BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues

Samuel Albanie[1⋆], Gül Varol[1⋆], Liliane Momeni[1], Triantafyllos Afouras[1], Joon Son Chung[1,2], Neil Fox[3], and Andrew Zisserman[1]

[1]Visual Geometry Group, University of Oxford, UK
[2]Naver Corporation, Seoul, South Korea
[3]Deafness, Cognition and Language Research Centre, University College London, UK
{albanie,gul,liliane,afourast,joon,az}@robots.ox.ac.uk;
neil.fox@ucl.ac.uk

Abstract. Recent progress in fine-grained gesture and action classification, and machine translation, point to the possibility of automated sign language recognition becoming a reality. A key stumbling block in making progress towards this goal is a lack of appropriate training data, stemming from the high complexity of sign annotation and a limited supply of qualified annotators. In this work, we introduce a new scalable approach to data collection for sign recognition in continuous videos. We make use of weakly-aligned subtitles for broadcast footage together with a keyword spotting method to automatically localise sign-instances for a vocabulary of 1,000 signs in 1,000 hours of video. We make the following contributions: (1) We show how to use mouthing cues from signers to obtain high-quality annotations from video data—the result is the BSL-1K dataset, a collection of British Sign Language (BSL) signs of unprecedented scale; (2) We show that we can use BSL-1K to train strong sign recognition models for co-articulated signs in BSL and that these models additionally form excellent pretraining for other sign languages and benchmarks—we exceed the state of the art on both the MSASL and WLASL benchmarks. Finally, (3) we propose new large-scale evaluation sets for the tasks of *sign recognition* and *sign spotting* and provide baselines which we hope will serve to stimulate research in this area.

Keywords: Sign Language Recognition, Visual Keyword Spotting

1 Introduction

With the continual increase in the performance of human action recognition there has been a renewed interest in the challenge of recognising sign languages such as American Sign Language (ASL), British Sign Language (BSL), and Chinese Sign Language (CSL). Although in the past isolated sign recognition has seen some progress, recognition of continuous sign language remains extremely challenging [10]. Isolated signs, as in dictionary examples, do not suffer from the

⋆ Equal contribution

naturally occurring complication of co-articulation (i.e. transition motions) between preceding and subsequent signs, making them visually very different from continuous signing. If we are to recognise ASL and BSL performed *naturally* by signers, then we need to recognise co-articulated signs.

Similar problems were faced by Automatic Speech Recognition (ASR) and the solution, as always, was to learn from very large scale datasets, using a parallel corpus of speech and text. In the vision community, a related path was taken with the modern development of automatic lip reading: first isolated words were recognised [16], and later sentences were recognised [15]—in both cases tied to the release of large datasets. The objective of this paper is to design a scalable *method* to generate large-scale datasets of continuous signing, for training and testing sign language recognition, and we demonstrate this for BSL. We start from the perhaps counter-intuitive observation that signers often mouth the word they sign simultaneously, as an additional signal [5, 53, 54], performing similar lip movements as for the spoken word. This differs from mouth gestures which are not derived from the spoken language [21]. The mouthing helps disambiguate between different meanings of the same manual sign [60] or in some cases simply provides redundancy. In this way, a sign is not only defined by the hand movements and hand shapes, but also by facial expressions and mouth movements [20].

We harness word mouthings to provide a method of automatically annotating continuous signing. The key idea is to exploit the readily available and abundant supply of sign-language translated TV broadcasts that consist of an overlaid interpreter performing signs and subtitles that correspond to the audio content. The availability of subtitles means that the annotation task is in essence one of alignment between the words in the subtitle and the mouthings of the overlaid signer. Nevertheless, this is a *very* challenging task: a continuous sign may last for only a fraction (e.g. 0.5) of a second, whilst the subtitles may last for several seconds and are not synchronised with the signs produced by the signer; the word order of the English need not be the same as the word order of the sign language; the sign may not be mouthed; and furthermore, words may not be signed or may be signed in different ways depending on the context. For example, the word "fish" has a different visual sign depending on referring to the animal or the food, introducing additional challenges when associating subtitle words to signs.

To detect the mouthings we use *visual keyword spotting*—the task of determining *whether* and *when* a keyword of interest is uttered by a talking face using *only* visual information—to address the alignment problem described above. Two factors motivate its use: (1) direct lip reading of arbitrary isolated mouthings is a fundamentally difficult task, but searching for a particular known word within a short temporal window is considerably less challenging; (2) the recent availability of large scale video datasets with aligned audio transcriptions [1, 17] now allows for the training of powerful visual keyword spotting models [32, 51, 62] that, as we show in the experiments, work well for this application.

We make the following contributions: (1) we show how to use visual keyword spotting to recognise the mouthing cues from signers to obtain high-quality

Table 1. Summary of previous public sign language datasets: The BSL-1K dataset contains, to the best of our knowledge, the largest source of annotated sign data in any dataset. It comprises of co-articulated signs outside a lab setting.

Dataset	lang	co-articulated	#signs	#annos (avg. per sign)	#signers	source
ASLLVD [4]	ASL	✗	2742	9K (3)	6	lab
Devisign [14]	CSL	✗	2000	24K (12)	8	lab
MSASL [33]	ASL	✗	1000	25K (25)	222	lexicons, web
WLASL [39]	ASL	✗	2000	21K (11)	119	lexicons, web
S-pot [57]	FinSL	✓	1211	4K (3)	5	lab
Purdue RVL-SLLL [59]	ASL	✓	104	2K (19)	14	lab
Video-based CSL [31]	CSL	✓	178	25K (140)	50	lab
SIGNUM [58]	DGS	✓	455	3K (7)	25	lab
RWTH-Phoenix [10, 34]	DGS	✓	1081	65K (60)	9	TV
BSL Corpus [50]	BSL	✓	5K	50K (10)	249	lab
BSL-1K	BSL	✓	1064	273K (257)	40	TV

annotations from video data—the result is the BSL-1K dataset, a large-scale collection of BSL (British Sign Language) signs with a 1K sign vocabulary; (2) We show the value of BSL-1K by using it to train strong sign recognition models for co-articulated signs in BSL and demonstrate that these models additionally form excellent pretraining for other sign languages and benchmarks—we exceed the state of the art on both the MSASL and WLASL benchmarks with this approach; (3) We propose new evaluation datasets for *sign recognition* and *sign spotting* and provide baselines for each of these tasks to provide a foundation for future research[1].

