MSc in Artificial Intelligence

“Multimodal machine learning techniques for sign language recognition”

Zidianakis Matthaios - MTN2008

Papadopoulos Georgios - MTN2025

Overview

- Sign language recognition refers to:
  - Classification problem of a sequence of video frames
  - Subcategory of gesture recognition in motion

- Dataset: *GSL* (Greek Sign Language)
  - Non multilabel (Isolated words)
  - 7 signers
  - 5 repetitions of each signers
  - 5 scenarios
  - 310 individuals words
  - 30 FPS - RGB (+Depth) - 848×480

Not used here

Tools

- Keypoints extraction MediaPipe
- Deep learning methods PyTorch
- Classical Machine learning (SVM) scikit learn

Dataset distribution

- Predefined *train*, *validation*, *test* sets → **not mixed** signers
- 4 flavours of the same dataset:
  1. Keeping the first **10** unique words.
  2. Keeping the first **12** unique words which have more than 500 samples (undersampled **exactly to 500**).
  3. Keeping the first **10** unique words containing **more than 500** samples for each words.
  4. Keeping all the glosses (**310**).

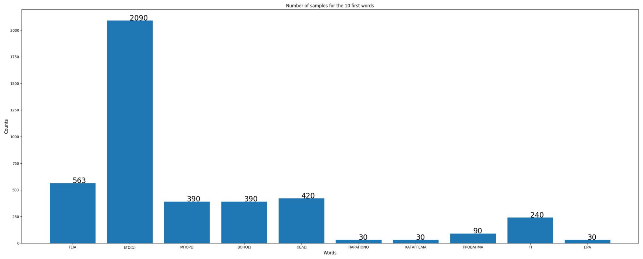


Figure 1: Word distribution plot for dataset 1 of train set.

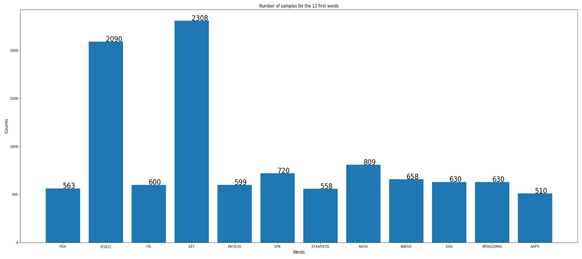
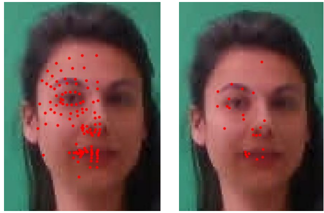


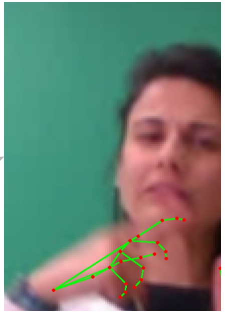
Figure 2: Word distribution plot for dataset 2 of train set before undersampling.

Data preprocessing



Sampling every 5 face points

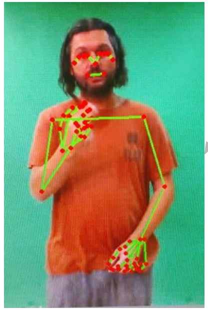
Blurring, 400% rescale and interpolation



Magnitude: $M_t^N = \sqrt{(D_{N_{x(t+1)}} - D_{N_{x(t)}})^2 + (D_{N_{y(t+1)}} - D_{N_{y(t)}})^2}$

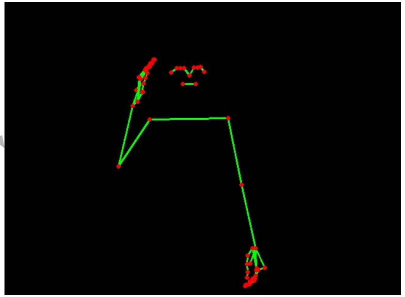
Preprocessing involves points selection:

Angles: $A_t^N = tanh^{-1}\left(\frac{D_{N_{y(t+1)}} - D_{N_{y(t)}}}{D_{N_{x(t+1)}} - D_{N_{x(t)}}}\right)$



