DATA DISCOVERY AND DOMAIN ADAPTATION FOR ISOLATED SIGN LANGUAGE RECOGNITION

by

Özgür Kara

A report submitted for EE492 senior design project class
in partial fulfillment of the requirements for the degree of
Bachelor of Science
(Department of Electrical and Electronics Engineering)
in Boğaziçi University

June 9th, 2022

Principal Investigator:
Dr. Murat Saraçlar

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisors Dr. Murat Saraçlar and Dr. Lale Akarun for their continuous support on my undergraduate research and thesis. Besides my advisors, I would like to thank to Alp Kındıroğlu for his insightful comments, encouragement and technical support on this project.

ABSTRACT

Sign language is a way used by deaf people to communicate with other peoples via hand motions and gestures. Sign Language Recognition (SLR) has the goal of bridging the gap between sign language users and others by automatically recognizing signs. Recently, for SLR, researchers have been attracted by skeleton based action recognition since the background and the subject becomes independent. Some methods have been proposed to use pose estimators to obtain better models, however one of the latest works reveals the superiority of Graph Convolutional Networks. Research on sign language recognition with current state of the art methods yields very high recognition accuracy. However, this is not always the case, when working with small datasets or when moving to real-time applications from large, high quality datasets. The limited amount of training data for the sign recognition task may lead to overfitting or otherwise restrict the performance of SLR models in real-world scenarios. In this paper, we present a selection of identical & similar signs in two public isolated Sign Language datasets to create a dataset where we can use supervised domain adaptation methods to measure transferrablity of learned classification models. We perform an in-depth analysis on the effectiveness of domain adaptation techniques. We also do an in-depth analysis on sign properties to analyze how much they effect the performance of the models.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Sign Language Recognition

Sign language is a way for deaf people to communicate with each other. Sign language is different from other spoken languages because signers convey information visually, that is, using the movement of their hands and body. Generally, isolated sign recognition and continuous sign recognition are the two types of sign language recognition.

Researchers have been studying to automate the recognition of sign language utilizing the advancement in computer vision. Traditional Sign Language Recognition (SLR) methods mainly employ feature extraction and localization methods such as HOG [1] and SIFT [2] associated with linear classifiers such as SVM [3]. As deep learning technology has been advancing and outperforms previous approaches, researchers have been attracted by this progress and oriented their attraction towards deep learning based methods. For SLR, general video representation learning approaches such as RNN and LSTMs, and 3D CNNs are widely used to achieve good performance [4,5]. To develop a more effective approach, local motion information and attention frameworks are incorporated for better results in terms of accuracy.

Some of the recent works have developed skeleton-based methods, which provides complementary information to RGB formatted image inputs and resulted in even better performances [6.7.8]. One of the recent work, SL-GCN [9], has utilized Graph Convolution Network (GCN), inspired by the state-of-the-art body pose estimation studies. They have shown very good results in terms of both its accuracy and fastness.

In this project, we propose to enhance the performance of the state-of-the-art deep learning method for isolated sign language recognition on BosphorusSign22K dataset using transfer learning methods. Transfer learning approaches aim to use data from a data-rich source task to improve task performance on a data-poor target task. We establish a sign language recognition baseline for cross-dataset sign language recognition using 2 public isolated Turkish Sign Language datasets.

1.2 Domain Adaptation and Transfer Learning

Domain adaptation is a field that aims to train a deep learning network on a "source" data distribution, which generally consists of many amount of labeled-data, such that it can transfer its knowled to perform well on a different "target" data distribution, which in general lacks labeled data. As an example, let us assume we have two car datasets, each of which has 10 different classes and each car in the first dataset is blue-colored, whereas it is red-colored in the second dataset. In other words, the data distribution is different. However, suppose, the second dataset does not contain the class information. Hence, one approach is to train a model using the first dataset in a supervised setting, then test and use it on the second dataset, where we do not have the classes. Domain adaptation is a type of transfer learning which proposes solutions to the above-described problem.

