2008

# Motion classification using Dynamic Time Warping

Kevin Adistambha
*University of Wollongong*, ka07@uowmail.edu.au

Christian Ritz
*University of Wollongong*, critz@uow.edu.au

Ian Burnett
*Royal Melbourne Institute of Technology*, ianb@uow.edu.au

# Motion classification using Dynamic Time Warping


## Abstract

Automatic generation of metadata is an important component of multimedia search-by-content systems as it both avoids the need for manual annotation as well as minimising subjective descriptions and human errors. This paper explores the automatic attachment of basic descriptions (or dasiaTagspsila) to human motion held in a motion-capture database on the basis of a dynamic time warping (DTW) approach. The captured motion is held in the Acclaim ASF/AMC format commonly used in game and movie motion capture work and the approach allows for the comparison and classification of motion from different subjects. The work analyses the bone rotations important to a small set of movements and results indicate that only a small set of examples is required to perform reliable motion classification.



## Disciplines

Physical Sciences and Mathematics

## Publication Details

K. Adistambha, C. H. Ritz & I. S. Burnett, "Motion classification using Dynamic Time Warping," in International Workshop on Multimedia Signal Processing, 2008, pp. 622-627.

# Motion Classification Using Dynamic Time Warping

Kevin Adistambha[1], Christian H. Ritz[1], and Ian S. Burnett[2]

[1]*Whisper Labs/TITR, School of Electrical and Telecommunications Engineering*
*University of Wollongong, Australia*
{ka07, critz}@uow.edu.au

[2]*School of School of Electrical & Computer Engineering, RMIT University, Australia*
ian.burnett@rmit.edu.au

***Abstract*—Automatic generation of metadata is an important component of multimedia search-by-content systems as it both avoids the need for manual annotation as well as minimising subjective descriptions and human errors. This paper explores the automatic attachment of basic descriptions (or 'Tags') to human motion held in a motion-capture database on the basis of a Dynamic Time Warping (DTW) approach. The captured motion is held in the Acclaim ASF/AMC format commonly used in game and movie motion capture work and the approach allows for the comparison and classification of motion from different subjects. The work analyses the bone rotations important to a small set of movements and results indicate that only a small set of examples is required to perform reliable motion classification.**

## I. INTRODUCTION

The use of human motion capture in games and movies is increasingly common and has been significantly simplified by the advent of powerful real-time rendering capabilities on even low-end computers. The availability of motion-capture data now makes it possible to find or generate high quality sequences of required motions relatively easily. However, one significant and outstanding issue in the area is the efficient search and accompanying automatic classification of motion capture sequences.

In the literature, there are several suggested solutions for fast and accurate motion matching within a large motion capture database. Notable examples include [1] and [2], where the focus is on providing fast retrieval methods for *visually* similar motions in large motion capture databases. However, the metric of "visually similar motion" is subjective and the general motion searching method requires the use of example motions for matching.

In light of recent advancement in metadata technologies, we suggest that the automated tagging of motion sequences would enable a useful initial step in motion capture database searches. The aim is to reduce the set of possible motion sequences on the basis of a prior grouping of the sequences according to similarity. This allows a user to select an example motion sequence from amongst an appropriate subset of sequences chosen on the basis of metadata. The second stage, full search using the techniques proposed by [1] and [2], can then proceed on the basis of the tagged example. Of course, it may be that users can select motion sequences directly on the basis of the tagging but in general, we expect further searching to be useful.

This paper will thus focus on automatic tagging of an unknown motion sequence using known motion sequences from a database. The goal is not to provide a set of measurements to assist in motion similarity search as performed in [1] and [2], but to automatically provide a set of metadata classifications and descriptions. These might usefully be taken from existing standards such as MPEG-7 [3]. One of the aims of the work is to understand the important features required for general 'tagging' of motion sequences. While full visual similarity matching of a motion sequence would require consideration of all the major bones in a body (i.e. the spine, both arms and both legs), the work investigates whether lower complexity models can be used for reliable and useful tag classification. In particular, this work aims to discover how many and which bones result in good, 'tagged' similarity matching when comparing an unknown motion against a set of reference motions. The paper introduces Dynamic Time Warping (DTW) to overcome time and speed differences in the sequences, and the underlying bone movements.

