ECCV'22 SLRTP Challenge

https://slrtp-2022.github.io/

Table of Contents

Track 1: Sign Recognition on BOBSL

Codalab competition page: https://codalab.lisn.upsaclay.fr/competitions/6724

The task is, given a point in time within a co-articulated sign language video, to classify the sign into one of the categories in the vocabulary.

➢ Download
Development phase: slrtp22-track1-development.zip
- `bobsl_vocab_2281.json`
- `dev.json`

Test phase: slrtp22-track1-test.zip (available after September 7)
- `reference_dev.json`
- `test.json`

➢ Development and test data
There are two files that define the data samples for the development and test sets. The development set is intended to be used to tune hyperparameters.

During the development phase, the participants are allowed to submit up to 10 submissions per day, and a total of 300 submissions to the Codalab evaluation server. After the test phase begins, the development annotations will be made available.

During the test phase, the participants are allowed to submit a total of 5 submissions. The submission with the best performance will be used in the leaderboard.

We provide two files:
`dev.json` with 24634 development samples,
`test.json` with 35231 test samples.

The content looks like the following, where the key represents the index of the sample (e.g. 0, 1, 2, …, 24633 for development), the `name` refers to the video file name in BOBSL, and the `anno_time_mid` refers to the point in time when the sign occurs in the video (note that the temporal boundaries of the sign are not given).


```
{
    "0": {
        "name": "5213374327398864506.mp4",
        "anno_time_mid": 998.704
    },
    "1": {
        "name": "5220442983744416655.mp4",
        "anno_time_mid": 1092.46
    },
    "2": {
        "name": "6023578245784400210.mp4",
        "anno_time_mid": 275.706
```

```
    },
    ...
}
```

➢ Vocabulary
The vocabulary consists of 2281 classes defined in `bobsl_vocab_2281.json`, where the key represents the index of the class (e.g., 0, 1, 2, …, 2280), and the value corresponds to an English representation for the sign category. Note that the vocabulary is automatically constructed, and there may be categories that share similar manual features, or a single category containing multiple manual features.

```
{
"0": "abandon",
"1": "aberdeen",
"2": "ability",
...
}
```

➢ Submission format
To submit predictions to the Codalab server, participants should prepare a `submission.csv` file in the following format. Note that the file name should exactly match `submission.csv`, which should then be zipped (with any file name, e.g., `submission.csv.zip`) to be then uploaded to Codalab.

Each row should contain 6 comma separated integers, where the first integer denotes the data index to refer to `dev.json` or `test.json` samples. The remaining 5 integers correspond to the top-5 predicted classes, ordered so that the most likely class appears first. The class index refers to the vocabulary defined in `bobsl_vocab_2281.json`.

```
0,816,824,1410,815,740
1,1244,1504,867,484,866
2,1762,1063,1769,2098,912
...
```

➢ Evaluation metric
The evaluation server reports 4 metrics in the leaderboard: top-1 and top-5 classification accuracies either per-instance (without class-balancing), or per-class (with class-balancing). Participants are ranked using the top-1 per-class metric.

Top-k per-instance accuracy records the percentage of correctly classified predictions where a sample is counted as correct if the ground truth label appears in the top k predicted classes.

Per-class accuracy computes an individual accuracy for each class, and averages accuracies over the classes present in the evaluation set. We use this metric due to the unbalanced nature of the evaluation sets.

➢ Baseline
We provide a baseline from https://arxiv.org/pdf/2111.03635.pdf where we use the RGB-I3D recognition model described in Section 4.1, whose performance is reported in Table 5. This model is released at https://www.robots.ox.ac.uk/~vgg/data/bobsl/rgb-i3d-recognition.pth.tar and can be used along with the code repository https://github.com/gulvarol/bsl1k.

Track 2: Subtitle Alignment on BOBSL

Codalab competition page: https://codalab.lisn.upsaclay.fr/competitions/6790

➢ Description
In the BOBSL dataset, like in many TV programmes with sign language interpretation, the subtitles are not aligned to the signing but to the spoken audio. This challenge is about aligning subtitle text to sign language, in order to create pairs of written sentences and their sign language interpretation. There is an average lag of around 2.7s between the audio-aligned subtitles and the corresponding sign language segments. Correcting for this lag gives a good approximation for the ground truth signing-aligned subtitles. This approximate alignment can be further improved by learning semantic cues from sign language (e.g. matching words in the subtitle text to signs) and learning prosodic content (e.g. short pauses in signing at sentence boundaries). A subset of BOBSL contains manually-annotated signing-aligned subtitles, which may be used for strong supervision. See https://arxiv.org/abs/2105.02877 for more details on subtitle alignment.

