ECCV 2022 Sign Spotting Challenge: Dataset, Design and Results

Manuel Vázquez Enríquez[1](✉), José L. Alba Castro[1], Laura Docio Fernandez[1], Julio C. S. Jacques Junior[2], and Sergio Escalera[2,3]

[1] University of Vigo, Vigo, Spain
{mvazquez,jalba,ldocio}@gts.uvigo.es
[2] Computer Vision Center, Bellaterra, Spain
jjacques@cvc.uab.cat
[3] University of Barcelona, Barcelona, Spain
sergio@maia.ub.es

Abstract. The ECCV 2022 Sign Spotting Challenge focused on the problem of fine-grain sign spotting for continuous sign language recognition. We have released and made publicly available a new dataset of Spanish sign language of around 10 h of video data in the health domain performed by 7 deaf people and 3 interpreters. The added value of this dataset over existing ones is the frame-level precise annotation of 100 signs with their corresponding glosses and variants made by sign language experts. This paper summarizes the design and results of the challenge, which attracted 79 participants, contextualizing the problem and defining the dataset, protocols and baseline models, as well as discussing top-winning solutions and future directions on the topic.

Keywords: Sign spotting · Sign language · 3DCNN · ST-GCN

1 Introduction

In recent years, the scientific community has accelerated research in automatic Sign Language Recognition (SLR), supported by the latest advances in deep learning models with flexible spatial-temporal representation capacities [27]. It is well known that these methods need to be fed with huge amounts of data and this is one of the main drawbacks for a faster progress of automatic SLR. Sign languages are purely visual languages lacking a fully accepted writing system. Transcribing the visual information of continuous signing to glosses or any of the few alternative codes (e.g., HamNoSys [26]) can only be made by few specialists and it is a extremely time-consuming task.

SLR can be roughly classified into Isolated (ISLR) and Continuous (CSLR) sign language recognition. ISLR is the most studied scenario by the scientific community and for which more annotated datasets are available, because of the relative easy annotation protocols of a discrete and predefined set of signs. CSLR, however, is much more complex due to three main factors. First, the co-articulation between signs drastically increases the variability of the sign realization with respect to isolated signs and reflects the signing style of each person.

L. Karlinsky et al. (Eds.): ECCV 2022 Workshops, LNCS 13808, pp. 225–242, 2023.
https://doi.org/10.1007/978-3-031-25085-9_13

Second, the speed is much higher than in the examples of isolated signs. Another point is the interplay of non-manual components: torso, head, eyebrows, eyes, mouth, lips, and even tongue [9]. The third factor is the already mentioned shortage of experts capable of annotating signs in sentences.

The ultimate goal consist in Sign Language Translation (SLT) from a Sign Language (SL) to a spoken language. Therefore, available corpora composed of a visual sign language and the corresponding transcription in glosses are scarce. Recent works [1,4,36] are leveraging captioned and signed broadcast media to develop direct SLT without passing explicitly through any intermediate transcribed representation. Nevertheless, in those restricted-domain cases where the transcription to glosses is available, either decoupling SLR to glosses and translation from glosses to spoken language, or using the glosses as a training guidance for the end-to-end translation process yields better performance [5].

Sign Spotting is a special case of CSLR where the specific grammar of each SL is not taken into account but only delimiting the localization of a particular sign. It allows the development of a wide number of applications such as indexing SL content, enabling efficient search and "intelligent fast-forward" to topics of interest, helping to linguistic studies or even learning sign language. Sign spotting can also be reliably used for collecting co-articulated samples of the query sign to improve an ISLR model without the need of extensive expert annotations [20].

2 Related Work

Early works on sign/gesture localization in video were supported by time-series pattern spotting techniques like Dynamic Time Warping [35], Global Alignment Kernel [25], Hidden Markov Models [2], Conditional Random Fields [39] or Hierarchical Sequential interval Pattern Trees [24] to build models over hand-crafted features usually directed by linguistic knowledge. Most of these approaches were tested over ad-hoc datasets of isolated signs in lab conditions or continuous sign language with limited variability of signers or annotated signs. The statistical approach used in speech recognition [18] was adapted to SL and tested on the "new" RWTH-Phoenix-Weather database [12], which set a milestone in CSLR.

With the release of the RWTH-Phoenix-Weather database of German continuous SL and the advent of deep learning, CSLR started to attract the attention of many research groups. Cui et al. [10] proposed the first CSLR system completely trained with deep networks, based on weakly supervision and end-to-end training. They used a convolutional neural network (CNN) with temporal convolution and pooling for spatio-temporal representation learning, and a RNN model with a long short-term memory (LSTM) module to learn the mapping of feature sequences to sequences of glosses. The trend shifted to learn short-term spatial-temporal features of signs directly from the RGB(+D) sequences or modalities derived from it (e.g., optical flow or skeletal data), mostly pushed by the success of 3D Convolutional Networks and the two-stream Inflated 3D ConvNet (I3D model) used in action classification [7]. Spatio-temporal Graph Convolutional Networks (ST-GCNs) [38], inherited from the action recognition field, also helped to stimulate research on ISLR and CSLR, using skeleton data from RGB body sequences [6,8] to learn spatio-temporal patterns.

