12TH WORKSHOP ON THE REPRESENTATION AND PROCESSING OF SIGN LANGUAGES: LANGUAGE IN MOTION · LREC 2026

A Comparative Analysis of Traditional and Contemporary Visual Features for Computational Annotation of Irish Sign Language

Sarmad Khan · Simon D. McLoughlin · Irene Murtagh

ADAPT Centre, AI-Driven Digital Content Technology · Technological University Dublin

sarmad.khan@adaptcentre.ie · Simon.D.McLoughlin@tudublin.ie · Irene.Murtagh@adaptcentre.ie

1. Introduction: the annotation bottleneck

Detailed gloss-level annotation of sign language video typically takes **several hours of expert work per minute of recording** .

Throughput. Annotation is highly time-consuming and requires expert linguistic and Deaf-community knowledge; trained annotators are scarce.

Data scarcity. Current annotation workflows limit the scalable production of annotated datasets required for sign-language processing and NLP research.

Scalable sign-language processing requires substantially larger annotated datasets than current annotation workflows can efficiently produce.

2. Background: Why Irish Sign Language

- Recognised under the **Irish Sign Language Act 2017**.
- Approximately **5,000 Deaf primary users** ; 40,000 broader users in Ireland.
- Distinct grammar, gender-based variation, visual gestural articulation.
- Substantially fewer annotated resources than ASL or BSL.

Key figures

Recognised	**2017 (ISL Act, Republic of Ireland)**
Primary users	5,000 Deaf first-language signers
Wider community	40,000 ISL users
Annotated corpora	Few, vs. ASL / BSL benchmarks

3. Dataset: Signs of Ireland Corpus

- **10 hours** of naturalistic narrative ISL video.
- **40 Deaf signers** (24 female, 16 male).
- **4 age groups**, **5 regions**: Dublin, Galway, Wexford, Waterford, and Cork.
- Time-aligned **ELAN** XML gloss annotations.
- Collected at Trinity College Dublin (Leeson, 2004).

ELAN annotation

$A_i = \langle ID_i, t_{start}, t_{end}, g_i \rangle$

Each gloss has precise temporal boundaries suitable for frame-level segmentation.

10	40	5	4
HOURS	SIGNERS	REGIONS	AGE GROUPS

4. Dataset: Data curation

From **3,865** raw gloss instances to **55** curated lexical glosses (1,950 instances).

Frequency

Keep glosses with ≥ 20 occurrences for stable learning.

Normalise

UPPERCASE labels , remove non-linguistic characters, unify lexical variants.

Exclude

Remove productive lexicon items, indices, name signs, emblems, and gestures.

Review

Manual validation; prioritizing established lexical forms (Sutton-Spence, 2007).

Result

55 ISL glosses · 1,950 instances · controlled benchmark.

Stage	Instances	Vocabulary
Raw	3,865	—
After curation	1,950	55

1,950 instances × 20 frames, 39,000 sampled frames · 448×448 · 80/20 stratified split.

5. Methodology: Two feature paradigms

Traditional pose-based

Track hands, body and face explicitly; model articulator trajectories as geometry.

- Interpretable, lightweight
- Fragile to occlusion and viewpoint variation
- Cascaded tracker errors propagate

e.g., MediaPipe Holistic

Contemporary self-supervised vision

Learn rich visual representations from raw video without explicit key point design.

- Captures motion, context, appearance
- More robust across signers and settings
- Less interpretable than pose-based approaches

e.g., DINOv2 (ViT-S/14)

Which paradigm more effectively supports the scalable expansion of annotation tasks for ISL?

6. Methodology: Feature extraction

Visual: DINOv2

768-D

Backbone	ViT-S/14, self-supervised
CLS token	384-D global descriptor
Patch tokens	384-D, PCA-foreground mean-pool
Input	448 × 448, normalised [-1, 1]

Pose: Media Pipe Holistic

258-D

Body	33 landmarks · (x, y, z, v) → 132
Hands	2 × 21 landmarks · (x, y, z) → 126
Encodes	articulator geometry, inter-limb relations
Trade-off	interpretable but cascaded errors

Multi-modal fusion

1,026-D

Early fusion: concatenate $[\, f_{visual} \,;\, f_{pose} \,]$

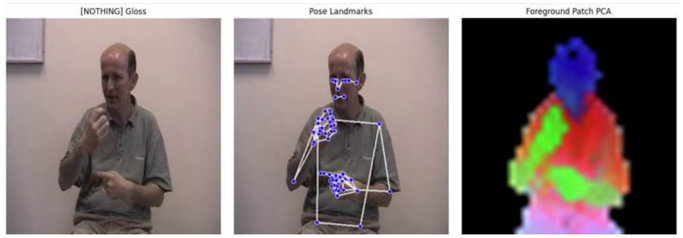
Tests whether explicit kinematics complement self-supervised semantics.

