ChaLearn LAP Large Scale Signer Independent Isolated Sign Language Recognition Challenge: Design, Results and Future Research

Ozge Mercanoglu Sincan
Ankara University, Turkey
omercanoglu@ankara.edu.tr

Julio C. S. Jacques Junior
Computer Vision Center, Spain
jjacques@cvc.uab.cat

Sergio Escalera
University of Barcelona, Spain
Computer Vision Center, Spain
sescalera@ub.edu

Hacer Yalim Keles
Ankara University, Turkey
hkeles@ankara.edu.tr

Abstract

The performances of Sign Language Recognition (SLR) systems have improved considerably in recent years. However, several open challenges still need to be solved to allow SLR to be useful in practice. The research in the field is in its infancy in regards to the robustness of the models to a large diversity of signs and signers, and to fairness of the models to performers from different demographics. This work summarises the ChaLearn LAP Large Scale Signer Independent Isolated SLR Challenge, organised at CVPR 2021 with the goal of overcoming some of the aforementioned challenges. We analyse and discuss the challenge design, top winning solutions and suggestions for future research. The challenge attracted 132 participants in the RGB track and 59 in the RGB+Depth track, receiving more than 1.5K submissions in total. Participants were evaluated using a new large-scale multi-modal Turkish Sign Language (AUTSL) dataset, consisting of 226 sign labels and 36,302 isolated sign video samples performed by 43 different signers. Winning teams achieved more than 96% recognition rate, and their approaches benefited from pose/hand/face estimation, transfer learning, external data, fusion/ensemble of modalities and different strategies to model spatio-temporal information. However, methods still fail to distinguish among very similar signs, in particular those sharing similar hand trajectories.

1. Introduction

Sign/gesture recognition in the context of sign languages is a challenging research domain in computer vision, where signs are identified by simultaneous local and global articulations of multiple manual and non-manual sources, i.e. hand shape and orientation, hand motion, body posture, and facial expressions. Although the nature of the problems in this field is primarily similar to the action recognition domain, some peculiarities of sign languages make this domain specially challenging; for instance, for some pairs of signs, hand motion trajectories look very similar, yet local hand gestures look slightly different. On the other hand, for some pairs, hand gestures look almost the same, and signs are identified only by the differences in the non-manual features, i.e. facial expressions. In some cases, a very similar hand gesture can impose a different meaning depending on the number of repetitions. Another challenge is the variation of signs when performed by different signers, i.e. body and pose variations, duration variations etc. Also, variation in the illumination and background makes the problem harder, which is inherently problematic in computer vision.

The performance of sign recognition algorithms have been improved considerably in recent years, mainly thanks to the release of associated datasets [20] and the development of new deep learning methodologies. Past works used to deal with data obtained in controlled lab environments, with a limited number of signers and signs. Recent works are dealing with more realistic and unconstrained settings and large scale datasets. In parallel, recent advancements in the domains of machine learning and computer vision, in particular deep learning, have pushed the state-of-the-art on the field substantially. Still, several open challenges need to be solved to allow recognition systems to be useful in sign language, including signer independent evaluation, continuous sign recognition, fine-grain hand analyses, combination with face and body contextual cues, sign production, as well as model generalisation to different sign languages and demographics.

To motivate research in the field, we challenged researchers with a signer independent classification task us-

ing a novel large-scale, isolated Turkish Sign Language Dataset, named AUTSL [31]. The video samples in AUTSL, containing variations in background and lighting, are performed by 43 signers. The challenge attracted a total of 191 participants, who made more than 1.5K submissions in total for the two challenge tracks. The RGB and RGB+Depth tracks (detailed in Sec. 3) received 1374 and 209 submissions, respectively, suggesting that the research community on SLR is currently paying more attention to RGB data, compared to RGB+Depth information. Moreover, top-winning solutions employed a wide variety of methods, such as the use of body/face/hand estimation/segmentation, different fusion/ensemble strategies and spatio-temporal modelling, external data and/or transfer learning, among others.

The rest of this paper is organised as follows: in Sec. 2, we provide a short literature review. In Sec. 3, we present the challenge design, evaluation protocol, dataset and baseline. Challenge results and top-winning methods are discussed in Sec. 4. Finally, in Sec. 5, we conclude the paper with a discussion and suggestions for future research.

2. Related Work

Automatic sign language recognition has been an active area of research since early 90s. Early studies relied on using colored gloves or haptic sensors to segment and track hands [12, 13, 26]. However, intrusive methods that require wearing external gloves with some probes create practical difficulties in daily life and often limit the movements of the signers. Therefore, recent studies focus more on computer vision based solutions that use only cameras as the primary equipment for a solution.