Sign spotting has benefited from the advances in spatial-temporal representation and sequence learning from weakly labeled signs. Sign spotting has a specific domain mismatch problem between the query (usually isolated signs) and target (co-articulated signs). Jiang et al. [16] designed a one-shot sign spotting approach using I3D for learning the spatial-temporal representation to address the temporal scale discrepancies between the query and the target videos, building multiple queries from a single video using different frame-level strides. They proposed a transformer-based network that exploits self attention and mutual attention between the query and target features to learn the self and mutual temporal dependency. Their results on the BSLCORPUS [30] (continuous) using SignBank [11] (isolated) as queries outperformed previous approaches. The recently released BSL-1K [1] dataset, with weakly-aligned subtitles from broadcast footage and a vocabulary of 1000 signs automatically located in 1000 h hours of video allows training larger CSLR models and testing sign spotting techniques. In [34], the authors proposed an approach to one/few shot sign spotting through three supervision cues: mouthing, subtitled words and isolated sign dictionary. They used a unified learning framework using the principles of Noise Contrastive Estimation and Multiple Instance Learning [22], which allows the learning of representations from weakly-aligned subtitles while exploiting sparse labels from mouthings [1] and explicitly accounts for sign variations. The approach is based on I3D and was evaluated for sign spotting on BSL-1K using the isolated signs dataset collected by the same research group, BSLDICT [23].

3 Challenge Design

The ECCV 2022 Sign Spotting Challenge[1] was aimed to attract attention on the strengths and limitations of the existing approaches, and help to define the future directions of the field. It was divided into two competition tracks:

- **MSSL (multiple shot supervised learning).** MSSL is a classical machine learning track where signs to be spotted are the same in training, validation and test sets. The three sets contain samples of signs cropped from the continuous stream of Spanish sign language, meaning that all of them have co-articulation influence. The training set contains the begin-end timestamps (in milisecs) annotated by a deaf person and a SL-interpreter (with an homogeneous criteria) of multiple instances for each of the query signs. Participants needed to spot those signs in a set of test videos.
- **OSLWL (one shot learning and weak labels).** OSLWL is a realistic variation of a one-shot learning adapted to the sign language specific problem, where it is relatively easy to obtain a couple of examples of a sign using just a sign language dictionary, but it is much more difficult to find co-articulated

[1] Challenge - https://chalearnlap.cvc.uab.cat/challenge/49/description/.

versions of that specific sign. When subtitles are available, as in broadcast-based datasets, the typical approach consists in using the text to predict a likely interval where the sign might be performed. In this track, we simulated that case by providing a set of queries (20 isolated signs performed by a signer of the dataset and 20 by an external one) and a set of 4 sec video intervals around each and every co-articulated instance of the queries. Intervals with no instances of the queries were provided to simulate the typical case where the subtitle or translated text shows a word that the signer selected not to perform or even a subtitle error. Participants needed to spot the exact location of the sign instances in the provided video intervals. In this track, only one sign needs to be located for each video.

The participants were free to join any of these challenge tracks. Each track was composed of two phases, i.e., development and test phase. At the development phase, public train data was released and participants needed to submit their predictions with respect to a validation set. At the test phase, participants needed to submit their results with respect to the test data. Participants were ranked, at the end of the challenge, using the test data. Note that this competition involved the submission of results (and not code). Therefore, participants were required to share their codes and trained models after the end of the challenge so that the organizers could reproduce their results in a "code verification stage". At the end of the challenge, top ranked methods that passed the code verification stage (discussed in Sect. 4) were considered as valid submissions.

3.1 The Dataset

LSE_eSaude_UVIGO, released for the ECCV 2022 Sign Spotting Challenge, is a dataset of Spanish Sign Language (LSE: "Lengua de Signos Española") in the health domain (around 10 h of video data), signed by 10 people (7 deaf and 3 interpreters) partially annotated with the exact location of 100 signs. This dataset has been collected under different conditions than the typical Continuous Sign Language datasets, which are mainly based on subtitled broadcast and real-time translation. In our case, the signs are performed to explain health contents by translating printed cards, so reliance on signers is large due to the richer expressivity and naturalness. The dataset was acquired in studio conditions with blue chroma-key, no shadow effects and uniform illumination, at 25 FPS and a resolution of 1080×720. The added value of the dataset is the rich and rigorous hand-made annotations. Experts interpreters and deaf people were in charge of annotating the location of the selected signs. Additional details about the dataset, data split and annotation protocol can be found in the dataset webpage[2].

The signers in the test set can be the same or different to the training and validation set. Signers are men, women, right and left-handed. The amount of samples per sign was not uniform, as some signs are common terms but others are related to a specific health topic. The total number of hand annotations

[2] Dataset - https://chalearnlap.cvc.uab.cat/dataset/42/description/.

are 4822 in MSSL track and 1921 in OSLWL track. The duration of the co-articulated signs is not a Gaussian variable, starting from 120 ms (3 frames) up to 2 s (50 frames), with a mean duration of 520 ms (13 frames).

3.2 Evaluation Protocol

The evaluation protocol is the same for both tracks, based on the F1 score. Matching score per sign instance is evaluated as the Intersection over Unit (IoU) of the ground-truth interval and the predicted interval. In order to allow relaxed locations, the IoU threshold is swept from $t = 0.2$ to 0.8 using 0.05 as step size.

Given a specific video file n and query sign i to spot, the submitted solutions are evaluated as follows. First, let us define $Ret_i(k, n)$ as the k^{th} interval in video file n retrieved as a prediction of sign i, and $Rel_i(n)$ an interval annotated for an instance of sign i. Then, $IoU_i(k, n)$ is obtained as:

$$IoU_i(k,n) = \frac{Ret_i(k,n) \cap Rel_i(n)}{Ret_i(k,n) \cup Rel_i(n)}. \tag{1}$$

A True Positive (TP) occurs if $IoU_i(k, n) \geq t$. Note that IoU is calculated only between intervals with prediction and ground-truth matching. Moreover, no overlap is allowed among Ret_i intervals. Then, for all $Ret_i(k, n)$ reported from each and every video file n containing L_n ground-truth instances, we compute $TP(t) = \sum_n \sum_i \sum_k (IoU_i(k, n) \geq t)$, $FP(t) = (\sum_n \sum_i \sum_k 1) - TP(t)$, and $FN(t) = (\sum_n L_n) - TP(t)$ for each IoU threshold t. Then, the accumulated over all t values as $TP = \sum_t TP(t)$, $FP = \sum_t FP(t)$, and $FN = \sum_t FN(t)$.