Let X and Y be our input and label spaces. What machine learning model does is to approximate a function f that maps X to Y, i.e. $f: X \rightarrow Y$. This model is learned in a data-driven fashion, where each sample can be represented as the following set: $S = \{(x_0, y_0), (x_1, y_1), \dots, (x_m, y_m)\}$ where m is the sample size and $(x_i, y_i) \in X \times Y$. This is the formulation for supervised learning. The difference between supervised learning and domain adaptation is that in domain adaptation, we have 2 sample distributions namely source and target denoted as D_S and D_T on $X \times Y$. The domain adaptation algorithms attempt to transfer the knowledge from the source domain to target domain, and the objective is to learn function f such that we get good results for target domain.

Compared to image-based domain adaptation (DA), Video-based DA attracted fewer studies. In video classification, models have to take into account the temporal variations on top of variations in the image space. Only a few works focus on small-scale video DA with only a few overlapping categories [21,22]. Several methods aim to use image-based domain adaptation methods with 3D classification networks, such as adding a gradient reversal layer for domain invariance [23]. The TA3N study [24] proposes a method using domain-specific attention while learning to align frames across domains. They demonstrate the improvement in performance by introducing the UCF-HMDB full and Kinetics-Gameplay datasets, providing a benchmark for cross-dataset transfer learning in action recognition.

In this project we establish a sign language recognition baseline for cross-dataset sign language recognition using 2 public isolated Turkish Sign Language datasets. These datasets share 57 common gestures. However not all gestures are identical as their performance varies from dataset to dataset due to differences in interpretation and style.

CHAPTER 2

METHODOLOGY

2.1 Domain Adaptation Methods

2.1.1 Domain Adversarial Training of Neural Networks (DANN) [10]

The first approach is “domain adversarial training of neural networks” method. This approach combines representation learning and unsupervised domain adaptation and proposes an end-to-end training model. It implements the idea that the model should not learn to discriminate between the source and target domains. To achieve this, the model jointly optimizes three training processes: 1- minimizing the loss coming from label classifier (Ly), 2- maximizing the loss of a domain classifier so that model does not learn how to discriminate them (they achieve this by adding a gradient reversal) 3- optimizing the feature extractor.

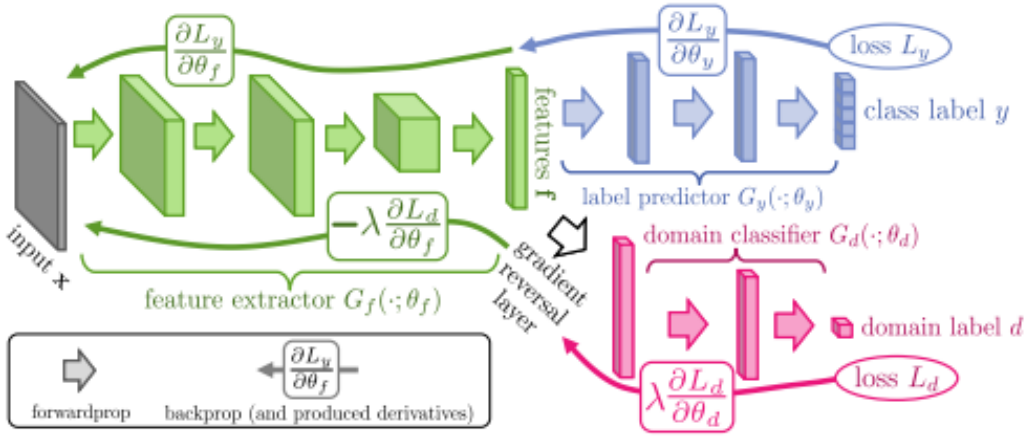

Figure 1. The proposed framework for DANN

2.1.2 Minimum Class Confusion (MCC) [11]

The second approach introduces a loss function that is calculated using the predictions in an unsupervised manner. The class confusion tries to explore the missing piece in

existing methods, the tendency that a model confuses the predictions between the correct and ambiguous classes for target examples. Inspired by this fact, they propose a general loss function, which can be characterized as a non-adversarial domain adaptation method with a high convergence speed. It calculates the class correlation of the predictions, and tries to minimize the non-diagonal elements, equivalent to maximizing the diagonal elements to prevent confusion.