Early studies were trained and evaluated on small-scale datasets in terms of number of signs and signers, e.g., Purdue RVL-SLLL [25], RWTH BOSTON50 [40]. In these studies, hand-crafted features, such as scale invariant feature transform (SIFT), histogram of oriented gradients (HOG) [8, 14], were frequently used. After feature extraction, support vector machine (SVM) models or sequence models, such as Hidden Markov Models (HMMs) [8, 41], were used for classification. Similar to earlier works, some studies segmented hand regions before extracting the features, yet this time utilising computer vision based methods, like skin color detection, hand motion detection and trajectory estimation [14, 39].

The emergence of Microsoft Kinect technology in 2010 enabled obtaining new data modalities, such as depth and skeleton, alongside RGB data sequence. New sets of small-scale multi-modal datasets (with less than 50 signs and 15 signers) were created using Kinect, such as DGS [7], GSL [7] and PSL [19]. In ChaLearn Looking at People (LAP) 2013 challenge, a multi-modal Italian gesture dataset, Montalbano V1, was released [10], including RGB, depth, user mask, skeletal model, and audio. It contains 20 gestures and approximately 14,000 samples performed by 27 different signers in total. In ChaLearn LAP 2014 challenge, an enhanced version of the dataset, i.e. Montalbano V2, was released [9]. Although there is only 20 different signs, Montalbano gesture dataset contains more samples and more variance in the video recordings than previously released datasets. In 2014, a large scale isolated Chinese Sign Language that is named as DEVISIGN was released [4]. It consists of 2,000 signs that are performed by 8 signers. The videos were recorded in a lab environment with a controlled background. With the emergence of multiple modalities, researchers worked on different fusing techniques using the features extracted from these modalities, e.g., early, intermediate or late fusion, to get more robust results [36, 27, 28, 38]. Moreover, recent advances prompted researchers to extract features using deep learning based models, instead of using hand-crafted features. Some works preferred using both manually extracted features and deep learning based features together [27, 38].

In 2016, Chalearn LAP RGB-D Isolated Gesture Recognition (IsoGD) dataset was released [37]. It was planned to challenge researchers for high performance automatic classification in "large-scale" and "signer independent" evaluation settings, which means that the samples in the test set are performed by different signers from the train set. In this dataset, there are 249 gestures that are performed by 21 different signers; each class contains approximately 200 RGB and depth videos. In the related years, commonly, 2D-CNN based models were used for feature extraction and sequence models, such as RNN, LSTM, GRU, HMM, were used for encoding temporal information [32, 21, 29, 35]. Recent developments in action recognition have also contributed significantly to the recognition of signs in sign languages. Using and fine-tuning 3D-CNN models, e.g., C3D [34], I3D [3], pre-trained on large action recognition datasets helped achieving higher accuracy rates compared to 2D-CNNs [23, 18, 15, 1].

In recent years, a number of large-scale isolated sign language datasets have been released, with large vocabulary sizes, large number of samples performed by many signers, e.g., MS-ASL [18], CSL [15] and WLASL [22]. MS-ASL provided 1,000 signs with 222 signers in signer independent setting. It was collected from a public video sharing platform. CSL is a multi-modal Chinese Sign Language dataset that consists of 500 signs performed by 50 different signers, arranged for signer independent evaluations. It contains RGB, depth, and skeleton data modalities. WLASL consists of 2,000 signs performed by 119 signers. It was collected from sign language websites. Although each of these datasets has several different challenges, video samples usually have plain backgrounds and data is collected in a controlled setting. Table 1 provides an overview of the

Table 1. Overview of isolated sign language/gesture datasets.

Datasets	Year	Signer independent	Modalities	#Signs	#Signers	#Samples
RWTH BOSTON50 [40]	2005	No	RGB	50	3	483
DGS [7]	2012	No	RGB, depth	40	15	3,000
GSL [7]	2012	No	RGB	20	6	840
Montalbano V1, V2 [10, 9]	2014	No	RGB, depth, audio, user mask, skeleton	20	27	13,858
DEVISIGN[4]	2014	No	RGB, depth	2,000	8	24,000
PSL [19]	2015	No	RGB, depth	30	1	300
LSA64[30]	2016	No	RGB	64	10	3,200
isoGD [40]	2016	Yes	RGB, depth	249	21	47,933
MS-ASL [18]	2019	Yes	RGB	1,000	222	25,513
CSL [15]	2019	Yes	RGB, depth, skeleton	500	50	125,000
WLASL [22]	2020	No	RGB	2,000	119	21,083
AUTSL [31]	2020	Yes	RGB, depth, user mask, skeleton	226	43	36,302

available isolated sign language/gesture datasets.