Precision (P) and Recall (R) averages the amounts from different thresholds as $P = TP/(TP + FP)$ and $R = TP/(TP + FN)$. Finally, the F1 score is obtained using in Eq. 2.

$$F1 = 2 * (P * R)/(P + R). \tag{2}$$

3.3 The Baseline

The baseline model is similar for both tracks, with a common pipeline and two branches that take into account the requirements of MSSL and OSLWL tracks. First, a person detection model [28] locates the position of the signer and a 512 × 512 ROI is extracted from the original footage. Then, the Mediapipe Holistic keypoint detector [13] extracts $11 + 21 \times 2 = 53$ coordinate points each 40ms. The core of the model is a multi-scale ST-GCN, MSG3D [21], trained with upper body and hands skeleton [37]. Two models are independently trained and averaged, one with coordinates as input features (joints), and the other with distances between connected coordinates (bones). No explicit RGB or motion information is used for spatial-temporal modeling. The model was pre-trained on the AUTSL [31] dataset, fine-tuned using 3 signers of MSSL train set and validated over the other 2. Ground-truth annotations of the 60 class-signs were used with 2 different context windows: 400 ms to train a short-context model and 1000 ms to train a long-context model. Small random shifts around the ground-truth interval in the training window allowed data augmentation.

At inference stage, the short- and long-context models are applied in a sequential decision pipeline. The short-context model is applied first with a stride of one video-frame (40 ms). It yields one class decision per frame if the wining class surpasses a threshold of 0.75. As this output is noisy, a non-linear filter designed as a morphological operator eliminates potential short-time false positives and false negatives by building convex-like sign outputs. The isolated candidate signs are then fed to the long-context model in charge of eliminating false positive candidates with a stricter threshold of 0.8. Figure 1 shows the main blocks of the pipeline for the MSSL (upper branch) and OSLWL (lower branch) tracks.

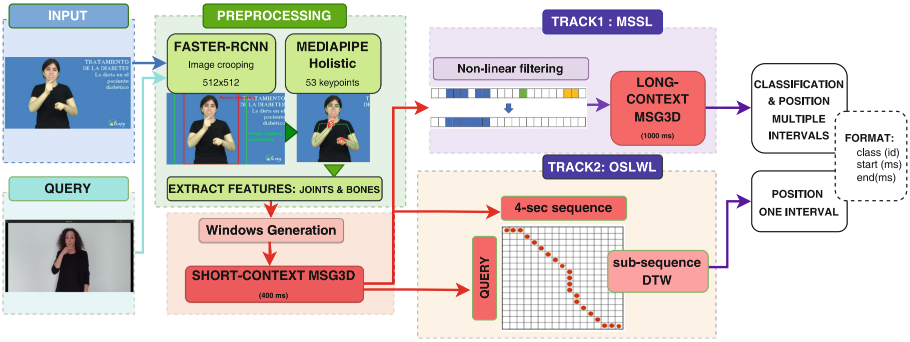


Fig. 1. Pipeline of the baseline solution for MSSL and OSLWL tracks.

The solution for OSLWL track had no training at all. The inference pipeline shares all the blocks with MSSL up to the short-context model. At this point, the baseline extracts an internal embedding vector of 384D just before the last FC layer. For every 4 s interval a specific sign can appear at any place (or not appear at all), so a sub-sequence Dynamic Time Warping (DTW) is applied with cosine distance between the specific query of the searched sign and the 4 s 384D-sequence. A fixed threshold of 0.2 was set for discarding low-valued matching scores. So, the result of the sub-sequence DTW is the time-interval of the spotted sign only if the score surpasses the threshold.

4 Challenge Results and Winning Methods

The challenge ran from 21 April 2022 to 24 June 2022 through Codalab[3], a powerful open source framework for running competitions. It attracted a total of 79 participants, 37 in MSSL track and 42 in OSLWL track, suggesting where the research community is paying more attention, given the two challenge tracks.

[3] Codalab - https://codalab.lisn.upsaclay.fr.

4.1 The Leaderboard

The leaderboard at the test phase of both tracks, for the submissions that passed the code verification and improved the baseline scores, are shown in Table 1, ranked by average F1 using multiple IOU thresholds from 0.2 to 0.8 with a stride 0.05. As it can be observed, the top-3 winning methods improved the baseline scores on both tracks by a large margin.

Table 1. Leaderboard of MSSL and OSLWL tracks at the test phase.

MSSL		
Rank	Participant	Avg F1
1	ryanwong	0.606554
2	th	0.566752
3	Random_guess	0.564260
4	Baseline	0.300123

OSLWL		
Rank	Participant	Avg F1
1	th	0.595802
2	Mikedddd	0.559295
3	ryanwong	0.514309
4	Baseline	0.395083

Table 2 shows the scores for different thresholds where the performance on the relaxed to strict localization requirements can be observed. It is interesting to highlight three points: i) the top-3 winning methods surpassed the baseline regardless the IoU threshold; ii) the reduction of F1 when requiring strict spatial spotting; iii) the *th* team would win both tracks under the stricter IoU threshold.

Table 2. IoU score for different threshold values (0.2, 0.5, 0.8). Bold values highlight the overall winning method. In italic, when swapped leadeboard position.

