Sign Language Video Retrieval with Free-Form Textual Queries

Amanda Duarte[1,2]   Samuel Albanie[3]   Xavier Giró-i-Nieto[1,4]   Gül Varol[5]

[1]*Universitat Politècnica de Catalunya, Spain*   [2]*Barcelona Supercomputing Center, Spain*
[3]*Department of Engineering, University of Cambridge, UK*
[4]*Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Spain*
[5]*LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France*

https://imatge-upc.github.io/sl_retrieval/

Abstract

Systems that can efficiently search collections of sign language videos have been highlighted as a useful application of sign language technology. However, the problem of searching videos beyond individual keywords has received limited attention in the literature. To address this gap, in this work we introduce the task of sign language retrieval with free-form[1] textual queries: given a written query (e.g. a sentence) and a large collection of sign language videos, the objective is to find the signing video that best matches the written query. We propose to tackle this task by learning cross-modal embeddings on the recently introduced large-scale How2Sign dataset of American Sign Language (ASL). We identify that a key bottleneck in the performance of the system is the quality of the sign video embedding which suffers from a scarcity of labelled training data. We, therefore, propose SPOT-ALIGN*, a framework for interleaving iterative rounds of sign spotting and feature alignment to expand the scope and scale of available training data. We validate the effectiveness of* SPOT-ALIGN *for learning a robust sign video embedding through improvements in both sign recognition and the proposed video retrieval task.*

1. Introduction

Sign languages are the primary means of communication among deaf communities. They are visual, complex, evolved languages that employ combinations of *manual* and *non-manual markers* such as movements of the face, body and hands to convey information [56].

Recent developments in automatic speech recognition (ASR) for spoken languages [14,15,65,70] have enabled automatic captioning of vast swathes of video content hosted
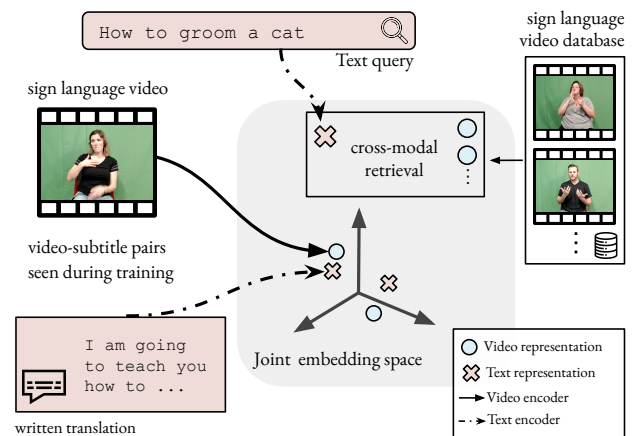

Figure 1. **Text-based sign language video retrieval:** In this work we introduce *sign language video retrieval with free-form textual queries*, the task of searching collections of sign language videos to find the best match for a free-form textual query, going beyond individual keyword search.

on platforms such as YouTube. In addition to rendering the videos more accessible, this captioning yields a second important benefit: it allows the content of the videos to be indexed and efficiently searched with text queries. By contrast, the same automatic captioning capability (and hence searchability) does not exist for sign language content. Indeed, recent work has drawn attention to the pressing need to develop systems that can index archives of sign language videos to render them searchable [6]. Without these tools, sign language video creators must type the spoken language translation of their content if they want to reach the same discoverability as their spoken language counterparts.

One solution might appear to be to use sign language translation systems to perform video captioning, analogous to ASR cascading in spoken content retrieval [34]. Unfortunately, while promising translation results have been demonstrated in constrained domains of discourse (such as

---

[1]The terminology "natural language query" is commonly used to describe unconstrained textual queries in spoken languages. However, since sign languages are also natural languages, we adopt for the term "free-form textual query" instead.

weather forecasts) [9, 10, 37], it has been widely observed that these systems are unable to achieve functional performance across the multiple domains of discourse [6, 29, 62] required for open-vocabulary video indexing (see Appendix D). An alternative solution would be to employ existing methods for *sign spotting* to perform keyword search. However, such approaches are fundamentally brittle—they work best when the user knows exactly which signs of interest were used in the video. Moreover, to build an accurate index of such signs using recent sign spotting techniques [1, 26, 46] requires a list of appropriate query candidates, which to date have often been obtained from subtitles corresponding to speech transcriptions of the translation, for example from an ASR engine. We focus on sign language videos produced by and for signers, that may not contain any speech track, so producing such speech transcriptions is not an option.

In this work, we address the task of sign language video retrieval with free-form textual queries by learning a joint embedding space between text and video as illustrated in Fig. 1. Cross-modal embeddings target only the task necessary to enable search (i.e. ranking a finite pool of sign language videos), rather than the more involved task of full sign language translation. As we demonstrate through experiments, this renders their practical application even across multiple topics. Moreover, cross-modal embeddings enable extremely efficient search (with the potential to scale up to collections of billions of videos thanks to mature approximate nearest neighbour algorithms for embedding spaces [27]).

The task of sign language video retrieval is extremely challenging for several reasons: (1) Translation mappings between sign languages and spoken languages are highly complex [57], with differing modalities and grammar structures (ordering is typically not preserved between signed and spoken languages, for example); (2) In contrast to the datasets used to train text-video retrieval models (millions of paired examples of videos with corresponding sentences [5, 41]) sign language datasets are orders of magnitude smaller in scale; (3) In addition to a paucity of paired data, the annotated data available for learning robust sign embeddings is also extremely scarce (with sign recognition datasets also considerably smaller than their counterparts for action recognition [11, 25], for example).

In this work, we propose to study sign language video retrieval on the recently released How2Sign American Sign Language (ASL) dataset [20]. To the best of our knowledge, this dataset represents the largest public source of sign language videos with aligned captions. In order to address the first and second challenges highlighted above, we construct cross-modal embeddings that leverage pretrained language models to reduce the burden of data required to learn the mapping between signing sequences and sentences. To address the third annotation scarcity challenge, we propose SPOT-ALIGN, a framework for automatic annotation that integrates multiple sign spotting methods to automatically annotate significant fractions of the How2Sign dataset. By training on the resulting annotations, we obtain more robust sign embeddings for the downstream retrieval task.