In the context of this challenge, a new large-scale, multimodal Turkish Sign Language dataset, AUTSL [31], is utilized in a signer independent evaluation setting. Different from the other large-scale datasets, it contains a variety of 20 different backgrounds obtained from indoor and outdoor environments, with several challenges (detailed in Sec. 3.1).

3. Challenge Design

The challenge[1] focused on isolated Sign Language Recognition (SLR) from signer independent non-controlled RGB+D (depth) data, involving a large number of sign categories (>200, detailed in Sec. 3.1). It was divided into two different competition tracks, i.e., RGB[2] and multimodal RGB+D[3]. The only restriction was that depth data was not allowed in any format and stage of training in RGB track. The participants were free to join any of these tracks. Both modalities have been temporally and spatially aligned. Each track was composed of two phases, i.e., development and test phase. At the development phase, public train data was released and participants submitted their predictions with respect to a validation set. At the test (final) phase, participants were requested to submit their results with respect to the test data. Participants were ranked, at the end of the challenge, using the test data.

The challenge ran from 22 December 2020 to 11 March 2021 through Codalab[4], a powerful open source framework for running competitions that involve result or code submission. It attracted a total of 191 registered participants, 132 in RGB track and 59 in RGB+D track. During development phase we received 1317 submissions from 39 teams in the RGB track, and 176 submissions from 15 teams in the RGB+D track. At the test (final) phase, we received 57 submissions from 23 teams in the RGB track, and 33 submissions from 14 teams in the RGB+D track. The reduction in the number of submissions from the development to the test phase is explained by the fact that the maximum number of submissions per participant on the final phase was limited to 3, to minimise the change of participants to improve their results by try and error.

It is important to note that the challenge was designed to deal with the submission of results (and not code). Participants submitted only their prediction files containing one label for each video. Therefore, participants were required to share their codes after the end of the challenge so that the organisers could validate their results in a "code verification stage". At the end of the challenge, top ranked methods (discussed in Sec. 4.2) passing the code verification stage (e.g., they publicly released their codes and the organisers were able to reproduce the results) were announced as top winning solutions.

3.1. The Dataset

AUTSL [31] is a large-scale, signer independent, multimodal dataset that contains isolated Turkish sign videos. It contains 226 signs that are performed by 43 different signers. The dataset is recorded with Microsoft Kinect V2 and contains RGB, depth, user mask, and skeleton data. Only RGB and depth modalities are released within the scope of the challenge. Some clipping and resizing operations are applied to RGB and depth data and video frames are resized to 512 x 512 pixel resolution. The average number of frames per video is 60 ± 10, while the frame rate for video is 30 frame per second (fps).

In the associated publication [31], while AUTSL test split was designed as a signer independent set; validation split was generated using a random split, i.e. 15% of the train data. For this challenge, we split the training set to create a signer independent validation set, making all sets signer independent. We selected 31 signers for training, 6 signers for validation, and 6 signers for testing. In this

[1] Challenge Webpage: http://chalearnlap.cvc.uab.es/challenge/43/description/

[2] https://competitions.codalab.org/competitions/27901

[3] https://competitions.codalab.org/competitions/27902

[4] https://codalab.org/

Figure 1. Some screenshots from the AUTSL [31] dataset.

setting, training set contains 28,142, validation set contains 4,418, and test set contains 3,742 video samples. AUTSL is a balanced dataset according to the sign distribution, i.e. each sign contains approximately the same number of samples (∼160). The train, validation and test set contain approximately 124, 19, and 17 samples per sign, respectively. Signs are selected from the daily spoken vocabulary. They cover a wide variety in terms of hand shape and hand movements; some signs are performed only with one hand while some with both hands, in some signs hands occlude each other or parts of the face. We depict examples of different backgrounds and signers from the dataset in Fig. 1.

Challenges: The dataset has various challenges, including lighting variability, different postures of signers, dynamic backgrounds, such as moving trees, or moving people behind the signer, high intra-class variability and inter-class similarities. In order to provide a basis for signer independent recognition systems, train, val, and test splits include different signers. The dataset contains 20 different backgrounds with several challenges. The test set contains 8 different backgrounds, 3 of which are not included in the training or validation sets. Another challenge is the inter-class similarity of signs; some signs contain exactly the same hand gesture, but differing only by the number of repetitions of the same gesture. Also, some signs are quite similar in terms of hand shape, hand orientation, hand position or hand movement; there is only subtle differences.

Limitations: The fact that the society is right-handed in general is also reflected in the distribution in AUTSL. Only 2 of the signers are left-handed out of 43 signers. Therefore, there is a bias towards the right handed signers in the dataset. Furthermore, female signers are more dominant, almost 3:1 ratio, in the dataset; 10 of the signers are men and 33 are women. The ages of our signers range from 19 to 50, and the average age of all signers is 31. In other words, there are no child or elderly signers. Another point that can be considered a source of bias in the dataset is the distribution of skin color, as there is no signer with dark skin. Although these limitations exist, we believe the challenge we have organised can help to advance the state-of-the-art on the field, as well as to promote either the design of new dataset or the development of novel methodologies that can deal with the aforementioned limitations.