MSSL			
Participant	1@20	F1@50	F1@80
ryanwong	**0.744**	**0.677**	0.280
th	0.660	0.626	**0.282**
Random_guess	0.715	0.652	0.204
Baseline	0.465	0.339	0.056

OSLWL			
Participant	F1@20	F1@50	F1@80
th	0.784	**0.647**	**0.269**
Mikedddd	**0.809**	0.594	0.160
ryanwong	0.744	0.552	*0.164*
Baseline	0.621	0.437	0.080

Table 3 shows general information about the top-3 winning methods. As can be seen, common strategies employed by top-winning solutions are the use of pre-trained models (most of them for feature extraction, as detailed in the next section), some face, hand, body detection, alignment or segmentation strategy, combined with pose estimation and/or spatio-temporal information modeling.

Table 3. General information about the top-3 winning approaches.

Track	MSSL			**OSLWL**		
Rank	1	2	3	1	2	3
Participant	**ryanwong**	**th**	**Random_guess**	**th**	**Mikedddd**	**ryanwong**
Pre-trained models	✓	✓	✓	✓	✓	✓
External data	✗	✗	✗	✗	✗	✗
Any kind of depth information (e.g., 3D pose estimation)	✗	✗	✓	✗	✓	✗
Use of validation set as part of the training data (at test stage)	✓	✓	✓	✗	✓	✗
Handcrafted features	✗	✗	✗	✗	✗	✗
Face / hand / body detection, alignment or segmentation	✓	✓	✗	✓	✓	✓
Use of any pose estimation method	✗	✓	✓	✓	✓	✗
Spatio-temporal feature extraction	✓	✓	✓	✓	✓	✓
Explicitly classify any attribute (e.g., gender/handedness)	✗	✗	✗	✗	✗	✗
Bias mitigation technique (e.g. rebalancing training data)	✓	✗	✓	✗	✗	✗

Next, we briefly introduce the top winning methods based on the information provided by the authors. For a detailed information, we refer the reader to the associated fact sheets, available for download in the challenge webpage(See footnote 4).

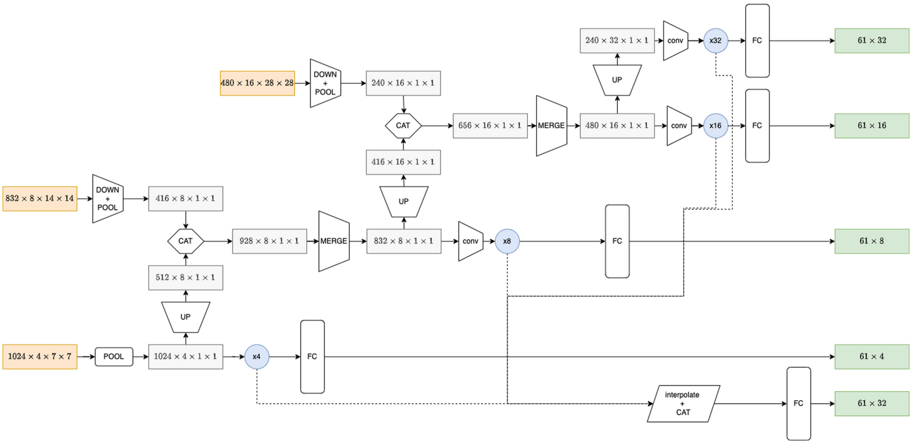
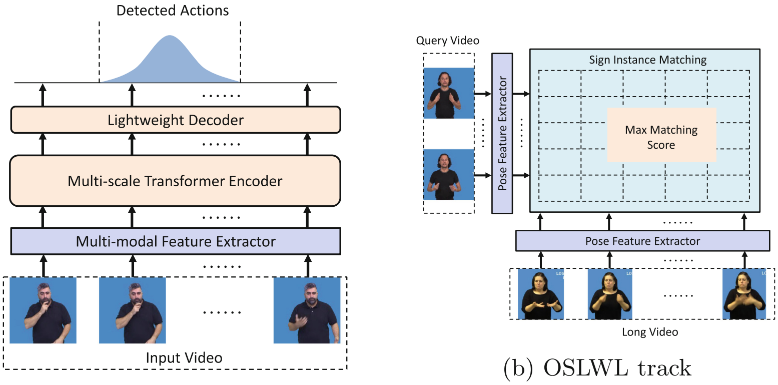


Fig. 2. Top-1, MSSL track (*ryanwong* team): proposed pipeline.

4.2 Top Winning Approaches: MSSL Track

Top-1: *ryanwong* team. The *ryanwong* team proposed to modify and use an I3D [7] model from [1], pretrained on WLASL [19] dataset. Originally, the I3D model outputs a single sign prediction (i.e., for a given region in time), given a sign video sequence of 32 frames, which limits the network to coarse temporal predictions. Thus, they proposed hierarchical I3D model, which can predict signs at frame level. Instead of taking the output at the final layer with spatial temporal global average pooling and applying a FC layer for class predictions, they take the output before global average pooling and additional feature outputs before the 3D max pool layers in the I3D model, obtaining 3 feature outputs each with a higher temporal resolution. For a given sequence length of 32 frames of dimensions 224 × 224, the base I3D model outputs the following features $1024 \times 4 \times 7 \times 7$, $832 \times 8 \times 14 \times 14$ and , $480 \times 16 \times 28 \times 28$, with a temporal resolution of 4, 8 and 16, respectively. The proposed network, illustrated in Fig. 2, uses these inputs to output coarse-to-fine temporal predictions ranging from 4, 8, 16 and 32 temporally aligned predictions. That is, making 1 prediction every 8, 4, 2 and 1 frame(s), respectively. Cross Entropy loss is used to predict the sign at each time segment for the coarse-to-fine predictions. A trade off between precision and recall is obtained with different random sampling probabilities were instead of selecting only frame regions around only known sign classes, they randomly select frame regions from other areas. The final predictions are based on temporally interpolating the softmax of the logit features for each of the predictions to the original sequence length (32) and averaging the 5 output results to obtain the probabilities for each class prediction at frame level.