7. Methodology: Feature Visualisation

Gloss

NOTHING

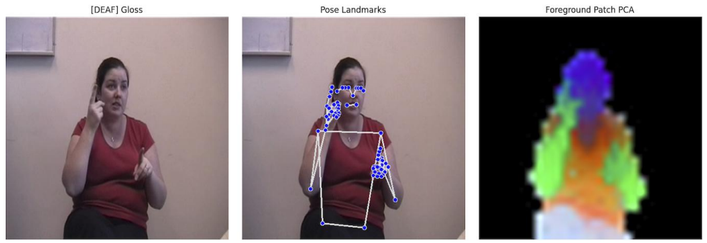
male signer



Gloss

DEAF

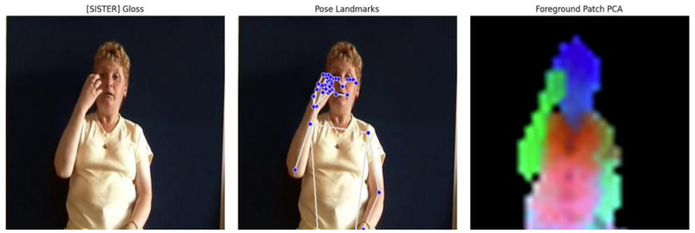
female signer



Gloss

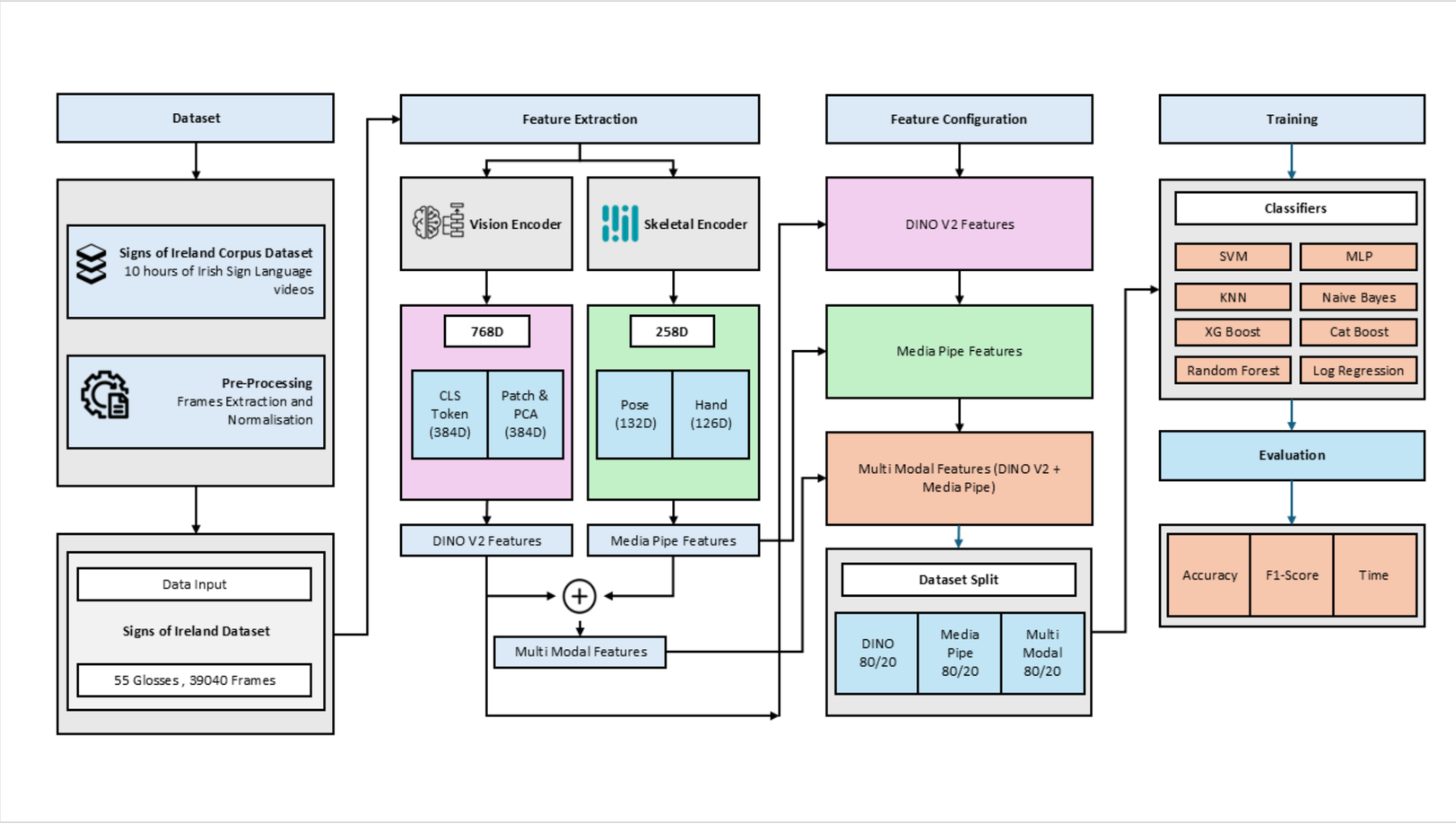
SISTER

female signer



Each strip: **Original frame** · **MediaPipe pose** · **DINOv2 patch PCA**. The DINOv2 PCA projection groups patches with similar visual semantics, separating hands, head, torso and signing space without supervision.

8. Methodology: Annotation framework



Vision encoder (DINOv2 ViT-S/14) and skeletal encoder (MediaPipe Holistic) feed three feature configurations into eight classifiers, each evaluated on accuracy, F1, and training time.

9. Results: Experimental design

3 feature configurations × 8 classifiers = 24 experiments

Classifiers

Spans ensemble, kernel, neural, linear, instance-based, and probabilistic families model-agnostic comparison.

XGBoost	**CatBoost**	**Random Forest**	**MLP**
SVM (RBF)	**Logistic Regression**	**KNN (k=7)**	**Naive Bayes**

ensemble · neural · kernel · linear · instance · probabilistic

SPLIT
80 / 20 stratified by gloss

SEED
random_state = 42

STOPPING
Early stopping on boosters

Evaluation metrics

Reported comparison focuses on accuracy, F1, and efficiency. Top-K is included as an annotation-oriented metric.

1	**Accuracy**	overall correctness
2	**Top-K**	gloss appears in top K predictions
3	**Macro / weighted F1**	class-balanced