2.1.3 Domain Specific Batch Normalization (DSBN) [16]

The DSBN approach shares all model parameters except batch normalization layers within the feature encoder. The batch normalization layer in a neural network regularizes feature representations from different domains without taking into account class or domain information. However, when domain discrepancy is significant, the effect of batch normalization is diminished. In this approach, individual batch normalization layers keep track of unique normalization parameters and batch statistics values for each domain.

2.1.4 Joint Adaptation Network (JAN) [17]

The JAN approach proposed in [17] maps features from different domains into a new data space where inter-class features have a more significant similarity. The method proposed, named joint maximum mean discrepancy (JMMD), minimizes the joint probability distribution distance of the source and target class-specific layers. The approach adds task-specific layers on top of the base SL-GCN network to learn mapping a common domain.

2.2 Datasets

2.2.1 BSIGN22K [12]

The dataset on which we mainly conduct our experiments is the BosphorusSign22k benchmark, which contains a vocabulary of 744 sign classes, performed by 6 native signers. Overall, the total number of videos is 22.542, which are recorded with Microsoft Kinect v2 – 30 FPS. This dataset is used as our target dataset.

We selected the samples belonging to User 4 as our test set as selected in the original paper. However, to better understand how well domain adaptation networks work, we also added the samples belonging to User 3 to our test set. Figure 2 shows the learning curves when our network is trained only with the samples from corresponding (title) indexed User samples and tested on Bsign22k User 4. As a result of the very low performance of User 3, we concluded that this user is the most different one, that is why we perform our testings on this user's samples for all other experiments.

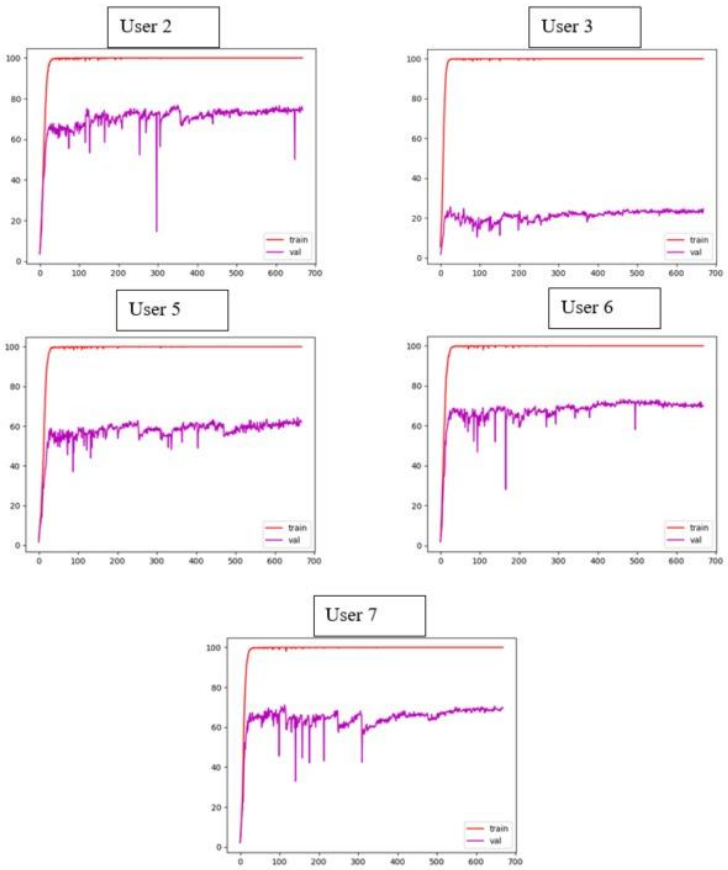

Figure 2. Learning curves for the test sets composed of each user at each subplot

2.2.2 AUTSL [13]

The source dataset that we use is Ankara University Turkish Sign Language Dataset (AUTSL) , a comparably large multimode dataset, and is composed of isolated Turkish sign videos. It consists of 38.336 isolated sign video samples, including 226

signs, which are performed by 43 different signers in total. Samples are recorded with Microsoft Kinect v2, and dataset provides RGB, depth and skeleton data. Note that these two datasets have 57 common classes.

2.3 Backbone Network

2.3.1 SL-GCN

In order to solve the multi-class classification task, we utilized Graph Convolution Networks due to their speed and high performance. For each sample video, we extract the whole body skeleton graph and feed the GCN with this graph that are composed of keypoints. The SL-GCN block is consturced with spatial convolutional network, self-attention and graph dropping module.

