Sign Language Detection “in the Wild” with Recurrent Neural Networks

Mark Borg
Kenneth P. Camilleri

Systems and Control Engineering,
University of Malta

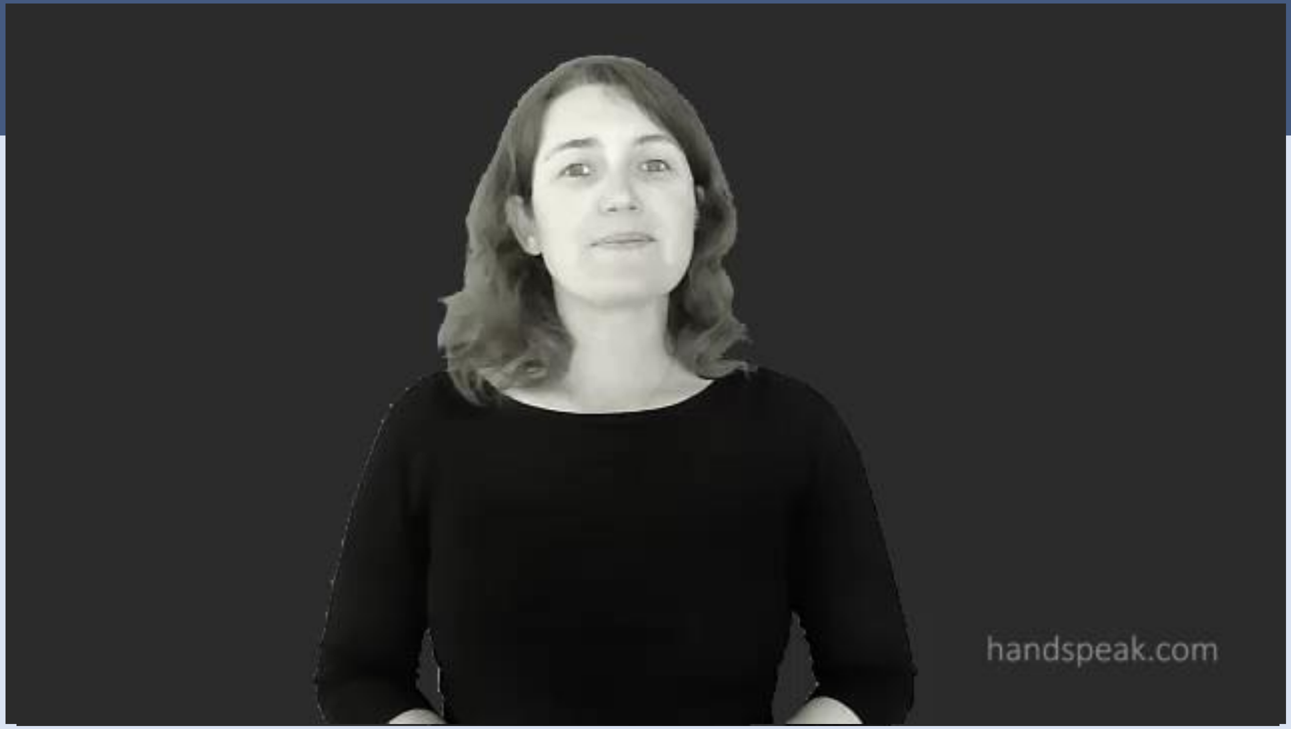
IEEE 2019 ICASSP
International Conference on Acoustics, Speech and Signal Processing

Sign Languages

- **Visual languages**

- Multi-modal
- Concurrent modalities

- Articulators:
  - Manual
    - Hand motion
    - Hand shapes
    - Place of articulation
  - Non-Manual
    - Mouth patterns
    - Facial Expressions
    - Body posture