This paper is organized as follows: Section 2 will provide details on Dynamic Time Warping for motion classification and the cost measure utilized; Section 3 will briefly background the Acclaim ASF/AMC format before describing the database used and the process of classification; Section 4 will provide the experimental results; Section 5 will provide discussions and Section 6 will provide the conclusions and future work.

## II. DYNAMIC TIME WARPING

The purpose of Dynamic Time Warping (DTW) is to compare two different length sequences of values and allow the evaluation of an error measurement on the basis of the match between the two sequences. DTW has been used extensively and successfully in speech recognition [4] to take into account different speeds of utterance of the same phrases and words.

The core of DTW is to find the path through the observations that would lead to the minimum global cost by minimizing the local cost. By continually minimizing the local cost by using dynamic programming, a global minimal error measurement is achieved. In mathematical terms, the global cost matrix $D$ between two sequences is created by the equation:

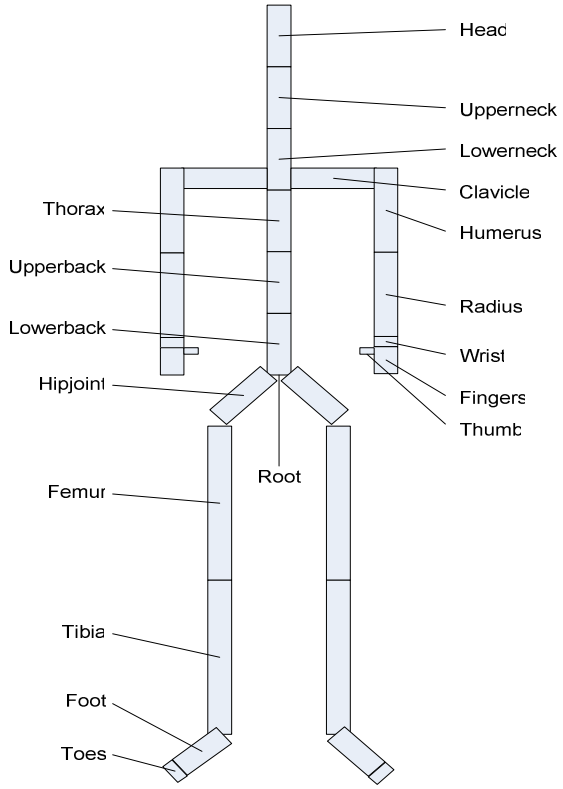


Fig 1. The names and locations of the bones as per the database used in this work.

$$D(i,j) = d(i,j) + \min_{p(i,j)} \{D[p(i,j)] + T[(i,j), p(i,j)]\} \quad (1)$$

Where $d(i,j)$ is the local cost between frame $i$ of the first sequence and frame $j$ of the second sequence, $p(i,j)$ is the set of possible previous costs to $i,j$ and $T$ is the cost function [4]. Each element in matrix $D$ then contains the minimum error between frames $i$ and $j$ based on adding the local cost $d(i,j)$ to the minimum error of frames $(i-1,j)$, $(i,j-1)$, and $(i-1,j-1)$. Hence, the bottom-right value of the matrix $D$ would yield the minimum global error between the two sequences and that is reached by minimizing the local errors between the two sequences.