In summary, we make the following contributions: (1) We introduce the task of *sign language video retrieval with free-form textual queries*; (2) We provide several baselines for this task, demonstrating the value of cross modal embeddings and the benefits of incorporating additional retrieval cues from a sign recognition method (whose predictions provide a basis for text-based similarity search) on the How2Sign and PHOENIX2014T datasets; (3) We propose the SPOT-ALIGN framework for automatic annotation and demonstrate its efficacy in producing more robust sign embeddings; (4) We contribute a new manually annotated test set for the How2Sign benchmark.

2. Related Work

Our work relates primarily to existing research in *text-video embeddings for video retrieval*, *sign language video retrieval* and *automatic annotation of sign language videos with auxiliary cues*, discussed next.

Text-video embeddings for video retrieval. Recently, there has been extensive research interest in enabling video content search with textual queries via cross-modal embeddings. Following the seminal DeViSE model [23] that demonstrated the strength of this approach for images and text, a wide array of text-video embeddings have been explored [5, 17, 19, 24, 39, 40, 43, 44, 48, 64, 67]. Differently from these works which target the retrieval of *describable events*, our work focuses on retrieving *signing content* that matches a spoken language query formulated with text. As noted in Sec. 1, a key challenge that arises from this distinction is the relative paucity of training data available to learn a robust sign video embedding—in this work, we propose SPOT-ALIGN (introduced in Sec. 3) to explicitly address this challenge.

Sign language video retrieval. The task of sign language video retrieval has primarily been investigated under the query-by-example search paradigm, in constrained domains and with small datasets. In this formulation, a user query consists of an example of the sign(s) of interest, similarly how most keyword-based search engines deal with text databases. Two particular variants of this problem have received attention for sign language video retrieval: searching *visual dictionaries of isolated signs*, and searching *continuous sign language datasets*, discussed next.

Sign language dictionaries are video repositories with recordings of individual signs suitable for learners. To search such videos, Athitsos et al. [4] coupled hand motion cues with Dynamic Time Warping (DTW) to enable signer-independent search of an American Sign Language (ASL) dictionary containing 3k signs and testing with 921 queries. For *continuous sign language datasets*, the goal is retriev-

ing all occurrences of a demonstrated query sign in a target video. Different techniques have been proposed for this purpose, including hand features with CRFs [66], hand motion with sequence matching [69], hand and head centroids [35], per-frame geometric features coupled with HMMs [68], and non-face skin distribution matching [63].

As an alternative to querying by example, a number of works have investigated *sign spotting* with learned classifiers. Ong et al. [47] tackled this problem with HSP-Trees, a hierarchical data structure built upon Sequential Interval Patterns. Later work combined human pose estimation with temporal attention mechanisms to detect (but not localise) the presence of a set of glosses among signing sequences [60]. This work was later extended to enable search for individual words [58] and further extended to additionally incorporate hand-shape features, improving performance [59]. More recently, Jiang et al. [26] showed the effectiveness of the transformer architecture for the sign spotting task, achieving promising results on the BSLCOR-PUS [53] and Phoenix2014 [31] datasets.

However, to the best of our knowledge, no prior sign language retrieval literature has considered the task that we propose in our work, namely *retrieving sign language videos with free-form textual queries*.

Automatic annotation of sign language with auxiliary cues. The abundance of *audio-aligned subtitles* in broadcast data with sign language interpreters has motivated a rich body of work that has sought to use them as an auxiliary cue to annotate signs. Cooper and Bowden [16] propose to use a priori mining to establish correspondences between subtitles and signs in news broadcasts. Alternative approaches investigate the use of Multiple Instance Learning [7, 28, 49]. Other recent contributions leverage words from audio-aligned subtitles with keyword spotting methods based on mouthing cues [1], dictionaries [46] and attention maps generated by transformers [62] to annotate large numbers of signs, as well as to learn domain invariant features for improved sign recognition through joint training [38].

Similarly to these works, we also aim to automatically annotate sign language videos by making use of audio-aligned subtitles. To this end, we make use of prior keyword spotting methods [1, 46]. However, differently from all the other methods mentioned above we propose an *iterative* approach, SPOT-ALIGN, that alternates between repeated sign spotting (to obtain more annotations) and jointly training on the resulting annotations together with dictionary exemplars (to obtain better features for spotting). We note that iterative labelling frameworks have previously seen success in the context of dense sequence re-alignment methods [30,32,50] (differently, we target the sparse annotation problem). We show that our methodology significantly increases the automatic annotation yield, and we demonstrate the value of these additional annotations by using them to learn better representations for downstream tasks.

3. Sign Language Retrieval

In this section, we first formulate the task of sign language video retrieval with free-form textual queries (Sec. 3.1). Next, we describe the cross-modal (CM) learning formulation considered in this work (Sec. 3.2), before introducing our SPOT-ALIGN framework for annotation enhancement (Sec. 3.3). Finally, we present our text-based retrieval through our sign recognition (SR) model (Sec. 3.4). Further implementation details are provided in Appendix B.

3.1. Retrieval task formulation

Let $\mathcal{V}$ denote a *dataset* of sign language videos of interest, and let t denote a free-form textual user query. The objective of the *sign language video retrieval with textual queries* task is to find the signing video $v \in \mathcal{V}$ whose signing content best matches the query t. We use *text-to-sign-video* (T2V) as notation to refer to this task. Analogously to the symmetric formulations considered in the existing cross-modal retrieval literature [18, 40], we also consider the reverse *sign-video-to-text* (V2T) task, in which a signing video, v, is used to query a collection of text, $\mathcal{T}$.

3.2. Cross modal retrieval embeddings

To address the retrieval task defined above, we assume access to a parallel corpus of signing videos with corresponding written translations. We aim to learn a pair of encoders, ϕ_v and ϕ_T, which map each signing video v and text t into a common real-valued embedding space, $\phi_v(v), \phi_T(t) \in \mathbb{R}^C$, such that $\phi_v(v)$ and $\phi_T(t)$ are close if and only if t corresponds to the content of the signing in v. Here C denotes the dimensionality of the common embedding space.