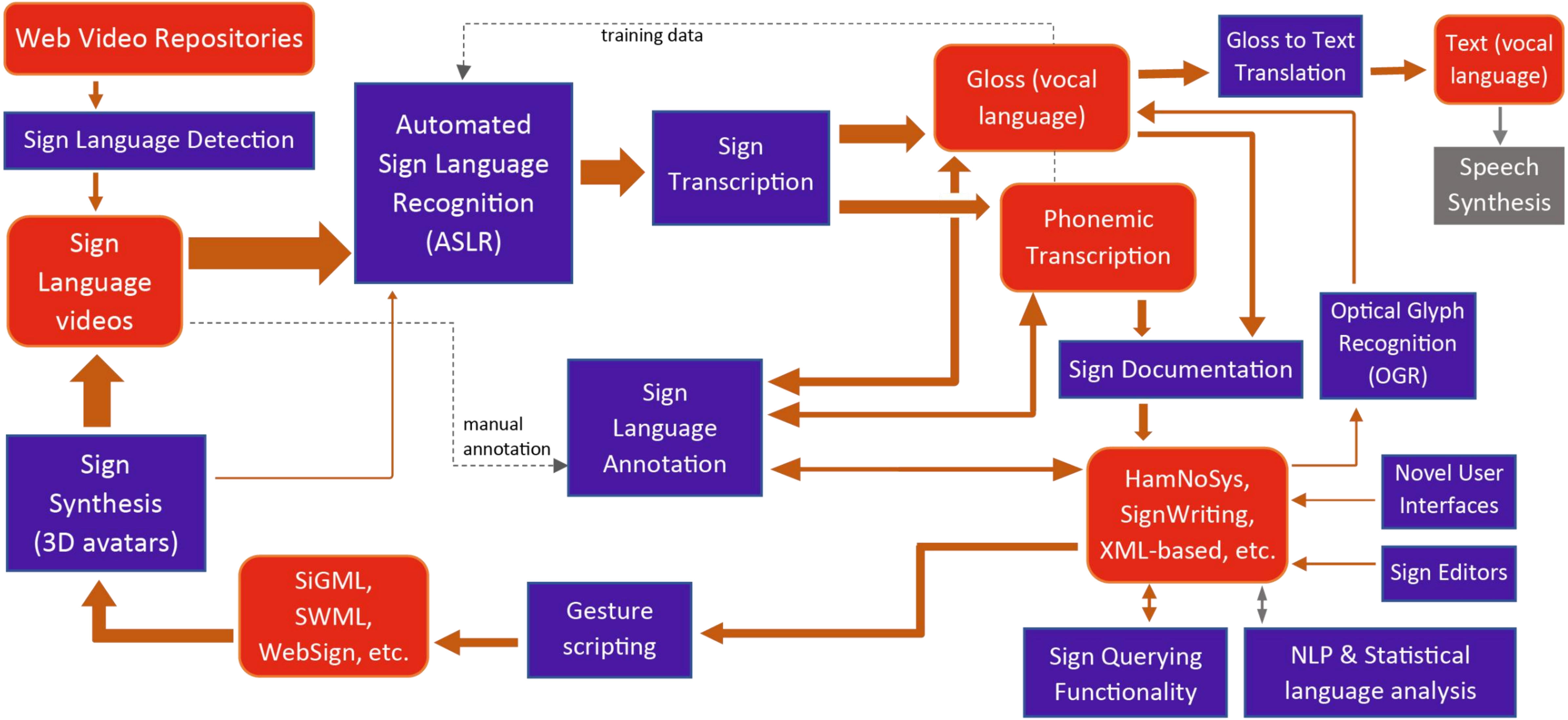


American Sign Language (ASL): **PERSON WHATEVER IX-they JUDGE IX1 REALLY WASTE, MINUTE [shook-head] INSTEAD-OF MINUTE IX WHAT-conj LOVE ACCEPT WHO**

English equivalent: **For every minute we judge, we have squandered a minute we could have used to accept and love someone.**

