*Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages*, pages 181–188
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

# BosphorusSign22k Sign Language Recognition Dataset

**Oğulcan Özdemir[a], Ahmet Alp Kındıroğlu[a], Necati Cihan Camgöz[b], Lale Akarun[a]**
[a]Boğaziçi University, Computer Engineering Department, Istanbul, Turkey
[b]CVSSP, University of Surrey, Guildford, United Kingdom
{ogulcan.ozdemir, alp.kindiroglu, akarun}@boun.edu.tr, n.camgoz@surrey.ac.uk

## Abstract

Sign Language Recognition is a challenging research domain. It has recently seen several advancements with the increased availability of data. In this paper, we introduce the BosphorusSign**22k**, a publicly available large scale sign language dataset aimed at computer vision, video recognition and deep learning research communities. The primary objective of this dataset is to serve as a new benchmark in Turkish Sign Language Recognition for its vast lexicon, the high number of repetitions by native signers, high recording quality, and the unique syntactic properties of the signs it encompasses. We also provide state-of-the-art human pose estimates to encourage other tasks such as Sign Language Production. We survey other publicly available datasets and expand on how BosphorusSign**22k** can contribute to future research that is being made possible through the widespread availability of similar Sign Language resources. We have conducted extensive experiments and present baseline results to underpin future research on our dataset.

**Keywords:** Turkish Sign Language (TID), Sign Language Recognition, Deep Learning

## 1. Introduction

As native languages of the Deaf, Sign Languages (SL) are visio-temporal constructs which convey meaning through hand gestures, upper body motion, facial expressions and mouthings. Automatic Sign Language Recognition (ASLR) is a challenging task and an active research field with the aim of reducing the dependency of sign language interpreters in the daily lives of the Deaf.

Among the many similar problems attempted by deep learning researchers, sign language recognition bears a resemblance to video-based action recognition because of its shared medium of information (Varol et al., 2017), and to speech recognition and machine translation problems (Bahar et al., 2019; Bahdanau et al., 2017), due to its linguistic nature. However, there are certain aspects of ASLR that makes the task more challenging, one of which is the asynchronous multi-articulatory nature of the sign (Sutton-Spence and Woll, 1999). Furthermore, the lack of large databases aimed at computer vision communities and the difficulty of annotating them has been an inhibiting factor in ASLR research (Hanke et al., 2010; Schembri et al., 2013).

In this paper, we present BosphorusSign**22k**, an isolated SL dataset, for benchmarking repeatable deep learning experiments on SLR. The dataset was derived from BosphorusSign (Camgoz et al., 2016c), which has high-quality recordings collected from Deaf users of Turkish Sign Language (TID). The BosphorusSign Dataset was categorized linguistically, where sign glosses with the same meaning but a different set of morphemes were considered belonging to the same class. Although this annotation scheme was aimed to be useful in a Q&A based interaction system, i.e., banking or hospital desk applications (Suzgun et al., 2015), it is not well-suited for sign language recognition and production systems, where distinguishing between instances of similar sign classes with similar manual and non-manual features is essential.

Although, BosphorusSign is publicly available, there are no benchmarks reported on this dataset. Furthermore, BosphorusSign does not have an evaluation protocol, making future research conducted using BosphorusSign dataset incomparable with one another.

Moreover, BosphorusSign dataset only provided skeleton information obtained from the Kinect V2 SDK. Although real-time depth-based skeleton estimation was state-of-the-art at the time of BosphorusSign's creation, it is jittery and lacks the crucial hand pose information, making the skeleton information provided in the BosphorusSign dataset inadequate for training human pose based sign recognition, translation and production models (Stoll et al., 2018; Stoll et al., 2020).

To address these issues, in this paper, we have enhanced and refined the BosphorusSign dataset, to help future research in the fields of sign language recognition and production. The contributions of this work are listed as;

- We have visually reviewed and cleaned up the dataset and removed all erroneous sign performances.
- We have revisited the labeling scheme and converted the linguistic categorization into a form more suitable for recognition and production research where each class shares the same manual features.
- We provide OpenPose (Cao et al., 2018) body and finger coordinates in addition to Kinect V2 skeleton information.
- We have proposed an evaluation protocol and reported two benchmark results using 3D ResNets (Tran et al., 2018) and IDT (Wang and Schmid, 2013) to underpin future research on this dataset.

