Description of the MSRC-12 Gesture Dataset

Contributors to this Document

Pushmeet Kohli, Sebastian Nowozin and Helena Mentis
Microsoft Research Cambridge

Simon Fothergill
University of Cambridge

Introduction

The Microsoft Research Cambridge-12 (MSRC-12) gesture dataset consists of sequences of human skeletal body part movements (represented as body part locations) and the associated meaning that needs to be recognized by the system. The dataset was collected at MSR Cambridge as part of a research project which is described in:

Simon Fothergill, Helena Mentis, Pushmeet Kohli, Sebastian Nowozin
"Instructing People for Training Gestural Interactive Systems"
ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2012)

If you use this dataset, you should cite the aforementioned paper in any resulting publication. The dataset is accompanied with a set of MATLAB files (written by Sebastian Nowozin) which can be used to verify and visualize the sequences contained in the dataset.

Description of the Dataset

The dataset comprises of 594 sequences, 719359 frames (approx. 6h40m) collected from 30 people performing 12 gestures. In total, there are 6244 gesture instances.

The motion files contain tracks of 20 joints estimated using the Kinect Pose Estimation pipeline. The body pose is captured at a sample rate of 30Hz with ~2cm accuracy in joint positions.

Participants

The participants were recruited at MSR Cambridge and the Computer Laboratory at University of Cambridge. Although some of the participants were familiar with the domain of machine learning and computer vision, none of the participants were privy to the workings of the machine learning algorithm of the study we were conducting. The demographics of the participants were 60% male, 93% right-handed, 5’0”-6’6” tall with an average of 5’8”, and 22-65 with an average of 31 years of age.

Gesture Types

The gestures can be categorized into two abstract categories: **Iconic gestures -** those that imbue a correspondence between the gesture and the reference, and **Metaphoric gestures** - those that represent an abstract concept. A list of all gestures along with the number of instances is given below:

A. **Iconic Gestures**

   a. **Crouch or hide (G2) [500]**
   b. **Shoot a pistol (G6) [511]**
   c. **Throw an object (G8) [515]**
   d. **Change weapon (G10) [498]**
   e. **Kick (G12) [502]**
   f. **Put on night vision goggles (G4) [508]**

B. **Metaphoric Gestures**

   a. **Start Music/Raise Volume (of music) (G1) [508]**
   b. **Navigate to next menu (G3) [522]**
   c. **Wind up the music (G5) [649]**
   d. **Take a Bow to end music session (G7) [507]**
   e. **Protest the music (G9) [508]**
   f. **Move up the tempo of the song (G11) [516]**

Instruction method

To study how the instruction modality affected the movements of the subjects, we collected data by giving different types of instructions to our participants.

We chose to provide participants with three familiar, easy to prepare instruction modalities and their combinations that did not require the participant to have any sophisticated knowledge. The three were (1) descriptive text breaking down the performance kinematics, (2) an ordered series of static images of a person performing the gesture with arrows annotating as appropriate, and (3) video (dynamic images) of a person performing the gesture. We wanted mediums to be transparent so they fulfil their primary function of conveying the kinematics. More concretely, for each gesture we roughly have data using: Text (10 people), Images (10 people), Video (10 people), Video with text (10 people), Images with text (10 people).

File Naming Convention

There are 594 pose sequences (.csv files) along with 594 associated “tagstream” files which contain information about when a particular gesture from the list above should be detected. The format of the file names is: P[#A]_[#B]_[#C]_P[#D]. [csv][tagstream]

[#A] and [#B] encode the instruction that was given to the subject (see table below and the paper). The presence of the Alphabet “A” after #B indicates that movements generated when two instructions were given (eg. text + images, or text + video).

[#C] denotes the identifier of the gesture being performed in the file. This is redundant information since this information is also present in the tag stream file.

[#D] denotes the identifier for the human subject whose movements are captured in the file.

Gestures	C1=Video	C2=Images	C3=Written description	C4=Video + written description	C5=Images + written description
G1=Start system	P1_1+P1_2	P2_1+P2_2	P3_1+P3_2	P1_1+P3_1	P2_1+P3_2
G2=Duck	P3_1+P3_2	P1_1+P1_2	P2_1+P2_2	P3_1+P2_1	P1_1+P2_2
G3=Push right	P2_1+P2_2	P3_1+P3_2	P1_1+P1_2	P2_1+P1_1	P3_1+P1_2
G4=Goggles	P1_1+P1_2	P2_1+P2_2	P3_1+P3_2	P1_2+P3_2	P2_2+P3_1
G5=Wind it up	P3_1+P3_2	P1_1+P1_2	P2_1+P2_2	P3_2+P2_2	P1_2+P2_1
G6=Shoot	P2_1+P2_2	P3_1+P3_2	P1_1+P1_2	P2_2+P1_2	P3_2+P1_1
G7=Bow	P1_1+P1_2	P2_1+P2_2	P3_1+P3_2	P1_1+P3_1	P2_1+P3_2
G8=Throw	P3_1+P3_2	P1_1+P1_2	P2_1+P2_2	P3_1+P2_1	P1_1+P2_2
G9=Had enough	P2_1+P2_2	P3_1+P3_2	P1_1+P1_2	P2_1+P1_1	P3_1+P1_2
G10=Change weapon	P1_1+P1_2	P2_1+P2_2	P3_1+P3_2	P1_2+P3_2	P2_2+P3_1
G11=Beat both	P3_1+P3_2	P1_1+P1_2	P2_1+P2_2	P3_2+P2_2	P1_2+P2_1
G12=Kick	P2_1+P2_2	P3_1+P3_2	P1_1+P1_2	P2_2+P1_2	P3_2+P1_1

MATLAB files

The dataset comes with a set of Matlab scripts (written by Sebastian Nowozin) which allow you to verify the content and to visualize the data. The basic description of the files is as follows:

1. LOAD_FILE -- Load gesture recognition sequence
2. SKEL_VIS -- Visualize a skeleton in 3D coordinates.
3. Demo 1: Visualize a sequence.
4. Verify integrity of the dataset

Acknowledgements

We would like to thank the staff, students and interns of Microsoft Research, Cambridge and the University of Cambridge for participating in our user study. We would also like to thank Olivia Nicell for help with data gathering and tabulation.