Top-2: *th* team. The *th* team proposed a two-stage framework, which consists of feature extraction and temporal sign action localization, illustrated in Fig. 3a.



(a) MSSL track

(b) OSLWL track

Fig. 3. a) Top-2 and b) Top-1 winning solutions on MSSL and OSLWL track, respectively, both from *th* team.

First, multiple modalities (RGB, optical flow [33] and pose) are used to extract robust spatio-temporal feature representation. Then, a Transformer backbone is adopted to identify actions in time and recognize their categories. As pre-processing, MMDetection [8] is employed to detect the signer spatial location, followed by MMPose [6] (from OpenMMLab project) for extracting body and hand poses, used to crop the upper-body patch of the signer. Finally, three types of data are generated, one for each modality. Different backbones are used for feature extraction, given their complementary representations. RGB modality is fed into the BSL-1k [1] pre-trained I3D [7]. Flow modality is processed by the AUTSL [31] pre-trained I3D [7]. For the pose modality, pre-trained GCN [3,14] is used to extract body and hand cues. The extracted features are concatenated for the next localization stage, where Transformer is used to perform localization similar to [40]. Specifically, it combines a multiscale feature representation with local self-attention, using a light-weighted decoder to classify every moment in time and to estimate the corresponding action boundaries. During training, focal loss for sign action classification and generalized IoU loss for distance regression are adopted. At the inference stage, the output of every time step is taken, organized as the triplet sign action confidence score, onset and offset of the action. These candidates are further processed via NMS to remove highly overlapping instances, which leads to the final localization outputs.

Top-3: *Random_guess* team. The *Random_guess* team proposed a two-stage pipeline for sign spotting. The aim of the first stage is to condense sign language relevant information from multiple domain experts into a compact representation for sign spotting, while the goal of the second stage is to spot signs from longer video with a more powerful temporal module. Their workflow diagram is presented in Fig. 4. At the inference stage, the processed video clips (cropped video, masked video, optical flow, and 3D skeleton) are fed into four trained expert modules. The extracted features of each clip are concatenated into a vector to represent sign information for spotting. The temporal module takes vectors as input and extracts contextual information. It is composed of a two-

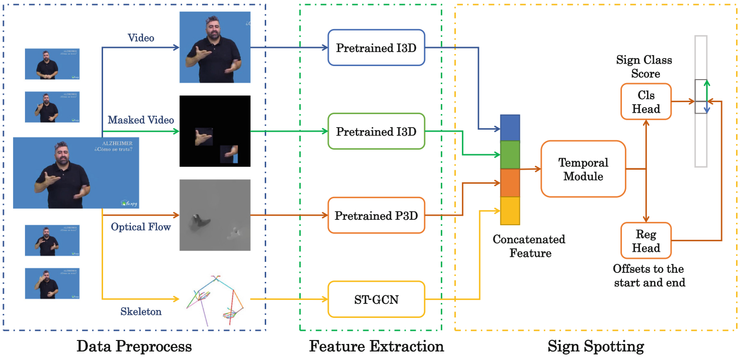


Fig. 4. Top-3, MSSL track (*Random_guess* team): workflow diagram.

layer 1D convolutional and a two-layer BiLSTM for local and contextual temporal information extraction, respectively. Their outputs are integrated through a convolutional layer. The integrated features are fed into the classification head and the regression head, separately, to predict the corresponding class and the offsets to its start-end. Then, a vote-based method is applied to provide better localization results. During training, they first train the backbone for feature extraction. Then, the whole model is fine tuned for sign spotting.

4.3 Top Winning Approaches: OSLWL Track

Top-1: *th* team. The *th* team presented a two-stage framework, consisting of feature extraction and sign instance matching, illustrated in Fig. 3b. First, they extract frame-level feature representation from pose modality. Then, they build a similarity graph to perform frame-wise matching between the query isolated sign and long video. MMDetection [8] is used to detect the signer spatial location, and MMPose for extracting body and hand poses [6]. The compact pose is used to indicate the gesture state of a certain frame, and for trimming the effective time span of the query sign. Pre-trained GCN [3,14] is used to extract body and hand representations in a frame-wise manner. Given the extracted feature of the query and long video, an ad-hoc similarity graph is built to perform sign instance matching. Authors suggest incorporating non-manual cues (e.g., facial expressions) as a possible way to further boost the performance.