The rest of this paper is organized as follows: In Section 2., we give an overview of the SLR literature and other publicly available Sign Language Recognition (SLR) datasets. In Section 3., we introduce the new BosphorusSign**22k** dataset. We then describe the baseline methods and share our experimental results in Section 4.. Finally, we conclude this paper in Sections 5. and 6. by analyzing our baseline results, discussion and future work.

## 2. Related Work

Since the work of Starner et al. (1998), there have been numerous studies on the isolated SLR task. More recently, utilization of state-of-the-art deep learning models (Koller et al., 2019; Zhang et al., 2016) have resulted in better representation learning that is capable of achieving high accuracies over hundreds of unique sign glosses. Because of the spatio-temporal characteristics of the ASLR problem, popular methods from video (human action & activity recognition) (Wang and Schmid, 2013; Carreira and Zisserman, 2017) and speech recognition (Graves and Schmidhuber, 2005) fields have been widely applied with success to the SLR problem. Since the focus of this paper is on computer vision based ASLR methods, the main variations among solutions proposed to this solution lie in their methods of data representation and temporal sign modeling.

Of the present methods in the literature, a large majority use sequences of RGB video frames and/or depth information (Cui et al., 2017; Wang et al., 2016; Wang et al., 2019). Some methods extract additional features from these input sources such as optical flow and coordinates calculated by human pose capture methods such as Kinect (Shotton et al., 2012) and OpenPose (Cao et al., 2018). A large number of methods extract state of the art features such as 2D (Koller et al., 2016a) and 3D Convolutional Neural Network (CNN) outputs (Joze and Koller, 2018; Huang et al., 2015; Camgoz et al., 2016a), Improved Dense Trajectory (IDT) features (Özdemir et al., 2016), hand appearance and trajectory features (Özdemir et al., 2018; He et al., 2016; Metaxas et al., 2018). The use of spatial attention as in Yuan et al. (2019) to focus learning on the signing space and temporal attention as in Camgoz et al. (2017) and Camgoz et al. (2018; Camgoz et al. (2020) are also among current popular research directions.

In terms of temporal segmentation and modeling, isolated videos of sign glosses often consist of varying length and complexity, requiring the use of temporal modeling. In Aran (2008) and Zhang et al. (2016), Hidden Markov Models (HMMs) are used with hand shape features and trajectories, while Koller et al. (2016b) uses HMMs to train hand shape classifiers from weakly labeled sign videos. In Liu et al. (2016), Long Short-Term Memory Networks (LSTMs) are used with gradient histograms while in Camgoz et al. (2017) they are used with Connectionist Temporal Classification (CTC) and neural network features to learn sign languages. Based on the works of Joze and Koller (2018) and Camgoz et al. (2016a) and the results of popular gesture recognition challenges such as Chalearn LAP (Wan et al., 2017), SLR studies with 3D CNNs currently tend to show higher performances in large datasets compared to other deep learning approaches utilizing LSTMs and other popular methods. This generalization loses validity in the case of continuous SLR where the average clip length exceeds a few seconds.

In isolated Sign Language Recognition, the difficulties in obtaining high quality annotated videos has limited the amount of available public datasets. To the best of our knowledge, currently there exist four similar public available large scale isolated SLR datasets, which can be seen in Table 1, while Chinese Sign Language (CSL) recognition dataset (Zhang et al., 2016) and MS-ASL (Joze and Koller, 2018) being the most recent ones. BosphorusSign**22k** differentiates from these datasets as follows: Contrary to Chinese Sign Language (CSL) recognition dataset, which was recorded in front of a white background, BosphorusSign**22k** dataset was captured using a Chroma Key background, which we believe will be beneficial for researchers who would like to utilize data augmentation techniques in their pipelines to improve their models generalization capabilities. We acknowledge the fact that MS-ASL is one of the new frontiers in sign language research as it initiated large scale “in the wild” isolated sign language recognition. However, the dataset is composed of publicly available YouTube videos. As Microsoft does not own the copyright of these videos, the availability of the data is not guaranteed. As of the submission of this paper, 290 videos are no longer publicly available which will potentially increase in the future, making comparison of future research against previously reported benchmarks harder. Additionally, there is the new SMILE DSGS corpus (Ebling et al., 2018), which hasn’t been fully publicly available and its evaluation protocol is yet to be defined.

