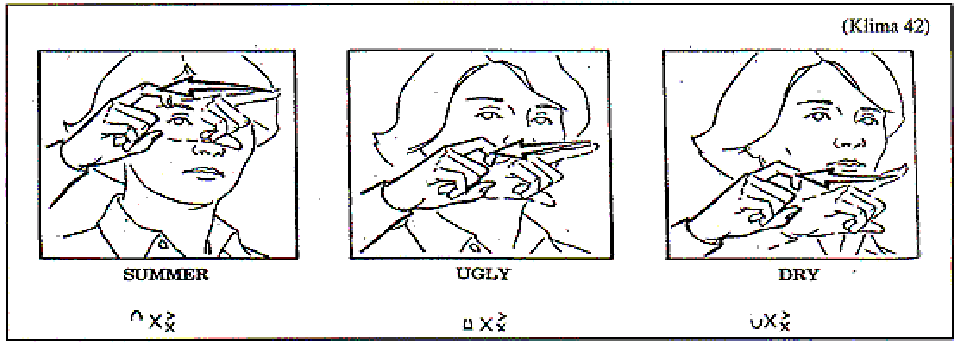
Towards a Transcription System of Sign Language Video Resources via Motion Trajectory Factorisation

Mark Borg
Kenneth P. Camilleri

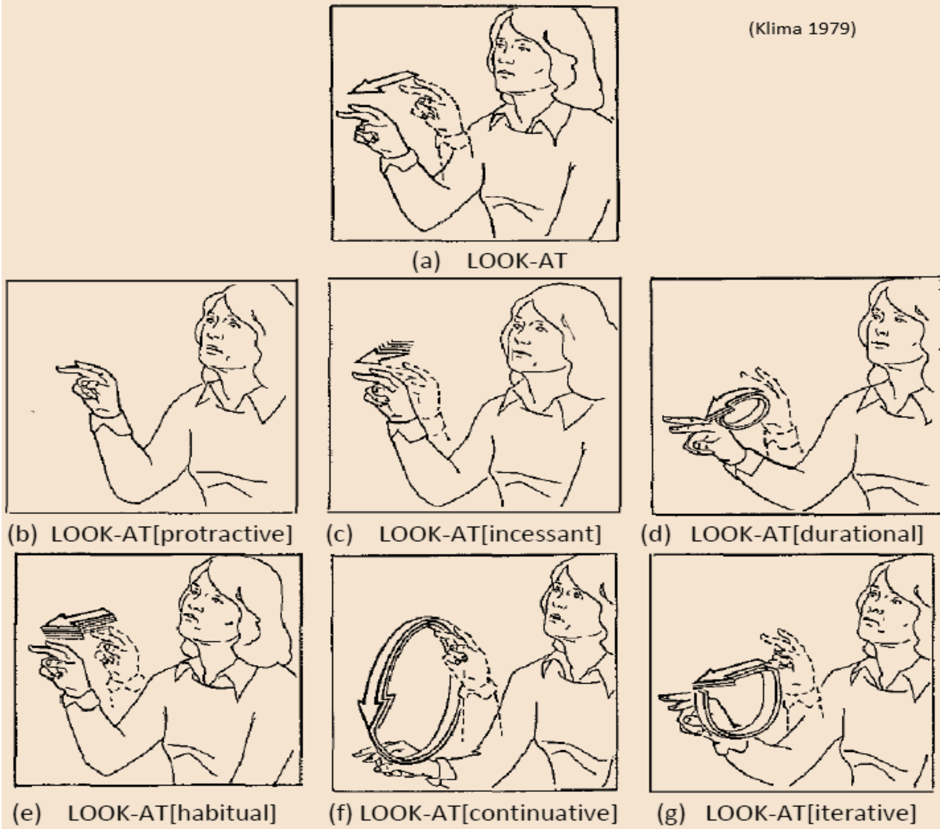
SCE: Systems and Control Engineering,
University of Malta

DocEng 2017, Valletta, Malta
4 – 7 September 2017

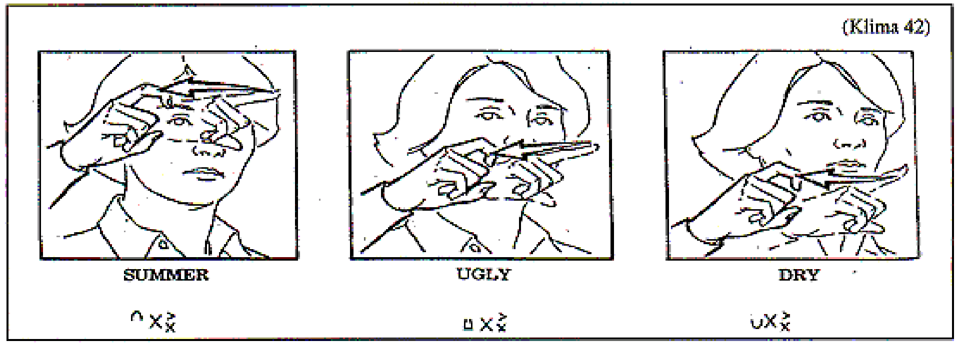
Sign Languages



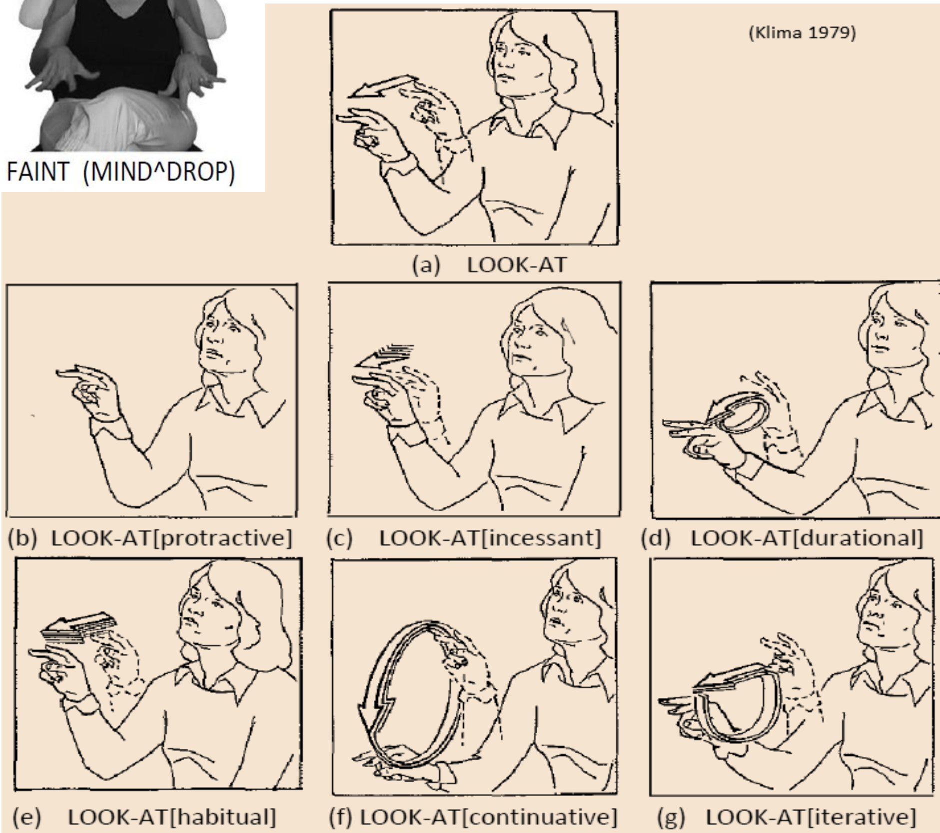
- Visual languages
- Articulators
  - Hand motion
  - Hand shapes
  - Place of articulation
  - Non-manual gestures:
    - Mouthings, facial expressions, body postures, …



Sign Languages



- Visual languages
- Articulators
  - Hand motion
  - Hand shapes
  - Place of articulation
  - Non-manual gestures:
    - Mouthings, facial expressions, body postures, …
- Sign Languages are complex
  - Non-Sequentiality
    - Parallel use of articulators, layering of meaning (sign inflection), composite signs, …
- Fully-fledged languages

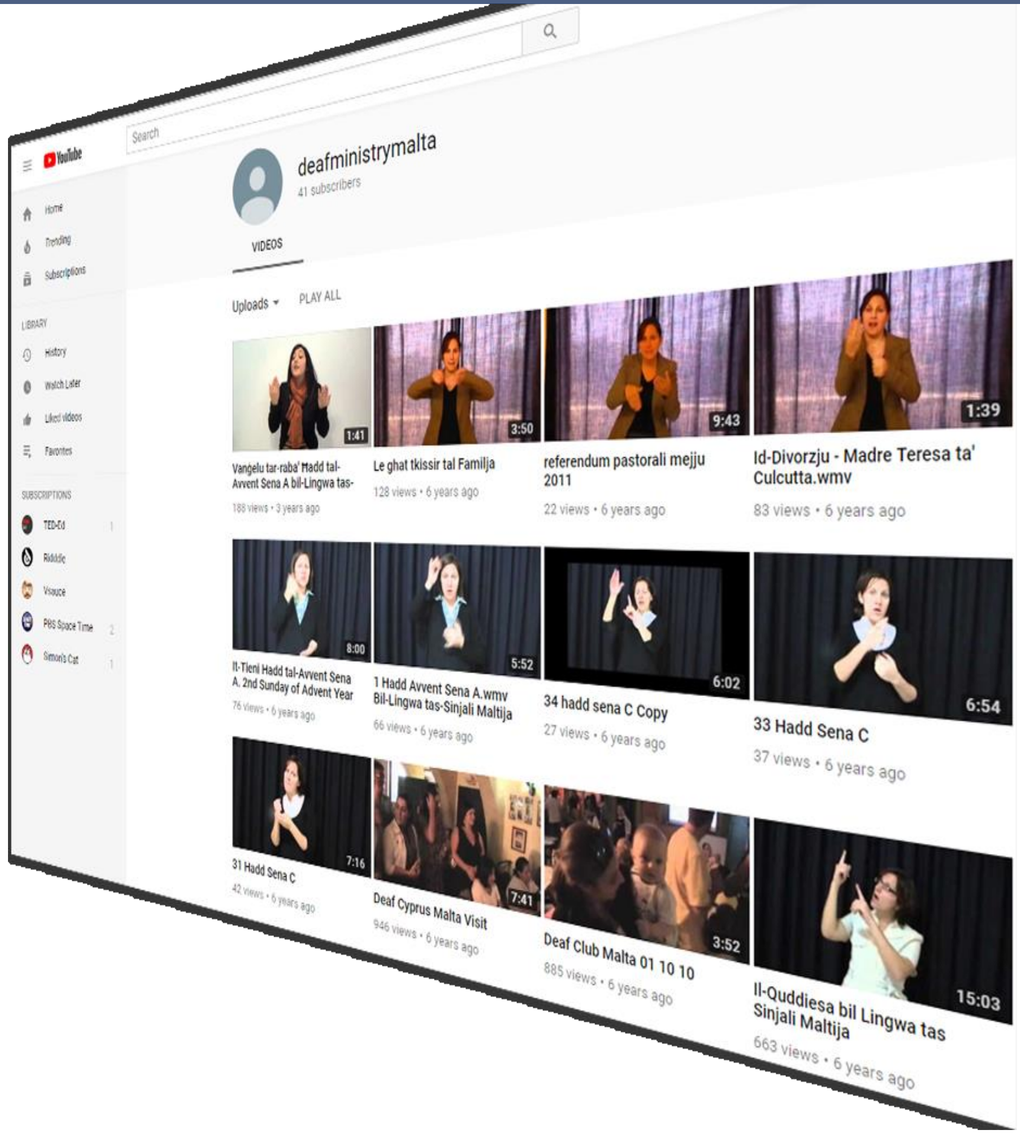


Sign Languages

- Communication barrier

Sign Languages

- Communication barrier
- Video blogs (*vlogs*) [1,2]



