EGGNOG: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels

Isaac Wang[1], Mohtadi Ben Fraj[2], Pradyumna Narayana[2], Dhruva Patil[2], Gururaj Mulay[2],
Rahul Bangar[2], J. Ross Beveridge[2], Bruce A. Draper[2], and Jaime Ruiz[1]
[1] Department of Computer Science, University of Florida, Gainesville, FL, USA
[2] Department of Computer Science, Colorado State University, Fort Collins, CO, USA

***Abstract*—People communicate through words and gestures, but current voice-based computer interfaces such as Siri exploit only words. This is a shame: human-computer interfaces would be natural if they incorporated gestures as well as words. To support this goal, we present a new dataset of naturally occurring gestures made by people working collaboratively on blocks world tasks. The dataset, called EGGNOG, contains over 8 hours of RGB video, depth video, and Kinect v2 body position data of 40 subjects. The data has been semi-automatically segmented into 24,503 movements, each of which has been labeled according to (1) its physical motion and (2) the intent of the participant. We believe this dataset will stimulate research into natural and gestural human-computer interfaces.**

I. INTRODUCTION

Modern voice-based interfaces such as Siri [3], Alexa[1], and Cortana[2] are popular because they make interacting with technology natural and intuitive. They let us talk to our applications much as we would another person. At the same time, users of these interfaces quickly learn that the communication channel is limited. Voice-based interfaces interpret our words, but not our gestures or expressions. If communicating with computers is to become truly natural, interfaces will need to recognize gestures and expressions as well as words.

This paper encourages the development of systems for recognizing communicative gestures by presenting a new data set of continuous, task-based conversations with RGB video, depth video, and frame-by-frame body positions. The data set also includes hand-generated ground truth gesture labels. As discussed in the Related Work section below, this is by no means the first gesture data set, and gesture recognition is already an active field. Shared data sets, however, strongly influence research directions, and existing data sets do not target the types of rapid and subtle gestures that occur spontaneously between people.

More precisely, this paper introduces a gesture data set collected while two people perform a shared, physical task. The participants are in separate rooms, connected by video and audio links. Both participants are standing in front of tables, as shown in Figure 1. One participant (the *actor*) is given a set of blocks; the other (the *signaler*) is given a picture of blocks in a specific arrangement. The task is for the signaler to get the actor to recreate the pattern of blocks. On some trials there is no audio channel, so gestures and expressions are the only modes of communication. Other trials include an audio channel, so gestures supplement rather than replace spoken words. The data set contains 360 trials with 40 participants, for a total of eight hours of multi-modal recordings of gestures. Section III describes the data and how it was collected in more detail.

The data set, called EGGNOG[3], captures aspects of communication through gestures not featured in previous data sets. Among the most important are:

1) **Naturally Occurring Gestures.** The participants were given no instructions about gesturing. Consequently, the gestures are those that occur naturally in the course of communication. Although some gestures are idiosyncratic, many gestures are used spontaneously by different subjects. These gestures are rapid and sometimes subtle, and different from the stylized gestures found in sign languages or video games.
2) **Continuous Data.** Gestures occur in the context of tasks that average around one minute (with sound) or one and a half minutes (without sound) to complete. The average gesture takes about one second, and of course not all movements are meaningful gestures. In fact, most motions have no discernable semantics. As a result, gestures recognition requires detection as well as classification.
3) **High-Quality Multi-Modal Data.** The advent of inexpensive depth sensors has broadened the scope of gesture recognition research. Gestures can be recognized in RGB video data, depth video data, or in 3D body position data (sometimes called *skeleton* data, and similar to motion capture data). Systems like the Microsoft Kinect [28] and Asus Xtion[4] capture real-time body position data, while sensors like the Intel RealSense[5] or LeapMotion[6] track hand position data. EGGNOG includes registered video, depth and body position data from a Microsoft Kinect v2 sensor. The Kinect v2 is significantly more accurate than the original Kinect, and

This work was supported by DARPA and ARO through grant W911NF-15-1-0459.

[1] https://developer.amazon.com/public/solutions/alexa/alexa-voice-service/rest-overview

[2] https://en.wikipedia.org/wiki/Cortana_(software)

[3] EGGNOG: *E*licited *G*iant *G*allery of *N*aturally *O*ccurring *G*estures

[4] https://www.asus.com/3D-Sensor/Xtion_PRO/

[5] http://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html

[6] https://en.wikipedia.org/wiki/Leap_Motion

978-1-5090-4023-0/17 $31.00 © 2017 IEEE
DOI 10.1109/FG.2017.145

we believe this added accuracy is necessary to detect naturally occurring gestures.