## 3. BosphorusSign22k Dataset

In this study, we present BosphorusSign**22k**[1], a new benchmark dataset for vision-based user-independent isolated SLR. Our dataset is based on the BosphorusSign (Camgoz et al., 2016c) corpus which was collected with the purpose of helping both linguistic and computer science communities. It contains isolated videos of Turkish Sign Language glosses from three different domains: Health, finance and commonly used everyday signs. Videos in this dataset were performed by six native signers, as shown in Figure 1, which makes this dataset valuable for user independent sign language studies.

All of the sign video recordings in the dataset were captured using Microsoft Kinect v2 (Zhang, 2012) with 1080p (1920x1080 pixels) video resolution at 30 frames per second. We believe having a higher resolution is essential for sign language recognition when interpreting the appearance information related to hand shape and movements. All of the videos share the same recording setup where signers stood in front of a Chroma-Key background which is 1.5 meter far away from the camera.

Specifications of the BosphorusSign**22k** dataset can be seen in Table 2. Since the dataset was collected using Microsoft Kinect v2, we provide RGB video, depth map and skeleton information of the signer for all sign videos in the dataset. Moreover, we also provide OpenPose (Cao et al., 2018) joints, which include facial landmarks and hand joint positions in addition to body pose information. An example of provided modalities of the BosphorusSign**22k** dataset can be seen in Figure 2.

BosphorusSign**22k** has a vocabulary of 744 sign glosses; 428 in Health while having 163 in Finance domains as well as another 174 commonly used sign glosses. Properties of the proposed dataset and how it differentiates from the BosphorusSign corpus can be found in Table 3.

[1] `https://www.bosphorussign.com`

Table 1: Publicly available Isolated Sign Language Recognition datasets

| Dataset | Sign Language | #Signers | Lexicon | Repetitions | #Clips | All Native Signers | Data Source |
|---|---|---|---|---|---|---|---|
| ASLLVD (Neidle et al., 2012) | American | 6 | 2,742 | arbitrary | 9,794 | Yes | RGB |
| Devisign (Chai et al., 2014) | Chinese | 8 | 2,000 | 1-2 | 24,000 | No | Kinect v1 |
| BosphorusSign (Camgoz et al., 2016c) | Turkish | 6 | 855 | 4+ | 22,670 | Yes | Kinect v2 |
| CSL (Zhang et al., 2016) | Chinese | 50 | 500 | 5 | 125,000 | No | Kinect v2 |
| MS-ASL (Joze and Koller, 2018) | American | 222 | 1,000 | arbitrary | 25,513 | Yes | RGB |
| **BosphorusSign22k** | **Turkish** | **6** | **744** | **4+** | **22,542** | **Yes** | **Kinect v2** |

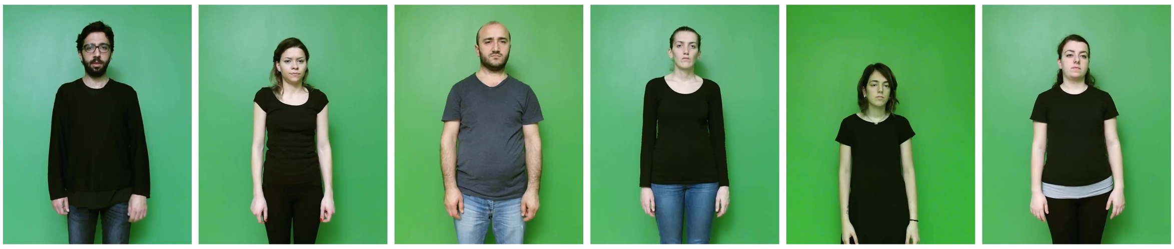

Figure 1: Native signer participants of the BosphorusSign**22k** dataset.
(We propose using the left-most five signers as the training set and keep the remaining for evaluation.)