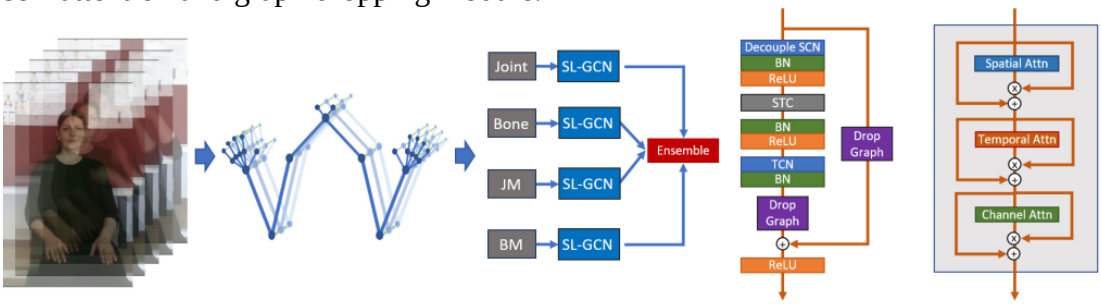

Figure 3. SL-GCN Framework

Instead of 3D CNNs [14], we performed our experiments with SL-GCN because Table 1 shows the effectiveness of SL-GCN.

Dataset	AUTSL	Bsign22k	AUTSL57	Bsign57
SL-GCN	91.22	89.25	97.78	92.97
Resnet 2+1D	84.65	81.2	92.3	88.26

Table 1. Top-1 Accuracy results obtained by using each datasets corresponding training and test sets are reported.

In Table 1, single domain accuracy results from each dataset are reported using both SL-GCN and 3D Convolution Based Neural Networks. In nearly all experiments, SL-GCN outperforms rgb based 3D Neural Network based experiments. For this reason, the remainder of this paper uses the SL-GCN baseline method to report accuracy numbers on Bsign57.

For isolated sign language recognition, we have a set of N training examples from each sign language dataset T as D_i^t, L_i^t, where D_i^t is a set of coordinates D in (J,T,3) extracted from a single isolated SLR video clip, J is the number of joints, T is clip length, and the 3 values are the detected X, Y coordinates and the detection confidence values for each joint at each timestep. We obtain J=30 joints belonging to fingertips, finger bases, wrists, arms, neck, mouth, nose, and eyes for each frame of a sign language video.

Having obtained the joints for each gesture, we apply several normalizations and augmentations to make our models more robust to small changes in user performance. A typical property of the isolated SLR datasets we use is that they both share the same rest pose where hands rest to the side of the users' legs. Frames that contain stationary hands in this rest pose are trimmed from the beginning and end of each video segment. In addition, further sampling is done from the sampled frames to bring the number of total frames sampled from each video to J frames. This was done using a random uniform sampling from a set of fixed intervals. This approach created duplicates of frames for shorter clips and provided temporal variation when sampling from longer clips.

For each coordinate, we apply a spatial normalization where the origin of the 2D coordinate system is moved to the temporal mean of the neck joint for that gesture. In addition, random horizontal mirroring and random spatial coordinate translation augmentations were found to improve recognition performance and were thus added to our pipeline.

The classifier base of our model is the SL-GCN model proposed by Jiang et. al [15]. The model takes input as a sequence of fixed length coordinates and outputs class prediction probabilities for sign language gestures.

We model the problem so that nodes of the graph correspond to joint landmark locations. The spatio-temporal graph adjacency matrix A is constructed in the spatial domain according to anatomical spatial ordering, where neighboring joints are assigned a value of 1 and all other joints are assigned a value of 0. In the temporal domain, all the joints are only connected to themselves.

The architecture of the model consists of ten SL-GCN blocks for node and edge processing. Each SL-GCN block consists of a spatial convolution layer, multiplication with the adjacency matrix, and temporal convolution layer. The model includes decoupling with drop graph module proposed in [19] and spatio temporal channel (STC) attention modules proposed in [20]. The decoupling adds increased recognition power by dropping random joints along with their neighbors closer than an adjacency of K in A. Similarly, the STC attention module increases recognition power by focusing more on important joints, frames, and coordinates for certain gestures.