Top-2: *Mikedddd* team. The *Mikedddd* team proposed a multi-modal framework for extracting sign language features from RGB images using I3D-MLP [7], 2D-pose [32] and 3D-pose [29] based on SL-GCN [15]. They introduced a novel sign spotting loss function by combining the triplet loss and cross-entropy loss to obtain more discriminative feature representations and training each model separately. At inference stage, the three models are employed to extract visual features before multi-modality features fusion for the sign spotting task. The pipeline of the proposed framework is shown in Fig. 5. More precisely, to achieve the goal of the challenge they proposed the following steps: 1) first, they observed there is an obvious domain gap between isolated signs and continuous signs. Therefore, they trimmed the isolated sign videos by removing the video frames before hands up and after hands down. Then, they fed the gallery sign videos into the model to generate feature representations. As a result, 20 feature representations are obtained in total, each of which represents an isolated sign; 2) For

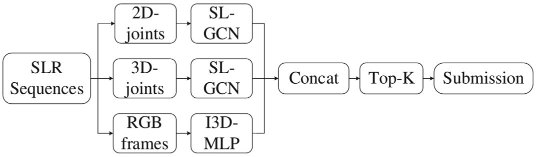


Fig. 5. *Mikedddd* team (top-2, OSLWL track): the inference pipeline.

each video from the query set they use a sliding window to crop video clips from the beginning to the end. Then, they input all clips into the proposed model to extract feature representations; 3) All feature representations from this video are used to calculate the Cosine distance with respect to the corresponding feature representation of the sign provided in the gallery set. The clip that has the maximal similarity with the gallery sign representation is regarded as their retrieval result; 4) They propose a $top - K$ transferring technique to address the domain gap between the gallery set and the query set. After the second step, they obtain several retrieved results for each isolated sign, which are sorted given the computed distance to find the most similar K clips. Each feature representation from the gallery set will be updated by the average of the feature representations of the $top - K$ retrieved clips. Iteratively, the updated feature representation for a certain sign will be used to retrieve signs from continuous sign videos again.

Top-3: ***ryanwong*** **team.** The *ryanwong* team presented a method for sign spotting using existing I3D [7] models pretrained on sign language datasets. First, they show how these models can be used for identifying important frames from isolated sign dictionaries. The goal at this stage is to discard the most common frames where the signer is in their resting pose or those frames with irrelevant information. To this end, feature vectors are obtained for each frame using I3D. Then, the Cosine similarity between each of the frame features across all of the isolated sign videos is computed, which creates a cosine similarity matrix that is used to determine how common the sign frame is within the sign video dictionary. The "key" frames are kept for each isolated sign sequence based on the obtained similarity matrix and predefined thresholds. Next, for each isolated query video they randomly select 8 frames (sorted by indices) of the important "key" frames previously identified. This query sequence is used as input into an I3D model for feature extraction, where a feature vector q is obtained. Similarly, the co-articulated sign video is used as input into the I3D model for feature extraction, and a feature vector k is obtained. The cosine similarity between q and k is calculated to obtain the similarity matrix s. This process is repeated with 64 different combinations of query sequences and 64 different co-articulated sign video (with random data augmentation and random frame selection), obtaining 4096 similarity scores. The mean of all similarity scores at each time step is calculated to obtain the final similarity score s_f. Finally, they compute the "normalized similarity score" s_n by making the assumption that there exists at least 1 occurrence of the isolated sign in the co-articulated sequence by dividing s_f by the maximum value in the sequence. Any indices in s_n greater than a predefined threshold (0.9) is considered a match between the isolated sign video and the continuous sign segment. For each time index with a matched spotting they include the 8 subsequent frames as spotting matches and combine matching spottings if they are in range of 10 frames of each other. The final results are obtained by ensembling the results of 2 models [1] pretrained on WLASL [19] and MSASL [17], and taking the average between the normalized cosine similarities.

4.4 Performance on Marginal Distributions of the Test Set

In this section, we analyze the performance of the top-winning methods on marginal distributions of the test set. Table 4 shows model performance in MSSL track on signers seen during training (SD) and new signers (SI). As expected, top-winning methods performed worse on the signer independent (i.e., SI) scenario, with a performance improvement greater than 10% *w.r.t* the signer-dependent scenario. Note that all methods had models pre-trained with external sign language data. However, they used the challenge training data for fine-tuning. The *baseline* also performed worse on the SI scenario but at a smaller ratio if compared with its SD F1 score. This may be explained by the fact that it was tuned with skeleton data only. It is also worth noting that ranking would stay almost the same if looking just to SI performance, with a swap between the 2^{nd} and 3^{rd} position, not surprising as both had already really close F1 values.

Table 4. Signer dependency of the systems of MSSL track at test phase.

MSSL: *SD → p01, p05, SI → p03, p08*								
Rank	*Participant*	SD/SI	*TP*	*FP*	*FN*	*Pr*	*Re*	*F1*
1	ryanwong	SD	5803	2439	2985	0.704	0.660	0.682
		SI	3813	2102	4949	0.645	0.435	0.520
2	th	SD	5607	2973	3181	0.653	0.638	0.646
		SI	3448	2376	5314	0.592	0.394	0.473
3	Random_guess	SD	4982	2298	3806	0.684	0.567	0.620
		SI	3406	1495	5356	0.695	0.389	0.499
4	baseline	SD	2763	5076	6025	0.352	0.314	0.332
		SI	1638	2301	7124	0.416	0.187	0.258

Note that this results account for the 13^{th} IoU threshold (i.e., 0.8) tested.

Table 5. Per signer performance of MSSL track at test phase.

MSSL: *Signers → p05(interpreter), p01, p03, p08 (deaf)*								
Rank	*Participant*	*Signer*	*TP*	*FP*	*FN*	*Pr*	*Re*	*F1*
1	ryanwong	p05	4823	1573	1781	0.754	0.730	0.742
		p01	980	866	1204	0.531	0.449	0.486
		p03	3348	1865	4478	0.642	0.427	0.514
		p08	465	237	471	0.662	0.497	0.568
2	th	p05	4635	1852	1969	0.714	0.702	0.708
		p01	972	1121	1212	0.464	0.445	0.454
		p03	3067	2224	4759	0.580	0.392	0.468
		p08	381	152	555	0.715	0.407	0.518
3	Random_guess	p05	4143	1473	2461	0.738	0.627	0.678
		p01	839	825	1345	0.504	0.384	0.436
		p03	2953	1324	4873	0.690	0.377	0.488
		p08	453	171	483	0.725	0.484	0.581
4	Baseline	p05	2389	3864	4215	0.382	0.362	0.372
		p01	374	1212	1810	0.236	0.171	0.198
		p03	1354	2104	6472	0.391	0.173	0.240
		p08	284	197	652	0.590	0.303	0.401