Table 2: Specifications of the BosphorusSign**22k** dataset.

| Property | Description |
|---|---|
| Number of sign classes | 744 |
| Number of signers | 6 |
| Number of videos | 22,542 |
| Total Duration | ∼19 hours (∼2M frames) |
| RGB Resolution | 1920 x 1080 pixels |
| Depth Resolution | 512 x 424 pixels |
| Frame Rate | 30 frames/second |
| Body Pose Information (Kinect v2) | 25 x 3D Keypoints |
| Body Pose Information (OpenPose) | 25 x 2D Keypoints |
| Facial Landmarks (OpenPose) | 70 x 2D Keypoints |
| 2 x Hand Pose Information (OpenPose) | 21 x 2D Keypoints |

Table 3: Properties of the publicly available subsets of the BosphorusSign corpus and the proposed BosphorusSign**22k** datasets.

| Dataset | Lexicon | # Clips | # Repetitions |
|---|---|---|---|
| HospiSign (Camgoz et al., 2016b) | 33 | 1,257 | 6-8 |
| BosphorusSign (Camgoz et al., 2016c) | 855 | – | – |
| - Publicly Available | 595 | 22,670 | 4+ |
| BosphorusSign22k | 744 | 22,542 | 4+ |
| - General | 174 | 5,788 | 4+ |
| - Finance | 163 | 4,998 | 4+ |
| - Health | 428 | 11,756 | 4+ |

In this work, we changed several aspects of the BosphorusSign dataset. First of all to set a baseline that would extend over the whole dataset, we merged all subsets and conducted our experiments accordingly. Further details of our experimental protocol can be found in the Section 4.1. Secondly, we manually inspected all the sign videos and eliminated erroneous recordings. Furthermore, we split signs that were semantically same but morphologically different. We also collapsed signs with similar manual features. The goal of this change was to benchmark the capabilities of the state-of-the-art models on learning meaningful representations for manual aspects of the sign glosses. However, we will be also releasing an uncollapsed version of the dataset. The changes on the BosphorusSign dataset are mostly focused on improving and cleaning the dataset and defining an evaluation protocol. The dataset will be publicly available for research purposes upon submitting an EULA to the authors.

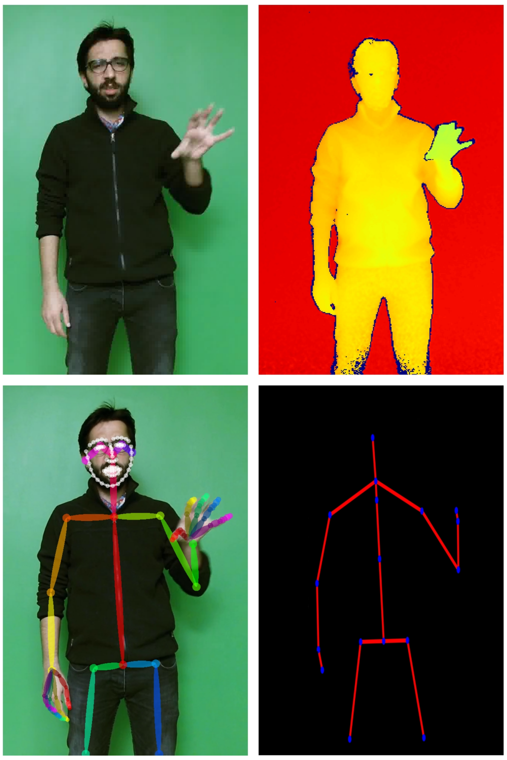

Figure 2: Modalities of BosphorusSign**22k**.
Top: (left) RGB frame and (right) depth image
Bottom: (left) OpenPose and (right) Kinect v2 outputs.

## 4. Baseline Recognition Methods and Experiments

In this section, we provide training and test protocols of the BosphorusSign**22k** dataset. We then describe baseline methods in detail and share our extensive experimental results.

### 4.1. Experiment Protocol

We defined our protocol by dividing the BosphorusSign**22k** dataset into training and test sets in a signer independent manner where we use one signer for test and others for training. This yielded us a test set of 4,524 sign samples and a training set containing 18,018 samples. To set a baseline on the new dataset and the evaluation protocols, we perform isolated sign language recognition experiments and report classification accuracies on the test set with only one signer.

