Distance Metrics for Finger Pose Similarity

Erin Walter
Mills College
ewalter@mills.edu

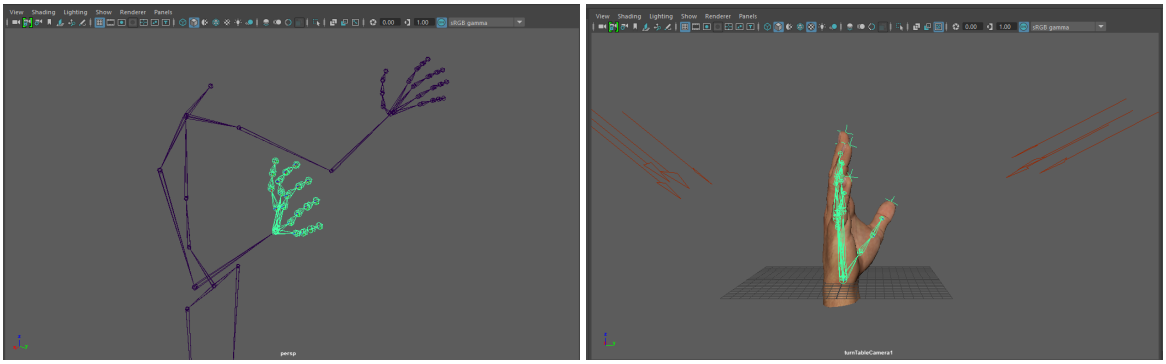


Figure 1: Exploring finger poses in *Maya*

1 REPORT

This report provides a summary of my research conducted during the Summer of 2017 at Clemson University's Computer Animation and Motion Perception Lab, as part of the Distributed Research Experiences for Undergraduates. At Clemson University's School of Computing, under supervision of Sophie Jörg, we seek to find the impacts of different types of errors on our perception of finger motions. This report adds to the body of work using joint rotation distances, by finding root mean distances between poses, and also adds supplemental modeling of different finger poses.

2 INTRODUCTION

In computer animation, hand and finger motions are typically created manually, which can be tedious. Hand and finger motions are extremely important in daily interactions, but they are difficult to capture. Clemson University is working to develop algorithms to automatically create finger motions, and synthesize body and finger motions using data-driven techniques. One part of that research is to find the impact of different types of errors on our perception of finger motions. Very slight errors in synchronizations in body and hand motions can be detected by the human eye [5]. Jöerg's research aims to improve the motion capture post-processing step by adding data-driven synthesis for hand gestures, adding to the efficiency of animations, and improving human perception of finger motions [5].

Different finger poses may be perceived as more or less similar, and this presents an interesting problem: How do we quantify differences in similarity, and how do we train a machine to learn what is similar or not? By exploring different methods of quantifying distances, and validating them by human perception, we are closer to solving this problem. In this report, I explore methods of motion capture, segmentation, finding distances between clips, and finding distances between finger poses. A glimpse of the skeletons and models used in animation and static poses respectively in *Maya* can be seen in Figure 1. At the beginning of the REU experience, we planned to add realistic textures and geometry to the hands of the left-hand skeleton in Figure 1, and conduct a perceptual experiment between animation clips. However, due to time constraints, I instead used a previously modeled hand and adjusted it to different static poses outside of a previously used database, and compared root mean distances between poses.

3 RELATED WORK

There currently exists a wide and growing foundation of different methods to find the similarity of body gestures, as well as methods for synthesizing finger motions. These findings include various methods for segmenting clips, determining distances between poses and distances between motions, and ultimately synthesizing animations.

Tang *et al.* [4] proposed a distance metric based on joint relative distance to judge whether two given postures appear similar or not, however this approach uses the full body to determine similarity of postures, and does not focus solely on joint distances of the fingers using Euclidean distance. Chen *et al.* [2] proposed *Relational Geometric Distance*, which accumulates the differences over a set of features that reflects the geometric relations between different body parts. Again, this approach focuses on the distances between all body

parts, while comparing *Joint Rotation Distance*, *Joint Position Distance*, and *Joint Relative Distance*, but does not detail hand and finger distances. It is known that using a whole-body motion, it is difficult to predict the exact motions of the hand. Because of this difficulty, work exists that finds correlations between wrist motion and selecting plausible finger motions.