Unlike many other datasets, EGGNOG is not tied to a single challenge problem. It provides over 8 hours of labeled data that supports many lines of research. EGGNOG supports gesture recognition research in RGB, depth, and/or body pose data, but it also supports research in gestural semantics, dialogue analysis and other related topics. It includes not only the raw data but labels and support tools as well. Every video is semi-automatically segmented, and the resulting segments are labeled according to (1) the physical motions and poses of the head, torso, arms and hands and (2) the intent of the signaler. The tools used to segment and label videos are provided, as are protocols and baseline results for one interpretation task (hand pose and orientation recognition). More information on the labeling scheme can be found in Section IV.

II. Related Work

In a 2014 survey, Ruffieux et al. [25] listed 15 publicly available gesture datasets. More recently, in a survey of RGBD datasets, Firman [8] added 5 more gesture datasets, and more recently still Ponce-López et al. [22] introduce the ChaLearn 2016 dataset. Of these 21 datasets, 7 concentrate exclusively on hands, with 4 showing hands against blank backgrounds [14], [23], [24], [18] while 3 focus on hands in more challenging settings [16], [5], [21]. 9 more datasets reflect full body gestures from predefined task-specific lexicons: 5 sign language lexicons [2], [6], [7], [13], [19], 2 video game lexicons [9], [4] and 2 military gesture lexicons [17], [27]. The ChaLearn 2013 dataset shows full body gestures from multiple lexicons, including but not limited to diving gestures and Italian social gestures [7]. The earlier ChaLearn 2011 [11] dataset is very large (about 50,000 gesture instances) and combines 100 different hand and arm gestures from a collection of lexicons.

Only two datasets allow participants to invent their own gestures. In the 3DIG dataset [26], subjects were shown 20 3D objects and asked to create their own object-specific gestures. The newer ChaLearn 2016 [22] is a behavior rather than gesture dataset. It contains 3,000 YouTube videos of people speaking to the camera, and the goal is to rate the speakers according to five personality traits (extraversion, agreeableness,conscientiousness, neuroticism and openness). The avoids predefining a lexicon of actions, but it means the videos are not labeled at the level of gestures, nor is their any need to recognize or assign meanings to specific gestures.

Since most of the datasets listed above contain only positive samples of gestures, Freeman et al. [10] provide a dataset of naturally occuring non-gestures for training systems that must distinguish between gestures and background motions.

With 21 gesture datasets available, why introduce another? Because datasets support and encourage research in specific tasks. Our goal is to stimulate research into gesture-based human computer interaction. People communicate not only through speech, but also through gestures. This is particularly true when people work cooperatively on physical tasks. We therefore present the EGGNOG dataset, which contains natural gestures used by people working in blocks world. EGGNOG is different from the datasets above in that it contains gestures elicited from human subjects (see Section III below). These gestures are often quite subtle and are qualitatively different from sign language or video game gestures, for example. Although the ChaLearn 2016 dataset records natural behavior, there is no task, there is no two-way communication, and the videos are not labeled at the level of gestures or actions. It is intended to support personality analysis, not communication through gestures.

EGGNOG is unique because it contains continuous data of people while they work cooperatively. People move a lot during these tasks. Sometimes they are meaningfully gesturing; other times they are just waving their arms while they talk. Communication requires more than a forced choice classification of pre-segmented samples or a general personality analysis. Specific gestures have to be detected against a background of other motions. Most of the datasets above only contain pre-segmented gestures. Freeman et al. [10] provide a data set of background motions as a way of enhancing training data for non-forced-choice scenarios, but EGGNOG is the only large, continuous gesture dataset. It should be noted, however, that there are continuous datasets of human actions, as opposed to gestures, for example [15].

Finally, with 24,503 labeled gesture instances, EGGNOG is large enough to support research that exploits learning-based techniques. This is not a unique feature – ChaLearn 2011 [11] is even larger, for example – but it is an important feature many other data sets lack.

III. Dataset Recording

A. Study Setup

To collect the EGGNOG dataset, we placed pairs of participants in separate rooms, each in front of a table. TV screens and Microsoft Kinect v2 sensors were set up facing each participant. Participants could see (and sometimes hear) each other as if they were at opposite ends of the table. We also used the Kinects for recording the studies. One participant, the *signaler*, had the task of directing the other participant, the *actor*, to construct a structure of wooden blocks, given an image showing the block layout. The signaler was not allowed to show the pattern to the actor. The actor was given twelve wooden blocks (4 inches on a side) to use in constructing the block layout. Figure 1 depicts the setup, while Figure 2 shows a snapshot of Kinect video of a signaler and actor. Since actors mostly move blocks, we describe the data from the signalers in this paper (although the actor data is also available).

