Sign Language Video Retrieval with Free-Form Textual Queries

Amanda Duarte\textsuperscript{1,2}  Samuel Albanie\textsuperscript{3}  Xavier Giró-i-Nieto\textsuperscript{1,4}  Gül Varol\textsuperscript{5}

\textsuperscript{1}Universitat Politècnica de Catalunya, Spain \textsuperscript{2}Barcelona Supercomputing Center, Spain \textsuperscript{3}Machine Intelligence Laboratory, University of Cambridge, UK \textsuperscript{4}Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Spain \textsuperscript{5}LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France

\url{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\textsuperscript{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 in the collection 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 \textsc{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 \textsc{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 \textit{manual} and \textit{non-manual markers} such as movements of the face, body and hands to convey information [54].

Recent developments in automatic speech recognition (ASR) for spoken languages [13,14,65,70] have enabled au-

\textsuperscript{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.

Figure 1. \textbf{Text-based sign language video retrieval}: In this work we introduce \textit{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 single keyword search.

Automatic captioning of vast swathes of video content hosted 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 [5]. 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 [32]. Unfortunately, while promising translation results have been
demonstrated in constrained domains of discourse (such as weather forecasts) [8,9,35], it has been widely observed that these systems are unable to achieve functional performance across the multiple domains of discourse [5,29,61] 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 [1,26,45] 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. Our focus is on sign language videos produced by and for signers, that do 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 [56], 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 [4,40]) 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 [10,25], 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. Our code, annotations and models will be made publicly available.

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 [4,16,19,24,38,39,42,43,47,64,67]. 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. [3] 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 retrieving 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 [33], 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. [46] 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 gestures among signing sequences [59]. This work was later extended to enable search for individual words [57] and further extended to additionally incorporate hand-shape features, improving performance [58]. More recently, Jiang et al. [26] showed the effectiveness of the transformer architecture for the sign spotting task, achieving promising results on the BSLCORPUS [51] and Phoenix2014 [30] 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 [15] 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 [6, 28, 48]. Other recent contributions leverage words from audio-aligned subtitles with keyword spotting methods based on mouthing cues [1], dictionaries [45] and attention maps generated by transformers [61] to annotate large numbers of signs, as well as to learn domain invariant features for improved sign recognition through joint training [36].

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, 45]. 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 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, 39], 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, $\phi_v$ and $\phi_t$, 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. [52]. Specifically, given paired samples $\{(v_n, t_n)\}_{n=1}^N$, we optimise a max-margin ranking loss:

$$
L = \frac{1}{B} \sum_{i=1}^B \left[ \eta_{ij} - \eta_{ii} + m \right]_+ + \left[ \eta_{ji} - \eta_{ii} + m \right]_+
$$

where $m$ denotes the margin hyperparameter, $[\cdot]_+$ denotes the hinge function $\max(\cdot, 0)$, $B$ denotes the size of minibatch sampled during training, and $\eta_{ij}$ 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, $\phi_v$ consists first of an initial sign video embedding, $\psi_v$, which we instantiate as a 1D neural network [11] over clips of 16 frames (motivated by its effectiveness for sign recognition tasks [1, 34, 60]). The output of $\psi_v$, 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 $\phi_T$, 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 [2] and projected via Gated Embedding Unit following the formulation of [39] 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, $\psi_v$, 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 $\phi_v$, and the language model employed by $\phi_T$.

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 SPOTALIGN framework (Fig. 2a) which we employ to obtain large numbers of automatic sign annotations on the How2Sign dataset. This dataset provides videos with corresponding written English translations but currently lacks individual sign 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 [45]. 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 mouthing in addition to head movements and manual gestures [56], employs the keyword spotting architecture of [53] with the improved P2G phone-to-grapheme keyword encoder proposed by Momeni et al. [44]. 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 [34]
or MSASL [60] 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 [45].** Next, we employ an exemplar-based sign spotting method similar to [45]. 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 [45], 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 $\psi_v$, 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 [34] and MSASL [60], 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 [61].

**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 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 $v_\psi$, 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, $\phi_v$, 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 $v_\psi$ 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 [32], except that sign recognition is significantly more difficult than speech recognition (in part, due to a lack of training data [3]). 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, welem-
matise both the query words and predicted words. We con-
strain 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 perform-
worse than the cross-modal retrieval approach, is com-
plementary 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 pro-
vide retrieval results on How2Sign dataset, conducting abla-
tion studies to evaluate the influence of different compo-
nents of our approach (Sec. 4.2). Then, we establish base-
line retrieval performances on the PHOENIX2014T dataset
(Sec. 4.3). Finally, we discuss limitations together with
qualitative analysis (Sec. 4.4), as well as societal impact
(Sec. 4.5).