,

CHAPTER 3

EXPERIMENTATION AND RESULTS

To make a comparison between domain adaptation methods and to understand whether the additional attribute labels are useful, we performed several experiments. Since it is a multi-class classification problem, we optimized our neural network model using categorical cross entropy loss. We used Adam optimizer having batch size of 128.

3.1 Experiment 1 – Training on Shared Classes

In the first experiment, we performed 5 different experiments namely target only, source only, combined, DANN and MCC. In all experiments, we only used the shared classes of these two datasets. In the target only, we only do training using our target dataset as the name suggests. In source only, at the training phase, we trained our model using samples from source dataset and tested it on our target dataset. In combined, we use both of the datasets in training phase and in DANN and MCC, we applied these adaptation methods to see how they effect the results. We used accuracy as our evaluation metric and measured on test set which is composed of the samples that belong to User 3 and User 4. Note that, we separated the training set of bsign57 dataset into 4 partitions namely Users 2, Users 2,5, Users 2,5,6, Users 2,5,6,7 where only the corresponding indexed users are used at the training.

	Train		**Test**	**Users**	**Users**	**Users**	**Users**
	Source	**Target**	**Target**	**2**	**2,5**	**2,5,6**	**2,5,6,7**
Target only	-	Bsign 57	Bsign 57	67.19	81.41	88.28	92.97
Source only	Autsl 57		Bsign 57	60.6			

Combined	Autsl 57	Bsign 57	Bsign 57	87.31	91.41	94.32	95.8
DANN	Autsl 57	Bsign 57	Bsign 57	**87.87**	**91.83**	94.04	96.54
MCC	Autsl 57	Bsign 57	Bsign 57	86.82	91.77	**95.68**	**97.15**

Table 2. Accuracy scores of domain adaptation methods when the samples from common classes is used for training

Of the experimented transfer learning methods, the Domain Adversarial Neural Network (DANN) approach yields the highest benefit with lower numbers of target samples, while minimum class confusion approach yields higher results when the amount of samples in the target class starts to increase. The benefit of closed set transfer learning is higher when the amount of samples in the target training set are minimal. This is the closest use case to improving real time recognition using transfer learning.

3.2 Experiment 2 – Training on Entired AUTSL

Secondly, we explore the transfer learning situation where a larger source dataset is used. The main difference from closed set transfer learning is that the source domain contains 57 common and 159 different classes. In this setting, the difference from transferring from an arbitrary sign language dataset is the fact that we are sure no gesture from the source dataset actually is identical to a gesture from the target dataset but has a different label.

	Train		**Test**	**Users**
	Source	**Target**	**Target**	**2,5,6,7**
Target only	-	Bsign 57	Bsign 57	92.97

Source only	Autsl 216		Bsign 57	71.23
Combined	Autsl 216	Bsign 57	Bsign 57	98.12
DANN	Autsl 216	Bsign 57	Bsign 57	98.19
MCC	Autsl 216	Bsign 57	Bsign 57	**98.63**

Table 3. Accuracy scores of domain adaptation methods when entire AUTSL is used for training

In Table 3, transfer learning from the larger AUTSL dataset is observed. The obtained accuracy numbers on Bsign57 reaches around %98 with the combined baseline and experimented transfer learning methods further improve upon this result.

3.3 Experiment 3 - Training Including Additional Attributes

In bsign22k dataset, we have extracted additional labels for each samples namely number of hands that are used while performing gestures, repetitiveness of the sign, and circular property of the sign. For each of the attributes, we have target labels of either 1 or 0, hence it can be considered as a multi-label classification problem. To enhance the performance of the network, we connect a second classifier layer to also make model predict these attributes, thus we added an additional loss function that takes multi labels. The loss function is selected as binary cross entropy and we performed experiments with different loss weights. In this experiment, we investigate the effects of adding these labels one-by-one. From Table 4 to Table 7, showing the test scores on target dataset when they are trained on bsign57 without using domain adapatation algorithms. It corresponds to “target-only” training procedure with additional multi-label classification.