In this work, DTW is used to provide a distance measurement between two motion sequences of different lengths in a similar approach to that used in speech recognition. In this work, a squared distance measure is used:

$$D = (V_A - V_B)^2 \quad (2)$$

Where $D$ is the squared distance between two values, and $V_A$ and $V_B$ are, for this work, the rotation of bones in degrees. The motion of each bone (see Figure 1) in the skeletal model can then be warped and then matched against the motions recorded for a sequence in the database. In this work, the individual x, y, and z rotations for bones (in some cases not all rotations exist due to bone movement limitations) were warped, however it is also possible to warp complete rotation

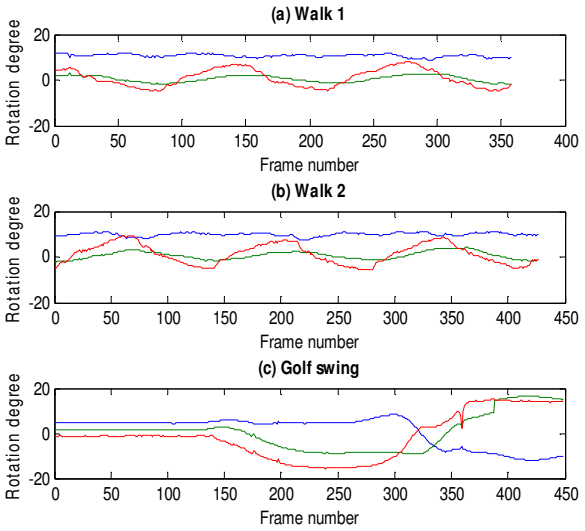


Fig 2. The plots of x, y, and z axis rotations of the lower back bone of two walking motions and a golf swing with different lengths. Each curve represents rotation of the back bones in the skeleton vs. time.

vector tracks, or even complete skeletal bone 'feature' vector tracks.

## III. Motion Matching and Classification

### A. *Description of the data*

#### 1) *Acclaim Motion Capture Format (ASF/AMC)*

The ASF/AMC format is a modern motion-capture format developed by the Acclaim Corporation for games [5]. In the format, the skeleton definition and specification is stored in the ASF (Acclaim Skeleton Format) file. The ASF file defines the skeleton in a hierarchical system and the AMC file details the movements of each of the bones from the ASF file. In the latter, each bone is defined as a child of another bone and the "root" point is defined as the point of origin.

The movement of the bones is described as the rotation of each bone relative to the parent of that bone. This relative coordinate system feature keeps the whole skeleton connected and avoids the problem of bones disconnecting from their parents; this would be a problem with an absolute coordinate system. Another advantage of using relative coordinates in the motion description and separating the bone motion from the skeletal description is that the motion description can then be made independent of bone lengths. This allows the format to generate similar descriptions for motions performed by people of different heights and build; this is vital to maximise the value of motion capture data.

#### 2) *The motion capture database*

The motion capture data was obtained from the freely available Carnegie-Mellon motion capture database, in the Acclaim ASF/AMC format [6]. The data consists of motion capture sequences for various activities such as sports, walking, running, dancing, and nursery rhyme actions. These