Upper body Pose points

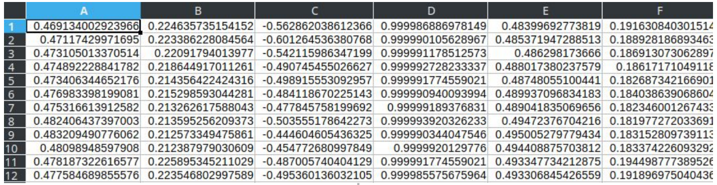
Pose skeleton in blank image



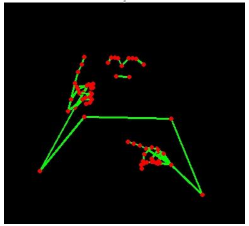
Methods: Deep learning models

L

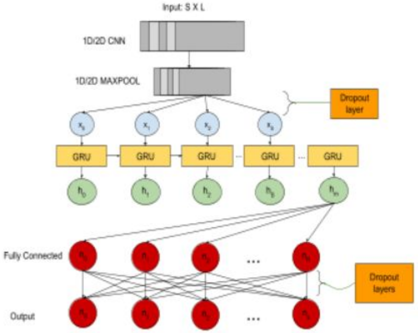
S



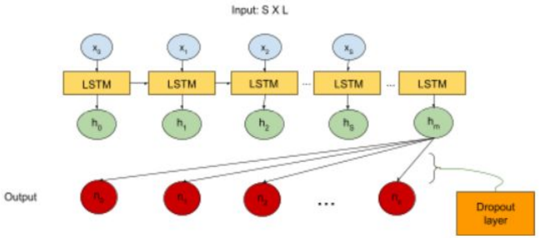
	A	B	C	D	E	F
1	0.469134002923966	0.224635735154152	-0.562862038612366	0.999986886978149	0.48399692773819	0.191630840301514
2	0.47117429971695	0.223386228084564	-0.601264536380768	0.999990105628967	0.485371947288513	0.188928186893463
3	0.473105013370514	0.22091794013977	-0.542115986347199	0.999991178512573	0.486298173666	0.186913073062897
4	0.474892228841782	0.218644917011261	-0.490745455026627	0.999992728233337	0.488017380237579	0.186171710491118
5	0.473406344652176	0.214356422424316	-0.498915553092957	0.999991774559021	0.48748055100441	0.182687342166901
6	0.476983398199081	0.215298593044281	-0.484118670225143	0.999990940093994	0.489937096834183	0.184038639068604
7	0.475316613912582	0.213262617588043	-0.477845758199692	0.99999189376831	0.489041835069656	0.182346001267433
8	0.482406437397003	0.213595256209373	-0.503555178642273	0.999993920326233	0.49472376704216	0.181977272033691
9	0.483209490776062	0.212573349475861	-0.444604605436325	0.999990344047546	0.495005279779434	0.183152809739113
10	0.480989485597908	0.212387979030609	-0.454772680997849	0.9999920129776	0.494408875703812	0.183374226093292
11	0.478187322616577	0.225895345211029	-0.487005740404129	0.999991774559021	0.493347734212875	0.194498777389526
12	0.477584689855576	0.223546802997589	-0.495360136032105	0.999985575675964	0.493306845426559	0.191896975040436



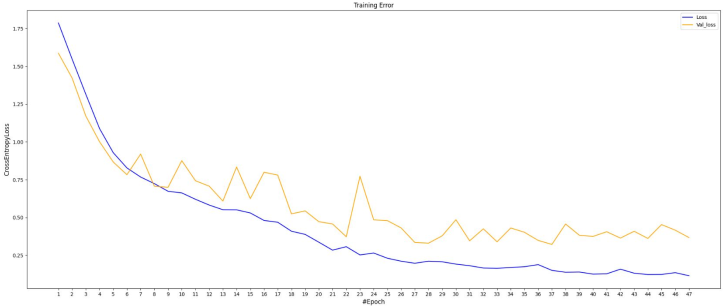
2D/1D CNN-GRU



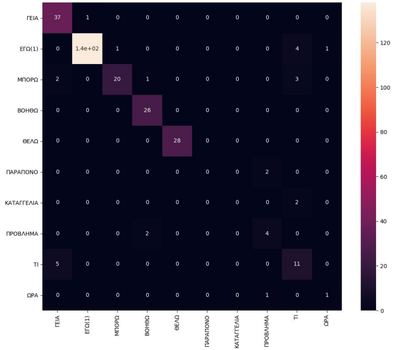
LSTM



Results - 1D CNN-GRU *dataset 1* (sampling face points)