To learn the encoders, we adopt the cross modal ranking learning objective proposed by Socher et al. [54]. Specifically, given paired samples $\{(v_n, t_n)\}_{n=1}^N$, we optimise a max-margin ranking loss:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1, i \neq j}^{B} [\eta_{ij} - \eta_{ii} + m]_+ + [\eta_{ji} - \eta_{ii} + m]_+ \quad (1)$$

where m denotes the margin hyperparameter, $[\cdot]_+$ denotes the hinge function $\max(\cdot, 0)$, B denotes the size of mini-batch sampled during training, and η_{ij} denotes the cosine similarity between signing video v_i and text t_j.

Once learned, the embeddings can be applied directly to both the T2V and V2T tasks. For the former, inference consists of simply computing the cosine similarity between the text query t and every indexed signing video $v \in \mathcal{V}$ to produce a ranking (and vice versa for the V2T task).

Encoder architectures. The sign video encoder, ϕ_v consists first of an initial *sign video embedding*, ψ_v, which we instantiate as an I3D neural network [12] over clips of 16 frames (motivated by its effectiveness for sign recognition
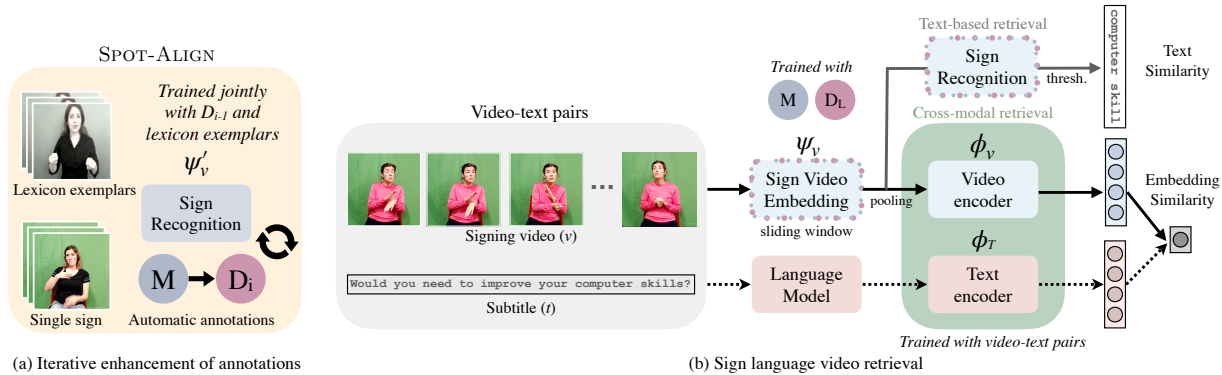
Figure 2. **Method overview:** *(a)* We propose SPOT-ALIGN, a framework for iteratively increasing annotation yield to obtain a good sign video embedding. At each iteration i, the current sign video embedding ψ'_v is trained for classification jointly on the How2Sign annotations from iteration $i-1$ and lexicon exemplars from the WLASL [36] and MSASL [61] datasets. The resulting improved embedding is then used to obtain a new set of sign spottings by re-querying How2Sign videos with lexicon exemplars. Our final sign video embedding, ψ_v, is obtained by training on the mouthing spottings (M) together with the last iteration of dictionary sign spottings, D_L (without joint training across lexicon exemplars). We provide a detailed sketch of this pipeline in Appendix B.1. *(b)* To perform cross modal retrieval, we employ ψ_v, together with a language model, to produce embeddings of videos and text. These are passed to a video encoder and text encoder, respectively, which are trained to project them into a joint space such that they are close if and only if the text matches the video. The embedding produced by ψ_v is additionally passed to a sign recognition model, providing the basis for text-based similarity search.

tasks [1, 36, 61]). The output of ψ_v is temporally aggregated to a fixed size vector, and then projected to the C-dimensional cross modal embedding space, $\phi_v(v) \in \mathbb{R}^C$.

To implement ϕ_T, each text sample, t, is first embedded through a language model that has been pretrained on large corpora of written text. The resulting sequence of word embeddings are then combined via NetVLAD [3] and projected via Gated Embedding Unit following the formulation of [40] to produce a fixed-size vector, $\phi_T(t) \in \mathbb{R}^C$.

In this work, we pay particular attention to the initial sign video embedding, ψ_s, which, as we show through experiments in Sec. 4, has a critical influence on performance. In Sec. 4, we also conduct experiments to evaluate suitable candidates for both the temporal aggregation mechanism on ϕ_V, and the language model employed by ϕ_T.

3.3. Iterative enhancement of video embeddings

As noted above, an effective cross modal embedding for our task requires a good sign video embedding. A key challenge in obtaining such an embedding is the relative paucity of annotated sign language data for training. For example, to the best of our knowledge, there are no large-scale public datasets of *continuous* signing with corresponding sign annotations in ASL.

To address this challenge, we propose the SPOT-ALIGN framework (Fig. 2a) which we use to obtain large numbers of *automatic* sign annotations on the How2Sign dataset. This dataset provides videos with corresponding written English translations but currently lacks such annotations.

In summary, we first obtain a collection of candidate sign annotations using sign spotting techniques proposed in recent works that employ mouthing cues [1] and dictionary

examples [46]. We supplement these sparse annotations: iteratively increasing the amount of dictionary-based annotations by retraining our sign video embeddings, and re-querying dictionary examples. Next, we describe each of these steps.