Jörg *et al.* [6] proposed a simple solution based on a metric for estimating the motion of a character's fingers by assigning weight factors to the wrist position and orientation, used as the input parameters in finding similar body motions in a database. Building on this work, Mousas *et al.* split the estimation process into sub-problems (gesture phase estimation, gesture type estimation, and motion retrieval), resulting in a decrease of the computational time that is required for retrieving the final motion [8]. Other works such as Ye and Liu [10] focus on contact, sliding, and relocation of finger joints.

Wang *et al.* [9] explores a semantic-correlation-strength based distant metric learning for 3D model retrieval, utilizing a Mahalanobic distance and Relevant Component Analysis (RCA). In practice, we often use Euclidean distances as similarity metric to calculate retrieval results of 3D models [9].

In my research, I use Euclidean distance to compute distance between frames, and root mean distance between poses, in order to explore average distances and visualize how they may relate to one another. The different gestures can then be perceived by humans to be more or less similar. These methods are explained further in the following sections.

4 METHODS

Motion Capture

During my research, I learned about motion capture, with an emphasis on post-processing. During post-processing, markers are labeled on joints of a skeleton, as seen in Figure 2, to represent body and finger joints, which move during animation. Data collected from the motion capture is applied to the joints via processing software such as *Vicon Blade* and *Vicon Nexus*, and movement is created as a result. The process of labeling finger joints may seem simple, however, due to label swapping and marker occlusion, it can be difficult to maintain proper position of the joints throughout the length of an animation. Due to the difficulty of this process, research in finger motion synthesis to a body's movement is a crucial and important problem to address.

Using the *Vicon Blade 3.4* software, I successfully labeled the body and finger motions of the skeleton from existing data from a motion capture session. Once a skeleton is successfully labeled, gaps in information can be filled either via the software or manually to create fluid motions. Many of these motions appear to 'jitter' due to movement noise or inaccuracies in interpolation. Once gaps have been filled, the last step is to export the skeleton figure in a .fbx format to be modeled and textured in a 3D software such as *Maya* or *Unity* for rendering.

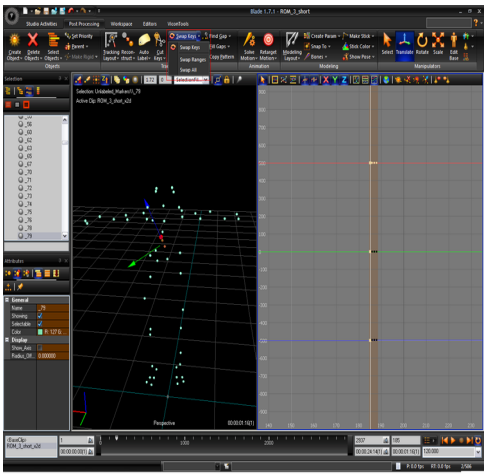


Figure 2: Post-processing marker labeling in *Vicon Blade 3.4* [1]

Segmentation

Segmentation is used to segment clips into different phases of a gesture. In my application of segmentation, I considered only finding the active phase of a gesture in order to obtain information regarding the finger motions and rotations, and calculate distances between joints and clips. Using a similar method presented by Jörg [6] for segmenting clips, where plausible finger motions can be inferred from the wrist motion, a threshold for wrist speed is chosen to determine which frames represent the active phase. I chose a speed of 40 as a threshold for average wrist speed, in order to maximize the number of frames present. Most active phases resulted in 30 frames per second.

Segmentation was calculated by first finding wrist speed by computing the difference in global wrist position throughout the length of the clip, and was visualized in *matplotlib* as seen in Figure 3. The visualization of wrist speed was helpful in determining which clips and gestures I wanted to use from a previously recorded gesture database, specifically the "Large Gesture Database" in Sophie Jörg's 2012 paper [6]. Then, segmentation code is applied to wrist speed, resulting in a list of possible frames for the active phase. The list of frames were validated by manually finding the start and end frames and verifying the phase. The active phase was then used as the range of frames in which joint rotations were extracted.