2 Related Work

Sign language datasets. We begin by briefly reviewing public benchmarks for studying automatic sign language recognition. Several benchmarks have been proposed for American [4, 33, 39, 59], German [34, 58], Chinese [14, 31], and Finnish [57] sign languages. BSL datasets, on the other hand, are scarce. One exception is the ongoing development of the linguistic corpus [49, 50] which provides fine-grained annotations for the atomic elements of sign production. Whilst its high annotation quality provides an excellent resource for sign linguists, the annotations span only a fraction of the source videos so it is less appropriate for training current state-of-the-art data-hungry computer vision pipelines.

Tab. 1 presents an overview of publicly available datasets, grouped according to their provision of *isolated* signs or *co-articulated* signs. Earlier datasets have been limited in the size of their video instances, vocabularies, and signers. Within the isolated sign datasets, Purdue RVL-SLLL [59] has a limited vocabulary of 104 signs (ASL comprises more than 3K signs in total [56]). ASLLVD [4] has only 6 signers. Recently, MSASL [33] and WLASL [39] large-vocabulary isolated

[1] The project page is at: https://www.robots.ox.ac.uk/~vgg/research/bsl1k/

sign datasets have been released with 1K and 2K signs, respectively. The videos are collected from lexicon databases and other instructional videos on the web.

Due to the difficulty of annotating co-articulated signs in long videos, continuous datasets have been limited in their vocabulary, and most of them have been recorded in lab settings [31, 58, 59]. RWTH-Phoenix [34] is one of the few realistic datasets that supports training complex models based on deep neural networks. A recent extension also allows studying sign language translation [10]. However, the videos in [10, 34] are only from weather broadcasts, restricting the domain of discourse. In summary, the main constraints of the previous datasets are one or more of the following: (i) they are limited in size, (ii) they have a large total vocabulary but only of isolated signs, or (iii) they consist of natural co-articulated signs but cover a limited domain of discourse. The BSL-1K dataset provides a considerably greater number of annotations than all previous public sign language datasets, and it does so in the co-articulated setting for a large domain of discourse.

Sign language recognition. Early work on sign language recognition focused on hand-crafted features computed for hand shape and motion [24, 25, 52, 55]. Upper body and hand pose have then been widely used as part of the recognition pipelines [7, 9, 19, 46, 48]. Non-manual features such as face [24, 34, 45], and mouth [3, 35, 37] shapes are relatively less considered. For sequence modelling of signs, HMMs [2, 23, 27, 52], and more recently LSTMs [9, 31, 63, 64], have been utilised. Koller et al. [38] present a hybrid approach based on CNN-RNN-HMM to iteratively re-align sign language videos to the sequence of sign annotations. More recently 3D CNNs have been adopted due to their representation capacity for spatio-temporal data [6, 8, 30, 33, 39]. Two recent concurrent works [33, 39] showed that I3D models [13] significantly outperform their pose-based counterparts. In this paper, we confirm the success of I3D models, while also showing improvements using pose distillation as pretraining. There have been efforts to use sequence-to-sequence translation models for sign language translation [10], though this has been limited to the weather discourse of RWTH-Phoenix, and the method is limited by the size of the training set. The recent work of [40] localises signs in continuous news footage to improve an isolated sign classifier.

In this work, we utilise mouthings to localise signs in weakly-supervised videos. Previous work [7, 17, 18, 48] has used weakly aligned subtitles as a source of training data, and both one-shot [48] (from a visual dictionary) and zero-shot [6] (from a textual description) have also been used. Though no previous work, to our knowledge, has put these ideas together. The sign spotting problem was formulated in [22, 57].

Using the mouth patterns. The mouth has several roles in sign language that can be grouped into spoken components (mouthings) and oral components (mouth gestures) [60]. Several works focus on recognising mouth shapes [3, 37] to recover mouth gestures. Few works [35, 36] attempt to recognise mouthings in sign language data by focusing on a few categories of visemes, i.e., visual correspondences of phonemes in the lip region [26]. Most closely related to our work, [47] similarly searches subtitles of broadcast footage and uses the mouth as a cue to improve alignment between the subtitles and the signing. Two key

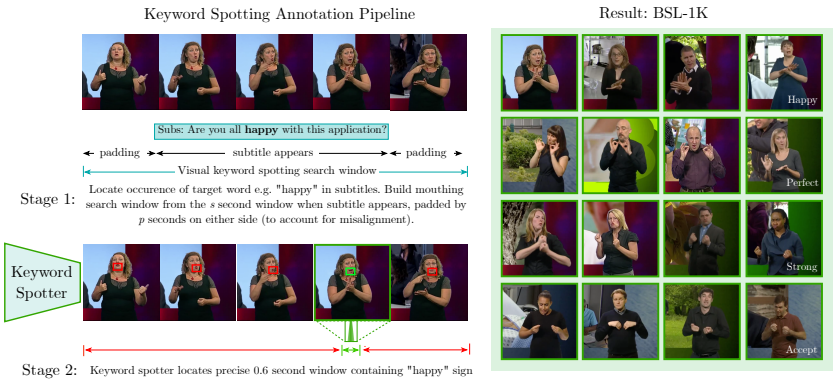

Fig. 1. Keyword-driven sign annotation: (Left, the annotation pipeline): Stage 1: for a given target sign (e.g. "happy") each occurrence of the word in the subtitles provides a candidate temporal window when the sign may occur (this is further padded by several seconds on either side to account for misalignment of subtitles and signs); Stage 2: a keyword spotter uses the mouthing of the signer to perform precise localisation of the sign within this window. (Right): Examples from the BSL-1K dataset—produced by applying keyword spotting for a vocabulary of 1K words.

differences between our work and theirs are: (1) we achieve precise localisation through keyword spotting, whereas they only use an open/closed mouth classifier to reduce the number of candidates for a given sign; (2) scale—we gather signs over 1,000 hours of signing (in contrast to the 30 hours considered in [47]).