Mouthing-based sign spotting [1]. First, we use the mouthing-based sign spotting framework of [1] to identify sign locations corresponding to words that appear in the written How2Sign translations. This approach, which relies on the observation that signing sometimes makes use of mouthings in addition to head movements and manual gestures [57], employs the keyword spotting architecture of [55] with the improved P2G phoneme-to-grapheme keyword encoder proposed by Momeni et al. [45]. We search for keywords from an initial candidate list of 12K words that result from applying text normalisation [22] to words that appear in How2Sign translations (to ensure that numbers and dates are converted to their written form, e.g. "7" becomes "seven") and filtering to retain only those words that contain at least four phonemes. Whenever the keyword spotting model localises a mouthing with a confidence over 0.5 (out of 1), we record an annotation. With this approach, we obtain approximately 37K training annotations from a vocabulary of 5K words. We filter these words to those that appear in the vocabulary of either WLASL [36] or MSASL [61] lexical datasets. The resulting 9K training annotations cover a vocabulary of 1079 words, which consists of our initial vocabulary for training a sign recognition model.

Dictionary-based sign spotting [46]. Next, we employ an exemplar-based sign spotting method similar to [46]. This approach considers a handful of video examples per sign

which are used as visual queries to compare against the continuous test video. The location is recorded as an automatic annotation for the queried sign at the time where the similarity is maximised. Such similarity measure between the query and the test videos requires a joint space. In [46], a complex two-stage contrastive training strategy is formulated. In this work, we opt for a simpler mechanism in which we *jointly* train a sign recognition model with an I3D backbone, denoted ψ'_v on the set of query videos (which are often from an isolated domain such as lexical dictionaries) and sign-annotated videos from our search domain (i.e. How2Sign sparse annotations obtained from the previous step of mouthing-based spotting). The latent features from this classification model (which are now approximately aligned between the two domains) are then used to compute cosine similarities.

Similarly to the mouthing method, we select candidate query words for each video based on the subtitles. However, when employing dictionary spotting, we look for both the original and the lemmatised (removing inflections) forms of the words, since the sign language lexicons we employ usually contain a single version of each word (e.g. 'run' instead of 'running').

As the source of sign exemplars from which we construct queries, we make use of the training sets of WLASL [36] and MSASL [61], two datasets of isolated ASL signing, with 2K and 1K vocabulary sizes, respectively. For joint training, we select samples from their training subsets that occur in the 1079-sign vocabulary from our previous mouthing annotations. However, we use the full training sets for querying, allowing us to automatically annotate signs outside of the initial 1079 signs. We record all annotations where the maximum similarity (over all exemplars per sign) is higher than 0.75 (out of 1), resulting in 59K training annotations from an expanded vocabulary of 1887 signs. We initialise the I3D classification from the pretrained BSL recognition model released by the authors of [62].

Iterative enhancement via SPOT-ALIGN. From the previous two methods, we obtain an initial set of automatic annotations. However, the yield of the dictionary-based spotting method is heavily limited by the *domain gap* between the videos of How2Sign and the datasets used to obtain the exemplars. It is therefore natural to ask whether we can improve the yield from dictionary-based spotting by achieving a better feature alignment between the dictionary exemplar and How2Sign domains. To this end, we introduce a retrain-and-requery framework, which we call SPOT-ALIGN, described next.

At iteration i, we employ the I3D latent features obtained by joint training between WLASL-MSASL lexicons and How2Sign automatic annotations provided by iteration $i - 1$. We observe a significant increase in the yield (e.g. 160K annotations in D_2 vs 59K annotations in D_1) despite using the *same exemplars* and *same subtitles* to construct our queries. The key difference is then the better aligned
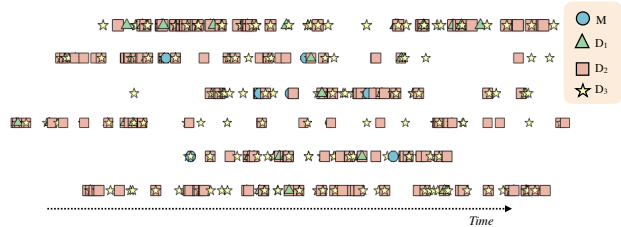

Figure 3. **Iterative enhancement of automatic annotations:** We illustrate sparse annotations generated by different iterations of the SPOT-ALIGN framework on six different video segments (rows) over a fixed-duration interval of 50 seconds each (x-axis).

embeddings with which we compare the exemplar and test videos. In Fig. 3, we illustrate the resulting sparse annotations over a continuous timeline for sample videos where we observe that the density of annotations significantly increases with SPOT-ALIGN iterations. We denote with D_i, the set of automatic training annotations after applying iteration i. An overview of this process is shown in Fig. 2a.

Given the annotations from the final iteration of this process, we train a new sign recognition model (trained only on the continuous dataset, i.e. How2Sign), from which we obtain our ultimate video sign embedding ψ_v using the (1024-dimensional) latent representation before the classification layer of 1887 signs. As shown in Fig. 2b, this embedding underpins the sign video encoder, ϕ_v, of our cross modal embedding, and is also used to classify individual signs to enable text-based retrieval, described next.

3.4. Text-based retrieval by sign recognition

The individual sign recognition model used to train the sign video embedding ψ_v can naturally be used to obtain a *sequence* of signs if applied in a sliding window manner on the long signing videos from v. While the performance of this model is not expected to be high (due to a lack of temporal modelling stemming from the lack of continuous annotations), the output list of predicted sign categories gives us a set of candidate words which can be used to check the overlap with the query text. This is analogous to cascading ASR for spoken content retrieval [34], except that sign recognition is significantly more difficult than speech recognition (in part, due to a lack of training data [6]). Since the order of signs do not necessarily follow word order in the translated text, we simply compute an Intersection over Union (IoU) to measure similarity between a query text and the recognised signs. Before we compute the IoU, we lemmatise both the query words and predicted words. We constrain the set of recognised signs by removing duplicates and removing classifications that have probabilities below a certain threshold (0.5 in our experiments). In Sec. 4, we show that this text-based retrieval approach, while performing worse than the cross modal retrieval approach, is complementary and can significantly boost overall performance. Implementation details are described in Appendix B.

4. Experiments

We first present the datasets, annotations and evaluation protocols used in our experiments (Sec. 4.1). Next, we provide retrieval results on How2Sign dataset, conducting ablation studies to evaluate the influence of different components of our approach (Sec. 4.2). Then, we establish baseline retrieval performances on the PHOENIX2014T dataset (Sec. 4.3). Finally, we discuss the limitations together with qualitative analysis and societal impact (Sec. 4.4).