### 4.2. Baseline Methods

As for benchmark methods, there is no agreed-upon state-of-the-art approach in the isolated SLR literature, and researchers use different methods on different datasets (Camgoz et al., 2017; Joze and Koller, 2018; Zhang et al., 2016). However, in the related field of action recognition, researches rely on benchmark datasets to compare their approaches against the state-of-the-art. Both 3D ResNets (Tran et al., 2018) and Improved Dense Trajectories - IDT (Wang and Schmid, 2013) are comparable state-of-the-art methods for action recognition, which have also yielded good performance on SLR (Özdemir et al., 2016). Therefore, we have chosen 3D ResNets and IDT as our baseline approaches to cover both deep learning based representation learning techniques as well as hand crafted feature based methods.

**Improved Dense Trajectories - IDT:** Although deep learning based models have become very popular recently, handcrafted approaches are still representative and competitive enough to be used in video recognition problems such as human action recognition and sign language recognition (Tran et al., 2018). Inspired by this and to also give further insight to the reader instead of just reporting baselines using a deep learning based approach, we have used Improved Dense Trajectories (IDT) (Wang and Schmid, 2013) which is one of the most successful handcrafted methods with competitive performance for human action recognition and was used in sign language and gesture recognition recently (Özdemir et al., 2016; Peng et al., 2015). IDT extracts local spatial features Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2005), and local temporal features Histograms of Optical Flow (HOF) (Laptev et al., 2008) and Motion Boundary Histograms (MBH) (Dalal et al., 2006) from the trajectories computed from dense optical flow field.

For recognizing sign language videos from the BosphorusSign**22k** dataset, we have used a recognition pipeline similar to the one proposed by Wang and Schmid (2013). We first extracted trajectories from every sign video. After extracting trajectories, we randomly sampled trajectories from the training set, assuming these trajectories represent the overarching distribution. Then, Principal Component Analysis (PCA) was applied to each component; namely HOG, HOF and MBH. Using PCA outputs, we performed Gaussian Mixture Model (GMM) to cluster each component of the trajectories. Finally, Fisher Vectors (FVs) were computed from each component of trajectory descriptors from each sign video using the parameters of PCA and GMM. Using these representations we trained a Linear Support Vector Machines (SVMs) using different combinations of concatenated FVs components.

**3D Residual Networks with Mixed Convolutions:** With the recent success of deep learning (Goodfellow et al., 2016) on tasks such as image and object recognition (Krizhevsky et al., 2012), researchers have also started to build deep architectures for human action recognition where both spatial and temporal dimension are exploited (Simonyan and Zisserman, 2014; Tran et al., 2015; Carreira and Zisserman, 2017). In this work, we have used a recently proposed video recognition method, which is based on 3D ResNet architecture with mixed 2D-3D convolutions, also called MC3 in Tran et al. (2018). The model consist of two residual blocks with 3D convolutions, three residual blocks with 2D convolutions and a fully connected layer as its classification layer. This network was built based on the hypothesis that learning temporal dynamics is beneficial in early layers while the higher levels semantic knowledge can be learnt in late layers (Tran et al., 2018).

As our second baseline, we trained MC3 models and investigated the effects of fine-tuning different residual blocks of the network using the Kinetics-400 dataset (Carreira and Zisserman, 2017). The proposed training method for this model used randomized clips of frames as inputs, which is not suitable for our problem because randomized clips may include different non-recurring parts with the same isolated sign gloss. Therefore, at the training phase, we randomly sampled batches of uniformly sampled frames from sign videos to give our networks sufficient coverage over the frames.

### 4.3. Implementation Details

To evaluate baseline methods on the BosphorusSign**22k** dataset, we have used the publicly available implementation in Wang and Schmid (2013) for extracting IDT features and PyTorch (Paszke et al., 2017) implementation of 3D ResNet model with mixed convolutions.

During training, we uniformly sampled 16 frames form each sign gloss with sizes of 112x112 pixels before feeding them to our networks. Sign clips are horizontally flipped a probability of 0.5 to be able to generalize over signers with different dominant hands. After preprocessing, we train our network using Adam optimizer (Kingma and Ba, 2014) with batch size of 32 on a NVIDIA Tesla V100 GPU. For testing, we performed the same preprocessing approach as in training except horizontal flipping of frame clips.