For each session, we ran up to ten trials per participant (sometimes less, depending on time constraints), after which the participants would switch roles, and repeat for up to another ten trials. Each trial used a different block layout, and layouts were not repeated within sessions. Ten sessions were completed where neither participant could speak (sound was muted for the entire session), and another ten sessions were completed where both verbal and non-verbal communication

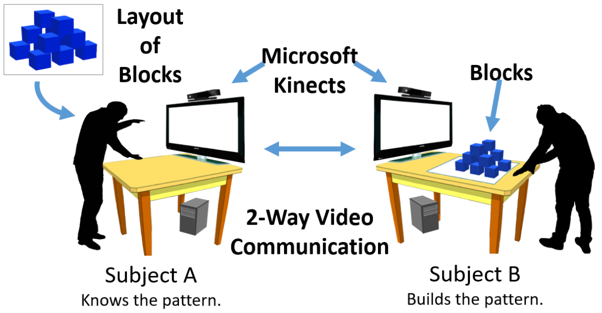


Fig. 1: The setup used to record the dataset. The EGGNOG data set focuses on data from the signalers (Subject A's above).

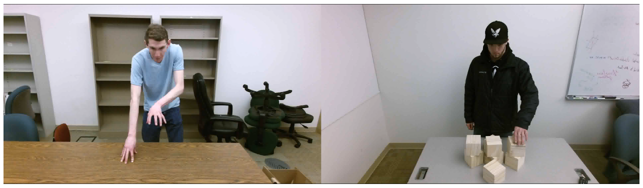


Fig. 2: A snapshot from videos recorded simultaneously from the signaler (left) and actor (right).

were allowed (i.e. participants could both see and hear each other).

In each trial, we presented the signaler with a picture of a block layout (mounted the the right of the signaler's screen, so the actor couldn't see it). The signaler then began to communicate with the actor and describe how to construct the layout. The actor would place and move blocks on the table until the layout was completed, at which point researchers would signal to both participants that the trial was complete. Each trial was recorded from when the layout was presented to when the layout was successfully replicated. In order to observe natural gestures that occur spontaneously, participants were not allowed to speak to each other or discuss strategies beforehand.

B. Participants

We recruited forty participants via email and visits to local computer science classes. Thirteen participants were female. Participant ages ranged from 19 to 64 (mean = 24.68, SD = 9.14). Only four participants were left-handed, and twenty-one participants had prior experience with gesture interaction systems such as the Microsoft Kinect or Nintendo Wii. Out of the twenty sessions (one pair of participants per session), ten pairs were acquainted. Participants each received a $10 Amazon gift card as compensation.

C. Data Collection

A total of 360 trials were recorded, with an average of roughly 18 trials completed per session. 3 more trials were discarded due to recording malfunctions. Collectively, the 360 trials contain 8:07:02 (h:mm:ss) of data. Trial lengths vary between 00:05.5 seconds and 13:36 (mm:ss), with an average time of 1:20 (SD = 1:25). We record RGB, depth, and body position data for all trials; sound is included when applicable. The RGB videos are recorded at a resolution of $1,920 \times 1,080$ pixels at 30 fps. The depth videos are recorded at 512×424 pixels, also at 30 fps. Figure 3 shows one frame from a video with its corresponding depth image and skeleton pose data.

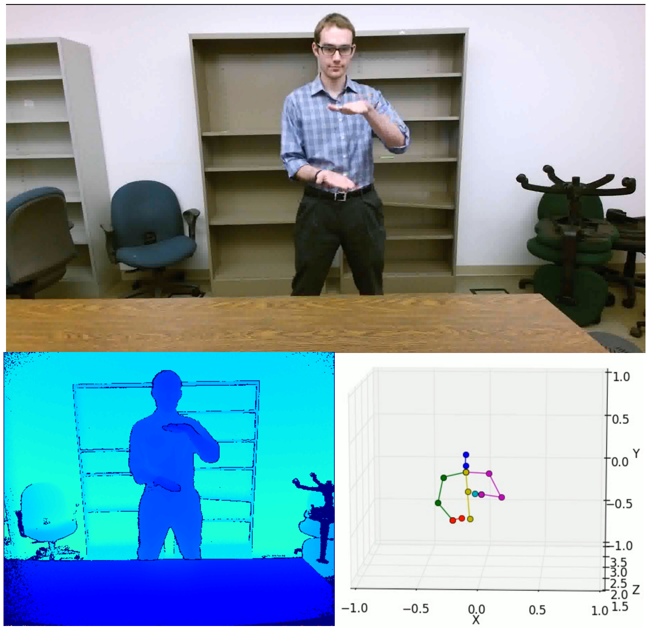


Fig. 3: A snapshot from an RGB video (top), with the matching frames from the depth image (bottom left) and skeleton data (bottom right).

The skeleton data we provide includes 3D positions for all 25 joints tracked by the Microsoft Kinect v2. Note, however, that the subjects are standing behind tables. Nothing below the waist is visible, so joint positions corresponding to the hips, legs and feet are unreliable and should be discarded as in Figure 3, resulting in 17 meaningfully tracked joints.