4.1. Data, annotation and evaluation protocols

Datasets. In this work, we primarily focus on the recently released **How2Sign** dataset [20], a multimodal open-vocabulary and subtitled dataset of about 80 hours of continuous sign language videos of American Sign Language translation of instructional videos. The recorded videos span a wide variety of topics. We use the videos and their temporally aligned subtitles for training and evaluating the retrieval model, taking subtitles as textual queries. There are 31075, 1739 and 2348 video-subtitle pairs in training, validation and test sets, respectively. Note that, we remove a small number of videos from the original splits, where the subtitle alignment is detected to fall outside the video duration (more details can be found in Appendix C). We use the validation set to tune parameters (i.e. training epoch), and report all results on the test set.

We also evaluate our sign language retrieval method to provide baselines on the **PHOENIX2014T** dataset [10], (although this is not our central focus due to its restricted domain of discourse). PHOENIX2014T contains German Sign Language (DGS) videos depicting weather forecast videos. The dataset consists of 7096, 519 and 642 training, validation and test video-text pairs, respectively. The benchmark is primarily used for sign language translation where promising results can be obtained due to the restricted vocabulary size of 3K German words. Here, we re-purpose it for retrieval, providing baselines using both our cross-modal embedding approach, and a text-based retrieval by sign language translation [10].

Annotations. For sign recognition, we train using the automatic sparse annotations produced by the SPOT-ALIGN framework. Summary statistics obtained across multiple iterations of sign spotting are illustrated in Fig. 4 (left), where we observe a significant increase in yield across consecutive iterations. To enable evaluation of sign recognition performance, we construct a manually verified test set. This is done by providing annotators proficient in ASL with sign spotting candidates using the VIA annotation tool [21]. This results in a recognition test set of 2,212 individual sign video-category pairs, available at our project page.

Evaluation metrics. To evaluate retrieval performance, we follow the existing retrieval literature [24, 39, 40] and report standard metrics R@K (recall at rank K, higher is better) and MedR (median rank, lower is better). For sign recog-
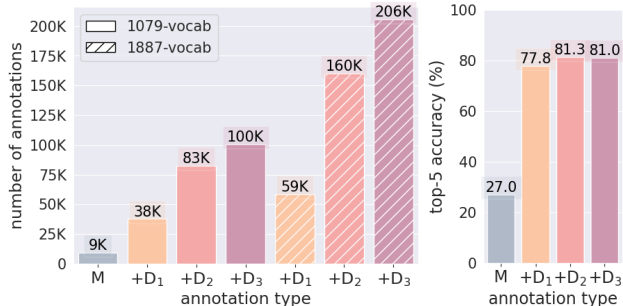

Figure 4. **Iteratively increasing the sign annotations:** Starting from a small set of mouthing annotations, we apply sign spotting through dictionaries several times, by retraining our I3D backbone on the previous set of automatic annotations. The left plot demonstrates the significant increase in the number of annotations, for both the restricted (1079) and the full (1887) set of categories. The right plot reports individual sign recognition (1079-way classification) results on the manually verified test set.

Table 1. **Effect of sign video embeddings:** The iterative increase of sign annotations with mouthing- (M) and dictionary-based (D) spotting improves the performance for sign video retrieval tasks with both sign recognition and cross-modal embeddings. The embeddings for the last seven rows are obtained from How2Sign trainings, pretrained on BOBSL (second row), which itself was pretrained on Kinetics (first row).

Sign-Vid-Emb	Vocab	Sign Recognition				Cross-Modal Embeddings			
		R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
Kinetics [12]	-	-	-	-	-	$1.0_{0.1}$	$4.4_{0.4}$	$6.9_{0.6}$	$296.8_{12.5}$
BOBSL [2]	-	-	-	-	-	$17.2_{0.6}$	$32.5_{0.7}$	$39.5_{1.3}$	$30.5_{2.2}$
M	1079	0.6	2.3	4.4	1174.5	$16.4_{1.2}$	$31.1_{0.8}$	$38.2_{0.8}$	$32.7_{3.1}$
M+D$_1$	1079	10.2	21.2	26.5	136.3	$20.6_{1.1}$	$36.7_{0.6}$	$43.3_{0.9}$	$22.0_{2.6}$
M+D$_2$	1079	15.6	29.0	33.9	92.0	$21.8_{0.4}$	$38.0_{0.6}$	$44.6_{0.8}$	$18.2_{2.0}$
M+D$_3$	1079	16.7	29.1	33.3	95.3	$21.9_{1.2}$	$38.2_{0.7}$	$44.8_{0.5}$	$18.7_{0.6}$
M+D$_1$	1887	14.1	26.1	31.4	88.0	$20.4_{0.6}$	$36.4_{0.3}$	$43.5_{0.7}$	$20.0_{1.0}$
M+D$_2$	1887	18.3	31.3	35.8	69.8	$23.7_{0.5}$	$40.8_{0.1}$	$47.1_{0.2}$	$14.7_{0.6}$
M+D$_3$	1887	**18.4**	**32.2**	**36.6**	**68.0**	$\mathbf{24.5}_{0.2}$	$40.7_{1.1}$	$46.7_{0.7}$	$15.7_{1.5}$

nition baselines (for which the time required to retrain ψ_v is more extensive), we report the results of a single model. For cross modal embedding ablations (for which the sign video embedding ψ_v is frozen, and only the text encoder, ϕ_T, and video encoder, ϕ_V, are trained), we report the mean and standard deviation over three randomly seeded runs.

4.2. Retrieval results on How2Sign

In this section, we present ablation studies experimenting with: (i) different sign video embeddings, (ii) video embedding aggregation mechanisms, and (iii) text embeddings. We further study (iv) the probability threshold hyperparameter for text-based retrieval via sign recognition. We also highlight (v) the importance of having a sign language aligned subtitle data by experimenting with using the original speech-aligned timings provided by [20]. Finally, we demonstrate the advantages of (vi) combining our cross-modal embedding similarities with text-based similarities via sign recognition.