3.2. Evaluation Protocol

To evaluate the performances of the models, we use the recognition rate, r, as defined in previous ChaLearn LAP challenges [37].

$$r = \frac{1}{n}\sum_{i=1}^{n} f(p_i, y_i), \tag{1}$$

where n is the total number of samples; p_i is the predicted label for the i^{th} sample; y_i is the true label for the i^{th} sample; $f(.)$ is 1 when $p_i = y_i$, 0 otherwise.

3.3. The Baseline

In order to set a baseline, several deep learning based models are trained and evaluated on AUTSL dataset. In the baseline method [31], 2D-CNNs are used to extract spatial features. Then, a Feature Pooling Module (FPM) [32] is placed on top of the last CNN layer. The idea behind FPM layers is to increase the field-of-views by using different dilated convolutions. In order to capture temporal information bidirectional LSTM (BLSTM) is used. A temporal attention mechanism is integrated to BLSTM in order to select the most effective video frames in classification.

The methods used in RGB and RGB+D track are basically the same, with minor modifications. Since the depth data is represented as a single channel gray-scale image for each frame, the same depth data is repeated into three color channels. Then, RGB and depth modalities are given as inputs to the two parallel CNN models that share the same parameters. After generating two feature matrices, i.e. one for the RGB data and one for the depth data, these feature matrices are concatenated at the end of the FPM layer.

In contrast to some top-winning solutions (detailed in Sec. 4), our baseline did not consider any face/hand/body detection or segmentation technique, nor additional modalities such as pose keypoints, optical flow, nor external data.

4. Challenge Results and Winning Methods

4.1. The Leaderboard

Results obtained by the top-10 winning solutions (in addition to the *Baseline*) at the test phase, for the RGB and RGB+D tracks, are reported in Table 2. The main observation from the table is that RGB+D and RGB results do not significantly differ, suggesting that state-of-the-art methods are obtaining highly accurate results without the use of depth information, at least on the adopted AUTSL [31] dataset. Later in Sec. 4.3, we analyse possible causes why some samples were not properly recognised by the top winning solutions, which could be used to guide future research directions on the field.

Table 2. Codalab leaderboards of RGB and RGB+D Tracks. Methods that passed the code verification stage are highlighted in bold.

RGB Track			**RGB+D Track**		
Rank	Participant	Rec. Rate	Rank	Participant	Rec. Rate
1	**smilelab2021**	0.9842	1	**smilelab2021**	0.9853
2	wz	0.9834	2	wz	0.9834
3	**rhythmblue**	0.9762	3	**rhythmblue**	0.9765
4	_Bo_	0.9743	4	**wenbinwuee**	0.9669
5	**wenbinwuee**	0.9655	5	lin.honghui	0.9567
6	deneme4	0.9626	6	ly59782	0.9548
7	**jalba**	0.9615	7	Bugmaker	0.9396
8	xz	0.9596	8	**m-decoster**	0.9332
9	wuyongfa	0.9580	9	papastrat	0.9172
10	adama	0.9578	10	xduyzy	0.9086
23	*Baseline*	0.4923	14	*Baseline*	0.6203

Fig. 2 illustrates the evolution of the challenge with respect to the number of submissions and highest score obtained for each day and competition track. Different observations can be made from these plots: 1) the participants were much more active in the RGB track, also reinforced by the number of registered participant on this track, suggesting that the research community is paying more attention on SLR from RGB data if compared to RGB+D information; 2) the number of submissions increases close to the end of each phase (development phase finished on 3rd of March and the test phase finished on 11th of March), suggesting that participants were struggling to improve their results to obtain a better rank position.

4.2. Top Winning Approaches

This section briefly presents the top winning approaches of both tracks. More concretely, the top-3 methods that passed the code verification stage (see Table 2). Table 3 shows some general information about the top-3 winning approaches. As it can be seen from Table 3, common strategies employed by top-winning solutions are transfer learning, external data, face/hand detection and pose estimation, fusion of modalities and ensemble models as well as different strategies to model spatio-temporal information.

4.2.1 Top-1: smilelab2021

Inspired by the recent development of whole-body pose estimation [17], the *smilelab2021*[5] team proposed to recognise sign language based on the whole-body key points and features. The recognition results are further ensembled with other modalities of RGB and optical flows to further improve the accuracy. The top-1 winning solution [16] proposed to use whole-body pose keypoints to recognise sign language via a multi-stream Graph Convolutional Network (GCN) model.