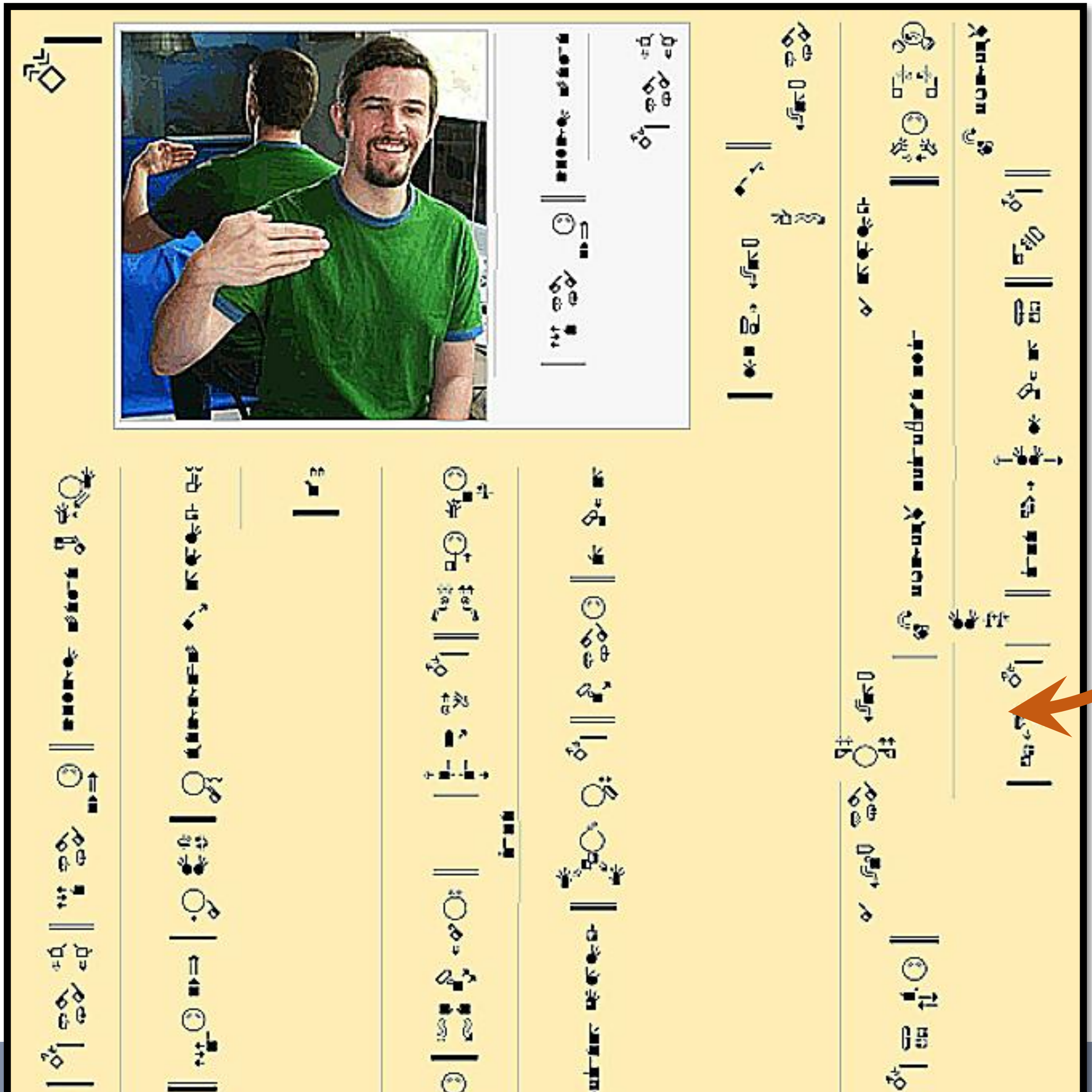
[1] www.deafVideo.tv
[2] www.deafread.com/vlogs/

Sign Languages

- Communication barrier
- Video blogs (*vlogs*) [1,2]
- Challenges:
  - How to extract content from sign language videos?
  - Documentation & representation?
  - Video-based search?
  - Preserving the cultural memory of the Deaf community

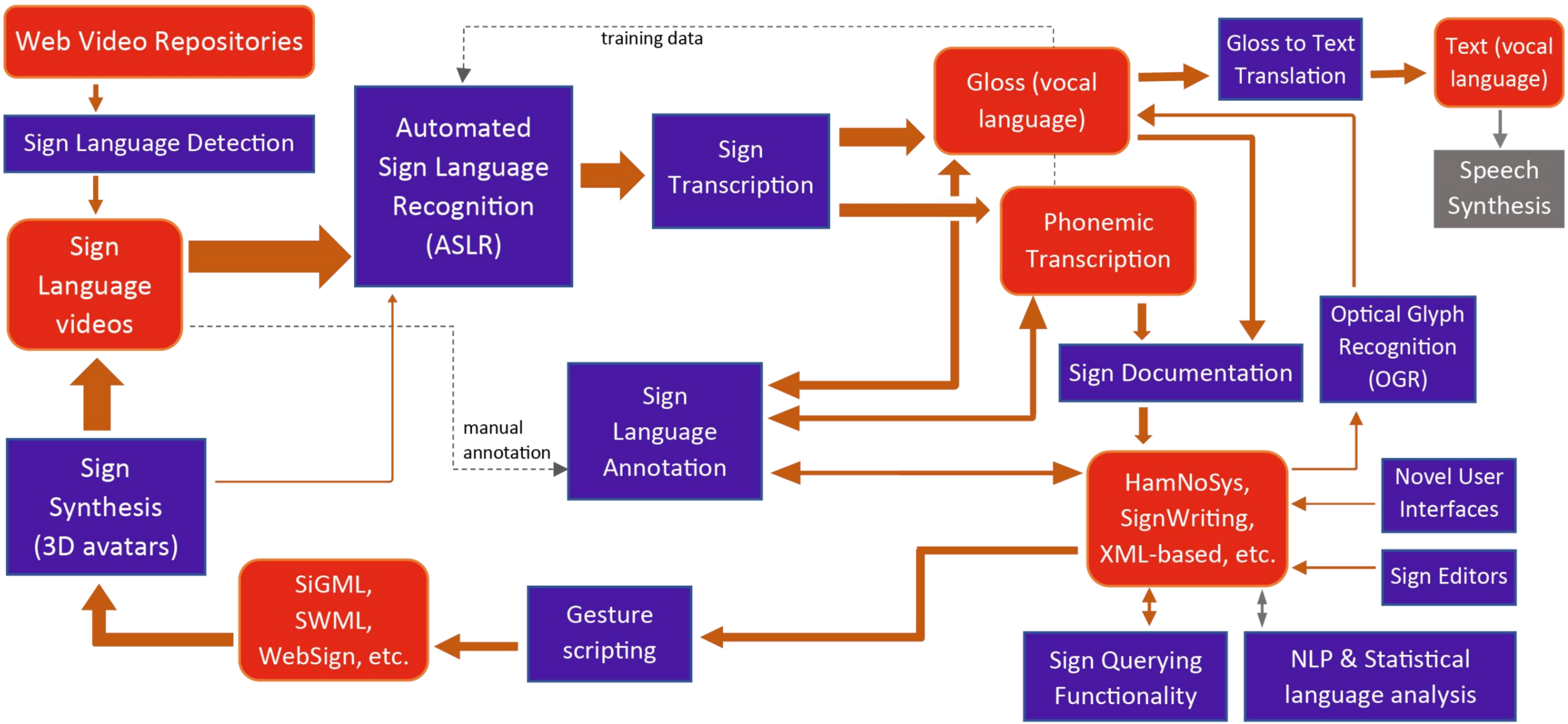
[1] www.deafVideo.tv
[2] www.deafread.com/vlogs/