(i) Comparison of sign video embeddings. Our main re-

Text query | **Sign video retrieval**

"OK, we're going to make some lidded jars today and first thing you want to start off with obviously is your clay." (GT rank: 1)

Similarity 0.36 — "OK, we're going to make some lidded jars today and first thing..."

Similarity 0.34 — "It's first off books I bought a quite of few books I've gotten toy books..."

"I hope you're having fun." (GT rank: 3)

Similarity 0.28 — "Cheers!"

Similarity 0.27 — "So just be relaxed and have, just do whatever you need to do to make your guest have a good"

Figure 5. **Qualitative results on text to sign language retrieval:** For each query, we show frames from the top two ranked videos as well as their corresponding sentences (these are not used during retrieval, and are provided for visualisation purposes). The top row shows a success case. The bottom row shows a failure case in which the retrieval model struggles with a less detailed query. More examples can be seen in Appendix A.

sults on sign language retrieval are summarised in Tab. 1. Here, we assess the quality of our end-to-end video classification model to obtain sign video embeddings from the last layer of the I3D model. We report both the sign language retrieval from the sign recognition outputs (using text-to-text matching, as described in Sec. 3.4) on the left, and the learned cross-modal embeddings (text-to-video matching) on the right.

We first observe that our cross-modal embeddings (which can potentially capture cues beyond the limited categories of the sign recognition model) perform significantly better than their text-based counterparts. Next, we compare various choices of backbone sign video embeddings to evaluate the effectiveness of our proposed SPOT-ALIGN framework. As a first baseline, we experiment with using standard Kinetics [12] training—we observe that this produces video embeddings that (as expected) perform poorly for our task. We also include as a baseline the model from [2] (pretrained on the BOBSL data) that was used to initialise our I3D sign video embedding. The sensitivity of our model to different initialisations is demonstrated in Appendix D. While strongly outperforming Kinetics features, BOBSL features remain substantially weaker than the end-to-end ASL sign recognition training on How2Sign enabled by SPOT-ALIGN.

We observe improvements from each of our SPOT-ALIGN iterations, instantiated from mouthing-only (M) annotations, expanded first in the number of annotations within the same vocabulary size of 1079, then expanded in the number of sign categories with 1887-way classification. The corresponding statistics for the training size are illustrated in Fig. 4 (left), and the sign recognition performance of the corresponding models on the manually verified test set can be seen in Fig. 4 (right). In light of their superior performance, we use sign video embeddings trained with $M+D_3$ annotations from the 1887 large-vocabulary for the rest of our experiments on How2Sign

Table 2. **Influence of sign video embedding aggregation strategy:** We compare temporal pooling strategies on the How2Sign retrieval benchmark. Performance metrics are reported as means and standard deviations over three randomly seeded runs.

Aggregation method	R@1↑	R@5↑	R@10↑	MedR↓
Max pooling	$23.3_{0.3}$	$39.7_{0.5}$	$46.3_{0.6}$	$15.3_{0.6}$
Avg. pooling	$24.5_{0.2}$	$40.7_{1.1}$	$46.7_{0.7}$	$15.7_{1.5}$

Table 3. **Influence of the text embedding:** We compare a variety of text embeddings on the How2Sign retrieval benchmark. Performance metrics are reported as means and standard deviations over three randomly seeded runs.

Text Embedding	R@1↑	R@5↑	R@10↑	MedR↓
GPT [51]	$15.4_{0.4}$	$30.5_{0.4}$	$37.6_{0.4}$	$30.2_{1.3}$
GPT-2-xl [52]	$17.0_{0.3}$	$32.5_{0.4}$	$39.6_{0.4}$	$25.7_{1.2}$
Albert-XL [33]	$19.7_{0.3}$	$36.7_{0.3}$	$43.8_{0.4}$	$19.2_{0.8}$
W2V [42]	$24.2_{0.4}$	$40.0_{0.4}$	$46.7_{0.2}$	$14.8_{0.3}$
GrOVLE [8]	$24.5_{0.2}$	$40.7_{1.1}$	$46.7_{0.7}$	$15.7_{1.5}$

(ii) Video embedding aggregation. Next, we compare the use of different temporal pooling strategies on the sequence of sign video embeddings for a given sign language video. While more sophisticated temporal aggregations are possible, in this work, we opt for a simple and efficient average pooling mechanism, which has widely been shown to be effective for text-video retrieval tasks [17, 40]. In Tab. 2, we compare average pooling with maximum pooling over the temporal axis for each feature dimension. We observe that average pooling performs best.

(iii) Text embedding. We then compare several choices of text embedding for the training of cross modal embeddings. We report the results in Tab. 3. We observe that word2vec [42] and GrOVLE [8] obtain competitive performance, outperforming higher capacity alternatives [33, 51, 52]. This phenomenon is also observed in [17], where the authors show that for a number of source text distributions, simpler word embeddings can outperform their "heavyweight" counterparts. We leave the end-to-end fine-tuning of the language models with our sign language translations to future work, which can potentially provide further improvements, and use GrOVLE embeddings for the rest of the experiments.

(iv) Sign recognition probabilities. Here, we ablate the text-based retrieval approach which employs a sign recognition classifier. Since the sliding window is applied at each frame densely, we obtain one sign prediction per frame (which can be very noisy). Consequently, an important hyperparameter for this method is the selection of which classification outputs to consider in our set of predicted words (which will in turn guides the text-based retrieval). Concretely, the hyperparameter we vary is the confidence threshold at which predictions are included as text tokens. We explore several threshold values in Tab. 4 and report retrieval performance. We observe that 0.5 performs best— we adopt this value for our remaining experiments.

Table 4. **Thresholding sign recognition probabilities:** We investigate the influence of the confidence threshold hyperparameter on How2Sign retrieval performance, and observe that 0.5 works best.

