SIGNER DIARISATION IN THE WILD

Samuel Albanie1⋆ Gül Varol1,2⋆ Liliane Moment1⋆ Triantafyllos Afouras1
Andrew Brown1 Chuhan Zhang1 Ernesto Coto1 Necati Cihan Camgöz3 Ben Saunders3
Abhishek Dutta1 Neil Fox4 Richard Bowden3 Bencie Woll4 Andrew Zisserman1

1 Visual Geometry Group, University of Oxford, UK
2 LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France
3 CVSSP, University of Surrey, Guildford, UK
4 Deafness, Cognition and Language Research Centre, University College London, UK

https://www.robots.ox.ac.uk/˜vgg/research/signer_diarisation/

ABSTRACT

In this work, we propose a framework that enables collection of large-scale, diverse sign language datasets that can be used to train automatic sign language recognition models. The first contribution of this work is SDT RACK, a generic method for signer tracking and diarisation in the wild. Our second contribution is to show how SDT RACK can be used to automatically annotate 90 hours of British Sign Language (BSL) content featuring a wide range of signers, and including interviews, monologues and debates. Using SDT RACK, this data is annotated with 35K active signing tracks, with corresponding video-level signer identifiers and subtitles, and 40K automatically localised sign labels.

Index Terms— Signer Diarisation

1. INTRODUCTION

Sign languages represent the natural means of communication of deaf communities. They are visual languages with grammatical structures that differ considerably from those of spoken languages; there is no universal sign language and unrelated sign languages are mutually unintelligible [37]. While there has been considerable progress in machine comprehension of spoken languages in recent years [18], automatic recognition of sign languages remains challenging [23]. A key obstacle is the scarcity of appropriate training data: sign language datasets are typically (i) orders of magnitude smaller than their spoken counterparts [3], and (ii) collected in constrained settings (e.g. lab conditions, TV interpreters), that do not reflect the complexity and diversity of “real-life” sign language interactions. For example, multiple signers often engage in conversation [9], which introduces the additional challenge of identifying who is signing when.

In this work, we address this issue by proposing a method for automatically annotating a large-scale, diverse dataset of signing under a broad range of conditions to allow training of strong sign language recognition models. This requires identifying (i) temporal segments when signing occurs, and (ii) who is signing as there may be several people present. Our first contribution is therefore to propose SDT RACK, a generic framework for signer detection, tracking and diarisation in the wild (Sec. 3). Our second contribution is to employ SDT RACK, together with a recently proposed automatic annotation technique [1] to annotate TV broadcasts of the programme SeeHear, which is presented entirely in BSL. The show includes signing in diverse conditions (both indoors and outdoors in a variety of scenes) and contains a rich variety of interviews, monologues and debates. The signing content is accompanied by written English translations, obtained from broadcast subtitles. It was the first magazine programme series for the British deaf community, launched in 1981, and continues to be broadcast. The SeeHear data is annotated with active signing tracks, along with corresponding signer identities and subtitles, in addition to localised sign labels (Sec. 4). Finally, we evaluate performance for the tasks of signer diarisation and sign recognition using this data (Sec. 5).

2. RELATED WORK

Signer diarisation. Signer diarisation has received limited attention in the computational literature. Existing works have considered small-scale experiments (i.e. on four short videos in constrained conditions) using low-level motion cues [16,
Signer re-identification
Multi-person tracker
Active signer detector

Fig. 2. The SDTRACK framework for diarisation comprises three modules: (left) Multi-person tracker forms person tracks and estimates the pose of each person, (middle) Active signer detector determines which (if any) of the tracked persons are actively signing, (right) Signer re-identification associates identities to each tracked person. See Sec. 3 for more details.

Several methods have been proposed for active signer detection (which aims to estimate whether a signer is active in a given video frame): [30] proposes an efficient online system appropriate for videoconferencing software, while [2] proposes a multi-layer RNN model which performs frame-level active signer detection. To the best of our knowledge, no prior work has considered signer diarisation at large-scale in unconstrained environments.

Sign language datasets. A number of sign language datasets have been proposed in the literature (see [23] for a survey of the predominant benchmarks). Key shortcomings of existing datasets include: a lack of diversity (in terms of signing environment, number of signer identities, or both), restricted domain of discourse [5, 24] (for example, weather broadcasts) and limited scale [3]. MSASL [22], WLASL [25], BSL-DICT [29] and BSL SignBank [14] cover a wide vocabulary, but are restricted to isolated signs. BSLCORPUS [35] provides fine-grained linguistic annotation of conversations and narratives, but is limited to pairs of signers under lab conditions. BSL-1K [1] provides a large-scale dataset of continuous signs, but is derived entirely from crops of inset public broadcast interpreters (occupying a fixed region of the screen) and lacks scene diversity and interaction.

3. DIARISATION

Task formulation. Given an unconstrained video containing an arbitrary number of signers, our objective is to answer the question: “who is signing when?” by parsing the video into spatio-temporal tracks of active signers together with their identities. Our task therefore represents the combination of two techniques: (i) active signer detection which seeks to answer the question “which person is signing?”, and (ii) signer identification which addresses the question “who is this person?”