Loss Weight	**No attributes**	**# of Hands**	**Repetitive**	**Circular**
1	92.97	90.24	90.41	85.51

5	92.97	89.7	87.66	82.92
10	92.97	89.34	86.11	83.9

Table 4. Trained on Users 2,5,6,7

Loss Weight	**No attributes**	**# of Hands**	**Repetitive**	**Circular**
1	88.28	85.39	86.44	85.02
5	88.28	88.1	83.46	79.09
10	88.28	83.83	83.48	78.75

Table 5. Trained on Users 2,5,6

Loss Weight	**No attributes**	**# of Hands**	**Repetitive**	**Circular**
1	81.41	79.77	80.73	75.85
5	81.41	78.43	76.6	72.7
10	81.41	78.24	71.79	72.28

Table 6. Trained on Users 2,5

Loss Weight	**No attributes**	**# of Hands**	**Repetitive**	**Circular**

1	67.19	69.78	67.35	69.84
5	67.19	68.01	67.4	61.49
10	67.19	68.52	66.69	62.64

Table 7. Trained on Users 2

From Table 4 to Table 7, we observe that when the training data gets smaller, then the additional attributes play an important role. In Table 7, number of hands attribute helps model learn better representations whereas when the training data gets larger, they worsen the performance of the model.

3.4 Experiment 4 - Fusion of Transfer Learning Methods

In this section, we explore the combinations of these classifiers with finetuning and each other on the BSign22k shared dataset. The fusion of these algorithms is achieved by initializing the feature extractor and classifier layers of the algorithm, freezing them for the first five epochs, and then applying respective model architecture loss combinations to a single model. Of the attempted methods, finetuning and MCC approaches yield the most promising results. In a greedy fashion, we combined this method with several other methods, which gave us 98.8\% accuracy on the BSign22k shared task of the dataset. Note that in this experiment, we selected 4 users (2,5,6,7) as our training set.

	Training		Validation
	Target	**Target**	$\text{BSign22k}_{\text{shared}}$
finetuning + DANN	AUTSL	$\text{BSign22k}_{\text{shared}}$	97.14
finetuning + MCC	AUTSL	$\text{BSign22k}_{\text{shared}}$	**98.59**
finetuning + JAN	AUTSL	$\text{BSign22k}_{\text{shared}}$	96.91
finetuning + DSBN	AUTSL	$\text{BSign22k}_{\text{shared}}$	98.19
finetuning + DANN + MCC	AUTSL	$\text{BSign22k}_{\text{shared}}$	92.57
finetuning + JAN + MCC	AUTSL	$\text{BSign22k}_{\text{shared}}$	84.33
finetuning + DSBN + MCC	AUTSL	$\text{BSign22k}_{\text{shared}}$	**98.8**

Table 8. Top-1 Accuracy results for fusion of transfer learning methods.

CHAPTER 4

CONCLUSION

4.1 Discussion and Future Work

Overall, we performed some experiments to understand how domain adaptation works in Turkish sign language datasets and we observed that DANN and MCC outperforms even combined approach. As the number of training samples for target dataset increases, we get better results for domain adaptation methods. Also, our experiments highlights the importance of additional attributes when training data is inadequate. We see that, as the number of training samples are decreasing, the additional attributes start playing an important role for the accuracy performance of the model.

Future work entails running additional experiments to understand how well the extra attributes are learned by the model when domain adaptation methods are applied. Furthermore, we want to investigate the performance when more than one domain adaptation methods are incorporated. By this way, using the different beneficial components of different domain adaptation algorithms, we aim to boost the performance of the model even better.

4.2 Social, Environmental and Economical Impact

As deaf people are not able to communicate with spoken language, sign language is a means of conveying information for them. To communicate with a deaf person is difficult for a hearing person as being proficient in sign language is difficult. As a result, there occurs a need for automatic interpretation of sign languages to solve this unfairness. Doing research and developing existing models for recognition of sign language for different cultures facilitate the communication between the hearing and the deaf individuals. It is one of the social impact of this project.

4.3 Cost Analysis

I will do the code implementations in my local PC. Therefore, I do not need to pay for any electrical or mechanical parts and devices. I am planning to work 15+ hours per week, if we consider the average hourly wage of a research and development engineer in USA, which is $53, then my salary would be $795 per week.