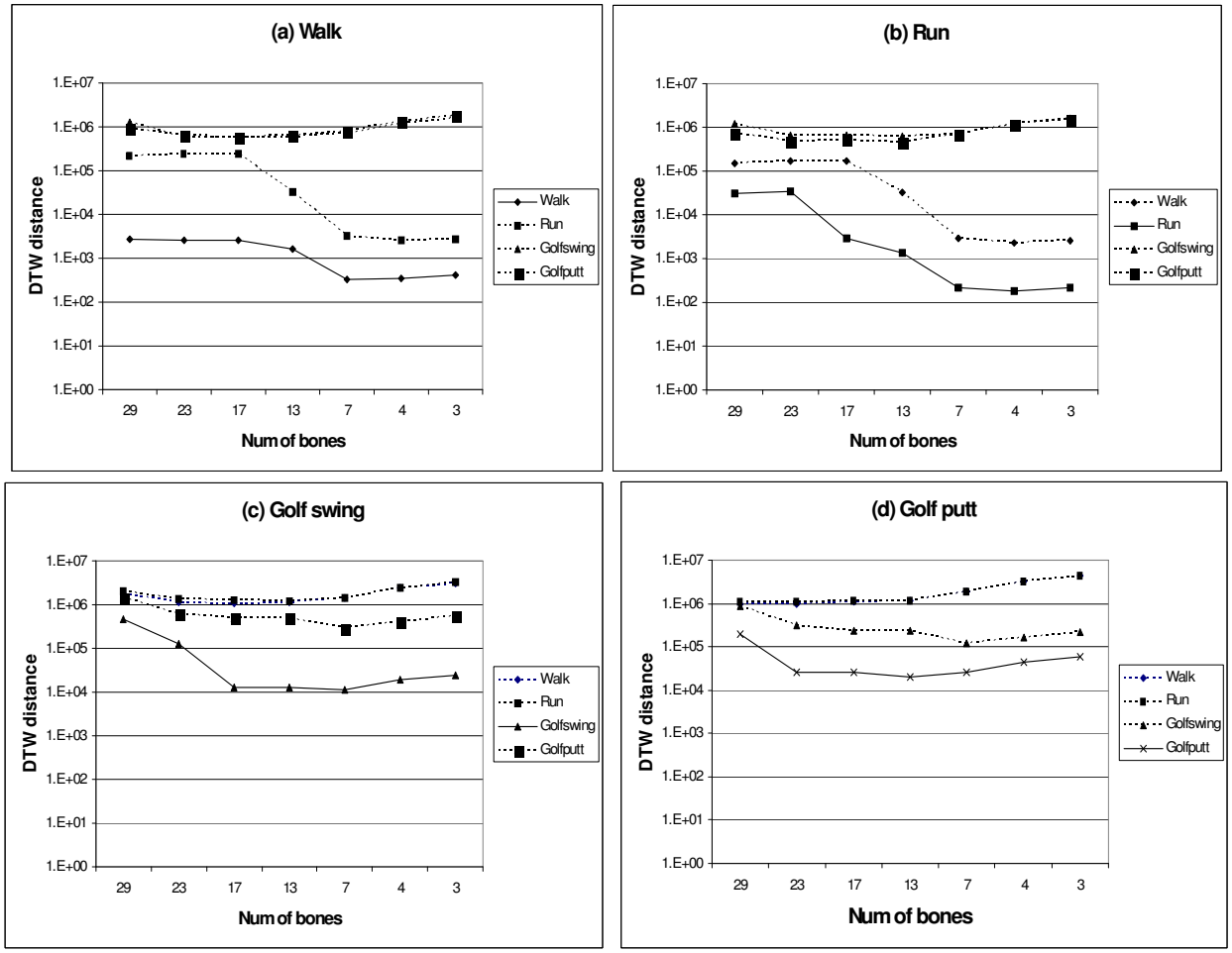


Fig 3. The result plots of DTW classification using 29, 23, 17, 13 7, 4 and 3 bones for (a) walk, (b)run, (c) golfswing and (d) golfputt codebooks. All Y axes are using log scale. Lower DTW distances are better, e.g. the motion to be classified is closer to the reference, and hence they are similar.

are captured at a rate of 120 frames per second. For each frame, the *x, y* and *z* axis rotation for each bone is recorded with respect to the degree of freedom available for that bone, e.g. the upper arm (humerus) has *x, y* and *z* rotations while the forearm (radius) has only *x*-axis rotation from the elbow.

In total, there are 28 bones in the model as shown in Fig. 1, with the 29th bone (root point) representing the rotation and translation of the whole body. This root point serves as the point of origin for the whole skeleton and is situated between the lowerback, left hipjoint and right hipjoint, as illustrated in Fig. 1.