➢ **Download**
Development phase: slrtp22-track2-development.zip
- `dev.json`

Test phase: slrtp22-track2-test.zip (available after September 7)
- `reference_dev.json`
- `test.json`

➢ **Development and test data**
There are two files that define the data samples for the development and test sets. The development set is intended to be used to tune hyperparameters.

During the development phase, the participants are allowed to submit up to 10 submissions per day, and a total of 300 submissions to the Codalab evaluation server. After the test phase begins, the development annotations will be made available.

During the test phase, the participants are allowed to submit a total of 5 submissions. The submission with the best performance will be used in the leaderboard.

We provide two files:
`dev.json` with 11135 development samples,
`test.json` with TBA test samples.

The content looks like the following, where the key represents the index of the sample (e.g. 0, 1, 2, …, 11134 for development), the `name` refers to the video file name in BOBSL.

```
{
    "0": {
        "name": "5952613360146027617.mp4",
        "subtitle_text": "This land of ours, its mountains and valleys, fields and forests,
a place to live, a place to work, a place to enjoy.",
        "subtitle_index": 0
    },
    "1": {
        "name": "5952613360146027617.mp4",
        "subtitle_text": "But it's much more than that.",
        "subtitle_index": 1
    },
    "2": {
        "name": "5952613360146027617.mp4",
        "subtitle_text": "Our landscape teaches us things as well.",
        "subtitle_index": 2
    },
    ...
}
```

➢ **Submission format**
To submit predictions to the Codalab server, participants should prepare a `submission.csv` file in the following format.

Each row should contain 3 comma separated numbers (one integer, two floats), where the first integer denotes the data index to refer to `dev.json` or `test.json` samples. The remaining 2 numbers correspond to the predicted start and end times of the signing-aligned subtitles.

```
0,42.36,42.52
1,42.52,53.4
2,53.4,56.76
…
```

Note that the file name should exactly match `submission.csv`, which should then be zipped (with any file name, e.g., `submission.csv.zip`) to be then uploaded to Codalab.

➢ **Evaluation metrics**

The evaluation server reports 4 metrics: frame accuracy, and F1-scores at three different IoU thresholds. For the F1-score, hits and misses of subtitle alignment to sign language video are counted over three temporal overlap thresholds (IoU = {0.10,0.25,0.50}) between predicted signing-aligned subtitles and manually annotated signing-aligned subtitles. The main evaluation metric is considered to be F1@0.50.

➤ Baseline
See https://github.com/hannahbull/subtitle_align for a baseline model. Alternatively, a simple baseline can be made by shifting the audio-aligned subtitles by their average lead of 2.7s.

```
import webvtt
import pandas as pd

dict_csv = {'subtitle_start': [], 'subtitle_end': []}
for vtt_file in list_vtt_files: # put list of audio-aligned vtt files here
      subs = webvtt.read(vtt_file)
      for idx, sub in enumerate(subs):
            dict_csv['subtitle_start'].append(sub._start + 2.7)
            dict_csv['subtitle_end'].append(sub._end + 2.7)

df = pd.DataFrame.from_dict(dict_csv)
df.to_csv(f'res/submission_{dev_test}_shift.csv', header=False)
```

Track 3: Sign Spotting on BSL Corpus

Codalab competition page: https://codalab.lisn.upsaclay.fr/competitions/6803

The task is, given a query sign gloss and a short video of continuous sign language, to determine whether or not the video contains the sign gloss and, if so, to localise the sign gloss.

➤ Download
Development phase: slrtp22-track3-development. Please note that this data may be used for academic non-commercial research only.
- `bslcp_vocab_981.json`
- `dev.json`
- `bslcp_challenge_data/train`
- `bslcp_challenge_data/dev`
- `bslcp_challenge_logits/train`
- `bslcp_challenge_logits/dev`
- `bslcp_challenge_data_bobsl/train`
- `bslcp_challenge_data_bobsl/dev`
- `bslcp_challenge_labels/train`

Test phase: slrtp22-track3-test (available after September 7). Please note that this data may be used for academic non-commercial research only.
- `reference_dev.json`
- `test.json`
- `bslcp_challenge_data/test`
- `bslcp_challenge_logits/test`
- `bslcp_challenge_data_bobsl/test`
- `bslcp_challenge_labels/dev`

➤ Development and test data
The development set is intended to be used to tune hyperparameters.