Concretely, given a sequence of video frames \(x_{1:T} = \{x_1, \ldots, x_T\} \) containing signing content, the objective is to identify active signers at each frame, \(x_t\), and to provide bounding boxes representing projections of each individuals’ 3D signing space on the image plane.

Proposed Framework: SDTRACK. Our approach, SD-TRACK (an abbreviation of Signer Diarisation and Tracking) comprises three modules: (1) a multi-person tracker, (2) an active signer detector, and (3) a signer re-identification model (see Fig. 2 for an overview). The role of each of these modules is described next.

Multi-Person Tracking. The objective of the tracker is to localise each person appearing in the footage and track their trajectories across frames. The tracker output consists of a collection of spatio-temporal bounding box trajectories \(T = \{\tau_i : \tau_i \in \mathbb{R}^{l_i \times 4}, i \in \{1, \ldots, N\}\} \), where \(N\) represents the total number of trajectories, \(l_i\) denotes the length of trajectory \(i\), and \(\tau_i\) specifies an array of bounding box locations in pixel coordinates.

Active Signer Detector. For each trajectory \(\tau_i\) delivered by the tracker, the role of the active signer detector is to produce a vector, \(y_i = \{0, 1\}^{l_i}\), which associates to each trajectory bounding box a label of 1 to indicate that the person contained within it is actively signing, and 0 otherwise.

Tracked person Re-identification (Re-ID) module. The role of this module is to group the trajectories \(T\) delivered by the tracker by their identity. To each track, \(\tau_i\), the model assigns an identity label \(\eta_i \in [0, \ldots, M_T]\), where \(M_T\) denotes the estimated total number of identities in \(T\).

Finally, the modules are combined by segmenting each person track \(\tau_i\) into the sequences of contiguous active signing frames indicated by \(y_i\) and assigning to each segment its corresponding track identity label \(\eta_i\).

Implementation. We instantiate the multi-person tracking module with the robust pose-based tracker recently proposed by [32]. In brief, this method employs a YOLOv3 [33] object detector, trained for person detection on the MS COCO dataset [26] to produce initial bounding box proposals at a fixed frame interval. Human keypoint estimation is then performed within the proposals over consecutive frames by recursively updating the bounding box such that it encloses the keypoints. These keypoints are fed to a two-layer Siamese graph convolutional network (which is trained to estimate the similarity of two poses via a learned metric in 128-dimensional feature space) that is used together with spatial consistency constraints to link box-enclosed poses into tracks.

While SDTRACK is agnostic to the implementation of the tracker, the pose-based tracking described above has the auxiliary benefit of estimating a set of human keypoints for each trajectory bounding box which can be efficiently re-used by the active signer detector module. We therefore implement the active signer detector as a binary classifier which takes as input the keypoint trajectories for a track and predicts its labels \(y_i\). We explore two variants of the active signer detector: (1) a simple frame-level decision tree heuristic which predicts an active signer whenever the estimated forearm-to-upper-arm angle is less than a fixed threshold (1.1 radians—determined through visual inspection on example data); (2) An eight-layer convolutional network which consumes temporal sequences of estimated keypoint confidences and lo-
ations (normalised with respect to their enclosing bounding box) and produces frame-level active signer labels. The latter model, implemented with the framework provided by [13], is “bootstrapped” by training on the predictions of the former across unlabelled data.

The Signer Re-ID module employs face verification for inter-track Re-ID (linking identities across tracks) and relies on the person tracker for intra-track Re-ID (identity is assumed consistent within a person track). Inter-track Re-ID comprises a three stage pipeline. (1) Faces are detected using the RetinaFace architecture [11], with Mobilenet0.25 [20] as the backbone network, trained on the WIDER FACE dataset [41]. (2) Face-identity discriminating embeddings are extracted from the face detections using an SENet-50 [21] architecture (pre-trained on the MS-Celeb-1M [19], and fine-tuned on VGGFace2 [6]). (3) As with the person tracker, face detections of the same identity in consecutive frames are linked together into face-tracks, using a combination of IoU (intersection over union) and similarity between face embeddings to link detections. For each face-track, the embeddings from the constituent detections are average-pooled across the track and L2-normalised, giving a single face-track embedding.

We evaluate these design choices and ablate their effect on diarisation performance in Sec. 5.

4. ANNOTATION PIPELINE

Data Description. The data that we annotate in this work, which we refer to as the SeeHear dataset, comprises 90 hours of British Sign Language (BSL) content produced by a wide variety of signers. It is sourced from public broadcast archive footage of the long-running UK BBC show, SeeHear. We annotate 35K active signer tracks and 40K automatically localised instances of specific signs for this data.

The partitions of the dataset are shown in Tab. 1. We note that in a number of episodes for the first several seasons of the show, presenters used “Sign-Supported English” (SSE)\(^3\), which differs from true BSL in terms of grammar and form of sign production [39, 40]. These data are excluded in this work from algorithmic performance evaluation but may nevertheless be of interest to the sign recognition community. We therefore estimate the presence of SSE in each signing track automatically with the method of [8].