4.4 Standards

My project is related to the following IEEE standard:

“P3110 - Standard for Computer Vision (CV) - Algorithms, Application Programming Interfaces (API), and Technical Requirements for Deep Learning Framework: This standard establishes the application programming interfaces (API) model of the computer vision systems and specifies the functional and technical requirements of the API

between the computer vision algorithm, deep-learning framework, and the data set in the process of algorithm training phase. This standard is suitable for the adaptation and invocation of computer vision algorithms using deep learning frameworks.”

The definition clearly illustrates the points why this standard is a relevant one. I propose to solve a computer vision problem, which is isolated sign language recognition, using deep learning frameworks and dataset analysis. In addition, I follow the code of conduct during my research.

BIBLIOGRAPHY

[1] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, and Shai Avidan. Fast human detection using a cascade of histograms of oriented gradients. In Proceedings of IEEE Computer Vision and Pattern Recognition, volume 2, pages 1491–1498, 2006

[2] David G Lowe. Object recognition from local scale-invariant features. In Proceedings of IEEE International Conference on Computer Vision, volume 2, pages 1150–1157, 1999

[3] Abbas Memis¸ and Songul Albayrak. A Kinect based sign ¨ language recognition system using spatio-temporal features. In Proceedings of International Conference on Machine Vision, volume 9067, page 90670X. International Society for Optics and Photonics, 2013

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

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

[6] Jinmiao Cai, Nianjuan Jiang, Xiaoguang Han, Kui Jia, and Jiangbo Lu. Jolo-gcn: Mining joint-centered light-weight information for skeleton-based action recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2735–2744, 2021

[7] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1110–1118, 2015

[8] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3595–3603, 2019.

[9] Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y. (2021). Skeleton Aware Multi-modal Sign Language Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[10] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V.S. (2016). Domain-Adversarial Training of Neural Networks. J. Mach. Learn. Res..

[11] Jin, Y., Wang, X., Long, M., & Wang, J. (2020, August). Minimum class confusion for versatile domain adaptation. In *European Conference on Computer Vision* (pp. 464-480). Springer, Cham.

[12] "Ozdemir, O., Kindiroglu, A., Cihan Camgoz, N., & Akarun, L. (2020). BosphorusSign22k Sign Language Recognition Dataset. In Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives.

[13] Sincan, O. M., Keles, H. Y. "AUTSL: A Large Scale Multi-modal Turkish Sign Language Dataset and Baseline Methods". IEEE Access, vol. 8, pp. 181340-181355, 2020.

[14] Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. 2015 IEEE International Conference on Computer Vision (ICCV), 4489-4497.

[15] Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y. (2021). Skeleton aware multi-modal sign language recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3413–3423).

[16] Chang, W., You, T., Seo, S., Kwak, S., & Han, B. (2019). Domain-Specific Batch Normalization for Unsupervised Domain Adaptation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7346-7354.

[17] Long, M., Zhu, H., Wang, J., & Jordan, M. (2017). Deep transfer learning with joint adaptation networks. In International conference on machine learning (pp. 2208–2217).

[18] Long, M., Zhu, H., Wang, J., & Jordan, M. (2017). Deep transfer learning with joint adaptation networks. In International conference on machine learning (pp. 2208–2217).

[19] Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., & Lu, H. (2020). Decoupling gcn with dropgraph module for skeleton-based action recognition. In European Conference on Computer Vision (pp. 536–553).

[20] Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2020). Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Transactions on Image Processing, 29, 9532–9545.

[21] Chen, M.H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., & Zheng, J. (2019). " alignment for large-scale video domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6321–6330).
[22] Xu, Y., Yang, J., Cao, H., Mao, K., Yin, J., & See, S. (2021). Aligning Correlation Information for Domain Adaptation in Action Recognition. arXiv preprint arXiv:2107.04932
[23] Bellitto, G., Proietto Salanitri, F., Palazzo, S., Rundo, F., Giordano, D., & Spampinato, C. (2021). Hierarchical domain-adapted feature learning for video saliency prediction. International Journal of Computer Vision, 129(12), 3216–3232.
[24] Chen, M.H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., & Zheng, J. (2019). Temporal attentive alignment for large-scale video domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6321–6330).