IV. DATASET LABELING

A. Labeling Language

To determine ground truth, as well as to aid researchers who may be interested in classification but not segmentation or detection, EGGNOG trials are semi-automatically segmented into motions. The segmentation algorithm divides videos at curvature maxima of the curve traced out by the subject's joint positions over time, using the algorithm described in [1]. Once a video is segmented, human annotators are given the option of moving, inserting, or deleting segment boundaries. They then label every segment twice: one label describes the physical gesture in terms of moving body parts, while the other label conveys the intent behind the gesture. This section presents the description languages used for these labels, while the next section provides label statistics over the dataset.

An important distinction between EGGNOG and other datasets is that the set of gestures was not mandated beforehand. Participants were asked to complete a task, and used whatever gestures came naturally to them. Consequently, the gesture labels and label format was created after the data was collected, in order to describe the observed gestures.

1) Physical Gesture Labels: Gesture labels follow a general format of *body part: description*, for example *head: nod* or *right hand: fist*. Body part descriptions may be combined into one label separated by semicolons, such as *head: nod; right hand: point, down.* Each description contains one or more terms describing aspects of the gesture, namely motion, relation, and/or pose. Motions, relations and poses in turn have orientations. In the description below, aspects are grouped with square brackets (and are considered optional), with individual terms in angle brackets:

```
<body part>:
    [<motion>, <orientation>,]
    [<relation>, <orientation>,]
    [<pose>, <orientation>];
```

Each label begins with one of the following body parts:

1) Body: Used to describe large motions of the person's upper body, such as stepping forward or back.
2) Head: Used to describe head motions (e.g. shake).
3) Arms: Either *arms* (plural), *LA* (left arm) or *RA* (right arm). Note that minor, unintentional arm motions are ignored.
4) Hands: Either *hands*, *LH* (left hand), or *RH* (right hand). Used to describe hand orientations, poses and motions. The palm is denoted as the 'front' of the hand, while the other side is the back.

Since the participants are standing behind a table, only the head, torso, and arms are visible. Legs and feet are therefore omitted. Shoulder shrugs are the one observed gesture not expressible in terms of the body parts above, and may be added as a special label in the future.

Each body part description contains a motion, relation, and/or pose. These are described as:

1) Motion: Terms for the complex motions of body parts. *Beckon*, for example, is the motion of curling arms and/or hands toward the body in a beckoning motion. In the case of plural body parts (i.e. arms or hands), the motion may be relative, as in *together*. The complete set of motion terms is: *apart, beckon, enough, move, nod, pivot, rotate, rub, shake, still, stack, surround, tap* and *together*. Some terms apply to any body part (e.g. *shake*), while others are limited to specific body parts (only hands *tap*, for example).
2) Pose: Describes detailed hand positions. There are hand poses for every number from *one* through *five*. In addition, there are other hand poses: *claw, closed, fist, inch, L, open, point,* and *thumbs*. Figure 4 shows examples of each of the non-numeric poses. Note that poses also have orientations, for example to distinguish between *thumbs, up* and *thumbs, down*.

