1 FSboard

FSboard is an American Sign Language fingerspelling dataset consisting of 3.2 million characters and 266 hours of video. It is collected from 147 paid and consented Deaf signers using the selfie camera of the Pixel 4A smartphone in a variety of environments. Using a variant of the open source smartphone sign collection app "Record These Hands," participants were prompted to fingerspell short phrases, person names, addresses, URLs, and phone numbers. Video is captured at 1944x2592 pixels at 30 frames per second (though sometimes the resolution varied due to participants accidentally changing settings).

This dataset is intended for the creation of fingerspelling recognition systems for text entry or educational software for learning fingerspelling, As conversing in ASL can be 12-35% fingerspelling, it may also assist in creating a more complete ASL recognition system or ASL generation system. Given the sociohistorical context surrounding sign language technology and perceptions of fingerspelling, it is important to emphasize that fingerspelling recognition/transcription *is not* sign language translation. Fingerspelling is an important part of ASL, but ultimately *just a part* of the language. Please do not exaggerate the scope of this dataset or task in any subsequent work.

Dataset Link

We release the dataset under a CC BY 4.0 license. Note that Creative Commons licenses do not supersede other rights such as right of publicity.

Data Card Author(s)

- Thad Starner (Owner)
- Garrett Tanzer (Contributor)

1.1 Authorship

1.1.1 Publishers

Publishing Organization(s)

- Google
- Deaf Professional Arts Network

Industry Type(s)

- Corporate - Tech
- Not-for-profit - Tech

1.1.2 Dataset Owners

Contact Detail(s)

- **Dataset Owner(s):** Manfred Georg, Thad Starner
- **Affiliation:** Google
- **Contact:** mgeorg@google.com, thadstarner@google.com

Author(s)

- Manfred Georg, Google
- Garrett Tanzer, Google
- Esha Uboweja, Google
- Saad Hassan, Tulane University
- Maximus Shengelia, Rochester Institute of Technology
- Sam Sepah, Google
- Sean Forbes, Deaf Professional Arts Network
- Thad Starner, Google

1.1.3 Funding Sources

Institution(s)

Google

Funding or Grant Summary(ies)

Google funded the Deaf Professional Arts Network to collect the dataset with the expectation of the dataset becoming public.

Additional Notes: None

1.2 Dataset Overview

Data Subject(s)

- Sensitive data about people (videos of people)
- Non-sensitive data about people
- Synthetically generated text data

Content Description

This dataset was collected from August 2021 to March 2023.

Dataset Snapshot

Size of Dataset: 266 hours

Number of Signers: 147

Number of Labels: 3.2 million characters

Number of Instances: 151,000 phrases

Average Labels Per Instance: 21.2 characters average per phrase (28.8 MacKenzie phrase; 18.6 addresses/phone/URL/etc.)

Types of Instances: SMS-style phrase (MacKenzie), randomly generated URLs, street addresses, common names (in the United States), phone numbers

Labeled Classes: 74 consisting of 10 digits, space, 11 symbols, 26 letters, 26 capitals (optionally indicated by signers)

1.2.1 Sensitivity of Data

Sensitivity Type(s)

- User Content
- User Metadata
- User Activity Data
- Identifiable Data
- PII

Field(s) with Sensitive Data

Intentional Collected Sensitive Data

(PII were collected as a part of the dataset creation process.)

Participant Video: Video of participant (upper body captured)

Participant Sign: Video of participant performing continuous fingerspelling

Security and Privacy Handling

Method: Participants were given a consent form. They were only allowed to record after providing consent for the following:

"The app will collect video and photographic images of Your face, torso, hands, and whatever is in view of the camera(s) along with associated camera metadata (such as color correction, focal length, etc.). ... Beyond the video, the following data may be recorded:

- The details of each Task, such as the type of Task that was done, signing certain words, or performing specific actions as instructed
- Date and time information associated with the Tasks
- Self identified gender
- Self identified age range
- Self-identified ethnicity
- Self assessed sign language proficiency
- Signing style information (such as general location where You learned, type of sign learned, age range when you started learning, signing community You are most closely associated with, etc)

As described earlier, if you consent, we will use photos or video clips where your face can be identified. We may use identifiable photos or video clips of you in written or oral presentations about this work and in publicly available on-line databases."

Risk Type(s)

- Direct Risk
- Residual Risk

Risk(s) and Mitigation(s)

The direct risk involves participants' visual features (their face and body) being linked to their full name. To mitigate this risk, we use anonymized user IDs to identify users. There is still some residual risk. Participants may still be identified using their faces alone. This risk is unavoidable with video data. We have participants sign consent forms acknowledging that they are creating a dataset intended for public use.

1.2.2 Dataset Version and Maintenance

Maintenance Status

Regularly Updated - New versions of the dataset have been or will continue to be made available.

Version Details

Current Version: 1.0

Release Date: 11/2024

Maintenance Plan

Versioning: Major updates will be released as a new version, incremented to the nearest tenth from the previous version. For example, if the current version is between 1.0 and 1.09, then a major update will be released as version 1.1. Major updates include the addition of new users and/or new categories of phrases. Minor updates are covered below.