Sign Languages

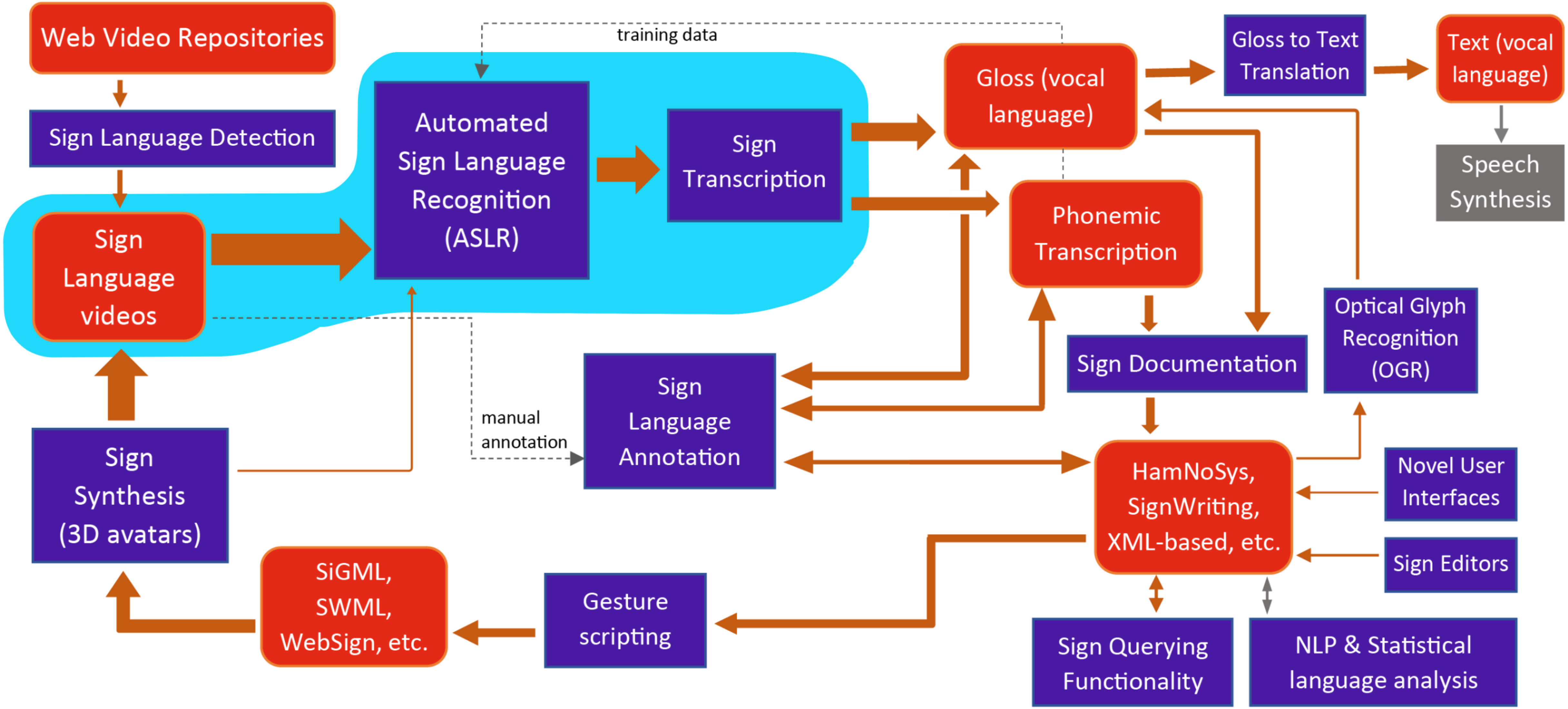


signwriting.org

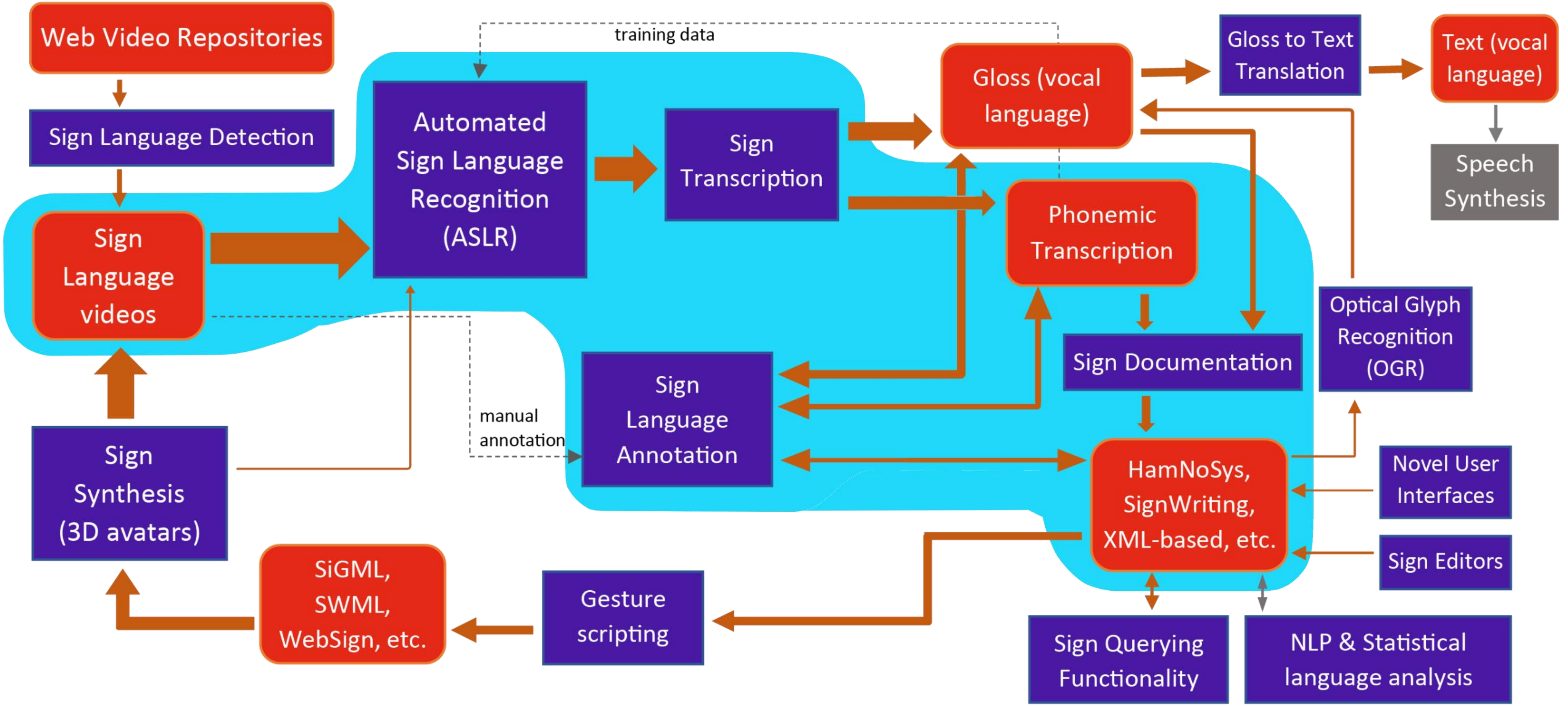
Sign Language technologies



Sign Language technologies

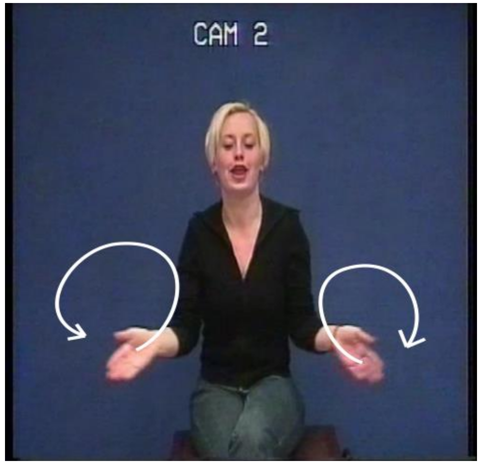


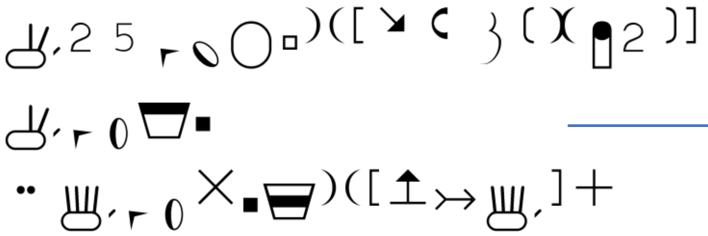
Sign Language technologies

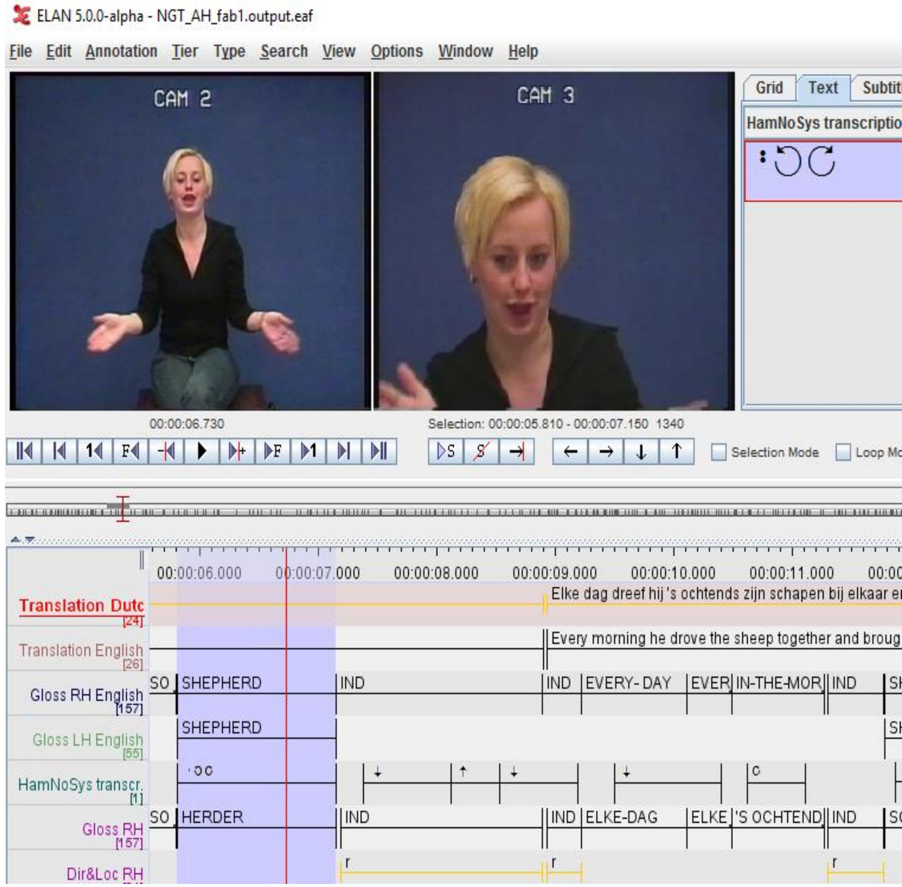


Outline

- Vision-based ASLR
- Hand motion classification
- HamNoSys transcription
- ELAN annotation tool
- Experiments and Results
- Conclusion
- Future work







<TIER DEFAULT_LOCALE="en" LINGUISTIC_TYPE_REF="Default" TIER_ID="HamNoSys transcr.">
<ANNOTATION>
<ALIGNABLE_ANNOTATION ANNOTATION_ID="a1009" TIME_SLOT_REF1="ts113" TIME_SLOT_REF2="ts124">
<ANNOTATION_VALUE>⊠⊠</ANNOTATION_VALUE>
</ALIGNABLE_ANNOTATION> <ALIGNABLE_ANNOTATION ANNOTATION_ID="a1010" TIME_SLOT_REF1="ts119" TI
<ANNOTATION_VALUE>⊠⊠</ANNOTATION_VALUE>
</ALIGNABLE_ANNOTATION>
</ANNOTATION>
</TIER>
<LINGUISTIC_TYPE GRAPHIC_REFERENCES="false" LINGUISTIC_TYPE_ID="Default" TIME_ALIGNABLE="true"/>

Computer Vision challenges

- Identical articulators (2 hands, fingertips)
- Fast motion (blurring)
- Non-rigid transformations
- Frequent and persistent occlusion, self-occlusion

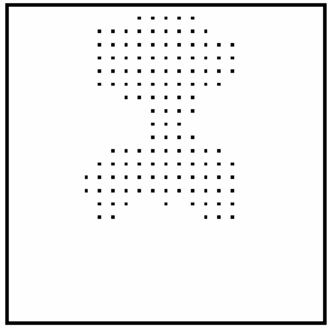
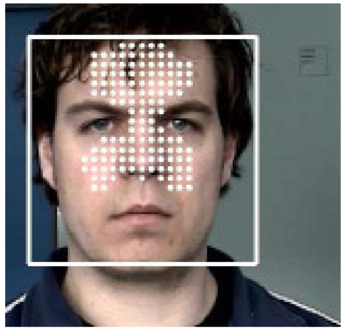
- Gesture recognition challenges
  - Continuous signing, sign spotting, coarticulation (blending), movement epenthesis, multi-modality, …

Datasets & Limitations

- We focus:
  - Gross motion of the hands (hand trajectories)

- Datasets:
  - BBC pose dataset (Oxford University)
  - ECHO NGT corpus (ECHO project, Radboud University)