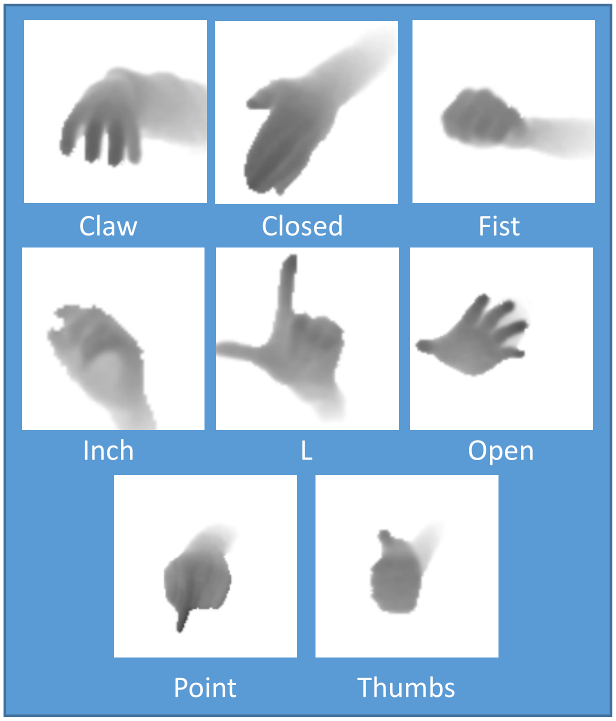


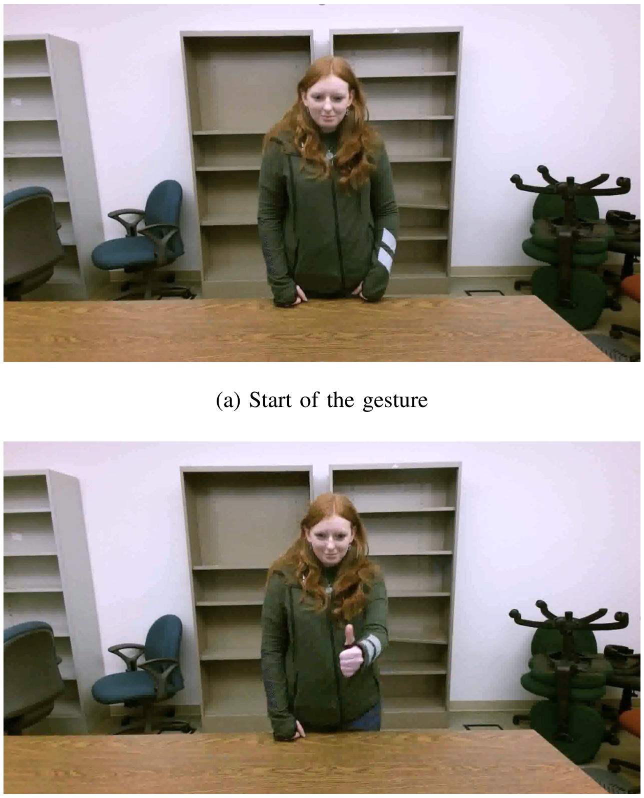
Fig. 4: Examples of non-numeric hand poses, extracted from depth data (similar to the bottom left of Figure 3) using the algorithm of Zhou et al. [29]. "Claw" mimics the hand shape needed to pick up a block. "Closed" and "Open" represent shapes where the hand is flat; in "Closed" the fingers are together, while in "Open" they are spread out. The numbers one through five add five more hand poses, for a total of 13.

3) Relation: Describes relations between body parts. For instance, people may put their hands together, forming a *contact* relation. Relations may also describe hand positions relative to each other, such as *facing* or *opposed* palms. The relation terms are: *contact, crossed, facing, gap, hold, interleaved, opposed, (hands) to face* and *(hands) to hips.*

Motions, poses and relations have optional orientations. We use egocentric coordinates: *up, down, left, right, front,* and *back*. For unsigned motions, for example the direction of a *gap*, we use the first term in every pair: *up, left* or *front*. To see how body parts, motions, poses and relations fit together, Figure 5 shows the first and last frames from an instance of *right arm: move, up; right hand: into thumbs, up*.

2) Intent Labels: The intent track records the inferred intention of the participant, within the context of communicating a block layout. Intents are a higher level description of communication, and the goal was to make intent descriptions as close as possible to English.

Multiple physical gestures often group into a single intent label. For example, a person might move their arm to the left, point to the left, and then move their arm back to their body. These gestures are grouped and given a single intent label,



(a) Start of the gesture

(b) End of the gesture

Fig. 5: First and last frame of an instance of the *Right arm:move, up; right hand: thumbs, up* gesture.

new block, because in the context of the experiment unused blocks are on the left side of the table.

Annotators used combinations of 41 intent labels to describe the data. Some are numbers (*one* through *five*), while some indicate spatial positions (*here, there*) or relative objects (*this, these, that, those*). Some labels are actions (*new, slide, stack*), while others are objects (*yourself, single blocks, gaps between blocks,* or *rows, columns, stacks, layers or pyramids of blocks*). Others are directions (*left, right, forward, backward, middle, between, top, down,* or *diagonal*). Still more pertain to the state of the conversation (*start, wait, done, ok, yes, no, stop*).

Although the structure of intent labels is less rigid than the structure of physical gesture labels, there are terms that modify other terms:

1) *Servo*: Used in conjunction with an action to denote that the action is to continue until a stop command is given. For instance, one may repeat a beckoning gesture until the other person moves the block to the correct position, at which point one either stops gesturing or an explicit stop gesture is given. This example would be labeled as a *servo slide back*. Intents without this word are assumed to be command/response.
2) *Relative*: Used when the direction and/or location is relative to a previous reference point. Intents without this word are in respect to the table (global). For example, one may point to a spot on the table (setting a reference point) and then point to a spot to the right of that. The second gesture would then have an intent of *here relative right*.

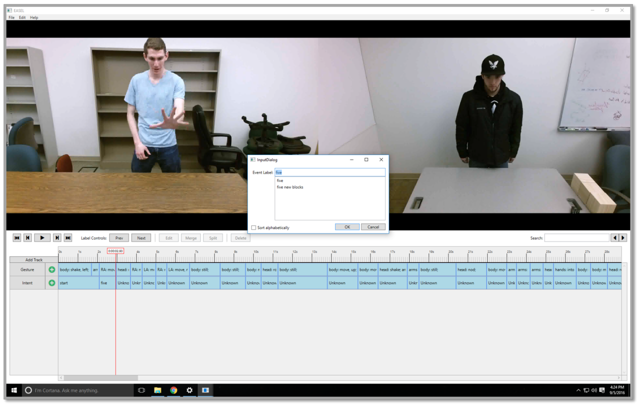


Fig. 6: The labeling tool

As our dataset has two types of videos (with sound and without sound), we did not use sound when determining intent. Instead, we describe the intent of the gesture alone. Gestures whose intent is not clear are labeled *unknown*.