Threshold	R@1↑	R@5↑	R@10↑	MedR↓
0.00	13.1	26.5	32.0	75.5
0.10	13.4	26.4	32.4	74.0
0.25	17.5	30.9	35.4	56.5
0.50	**18.4**	**32.2**	**36.6**	**68.0**
0.75	15.0	27.9	32.4	91.0

Table 5. **Effect of subtitle alignment:** We report retrieval performance on How2Sign for models trained and evaluated on subtitles aligned to speech and to signing. We observe a significant drop in performance when using speech-aligned subtitles.

Alignment	Sign Recognition				Cross-Modal Embeddings			
	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
Speech	9.5	16.1	19.0	418.0	$5.9_{0.6}$	$13.6_{0.6}$	$18.0_{0.2}$	$483.5_{17.9}$
Signing	18.4	32.2	36.6	68.0	$24.5_{0.2}$	$40.7_{1.1}$	$46.7_{0.7}$	$15.7_{1.5}$

Table 6. **Combination of models:** We report our final benchmark performance on How2Sign for the retrieval models based on sign recognition (SR) and cross modal (CM) embeddings. We observe that the two approaches are highly complementary.

Models	T2V				V2T			
	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
SR	18.4	32.2	36.5	68.0	11.5	27.9	33.3	66.0
CM	24.7	39.6	46.0	17.0	17.9	40.8	46.6	15.0
SR + CM	**32.8**	**47.7**	**52.9**	**7.0**	**23.3**	**48.5**	**53.7**	**7.0**

Table 7. **Retrieval performance on the PHOENIX2014T dataset:** We report baseline performances for cross modal embeddings, as well as text-based retrieval by sign language translation on the 642 sign-sentence pairs of the test set.

Text Embedding	T2V				V2T			
	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
Translation [10]	30.2	53.1	63.4	4.5	28.8	52.0	60.8	56.1
Cross-modal	48.6	76.5	84.6	2.0	50.3	78.4	84.4	**1.0**
Combination	**55.8**	**79.6**	**87.2**	**1.0**	**53.1**	**79.4**	**86.1**	**1.0**

(v) Effect of alignment. We next highlight the importance of having aligned video-sentence pairs, motivating our selection of How2Sign (which contains the largest public source of aligned video-sentence pairs) as a test-bed for our study of large-vocabulary retrieval. We retrain cross modal embeddings on speech-aligned training data, and evaluate both the recognition and cross modal embedding models on the speech-aligned test data of How2Sign to compare with signing-aligned version. The results are reported in Tab. 5, where we observe that speech-aligned subtitles significantly damage retrieval performance.

(vi) Combining several cues. Finally, in Tab. 6, we combine our two types of models based on sign recognition (SR) and cross-modal embeddings (CM). We perform *late fusion* (averaging the similarities, with equal weights) computed with individual models. We conclude that sign recognition provides complementary cues to our cross-modal embedding training, significantly boosting the final performance. Tab. 6 presents both T2V and V2T performances establishing a new benchmark on the task of retrieval for the recent How2Sign dataset. Some qualitative examples of videos retrieved by our system are provided in Fig. 5.

4.3. Retrieval results on PHOENIX2014T

In addition to the How2Sign ASL dataset that formed the primary basis of our study, we also provide retrieval baselines on the PHOENIX2014T dataset [9, 31]. For cross modal embedding training, we employ a text embedding model trained on German language corpora, GPT-2 [52] released by Chan et al. [13]. For text-based retrieval, here we incorporate a state-of-the-art sign language translation model [10], with which we compute an IoU similarity measure. Note that sign language translation performance is high on this dataset due to its restricted domain of discourse, which is the reason why we opt for a translation-based approach instead of the recognition-based retrieval as in Sec. 3.4. In both experiments we use the video features provided by [10, 30]. The results are reported in Tab. 7. We observe that our cross-modal embeddings strongly outperform the translation-based retrieval. Their combination performs best (as in Tab. 6).

4.4. Limitations and societal impact

One limitation of SPOT-ALIGN method is that it is not able to discover new signs outside of the vocabulary of queried lexicons. Qualitatively, we observe failure cases of our cross-modal retrieval model (illustrated in Fig. 5), when using more generic queries that lack precise detail. It is also worth noting that all datasets used in our experiments are interpreted sign language rather than conversational (e.g. conversations between native signers – see [6] for a broader discussion on how this can limit models trained on such data). **Societal Impact.** The ability to efficiently search sign language videos has a number of useful applications for content creators and researchers in the deaf community. However, by providing this technical capability, it also potentially brings risk of surveillance of signers, since large volumes of signing content can be searched automatically.

5. Conclusion

In this work, we introduced the task of sign language video retrieval with free-form textual queries. We provided baselines for this task on the How2Sign and PHOENIX2014T datasets. We also proposed the SPOT-ALIGN framework to obtain automatic annotations, and demonstrated their value in producing effective sign video embeddings for retrieval.

References

[1] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 2, 3, 4

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*, 2021. 6, 7

[3] Relja Arandjelović, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1437–1451, 2018. 4

[4] Vassilis Athitsos, C. Neidle, Stan Sclaroff, Joan P. Nash, Alexandra Stefan, Ashwin Thangali, Haijing Wang, and Quan Yuan. Large lexicon project: American Sign Language video corpus and sign language indexing / retrieval algorithms. In *Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT)*, 2010. 2

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[6] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st international ACM SIGACCESS conference on computers and accessibility*, pages 16–31, 2019. 1, 2, 5, 8

[7] Patrick Buehler, Andrew Zisserman, and Mark Everingham. Learning sign language by watching tv (using weakly aligned subtitles). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2968. IEEE, 2009. 3

[8] Andrea Burns, Reuben Tan, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. Language Features Matter: Effective language representations for vision-language tasks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 7

[9] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018. 2, 8

[10] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033, 2020. 2, 6, 8

[11] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 2

[12] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3, 6, 7

[13] Branden Chan, Stefan Schweter, and Timo Möller. German's next language model. *ArXiv*, abs/2010.10906, 2020. 8

[14] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE, 2016. 1

[15] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 577–585, 2015. 1

[16] Helen Cooper and Richard Bowden. Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2568–2574. IEEE, 2009. 3