Hand Tracking system

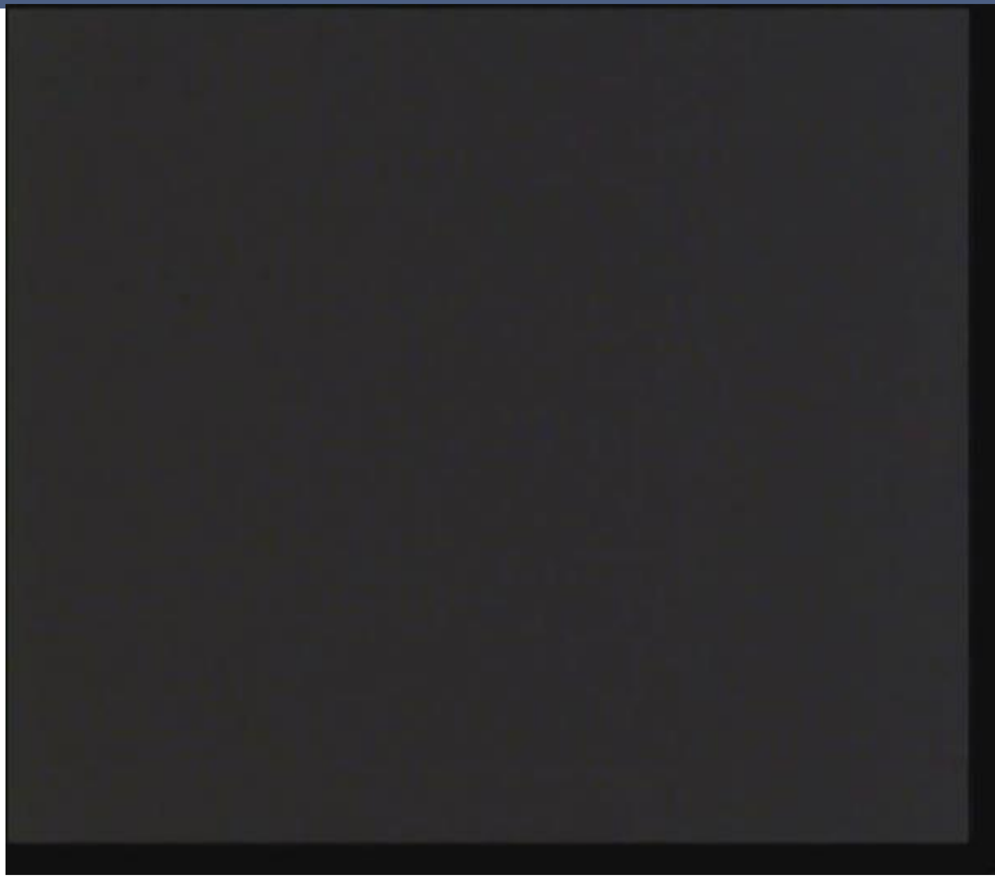
- Haar-based face detection [1]
- Adaptive skin colour model [2]
- KLT (Kanade-Lucas-Tomasi) features [3]
- Candidate hand regions
- MHT (Multiple Hypothesis Tracking) framework [4]

[1] Viola and Jones (2001) Robust Real-time Object Detection
[2] Wimmer and Radig (2005) Adaptive skin color classificator
[3] Shi and Tomasi (1994) Good Features to Track
[4] Antunes et al (2011) A Library for Implementing MHT

Borg and Camilleri (2015) Multiple Hypothesis Tracking with Sign Language Hand Motion Constraints

The Factorisation Method

- **SfM** – Structure from Motion technique
- The Factorisation Method [1]
- Looks for camera/object motion and 3D structure that best explains the image data
- A **model-free** approach that exploits the **complete** image data of the object's shape
- An elegant and simple solution based on matrix factorization (SVD)



Gunnar Johansson, James Maas (1971)

[1] Carlo Tomasi and Takeo Kanade (1992) Shape and Motion from Image Streams Under Orthography: A Factorization Method

The Factorisation Method

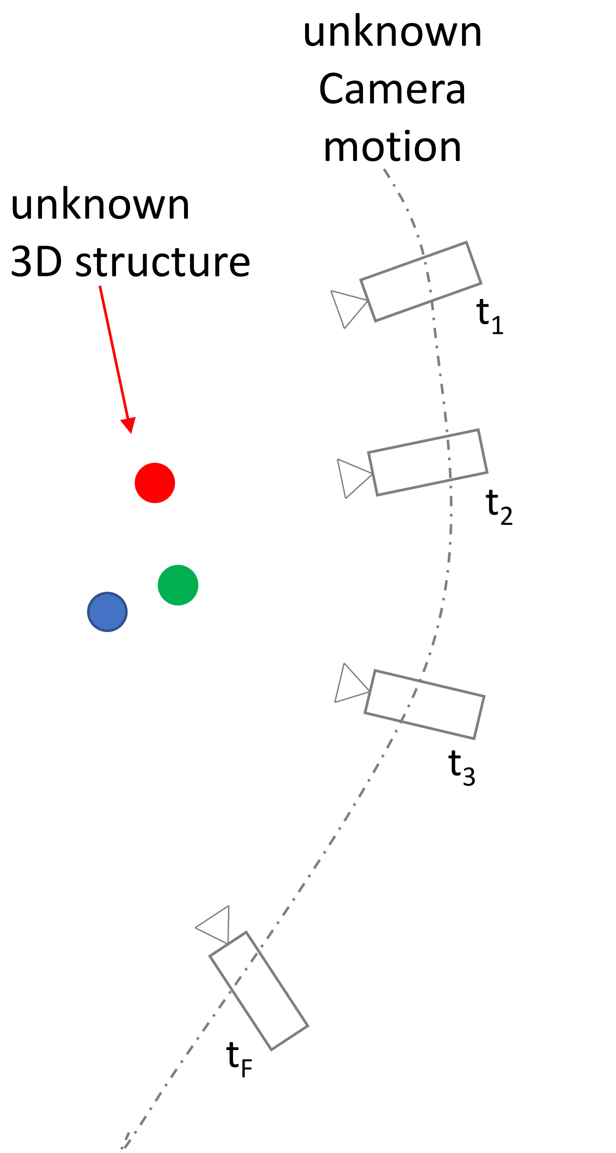


Image data:

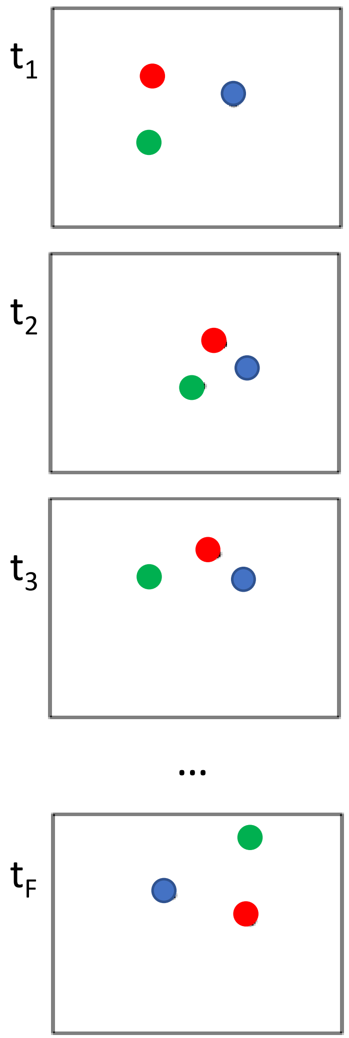


Image positions of point 1

3D position of point 1

$$\underbrace{\mathbf{W}}_{2F \times P} = \begin{bmatrix} u_{1,1} & \\ v_{1,1} & \\ u_{2,1} & \\ v_{2,1} & \cdots \\ \vdots & \\ u_{F,1} & \\ v_{F,1} & \end{bmatrix} = \underbrace{\mathbf{M}}_{2F \times 3} \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_F \end{bmatrix} \underbrace{\mathbf{S}}_{3 \times P} \begin{bmatrix} X_1 & \\ Y_1 & \cdots \\ Z_1 & \end{bmatrix}$$

$2F \times P$ $2F \times 3$ $3 \times P$

The Factorisation Method

- Trajectory matrix W has a lot of redundancy
- The trajectories reside in a low-dimensional **subspace**
  - 3D for orthographic
- Reflected in the **rank** of matrix W
  - Thus trajectory matrix W has rank 3 (**rank deficient**)

The Factorisation Method

- Using **SVD** (**<u>s</u>ingular <u>v</u>alue <u>d</u>ecomposition**)

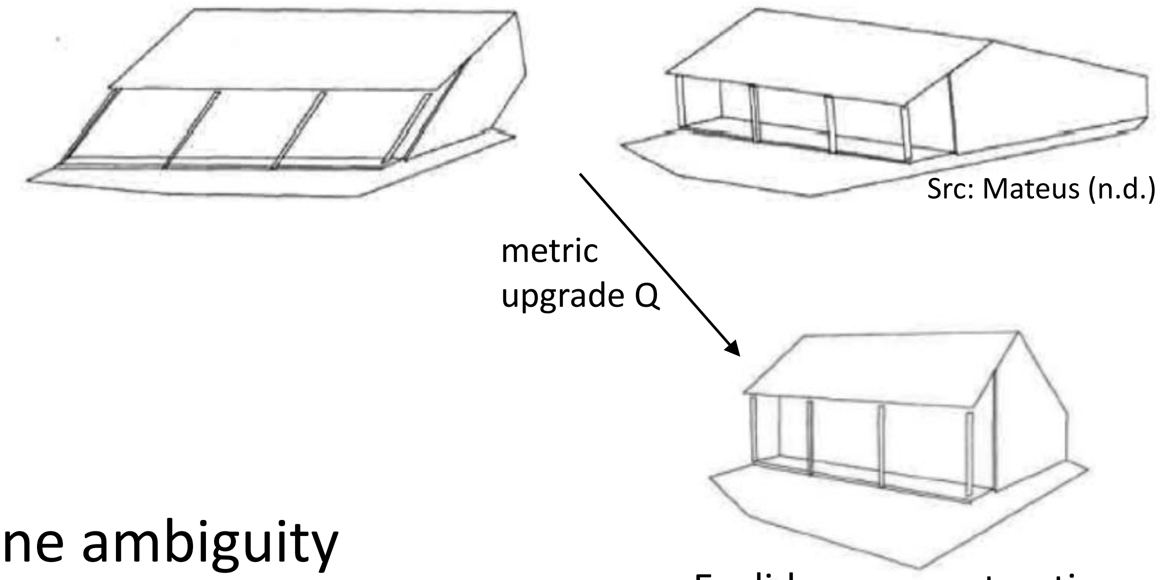
$$\underset{2F\times P}{W} \overset{SVD}{=} \underset{2F\times 2F}{U} \; \underset{2F\times P}{\Sigma} \; \underset{P\times P}{V^T}$$

$$\underset{2F\times P}{W} \overset{SVD}{=} \underset{2F\times 3}{U'} \; \underset{3\times 3}{\Sigma'} \; \underset{3\times P}{V'^T}$$ ...reduced rank 3

$$\underset{2F\times P}{W} \overset{SVD}{=} \underbrace{\underset{2F\times 3}{U'} \; \underset{3\times 3}{\Sigma'^{\frac{1}{2}}}} \; \underbrace{\underset{3\times 3}{\Sigma'^{\frac{1}{2}}} \; \underset{3\times P}{V'^T}}$$

$$W = M_{\text{affine}} \, S_{\text{affine}}$$ ...unique up to an affine transformation

Affine reconstructions



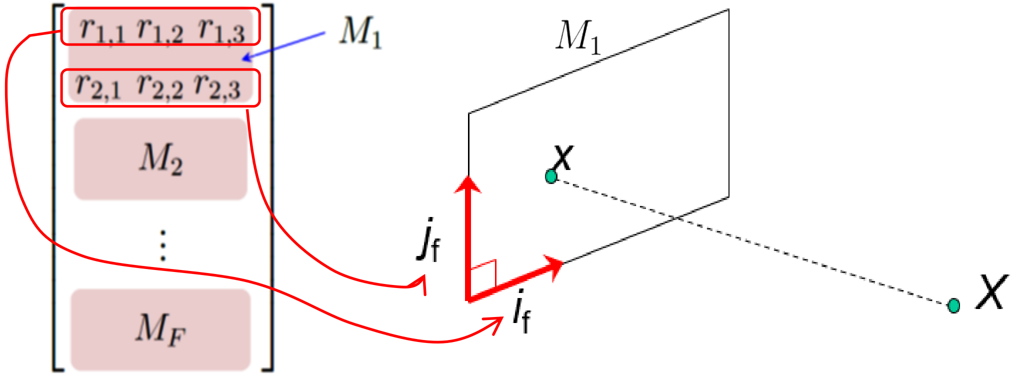
Euclidean reconstruction

- Need to find an upgrading matrix Q to remove affine ambiguity
- Imposing the **metric constraints**:

$$\mathbf{i}_f^T \mathbf{i}_f = \mathbf{i}_f Q Q^T \mathbf{i}_f^T = 1$$
$$\mathbf{j}_f^T \mathbf{j}_f = \mathbf{j}_f Q Q^T \mathbf{j}_f^T = 1$$
$$\mathbf{i}_f^T \mathbf{j}_f = \mathbf{i}_f Q Q^T \mathbf{j}_f^T = 0$$