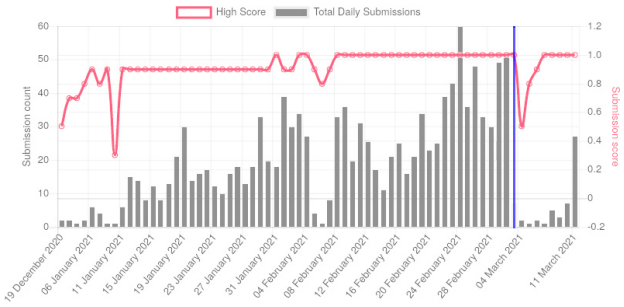
A total of 133-points including face, hand, body and foot are extracted from the input images. They are used in their GCN network as skeleton modality. Features extracted from pretrained whole-body pose estimation are used as another modality. The keypoints are also used to crop frames in other modalities (RGB and optical flow). As base model, Resnet2+1d pretrained on Kinetics [2] dataset is used. For RGB modality, they pre-trained their models on Chinese Sign Language dataset [42] before training on the challenge dataset. Label smoothing and weight decay were used as regularization during training on both tracks. In the RGB+D track, they extracted HHA [16] features from depth video as another modality, referred to as handcraft features. HHA features encode depth information and are generated using a RGB-like 3-channel output, where HHA stand for "Horizontal disparity", "Height above the ground", and "Angle normal makes with". Since multiple modalities are considered (skeleton keypoints, skeleton features, RGB and optical flow - and HHA and depth flow in the case of RGB+D track), they adopted a late fusion technique where the output of the last fully-connected layers is kept, before softmax, associating weights to them and sum them up with weights as a final predicted score. Those weights serve as hyper-parameters and are tuned based on the accuracy on validation set.

4.2.2 Top-2: rhythmblue

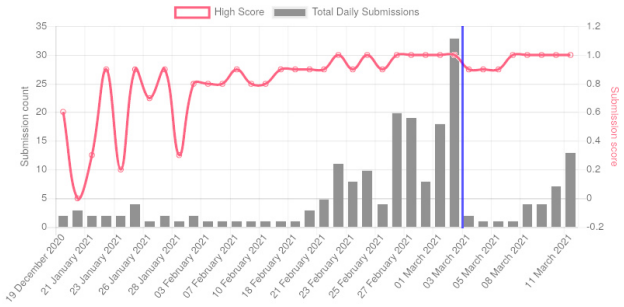
The *rhythmblue*[6] team proposed an ensemble framework composed of multiple neural networks (e.g., I3D, SGN) to conduct isolated sign language recognition, also taking into account pose, hand and face patch-based information. The

[5]Code: https://github.com/jackyjsy/CVPR21Chal-SLR

[6]Code: https://github.com/ustc-slr/ChaLearn-2021-ISLR-Challenge



(a) RGB Track



(b) RGB+D Track

Figure 2. Challenge evolution with respect to the number of submissions and best obtained score, per day and per track. The blue line indicates the end of the development phase and the start of the test phase.

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

Participant	**top-1: smilelab2021**		**top-2: rhythmblue**		**top-3: wenbinwuee**	
Feature/Track	*RGB*	*RGB+D*	*RGB*	*RGB+D*	*RGB*	*RGB+D*
Depth information either during training or testing stage	-	√	-	√	-	√
Pre-trained models	√	√	√	√	√	√
External data	√	√	-	-	-	-
Regularization strategies/terms	√	√	-	-	-	-
Handcrafted features	-	√	-	-	-	-
Face/hand/body detection, alignment or segmentation	-	-	√	√	√	√
Pose estimation	√	√	√	√	-	-
Fusion of modalities	√	√	√	√	-	-
Ensemble models	√	√	√	√	√	√
Spatio-temporal feature extraction	√	√	√	√	√	√
Explicitly classify any attribute (e.g. gender)	-	-	-	-	-	-
Bias mitigation technique (e.g. rebalancing training data)	-	-	-	-	-	-

networks are trained separately for different cues. For patch sequence of full-frame, hands and face, 3D-CNNs are used to model the spatio-temporal information. For pose data, GCN-based method is selected to capture the skeleton correlation. During ensemble stage, late fusion is adopted for final prediction.

More concretely, an upper-body patch is obtained according to a bounding box estimation of the signer, given by MMDetection [5] and the joint positions estimated by HRNet [33]. Full body joint positions are extracted with MMPose [6]. Hand and face regions are obtained from keypoint positions. Five types of data are generated, i.e., full-body patch, left-hand patch, right-hand patch, face patch and full-body pose. To process full-body patch, left-hand patch and right-hand patch, separately I3D [3] networks are used. A SlowFast [11] Network is used for full-body patch. To process full-body pose, SGN [43] is used. During inference, all outputs are summed before the SoftMax layers of the above networks with weights. Then, the category with the largest activation is selected. In the case of RGB+D track, I3D-Depth is used, with a pretrained model provided by I3D-Kinectics-Flow [3].