3 Learning Sign Recognition with Automatic Labels

In this section, we describe the process used to collect BSL-1K, a large-scale dataset of BSL signs. An overview of the approach is provided in Fig. 1. In Sec. 3.1, we describe how large numbers of video clips that are likely to contain a given sign are sourced from public broadcast footage using subtitles; in Sec. 3.2, we show how automatic keyword spotting can be used to precisely localise specific signs to within a fraction of a second; in Sec. 3.3, we apply this technique to efficiently annotate a large-scale dataset with a vocabulary of 1K signs.

3.1 Finding probable signing windows in public broadcast footage

The source material for the dataset comprises 1,412 episodes of publicly broadcast TV programs produced by the BBC which contains 1,060 hours of continuous BSL signing. The episodes cover a wide range of topics: medical dramas, history and nature documentaries, cooking shows and programs covering gardening, business and travel. The signing represents a translation (rather than a transcription) of the content and is produced by a total of forty professional BSL interpreters. The signer occupies a fixed region of the screen and is cropped

directly from the footage. A full list of the TV shows that form BSL-1K can be found in the appendix. In addition to videos, these episodes are accompanied by subtitles (numbering approximately 9.5 million words in total). To locate temporal windows in which instances of signs are likely to occur within the source footage, we first identify a candidate list of words that: (i) are present in the subtitles; (ii) have entries in both BSL signbank[2] and sign BSL[3], two online dictionaries of isolated signs (to ensure that we query words that have valid mappings to signs). The result is an initial vocabulary of 1,350 words, which are used as queries for the keyword spotting model to perform sign localisation—this process is described next.

3.2 Precise sign localisation through visual keyword spotting

By searching the content of the subtitle tracks for instances of words in the initial vocabulary, we obtain a set of candidate temporal windows in which instances of signs may occur. However, two factors render these temporal proposals extremely noisy: (1) the presence of a word in the subtitles does not guarantee its presence in the signing; (2) even for subtitled words that are signed, we find through inspection that their appearance in the subtitles can be misaligned with the sign itself by several seconds.

To address this challenge, we turn to *visual keyword spotting*. Our goal is to detect and precisely localise the presence of a sign by identifying its "spoken components" [54] within a temporal sequence of mouthing patterns. Two hypotheses underpin this approach: (a) that mouthing provides a strong localisation signal for signs as they are produced; (b) that this mouthing occurs with sufficient frequency to form a useful localisation cue. Our method is motivated by studies in the Sign Linguistics literature which find that spoken components frequently serve to identify signs—this occurs most prominently when the mouth pattern is used to distinguish between manual homonyms[4] (see [54] for a detailed discussion). However, even if these hypotheses hold, the task remains extremely challenging—signers typically do not mouth continuously and the mouthings that are produced may only correspond to a portion of the word [54]. For this reason, existing lip reading approaches cannot be used directly (indeed, an initial exploratory experiment we conducted with the state-of-the-art lip reading model of [1] achieved zero recall on five-hundred randomly sampled sentences of signer mouthings from the BBC source footage).

The key to the effectiveness of visual keyword spotting is that rather than solving the general problem of lip reading, it solves the much easier problem of identifying a single token from a small collection of candidates within a short temporal window. In this work, we use the subtitles to construct such windows. The pipeline for automatic sign annotations therefore consists of two stages (Fig. 1, left): (1) For a given target sign e.g. "happy", determine the times of

[2] https://bslsignbank.ucl.ac.uk/

[3] https://www.signbsl.com/

[4] These are signs that use identical hand movements (e.g. "king" and "queen") whose meanings are distinguished by mouthings.

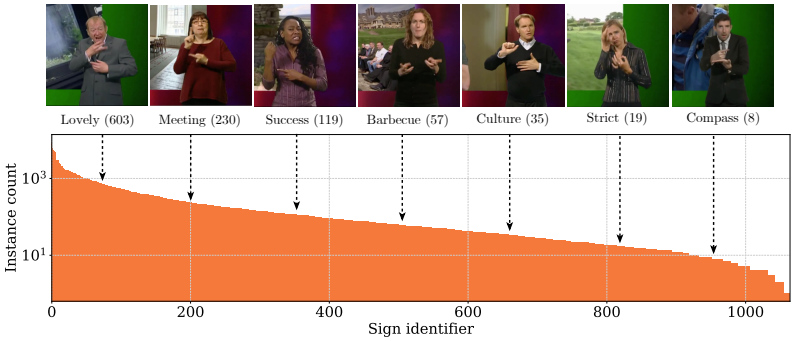

Fig. 2. BSL-1K sign frequencies: Log-histogram of instance counts for the 1,064 words constituting the BSL-1K vocabulary, together with example signs. The long-tail distribution reflects the *real* setting in which some signs are more frequent than others.

all occurrences of this sign in the subtitles accompanying the video footage. The subtitle time provides a short window during which the word was spoken, but not necessarily when its corresponding sign is produced in the translation. We therefore extend this candidate window by several seconds to increase the likelihood that the sign is present in the sequence. We include ablations to assess the influence of this padding process in Sec. 5 and determine empirically that padding by four seconds on each side of the subtitle represents a good choice. (2) The resulting temporal window is then provided, together with the target word, to a keyword spotting model (described in detail in Sec. 4.1) which estimates the probability that the sign was mouthed at each time step (we apply the keyword spotter with a stride of 0.04 seconds—this choice is motivated by the fact that the source footage has a frame rate of 25fps). When the keyword spotter asserts with high confidence that it has located a sign, we take the location of the peak posterior probability as an anchoring point for one endpoint of a 0.6 second window (this value was determined by visual inspection to be sufficient for capturing individual signs). The peak probability is then converted into a decision about whether a sign is present using a threshold parameter. To build the BSL-1K dataset, we select a value of 0.5 for this parameter after conducting experiments (reported in Tab. 3) to assess its influence on the downstream task of sign recognition performance.