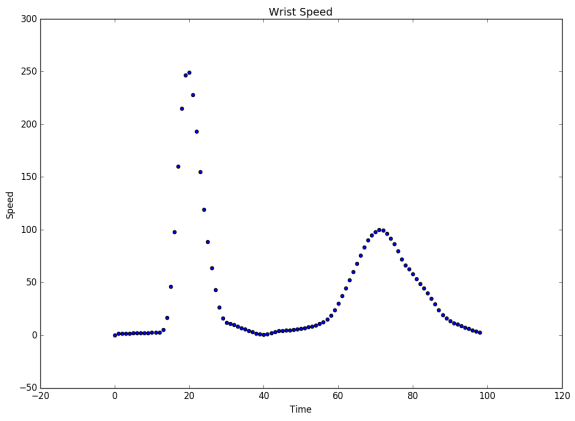


Figure 3: Wrist speed curve for 'Attention' gesture

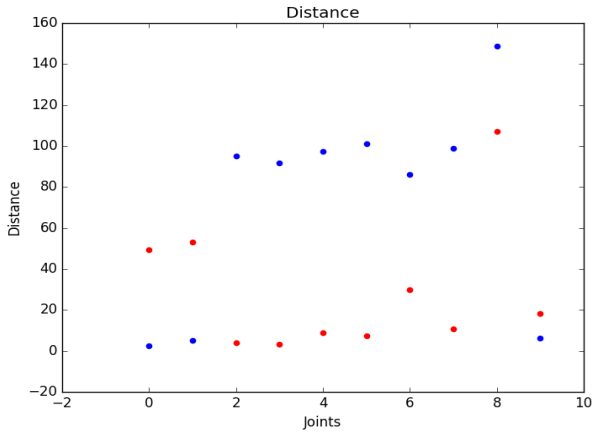


Figure 4: Distances between joint rotations: 'Ok' vs. 'Small'

Distance Metric for Animations

For computing distances between clips, clips were first segmented, in order to pinpoint different phases of a gesture. Then, using the determined active phase, the rotations for finger joints were extracted for that phase via *Maya* Scripting Window using Python and MEL commands. The differences between joints uses a Euclidean distance formula summarized in Du Q. Huynh's 2009 paper, *Metrics for 3D rotations: Comparison and Analysis* [3].

Minimum x, y, and z. Values for x, y, and z are found by finding the lesser value of:

$$(2 * \pi) - abs(x_1 - x_2)$$

or

$$abs(x_1 - x_2)$$

Euclidean Distance. Euclidean distance is calculated as follows from x, y, and z location joint rotations per frame:

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

My method for using scripting vs. manually interfacing *Maya's* UI was influenced partially by my unfamiliarity with the *Maya* UI, but primarily to obtain a reproducible method for extracting rotations, which can be applied to multiple clips. Differences between frames were found using the Euclidean distances from one gesture's joint rotations to another. Differences between clips that were slowed down and sped up, as well as distances between clips where rotation offsets have been added to a gesture, were also found. Time did not permit for calculating the mean distances of those differences. [1]

[1]Supplemental data: https://github.com/erinbagel/dreu.git

Distance Metric for Finger Poses

Code for extracting local rotations was applied to each joint, the metacarpophalangeal (MP), the proximal interphalangeal (PIP), and the distal interphalangeal (DIP) [7] on each finger: index, middle, ring, pinky, and thumb. Euclidean distance is used to find the distance between each of the joints. Including the thumb, this results in 10 distances consisting of: thumb MP to PIP, thumb PIP to DIP, index MP to PIP, index PIP to DIP, middle MP to PIP, index PIP to DIP, ring MP to PIP, ring PIP to DIP, pinky MP to PIP, and finally pinky PIP to DIP. Root mean square average between different finger poses as seen in Figure 7 are then calculated.