B. Video Labeling

1) Segmentation: The labeling process begins by automatically segmenting videos into gesture instances. We use an automatic segmentation algorithm that models body poses as points in a 51 dimensional pose space (because there are 17 meaningful 3D joint positions). Motions are curves in this pose space, and we segment the videos at local curvature maxima. This segmentation is based on the work of Arn, et al. [1].

The automatic segmentation algorithm is not perfect, and human inspection is still needed. In some cases, the start or end of gestures are miss-timed, and in other cases, multiple gestures are grouped into a single gesture. Using the tool described below, human annotators split, merge, and otherwise correct segments as needed.

2) Labeling Support: To assist with the labeling process, we used a labeling tool called *Easel*. Easel (see Figure 6) features a graphical user interface with a ”piano roll” notation similar to that found in video editing software. When loading a video for annotation, the tool segments the file into gesture instances as described above. Annotators then assign physical gesture and intent labels with the help of pull-down menus, side-by-side video playback, and auto-completion.

3) Data Format: Annotations are stored in XML. We use a hierarchical XML format consisting of Session, Track, and Annotation elements, where a Session represents a video and contains one or more Tracks, a Track is a stream of Annotations over the length of the video and can be of either type Gesture or Intent, and an Annotation is a label for a specific gesture over a defined period of time. Each of these elements contains meta-data such as names, descriptions, etc. This hierarchy is depicted below:

```
<Session>
    <Tracks>
        <Track>
            <Annotations>
                <Annotation />
                ...
            </Annotations>
        </Track>
        ...
    </Tracks>
</Session>
```

Each annotation contains the start time, duration, and labels of a gesture instance. Start/end frames are also provided for precise analysis. Time follows a hh:mm:ss format, with fractional seconds to tick (100 ns) resolution. A sample annotation is below:

```
<Annotation>
    <Time>00:00:02.5342835</Time>
    <Duration>00:00:01.1000034</Duration>
    <StartFrame>69</StartFrame>
    <EndFrame>99</EndFrame>
    <Label>body: still;</Label>
</Annotation>
```

V. DATASET STATISTICS

All together, EGGNOG has 40 subjects performing 360 trials, which in turn are segmented into 24,503 individual motions. As described above, physical gesture labels contain (possibly multiple) body parts, poses, motions, relations and directions. We observed 5,060 distinct labels, i.e. combinations of parts, poses, motions, relations and directions. 330 of these multi-part labels occur 6 or more times.

Breaking down gesture labels into components, we find that *body* was the body part most commonly labeled in gestures, appearing in 6,718 labels. This is misleading, however, because people tend to stand still while waiting for their partners to complete an action. As a result, the label *body: still* appears 4,307 times, leaving only 2,411 gestures that include a moving body. In comparison, right arm motions are labeled 4,482 times, joint motions of arms are labeled 4,318 times, and left arm motions are labeled 2,484 times. The ratio of right to left arm motions is probably because 36 of 40 participants were right handed. Table I lists the most often labeled body parts in order of the number of occurrences.

Table I shows that collectively right, left and combined arm motions account for 11,284 segmented motions. This is over half of all motions, if we ignore *body: still*. Some of these motions are communicative gestures, but most simply reflect the tendency of people to punctuate their speech with arm motions (what psychologists call *beats* [20]). Hand poses often determine whether an arm motion is meaningful. Table II shows the frequencies of hand poses.

Subpart	Occurrences
body	6718
RA (Right Arm)	4482
arms	4318
LA (Left Arm)	2484
head	2285
RH (Right Hand)	1206
hands	1019
LH (Left Hand)	523

TABLE I: Counts of gesture occurrences by body part

Pose:Orientation	Occurrences
Claw:down	1398
Thumbs:up	855
Open:down	665
Point:front	664
Fist	470
Point:down	396
Closed:down	309
Closed:facing	248

TABLE II: Counts of most common hand poses, with directions where appropriate. (In our labeling scheme, Fists do not have orientations.)

We provide multi-part labels because it is usually a combination of factors that determine the meaning of a gesture. The right arm moving forward, for example, is more meaningful if the right hand is in a pointing pose. Table III shows the 15 most common gestures. We have already discussed *body: still*. *Head: rotate* is common because people look from the table to the pattern and then back to the table quite often. It's a head motion, but not really a gesture. The same might be said of moving the right arm or both arms down. Arguably, the most commonly occurring communicative gesture is *head: nod*, a sign of agreement. The next might be *arms: apart, left*. This is where the subject moves their arms apart, generally as a signal to increase the space between two objects. In general, as we move down Table III, the gestures become less common but more communicative. One of the challenges of gesture recognition is to recognize meaningful gestures among all the other motions.