3.3 BSL-1K dataset construction and validation

Following the sign localisation process described above, we obtain approximately 280k localised signs from a set of 2.4 million candidate subtitles. To ensure that the dataset supports study of signer-independent sign recognition, we then compute face embeddings (using an SENet-50 [29] architecture trained for verification on the VGGFace2 dataset [11]) to group the episodes according to which of the forty signers they were translated by. We partition the data into three splits, assigning thirty-two signers for training, four signers for validation and

Table 2. Statistics of the proposed BSL-1K dataset: The *Test-(manually verified)* split represents a sample from the Test-(automatic) split annotations that have been verified by human annotators (see Sec. 3.3 for details).

Set	sign vocabulary	sign annotations	number of signers
Train	1,064	173K	32
Val	1,049	36K	4
Test-(automatic)	1,059	63K	4
Test-(manually verified)	334	2103	4

four signers for testing. We further sought to include an equal number of hearing and non-hearing signers (the validation and test sets both contain an equal number of each, the training set is approximately balanced with 13 hearing, 17 non-hearing and 2 signers whose deafness is unknown). We then perform a further filtering step on the vocabulary to ensure that each word included in the dataset is represented with high confidence (at least one instance with confidence 0.8) in the training partition, which produces a final dataset vocabulary of 1,064 words (see Fig. 2 for the distribution and the appendix for the full word list).

Validating the automatic annotation pipeline. One of the key hypotheses underpinning this work is that keyword spotting is capable of correctly locating signs. We first verify this hypothesis by presenting a randomly sampled subset of the test partition to a native BSL signer, who was asked to assess whether the short temporal windows produced by the keyword spotting model with high confidence (each 0.6 seconds in duration) contained correct instances of the target sign. A screenshot of the annotation tool developed for this task is provided in the appendix. A total of 1k signs were included in this initial assessment, of which 70% were marked as correct, 28% were marked as incorrect and 2% were marked as uncertain, validating the key idea behind the annotation pipeline. Possible reasons for incorrect marks include: BSL mouthing patterns are not always identical to spoken English and mouthings many times do not represent the full word (e.g., "fsh" for "finish") [54].

Constructing a manually verified test set. To construct a high quality, human verified test set and to maximise yield from the annotators, we started from a collection of sign predictions where the keyword model was highly confident (assigning a peak probability of greater than 0.9) yielding 5,826 sign predictions. Then, in addition to the validated 980 signs (corrections were provided as labels for the signs marked as incorrect and uncertain signs were removed), we further expanded the verified test set with non-native (BSL level 2 or above) signers who annotated a further 2k signs. We found that signers with lower levels of fluency were able to confidently assert that a sign was correct for a portion of the signs (at a rate of around 60%), but also annotated a large number of signs as "unsure", making it challenging to use these annotations as part of the validation test for the effectiveness of the pipeline. Only signs marked as correct were included into the final verified test set, which ultimately comprised 2,103 annotations covering 334 signs from the 1,064 sign vocabulary. The statistics of each partition of the dataset are provided in Tab. 2. All experimental test set

results in this paper refer to performance on the verified test set (but we retain the full automatic test set, which we found to be useful for development).

In addition to the keyword spotting approach described above, we explore techniques for further dataset expansion based on other cues in the appendix.

4 Models and Implementation Details

In this section, we first describe the visual keyword spotting model used to collect signs from mouthings (Sec. 4.1). Next, we provide details of the model architecture for sign recognition and spotting (Sec. 4.2). Lastly, we describe a method for obtaining a good initialisation for the sign recognition model (Sec. 4.3).

4.1 Visual keyword spotting model

We use the improved visual-only keyword spotting model of Stafylakis et al. [51] from [44] (referred to in their paper as "P2G [51] baseline"), provided by the authors. The model of [51] combines visual features with a fixed-length keyword embedding to determine whether a user-defined keyword is present in an input video clip. The performance of [51] is improved in [44] by switching the keyword encoder-decoder from grapheme-to-phoneme (G2P) to phoneme-to-grapheme (P2G).

In more detail, the model consists of four stages: (i) visual features are first extracted from the sequence of face-cropped image frames from a clip (this is performed using a 512×512 SSD architecture [42] trained for face detection on WIDER faces [61]), (ii) a fixed-length keyword representation is built using a P2G encoder-decoder, (iii) the visual and keyword embeddings are concatenated and passed through BiLSTMs, (iv) finally, a sigmoid activation is applied on the output to approximate the posterior probability that the keyword occurs in the video clip for each input frame. If the maximum posterior over all frames is greater than a threshold, the clip is predicted to contain the keyword. The predicted location of the keyword is the position of the maximum posterior. Finally, non-maximum suppression is run with a temporal window of 0.6 seconds over the untrimmed source videos to remove duplicates.