### 4.4. Experimental Results

We start our experiments by training several IDT based model with varying feature components, results of which can be seen in Table 4. Our results have shown that using motion features separately, HOF and MBH, has yielded better results (83.29% and 86.63%) than using only appearance features, trajectory information (63.68%) and HOG

Table 4: Baseline IDT results on the BosphorusSign**22k** dataset

| Method | Top-1 Acc (%) |
|---|---|
| TRAJ | 63.68 |
| HOG | 76.59 |
| HOF | 83.29 |
| MBH | 86.63 |
| HOG + HOF | 86.89 |
| HOG + MBH | 86.98 |
| HOF + MBH | 87.33 |
| HOG + HOF + MBH | **88.53** |
| TRAJ + HOG + HOF + MBH | 87.86 |

(76.59%). Since the trajectory information only contains the position of the trajectory (mostly the position of hand regions in our case), it is expected that it cannot fully represent motion or appearance based features which have higher dimensionality and cannot encode more complex information about the sign or hand shape.

Moreover, in the case of fusion of the trajectory components, our experiments have shown that using only HOG, HOF and MBH features together improves our recognition accuracy (88.53%), while adding the trajectory (TRAJ) information sightly decreases the performance of our system (87.86%). Although the performance of our system is very close in the case of fusing multiple components, we can see that using HOG features with other features has improved our recognition performance in all cases. This supports the idea that appearance representation obtained using hand crafted features, such as HOG, is useful along with the temporal information when recognizing signs where the manual features of the sign are the main differentiating aspect between target classes.

As for our deep learning baseline, we performed several experiments on fine-tuning different residual blocks of the MC3 network (Tran et al., 2018). In our first experiment, we first compared training our networks from scratch against using a pre-trained network (on Kinetics-400 dataset) and only training the final fully connected (fc) layer.

Table 5: Baseline 3D ResNet results on the BosphorusSign**22k** dataset

| Method | Top-1 Acc (%) | Top-5 Acc (%) |
|---|---|---|
| Training from scratch | 57.76 | 84.22 |
| Training only the last fc | 55.03 | 81.98 |
| Fine-tuning last 2 blocks | 75.38 | 94.16 |
| Fine-tuning last 3 blocks | **78.85** | **94.76** |
| Fine-tuning last 4 blocks | 63.88 | 88.66 |
| Fine-tuning all layers | 71.02 | 92.51 |

As it can be seen in Table 5, training the network from scratch performed slightly better. We believe this is due to the fact that pre-trained network has never seen any sign samples, hence some of the essential spatio-temporal information that forms the sign is lost until it reaches the final fully connected layer. Using this insight, we decided to fine-tune other layers of the pre-trained network in additional to the last fully connected layer. We found that fine tuning the last 3 blocks to yield the best results for our task, an Top-1 accuracy of 78.85% and and a Top-5 accuracy of 94.76% on the test set.

## 5. Analysis of Results and Discussion

Although the 88.53% Top-1 accuracy achieved by IDT is quite high for a 744 class problem, there is certainly still room for improvement. On the other hand, 3D ResNets, which are general-purpose video classification algorithms perform worse. We believe this is due to their inability to model longer-range temporal characteristics. Possible improvements include additional modalities and better temporal modeling.

We further investigated false classifications to gain further insight. For example, INSURANCE, INTERNET and COLD sign glosses are commonly confused with FUND, TEACHING and FACE respectively. Upon investigation we discovered that this is due to baseline methods' inability to model fine grained hand shapes. As seen in Figure 3, while our models were able to distinguish the sign by its motion, it fails to discriminate it using the similar handshapes. We believe this is due to the image resolution our networks are trained for in the case of MC3 model and the representation limitations of the HOG features for our IDT baseline. One way to tackle this problem could be to utilize specialized networks, such as Deep Hand proposed by Koller et al. (2016a), and use it as another modality in our recognition pipeline.

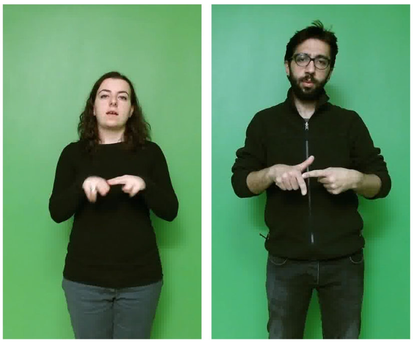

Figure 3: Test sample of INSURANCE sign gloss (left) misclassified as FUND (right).