For the purposes of this work, four motion classes performed by two subjects were used: golf swing, golf putt, running and walking. The motions were chosen to provide visually similar motions (walking and running) and then visually dissimilar movements but which utilised a similar set of bones (golf swing and golf putt). Golf swing and golf putt involve a relatively steady head and a rotation of the spine e.g. the thorax, neck and back.

A plot showing an example of the dataset is shown in Fig. 2. In Fig. 2, the x-axis represents the frame number of the motion and the y-axis represents the degree of rotation applied to each bone in the skeleton. Fig. 2 shows the x, y, and z axis rotation of the lowerback bone for two walking motions and a golf swing.

### *B. The classification process*

#### *1) Model creation*

Three randomly selected motion examples from each motion class were used to generate a model of the action by averaging the time-warped movements for all bones. The exact examples chosen was not found to be significant, however the CMU database motion sequences are carefully regulated to be similar. For testing we used a separate testing portion of the dataset: 21 motions for walking, 7 for running, 7 for the golf swing and 2 for the golf putt. The reason for this choice of subset is primarily the limited number of motions available in the CMU database that represents the four motion classes explored in this work. Future work will extend the techniques to broader motion capture sequences which also use a different capture process.

For every bone $k$ in motion $l$ and example $m$, the time warped representation of the example motion $V_{l,m}^k$ is obtained by:

$$V'^k_l = DTW(V_l^k) \quad (3)$$

Each code vector is then formed by averaging the bone motion vectors of the three randomly selected examples. This can be described as:

$$C_l^k = \frac{1}{N}\sum_{m=1}^{N} V'^k_{l,m} \quad (4)$$

Where $N$ is the model set size (we used $N=3$) and the length of each vector $C_l^k$ will vary depending on the motion sequences used in the model creation. The model for a motion is then the collection of vector tracks $C_l^k$ which serve as the representative model and temporal description of the motion.

*2) Classification using DTW*

Each model was then matched against the test set, consisting of the same type of motion and also motions of different types. The lower the distance between the model motion and the time warped test motion sequence, the higher the probability that the motion tested corresponds to that model.

The sets of bones used in the matching process in addition to the original 29 bones described in Section IIIA were:

- 23 bones: root (point of origin; see Fig. 1), lowerback, upperback, thorax, lowerneck, upperneck, head, left and right clavicle, left and right humerus, left and right femur, left and right radius, left and right tibia, left and right wrist, left and right hand, left and right foot.
  17 bones: root, lowerback, upperback, thorax, lowerneck, upperneck, head, left and right clavicle, left and right humerus, left and right femur, left and right radius, left and right tibia.
- 13 bones: root, lowerback, upperback, thorax, lowerneck, upperneck, head, left and right clavicle, left and right humerus, left and right femur.
- 7 bones: root, lowerback, upperback, thorax, lowerneck, upperneck, head.
- 4 bones: root, lowerback, upperback, thorax.
- 3 bones: root, lowerback, upperback.

The selection of bones for each experiment was made subjectively on the basis of observations that bones closer to the spine provide the low-frequency detail of a motion while bones farther from the spine (e.g. the fingers and toes) provide the high-frequency details. The authors are currently analysing capture data across the CMU motion capture dataset to verify these observations. This work thus tests the relevance of bones to the classification process without the use of weighting (i.e. instead, a binary weighting system for bones is employed). For the shift from 29 to 3 bones, bones are progressively removed from the skeleton according to the observed low-pass criterion and, in terms of the model vector creation of Eq. (4):

$$C_l^k = W^k \frac{1}{N}\sum_{m=1}^{N} V'^k_{l,m} \quad (5)$$

where $W \in \{0,1\}$ and $N$ is the model set size.