4.2 Sign recognition model

We employ a spatio-temporal convolutional neural network architecture that takes a multiple-frame video as input, and outputs class probabilities over sign categories. Specifically, we follow the I3D architecture [13] due to its success on action recognition benchmarks, as well as its recently observed success on sign recognition datasets [33, 39]. To retain computational efficiency, we only use an RGB stream. The model is trained on 16-frame consecutive frames (i.e., 0.64 sec for 25fps), as [7, 47, 57] observed that co-articulated signs last roughly for 13 frames. We resize our videos to have a spatial resolution of 224×224. For training, we randomly subsample a fixed-size, temporally contiguous input from the spatio-temporal volume to have $16 \times 224 \times 224$ resolution in terms of number of

frames, width, and height, respectively. We minimise the cross-entropy loss using SGD with momentum (0.9) with mini-batches of size 4, and an initial learning rate of 10^{-2} with a fixed schedule. The learning rate is decreased twice with a factor of 10^{-1} at epochs 20 and 40. We train for 50 epochs. Colour, scale, and horizontal flip augmentations are applied on the input video. When pretraining is used (e.g. on Kinetics-400 [13] or on other data where specified), we replace the last linear layer with the dimensionality of our classes, and fine-tune all network parameters (we observed that freezing part of the model is suboptimal). Finally, we apply dropout on the classification layer with a probability of 0.5.

At test time, we perform centre-cropping and apply a sliding window with a stride of 8 frames before averaging the classification scores to obtain a video-level prediction.

4.3 Video pose distillation

Given the significant focus on pose estimation in the sign language recognition literature, we investigate how explicit pose modelling can be used to improve the I3D model. To this end, we define a *pose distillation* network that takes in a sequence of 16 consecutive frames, but rather than predicting sign categories, the 1024-dimensional (following average pooling) embedding produced by the network is used to regress the poses of individuals appearing in each of the frames of its input. In more detail, we assume a single individual per-frame (as is the case in cropped sign translation footage) and task the network with predicting 130 human pose keypoints (18 body, 21 per hand, and 70 facial) produced by an OpenPose [12] model (trained on COCO [41]) that is evaluated per-frame. The key idea is that, in order to effectively predict pose across multiple frames from a single video embedding, the model is encouraged to encode information not only about pose, but also descriptions of relevant dynamic gestures. The model is trained on a portion of the BSL-1K training set (due to space constraints, further details of the model architecture and training procedure are provided in the appendix).

5 Experiments

We first provide several ablations on our sign recognition model to answer questions such as which cues are important, and how to best use human pose. Then, we present baseline results for sign recognition and sign spotting, with our best model. Finally, we compare to the state of the art on ASL benchmarks to illustrate the benefits of pretraining on our data.

5.1 Ablations for the sign recognition model

In this section, we evaluate our sign language recognition approach and investigate (i) the effect of mouthing score threshold, (ii) the comparison to pose-based approaches, (iii) the contribution of multi-modal cues, and (iv) the video pose distillation. Additional ablations about the influence of the temporal extent of

Table 3. Trade-off between training noise vs. size: Training (with Kinetics initialisation) on the full training set BSL-1K$_{m.5}$ versus the subset BSL-1K$_{m.8}$, which correspond to a mouthing score threshold of 0.5 and 0.8, respectively. Even when noisy, with the 0.5 threshold, mouthings provide automatic annotations that allow supervised training at scale, resulting in 70.61% accuracy on the manually validated test set.

Training data	#videos	per-instance top-1	per-instance top-5	per-class top-1	per-class top-5
BSL-1K$_{m.8}$ (mouthing≥0.8)	39K	69.00	83.79	45.86	64.42
BSL-1K$_{m.5}$ (mouthing≥0.5)	173K	**70.61**	**85.26**	**47.47**	**68.13**

the automatic annotations and the search window size for the keyword spotting can be found in the appendix.

Evaluation metrics. Following [33, 39], we report both top-1 and top-5 classification accuracy, mainly due to ambiguities in signs which can be resolved in context. Furthermore, we adopt both per-instance and per-class accuracy metrics. Per-instance accuracy is computed over all test instances. Per-class accuracy refers to the average over the sign categories present in the test set. We use this metric due to the unbalanced nature of the datasets.

The effect of the mouthing score threshold. The keyword spotting method, being a binary classification model, provides a confidence score, which we threshold to obtain our automatically annotated video clips. Reducing this threshold yields an increased number of sign instances at the cost of a potentially noisier set of annotations. We denote the training set defined by a mouthing threshold 0.8 as BSL-1K$_{m.8}$. In Tab. 3, we show the effect of changing this hyper-parameter between a low- and high-confidence model with 0.5 and 0.8 mouthing thresholds, respectively. The larger set of training samples obtained with a threshold of 0.5 provide the best performance. For the remaining ablations, we use the smaller BSL-1K$_{m.8}$ training set for faster iterations, and return to the larger BSL-1K$_{m.5}$ set for training the final model.

Pose-based model versus I3D. We next verify that I3D is a suitable model for sign language recognition by comparing it to a pose-based approach. We implement Pose→Sign, which follows a 2D ResNet architecture [28] that operates on $3 \times 16 \times P$ dimensional dynamic pose images, where P is the number of keypoints. In our experiments, we use OpenPose [12] (pretrained on COCO [41]) to extract 18 body, 21 per hand, and 70 facial keypoints. We use 16-frame inputs to make it comparable to the I3D counterpart. We concatenate the estimated normalised xy coordinates of a keypoint with its confidence score to obtain the 3 channels. In Tab. 4, we see that I3D significantly outperforms the explicit 2D pose-based method (65.57% vs 49.66% per-instance accuracy). This conclusion is in accordance with the recent findings of [33, 39].