4.2.3 Top-3: wenbinwuee

The *wenbinwuee*[7] team used RGB, Depth information (in RGB+D track), optical flow and human segmentation data to train several models using SlowFast [11], SlowOnly [11] and TSM [24]. Results are late fused to get a final prediction. For the RGB+D track, results are obtained by fusing the RGB+D models prediction scores with the RGB track results.

4.3. What challenge the models the most?

In this section, we analyse the prediction files submitted at the test phase on both competition tracks and for the top-10 teams shown in Table 2, and discuss some particularities that challenge the methods the most. For instance, we show the samples that were more frequently wrongly classified given the signer or sign IDs, which could indicate a weakness of the evaluated methods or any issue in the adopted dataset (e.g., high inter-class similarity), that could suggest future research.

Fig. 3 shows misclassification rates in the RGB track for

[7]Code: https://github.com/Koooko96/Chalearn2021code

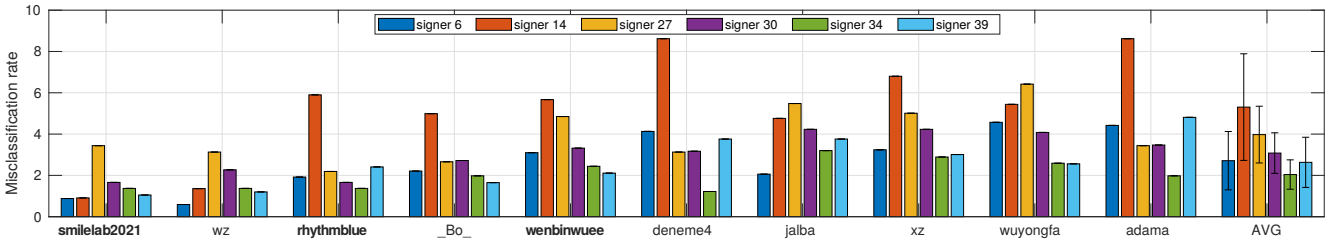
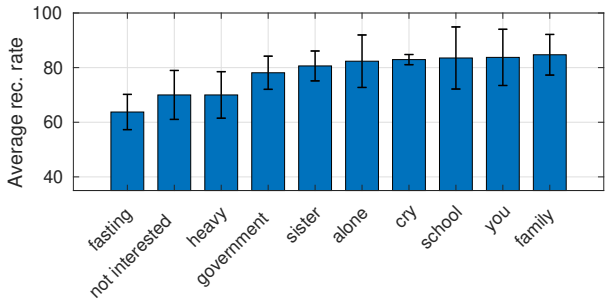
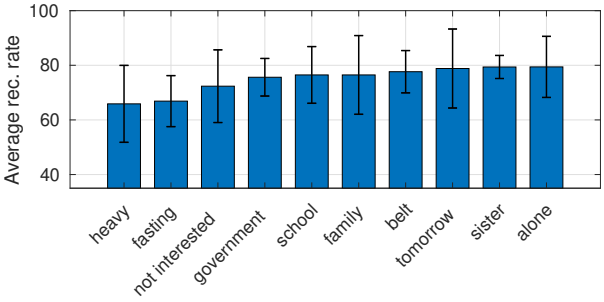


Figure 3. Misclassification rate (%) per signer on the test set, given the top-10 teams (shown in Table 2) in the RGB track. Average misclassification rate is also reported. In bold, the top-3 winning solutions that passed the code verification stage.

each signer in the test set. No significant differences were observed for the results that are obtained for RDB+D track. The first 10 segments show the results of the top-10 teams shown in Table 2, and the last segment shows the average misclassification rates per signer. As it can be seen, different teams misclassified different signers without a clear pattern, suggesting that there was not a particular signer (or set of signers) that could be considered an outlier. Nevertheless, if we take the average misclassification as reference, we can observe that Signer 14 and 27 were the ones more frequently misclasified (which was not the case of the top-1 winning solution, at least for Signer 14). One possible explanation for these cases is the high intra-class variability, which imposes an additional challenge for generalisation.



(a) RGB Track



(b) RGB+D Track

Figure 4. Average recognition rate (%) of top-10 misclassified signs (test set) given the top-10 teams (shown in Table 2).

In Fig. 4, we show the average recognition rate of top-10 misclassified signs on the test phase, given the submitted files of top-10 teams (shown in Table 2). As it can be seen, the set of signs with lower recognition rates (e.g., top-5) are more or less the same in both tracks, suggesting that RGB and RGB+depth are providing similar information to solve the task on those cases, or maybe that the complementarity of RGB and depth are not being fully exploited.