- Thus:

$$W = \underbrace{M_{\text{affine}} \, Q} \; \underbrace{Q^{-1} \, S_{\text{affine}}}$$
$$W = M \quad S$$



Rigid-body Structure from Motion (SfM)

Image positions of point 1

3D position of point 1

$$
\underbrace{\begin{bmatrix} u_{1,1} & \\ v_{1,1} & \\ u_{2,1} & \\ v_{2,1} & \\ \vdots & \cdots \\ u_{F,1} & \\ v_{F,1} & \end{bmatrix}}_{\textstyle 2F \times P}^{\textstyle \boldsymbol{W}}
=
\underbrace{\begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_F \end{bmatrix}}_{\textstyle 2F \times 3}^{\textstyle \boldsymbol{M}}
\underbrace{\begin{bmatrix} X_1 & \\ Y_1 & \cdots \\ Z_1 & \end{bmatrix}}_{\textstyle 3 \times P}^{\textstyle \boldsymbol{S}}
$$

Rigid body, orthographic

Carlo Tomasi and Takeo Kanade (1992) Shape and Motion from Image Streams Under Orthography: A Factorization Method

Non-rigid Structure from Motion (NRSfM)

Image positions of point 1

3D positions of point 1 in *F* frames

$$\underbrace{\begin{bmatrix} u_{1,1} & \\ v_{1,1} & \\ u_{2,1} & \\ v_{2,1} & \\ \vdots & \cdots \\ u_{F,1} & \\ v_{F,1} & \end{bmatrix}}_{\boldsymbol{W}:\ 2F \times P} = \underbrace{\begin{bmatrix} M_1 & & & \\ & M_2 & & \\ & & \ddots & \\ & & & M_F \end{bmatrix}}_{\boldsymbol{M}:\ 2F \times 3F} \underbrace{\begin{bmatrix} X_1 & \\ Y_1 & \\ Z_1 & \\ X_2 & \\ Y_2 & \cdots \\ Z_2 & \\ \vdots & \\ X_F & \\ Y_F & \\ Z_F & \end{bmatrix}}_{\boldsymbol{S}:\ 3F \times P}$$

Non-rigid body, orthographic

Bregler et al. (2000) Recovering Non-Rigid 3D Shape from Image Streams

Non-rigid Structure from Motion (NRSfM)

Image positions of point 1

3D positions of point 1 in *F* frames

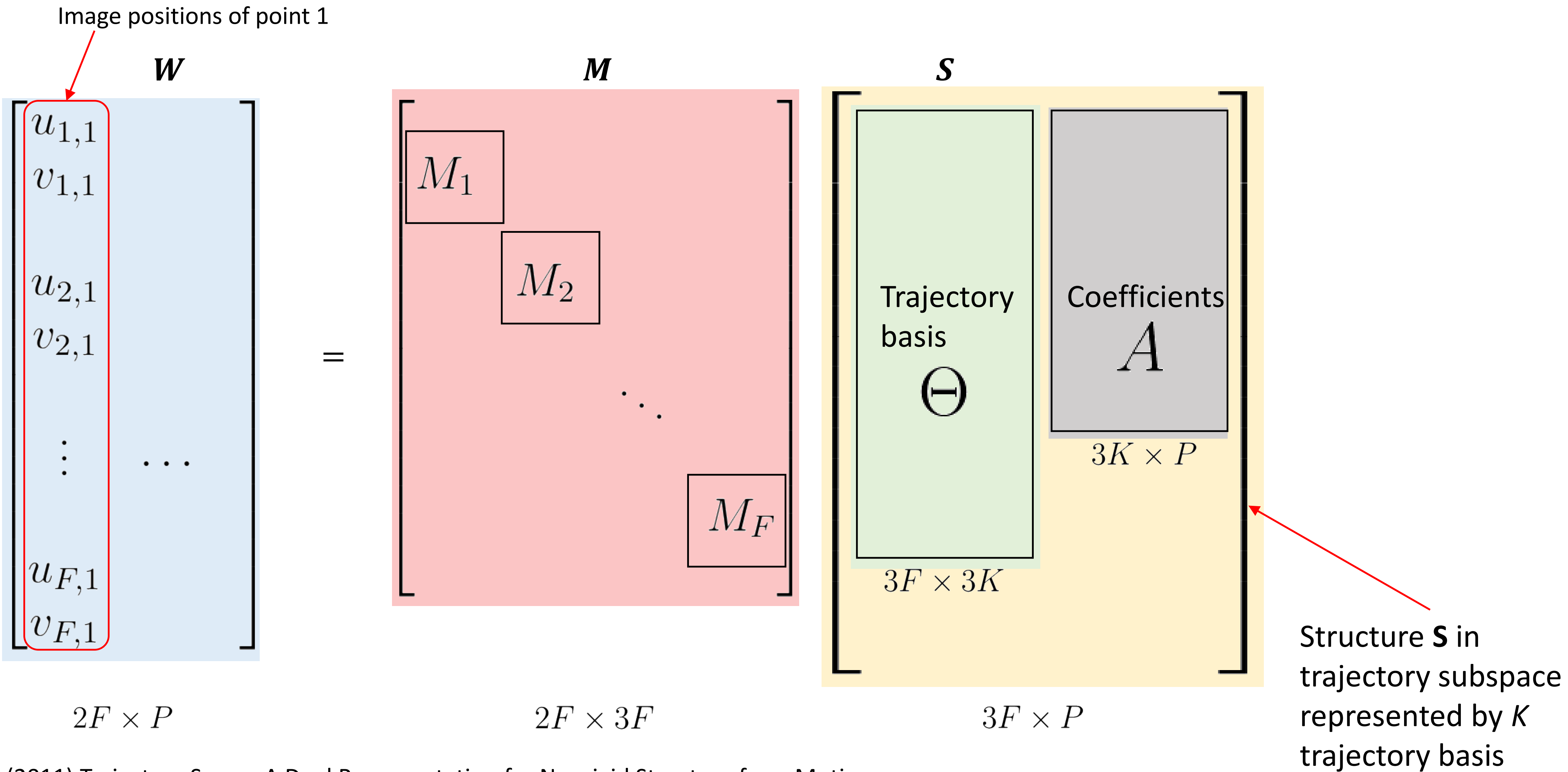
$$\underbrace{\begin{bmatrix} u_{1,1} & \\ v_{1,1} & \\ u_{2,1} & \\ v_{2,1} & \\ \vdots & \cdots \\ u_{F,1} & \\ v_{F,1} & \end{bmatrix}}_{\boldsymbol{W}:\ 2F \times P} = \underbrace{\begin{bmatrix} M_1 & & & \\ & M_2 & & \\ & & \ddots & \\ & & & M_F \end{bmatrix}}_{\boldsymbol{M}:\ 2F \times 3F} \underbrace{\begin{bmatrix} X_1 & \\ Y_1 & \\ Z_1 & \\ X_2 & \\ Y_2 & \cdots \\ Z_2 & \\ \vdots & \\ X_F & \\ Y_F & \\ Z_F & \end{bmatrix}}_{\boldsymbol{S}:\ 3F \times P}$$

More unknowns!
Harder to recover 3D structure

Non-rigid body, orthographic

Bregler et al. (2000) Recovering Non-Rigid 3D Shape from Image Streams

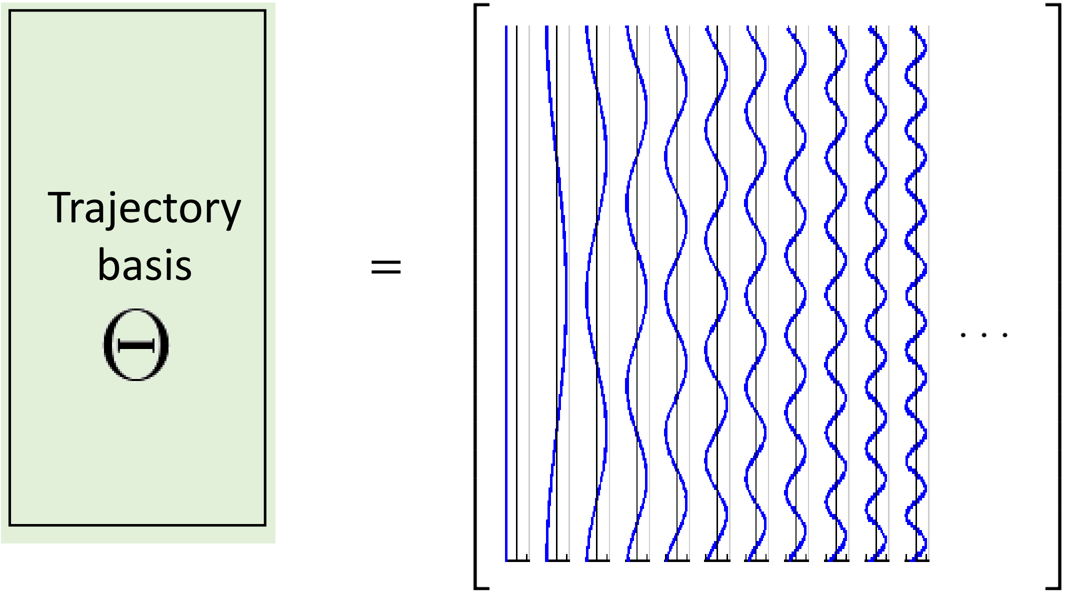
Trajectory Space Factorisation (NRSfM)



Akhter et al. (2011) Trajectory Space: A Dual Representation for Nonrigid Structure from Motion

Trajectory Space Factorisation (NRSfM)