Contribution of individual cues. We carry out two set of experiments to determine how much our sign recognition model relies on signals from the mouth and face region versus the manual features from the body and hands: (i) using Pose→Sign, which takes as input the 2D keypoint locations over several frames, (ii) using I3D, which takes as input raw video frames. For the pose-based model,

Table 4. Contribution of individual cues: We compare I3D (pretrained on Kinetics) with a keypoint-based baseline both trained and evaluated on a subset of BSL-1K$_{m.8}$, where we have the pose estimates. We also quantify the contribution of the body&hands and the face regions. We see that significant information can be attributed to both types of cues, and the combination performs the best.

			per-instance		per-class	
	body&hands	face	top-1	top-5	top-1	top-5
Pose→Sign (70 points)	✗	✓	24.41	47.59	9.74	25.99
Pose→Sign (60 points)	✓	✗	40.47	59.45	20.24	39.27
Pose→Sign (130 points)	✓	✓	**49.66**	**68.02**	**29.91**	**49.21**
I3D (face-crop)	✗	✓	42.23	69.70	21.66	50.51
I3D (mouth-masked)	✓	✗	46.75	66.34	25.85	48.02
I3D (full-frame)	✓	✓	**65.57**	**81.33**	**44.90**	**64.91**

Table 5. Effect of pretraining the I3D model on various tasks before fine-tuning for sign recognition on BSL-1K$_{m.8}$. Our dynamic pose features learned on 16-frame videos provide body-motion-aware cues and outperform other pretraining strategies.

Pretraining		per-instance		per-class	
Task	Data	top-1	top-5	top-1	top-5
Random init.	-	39.80	61.01	15.76	29.87
Gesture recognition	Jester [43]	46.93	65.95	19.59	36.44
Sign recognition	WLASL [39]	69.90	83.45	44.97	62.73
Action recognition	Kinetics [13]	69.00	83.79	45.86	64.42
Video pose distillation	Signers	**70.38**	**84.50**	**46.24**	**65.31**

we train with only 70 facial keypoints, 60 body&hand keypoints, or with the combination. For I3D, we use the pose estimations to mask the pixels outside of the face bounding box, to mask the mouth region, or use all the pixels from the videos. The results are summarised in Tab. 4. We observe that using only the face provides a strong baseline, suggesting that mouthing is a strong cue for recognising signs, e.g., 42.23% for I3D. However, using all the cues, including body and hands (65.57%), significantly outperforms using individual modalities.

Pretraining for sign recognition. Next we investigate different forms of pretraining for the I3D model. In Tab. 5, we compare the performance of a model trained with random initialisation (39.80%), fine-tuning from gesture recognition (46.93%), sign recognition (69.90%), and action recognition (69.00%). Video pose distillation provides a small boost over the other pretraining strategies (70.38%), suggesting that it is an effective way to force the I3D model to pay attention to the dynamics of the human keypoints, which is relevant for sign recognition.

5.2 Benchmarking sign recognition and sign spotting

Next, we combine the parameter choices suggested by each of our ablations to establish baseline performances on the BSL-1K dataset for two tasks: (i) sign recognition, (ii) sign spotting. Specifically, the model comprises an I3D architecture trained on BSL-1K$_{m.5}$ with pose-distillation as initialisation and random temporal offsets of up to 4 frames around the sign during training (the ablation studies for this temporal augmentation parameter are included in the appendix).

Fig. 3. Qualitative analysis: We present results of our sign recognition model on BSL-1K for success (top) and failure (bottom) cases, together with their confidence scores in parentheses. To the right of each example, we show a random training sample for the predicted sign (in small). We observe that failure modes are commonly due to high visual similarity in the gesture (bottom-left) and mouthing (bottom-right).

Table 6. Benchmarking: We benchmark our best sign recognition model (trained on BSL-1K$_{m.5}$, initialised with pose distillation, with 4-frame temporal offsets) for sign recognition and sign spotting task to establish strong baselines on BSL-1K.

	per-instance		per-class	
	top-1	top-5	top-1	top-5
Sign Recognition	75.51	88.83	52.76	72.14

	mAP (334 sign classes)
Sign Spotting	0.159

The sign recognition evaluation protocol follows the experiments conducted in the ablations, the sign spotting protocol is described next.

Sign spotting. Differently from sign recognition, in which the objective is to classify a pre-defined temporal segment into a category from a given vocabulary, *sign spotting* aims to locate all instances of a particular sign within long sequences of untrimmed footage, enabling applications such as content-based search and efficient indexing of signing videos for which subtitles are not available. The evaluation protocol for assessing sign spotting on BSL-1K is defined as follows: for each sign category present amongst the human-verified test set annotations (334 in total), windows of 0.6-second centred on each verified instance are marked as positive and all other times within the subset of episodes that contain at least one instance of the sign are marked as negative. To avoid false penalties at signs that were not discovered by the automatic annotation process, we exclude windows of 8 seconds of footage centred at each location in the original footage at which the target keyword appears in the subtitles, but was not detected by the visual keyword spotting pipeline. In aggregate this corresponds to locating approximately one positive instance of a sign in every 1.5 hours of continuous signing negatives. A sign is considered to have been correctly spotted if its temporal overlap with the model prediction exceeds an IoU (intersection-over-union) of 0.5, and we report mean Average Precision (mAP) over the 334 sign classes as the metric for performance.

Table 7. Transfer to ASL: Performance on American Sign Language (ASL) datasets with and without pretraining on our data. I3D results are reported from the original papers for MSASL [33] and WLASL [39]. I3D† denotes our implementation and training, adopting the hyper-parameters from [33]. We show that our features provide good initialisation, even if it is trained on BSL.

		WLASL [39]				MSASL [33]			
		per-instance		per-class		per-instance		per-class	
	pretraining	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
I3D [33]	Kinetics	-	-	-	-	-	-	57.69	81.08
I3D [39]	Kinetics	32.48	57.31	-	-	-	-	-	-
I3D†	Kinetics	40.85	74.10	39.06	73.33	60.45	82.05	57.17	80.02
I3D	BSL-1K	**46.82**	**79.36**	**44.72**	**78.47**	**64.71**	**85.59**	**61.55**	**84.43**

We report the performance of our strongest model for both the sign recognition and sign spotting benchmarks in Tab. 6. In Fig. 3, we provide some qualitative results from our sign recognition method and observe some modes of failure which are driven by strong visual similarity in sign production.