The Train, Val, Test and Reserved partitions (summarised in Tab. 1) are formed by assigning the earliest 80 episodes (for which SSE is more prevalent) to Train, then randomly assigning the remaining episodes across splits in the given proportions.

Dataset Pipeline. To construct the SeeHear dataset, SDTRACK was employed as part of a multi-stage dataset pipeline, described next.

1. Source footage was obtained from 344 episodes of the TV show, SeeHear, from UK university library sources [38], enabling its use for non-commercial research.

2. OCR subtitle extraction. Subtitle text is first detected using the PixelLink algorithm [10, 27], and then recognised via the matching method of [42] trained on synthetic text lines. For each frame sequence containing a subtitle text, the text recognised on each frame is then analysed with an English transformer language model [31] trained with the Fairseq sequence modelling toolkit. The text with the lowest perplexity is chosen as the subtitle for the whole frame sequence.

3. Signer tracking and diarisation. Using the SDTRACK framework, tracks were obtained from all episodes and divided into active signer segments. Each active signer is associated with the corresponding subtitles produced by the previous stage.

4. Annotations from mouthings. Within the active signer segments produced by SDTRACK, we apply the sign spotting method proposed by [1] using the improved visual-only keyword spotting model of Stafylakis et al. [36] from [28] (referred to in their paper as “P2G [36] baseline”). To generate the list of candidate signs for spotting, the subtitles are first text-normalised using the method of [15] (this converts non-standard words such as numbers and date abbreviations to a canonical written form, e.g. “35” is mapped to “thirty five”). Signing tracks are then queried with each word present in the normalised subtitles that possesses four or more phonemes.

Following [1], sparse annotations are retained for all spotting predictions which are assigned a posterior probability of 0.5 or greater by the visual keyword spotting model. These localised sign annotations allow us to use the SeeHear data to assess sign recognition performance.

The outputs of this fully automatic pipeline, which comprise SeeHear with active signing tracks along with subtitle sentences and sparse sign annotations, are summarised in Tab. 1. In addition to the above, the dataset also includes a set of 4,478 sparse annotations (distributed across all dataset partitions) that have been manually verified as correct by annotators (described in more detail in Sec. 5).

5. EXPERIMENTS

In this section, we report on using ground truth annotations to evaluate components of the SDTRACK framework, validate the automatic sparse annotations, and assess automatic sign
We next provide an assessment of the current active signer in the scene (to whom the content of the subtitles corresponds), highlighting the room for further improvement of the latter.

**Automatic sign recognition.** As described in Sec. 4, besides diarisation, SEEHEAR can be used for sign language recognition. Here, we provide results for classifying signs given short temporal intervals around sparse annotations within active signer tracks. Specifically, the input is a 20-frame video cropped around the active signer. We determine a vocabulary of 4321 signs from the train split, removing words for which all mouthing annotation confidences are below 0.8. For the training set (combining Train+SSE splits), we use 36K automatic mouthing annotations. For testing, we use the 307 manually verified signs which fall within the test partition spanning a vocabulary of 124 signs. Tab. 3 reports the recognition performance for several 13D [7] models with or without fine-tuning on the SEEHEAR training set. First, we apply the state-of-the-art model of [1], which is trained for 1064 sign categories on BSL-1K, covering our test vocabulary. We obtain 21.2% accuracy, which highlights the difficulty of generalisation to in-the-wild SEEHEAR data. Second, we finetune this model for the more challenging task of 4321-way classification and observe significant gains (67.4%). Third, similar to [1], training with a less noisy but smaller set of annotations by filtering the mouthing confidence at 0.7 (SEEHEAR_{m,7}) degrades the performance (63.5%). Finally, sign recognition pretraining on BSL-1K gives a significant boost over action recognition pretraining with Kinetics (52.4% vs 67.4%).

<table>
<thead>
<tr>
<th>Pretraining</th>
<th>Finetuning</th>
<th>per-instance</th>
<th>per-class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>top-1</td>
<td>top-5</td>
</tr>
<tr>
<td>BSL-1K [1]</td>
<td>-</td>
<td>21.2</td>
<td>38.4</td>
</tr>
<tr>
<td>Kinetics</td>
<td>SEEHEAR_{m,5}</td>
<td>52.4</td>
<td>68.7</td>
</tr>
<tr>
<td>BSL-1K</td>
<td>SEEHEAR_{m,7}</td>
<td>63.5</td>
<td>80.1</td>
</tr>
<tr>
<td>BSL-1K</td>
<td>SEEHEAR_{m,5}</td>
<td>67.4</td>
<td>82.1</td>
</tr>
</tbody>
</table>

**6. CONCLUSION**

In this work, we introduced the SDTRACK framework for signer tracking and diarisation in the wild. We used SDTRACK to annotate a large-scale collection of BSL signing content in diverse conditions. We also used the annotated data to evaluate several diarisation and recognition baselines to assess their effectiveness for sign language understanding in the wild.

**Acknowledgements.** This work was supported by EPSRC grant ExTol. The authors would like to thank Himel Chowdhury for his assistance with sign annotation.
References