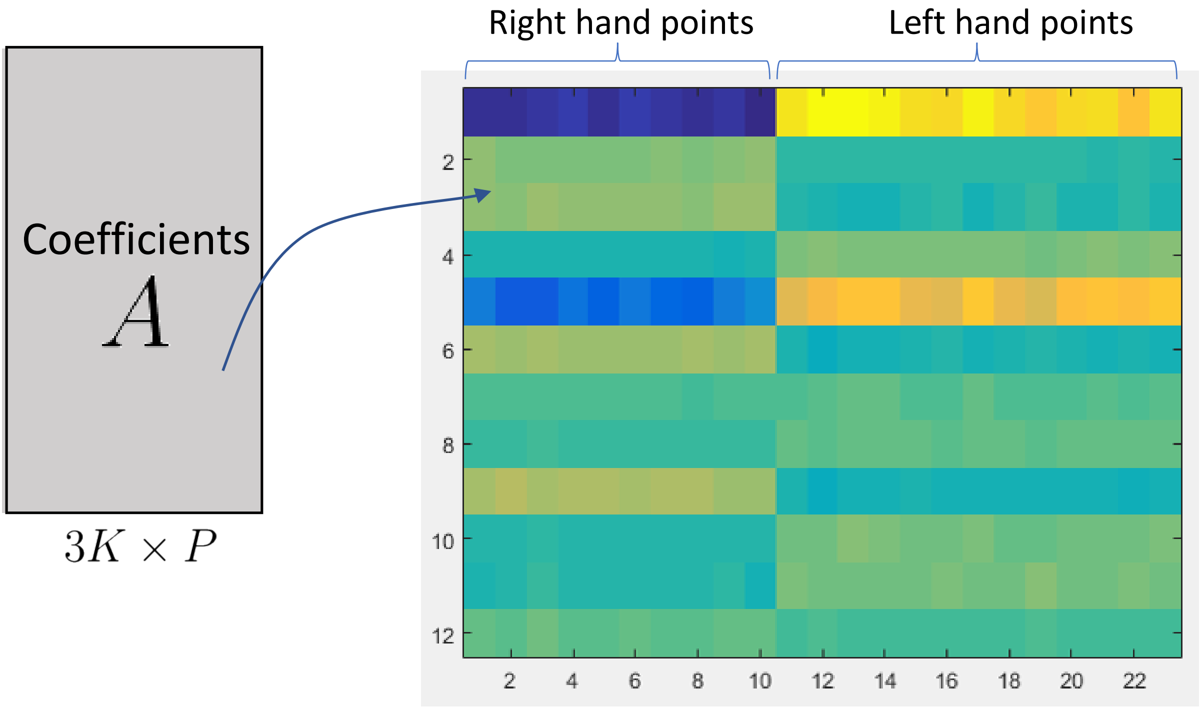
- DCT as Trajectory basis

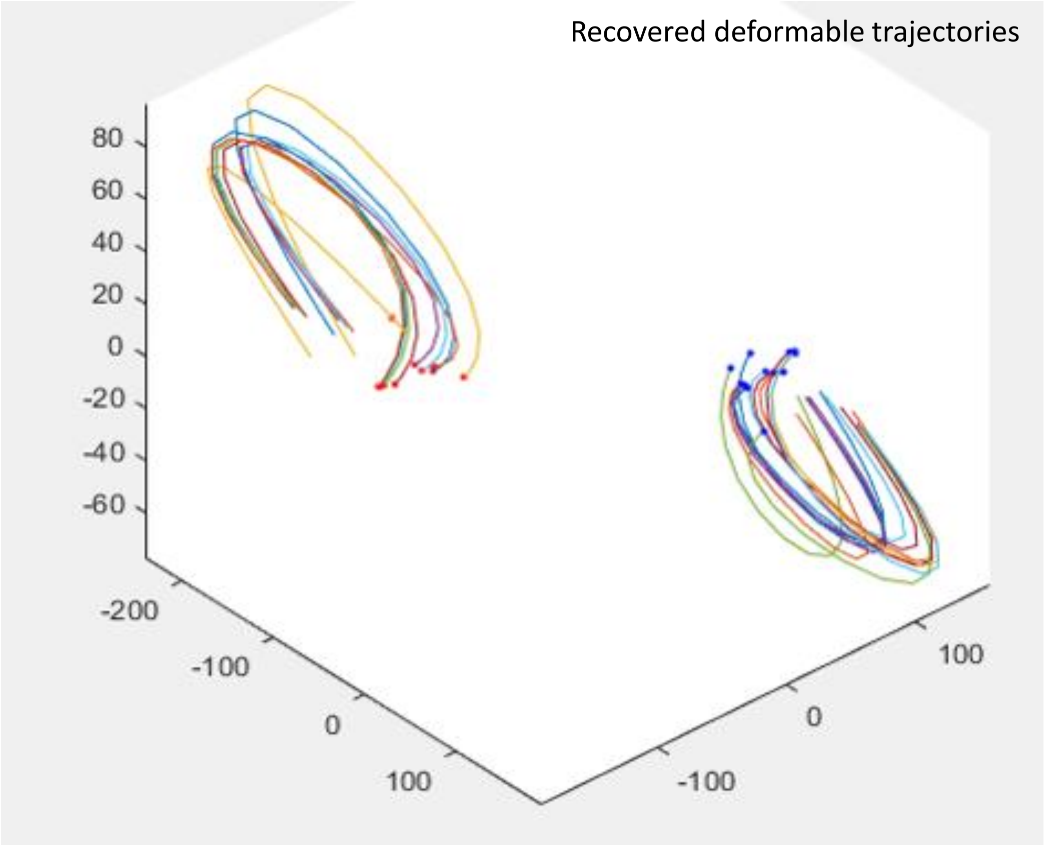


- Advantages of trajectory space factorisation:
  - Pre-defined basis
  - Less unknowns, hence easier metric upgrade
  - Trajectory basis are object independent
  - Trajectory basis can be ‘recycled’ across video sequences

Trajectory Space Factorisation applied to sign videos

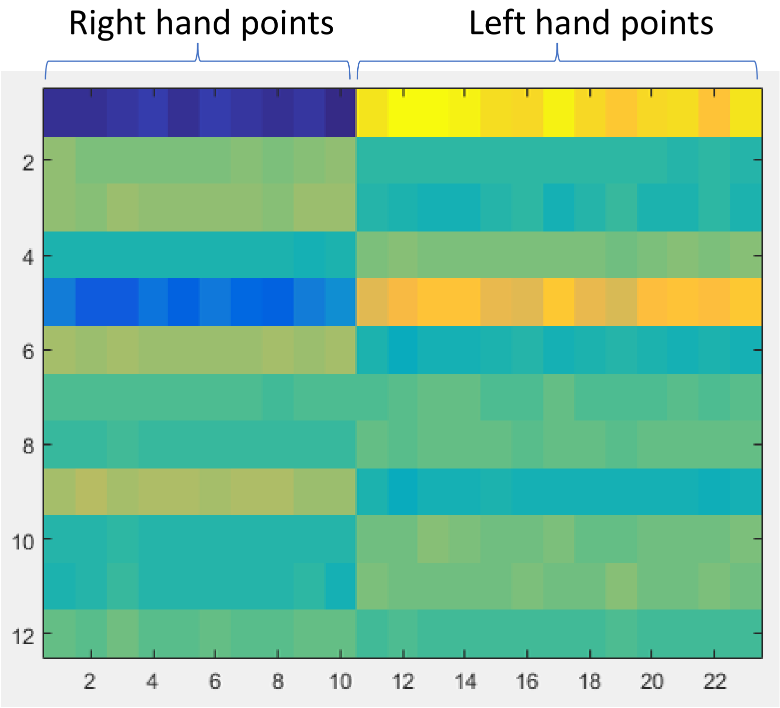
- Motion:
  - motion of signer (body rotations, body leanings, …)
- Non-rigid Shape:
  - the trajectories of the hands are the shape deformations with respect to the signer’s centroid





Trajectory Space Factorisation for ASLR

- Coefficient matrix A encodes useful information on the motion trajectories of the 2 hands
- We use the coefficients for recognising sign language phonemes



- Non-parametric statistical measures extracted from A:
  - Five-number summary statistics:
    - (median, 1st quartile q_1, 3rd quartile q_3, minimum, maximum)
  - Outlier removal:

$$[q_1 - 1.5 \times \mathrm{iqr} \cdots q_3 + 1.5 \times \mathrm{iqr}]$$

where the interquartile range is:

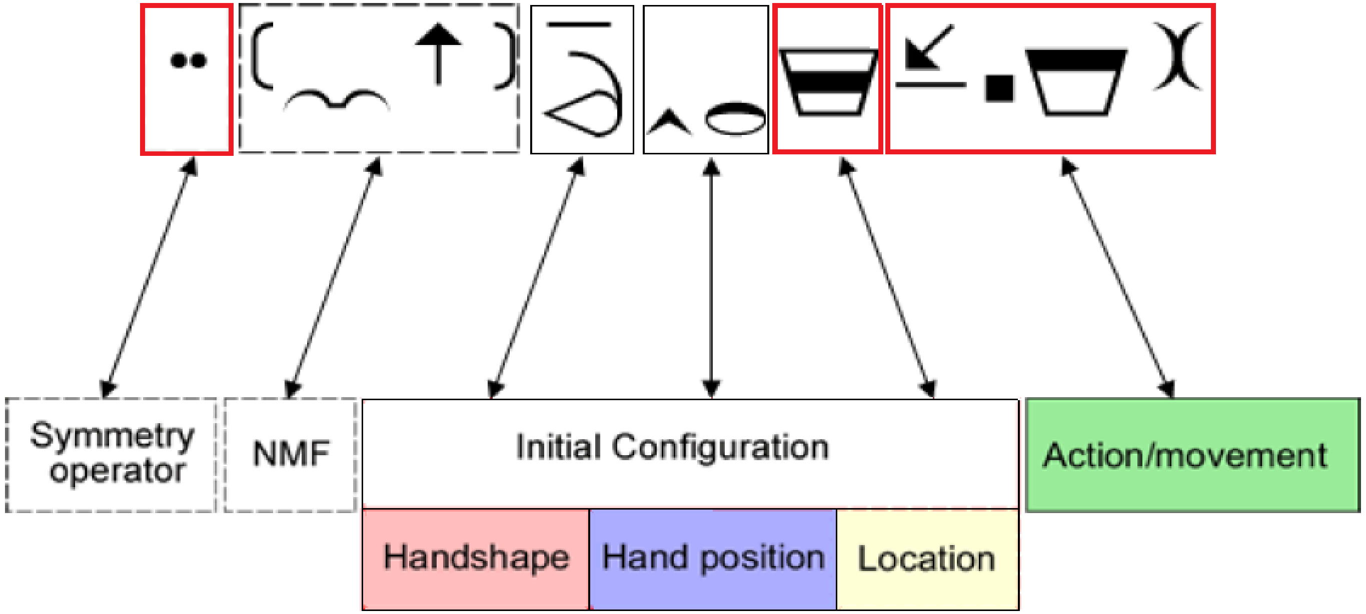
$$\mathrm{iqr} = q_3 - q_1$$

Hand Motion classifiers

Classifier	Class labels	Description
symmetry	asym	asymmetric hand motion
	sym	mirror symmetry with respect to midline
	sym	radial symmetry with respect to torso centroid
h1 stationary	moving	dominant hand (h1) is moving
	stationary	h1 is not moving
h2 stationary	moving	non-dominant hand (h2) is moving
	stationary	h2 is not moving
	at rest	h2 is not moving and is at its rest position (e.g. on signer's lap)
motion	0	no hand motion, small hand motions, or irregular motion
	mu	upward hand movement
	mul	up-left hand movement
	ml	left hand movement
	mdl	downard-left hand movement
	md	downward hand movement
	mdr	downward-right hand movement
	mr	right hand movement
	mur	upward-right hand movement
	cm	hand follows a clockwise rotational motion
	ccm	hand follows a counter-clockwise rotational motion

HamNoSys transcription

- Classifier – Phoneme correspondence
- Hamburg Notation System (HamNoSys) [1]
  - Detailed phonetic description of signs
- Loess filtering