Updates: If there are missing/extraneous/erroneous videos (error cases described below), any fixes will be released as a new version, incremented by 0.01. E.g. if the current version is 1.0, then any minor updates will be released as 1.01.

Errors: Errors in the dataset include incorrectly labeled videos, missing videos, or extraneous videos. Missing videos include videos that the participant recorded but were not included in the final release. Extraneous videos include videos that only have partial input or no sign at all, but were included in the final dataset.

Feedback: We will accept feedback to the authors' emails.

Next Planned Update(s)

Version affected: 1.0

Next data update: TBD

Next version: 1.1

Next version update: TBD

Expected Change(s)

Updates to Data: The current dataset includes capitalization but some signers did not indicate capitals explicitly. Often, the same symbol would be signed with a variety of hand shapes, which are not detailed in the annotation. Future annotation improvements may address these shortcomings. While the same 500 MacKenzie phrases were collected repeatedly, that was not the intention for the URLs, names, etc. About 1 million more characters were collected but are not released yet as they unintentionally repeated previous prompts. These may be released at a future date.

1.3 Example of Data Points

Primary Data Modality

- Video Data

Typical Data Point

A typical data point includes the signer fingerspelling a prompt, with little to no empty space (i.e. no motion) at the beginning or the end of the video. The average video is six seconds in length.

Atypical Data Point

An atypical data point may be improperly clipped/realigned, so the clip contains contains no fingerspelling, only part of the desired fingerspelled phrase, an erroneous version of the phrase and then a redo, or the wrong fingerspelled phrase.

1.4 Motivations & Intentions

1.4.1 Motivations

Purpose(s)

- Research
- Production
- Education

Domain(s) of Application

`Educational Technology`, `Accessibility`, `Sign Language Recognition`, `Machine Learning`, `Computer Vision`, `Fingerspelling Recognition`, `American Sign Language`

Motivating Factor(s)

- Text Entry
- Teaching Sign
- Developing Educational Technology
- Developing Sign Language Recognition

The MobileHCI 2023 paper “Tap to sign: Towards using american sign language for text entry on smartphones” demonstrated that Deaf signers fingerspelled significantly faster, on average, than using standard smartphone virtual keyboards when performing text entry tasks. Data collected here suggest that speeds could average as much as twice as fast. In addition, due to dexterity limitations, some older Deaf signers have difficulty with mobile phone text entry but can fingerspell fluently. Our dataset is collected on mobile phones and is designed to facilitate text entry on smartphones for the Deaf. Additionally, we foresee the use of this dataset for creating educational technology for helping those who wish to learn sign language gain speed and fluency when fingerspelling. Finally, fingerspelling is an integral part of sign in general, and this dataset may be useful in working towards a conversational sign language recognition or generation system.

1.4.2 Intended Use

Dataset Use(s)

- Safe for research use

Suitable Use Case(s)

Suitable Use Case: ASL fingerspelling recognition and generation

Suitable Use Case: ASL fingerspelling text entry

Suitable Use Case: ASL fingerspelling educational software

Suitable Use Case: Pretraining for fingerspelling tasks in other sign languages with a similar alphabet

In general, the data can be used for ASL fingerspelling recognition/generation and related downstream applications.

Unsuitable Use Case(s)

Unsuitable Use Case: Full Translation/Generation between ASL and English

Unsuitable Use Case: Face Recognition

By itself, this data is not sufficient for continuous sign language recognition/generation or sign to spoken language translation.

Research and Problem Space(s)

This dataset is primarily intended to create a fingerspelling recognition system suitable for text entry.

Citation Guidelines

Please cite FSboard as follows:

```
@misc{georg2024fsboard3millioncharacters, title={FSboard: Over 3 million characters of ASL fingerspelling collected via smartphones}, author={Manfred Georg and Garrett Tanzer and Saad Hassan and Maximus Shengelia and Esha Uboweja and Sam Sepah and Sean Forbes and Thad Starner}, year={2024}, eprint={2407.15806}, archivePrefix={arXiv}, primaryClass={cs.CV}, url={https://arxiv.org/abs/2407.15806}, }
```

1.5 Access, Rentention, & Wipeout

1.5.1 Access

Access Type

- External - Open Access

Documentation Link(s)

- Dataset Website: https://www.kaggle.com/datasets/googleai/fsboard

1.5.2 Retention

Duration

The dataset will be available for at least 5 years.

1.6 Provenance

1.6.1 Collection

Method(s) Used

We collected data from an open-source mobile recording app called “Record These Hands.” The app presents 10 phrases for recording in a single recording session. The entire session was captured on video, but the sign recordings happened during specific time intervals. The participants were presented a phrase to record. They then tapped a record button to record themselves signing and tapped again to finish signing. The timestamps corresponding to the recording intervals for each sign were saved in a separate file.

