SIHeDA-Net: Sensor to Image Heterogeneous Domain Adaptation for Sign Language Detection

Ishikaa Lunawat ISHIKAA.NITT@GMAIL.COM
Vignesh S VIGNESH.NITT10@GMAIL.COM
S P Sharan SPSHARAN2000@GMAIL.COM
Spider R&D, National Institute of Technology Tiruchirappalli

Abstract

The main advantage of wearable devices lies in enabling them to be tracked without external infrastructure. However, unlike vision (cameras), there is a dearth of large-scale training data to develop robust ML models for wearable devices. SIHeDA-Net (Sensor-Image Heterogeneous Domain Adaptation) uses training data from public images of American Sign Language (ASL) that can be used for inferences on sensors even with errors by bridging the domain gaps through latent space transfer. Our codes are open-sourced at: github.com/spider-tronix/SIHeDA-Net.

Keywords: Domain Adaptation, American Sign Language, Latent Space Transfer

1. Introduction

Motivated by the various applications of robotics and facilitated by recent advances in deep learning, there has been a surge of recent research in the field of gesture recognition. Gesture recognition can be purely visual or non-visual based, or a combination of the two. Nevertheless, one difficulty in gesture identification is making data-efficient predictions. Previous works on domain adaptation (Prabono et al., 2021), requires some commonality between domain-specific features which may not always be available. To this end, we present a novel label-efficient method of classification using heterogeneous domain transfer between sensor and image. Our concrete technical contributions can be summarized as follows:

- Improvement of performance on noisy/scarce data using heterogeneous domain transfer that captures similarities between data of different domains.
- A network, known as SIHeDA-Net, that learns to classify ASL alphabets, with noisy sensor data. This network is trained along with an image dataset (Sign-MNIST) for learning from the image domain to aid prediction on sensor data.
- Experiment to examine the effect of up-scaling the sensor data latent vectors rather than down-scaling image latent vectors by using a simple ANN auto encoder, called Sensor-AE.

2. Proposed Method

Sensor and Image Encoding: Figure 1 shows two encoder-decoder networks, Sensor-AE and the CNN-VAE. We use a CNN-VAE network that maps each sign language image $E_1 : \mathbb{I}^{64x64} \rightarrow \mathbb{Z}_i^{256x1}$. Here, we perform generic down-scaling of the image for computation purposes. To simulate errors encountered in real-time, we impart fault to sensor values at random samples in the sensor dataset and call that "corrupt" dataset. The second part of our model, Sensor-AE, up-scales the corrupt sensor data from $E_2 : \mathbb{S}_c^{16x1} \rightarrow \mathbb{Z}_{cs}^{256x1}$. The quality of the latent vectors generated by the Image Variational Autoencoder E_1 is ensured by checking the mean accuracy on the labels by passing the reconstructed samples through a ResNet-50 classifier trained on the Sign-MNIST dataset.

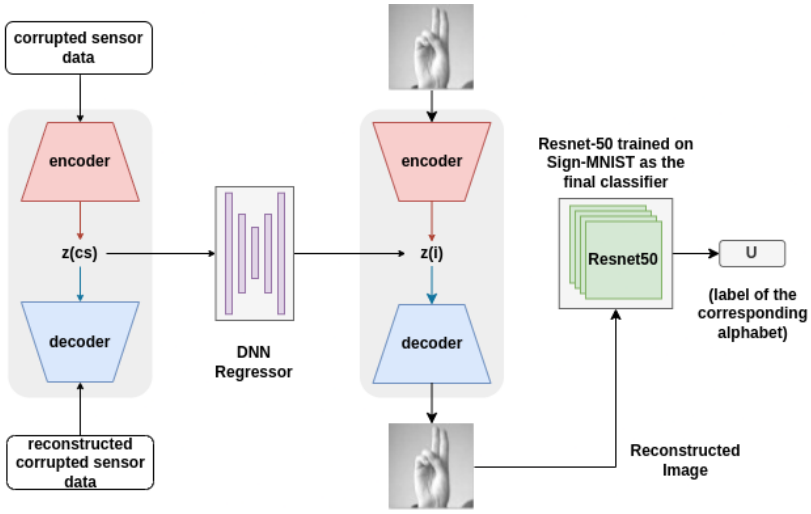

Figure 1: SIHeDA-Net architecture.