Table IV shows the 10 most common gestures that involve more than one body part. Without exception, these are com-

Gesture	Occurrences
body: still	3137
head: rotate	749
RA: move, down	610
arms: move, down;	438
head: nod	407
LA: move, down	250
RA: move, right	223
arms: apart, left;	220
RA: move, up	216
arms: move, back	212
RA: move, left	176
RH: rotate	163
RA: move, front	152
body:move, front	151

TABLE III: Top 20 most often occurring gestures.

Gesture	Occurrences
RA: move, up; RH: into thumbs, up	149
arms: move, up; hands: into thumbs, up	113
RA: move, front; RH: into point, front	91
RA: move, up; RH: into point, front	86
arms: move, down; hands: into claw, down	80
RA: move, down; RH: into claw, down	57
RA: move, up; RH: into claw, down	54
LA: move, up; LH: to face	51
RA: move, up; RH: to face	50

TABLE IV: Top 10 most often occurring gestures with more than one body part.

Intent	Occurrences
Wait	3431
Think	2033
Talk	1871
Here	1129
Yes	668
Here relative	507
These	354
No	326
Done	323
This	301
Rotate	299
There	287
Stack	266
Slide Left	250
That	248

TABLE V: Top 15 most often occurring intent labels.

binations of arm motions and hand poses, and almost all are meaningful gestures. In general, motions with multiple body parts are usually, although not always, important.

Finally, Table V shows the most common intent labels assigned to segments. It is clear that participants spent a lot of time not doing much: the top three labels are *wait, think*, and *talk*. The next most common category are deictic labels: *here, here relative, these, this, there*, and *that* all appear in the top 15 labels. Then come the social cues: *yes, no*, and *done* (*wait* and *think* are also social cues). Action verbs are less common. Only *rotate, stack* and *slide left* make the list, all near the bottom. Numbers and directions do not crack the top 15, although they do show up farther down the list.

VI. Baseline: Hand Pose Recognition

EGGNOG supports many lines of research, and we anticipate that researchers will extract many different data subsets for various recognition tasks. The EGGNOG web site will host these subsets when requested to. As an example, we begin by providing protocols and baseline results for the task of recognizing hand poses and orientations from depth data.

Hand label recognition in EGGNOG is challenging for two reasons. First, hand labels cannot be recognized in the body position data. The skeletons provided by the Kinect v2 contain 3D positions for only the palm and the tips of the index finger and thumb. None of the other fingers are represented, and the positions of the thumb and index finger are highly noisy. Second, as described in Section IV-A1, hand labels are complex, having both poses and orientations. For example, the hand pose for the number two can be made with either the palm or the back of the hand facing the camera. We have identified 25 different pose/orientation pairs that appear commonly in EGGNOG.

Fold	Normalized	total
	Accuracy	Accuracy
1	0.93	0.88
2	0.82	0.83
3	0.84	0.82
4	0.84	0.80
5	0.82	0.78
Avg	0.85	0.82

TABLE VI: Accuracy of ResNet-style [12] DCNN predicting hand poses and orientations from depth images. Since the number of samples per class is uneven, we report both the average class accuracy and the total accuracy for each fold, and the averages across all five folds.

As a protocol for testing hand label recognition, we have extracted $28,506$ samples of right hands across all 40 subjects, unevenly distributed across the 25 pose/orientation categories. The smallest category has 149 samples; the largest has $4,947$. The samples were identified based on EGGNOG's labels, and for each sample a 128×128 depth image was extracted, centered on the position of the right palm in the body position data. The task is to identify the orientation and pose of a hand, given the depth image. The samples are divided into 5 folds of 8 subjects each, such that no subject appears in more than one fold, to support cross-validation testing. Since some pose/orientations pairs are less common than others, the training set is augmented by creating artificial samples that are minor translations and rotations of the originals. These augmented samples are used to provide at least $4,000$ training samples per class, but the augmented samples are not used during testing. The original hand samples, augmented samples, and folds are all provided on the EGGNOG web site.

Our baseline algorithm for this task is a ResNet-style [12] Deep Convolutional Neural Network with 50 layers. A total of 5 nets were trained, with each net being trained on 4 data folds and tested on the fifth. The accuracy of the resulting nets is shown in Table VI, along with the average accuracies. Since the number of samples per class varies, we report both the average class accuracy and the average sample accuracy.

VII. CONCLUSION

People communicate through words and gestures. This paper presents a new data set (called EGGNOG) of naturally occurring gestures made by people working collaboratively on blocks world tasks. It contains over 8 hours of RGB video, depth video and Kinect v2 body position data of 40 subjects, and has been segmented into 24,503 distinct movements. Each movement has been labeled according to (1) its physical motion and (2) the intent of the signaler.