During the development phase, the participants are allowed to submit up to 10 submissions per day, and a total of 300 submissions to the Codalab evaluation server. After the test phase begins, the development annotations will be made available.

During the test phase, the participants are allowed to submit a total of 5 submissions.

We provide two files:
`dev.json` with 5886 development samples,

`test.json` with TBA test samples.

The content looks like the following, where the key represents the index of the sample (e.g. 0, 1, 2, …, 5885) for development), the `video` refers to the video file name, and the `query` refers to the vocabulary ID of the sign gloss.


```
{
    "0": {
        "video": "Conversation-Bristol-29+30-BL30c_000272-430_000277-450.npy",
        "query": 419
    },
    "1": {
        "video": "Conversation-Bristol-29+30-BL30c_000272-430_000277-450.npy",
        "query": 791
    },
    "2": {
        "video": "Conversation-Bristol-29+30-BL30c_000272-430_000277-450.npy",
        "query": 273
    },
    ...
}
```


➢ Features

- `bslcp_challenge_data/*`
  This folder contains features from an I3D model for sign classification pre-trained on Kinetics and fine-tuned on BSL Corpus, as described in https://arxiv.org/pdf/2011.12986.pdf. The checkpoints for this model `i3d_kinetics_bslcp.pth.tar` are released at https://github.com/RenzKa/sign-segmentation.
- `bslcp_challenge_logits/*`
  This folder contains the logits (981 classes) for the I3D model above.
- `bslcp_challenge_data_bobsl/*`
  This folder contains features from the same model as the BOBSL feature data release. See tracks 1 and 2 for details.

➢ Labels

Frame level sign gloss labels are provided in `bslcp_challenge_data_bobsl/*`

➢ Vocabulary

The vocabulary consists of 981 classes defined in `bslcp_vocab_981.json`, where the key represents the index of the class (e.g., 0, 1, 2, …, 980), and the value corresponds to an English representation for the sign category.


```
{
"0": "ABOUT",
"1": "ABOUT-NUMBER",
"2": "ACCEPT",
...
}
```


➢ Submission format

To submit predictions to the Codalab server, participants should prepare a `submission.csv` file in the following format. Note that the file name should exactly match `submission.csv`, which should then be zipped (with any file name, e.g., `submission.csv.zip`) to be then uploaded to Codalab.

Each row should contain 4 comma separated integers, where the first integer denotes the data index to refer to `dev.json` or `test.json` samples. The second index is a binary 0/1 variable, denoting whether or not the sample contains the query sign gloss (1=contains gloss, 0=does not contain gloss). The third and fourth integers denote the frame number of the start and end of each detected sign gloss (closed interval [start, end]). If the second integer is 0, the third and fourth integers should be equal to -1.

```
0,0,-1,-1
1,1,15,22
2,1,7,18
```

. . .

➢ Evaluation metrics
The evaluation server reports 7 metrics in the leaderboard. The main metric is Score @ 0.50 IoU. This is computed as the sum of 4 other metrics: Correct prediction, Correct absence, Incorrect prediction * -1, Incorrect absence * -1.

1. Correct predictions occur when the IoU between the predicted localisation of the sign gloss and the ground truth localisation is greater than or equal to 0.5.
2. Correct absences occur when a sign gloss is correctly predicted to be absent in the video clip.
3. Incorrect predictions occur when either a sign gloss is localised although absent in the ground truth, or the IoU between predicted localisation of the sign gloss and the ground truth localisation is lower than 0.5.
4. Incorrect absences occur when a sign gloss is predicted to be absent in the video clip, but it is actually present.

These metrics are divided by the number of samples in the evaluation set.
We additionally include the metrics Score @ 0.25 IoU and Score @ 0.10 IoU, which correspond to IoU thresholds of 0.25 and 0.1 respectively.

➢ Baseline
See https://github.com/hannahbull/slrtp2022_t3 for a simple baseline model to get started.