4.1. Data, annotation and evaluation protocols

Datasets. In this work, we primarily focus on the re-
cently released How2Sign dataset [20], a multimodal open-
vocabulary and subtitled dataset of about 80 hours of con-
tinuous 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 dura-
tion (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 [9], (al-
though 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 train-
ning, 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 re-
stricted 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 re-
trieval by sign language translation [9].

Annotations. For sign recognition, we train using the au-
tomatic 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 con-
secutive iterations. To enable evaluation of sign recogni-
tion 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 2212 individual sign
video-category pairs, which will be made available.

Evaluation metrics. To evaluate retrieval performance,
we follow the existing retrieval literature [24, 38, 39] and report
standard metrics R@K (recall at rank K, higher is better)
and MedR (median rank, lower is better). For sign recogni-
tion baselines (for which the time required to retrain ψv
is more extensive), we report the results of a single model.
For cross modul retrieval ablations (for which the sign
video embedding ψv is frozen, and only the text encoder,
φv, and video encoder, φv, are trained), we report the mean
and standard deviation over three randomly seeded runs.

![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.](image-url)

![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.](table-url)
is your clay.”

you want to start

and first thing

to make some

"OK, we’re going

having fun.”

I hope you’re

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.

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 results 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 [11] 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 [61] (pretrained on British Sign Language (BSL) data [1]) that was used to initialise our I3D sign video embedding. While strongly outperforming kinetics training, this model remains substantially weaker than the end-to-end ASL sign recognition training enabled by SPOT-ALIGN. We notice substantial 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+D3 annotations from the 1887 large-vocabulary for the rest of our experiments on How2Sign.

(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 [16, 39]. 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 [41] and GrOVLE [7] obtain competitive performance, outperforming higher capacity alternatives [31, 49, 50]. This phenomenon is also observed in [16], 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 trans-
Table 4. Thresholding sign recognition probabilities: We investigate the influence of the confidence threshold hyperparameter on How2Sign retrieval performance. We observe that a threshold of 0.5 works best.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>R@1↑</th>
<th>R@5↑</th>
<th>R@10↑</th>
<th>MedR↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>13.0</td>
<td>25.9</td>
<td>31.5</td>
<td>74.3</td>
</tr>
<tr>
<td>0.1</td>
<td>13.3</td>
<td>26.5</td>
<td>31.8</td>
<td>72.3</td>
</tr>
<tr>
<td>0.25</td>
<td>17.3</td>
<td>30.1</td>
<td>35.5</td>
<td>58.5</td>
</tr>
<tr>
<td>0.50</td>
<td>18.9</td>
<td>32.1</td>
<td>36.5</td>
<td>62.0</td>
</tr>
<tr>
<td>0.75</td>
<td>14.8</td>
<td>27.0</td>
<td>31.3</td>
<td>103.5</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Sign Recognition</th>
<th>Cross-Modal Embeddings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment</td>
<td>R@1↑</td>
</tr>
<tr>
<td>Speech</td>
<td>9.5</td>
</tr>
<tr>
<td>Signing</td>
<td>18.5</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Models</th>
<th>R@1↑</th>
<th>R@5↑</th>
<th>R@10↑</th>
<th>MedR↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR</td>
<td>18.9</td>
<td>32.1</td>
<td>36.5</td>
<td>62.0</td>
</tr>
<tr>
<td>CM</td>
<td>24.3</td>
<td>40.7</td>
<td>46.5</td>
<td>16.0</td>
</tr>
<tr>
<td>SR + CM</td>
<td>34.2</td>
<td>48.0</td>
<td>52.6</td>
<td>8.0</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Text Embedding</th>
<th>R@1↑</th>
<th>R@5↑</th>
<th>R@10↑</th>
<th>MedR↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation</td>
<td>30.2</td>
<td>53.1</td>
<td>63.4</td>
<td>4.5</td>
</tr>
<tr>
<td>Cross-modal</td>
<td>48.6</td>
<td>76.5</td>
<td>84.6</td>
<td>2.0</td>
</tr>
<tr>
<td>Combination</td>
<td>55.8</td>
<td>79.6</td>
<td>87.2</td>
<td>1.0</td>
</tr>
</tbody>
</table>

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 [8, 30]. For cross-modal embedding training, we employ a text embedding model trained on German language corpora, GPT-2 [50] released by Chan et al. [12]. For text-based retrieval, here we incorporate a state-of-the-art sign language translation model [9], 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. 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

One limitation of SPOT-ALIGN method 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.