[17] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 2, 7

[18] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Word2visualvec: Image and video to sentence matching by visual feature prediction. *arXiv preprint arXiv:1604.06838*, 2016. 3

[19] Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, and Dima Damen. Action modifiers: Learning from adverbs in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 868–878, 2020. 2

[20] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multi-modal dataset for continuous american sign language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 6

[21] Abhishek Dutta and Andrew Zisserman. The via annotation software for images, audio and video. In *Proceedings of the 27th ACM international conference on multimedia*, 2019. 6

[22] Emma Flint, Elliot Ford, Olivia Thomas, Andrew Caines, and Paula Buttery. A text normalisation system for non-standard english words. In *NUT@EMNLP*, 2017. 4

[23] Andrea Frome, Greg S Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: a deep visual-semantic embedding model. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 2121–2129, 2013. 2

[24] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer, 2020. 2, 6

[25] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition, pages 12046–12055, 2019. 2

[26] Tao Jiang, Necati Cihan Camgoz, and Richard Bowden. Looking for the signs: Identifying isolated sign instances in continuous video footage. *IEEE International Conferene on Automatic Face and Gesture Recognition*, 2021. 2, 3

[27] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. 2

[28] Daniel Kelly, John Mc Donald, and Charles Markham. Weakly supervised training of a sign language recognition system using multiple instance learning density matrices. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(2):526–541, 2010. 3

[29] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020. 2

[30] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2306–2320, 2019. 3, 8

[31] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015. 3, 8

[32] Oscar Koller, Sepehr Zargaran, and Hermann Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4297–4305, 2017. 3

[33] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942, 2020. 7

[34] Lin-shan Lee, James Glass, Hung-yi Lee, and Chun-an Chan. Spoken content retrieval—beyond cascading speech recognition with text retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9), 2015. 1, 5

[35] François Lefebvre-Albaret and Patrice Dalle. Video retrieval in sign language videos : How to model and compare signs? In *LREC*, 2010. 3

[36] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020. 4, 5

[37] Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Ben Swift, Hanna Suominen, and Hongdong Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *arXiv preprint arXiv:2010.05468*, 2020. 2

[38] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6204–6213, 2020. 3

[39] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019. 2, 6

[40] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 2, 3, 4, 6, 7

[41] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 2

[42] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013. 7

[43] Niluthpol Chowdhury Mithun, Juncheng Billy Li, Florian Metze, and Amit K. Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018. 2

[44] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2019. 2

[45] Liliane Momeni, Triantafyllos Afouras, Themos Stafylakis, Samuel Albanie, and Andrew Zisserman. Seeing wake words: Audio-visual keyword spotting. *BMVC*, 2020. 4

[46] Liliane Momeni, Gül Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Watch, read and lookup: learning to spot signs from multiple supervisors. In *Asian Conference on Computer Vision (ACCV)*, 2020. 2, 3, 4, 5

[47] Eng-Jon Ong, Oscar Koller, Nicolas Pugeault, and Richard Bowden. Sign spotting using hierarchical sequential patterns with temporal intervals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1923–1930, 2014. 3

[48] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *International Conference on Learning Representations*, 2020. 2

[49] Tomas Pfister, James Charles, and Andrew Zisserman. Large-scale learning of sign language by watching tv (using co-occurrences). In *BMVC*, 2013. 3

[50] Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4165–4174, 2019. 3

[51] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 7

[52] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. In *OpenAI blog*, 2019. 7, 8

[53] Adam Schembri, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. Building the british sign language corpus. *Language Documentation & Conservation*, 7:136–154, 2013. 3

[54] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and A. Ng. Grounded compositional semantics for finding and describing images with sentences.

Transactions of the Association for Computational Linguistics, 2:207–218, 2014. 3

[55] Themos Stafylakis and Georgios Tzimiropoulos. Zero-shot keyword spotting for visual speech recognition in-the-wild. In *European Conference on Computer Vision (ECCV)*, 2018. 4

[56] William C Stokoe Jr. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37, 2005. 1

[57] Rachel Sutton-Spence and Bencie Woll. *The linguistics of British Sign Language: an introduction*. Cambridge University Press, 1999. 2, 4

[58] Nazif Can Tamer and Murat Saraçlar. Cross-lingual keyword search for sign language. In *LREC 2020*, 2020. 3

[59] Nazif Can Tamer and Murat Saraçlar. Improving keyword search performance in sign language with hand shape features. In *ECCV Workshops*, 2020. 3

[60] Nazif Can Tamer and Murat Saraçlar. Keyword search for sign language. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8184–8188, 2020. 3

[61] Hamid Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *The British Machine Vision Conference (BMVC)*, 2019. 4, 5

[62] Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Read and attend: Temporal localisation in sign language videos. In *CVPR*, 2021. 2, 3, 5

[63] Ville Viitaniemi, Tommi Jantunen, Leena Savolainen, Matti Karppa, and Jorma Laaksonen. S-pot–a benchmark in spotting signs within continuous signing. In *Proceedings of the 9th international conference on Language Resources and Evaluation (LREC 2014), ISBN 978-2-9517408-8-4*. European Language Resources Association (LREC), 2014. 3

[64] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 450–459, 2019. 2

[65] Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. Self-training and pre-training are complementary for speech recognition. In *ICASSP 2021- 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3030–3034. IEEE, 2021. 1

[66] Hee-Deok Yang, Stan Sclaroff, and Seong-Whan Lee. Sign language spotting with a threshold model based on conditional random fields. *IEEE transactions on pattern analysis and machine intelligence*, 31(7):1264–1277, 2008. 3

[67] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 2

[68] Shilin Zhang and Bo Zhang. Using hmm to sign language video retrieval. *2010 Second International Conference on Computational Intelligence and Natural Computing*, 1:55–59, 2010. 3

[69] Shilin Zhang and Bo Zhang. Using revised string edit distance to sign language video retrieval. *2010 Second International Conference on Computational Intelligence and Natural Computing*, 1:45–49, 2010. 3

[70] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*, 2020. 1