4	**Per-class F1**	gloss-level diagnostics
5	**Log loss**	prediction confidence

Ground truth: expert ELAN annotations from Signs of Ireland.

10. Results: Headline Accuracy

DINOv2 achieves the best average accuracy and the strongest accuracy–efficiency trade-off.

11. Results: Cross-classifier consistency

Accuracy (%) by classifier × feature. Bold = best feature for that classifier; shaded row = peak overall.

Classifier	DINOv2	MediaPipe	Multi-modal
XGBoost peak	**95.91**	95.12	95.61
CatBoost	**95.84**	95.17	94.74
Random Forest	**95.43**	94.83	85.07
MLP	**95.26**	91.09	94.74
KNN (k = 7)	**95.35**	86.74	92.57
Logistic Regression	94.45	60.05	**94.56**
SVM (RBF)	**93.99**	76.96	91.61
Naive Bayes (independence assumption violated)	22.69	13.97	25.32

DINOv2 is best or tied in every viable classifier family. Pose-only collapses on linear / kernel classifiers (60% / 77%).

Peak: **XGBoost + DINOv2 = 95.91%** .

12. Results: Efficiency: accuracy vs. training time

Feature configuration	Avg accuracy	Avg train time	Note
DINOv2	**86.12%**	119 s	Best balance of accuracy & cost
MediaPipe	76.74%	61 s	Fastest, but accuracy ceiling is low
Multi-modal fusion	84.28%	**1,077 s**	9× DINOv2 training time, lower accuracy

Multi-modal fusion costs **9× more** classifier training time on average for **lower** average accuracy. This is classifier training time only (excluding feature extraction); even on this narrower axis, naive fusion is a poor trade-off.