4.5. 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 increased 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 SPOTALIGN framework to obtain automatic annotations, and demonstrated their value in producing effective sign video embeddings for retrieval.

Acknowledgements. The authors would like to thank M. Fischetti, C. Marsh and M. Dippold for their work on data annotation and also A. Dutta, A. Thandavan, J. Pujal, L. Ventura, E. Vincent, L. Tarres, P. Cabot, C. Punti and Y. Kalantidis for their help and valuable feedback. This work was supported in part by the project PID2020-117142GB-I00, funded by MCIN/AEI 10.13039/501100011033, as well as research gifts from Google and Adobe. A. Duarte has received support from la Caixa Foundation (ID 100010434) under the fellowship code LCF/BQ/IN18/11660029.

References


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


[37] Liyuan Liu, Haoaming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. International Conference on Learning Representations, 2019. 13

[38] 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


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


APPENDIX

This appendix provides additional qualitative analyses (Sec. A), implementation details (Sec. B), dataset details
(Sec. C), and an experiment demonstrating challenges of using text-based retrieval via sign language translation
(Sec. D). We further provide a supplemental webpage on the project page, described next.

A. Qualitative Analysis

Supplemental webpage. We qualitatively illustrate, with the supplemental webpage (https://imatge-upc.github.io/sl_retrieval/app-qualitative/index.html), our retrieval results using the best model on the How2Sign dataset (SR+CM combination from Tab. 6). For each query, we show the top three ranked videos as well as their corresponding topic category (see [20] for more details of video topic categories), signer ID and sentences (note that these are not used during retrieval, and are provided for visualisation purposes).

The top ten rows of the html page show cases in which our model is able to correctly retrieve the video corresponding to the textual query. The middle five rows of the html page show cases where the correct video is not retrieved successfully. For these failures, we nevertheless observe that the retrieval model makes reasonable mistakes (for instance, in the majority of cases, at least one of the top three ranked videos share the same topic category of the GT video). In the bottom five rows, we show examples of failure cases of our model.

Combination of cross-modal and sign recognition. We noticed that our strongest retrieval model combines similarities from the cross-modal embeddings and the sign recognition model (SR+CM combination from Tab. 6). In Fig. A.2, we illustrate two example queries for which the use of the sign recognition model substantially improves the performance of the cross-modal embeddings.

B. Implementation Details

In this section, we provide additional implementation details for the sign video embedding (Sec. B.1), text embedding (Sec. B.2) and cross-modal retrieval training (Sec. B.3).

B.1. Sign recognition and sign video embedding

Sign recognition training. As explained in Sec. 3.4 of the main paper, we train a sign recognition model, a 3D convolutional neural network instantiated with an I3D [11] architecture pretrained on BSL-1K [1, 61]. We finetune this model on the How2Sign dataset using our automatic sign spotting annotations. In the final setting with mouthing (M) and dictionary (D₃) spottings from a vocabulary of 1887 signs, we have 206K training video clips, each corresponding to a single sign. Since the spottings represent a point in time, rather than a segment with beginning-end times, we determine a fixed window for each video clip. For mouthing annotations, this window is defined as 15 frames before the annotation time and 4 frames after ([-15, 4]). For dictionary annotations, the window is similarly set to [-3, 22]. During training, we randomly sample 16 consecutive frames from this window, such that the RGB video input to the network becomes of dimension $16 \times 3 \times 224 \times 224$. We apply a similar spatial cropping randomly from 256 × 256 resolution. We further employ augmentations such as colour jittering, resizing and horizontal flipping.

We perform a total of 25 epochs on the training data, starting with a learning rate of 1e-2, reduced by a factor of 10 at epoch 20. We optimise using SGD with momentum (with a value of 0.9) and a minibatch of size 4.

At test time, for recognition, we apply a sliding window averaging in time, and center cropping in space. At test time, for text-based retrieval, we obtain the predicted class per 16-frame sliding window (with a stride of 1 frame), and record the corresponding word out of the 1887-vocabulary if the probability is above the 0.5 threshold. The resulting set of words are merged in case of repetitions, and are compared against the queried text to obtain an intersection over union (IOU) score, used as the similarity.