The goal of this data set is to support research into recognizing the types of gestures that occur during human communication. The data set, including video, depth, and body pose data, is publicly available at https://cwc.cs.colostate.edu/datasets, along with the corresponding segment boundaries and ground truth labels. The same web site will also maintain a list of results from papers using the data set. The segmentation and labeling tool is the subject of another paper (in preparation), but that too will be made available as soon its paper is published.

We are not proposing specific experimental protocols because this data can be used for so many purposes. We are currently studying the differences in the types of gestures people use with and without sound. This data can also be used, however, for studies in hand pose recognition, motion recognition, expression recognition, or combinations thereof. It can be used to study variations within or across subjects. Most importantly, it can be used for evaluating naturally occurring gesture recognition systems.

References

[1] R. Arn, B. Draper, M. Kirby, and C. Peterson. The frenet-serret apparatus and local singular value decomposition of curves in r^n. *arXiv preprint arXiv:1511.05008*, 2015.

[2] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali. The american sign language lexicon video dataset. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

[3] J. R. Bellegarda. Spoken language understanding for natural interaction: The siri experience. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 3–14. Springer, 2014.

[4] V. Bloom, V. Argyriou, and D. Makris. G3di: A gaming interaction dataset with a real time detection and evaluation framework. In *Workshop at the European Conference on Computer Vision*, pages 698–712. Springer, 2014.

[5] M. Chen, G. AlRegib, and B.-H. Juang. A new 6d motion gesture database and the benchmark results of feature-based statistical recognition. In *Emerging Signal Processing Applications (ESPA), 2012 IEEE International Conference on*, pages 131–134. IEEE, 2012.

[6] C. Conly, P. Doliotis, P. Jangyodsuk, R. Alonzo, and V. Athitsos. Toward a 3d body part detection video dataset and hand tracking benchmark. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, page 2. ACM, 2013.

[7] S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton, and I. Guyon. Chalearn looking at people challenge 2014: Dataset and results. In *Workshop at the European Conference on Computer Vision*, pages 459–473. Springer, 2014.

[8] M. Firman. Rgbd datasets: Past, present and future. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 661–673. IEEE, 2016.

[9] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1737–1746. ACM, 2012.

[10] D. Freeman, R. Jota, D. Vogel, D. Wigdor, and R. Balakrishnan. A dataset of naturally occurring, whole-body background activity to reduce gesture conflicts. *arXiv preprint arXiv:1509.06109*, 2015.

[11] I. Guyon, V. Athitsos, P. Jangyodsuk, H. J. Escalante, and B. Hamner. Results and analysis of the chalearn gesture challenge 2012. In *Advances in Depth Image Analysis and Applications*, pages 186–204. Springer, 2013.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[13] Y.-S. Hsiao, J. Sanchez-Riera, T. Lim, K.-L. Hua, and W.-H. Cheng. Lared: a large rgb-d extensible hand gesture dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference*, pages 53–58. ACM, 2014.

[14] T.-K. Kim, S.-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[15] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.

[16] A. Kurakin, Z. Zhang, and Z. Liu. A real time system for dynamic hand gesture recognition with a depth sensor. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 1975–1979. IEEE, 2012.

[17] Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *2009 IEEE 12th international conference on computer vision*, pages 444–451. IEEE, 2009.

[18] L. Liu and L. Shao. Learning discriminative representations from rgb-d video data. In *IJCAI*, volume 1, page 3, 2013.

[19] G. Marin, F. Dominio, and P. Zanuttigh. Hand gesture recognition with leap motion and kinect devices. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1565–1569. IEEE, 2014.

[20] D. McNeil. *Hand and Mind: What Gestures Reveal About Thought*. The University of Chicago Press, Chicago and London, 1992.

[21] L. Minto and P. Zanuttigh. Exploiting silhouette descriptors and synthetic data for hand gesture recognition. 2015.

[22] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera. Chalearn lap 2016: First round challenge on first impressions - dataset and results. In *ECCV*, pages 400–418, 2016.

[23] Z. Ren, J. Meng, J. Yuan, and Z. Zhang. Robust hand gesture recognition with kinect sensor. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 759–760. ACM, 2011.

[24] S. Ruffieux, D. Lalanne, and E. Mugellini. Chairgest: a challenge for multimodal mid-air gesture recognition for close hci. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 483–488. ACM, 2013.

[25] S. Ruffieux, D. Lalanne, E. Mugellini, and O. A. Khaled. A survey of datasets for human gesture recognition. In *International Conference on Human-Computer Interaction*, pages 337–348. Springer, 2014.

[26] A. Sadeghipour, L.-P. Morency, and S. Kopp. Gesture-based object recognition using histograms of guiding strokes. In *BMVC*, pages 1–11, 2012.

[27] Y. Song, D. Demirdjian, and R. Davis. Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 500–506. IEEE, 2011.

[28] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE MulitMedia*, 19:4–10, 2012.

[29] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei. Model-based deep hand pose estimation. In *IJCAI*, 2016.