F1-macro score		
	Validation set	Test set
With class weights	53%	59%
Without class weights	61%	67%

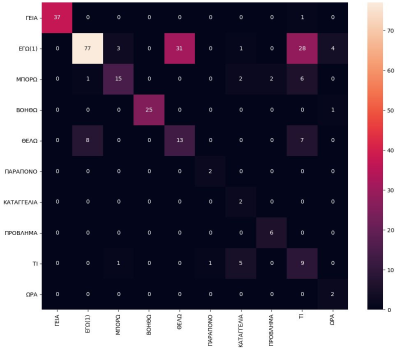


	ΓΕΙΑ	ΕΓΩ(1)	ΜΠΟΡΩ	ΒΟΗΘΩ	ΘΕΛΩ	ΠΑΡΑΠΟΝΟ	ΚΑΤΑΓΓΕΛΙΑ	ΠΡΟΒΛΗΜΑ	ΤΙ	ΩΡΑ
ΓΕΙΑ	37	1	0	0	0	0	0	0	0	0
ΕΓΩ(1)	0	1.4e+02	1	0	0	0	0	0	4	1
ΜΠΟΡΩ	2	0	20	1	0	0	0	0	3	0
ΒΟΗΘΩ	0	0	0	26	0	0	0	0	0	0
ΘΕΛΩ	0	0	0	0	28	0	0	0	0	0
ΠΑΡΑΠΟΝΟ	0	0	0	0	0	0	0	2	0	0
ΚΑΤΑΓΓΕΛΙΑ	0	0	0	0	0	0	0	0	2	0
ΠΡΟΒΛΗΜΑ	0	0	0	2	0	0	0	4	0	0
ΤΙ	5	0	0	0	0	0	0	0	11	0
ΩΡΑ	0	0	0	0	0	0	0	1	0	1

Totally lose 2 classes!

$$weight_i = 1/\left(|samples_i| / \sum_{i=1}^{n} |samples_i| \right)$$

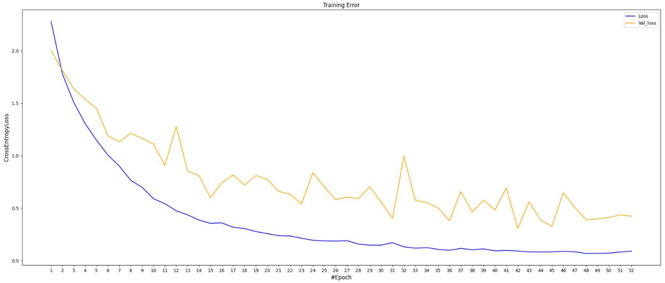
$$normalized\ weight_i = weight_i / \sum_{i=1}^{n} weight_i$$

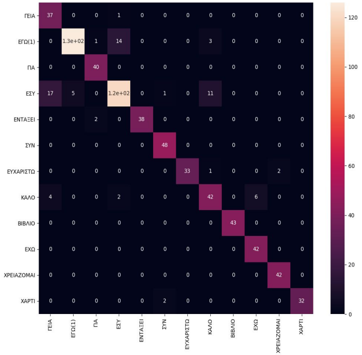


	ΓΕΙΑ	ΕΓΩ(1)	ΜΠΟΡΩ	ΒΟΗΘΩ	ΘΕΛΩ	ΠΑΡΑΠΟΝΟ	ΚΑΤΑΓΓΕΛΙΑ	ΠΡΟΒΛΗΜΑ	ΤΙ	ΩΡΑ
ΓΕΙΑ	37	0	0	0	0	0	0	0	1	0
ΕΓΩ(1)	0	77	3	0	31	0	1	0	28	4
ΜΠΟΡΩ	0	1	15	0	0	0	2	2	6	0
ΒΟΗΘΩ	0	0	0	25	0	0	0	0	0	1
ΘΕΛΩ	0	8	0	0	13	0	0	0	7	0
ΠΑΡΑΠΟΝΟ	0	0	0	0	0	2	0	0	0	0
ΚΑΤΑΓΓΕΛΙΑ	0	0	0	0	0	0	2	0	0	0
ΠΡΟΒΛΗΜΑ	0	0	0	0	0	0	0	6	0	0
ΤΙ	0	0	1	0	0	1	5	0	9	0
ΩΡΑ	0	0	0	0	0	0	0	0	0	2