Source: HandSpeak

Sign Language technologies

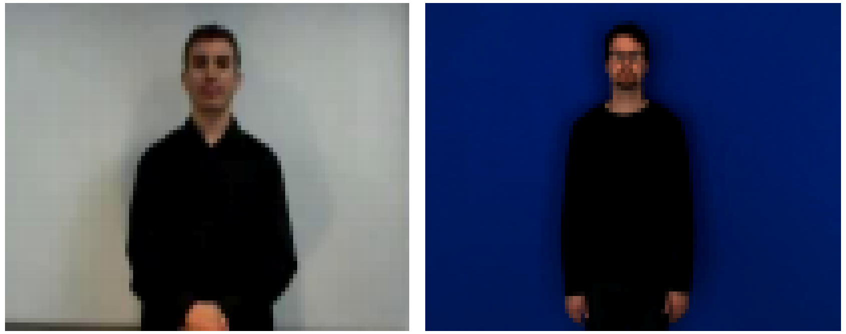


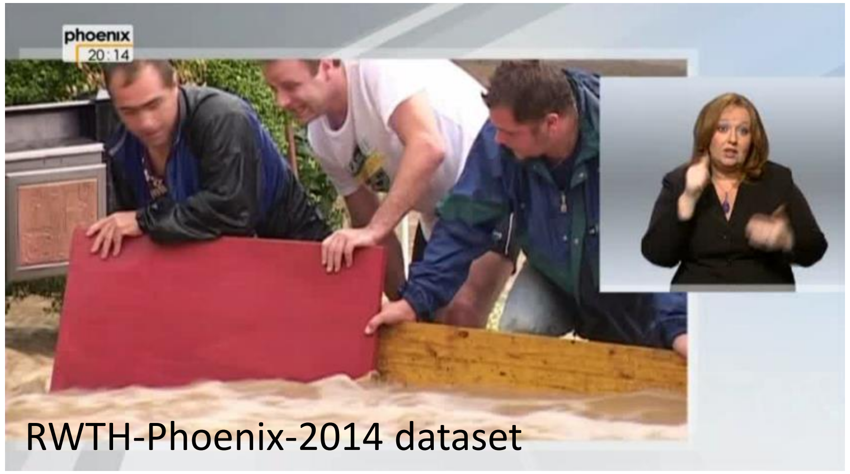
Sign Language Detection – state of the art

- Monteiro et al. (2012 SIGACCESS)
    - Face detection, background subtraction
    - Hand-crafted visual features: velocity-based
    - SVM
- Shipman et al. (2015 JCDL, 2017 SIGACCESS)
    - Face detection, background subtraction
    - Hand-crafted visual features: polar motion profiles
    - SVM
- Gebre et al. (2013 ICIP)
    - Face detection, skin detection
    - hand-crafted visual features
    - random forests
- Yanovich (2016 LREC)
    - Identification of major sign language constructs: fingerspellings, classifiers, …
    - Hand-crafted visual features
    - k-NN classifier
- Gebre et al. (2014 Comp. Ling.)
    - Identification of particular sign languages: BSL, DSL, FBSL, FSL, GSL, NGT
    - Sparse auto-encoder and 3D CNN

Datasets – the need for sign language detection datasets

- No signing in generic video action recognition datasets, like AVA, THUMOS, …
- Previous work in SL detection
  - Datasets not made publicly available
  - Small size (~200 videos)
- Sign Language Recognition (ASLR) datasets, Phoenix, SIGNUM, VGG BBC pose, …
  - Trimmed
  - Captured under constrained conditions





RWTH-Phoenix-2014 dataset

Src: Dreuw et al. (2010)

“Signing in the Wild” dataset

- Untrimmed videos
- Each video can include multiple signing and non-signing events
- Harvested from YouTube

3 categories: Signing Speaking Other

- **1120** video segments
- Each video segment:
  - Up to 6.6 minutes (sampled at 5 Hz)
  - Up to 2000 frames long
- **1.45 million frames** in total