5.3 Comparison with the state of the art on ASL benchmarks

BSL-1K, being significantly larger than the recent WLASL [39] and MSASL [33] benchmarks, can be used for pretraining I3D models to provide strong initialisation for other datasets. Here, we transfer the features from BSL to ASL, which are two distinct sign languages.

As none of the models from [33, 39] are publicly available, we first reproduce the I3D Kinetics pretraining baseline with our implementation to achieve fair comparisons. We use 64-frame inputs as isolated signs in these datasets are significantly slower than co-articulated signs. We then train I3D from BSL-1K pretrained features. Tab. 7 compares pretraining on Kinetics versus our BSL-1K data. BSL-1K provides a significant boost in the performance, outperforming the state-of-the-art results (46.82% and 64.71% top-1 accuracy). Find additional details, as well as similar experiments on co-articulated datasets in the appendix.

6 Conclusion

We have demonstrated the advantages of using visual keyword spotting to automatically annotate continuous sign language videos with weakly-aligned subtitles. We have presented BSL-1K, a large-scale dataset of co-articulated signs that, coupled with a 3D CNN training, allows high-performance recognition of signs from a large-vocabulary. Our model has further shown beneficial as initialisation for ASL benchmarks. Finally, we have provided ablations and baselines for sign recognition and sign spotting tasks. A potential future direction is leveraging our automatic annotations and recognition model for sign language translation.

Acknowledgements. This work was supported by EPSRC grant ExTol. We also thank T. Stafylakis, A. Brown, A. Dutta, L. Dunbar, A. Thandavan, C. Camgoz, O. Koller, H. V. Joze, O. Kopuklu for their help.

Bibliography

[1] Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 2, 6

[2] Agris, U., Zieren, J., Canzler, U., Bauer, B., Kraiss, K.F.: Recent developments in visual sign language recognition. Universal Access in the Information Society **6**, 323–362 (2008) 4

[3] Antonakos, E., Roussos, A., Zafeiriou, S.: A survey on mouth modeling and analysis for sign language recognition. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (2015) 4

[4] Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Quan Yuan, Thangali, A.: The american sign language lexicon video dataset. In: CVPRW (2008) 3

[5] Bank, R., Crasborn, O., Hout, R.: Variation in mouth actions with manual signs in sign language of the Netherlands (ngt). Sign Language & Linguistics **14**, 248–270 (2011) 2

[6] Bilge, Y.C., Ikizler, N., Cinbis, R.: Zero-shot sign language recognition: Can textual data uncover sign languages? In: BMVC (2019) 4

[7] Buehler, P., Zisserman, A., Everingham, M.: Learning sign language by watching TV (using weakly aligned subtitles). In: CVPR (2009) 4, 9

[8] Camgoz, N.C., Hadfield, S., Koller, O., Bowden, R.: Using convolutional 3D neural networks for user-independent continuous gesture recognition. In: IEEE International Conference of Pattern Recognition, ChaLearn Workshop (2016) 4

[9] Camgoz, N.C., Hadfield, S., Koller, O., Bowden, R.: SubUNets: End-to-end hand shape and continuous sign language recognition. In: ICCV (2017) 4

[10] Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: CVPR (2018) 1, 3, 4

[11] Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: A dataset for recognising faces across pose and age. In: International Conference on Automatic Face and Gesture Recognition (2018) 7

[12] Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: arXiv preprint arXiv:1812.08008 (2018) 10, 11

[13] Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the Kinetics dataset. In: CVPR (2017) 4, 9, 10, 12

[14] Chai, X., Wang, H., Chen, X.: The devisign large vocabulary of chinese sign language database and baseline evaluations. Technical report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS (2014) 3

[15] Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: CVPR (2017) 2

[16] Chung, J.S., Zisserman, A.: Lip reading in the wild. In: ACCV (2016) 2
[17] Chung, J.S., Zisserman, A.: Signs in time: Encoding human motion as a temporal image. In: Workshop on Brave New Ideas for Motion Representations, ECCV (2016) 2, 4
[18] Cooper, H., Bowden, R.: Learning signs from subtitles: A weakly supervised approach to sign language recognition. In: CVPR (2009) 4
[19] Cooper, H., Pugeault, N., Bowden, R.: Reading the signs: A video based sign dictionary. In: ICCVW (2011) 4
[20] Cooper, H., Holt, B., Bowden, R.: Sign language recognition. In: Visual Analysis of Humans: Looking at People, chap. 27, pp. 539 – 562. Springer (2011) 2
[21] Crasborn, O.A., Van Der Kooij, E., Waters, D., Woll, B., Mesch, J.: Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. Sign Language & Linguistics (2008) 2
[22] Eng-Jon Ong, Koller, O., Pugeault, N., Bowden, R.: Sign spotting using hierarchical sequential patterns with temporal intervals. In: CVPR (2014) 4
[23] Farhadi, A., Forsyth, D.: Aligning ASL for statistical translation using a discriminative word model. In: CVPR (2006) 4
[24] Farhadi, A., Forsyth, D.A., White, R.: Transfer learning in sign language. In: CVPR (2007) 4
[25] Fillbrandt, H., Akyol, S., Kraiss, K..: Extraction of 3D hand shape and posture from image sequences for sign language recognition. In: IEEE International SOI Conference (2003) 4
[26] Fisher, C.G.: Confusions among visually perceived consonants. Journal of Speech and Hearing Research **11**(4), 796–804 (1968) 4
[27] Forster, J., Oberdörfer, C., Koller, O., Ney, H.: Modality combination techniques for continuous sign language recognition. In: Pattern Recognition and Image Analysis (2013) 4
[28] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 11
[29] Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 7
[30] Huang, J., Zhou, W., Li, H., Li, W.: Sign language recognition using 3D convolutional neural networks. In: International Conference on Multimedia and Expo (ICME) (2015) 4
[31] Huang, J., Zhou, W., Zhang, Q., Li, H., Li, W.: Video-based sign language recognition without temporal segmentation. In: AAAI (2018) 3, 4
[32] Jha, A., Namboodiri, V.P., Jawahar, C.V.: Word spotting in silent lip videos. In: WACV (2018) 2
[33] Joze, H.R.V., Koller, O.: MS-ASL: A large-scale data set and benchmark for understanding american sign language. In: BMVC (2019) 3, 4, 9, 11, 14
[34] Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding **141**, 108–125 (2015) 3, 4

