📄 **HW1.md** 5.6 KB

Basics

What is the dataset name?

Head Impact Kinematics and Brain Deformation Dataset

How was it collected? Describe instruments, experiments, subjects, etc.?

The data was collected using the Stanford Instrument Mouthguard which records multiple measures such as angular velocity. "The kinematics are generally measured by wearable sensors (but for this case, simulated by anthropomorphic dummy test FEM). The MPS is simulated by KTH model, one of the validated brain model to calculate the biomechanical responses of the brain (4124 brain elements in total)."

Where did it come from?

The data was recorded by researchers conducting experiments that simulate Traumatic Brain Impacts (TBI) in CamLab at Stanford.

If public, what is the URL to the dataset?

Not public

Is there a paper associated with this dataset? Provide the link if so.

Yes. https://arxiv.org/abs/2102.05020

Provide a short description of the dataset.

Data types: 1) Kinematics of head movement 2) brain deformation calculated by finite element modeling (FEM).
The two files of kinematics involve the kinematics measured by the mouthguard, including: linear acceleration at the brain center of gravity, angular velocity, angular acceleration and the time of the sampling points.

File Format

What file format is the data distributed in? Provide a few sentence description of this file format.

The Kinematics files are already processed into .mat files and are ready to work with in MatLab. There are 2 zipped files for Maximum Principle Strain that contain 2130 CSV files, one for each brain impact simulation.

Provide an overview of the metadata associated with this dataset.

The root folder is Camarillo, which is the name of the PI at CamLab. There are 2 folders within, one for Kinematics and one for Maximum Principle Strain, and a Data Description.txt which has some information on how the data is stored/what it represents.

Structure & Encoding

What is the structure of the dataset? Describe the dimensions, organization, hierarchies, etc. Be reasonably detailed so that a reader could reconstruct the structure of the dataset. If there are multiple data structures in the dataset, describe each one.

The data in the Maximum Principle Strain zip folders is represented as 2130 CSV files, one for each brain impact simulation. In each of these CSV files, the far left column is time incremented by +.001sec starting from 0sec to .069sec, so there is a time for each row of the 70 rows. Each 2nd to 4125th column represents 1 of the 4124 brain elements. At each time for each brain element, there is a brain deformation value (which is a measurement of how enlarged, strunken, or displaced the brain element is relative to its natural state).

What is the data encoding model of the dataset? Be descriptive enough so that the explanation goes down to level of the three primitive data classes (integers, characters, and floating-point numbers).

The data is encoded using a model of 64 bit double-precision floating points. Each brain element is uniquely labeled using 'H' character followed by 3 integers. Each brain deformation value is a floating-point number that goes up to the 1E-20 decimal place. Each time label is a floating-point number that goes up to 1E-3 decimal place.

Size

What is the size of the dataset as distributed?

4GB

Is the dataset distributed with compression? If so, what compression is used and what is the uncompressed size of the dataset.

The dataset is distributed with the Maximum Principle Stain files compressed into zip files. The uncompressed size of the data set is 16GB.

Approximate the uncompressed size of the dataset using your understanding of the dataset structure and the encoding model.

Based on the structure of the dataset, I would guess that the uncompressed size of the dataset would be at least 16GB because the files that I unzipped contained CVS files of lots of float point numbers which decompress to about 4 times their compressed size.

The math here must be at the level of the three primitive data classes.

Provide a web screenshot or console text snippet (use du -sch for calculating size of a directory) of the size of the distributed dataset

If compression is used, also show the uncompressed dataset size, as that is what is being estimated

Workup

What type of analysis might you hope to perform with this data?

I want to do some simple statistical analysis of the brain deformation across all 2130 simulated brain impacts. I am most interested in calculating the "time peak" of the impact when brain deformation occurs which will give me more insight as to the level of brain injury in a certain time span. This could be used to measure the effectiveness of TBI preventative equipement and allow manufacturer/designers/engineers to develop better protective gear.

Describe potential summary statistics or measures that may be appropriate to calculate

I could find the mean, median, and mode of the brain deformation across the 2130 simulations' time peaks. Along with this, I could find the area under the "time peak" curve to measure the total amount of deformation across the time span, and with this I can compare sharper to wider time peaks.

What types of visualizations might you create with this data?

I could create a graph for each simulation where the x axis is time and the y axis is brain deformation. Then I could compare the time peaks, when impact and deformation occurs, between the different graphs. I can also then plot the average, median, mode, graphs of all 2130 simulations on top of each other.