YDSP Young Defence Scientists Programme

DSTA Defence Science & Technology Agency

DSO National Laboratories

SIGN LANGUAGE TRANSCRIPTION WITH MACHINE LEARNING

Member:
Andrea Tan Kai Xuan
(Nanyang Girls' High School)

Mentors:
Woo Chin Jian, Ip Hei Man, Shen Bingquan
(DSO National Laboratories)

INTRODUCTION

This study investigates the feasibility of using classical Machine Learning (ML) methods to classify high-dimensional sign language videos into their corresponding gloss words, without the need for deep learning methods. This emphasises the potential of ML in Sign Language Transcription, fostering inclusivity for the Deaf and Hard-of-Hearing community.

Glossing

Glossing translates each morpheme in sign language into words, which are then translated into full sentences. However, annotating gloss words is labour-intensive, leading to gloss-free methods that bypass this step.
While effective, gloss-free approaches struggle with accurate video-to-gloss transcription due to the Representation Density Problem.

Representation Density Problem

Sign gestures with similar visual appearances but different meanings are closely located in the feature space could lead to MISINTERPRETATION especially in gloss-free methods, where models face difficulty in learning semantic boundaries in continuous sign videos, contributing to translation ambiguity.

METHODOLOGY

1 **T-distributed Stochastic Neighbour Embedding (t-SNE)** was used to visualise the Representation Density Problem by converting high-dimensional video embeddings of the Two-Stream Inflated 3D ConvNet (I3D) model (gloss-free) to a 2D space.

2 The **Sign Density Ratio (SDR)** was used to measure the severity of the Representation Density Problem. It computes the ratio of the Intra-Gloss Distance (distance within a single gloss) to the average Inter-Gloss Distance (distance of a gloss to all other glosses).

$$SDR(G_i) = \frac{D^{intra}_{G_i}}{avg.D^{inter}_{G_i}} = \frac{D(G_i)}{Mean_{j\neq i}(D(G_i, G_j))} \tag{1}$$

$$\mathrm{D}(G_i,\ G_j) = \frac{1}{|G_i|}\frac{1}{|G_j|}\sum_{x\in G_i,\, y\in G_j} d(x, y); \tag{2}$$

$$\mathrm{D}(G_i) = \frac{1}{|G_i|}\frac{1}{(|G_i|-1)}\sum_{x,y\in G_i,\, x\neq y} d(x, y); \tag{3}$$

3 Classical Machine Learning methods like the **Support Vector Machine (SVM)** and **Random Forest (RF)** classifiers were used to classify the video embeddings into their corresponding gloss words.

EXPERIMENTAL RESULTS

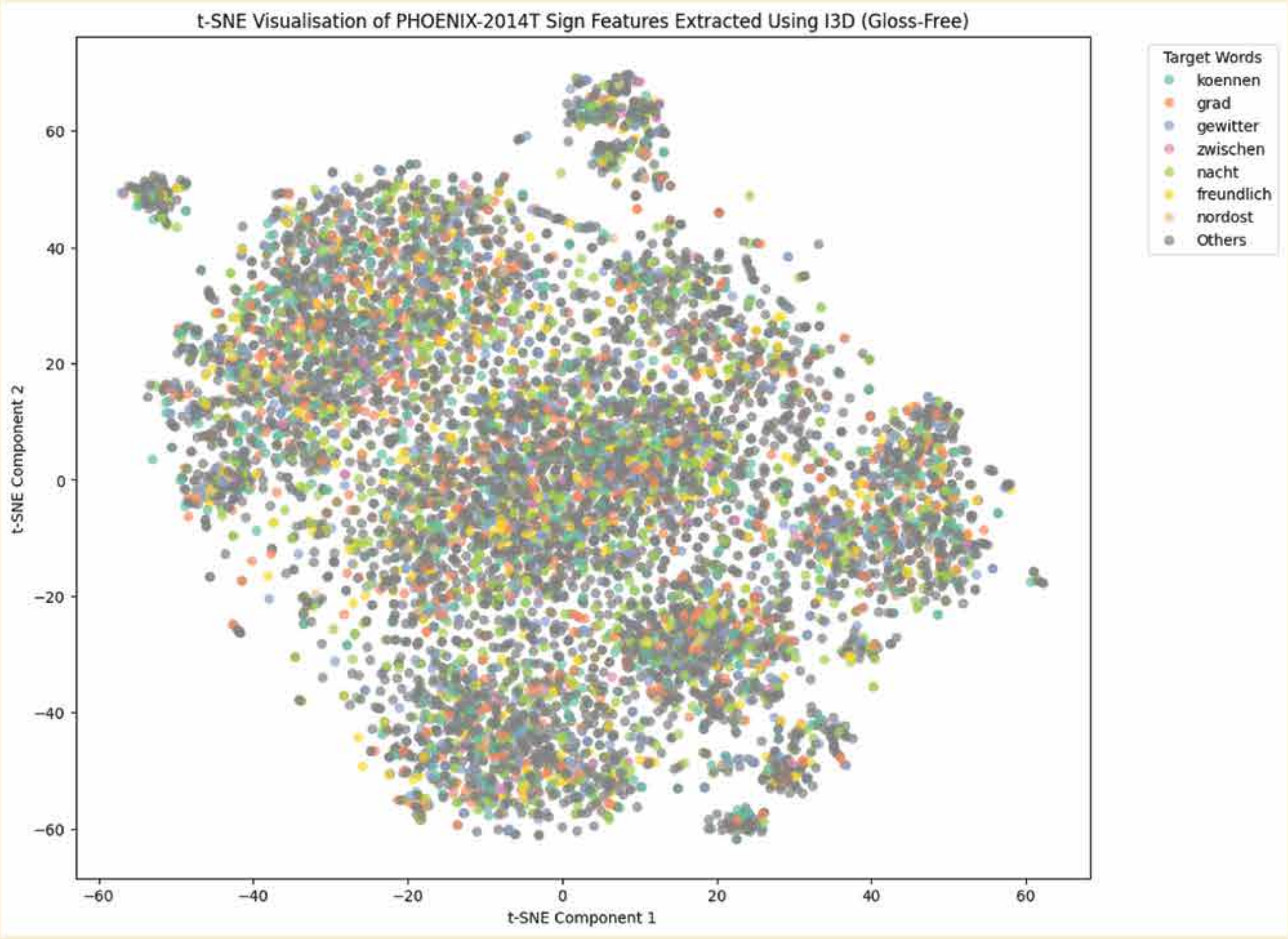

Figure 1: t-SNE Visualisation of PHOENIX-2014T sign features extracted using I3D

Overall SDR Value: 0.91
(i.e. Mean of SDR value of each target word)

Table 1: Overall evaluation of ML models for classifying sign videos into gloss

Model	Hyperparameter	Precision	Recall	F1-score	Accuracy
SVM	Linear	0.8529	0.8646	0.8567	0.97
	RBF	0.9524	0.9534	0.9524	0.97
RF	100 trees	0.9748	0.9749	0.9749	0.97

CONCLUSION

The t-SNE plot showed a **sparse distribution** of the video embeddings and the SDR value calculated was **high** compared to results obtained by Ye et al. (2024), revealing that target words were not distinctly clustered together in the gloss-free I3D model.

Despite this, classical Machine Learning methods yielded exceptional results, with the Random Forest classifier achieving the best **F1-score of 97.49%**. This proves that classical ML methods are **sufficient** in carrying out video-to-gloss sign language transcription.