Table 5 shows the results marginalized per signer. Signer *p05* is a sign language interpreter, the remainder in MSSL test set are native deaf signers. Participants methods obtained a +10% larger F1 on the interpreter videos, *w.r.t* deaf signers. A possible explanation is that interpreters display less naturalness, hence variability, while signing. This hypothesis should be contrasted further, as *p05* is also the signer with more training data. The F1 ranking among signers holds among the three competing methods ($p05 > p08 > p03 > p01$), so there is a clear dependency on signing style that makes it difficult to separate from the dependency of amount of training data (*p03* and *p08* not seen during training). We leave this important question open to further testing because many of the current larger datasets are based on broadcast video with interpreters' signing.

Table 6 shows per signer performance, precision (Pr), recall (Re) and F1 in OSLWL track. Note that OSLWL does not have explicit training signs/signers, but 2 of the analyzed methods, 2^{nd},*Mikedddd* and 4^{th}, *Baseline*, used MSSL data in the model that extracts feature vectors from the query and MSSL videos. *Mikedddd* also used the OSLWL validation data. We can make some observations from Table 6. First, performance on interpreters (*p05*, *p09*) is better than on deaf signers if the model used training data containing them, which is not observed on methods ranked 1^{st} and 3^{rd}, where deaf signer *p04* obtained better or similar performance. So, as in MSSL, performance is quite signer dependent. Second, there's not a better performance on signer *p03* who is the one that signs half of the queries. This observation shows that, even though the scenario and signer is the same, the domain change between isolated signs and co-articulated signs plays a more important role.

In short, the top-winning methods show a consistent performance on marginal distributions of the test set. Per signer evaluation shows a remarkable performance difference that can not be fully blamed to the amount of training data available for each signer but, probably, to factors related to signing style. It seems that interpreters' style is easier to learn that deafs' style, but there's not enough evidence to support this hypothesis, which deserves further investigation.

Table 6. Per signer performance of OSLWL track at test phase.

OSLWL: *Signers → p05, p09 (interpreters), p01, p03, p04, p08 (deaf)*										
Rank	*Participant*	*Pr*	*Re*	*F1*	*Signer*	*F1*	*Re*	*Pr*	*Participant*	*Rank*
1	th	0.588	0.596	0.592	p05	0.584	0.617	0.554	Mikedddd	2
		0.645	0.645	0.645	p09	0.651	0.680	0.624		
		0.507	0.497	0.502	p01	0.460	0.477	0.444		
		0.453	0.453	0.453	p03	0.462	0.510	0.421		
		0.702	0.728	0.715	p04	0.582	0.612	0.555		
		0.632	0.438	0.517	p08	0.361	0.361	0.361		
3	ryanwong	0.488	0.521	0.504	p05	0.422	0.372	0.487	Baseline	4
		0.562	0.576	0.569	p09	0.485	0.448	0.529		
		0.415	0.484	0.447	p01	0.264	0.239	0.294		
		0.474	0.568	0.517	p03	0.277	0.226	0.357		
		0.529	0.553	0.541	p04	0.423	0.393	0.457		
		0.405	0.467	0.434	p08	0.278	0.225	0.365		

5 Conclusions

This paper summarized the ECCV 2022 Sign Spotting Challenge. We analysed and discussed the challenge design, top winning solutions and results. The new released dataset allowed training and testing sign spotting methods under challenging conditions with deaf and interpreter signers. Although having around 10 h of video data, rich and rigorous hand-made annotations, the dataset is still limited by its small number of participants which have an impact on generalization. To address this problem, future research directions should move on the development of novel large-scale and public datasets and on the research and development of methods that are both fair and accurate, where people from different gender, age, demographics, sign style, among others, are considered.

Top winning solutions of this challenge shared similar pipeline blocks regarding spatial-temporal representation using pre-trained models on larger sign language datasets from different languages, both from 3DCNN and ST-GCN deep learning models. That is, they benefited from state-of-the-art approaches for feature extraction, modeling and/or fusion. Main differences are based on how the domain adaptation problem is handled and the use of the target dataset to train/fine-tune the models or their meta-parameters. The post-challenge experiments showed that signer style can affect the quality of the sign spotting performance, which can be affected by the amount of train data and data distribution previously mentioned. In this line, future work should consider paying more attention to explainability/interpretability, which are key to understand what part or components of the model are more relevant to solve a particular problem, or to explain possible sources of bias or misclassification.

Acknowledgments. This work has been supported by the Spanish projects PID2019-105093GB-I00 and RTI2018-101372-B-I00, by ICREA under the ICREA Academia programme and by the Xunta de Galicia and ERDF through the Consolidated Strategic Group AtlanTTic (2019–2022). Manuel Vázquez is funded by the Spanish Ministry of Science and Innovation through the predoc grant PRE2019-088146.

