Domain Adaptation with Contrastive Simultaneous Multi-loss Training for Hand Gesture Recognition

Joel Baptista [1] Vítor Santos [1] Filipe Silva [1] Diogo Pinho [2]

[1]Institute of Electronics and Informatics Engineering of Aveiro (IEETA)
[2]Bosch Termotecnologia, S. A.

CIÊNCIA 2024

Introduction

Hand gestures are an important aspect of human communication, serving several purposes such as enhancing spoken messages, signaling intentions or expressing emotions. Driven by technological advances, the process of classifying meaningful hand gestures, known as Hand Gesture Recognition (HGR), has received increasing attention in recent years.

The major application areas of gesture recognition include sign language translation, human-machine interaction, medical rehabilitation, and virtual reality. HGR systems also target robotic applications using a variety of input devices, among which stands out color cameras, depth sensors or gloves with embedded sensors. In this context, the ability of robots to recognize hand gestures seems very promising for progress in Human-Robot Collaboration (HRC), as these gestures are simple and intuitive for a human partner to produce. This, in turn, allows humans and robots to coexist and cooperate, improving task efficiency and safety.

However, HGR is an inherently challenging task due to the complex, non-rigid properties of the hand, such as its shape, color, and orientation. Vision-based HGR systems must also be robust to variations in lighting conditions, cluttered environments, complex backgrounds, and occlusions. In reality, the assumption that the training and test datasets are drawn from the same distribution rarely holds in practical situations due to domain shifts.

To tackle this issue, this paper proposes a domain adaptation technique for hand gesture recognition in human-robot collaboration scenarios. The proposed approach is based on a two-headed deep architecture that simultaneously adopts cross-entropy and a contrastive loss from different network branches to classify hand gestures.

Objectives

The main goal is to shed light on the impact of supervised contrastive learning (SCL) on the generalization ability of a trained deep model in gesture classification when faced with a distribution shift. For this purpose, the study:

- Contributes with a new RGB annotated dataset of hand gestures, with a complex background and multiple subjects.
- Compares the results from the proposed approach against baselines.

Dataset Description

This implementation uses four hand gesture classes inspired by American Sign Language; the symbols chosen are the "A", "F", "L" and "Y" signs. These signals have the advantage of being well known symbols with already real world applications, implying that they are easy to use. These specific signs were chosen also because they are relatively distinct.

To test the degree of generalization of the proposed method, two datasets were recorded. The first dataset was used to train the classification model. This dataset was recorded by one subject, which can be seen in Figure 1.

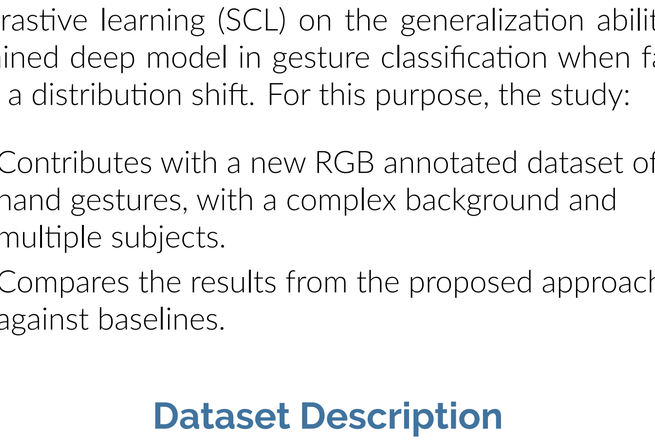

Figure 1. Examples of the training dataset of the four hand gesture classes in the unstructured environment and complex background.

The second dataset is used to test the model. This multi-user test dataset was recorded with three subjects who were not included in the training dataset. The dataset was recorded on a different day and at a different time of day, resulting in variation in luminosity. Figure 2 shows some samples that constitute the multi-user test dataset.

Figure 2. Examples of the test dataset with three different subjects and acquired in a different time of day in relation to the training dataset.

Table 1 shows the distribution of samples among all classes. Although the dataset is small when compared to the large-scale datasets, it has a distribution of samples per class similar to other static hand gesture datasets used in HGR. Additionally, we apply online data augmentation in the training phase that further help compensate for the reduced number of samples.

Dataset	A	F	L	Y	Total
Training dataset	6430	6148	5989	6044	**24611**
Multi-user test dataset	4183	4277	4276	4316	**17051**

Table 1. Number of images, per class, of the training dataset and the multi-user test dataset. The training dataset includes one subject and the test dataset includes three subjects.

Both of the datasets were recorded in the collaborative cell located in the Laboratory of Automation and Robotics at the University of Aveiro.

Methodology

This section describes the proposed contrastive domain adaptation technique for HGR. Our goal is to train a model on the source domain and, then, use it to make predictions on the target domain that has different characteristics, namely, different subjects and illumination conditions. We compare this approach with traditional transfer learning methods, that consists of the Inception-v3 pre-trained model that is repurposed for the specific hand gesture classification task involving four classes.

SCL has been shown to be effective for domain adaptation, because it can help the model to learn features that are invariant to domain shifts.

The idea behind the proposed contrastive domain adaptation technique is to use a network that branches twice after the encoder model (dual-branch head), allowing to train the representation model and the classification model simultaneously. Figure 3 shows the model architecture.

