniques for the intended application.

* An image where each pixel represents the distance to the first object encountered at that pixel.

In the final sections of the paper, the authors discuss the implications of their research, potential improvements, and the direction that they may move with future research. They thank the team of engineers that implemented their algorithm on the Xbox Kinect, and the collection of other researchers that helped them by participating in discussion and providing test data. Doing so, the authors give authority to their work by indicating their pedigree and status in the community.

Overall, the organization of the paper makes the algorithm very easy to understand. By first summarizing their work, then linearly moving though their algorithm, anyone with a basic understanding of the field can get a good idea of their research.

GRAPHICS

Almost every page of the article contains some form of graphic. Many of these are representations of the algorithms used. For example, on the fourth page, figure 3 displays a collection of synthesized character models generated with the same pose but varying height, weight, hair, and clothing. Due to the incredibly graphic nature of this research, these images are invaluable, since they directly correlate to the steps of the algorithm, and they allow the reader to visualize how the algorithm manipulates the data. An additional benefit of these images is that they brighten the otherwise black and white pages with a little color.

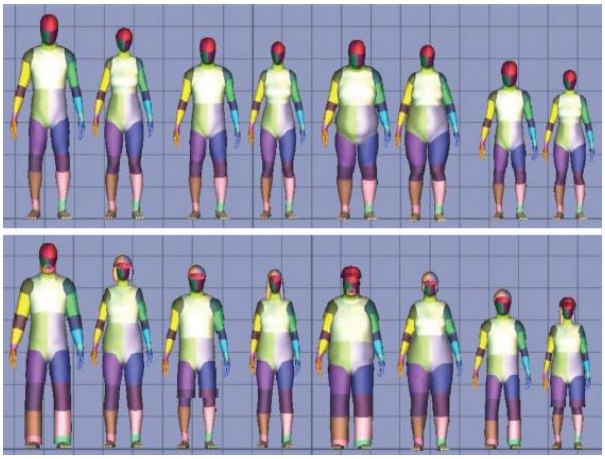

Figure 2: Figure 3 from paper displaying synthesized variations of a pose [1 (pg 119)].

Another form of graphic used fairly often are the mathematical equations used to interpret the data in the algorithm (figures 4-7 in paper). These equations are incredibly important, since they act as keys to the algorithm—without them it would work very differently. By breaking text around these equations and using clear mathematic notation, the authors clarify and draw attention to this valuable information.

Finally, in section 4, the authors use a series of line and bar graphs to compare the speed and accuracy of their research to other methods commonly used. Again, these graphs add color to the page, and they make otherwise abstract numeric notation easier to understand and use.

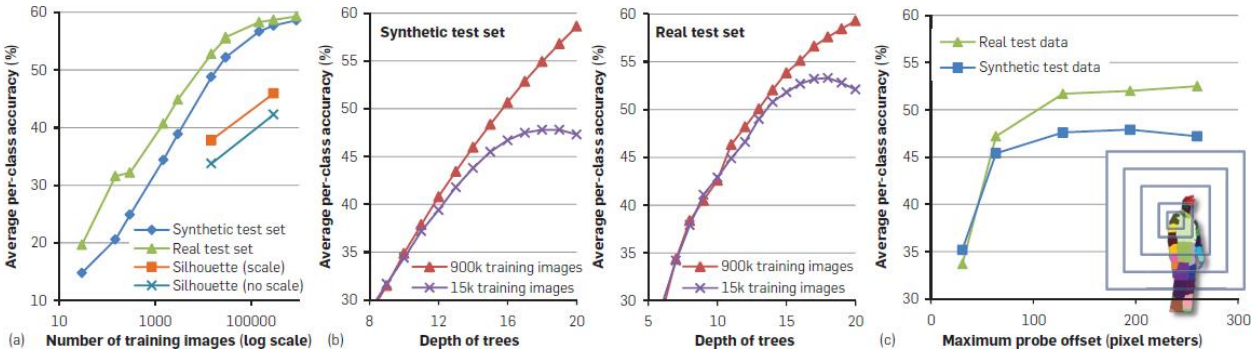

Figure 3: Figure 8 from paper illustrates the accuracy of the algorithm using various configurations [1 (pg 122)].

ASSUMED KNOWLEDGE/AUDIENCE/PURPOSE

This paper is very technical, since it is targeted towards other researchers and engineers in the field of computer vision with the intention that they should be able to use, implement, or build upon the algorithm after finishing the paper. Without a reasonable understanding of computer science, machine learning, computer vision, and mathematics (linear algebra, statistics), many of the concepts discussed are hard to grasp. For example, the paper's description of Random Decision Forests in section 3.3 is not intended to completely cover the concept, since the authors expect the intended audience, CV researchers, to have prior knowledge of such a well-known algorithm. They summarize the algorithm and provide the

equation they used with it; however, to completely understand how Random Decision Forests work, one would need to perform further research. This is also true of section 3.4, where the authors describe their use of "mean shift" to "generate reliable proposals for the positions of 3D skeletal joints [1(pg. 121)]." The information provided is clear, but without a general knowledge of mean shift, the step would be very difficult to implement.

REACTION

This paper meets my expectations of the field and conforms to the style of other papers that I have read from various fields in computer science. Although it was very technical, I was able to understand all but the most complicated ideas in the paper, and if I needed to implement it, I could track down the information I lack to fill the gaps in my understanding. I didn't encounter academic papers on my first co-op; however, my classes are preparing me to read and write papers of this genre, since they often require me to research, implement, and respond to them for homework. My current writing skills are more than sufficient to write papers like this. While the concepts are complicated, the language and organization are mundane. To write as an authority in this genre, it seems important that you are able to succinctly explain complicated processes and relate them to other research in the field. As I move further into my education and participate in research, I have little doubt that I will attain the literary and technical knowledge to achieve this.

CONCLUSION

"Real-Time Human Pose Recognition in Parts from Single Depth Images" is a highly technical research paper written for experts and researchers in the field of computer vision and machine learning. Although its subject is complicated, straight-forward organization and

colorful graphics present the information in the clearest manner possible. This allows the paper to be consumed and used easily by the rest of the community.

REFERENCES

1. Shotton J, Sharp T, Kipman A, Fitzgibbon A, Finocchio M, Blake A, Cook M, Moore R. Real-time human pose recognition in parts from single depth images [Internet]. Communications of the ACM. 2013 Jan [Cited 2015 Jan 20];56(1):116-124. Available from: http://doi.acm.org/10.1145/2398356.2398381