Root Mean Square Average. Root mean average is calculated as follows between two poses:

$$\sqrt{(i)^2 + (i+1)^2 + ... + (i+n)^2/(indices)}$$

These averages change between each pose comparison, because the representation of the local rotation of each joint changes. We can see the distances between joints in Figure 4, where 'Ok' and 'Small' are not very similar. In comparison, in Figure 5 the two gestures 'Big' and 'PP' are very similar perceptually and have seemingly smaller distances between joints. In Figure 6, we see the root mean or average distance between all finger poses compared against the 'Attention' gesture. Additional comparisons are explained in the section Evaluating Distances Between Poses.

Textures and geometry from a hand model were also applied to the finger poses in order to compare gestures in a perceptual experiment for future use, as seen in Figure 7. Animated turntables of each pose were also created. [2]

[2]Supplemental movies: http://bit.ly/2vK5B2M

	atten-tion	big	cow	fist	hitch hike	ok	peace	point	pp	pp bend	rock n roll	shoot	small	thumb up
atten-tion	0	60.38	39.62	35.72	33.43	62.94	46.35	20.30	53.66	47.62	24.06	18.72	19.72	39.92
big	60.38	0	59.46	53.69	51.76	21.33	51.80	53.62	8.42	17.55	55.75	51.70	66.86	59.74
cow	39.62	59.46	0	31.69	30.89	51.11	56.38	44.23	52.84	44.20	35.18	42.95	50.51	31.60
fist	35.72	53.69	31.69	0	34.50	46.7	34.99	35.78	47.83	38.97	44.58	40.27	47.19	38.82
hitch hike	33.43	51.76	30.89	34.50	0	45.26	52.97	28.26	45.57	40.69	37.10	26.84	42.37	14.48
ok	62.94	21.33	51.11	46.7	45.26	0	51.73	54.70	20.65	20.89	58.87	54.82	70.86	51.22
peace	46.35	51.80	56.38	34.99	52.97	51.73	0	38.33	47.44	42.69	54.84	45.46	52.64	57.33
point	20.30	53.62	44.23	35.78	28.26	54.70	38.33	0	47.18	42.48	31.87	15.02	30.67	36.68
pp	53.66	8.42	52.84	47.83	45.57	20.65	47.44	47.18	0	11.22	48.76	45.22	60.72	53.72
pp bend	47.62	17.55	44.20	38.97	40.69	20.89	42.69	42.48	11.22	0	42.23	42.13	56.71	49.45
rock n roll	24.06	55.75	35.18	44.58	37.10	58.87	54.84	31.87	48.76	42.23	0	29.67	36.59	45.43
shoot	18.72	51.70	42.95	40.27	26.84	54.82	45.46	15.02	45.22	42.13	29.67	0	23.90	33.46
small	19.72	66.86	50.51	47.19	42.37	70.86	52.64	30.67	60.72	56.71	36.59	23.90	0	44.82
thumb up	39.92	59.74	31.60	38.82	14.48	51.22	57.33	36.68	53.72	49.45	45.43	33.46	44.82	0

color	range
	1.0 - 15.0
	15.0 - 20.0
	20.0 - 25.0
	25.0 - 30.0
	30.0+

Table 1: Top 16 most similar gestures by root mean square average

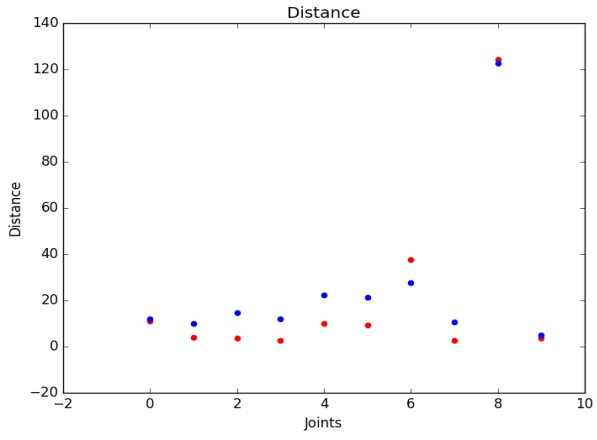


Figure 5: Distances between joint rotations: 'Big' vs. 'PP'