By analysing a confusion matrix of the signs and submitted predictions of both tracks, we observed that the most frequently confused sign pairs by participants methods in the AUTSL [31] dataset are: {heavy *vs.* lightweight}, {fasting *vs.* school}, {not interested *vs.* why}, {school *vs.* soup} and {government *vs.* Ataturk}. The reason behind such misclassifications are due to the high similarity of local and global hand gestures in these signs, illustrated in Fig. 5. In some signs, there is only a subtle difference in the position of the hand, e.g., government and Ataturk. While in the sign of government the index finger touches under the eye, in the sign of Ataturk it touches the cheek. In some signs, there is only a subtle difference in movement of hands, e.g., school and soup. In the sign of soup the hand moves a little more from the bottom up. In some signs, facial expression also contains an important clue for the meaning of the sign, e.g., heavy.

5. Conclusions

This work summarised the ChaLearn LAP Large Scale Signer Independent Isolated SLR Challenge. The challenge attracted more then 190 participants in two computational tracks (RGB and RGB+D), who made more than 1.5K submissions in total. We analysed and discussed the challenge design, top winning solutions and results. Interestingly, the challenge activity and results on the different tracks showed that the research community in this field is currently paying more attention to RGB information, compared to RGB+Depth data. Top winning methods combined hand/face detection, pose estimation, transfer learning, external data, fusion/ensemble of modalities and different strategies to model spatio-temporal information.

We believe that future research directions should move at least in two different lines, that is, on the development of novel large-scale and public datasets, and on the research and development of methods that are both fair and accurate. Fairness is an emergent topic in computer vision and

(a) {heavy *vs.* lightweight}

(b) {not interested *vs.* why}

(c) {fasting *vs.* school *vs.* soup}

(d) {government *vs.* Ataturk}

Figure 5. The most frequently confused sign pairs in both RGB and RGB+D track. Each row displays a sign video sample summarised in 4 frames.

machine learning, and new datasets should include people from different ages, gender, skin tones, demographics, among others, with the goal of having as much as possible balanced distributions given the different attributes. Moreover, continuous sign language seems to be a logical next stage in order to do research on begin-end of sign detection and to include of a higher level of language semantics in the recognition process. The inclusion and analysis of context and spatio-temporal attention mechanisms could be helpful for discriminating very similar signs. On the other hand, models are benefiting from state-of-the-art approaches developed for other purposes to achieve state-of-the-art performance. The fusion of different modalities and models seems to be a key to advance the research on this field. It should be noticed that the top-2 winning solutions benefited from Graph Convolutional Networks (GCN), which demonstrated to be very useful to model spatio-temporal information. Furthermore, up to date self-attention strategies have not been fully exploited in sign language, and its usage could benefit the spatio-temporal learning of signs.

Finally, future work should also consider paying more attention to explainability/interpretability, so that the results obtained by different models could be easily explained and interpreted. This is 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 partially supported by the Scientific and Technological Research Council of Turkey (TUBITAK) project 217E022, the Spanish project PID2019-105093GB-I00 (MINECO/FEDER, UE) and CERCA Programme/Generalitat de Catalunya, and ICREA under the ICREA Academia programme.

References

[1] Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. A comprehensive study on sign language recognition methods. *arXiv:2007.12530*, 2020. 2

[2] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv:1808.01340*, 2018. 5

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 2, 6

[4] X Chai, H Wanga, M Zhoub, G Wub, H Lic, and X Chena. Devisign: dataset and evaluation for 3d sign language recognition. Technical report, Beijing, Tech. Rep, 2015. 2, 3

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv:1906.07155*, 2019. 6

[6] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 6

[7] HM Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. Sign language recognition using sub-units. *Journal of Machine Learning Research*, 13:2205–2231, 2012. 2, 3

[8] Nasser H Dardas and Nicolas D Georganas. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Transactions on Instrumentation and measurement*, 60(11):3592–3607, 2011. 2

[9] Sergio Escalera, Xavier Baró, Jordi Gonzalez, Miguel A Bautista, Meysam Madadi, Miguel Reyes, Víctor Ponce-López, Hugo J Escalante, Jamie Shotton, and Isabelle Guyon. Chalearn looking at people challenge 2014: Dataset and results. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 459–473, 2014. 2, 3

[10] Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Isabelle Guyon, Vassilis Athitsos, Hugo Escalante, Leonid Sigal, Antonis Argyros, Cristian Sminchisescu, et al. Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 365–368, 2013. 2, 3

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *IEEE International Conference on Computer Vision (ICCV)*, October 2019. 6