In addition, our analysis have shown that PRICE sign gloss is confused with SHOPPING sign gloss (see Figure 4) because the number of repetitions of the same motion subunit is different in both signs. Although both signs have the same hand shape and movements, signers performing the SHOPPING sign gloss repeat the same motion sub-unit more than PRICE.

Furthermore, when looking at the Top-5 recognition accuracy on experiments with 3D ResNets, we can see that most of the misclassified signs are successfully classified among the top-5. Thus, we believe that focusing on problems mentioned above will help us to improve recognition performance. Baseline results also show that IDT, as a handcrafted approach, is still performing well on SLR as it can comprehensively model appearance and motion information obtained from the signer in the frame compared

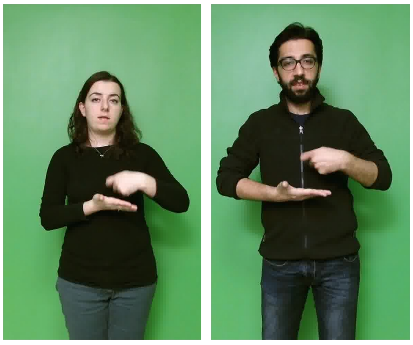

Figure 4: Test sample of PRICE sign (left) misclassified as SHOPPING (right)

to the 3D ResNet model trained without any specific guidance. Another factor contributing to this performance disparity is the input size which is 112x112 for MC3 networks and 640x360 for the IDT.

## 6. Conclusion

In this paper, we present BosphorusSign**22k**, a new signer-independent SLR evaluation benchmark. The dataset contains over 22k samples of isolated videos, of 744 unique Turkish Sign glosses performed by six native signers. To underpin future research, we applied two successful video recognition methods from the literature, namely IDT and 3D ResNets (MC3). We share our quantitative results as well as qualitative samples, providing further insight to the reader.

As shown by our experimental results, there is still room for improvement in signer-independent SLR for cases where the manual aspects of the sign subtlety differentiates from other classes. As future work we plan to exploit the capture setup of our dataset, namely its suitability for data augmentation, and extend our protocol to investigate environment independent SLR. BosphorusSign**22k** also enables further research into using the depth information to explore multi-modal fusion approaches.

## 7. Acknowledgements

This work has been supported by the TUBITAK Project #117E059 and TAM Project #2007K120610 under the Turkish Ministry of Development.

## 8. Bibliographical References

Aran, O. (2008). *Vision-based Sign Language Recognition: Modeling and Recognizing Isolated Signs with Manual and Non-Manual Components*. Ph.D. thesis, Bogazici University.

Bahar, P., Bieschke, T., and Ney, H. (2019). A comparative study on end-to-end speech to text translation. *arXiv preprint arXiv:1911.08870*.

Bahdanau, D., Bosc, T., Jastrzębski, S., Grefenstette, E., Vincent, P., and Bengio, Y. (2017). Learning to Compute Word Embeddings on the Fly. *arXiv:1706.00286*.

Camgoz, N. C., Hadfield, S., Koller, O., and Bowden, R. (2016a). Using convolutional 3d neural networks for user-independent continuous gesture recognition. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 49–54. IEEE.

Camgoz, N. C., Kindiroglu, A. A., and Akarun, L. (2016b). Sign Language Recognition for Assisting the Deaf in Hospitals. In *Proceedings of the International Workshop on Human Behavior Understanding (HBU)*.

Camgoz, N. C., Kindiroglu, A. A., Karabuklu, S., Kelepir, M., Ozsoy, A. S., and Akarun, L. (2016c). BosphorusSign: A Turkish Sign Language Recognition Corpus in Health and Finance Domains. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Camgoz, N. C., Hadfield, S., Koller, O., and Bowden, R. (2017). Subunets: End-to-end hand shape and continuous sign language recognition. In *IEEE International Conference on Computer Vision (ICCV)*.

Camgoz, N. C., Hadfield, S., Koller, O., Bowden, R., and Ney, H. (2018). Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020). Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

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

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

Cui, R., Liu, H., and Zhang, C. (2017). Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.

Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *Computer Vision – ECCV 2006*, pages 428–441. Springer Berlin Heidelberg.