References

1. Albanie, Samuel, et al.: BSL-1K: scaling up co-articulated sign language recognition using mouthing cues. In: Vedaldi, Andrea, Bischof, Horst, Brox, Thomas, Frahm, Jan-Michael. (eds.) ECCV 2020. LNCS, vol. 12356, pp. 35–53. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_3
2. Alon, J., Athitsos, V., Yuan, Q., Sclaroff, S.: A unified framework for gesture recognition and spatiotemporal gesture segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **31**(9), 1685–1699 (2009)
3. Cai, Y., et al.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: International Conference on Computer Vision (ICCV), pp. 2272–2281 (2019)
4. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR) (2018)

5. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: joint end-to-end sign language recognition and translation. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10023–10033 (2020)
6. Cao, Z., Hidalgo, G., Simon, T., Wei, S., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(01), 172–186 (2021)
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
8. Chen, K., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. CoRR abs/1906.07155 (2019)
9. Cooper, H., Holt, B., Bowden, R.: Sign language recognition. In: Moeslund, T., Hilton, A., Krüger, V., Sigal, L. (eds.) Visual Analysis of Humans, Springer, London, pp. 539–562 (2011). https://doi.org/10.1007/978-0-85729-997-0_27
10. Cui, R., Liu, H., Zhang, C.: Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1610–1618 (2017)
11. Fenlon, J.B., et al.: Bsl signbank: a lexical database and dictionary of British sign language 1st edn (2014)
12. Forster, J., Schmidt, C., Koller, O., Bellgardt, M., Ney, H.: Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pp. 1911–1916 (2014)
13. Grishchenko, I., Bazarevsky, V.: Mediapipe holistic - simultaneous face, hand and pose prediction, on device. https://ai.googleblog.com/2020/12/mediapipe-holisticsimultaneous-face.html (2022). Accessed 18 Jul 2022
14. Hu, H., Zhao, W., Zhou, W., Wang, Y., Li, H.: SignBERT: pre-training of hand-model-aware representation for sign language recognition. In: International Conference on Computer Vision (ICCV), pp. 11087–11096 (2021)
15. Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., Fu, Y.: Skeleton aware multi-modal sign language recognition. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3408–3418 (2021)
16. Jiang, T., Camgoz, N.C., Bowden, R.: Looking for the signs: identifying isolated sign instances in continuous video footage. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pp. 1–8 (2021)
17. Joze, H.R.V., Koller, O.: MS-ASL: a large-scale data set and benchmark for understanding American sign language. CoRR abs/1812.01053 (2018)
18. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: towards large vocabulary statistical recognition systems handling multiple signers. Comput. Vis. Image Understand. **141**, 108–125 (2015)
19. Li, D., Rodriguez, C., Yu, X., Li, H.: Word-level deep sign language recognition from video: a new large-scale dataset and methods comparison. In: Winter Conference on Applications of Computer Vision (WACV) (2020)
20. Li, D., Yu, X., Xu, C., Petersson, L., Li, H.: Transferring cross-domain knowledge for video sign language recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6204–6213 (2020)
21. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 140–149 (2020)

22. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
23. Momeni, L., Varol, G., Albanie, S., Afouras, T., Zisserman, A.: Watch, read and lookup: learning to spot signs from multiple supervisors. In: ACCV (2020)
24. Ong, E.J., Koller, O., Pugeault, N., Bowden, R.: Sign spotting using hierarchical sequential patterns with temporal intervals. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1931–1938 (2014)
25. Pfister, T., Charles, J., Zisserman, A.: Domain-adaptive discriminative one-shot learning of gestures. In: European Conference on Computer Vision (ECCV), vol. 8694, pp. 814–829 (2014)
26. Prillwitz, S.: HamNoSys Version 2.0. Hamburg Notation System for Sign Languages: An Introductory Guide. Intern. Arb. z. Gebärdensprache u. Kommunik, Signum Press, Dresden (1989)
27. Rastgoo, R., Kiani, K., Escalera, S.: Sign Language recognition: a deep Survey. Expert Syst. Appl. **164**, 113794 (2021)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28, pp. 91–99. Curran Associates, Inc. (2015)
29. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In: International Conference on Computer Vision Workshops (ICCVW), pp. 1749–1759 (2021)
30. Schembri, A.C., Fenlon, J.B., Rentelis, R., Reynolds, S., Cormier, K.: Building the British sign language corpus. Lang. Documentation Conserv. **7**, 136–154 (2013)
31. Sincan, O.M., Keles, H.Y.: AUTSL: a large scale multi-modal Turkish sign language dataset and baseline methods. IEEE Access **8**, 181340–181355 (2020)
32. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5686–5696 (2019)
33. Sánchez Pérez, J., Meinhardt-Llopis, E., Facciolo, G.: TV-L1 optical flow estimation. Image Process. Line **3**, 137–150 (2013)
34. Varol, G., Momeni, L., Albanie, S., Afouras, T., Zisserman, A.: Scaling up sign spotting through sign language dictionaries. Int. J. Comput. Vis. 1–24 (2022). https://doi.org/10.1007/s11263-022-01589-6
35. Viitaniemi, V., Jantunen, T., Savolainen, L., Karppa, M., Laaksonen, J.: S-pot - a benchmark in spotting signs within continuous signing. In: International Conference on Language Resources and Evaluation (LREC), pp. 1892–1897 (2014)
36. Voskou, A., Panousis, K.P., Kosmopoulos, D., Metaxas, D.N., Chatzis, S.: Stochastic transformer networks with linear competing units: application to end-to-end SL translation. In: International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021, pp. 11926–11935 (2021)
37. Vázquez-Enríquez, M., Alba-Castro, J.L., Docío-Fernández, L., Rodríguez-Banga, E.: Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2021)
38. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pp. 7444–7452 (2018)

39. Yang, H.D., Sclaroff, S., Lee, S.W.: Sign language spotting with a threshold model based on conditional random fields. IEEE Trans. Pattern Anal. Mach. Intell. **31**(7), 1264–1277 (2009)
40. Zhang, C., Wu, J., Li, Y.: Actionformer: localizing moments of actions with transformers. CoRR abs/2202.07925 (2022)