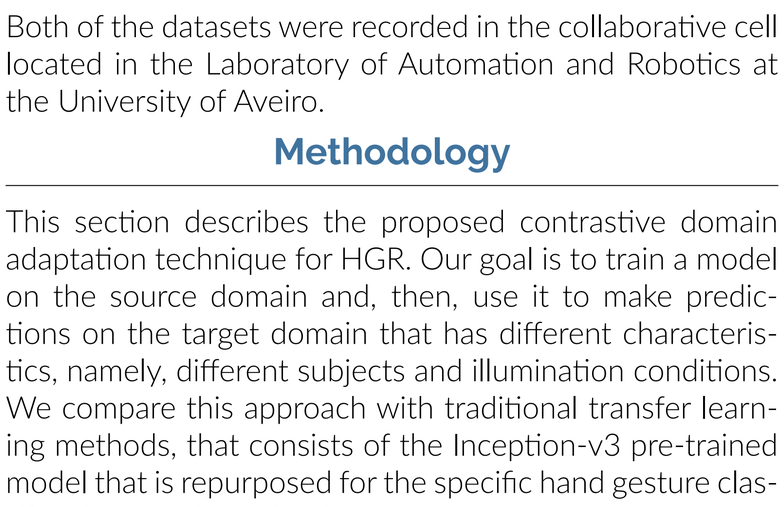

Figure 3. Model's architecture. The losses used are the contrastive loss and the Cross Entropy Loss.

Contrastive Learning

In Contrastive Learning we use a contrastive loss, defined by:

$$C_{loss} = \sum_{i \in B} \frac{-1}{p} \sum_{j \in P(i)} \log \frac{e^{(v_i \cdot v_j)/\tau}}{\sum_{a \in A(i)} e^{(v_i \cdot v_a)/\tau}}, \quad (1)$$

This loss function encourages the base model to produce similar feature vectors in the same class and dissimilar feature vectors in different classes.

Variables

Each batch B has v_i feature vectors. The set $P(i)$ represents the indexes of the positive samples j in relation to an anchor sample i, and it has a size of p. A sample is classified as positive when it belongs to the same class as the anchor. The set $A(i)$ includes all the indexes of B except i. The exponents exhibit the dot product between two feature vectors divided by a scalar temperature parameter τ.

Results

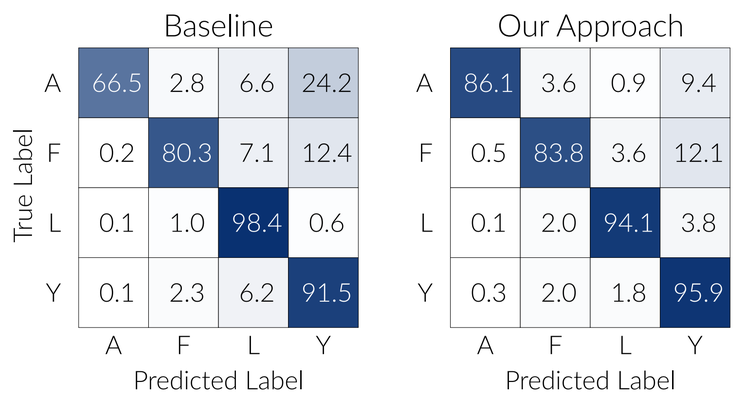

Figure 4. Examples of the test dataset with three different subjects and acquired in a different time of day in relation to the training dataset.

Model	Accuracy	Recall	Precision	F1
Baseline	84.23	84.15	86.82	84.07
Our approach	**90.03**	**89.97**	**90.96**	**90.12**

Table 2. Evaluation metrics for testing the both approaches in the multi-user test dataset. These values are the average of the metrics calculated for each class.

Conclusions

The focus of this study was the generalization capacity of the model, which was tested using the multi-user test dataset. In this testing phase, the results show that joining cross entropy loss and contrastive loss in a multi-loss training approach helps the model reach higher accuracy. In fact, this approach performed an increase of 6% in the accuracy of the model, compared to the traditional transfer learning method of training the model only with the CEL. This shows that contrastive learning is focused on learning task-specific and invariant features, being more effective to deal with the domain shift problem.

References

Rato, D., Oliveira, M., Santos, V., Gomes, M. and Sappa, A. (2022) 'A sensor-to-pattern calibration framework for multi-modal industrial collaborative cells', Journal of Manufacturing Systems, 64, pp. 497-507. doi: https://doi.org/10.1016/j.jmsy.2022.07.006.

Castro, A., Silva, F. and Santos, V. (2021) 'Trends of Human-Robot Collaboration in Industry Contexts: Handover, Learning, and Metrics', Sensors, 21(12), 4113. doi: 10.3390/s21124113.

Castro, A., Baptista, J., Silva, F. and Santos, V. (2023) 'Classification of handover interaction primitives in a COBOT–human context with a deep neural network', Journal of Manufacturing Systems, 68, pp. 289-302. doi: https://doi.org/10.1016/j.jmsy.2023.03.010.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C. and Krishnan, D. (2020) 'Supervised Contrastive Learning', CoRR, 2004.11362. Available at: https://arxiv.org/abs/2004.11362 (Accessed: 19 June 2024).

https://doi.org/10.3390/s23063332

joelbaptista@ua.pt