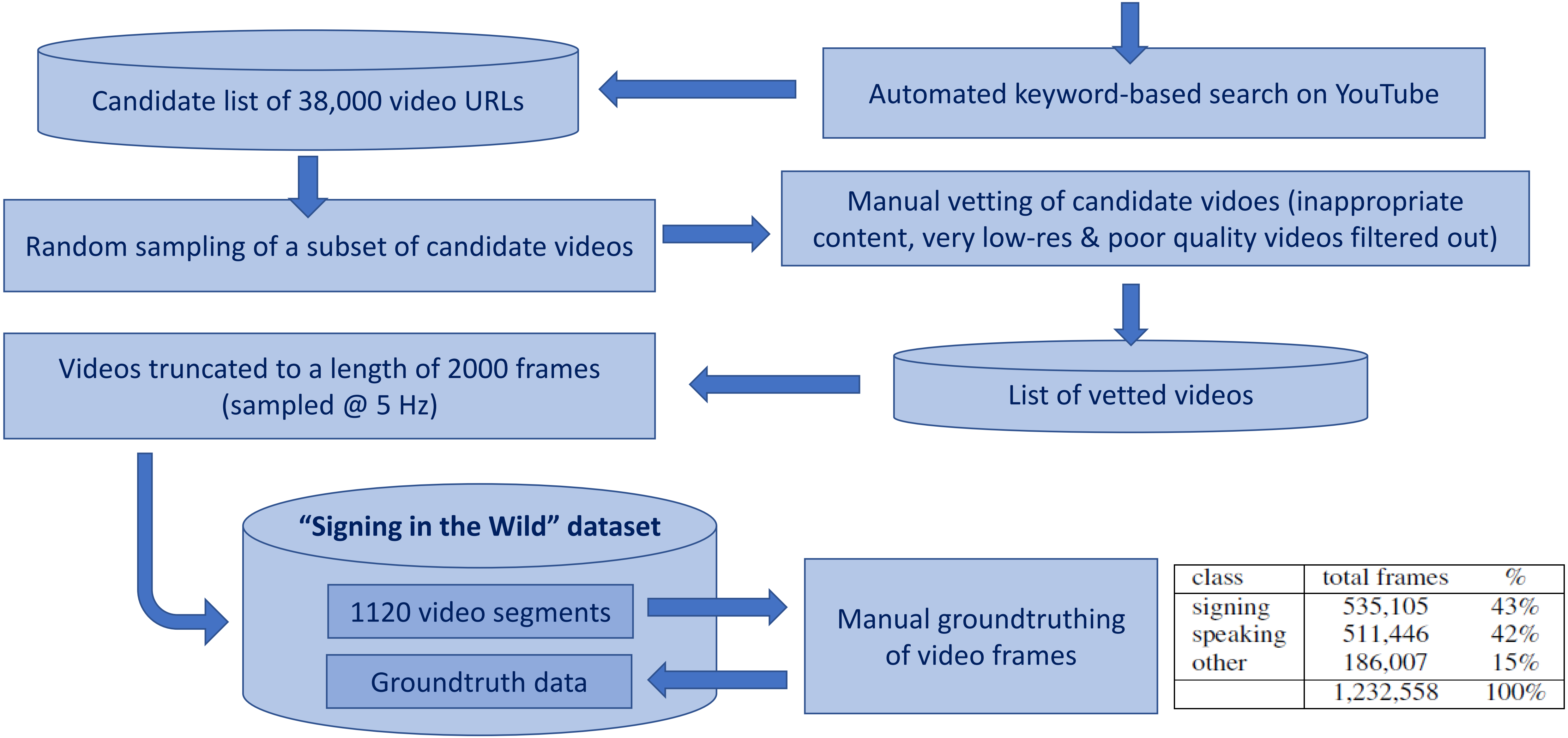
Groundtruthing:

- Frame-level
- 10-frame temporal context
- **1.23 million frames**

Publicly available:

- **IEEE DataPort**
- https://github.com/mark-borg/Signing-in-the-Wild-dataset

“Signing in the Wild” dataset



Automated keyword-based search on YouTube

Candidate list of 38,000 video URLs

Random sampling of a subset of candidate videos

Manual vetting of candidate vidoes (inappropriate content, very low-res & poor quality videos filtered out)