DTW (as per Eq. (1)) is then used with the error criteria in Eq. (2) to determine the similarity between an input motion and the model by time-warping and distance measurement between the model vector $C_l^k$ and the corresponding bone movements in the input motion. The distances across all the vectors are then averaged as the global distance of the input motion against the model according to:

$$S_{DTW} = \frac{1}{K}\sum_{k=1}^{K}\left(V_c^k - V_a^k\right)^2 \quad (6)$$

Where $S$ is the global DTW distance, $K$ is the total number of bones in the model, $V_c^k$ is the $k^{th}$ bone of the codebook and $V_a^k$ is the $k^{th}$ bone of the input motion to be classified.

## IV. RESULTS

From the experiment, it is evident that using only three motions to serve as the model motion allow the system to reliably differentiate between motions for the small, available dataset.

From Figure 3, it can be seen that the distances of an input motion to a reference motion are smallest if the motion in question is of the same class. This result was uniform across all bone set sizes from 29 to 3. However, best average distance results were achieved using 17 bones, 13 bones and 7 bones (as shown in Fig. 3). Therefore, in order to minimize the amount of data required to perform motion classification, only 7 bones consisting of the root, lowerback, upperback, thorax, lowerneck, upperneck and head are required. A sample plot of the degree of rotations of the 7 bones is shown in Fig. 4.

The results shown in Figure 3 indicate that using all 29 bones for matching and classification purposes is not necessary, while using too few bones provides some classification but not maximum performance. The best classification performance is reached if the number of bones used is 17, 13 or 7 for all motion types tested. This result agrees with the observation explained in Section IIIB-2: that a reliable classification could be performed with a certain mix of low-frequency and high-frequency detail.

In the case of the motions explored in this work, the 7 bones of the spine plus the head, e.g. head, upperneck, lowerneck, thorax, upperback and lowerback, provide the lowest distances (compared to all the other bones) to the reference motions. This is shown in the results of Fig. 5. The latter indicates the contribution of the various bones to the distance measures for the test motion sequences. The histograms are ordered such that the most important bones are first (to the left) in the plots.

In the case of walking motion, using 7 bones discounts the variability of starting with the left or the right foot. By concentrating on the spine (which would move in similar way without regard of the feet movement), a walking motion could be identified properly.

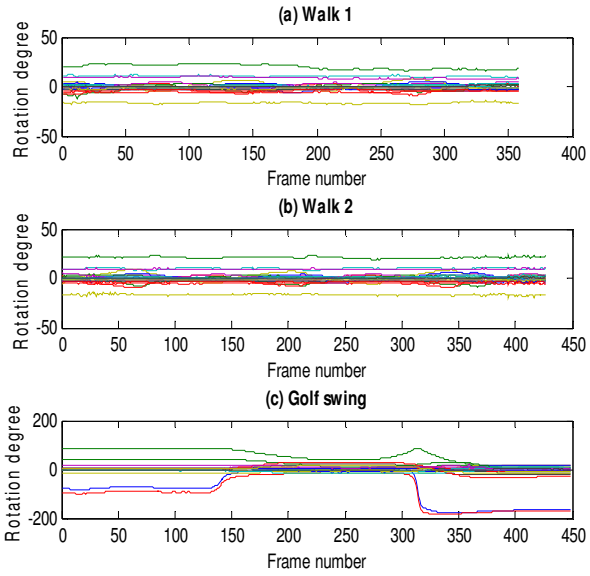


Fig 4. Plots of the degree of rotations of the 7 bones used for classification. The motions in this plot are the same motions shown in Fig. 2.

## V. Discussion

The 7 bones chosen in this work performed well for motion classification and this is confirmed from work in biomechanics (such as in [7]) which shows that in e.g. the walking motion, the spine provides a stable platform for vision. Although there is a slight bob in the head that follows a sinusoidal pattern, it is relatively stable compared to running. In running, there are moments where both feet are off the ground and it is quite a distinct movement class compared to walking [8]. Referring to Fig. 1, a person performing a walking motion would attempt to keep the lowerback, upperback, thorax, lowerneck and upperneck relatively steady. This fact is also evident in Fig. 5(a), where the head in test walking motions yields the lowest distance to the reference walking motion.