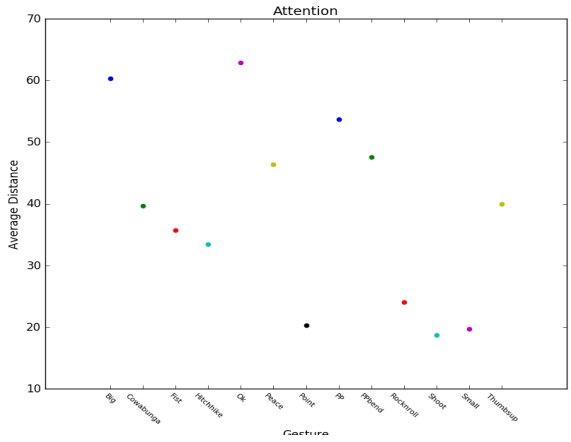
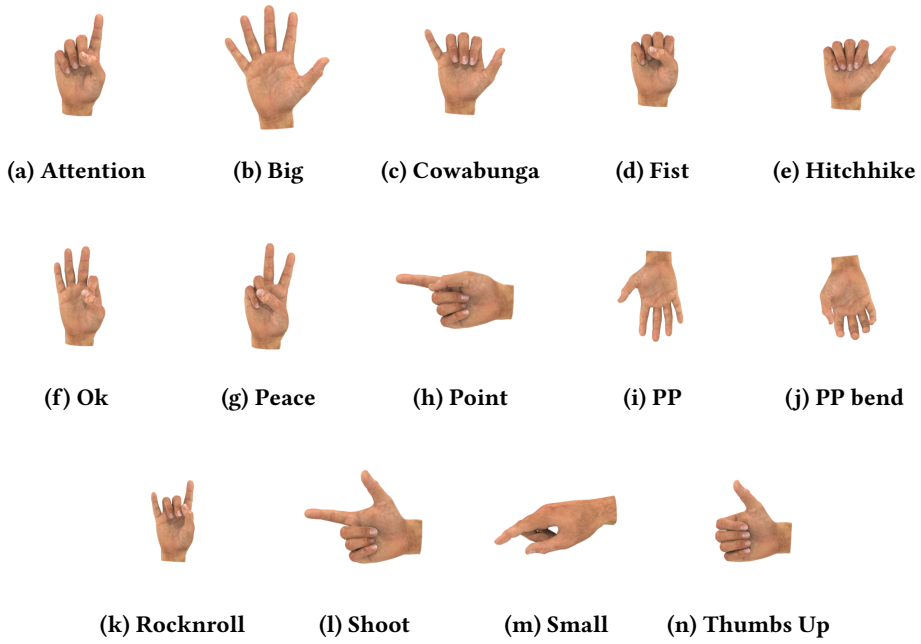


Figure 6: Root mean average comparisons for 'Attention'



(a) Attention (b) Big (c) Cowabunga (d) Fist (e) Hitchhike

(f) Ok (g) Peace (h) Point (i) PP (j) PP bend

(k) Rocknroll (l) Shoot (m) Small (n) Thumbs Up

Figure 7: Finger poses

5 EVALUATIONS AND RESULTS

Evaluating Distances Between Clips

Rather than finding Euclidean distances per joint and then differences per frame, it may have been more useful to find the root mean distance between clips. Instead of finding distances between individual joints for each gesture and comparing them, I could have calculated the root mean distance between frames. This could help visualize the differences in rotations much better, which proved to be useful in the previous section. Finding finger joint velocities as another feature in addition to rotations also may help find stronger correlations between distances.

Evaluating Distances Between Poses

Figure 5 gives an example of the visualizations of distances between joints, while Figure 6 visualizes root mean square comparisons between poses. We can learn how each gesture compares to another with this visualization, however an overall comparison in detail is visualized in Table 1. The results of the Euclidean Distance and root mean as a metric look promising, as seen in Table 1, however there are results that require further investigation. For example, the Top 16 most similar gestures based on a low average number may be validated as perceptually similar, however the gestures 'Cowabunga' and 'Hitchhike', which may rank as perceptually similar, do not have as low of a number as others like 'Shoot' and 'Hitchhike'. Although the number is still relatively low, it is unclear why certain gestures may be more or less similar based on average alone. Ideas for improving the results are discussed in the next section.