List of vetted videos

Videos truncated to a length of 2000 frames (sampled @ 5 Hz)

“Signing in the Wild” dataset

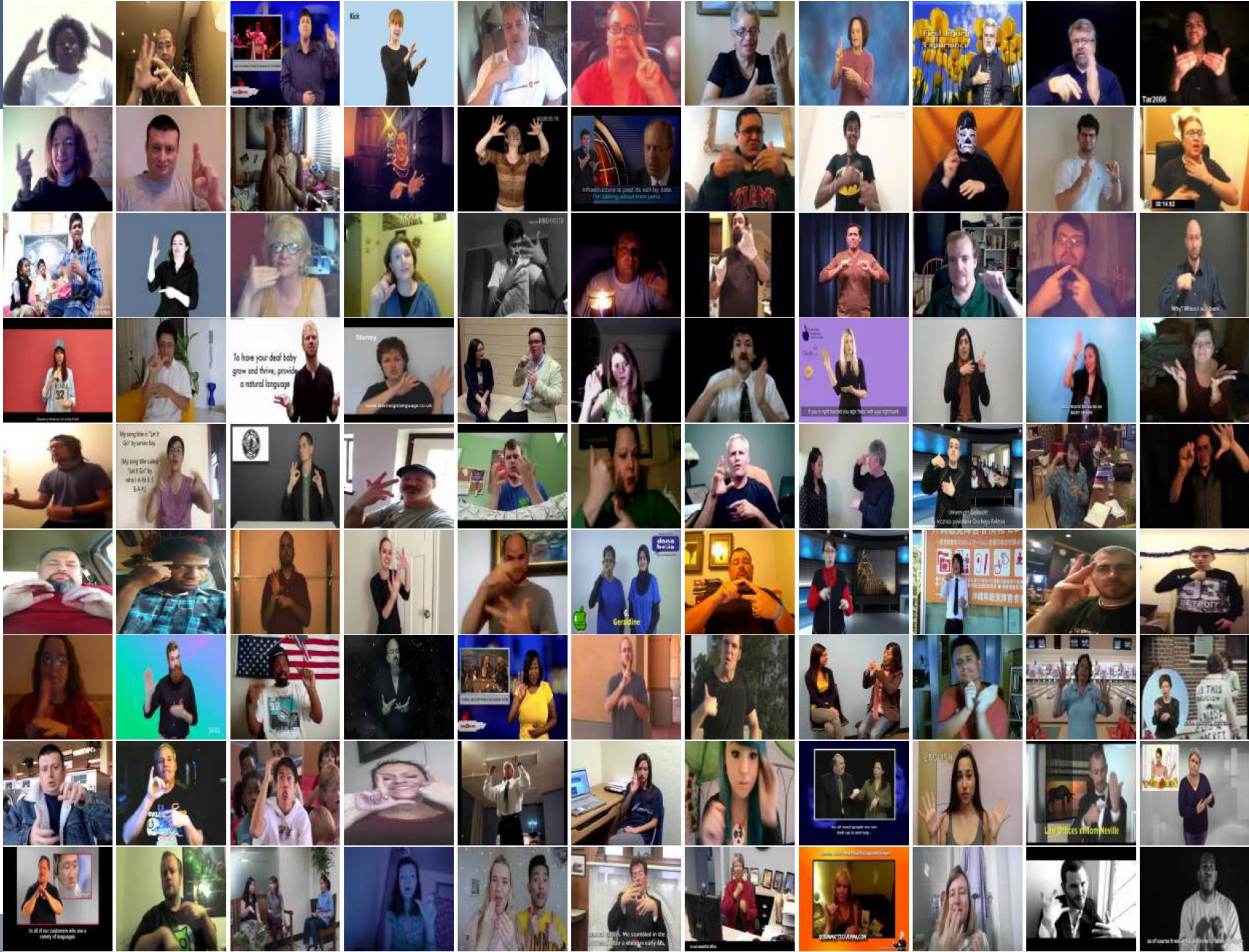
1120 video segments