Results - 2D/1D CNN-GRU (sampling face points)

1D CNN-GRU *dataset 2*





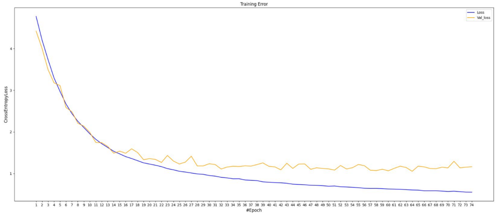
	ΓΕΙΑ	ΕΓΩ(1)	ΠΑ	ΕΣΥ	ΕΝΤΑΞΕΙ	ΣΥΝ	ΕΥΧΑΡΙΣΤΩ	ΚΑΛΟ	ΒΙΒΛΙΟ	ΕΧΩ	ΧΡΕΙΑΖΟΜΑΙ	ΧΑΡΤΙ
ΓΕΙΑ	37	0	0	1	0	0	0	0	0	0	0	0
ΕΓΩ(1)	0	1.3e+02	1	14	0	0	0	3	0	0	0	0
ΠΑ	0	0	40	0	0	0	0	0	0	0	0	0
ΕΣΥ	17	5	0	1.2e+02	0	1	0	11	0	0	0	0
ΕΝΤΑΞΕΙ	0	0	2	0	38	0	0	0	0	0	0	0
ΣΥΝ	0	0	0	0	0	48	0	0	0	0	0	0
ΕΥΧΑΡΙΣΤΩ	0	0	0	0	0	0	33	1	0	0	2	0
ΚΑΛΟ	4	0	0	2	0	0	0	42	0	6	0	0
ΒΙΒΛΙΟ	0	0	0	0	0	0	0	0	43	0	0	0
ΕΧΩ	0	0	0	0	0	0	0	0	0	42	0	0
ΧΡΕΙΑΖΟΜΑΙ	0	0	0	0	0	0	0	0	0	0	42	0
ΧΑΡΤΙ	0	0	0	0	0	2	0	0	0	0	0	32

F1-macro score	
Validation set	Test set
91.77%	93.89%

2D CNN-GRU *dataset 2 (skeletal)*

F1-macro score	
Validation set	Test set
42.1%	43.1%

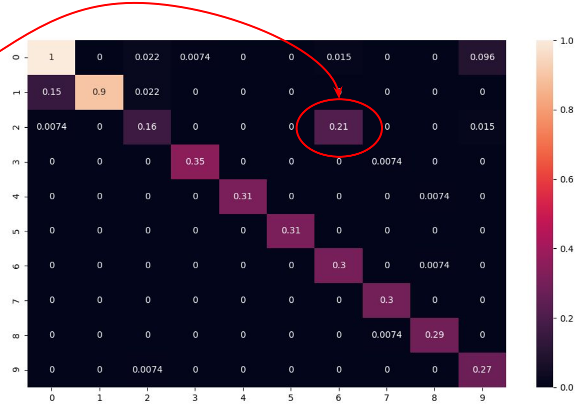
1D CNN-GRU *dataset 4 (all words)*



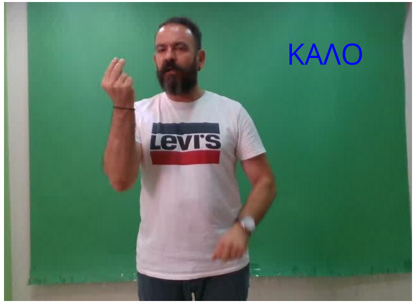
F1-macro score	
Validation set	Test set
56.81%	60.5%

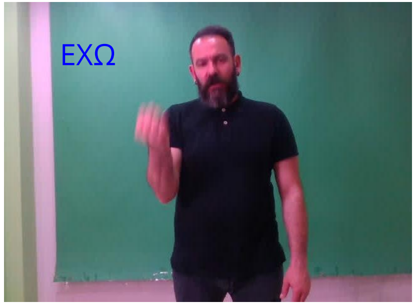
Results - Early experiments

- *LSTM* architecture
  - full head keypoints → noisy features
  - misclassified two similar words