[35] Koller, O., Ney, H., Bowden, R.: Read my lips: Continuous signer independent weakly supervised viseme recognition. In: ECCV (2014) 4
[36] Koller, O., Ney, H., Bowden, R.: Weakly supervised automatic transcription of mouthings for gloss-based sign language corpora. In: LREC Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel (2014) 4
[37] Koller, O., Ney, H., Bowden, R.: Deep learning of mouth shapes for sign language. In: Third Workshop on Assistive Computer Vision and Robotics, ICCV (2015) 4
[38] Koller, O., Zargaran, S., Ney, H.: Re-Sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs. In: CVPR (2017) 4
[39] Li, D., Opazo, C.R., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: WACV (2019) 3, 4, 9, 11, 12, 14
[40] Li, D., Yu, X., Xu, C., Petersson, L., Li, H.: Transferring cross-domain knowledge for video sign language recognition. In: CVPR (2020) 4
[41] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 10, 11
[42] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV (2016) 9
[43] Materzynska, J., Berger, G., Bax, I., Memisevic, R.: The Jester dataset: A large-scale video dataset of human gestures. In: ICCVW (2019) 12
[44] Momeni, L., Afouras, T., Stafylakis, T., Albanie, S., Zisserman, A.: Seeing wake words: Audio-visual keyword spotting. arXiv (2020) 9
[45] Nguyen, T.D., Ranganath, S.: Tracking facial features under occlusions and recognizing facial expressions in sign language. In: International Conference on Automatic Face and Gesture Recognition (2008) 4
[46] Ong, E., Cooper, H., Pugeault, N., Bowden, R.: Sign language recognition using sequential pattern trees. In: CVPR (2012) 4
[47] Pfister, T., Charles, J., Zisserman, A.: Large-scale learning of sign language by watching tv (using co-occurrences). In: BMVC (2013) 4, 5, 9
[48] Pfister, T., Charles, J., Zisserman, A.: Domain-adaptive discriminative one-shot learning of gestures. In: ECCV (2014) 4
[49] Schembri, A., Fenlon, J., Rentelis, R., Cormier, K.: British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2017 (Third Edition) (2017), http://www.bslcorpusproject.org 3
[50] Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., Cormier, K.: Building the British sign language corpus. Language Documentation & Conservation **7**, 136–154 (2013) 3
[51] Stafylakis, T., Tzimiropoulos, G.: Zero-shot keyword spotting for visual speech recognition in-the-wild. In: ECCV (2018) 2, 9
[52] Starner, T.: Visual Recognition of American Sign Language Using Hidden Markov Models. Master's thesis, Massachusetts Institute of Technology (1995) 4

[53] Sutton-Spence, R.: Mouthings and simultaneity in British sign language. In: Simultaneity in Signed Languages: Form and Function, pp. 147–162. John Benjamins (2007) 2
[54] Sutton-Spence, R., Woll, B.: The Linguistics of British Sign Language: An Introduction. Cambridge University Press (1999) 2, 6, 8
[55] Tamura, S., Kawasaki, S.: Recognition of sign language motion images. Pattern Recognition **21**(4), 343 – 353 (1988) 4
[56] Valli, C., University, G.: The Gallaudet Dictionary of American Sign Language. Gallaudet University Press (2005) 3
[57] Viitaniemi, V., Jantunen, T., Savolainen, L., Karppa, M., Laaksonen, J.: S-pot – a benchmark in spotting signs within continuous signing. In: LREC (2014) 3, 4, 9
[58] von Agris, U., Knorr, M., Kraiss, K.: The significance of facial features for automatic sign language recognition. In: 2008 8th IEEE International Conference on Automatic Face Gesture Recognition (2008) 3, 4
[59] Wilbur, R.B., Kak, A.C.: Purdue RVL-SLLL American sign language database. School of Electrical and Computer Engineering Technical Report, TR-06-12, Purdue University, W. Lafayette, IN 47906. (2006) 3, 4
[60] Woll, B.: The sign that dares to speak its name: Echo phonology in British sign language (BSL). In: Boyes-Braem, P., Sutton-Spence, R. (eds.) The hands are the head of the mouth: The mouth as articulator in sign languages, pp. 87–98. Hamburg: Signum Press (2001) 2, 4
[61] Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: A face detection benchmark. In: CVPR (2016) 9
[62] Yao, Y., Wang, T., Du, H., Zheng, L., Gedeon, T.: Spotting visual keywords from temporal sliding windows. In: Mandarin Audio-Visual Speech Recognition Challenge (2019) 2
[63] Ye, Y., Tian, Y., Huenerfauth, M., Liu, J.: Recognizing american sign language gestures from within continuous videos. In: CVPRW (2018) 4
[64] Zhou, H., Zhou, W., Zhou, Y., Li, H.: Spatial-temporal multi-cue network for continuous sign language recognition. CoRR **abs/2002.03187** (2020) 4