The same 7 bones are also useful in differentiating between a golf swing and golf putt, which although basically similar motions, differ in the area of spine rotation (lowerback, upperback, and thorax in Fig. 1). In a golf swing, the spine rotates rapidly while the head (lowerneck, upperneck and head in Fig. 1) remains more or less still, while in golf putting the head is still and the spine rotates in a smaller degree [9].

For the motions explored in this work, it is reasonable to conclude that head and spine movement alone provides a clear distinction between the four motions. However, the inclusion of the root point provides the translation and rotation of the whole body to further differentiate e.g. walking and climbing a stair, where there are considerably more vertical movement in climbing stairs compared to walking. Although visually similar, walking and climbing a stair are very different in biomechanical aspect [8].

## VI. Conclusions and Future Work

Motion classification can be effectively performed using only 7 bones which constitute primarily the head and the spine. Contrary to the application of searching for a visually similar motion by means of query-by-example, the motion classification performed in this work does not need high motion detail in order to perform reliably. The results shown in Fig. 5 indicate that the relative distances of each bone to the reference are in accordance with observations based on the analysis of the biomechanics.

DTW can be used effectively for motion classification to generate and match the temporal description tracks of each bone in a motion; classification can then be performed on the basis of distance to motion classes. A more efficient DTW algorithm could be utilized to speed up the classification process with a large codebook size. Since DTW performs an exhaustive search of the least cost, a possible improvement is to put constraints on the search space of the DTW matrix - by which a faster DTW performance can be realized. The risk of such an approach is that the path found by the DTW algorithm may not be the least-cost path and could potentially result in misclassification.

The use of more classes of motions, as used in the field of biomechanics, would also provide more insight into the validity of the 7 bones classifier across a wider range of motions. Also, since this work only explores the minimum set of bones required to reliably classify the tested motions, alternative weighting systems (i.e. non-binary) for the bones could be implemented to discover the proper weighting scheme by taking into account the movements of all the bones instead of a subset of them. Further, this work should also be extended to investigate alternative classification methods and a broader range of motion capture content. The authors are currently capturing motion on-site at the University of Wollongong using a newly acquired motion capture suit.

## Acknowledgment

The data used in this project was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.

## References

[1] Y. Lin, "Efficient motion search in large motion capture databases", *Lecture notes in computer science vol. 4291 pp. 151-160, Springer-Verlag*, 2006.
[2] M. Muller, T. Roder, M. Clausen, "Efficient content-based retrieval of motion capture data", *Proceedings of ACM SIGGRAPH*, 2005.
[3] ISO/IEC, "MPEG-7 Part 5: multimedia description schemes", ISO/IEC 15938-5:2001, 2001.
[4] B. Gold, N. Morgan, *Speech and Audio Signal Processing*, John Wiley & Sons, 2000.
[5] J. Lander, "Working with motion capture file formats", *Game Developer magazine,* January 1998.
[6] "Carnegie Mellon University – CMU Graphics Lab – motion capture library", http://mocap.cs.cmu.edu/
[7] M. Whittle, *Gait Analysis: an introduction*, Butterworth Heinemann Elsevier, 2007.
[8] J. Low, A. Reed, *Basic Biomechanics Explained*, Butterworth Heinemann Elsevier, 1996.
[9] G. K. Hung, J. M. Pallis (eds.), *Biomedical Engineering Principles in Sports*, Kluwer Academic, 2004.