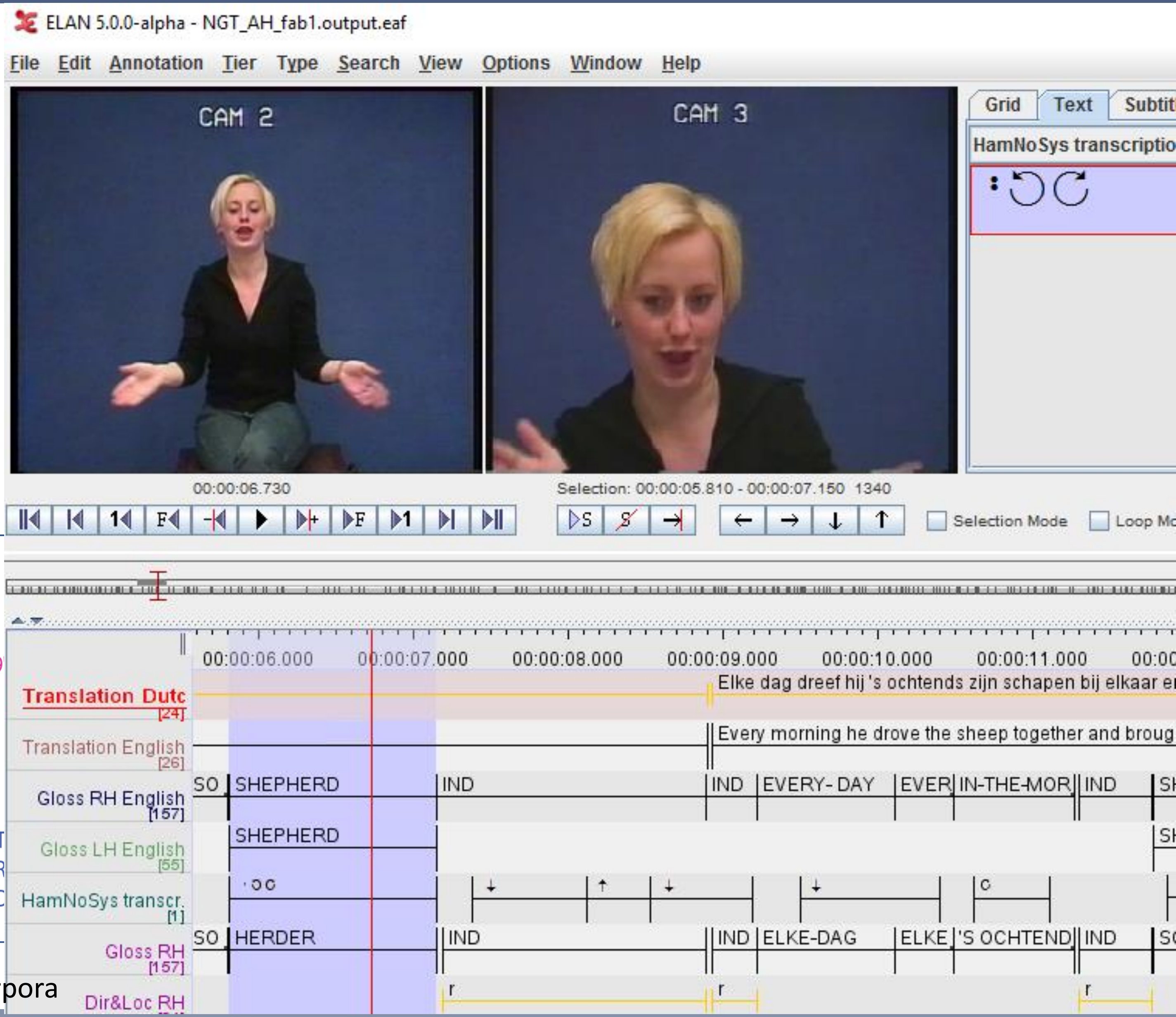


Classifier	Class labels	HamNoSys symbols
symmetry	asym	
	sym	••
	sym	•• •
h1 stationary	moving	
	stationary	
h2 stationary	moving	
	stationary	
	at rest	
motion	0	
	mu	↑
	mul	↗
	ml	↗
	mdl	↘
	md	↓
	mdr	↙
	mr	←
	mur	↖
	cm	↑⊂ ↻
	ccm	↑⊃ ↺

[1] Hanke (2004) HamNoSys – Representing Sign Language Data in Language Resources and Language Processing Contexts

Integration with Annotation Tools

- ELAN annotation tool [1]
  - Additional tier for HamNoSys
  - ELAN annotation file (EAF)
  - XML-based format
  - HamNoSys Unicode font




```
<TIER DEFAULT_LOCALE="en" LINGUISTIC_TYPE_REF="Default" TIER_ID="HamNoSys transcr.">
  <ANNOTATION>
    <ALIGNABLE_ANNOTATION ANNOTATION_ID="a1009" TIME_SLOT_REF1="ts113" TIME_SLOT_REF2="ts124">
      <ANNOTATION_VALUE>□□</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION> <ALIGNABLE_ANNOTATION ANNOTATION_ID="a1010" TIME_SLOT_REF1="ts119
      <ANNOTATION_VALUE>□□</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION>
</TIER>
<LINGUISTIC_TYPE GRAPHIC_REFERENCES="false" LINGUISTIC_TYPE_ID="Default" TIME_ALIGNABLE="true"/>
<LINGUISTIC_TYPE CONSTRAINTS="Symbolic_Association" GRAPHIC_REFERENCES="false" LINGUISTIC_TYPE_ID="T
<LINGUISTIC_TYPE CONSTRAINTS="Symbolic_Association" GRAPHIC_REFERENCES="false" LINGUISTIC_TYPE_ID="R
<LINGUISTIC_TYPE CONSTRAINTS="Symbolic_Association" GRAPHIC_REFERENCES="false" LINGUISTIC_TYPE_ID="D
```

[1] Crasborn et al. (2008) Enhanced ELAN functionality for sign language corpora

Experiments

- ECHO Sign Language (NGT) Corpus
- Temporal sliding window ($F = 15$)
- Trajectory factorization (trajectory basis $K = 4$)
- Hand motion classifiers:
  - K-nearest neighbor (k-NN)
  - Support Vector Machines (SVMs)
  - XGBoost

k-NN parameter
k = 3

SVM parameters
radial basis function, cost C = 10, γ = 0.1

XGBoost hyperparameter selection

Hyperparameter	Value	Tuning approach	Range
Number of trees	1000	Fixed	
Learning rate η	0.04	Fixed $\rightarrow$ Fine-tuned	0.02 $\rightarrow$ [0.02, 0.04, 0.06, 0.08, 0.1]
Row sampling	0.70	Grid Search	[0.5, 0.7, 0.75, 0.8, 1.0]
Column sampling	0.4	Grid Search	[0.3, 0.4, 0.5, 0.6, 0.8, 1.0]
Max tree depth	8	Grid Search	[4, 6, 8, 10]
Min leaf weight	1	Fixed $\rightarrow$ Fine-tuned	3 $\rightarrow$ [1, 5]
Min split gain γ	0	Fixed	

Results

- Classification accuracy:

Classifier	XGBoost	SVM	k-NN	baseline
h1 motion	89.49%	84.57%	70.74%	72.21%
h1 stationary	97.74%	96.94%	84.97%	96.54%
h2 stationary	86.97%	84.04%	63.16%	77.39%
symmetry	87.37%	76.99%	70.08%	58.51%

- Best classification results obtained with XGBoost

Results

- Confusion matrices for h1 hand motion classifier (**k-NN**)

Predicted \ Actual	0	mul	ml	mdl	md	mdr	mr	mur	mu	cm	ccm
0	327	0	0	0	0	0	0	0	0	0	0
mul	8	16	0	0	0	1	0	0	0	0	0
ml	26	0	28	0	0	0	1	0	0	0	0
mdl	8	0	0	5	0	0	0	0	0	0	0
md	37	0	0	0	25	0	0	0	0	0	0
mdr	1	0	0	0	0	7	0	0	0	0	0
mr	39	0	1	0	0	0	41	0	0	0	0
mur	7	0	0	0	0	0	0	7	0	0	1
mu	0	0	0	0	0	0	0	0	3	0	0
cm	5	0	0	0	0	0	0	0	0	3	0
ccm	85	0	0	0	0	0	0	0	0	0	70

Results

- Confusion matrices for h1 hand motion classifier (**SVM**)

Predicted \ Actual	0	mul	ml	mdl	md	mdr	mr	mur	mu	cm	ccm
0	431	0	0	0	0	0	0	0	0	0	0
mul	3	16	0	0	0	1	0	0	0	0	0
ml	6	0	29	0	0	0	1	0	0	0	0
mdl	6	0	0	5	0	0	0	0	0	0	0
md	13	0	0	0	25	0	0	0	0	0	0
mdr	1	0	0	0	0	7	0	0	0	0	0
mr	8	0	0	0	0	0	41	0	0	0	0
mur	5	0	0	0	0	0	0	7	0	0	2
mu	0	0	0	0	0	0	0	0	3	0	0
cm	2	0	0	0	0	0	0	0	0	3	0
ccm	68	0	0	0	0	0	0	0	0	0	69

Results

- Confusion matrices for h1 hand motion classifier (**XGBoost**)

Predicted \ Actual	0	mul	ml	mdl	md	mdr	mr	mur	mu	cm	ccm
0	466	0	0	0	0	0	0	0	0	0	0
mul	3	16	0	0	0	1	0	0	0	0	0
ml	13	0	29	0	1	0	0	0	0	0	0
mdl	2	0	0	5	0	0	0	0	0	0	0
md	14	0	0	0	24	0	0	0	0	0	0
mdr	2	0	0	0	0	7	0	0	0	0	0
mr	9	0	0	0	0	0	42	0	0	0	0
mur	0	0	0	0	0	0	0	7	0	0	0
mu	0	0	0	0	0	0	0	0	3	0	0
cm	2	0	0	0	0	0	0	0	0	3	0
ccm	32	0	0	0	0	0	0	0	0	0	71