Latent Space Translation: As in Wan et al. (2020), a translation network is used to map vector $\mathbb{Z}_{cs}$ to $\mathbb{Z}_i$ and close the domain gap between the sensor and image latent vector to improve predictions on the sensor data itself. In Figure 1, the mapping is done between *z(cs)* and *z(i)* by training a Dense Regressor, $D_r : \mathbb{Z}_{cs}^{256x1} \rightarrow \mathbb{Z}_i^{256x1}$. For a sensor latent vector z_{cs}, D_r outputs a corresponding vector in the image latent space corresponding to the input sensor latent vector of the same label, $z_{cs}^{'} \in \mathbb{Z}_{cs}^{'256x1}$. The key motive is to minimize the loss between $z_{cs}^{'}$ and z_i.

Classification: The mapped sensor latent vector $z_s^{'}$ is passed through the decoder network of the Image VAE which is sent to a ResNet-50 (trained on Sign-MNIST) to predict the ASL alphabet. The entire pipeline when used in real-time, can be summarized as:

$$\textit{corrupt sensor data} \rightarrow \mathbf{z_{cs}} \rightarrow \mathbf{z_{cs}^{'}} \rightarrow \textit{Label} \tag{1}$$

3. Experiments and Discussions

Dataset: The image dataset is called "Sign-MNIST" [1]. Each training and test sample is an image that belongs to an alphabetical label (0-24). It is to be noted that samples for $9 = J$ or $25 = Z$ are omitted due to the presence of motion gestures in their corresponding representations. The training data consists of $27,455$ images, and the test data consists of 7172 images of 28×28 pixels belonging to the grayscale colour grade. The sensor dataset, on the other hand, is created with the intention of simulating a mini-potentiometer whose values are constructed by varying the mean of each sensor under the range $\in (-24,\ 23)$ with a standard deviation of 0.5. The dataset has 500 samples per class (24 alphabets).The corrupt/faulty dataset is formed by setting values to low (0) or high (1) at random in randomly chosen samples, to represent *stuck at 0/1 faults*. This to simulate the real-life mis-alignments that a sensor brings when trying to model a system. This will allow us to generalize our SIHeDA-Net to be diverse and accurate.

1. https://www.kaggle.com/datasets/datamunge/sign-language-mnist

Table 1: Accuracy metrics to validate the sub-networks

Evaluation Networks	Dataset	MSE Loss	Mean Accuracy
Sensor Encoder - **AE**	Corrupted Sensor ($\mathbb{Z}_{cs}$)	10^{-5}	–
Image Encoder - **VAE**	Sign-MNIST ($\mathbb{I}$)	11.21	–
Sensor Latent Classifier - **DNN**	Uncorrupted Sensors	–	0.9852
Image Classifier - **ResNet-50**	Sign-MNIST ($\mathbb{I}$)	–	0.9944
Image Classifier - **ResNet-50**	Recon. Sign-MNIST ($\mathbb{I}$)	–	0.8423

Table 2: Performance Comparison on the Corrupted Sensor Dataset ($z_{cs} \rightarrow$ Label)

Models	Mean Accuracy
Artificial Neural Network - ANN	0.3813
SIHeDA-Net	**0.7083**

Preliminary Results: In Table 1, the accuracy of the ResNet-50 (trained on Sign-MNIST) tested on the reconstructed Sign-MNIST samples is **99.94%**. The Sensor-AE (corrupted dataset) and Image-VAE show a mean-squared error loss of **11.21** and $\mathbf{10^{-5}}$ respectively. Table 2 shows that our SIHeDA-Net has an accuracy of **70.83%** on predicting labels from corrupted sensor dataset while a simple ANN baseline (without transfer of domain knowledge) scores only **38.13%** (showing a 85.75% increase over naive classification).

4. Conclusions

Through SIHeDA-Net, we see a significant improvement of predictions, given insufficient and corrupted training data. Latent space transfer between heterogeneous domains is an emerging field and has a great scope for performing classification when data is both deficit and partially corrupt. In our work, we explore the scope of exploiting correlations between starkly different domains and implicitly interpolating within their corresponding latent spaces to achieve heterogeneous adaptation in the absence of data feature commonality. Through our results, it is evident that SIHeDA-Net outperforms our baselines by about 85%. Our future works include employing graph neural networks for better representational modelling of hand-gestures, and also transitioning into ultra label/data-efficient fine-tuning regimes.

References

Aria Ghora Prabono, Bernardo Nugroho Yahya, and Seok-Lyong Lee. Hybrid domain adaptation for sensor-based human activity recognition in a heterogeneous setup with feature commonalities. *Pattern Analysis and Applications*, 24(4):1501–1511, 2021.

Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2747–2757, 2020.