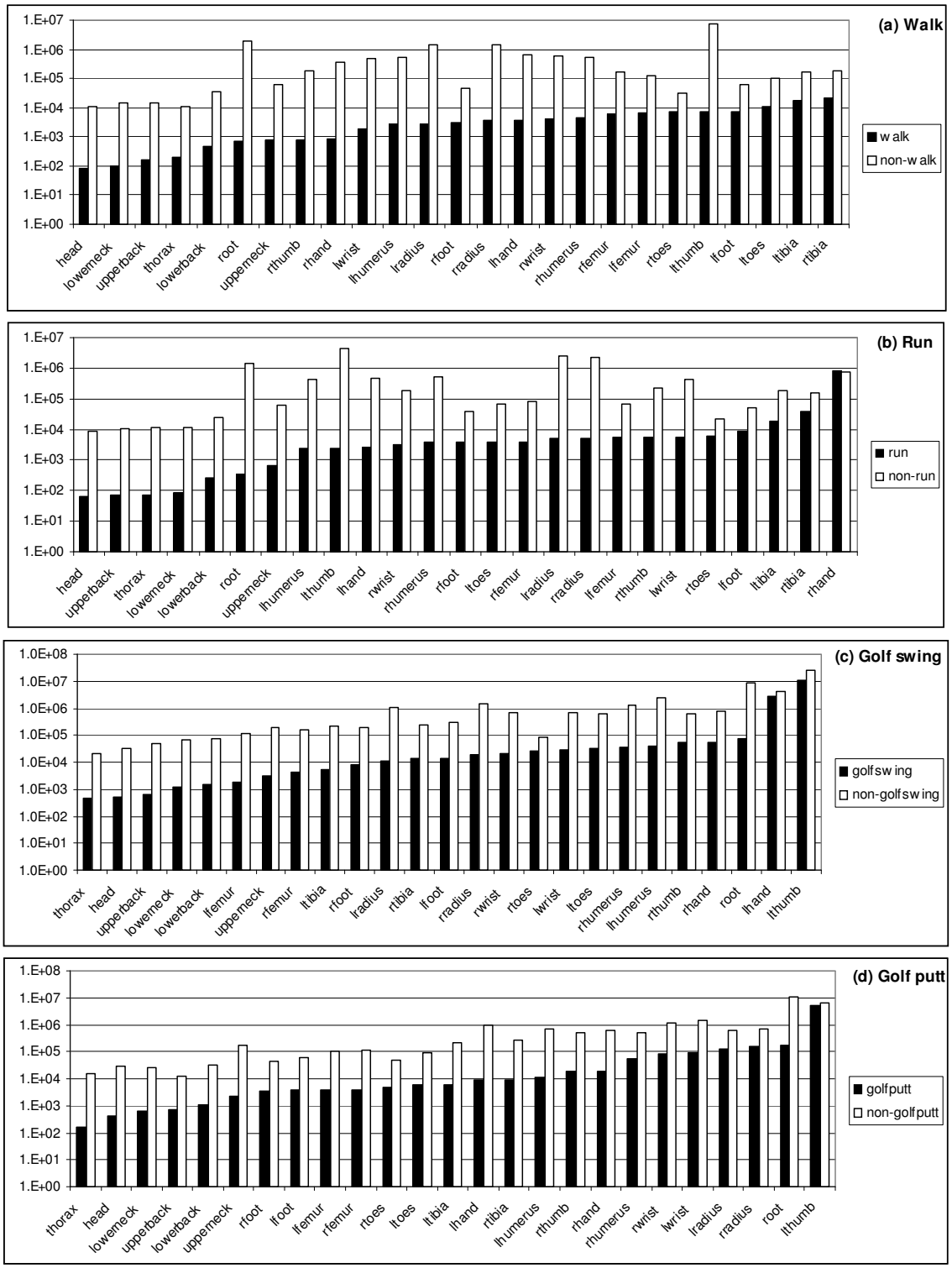


Fig 5. Plots of the average distances of each bone track using all 29 bones in the model compared to their respective reference motions. The plot shows that across all motions tested, lowest distances to the reference motion were generally achieved by bones of the head and the spine (thorax, lowerneck, upperneck, lowerback, upperback). All Y axes are using log scale. Black bars indicate tests against the respective reference motion sequences while white bars indicate results against other motion sequences.