Naive multi-modal fusion incurs substantially higher classifier training cost without corresponding accuracy gains.

13. Results: Per-class behavior

Per-class F1 distribution

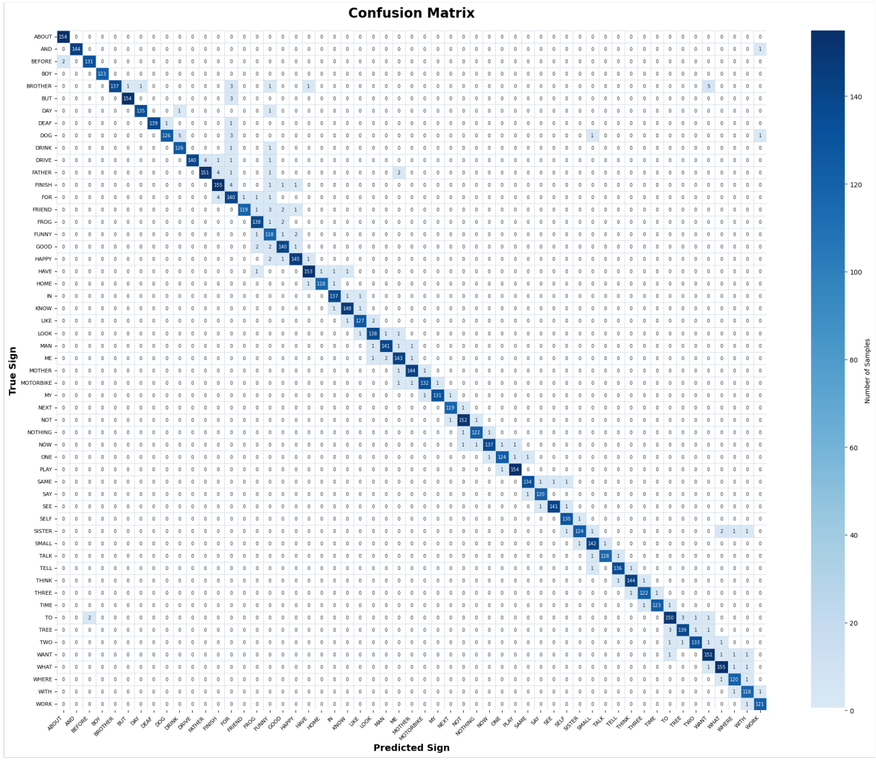
HIGHEST		LOWEST	
DRINK	0.99	SAME	0.90
SMALL	0.99	FINISH	0.91
MOTORBIKE	0.99	MY	0.93
PLAY	0.99	WHERE	0.93
DEAF	0.99	FATHER	0.93

58.2%
CLASSES ≥ 0.97 F1

10.9%
CLASSES < 0.95 F1

Even the worst gloss (SAME) lands at 0.90 performance is consistent across the long tail.

55 × 55 confusion



Strong diagonal concentration; off-diagonal mass is small and not systematic.

14. Conclusion

Summary:

- **Prioritise self-supervised visual features.** DINOv2 captures appearance, pose, spatial context and motion-relevant cues without explicit landmark extraction.
- **Treat pose estimation as optional.** Naive fusion adds cost and brittleness; use pose only when interpretability or kinematics are the goal.
- **Pair with efficient classifiers.** Gradient boosting on DINOv2 features reaches **95.91%** in minutes deployable inside annotator workflows.

NEXT STEPS

- Apply explainable AI methods to better understand model decisions.
- Scale beyond 55 high-frequency glosses into the long tail.
- Continuous signing & gloss-boundary detection.
- Integrate suggestions into ELAN annotator workflows.
- Evaluate signer-independent generalisation.

CONTACT

Sarmad Khan
sarmad.khan@adaptcentre.ie

Simon D. McLoughlin
simon.d.mcloughlin@tudublin.ie

Irene Murtagh
irene.murtagh@adaptcentre.ie

ADAPT Centre • TU Dublin

Engaging Content
Engaging People

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Thank you : Questions?