Groundtruth data

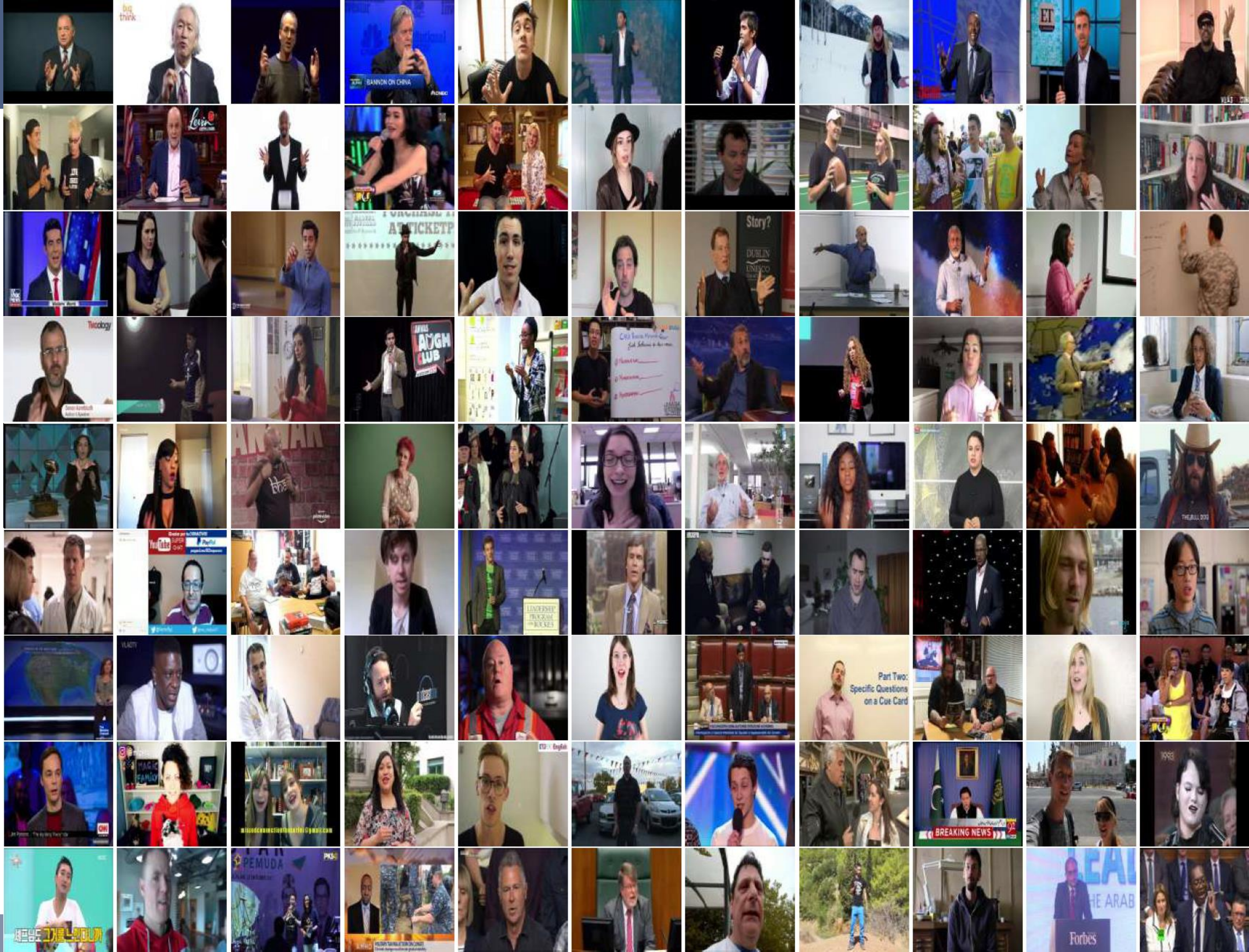
Manual groundtruthing of video frames

class	total frames	%
signing	535,105	43%
speaking	511,446	42%
other	186,007	15%
	1,232,558	100%