	0	1	2	3	4	5	6	7	8	9
0	1	0	0.022	0.0074	0	0	0.015	0	0	0.096
1	0.15	0.9	0.022	0	0	0	0	0	0	0
2	0.0074	0	0.16	0	0	0	0.21	0	0	0.015
3	0	0	0	0.35	0	0	0	0.0074	0	0
4	0	0	0	0	0.31	0	0	0	0.0074	0
5	0	0	0	0	0	0.31	0	0	0	0
6	0	0	0	0	0	0	0.3	0	0.0074	0
7	0	0	0	0	0	0	0	0.3	0	0
8	0	0	0	0	0	0	0	0.0074	0.29	0
9	0	0	0.0074	0	0	0	0	0	0	0.27

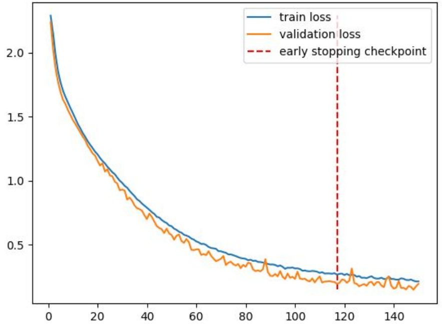


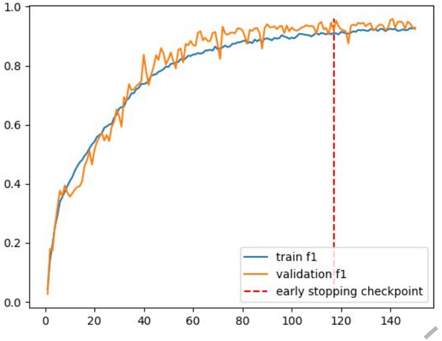


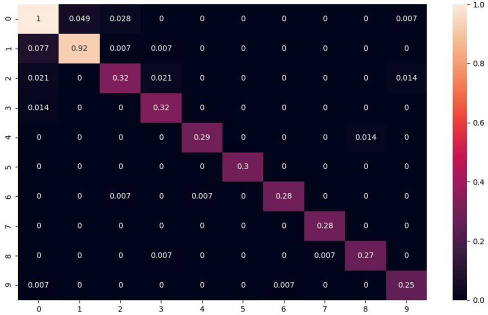
Results - LSTM

Training/ Validation

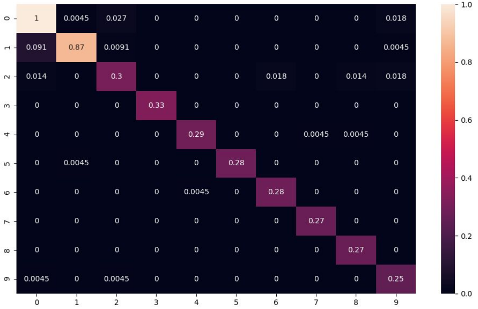
94%







- *F1 macro score*
- *L: upper body points*



0	1	2	3	4	5	6	7	8	9
ΕΣΥ	*ΕΓΩ (1)*	*ΚΑΛΟ*	*ΣΥΝ*	*ΒΙΒΛΙΟ*	*ΧΡΕΙΑΖΟΜΑΙ*	*ΕΧΩ*	*ΓΙΑ*	*ΕΝΤΑΞΕΙ*	*ΓΕΙΑ*

Test

95%

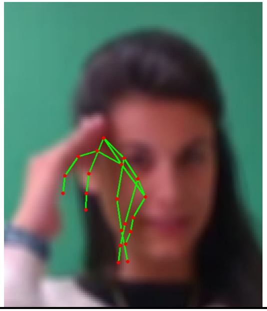
Results - Miscellaneous

SVM F1-macro scores (sampling face points)		
	Validation set	Test set
Dataset 1	78%	88%
Dataset 2	81%	82%
Dataset 4	39%	42%

Angles - Magnitude F1-macro score (*dataset 2*)		
	Validation set	Test set
1D CNN-GRU	70%	65%
SVM	45%	49%

Filters applied - F1-macro score (*dataset 2*)		
	Validation set	Test set
1D CNN-GRU	88.25%	91.86

False points recognition!



Conclusions

- *Mediapipe* → suitable for keypoints feature extraction
    - Especially for different deep learning methods
    - Adequate classical ML (SVM) performance
- Improve with **only** the representative points
- Applied filters were weaker
- Preprocessing proved to be significant!
- Future work:
    - Keypoints position relative to image → independent to captured image angle
        - Division by shoulder points coordinates
        - Video frames data augmentation
    - Examine *GANs* architecture → denoise video frames from motion blur

THANK YOU