Results

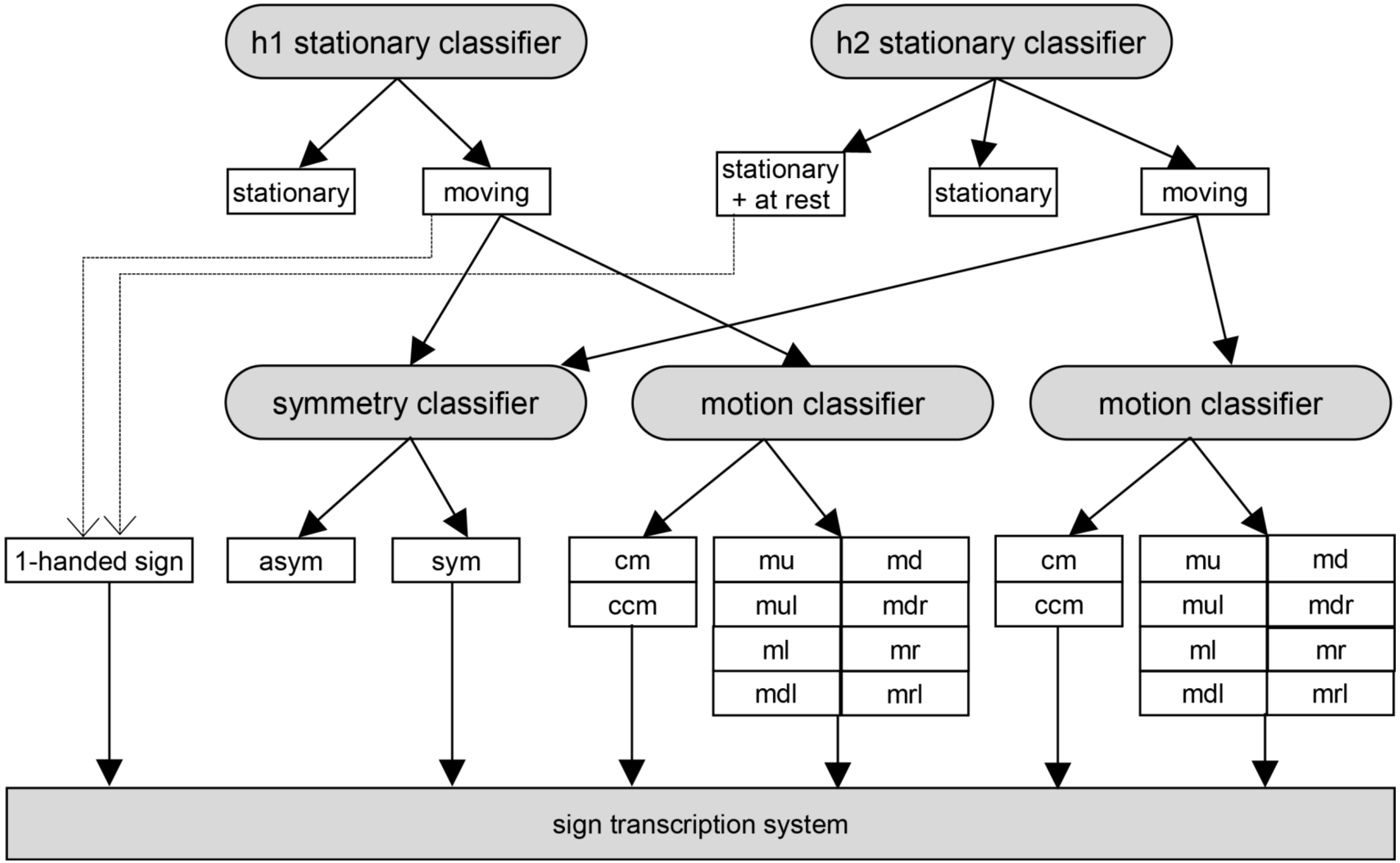
- Confusion matrices for h1 hand motion classifier (**XGBoost**)

Predicted \ Actual	0	mul	ml	mdl	md	mdr	mr	mur	mu	cm	ccm
0	466	0	0	0	0	0	0	0	0	0	0
mul	3	16	0	0	0	1	0	0	0	0	0
ml	13	0	29	0	1	0	0	0	0	0	0
mdl	2	0	0	5	0	0	0	0	0	0	0
md	14	0	0	0	24	0	0	0	0	0	0
mdr	2	0	0	0	0	7	0	0	0	0	0
mr	9	0	0	0	0	0	42	0	0	0	0
mur	0	0	0	0	0	0	0	7	0	0	0
mu	0	0	0	0	0	0	0	0	3	0	0
cm	2	0	0	0	0	0	0	0	0	3	0
ccm	32	0	0	0	0	0	0	0	0	0	71

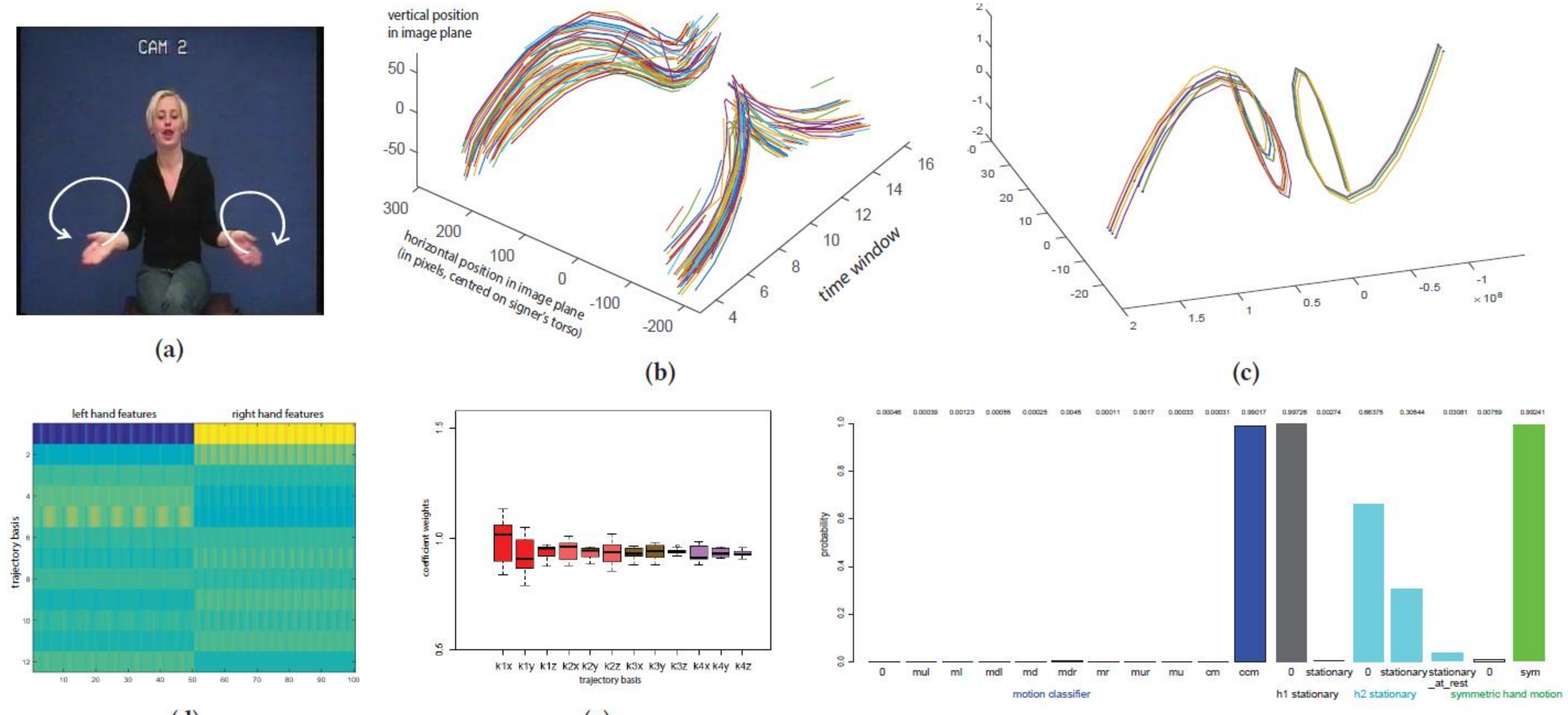
Small hand motions (stationary) or irregular motions.

Also coarticulation effects.

Hand Motion classifier hierarchy

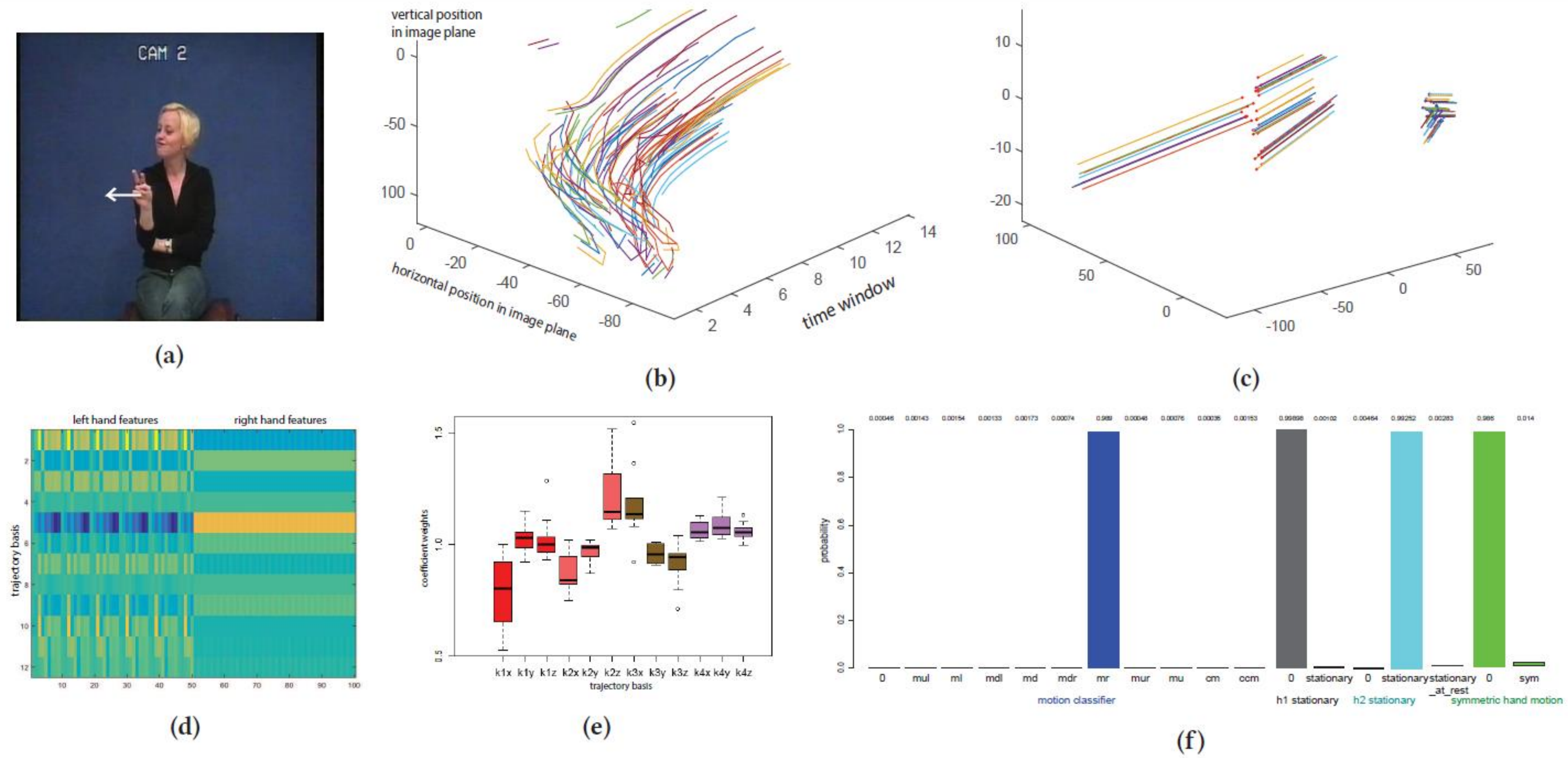


Results



(a) (b) (c) (d) (e) (f)

Results



(a) (b) (c) (d) (e) (f)

Conclusions

- Trajectory space factorization can be successfully applied to sign language videos
  - It is able to separate global signer motion from hand trajectory motion (posed as an NRSfM problem)
  - Coefficient matrix encodes rich information on hand trajectories – this can be used for hand motion classification
- Our XGBoost-based hand motion classification system achieves successful recognition rates for various hand motion types, like symmetric motion, circular motion, linear motion
- Explored how our hand motion classification system can be used for transcribing sign language (e.g. via the use of HamNoSys and the ELAN annotation tool)

Future Work

- Incorporating hand motion classifiers for more complex phonological elements, like zig-zag motions
- Investigate how our method can be used for deriving phonetically meaningful sub-units for training an ASLR system
- Further integration with sign language annotation tools, such as ELAN
- From phoneme classifiers to word-level HMMs

- Thank you for your attention