Sign video embedding. As noted in Sec. 3.2 of the main paper, we employ the I3D recognition model (described above) to instantiate our sign video embedding. More specifically, we use the outputs corresponding to the spatio-temporally pooled vector before the last (classification) layer. This produces a 1024-dimensional real-valued vector for each 16 consecutive RGB frames. We extract these features densely with a stride 1 from How2Sign sign language sentences to obtain the sequence of sign video embeddings.

B.2. Text embedding

We consider several text embeddings in this work. When conducting experiments on the How2Sign dataset, we explore the following English language embeddings:

GPT [49] is a 768-dimensional embedding that uses a Transformer decoder which is trained on the BookCorpus [71] dataset.

GPT-2-xl [50] is a 1600-dimensional embedding (employing 1558M parameters, also in a Transformer architecture [62]) that is trained on the WebText corpus (containing millions of pages of web text).

Albert-XL [31] is a 2048-dimensional embedding that builds on BERT [17] to increase its efficiency. It is trained with a loss that models inter-sentence coherence on the BookCorpus [71] and Wikipedia [17] datasets.

W2V [41] is a 300-dimensional word embedding,
trained on the Google News corpus (we use the GoogleNews-vectors-negative300.bin.gz model from https://code.google.com/archive/p/word2vec/).

GroVLE [7]. This is a 300-dimensional embedding that aims to be vision-centric: it is adapted from Word2Vec [41].

For experiments on the PHOENIX2014T dataset, we use a German language model:

German GPT-2 [12] (based on the original GPT-2 architecture of [49]) is a 768-dimensional embedding. The model is trained on the OSCAR [55] corpus, together with a blend of smaller German language data. We use the parameters made available at https://huggingface.co/dbmdz/german-gpt2.

B.3. Cross-modal retrieval

The dimensionality of the shared embedding space (denoted by the variable $C$ in Sec. 3.2) used in this work is 512. The margin hyperparameter, $m$, introduced in Eqn. 1, is set to 0.2, following [39]. All cross modal embeddings are trained for 40 epochs using the RAdam optimiser [37] with a learning rate of 0.001, a weight decay of $1E−5$ and a batch size of 128. For each experiment, the epoch achieving the highest geometric mean of $R@1$, $R@5$ and $R@10$ on the validation set was used to select the final model for test set evaluation. The NetVLAD [2] layer employed in the text encoder uses 20 clusters. Sign video embeddings (which form the input to the video encoder, $\phi_V$ described in Sec. 3.2) are extracted densely (i.e. with a temporal stride 1).

C. Dataset Details

To construct training, validation and test retrieval partitions from the How2Sign dataset, we select video segments with their corresponding manually aligned subtitles (released by the authors of [20]). This provides an initial pool of 31,164 training, 1,740 validation and 2,356 test videos with corresponding translations. After initial inspection, we found that while most annotations were produced to a high quality, a small number of the manually aligned subtitles were invalid (i.e. exhibited no temporal overlap with the video). We excluded these invalid subtitles from our retrieval benchmark, producing final splits of: 31,075 training, 1,739 validation and 2,348 test videos.

In Fig. A.1, we visualise the difference between the timings of the original subtitles versus the manually aligned subtitles. We note that the signing is on average behind the speech, constituting a misalignment when using the original subtitle timings. This misalignment explains the performance drop we demonstrated in Tab. 5 of the main paper when experimenting with the original subtitles instead of the manually aligned ones.

D. Text-based Retrieval Attempt through Sign Language Translation

As mentioned in Sec. 1 of the main paper, a text-based retrieval solution using sign language translation on videos is not a viable option due to unsatisfactory video-to-text translation performance of current state-of-the-art models [9] on open-vocabulary domains. Here, we provide a brief justification by training the encoder-decoder Transformer model of [9] on the How2Sign dataset using the same sign video embeddings as in our cross-modal retrieval setting, i.e. the densely extracted I3D features. We keep all the hyperparameters identical to the publicly available setting of [9] and obtain poor BLEU scores of 1.74 and 17.08 for BLEU-4 and BLEU-1, respectively. We provide in the project webpage (https://imatge-upc.github.io/sl_retrieval/app-translation/index.html) the ground truth (left) and the predicted (right) sentences on the validation set and observe that the predictions tend to be generic sentences that do not correspond to the input sign language video, with the exception that sometimes the model predicts one word right out of the entire sentence.
Figure A.2. **Qualitative results**: We show two samples where text-based retrieval using sign recognition (SR) helps retrieval when combined with cross-modal embeddings (CM). Top, middle and bottom rows show the retrieval results for the same query using the average of the similarities from SR and CM (Combination), Cross-Modal and Sign Recognition models, respectively.