Collection Method: Record These Hands App

Platform: Android, Google Pixel 4a

Dates of Collection: [09 2021 - 03 2023]

Primary modality of collection data:

- Video Data

Update Frequency for collected data:

- Static

Source Description(s)

Participants were recruited by DPAN. Participants are Deaf signers who use ASL as their primary language.

Collection Cadence

Static: Data was collected once or twice from single or multiple sources.

Data Processing

Collection Method: Record These Hands

Description: We split session recordings from Record These Hands using python. The resulting split videos were named following this convention: "<participant id>-<phrase id>-<recording start time>-<recording index>.mp4".

Tools or libraries: Python, FFmpeg

1.6.2 Collection Criteria

Data Selection

Due to bugs in an early version of the data collection app and some user errors, the timespans recorded above frequently did not capture the actual phrase. In order to generate reasonable clips a bootstrapping method was used with a ByT5 model pretrained on YouTube sign language data. First the model was trained on a large corpus of YouTube videos with captions. Next 5 models were finetuned to transcribe fingerspelling data using 4/5ths of the clips with lots of incorrect bounds. Since evaluation bias was irrelevant, the dataset splits were performed at the clip level such that every participant was in every split. Each model was then used to predict text for the remaining 1/5th that it had not yet seen. Where the model agreed with the clip boundaries and content, the clip was labeled as clean, otherwise the clip was labeled as noisy. The whole process was repeated 2 more times (starting each time with a fresh model) using only the clean clips from the previous round. A significant amount of manual editing and custom rules tailored to each participant were then used to further clean the clips. There are still significant issues with the clip boundaries in the dataset and it would benefit from further annotation.

Data Inclusion

See above.

Data Exclusion

See above.

1.6.3 Relationship to Source

Use & Utility(ies)

FSboard is intended to train a real-time fingerspelling text entry system for use on smartphones.

Benefit and Value(s)

A fingerspelling text entry system may prove significantly faster than currently smartphone text entry for the Deaf; may be preferred by Deaf signers; and may cause less discomfort.

Limitation(s) and Trade-Off(s)

FSboard only focuses on fingerspelling, not full sign language recognition.

1.7 Human and Other Sensitive Attributes

Sensitive Human Attribute(s)

- Language
- Culture
- Age
- Gender
- Ethnicity

Intentionality

Intentionally Collected Attributes

Language: The signers were selected to use ASL as their primary language.

Culture: The signers were selected to be culturally Deaf.

Unintentionally Collected Attributes

Human attributes were not explicitly collected as a part of the dataset creation process, but given that the data includes videos of the participants (including their face), attributes like age, gender, ethnicity, etc. can be predicted using additional methods.

Rationale

We intended to collect continuous fingerspelling data; hence videos of the participant's signing were collected. The collected attributes (both intentional and unintentional) may be inferred (though not always accurately) from the videos.

Source(s)

We had three professional raters rate the perceived gender and Monk skin tone for FSboard participants. The majority assignment was used. Please see the main paper for details and distributions.

Limitation(s) and Recommendation(s)

Limited scope: FSboard is not intended to address full sign language recognition/translation. Its focus is fingerspelling.

Re-identification: Do not combine FSboard with other datasets in order to re-identify participants. Blur faces when including images in publications or demonstration videos.

Risk(s) and Mitigation(s)

The direct risk with this type of video data is with the participant's identity being revealed. For this reason, we use anonymous identifiers. There is still some residual risk with the participant being identified through their faces alone. These participants have signed a consent form (given in the Data Sensitivity section) to address this concern.

1.7.1 Use in ML or AI Systems

Dataset Use(s)

- Training
- Testing
- Validation
- Development or Production Use
- Fine Tuning

1.8 Annotations Labeling

- Annotation Target in Data
- Machine-Generated
- Human Annotations (Expert)

Annotation Characteristic(s)

Besides the automatic affiliation of the phrase with the corresponding video of fingerspelling, we were interested in knowing whether we were collecting a diverse set of skin tones and genders.

We had three professional raters rate the perceived gender and Monk skin tone for FSboard participants. The majority assignment was used. Please see the main paper for details and distributions.

1.9 Known Applications & Benchmarks

Fingerspelling recognition

Evaluation Result(s)

See the paper for our baselines on fingerspelling recognition using MediaPipe Holistic and a fine-tuned ByT5 Small. We achieve 11.1% CER and 52.1% top-1 accuracy on the test set, which has nonoverlapping signers and phrases with respect to the training set.

Expected Performance and Known Caveats

FSboard has a relatively limited domain, so depending on the method models trained on it may not generalize to other domains.

1.10 Terms of Art

ASL

American Sign Language

Fingerspelling

A system within various sign languages for spelling out words using a manual alphabet. This is just one small component of sign languages.

1.11 Reflections on Data

Some of the randomly generated URLs collected in the dataset may be offensive to some parties. We do not include phrases that our automatic classifiers deem explicit in the main dataset metadata, but rather release it as a separate file so that it must be used consciously.