Ebling, S., Camgoz, N. C., Braem, P. B., Tissi, K., Sidler-Miserez, S., Stoll, S., Hadfield, S., Haug, T., Bowden, R., Tornay, S., Razavi, M., and Magimai-Doss, M. (2018). SMILE Swiss German Sign Language Dataset. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT press.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Hanke, T., König, L., Wagner, S., and Matthes, S. (2010). DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In *Proceedings of the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*.

He, J., Liu, Z., and Zhang, J. (2016). Chinese sign language recognition based on trajectory and hand shape features. In *2016 Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE.

Huang, J., Zhou, W., Li, H., and Li, W. (2015). Sign language recognition using 3d convolutional neural networks. In *2015 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE.

Joze, H. R. V. and Koller, O. (2018). MS-ASL: A large-scale data set and benchmark for understanding american sign language. *CoRR*, abs/1812.01053.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Koller, O., Ney, H., and Bowden, R. (2016a). Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Koller, O., Zargaran, S., Ney, H., and Bowden, R. (2016b). Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Koller, O., Camgoz, N. C., Bowden, R., and Ney, H. (2019). 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)*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Liu, T., Zhou, W., and Li, H. (2016). Sign language recognition with long short-term memory. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2871–2875. IEEE.

Metaxas, D., Dilsizian, M., and Neidle, C. (2018). Linguistically-driven framework for computationally efficient and scalable sign recognition. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Neidle, C., Thangali, A., and Sclaroff, S. (2012). Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC*. Citeseer.

Özdemir, O., Camgöz, N. C., and Akarun, L. (2016). Isolated sign language recognition using improved dense trajectories. In *2016 24th Signal Processing and Communication Application Conference (SIU)*, pages 1961–1964. IEEE.

Özdemir, O., Kindiroglu, A. A., and Akarun, L. (2018). Isolated sign language recognition with fast hand descriptors. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Peng, X., Wang, L., Cai, Z., and Qiao, Y. (2015). Action and gesture temporal spotting with super vector representation. In *Computer Vision - ECCV 2014 Workshops*, pages 518–527. Springer International Publishing.

Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., and Cormier, K. (2013). Building the British Sign Language Corpus. *Language Documentation & Conservation (LD&C)*, 7.

Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., et al. (2012). Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(12):2821–2840.

Simonyan, K. and Zisserman, A. (2014). Two-stream Convolutional Networks for Action Recognition in Videos. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

Starner, T., Weaver, J., and Pentland, A. (1998). Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on pattern analysis and machine intelligence*, 20(12):1371–1375.

Stoll, S., Camgoz, N. C., Hadfield, S., and Bowden, R. (2018). Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Stoll, S., Camgoz, N. C., Hadfield, S., and Bowden, R. (2020). Text2Sign: Towards Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision (IJCV)*.

Sutton-Spence, R. and Woll, B. (1999). *The Linguistics of British Sign Language: An Introduction*. Cambridge University Press.

Suzgun, M. M., Ozdemir, H., Camgoz, N. C., Kindiroglu, A., Basaran, D., Togay, C., and Akarun, L. (2015). Hospisign: An Interactive Sign Language Platform for Hearing Impaired. In *Proceedings of the International Conference on Computer Graphics, Animation and Gaming Technologies (Eurasia Graphics)*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Varol, G., Laptev, I., and Schmid, C. (2017). Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517.

Wan, J., Escalera, S., Anbarjafari, G., Jair Escalante, H., Baró, X., Guyon, I., Madadi, M., Allik, J., Gorbova, J., Lin, C., et al. (2017). Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3189–3197.

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. In *2013 IEEE International Conference on Computer Vision*, pages 3551–3558. IEEE.

Wang, H., Chai, X., Hong, X., Zhao, G., and Chen, X. (2016). Isolated Sign Language Recognition with Grassmann Covariance Matrices. *ACM Transactions on Accessible Computing*, 8(4).

Wang, H., Chai, X., and Chen, X. (2019). A novel sign language recognition framework using hierarchical grassmann covariance matrix. *IEEE Transactions on Multimedia*, 21(11):2806–2814.

Yuan, Q., Wan, J., Lin, C., Li, Y., Miao, Q., Li, S. Z., Wang, L., and Lu, Y. (2019). Global and local spatial-attention network for isolated gesture recognition. In *Chinese Conference on Biometric Recognition*, pages 84–93. Springer.

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

Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 19(2):4–10, April.