6 CONCLUSIONS AND FUTURE WORK

For finding a Distance Metric for Animations, the next phase was to add geometry to the skeleton, in order to run a perceptual experiment. Unfortunately, fixing the local orientation of each joint is time consuming, and this portion of the project ceased. In regards to collected distance data, the method for finding root mean distance for each frame may have been more helpful, rather than the method of finding Euclidean distances of joints for each frame and computing differences. Additionally, other distance formulas exist that are worth exploring, such as rotation axes using quaternions and deviations from identity matrix [3]. Other features would be worth including, such as classification or adding weights, to obtain a better correlation perceptually [4]. For example, maximum velocity of finger joints, or velocities of specific fingers may influence motion perception.

Regarding a Distance Metric for Finger Poses, joint distances and root mean average appears to provide perceptually validated results. For example, 'Big' and 'PP' gestures have a low average distance, and appear to be very similar. However, to learn which percentage of the results are perceptually validated, the planned human perception experiment would have been useful. This kind of experiment and further

exploration on the impact of errors is highly suggested for future work, in addition to other methods for computing distances.

ACKNOWLEDGMENTS

The author would like to thank Dr. Sophie Jörg for providing advising and inspiration, and the matlab code of the segmentation. The author would also like to thank Moshe Bitan and Isabela Figueira for help with Maya and Numpy.

The author would also like to thank the Computer Research Association for Women and the National Science Foundation for this opportunity.

REFERENCES

[1] Clemson Digital Production Arts. 2017. Motion Capture Pipeline. Wiki. (2017). Retrieved August 11, 2017 from http://wiki.fx.clemson.edu/mediawiki/index.php/Motion_Capture_Pipeline

[2] Jun Xiao Cheng Chen, Yueting Zhuang and Zhang Liang. 2009. Perceptual 3D pose distance estimation by boosting relational geometric features. *Computer Animation and Virtual Worlds* 20 (May 2009), 267–277. https://doi.org/10.1002/cav.297

[3] Du Q. Huynh. 2009. Metrics for 3D Rotations: Comparison and Analysis. *J. Math. Imaging Vis.* 35, 2 (Oct. 2009), 155–164. https://doi.org/10.1007/s10851-009-0161-2

[4] Taku Komura Jeff K. T. Tang, Howard Leung and Hubert P. H. Shum. 2008. Emulating Human Perception of Motion Similarity. (Aug. 2008), 211–221 pages. https://doi.org/10.1002/cav.260

[5] Sophie Jörg, Jessica Hodgins, and Carol O'Sullivan. 2010. The Perception of Finger Motions. In *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization (APGV 2010)*. ACM, New York, NY, USA, 129–133. https://doi.org/10.1145/1836248.1836273

[6] Sophie Jörg, Jessica K. Hodgins, and Alla Safonova. 2012. Data-driven Finger Motion Synthesis for Gesturing Characters. *ACM Transactions on Graphics* 31, 6, 189:1–189:7. https://doi.org/10.1145/2366145.2366208

[7] Sophie Jörg and Carol O'Sullivan. 2009. Exploring the Dimensionality of Finger Motion. In *Proceedings of the 9th Eurographics Ireland Workshop (EGIE 2009)*. 1–11.

[8] Christos Mousas, Christos-Nikolaos Anagnostopoulos, and Paul Newbury. 2015. Finger Motion Estimation and Synthesis for Gesturing Characters. In *Proceedings of the 31st Spring Conference on Computer Graphics (SCCG '15)*. ACM, New York, NY, USA, 97–104. https://doi.org/10.1145/2788539.2788552

[9] X. Wang, S. Wang, and H. Pang. 2011. Distance Metric Learning Based on Semantic Correlation Strength for 3D Model Retrieval. In *2011 International Conference on Multimedia and Signal Processing*, Vol. 1. 334–338. https://doi.org/10.1109/CMSP.2011.74

[10] Yuting Ye and C. Karen Liu. 2012. Synthesis of detailed hand manipulations using contact sampling. (July 2012), 10 pages. https://doi.org/10.1002/cav.260