Terms and Conditions

➢ Data licenses
Participants are responsible for obtaining the necessary licenses to get access to the datasets.

➢ Usage of other data sources and pre-trained models
Any other data sources and pre-trained models may be used for this challenge, as long as they are publicly available to other participants, i.e. available for non-commercial research purposes.

➢ Provide code
Winners agree to provide code at the end of the challenge, in order to ensure reproducibility of the winning models.

➢ Warranty and limitation on liability (modified from ChaLearn)
Datasets used in the challenge, if no other information is provided at each specific dataset webpage, are presented "AS IS". By downloading and using them, you acknowledge they may contain errors, and take full responsibility on any potential risk or damage. To the fullest extent provided by law, in no event will we, our affiliates, or our licensors, service providers, employees, agents, officers, or directors be liable for damages of any kind, under any legal theory, arising out of or in connection with the developer's use, or inability to use, the services, datasets, any content on the services or such other services, including any direct, indirect, special, incidental, consequential, or punitive damages, including but not limited to, personal injury, pain and suffering, emotional distress, loss of revenue, loss of profits, loss of business or anticipated savings, loss of use, loss of goodwill, loss of data, and whether caused by tort (including negligence), breach of contract, or otherwise, even if foreseeable.

➢ Disclaimer (modified from ChaLearn)
Although efforts have been devoted to the compilation and curation of resources, collected data including associated annotations and labels (obtained manually, automatically and semi-automatically) may not be representative samples of real application scenarios. The adopted data gathering and labelling methodologies may not include exhaustive and/or inclusive mechanisms that allow users to reach conclusive findings.

BOBSL dataset for Tracks 1 and 2

➢ About BOBSL
See https://arxiv.org/pdf/2111.03635.pdf for a detailed description of BOBSL. Note that there are a small number (22) of videos that are no longer available in BOBSLv1_2 (the version used for the challenge), and so the statistics for the train, validation and test sets differ slightly from the numbers presented in the article.

➢ Non-Commercial Research Only Licence
Through a data-sharing agreement with the BBC, BOBSL is available for non-commercial research usage. This challenge is thus restricted to participants from academic institutions. BSL translation services are currently supplied to the BBC by Red Bee Media Ltd. They have indicated that they and their staff are happy for their footage to be used for research purposes. However, if the position changes the dataset will need to be revised accordingly. Researchers should be

mindful of this, and should be aware that the 'Permission to Use' form they will need to sign obligates them to delete portions (or, indeed, the whole) of the dataset in the future, if so instructed. For full information on the BOBSL data license, please visit https://www.robots.ox.ac.uk/~vgg/data/bobsl/ and https://www.bbc.co.uk/rd/projects/extol-dataset.

➤ **Instructions for obtaining the challenge data**
To download the data to be used for the challenge, follow the instructions at https://www.robots.ox.ac.uk/~vgg/data/bobsl/.

➤ **Train split**
You may use any data from the public training split (1658 episodes in BOBSLv1_2) for training your models. You may not use any data from the public test split (250 episodes in BOBSLv1_2) for training models.

BSL Corpus for Track 3

➤ **About BSL Corpus**
The BSL Corpus is a linguistic corpus of British Sign Language with partial annotations, including sign gloss annotations. See https://bslcorpusproject.org/ for a detailed description of the BSL Corpus project.

➤ **Licence**
A subsection of the BSL Corpus is openly available for research and teaching purposes (https://bslcorpusproject.org/cava/open-access-data/). Notably, the subsection of the Track 3 challenge data containing Narrative elicitations can be directly downloaded from http://digital-collections.ucl.ac.uk/R/?local_base=BSLCP. Part of the BSL Corpus used in Track 3 is however only available for research purposes and access is granted upon agreement of the terms and conditions (https://bslcorpusproject.org/cava/restricted-access-data/).

➤ **Instructions for obtaining the challenge data**
We provide video features for the training data in BSL Corpus. It is possible to train models using these features and thus complete this challenge without requesting data (see Track 3 description). The original videos of a subset of the BSL Corpus data used for this task (Narrative elicitations) are directly available without requesting access. To request access for the remaining BSL Corpus videos used in this task, see https://bslcorpusproject.org/cava/restricted-access-data/.

➤ **Train split**
You may use any videos in the provided training sample (see Track 3 description) for training your models. You may not use any data from the test sample for training models.