[12] Sidney S Fels and Geoffrey E Hinton. Glove-talk: A neural network interface between a data-glove and a speech synthesizer. *IEEE transactions on Neural Networks*, 4(1):2–8, 1993. 2

[13] Kirsti Grobel and Marcell Assan. Isolated sign language recognition using hidden markov models. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 1, pages 162–167. IEEE, 1997. 2

[14] Junwei Han, George Awad, and Alistair Sutherland. Modelling and segmenting subunits for sign language recognition based on hand motion analysis. *Pattern Recognition Letters*, 30(6):623–633, 2009. 2

[15] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. Attention-based 3d-cnns for large-vocabulary sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9):2822–2832, 2018. 2, 3

[16] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021. 5

[17] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision (ECCV)*, pages 196–214, 2020. 5

[18] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv:1812.01053*, 2018. 2, 3

[19] Tomasz Kapuscinski, Mariusz Oszust, Marian Wysocki, and Dawid Warchol. Recognition of hand gestures observed by depth cameras. *International Journal of Advanced Robotic Systems*, 12(4):36, 2015. 2, 3

[20] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv:2008.09918*, 2020. 1

[21] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(9):2306–2320, 2019. 2

[22] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020. 2, 3

[23] Yunan Li, Qiguang Miao, Kuan Tian, Yingying Fan, Xin Xu, Rui Li, and Jianfeng Song. Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model. In *International Conference on Pattern Recognition (ICPR)*, pages 25–30. IEEE, 2016. 2

[24] J. Lin, C. Gan, and S. Han. TSM: Temporal shift module for efficient video understanding. In *International Conference on Computer Vision (ICCV)*, pages 7082–7092, 2019. 6

[25] Aleix M Martínez, Ronnie B Wilbur, Robin Shay, and Avinash C Kak. Purdue rvl-slll asl database for automatic recognition of american sign language. In *IEEE International Conference on Multimodal Interfaces*, pages 167–172, 2002. 2

[26] Syed Atif Mehdi and Yasir Niaz Khan. Sign language recognition using sensor gloves. In *International Conference on Neural Information Processing*, volume 5, pages 2204–2206. IEEE, 2002. 2

[27] Natalia Neverova, Christian Wolf, Graham W Taylor, and Florian Nebout. Multi-scale deep learning for gesture detection and localization. In *European Conference on Computer Vision (ECCV)*, pages 474–490, 2014. 2

[28] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. Sign language recognition using convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, pages 572–578, 2014. 2

[29] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Video-based isolated hand sign language recognition using a deep cascaded model. *Multimedia Tools and Applications*, 79:22965–22987, 2020. 2

[30] Franco Ronchetti, Facundo Quiroga, César Armando Estrebou, Laura Cristina Lanzarini, and Alejandro Rosete. Lsa64:

an argentinian sign language dataset. In *XXII Congreso Argentino de Ciencias de la Computación (CACIC)*, 2016. 3

[31] O. M. Sincan and H. Y. Keles. AUTSL: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020. 2, 3, 4, 5, 7

[32] Ozge Mercanoglu Sincan, Anil Osman Tur, and Hacer Yalim Keles. Isolated sign language recognition with multi-scale features using lstm. In *27th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2019. 2, 4

[33] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[34] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. 2

[35] Anil Osman Tur and Hacer Yalim Keles. Evaluation of hidden markov models using deep cnn features in isolated sign recognition. *Multimedia Tools and Applications*, pages 1–19, 2021. 2

[36] Jun Wan, Qiuqi Ruan, Wei Li, and Shuang Deng. One-shot learning gesture recognition from rgb-d data using bag of features. *The Journal of Machine Learning Research*, 14(1):2549–2582, 2013. 2

[37] Jun Wan, Yibing Zhao, Shuai Zhou, Isabelle Guyon, Sergio Escalera, and Stan Z Li. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–64, 2016. 2, 4

[38] Di Wu, Lionel Pigou, Pieter-Jan Kindermans, Nam Do-Hoang Le, Ling Shao, Joni Dambre, and Jean-Marc Odobez. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(8):1583–1597, 2016. 2

[39] Quan Yang. Chinese sign language recognition based on video sequence appearance modeling. In *2010 5th IEEE Conference on Industrial Electronics and Applications*, pages 1537–1542. IEEE, 2010. 2

[40] Morteza Zahedi, Daniel Keysers, Thomas Deselaers, and Hermann Ney. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Joint Pattern Recognition Symposium*, pages 401–408. Springer, 2005. 2, 3

[41] Mahmoud M Zaki and Samir I Shaheen. Sign language recognition using a combination of new vision based features. *Pattern Recognition Letters*, 32(4):572–577, 2011. 2

[42] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li. Chinese sign language recognition with adaptive hmm. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2016. 5

[43] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6