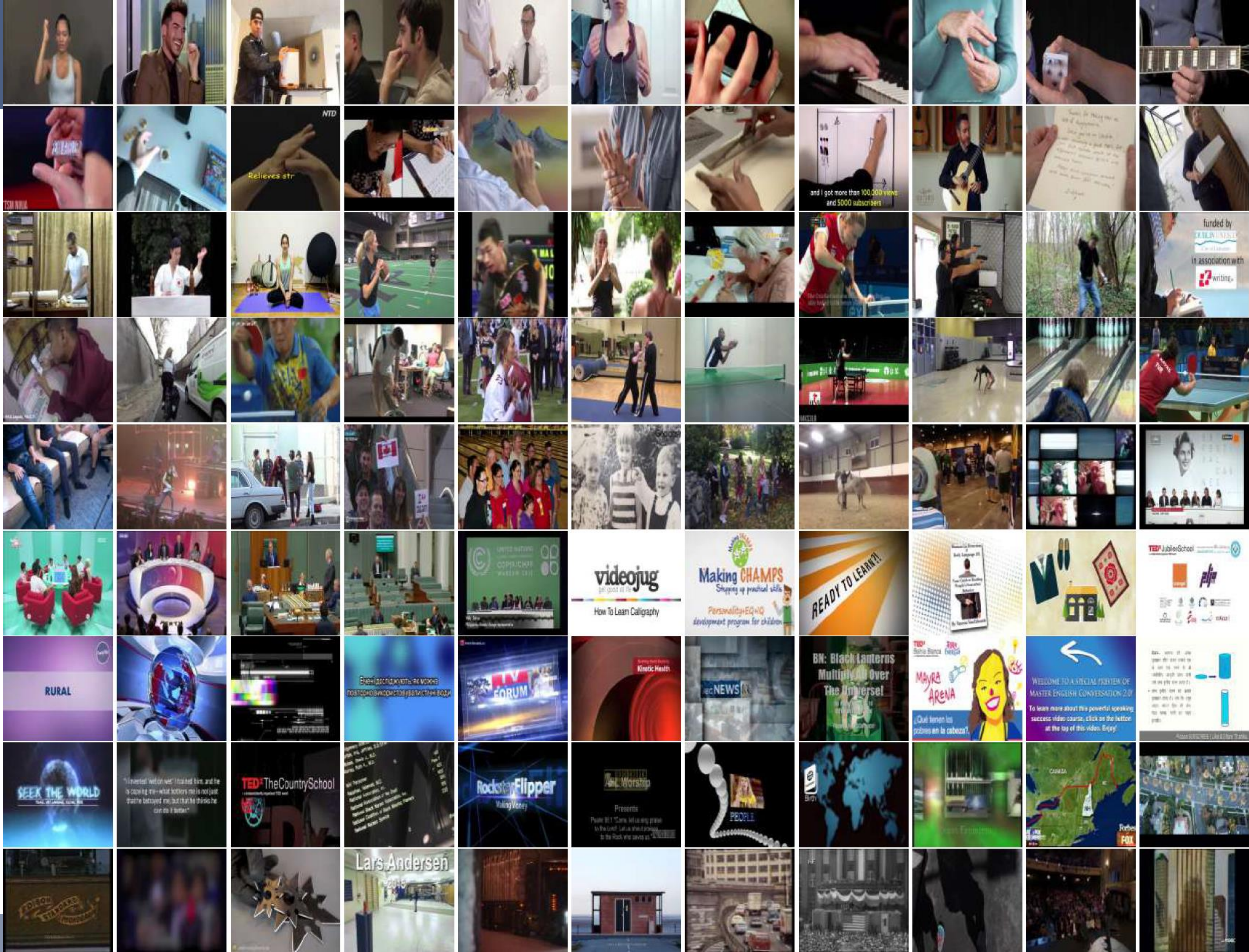
Example frames from class `signing`



Example frames from class `speaking`



Example frames from class `other`

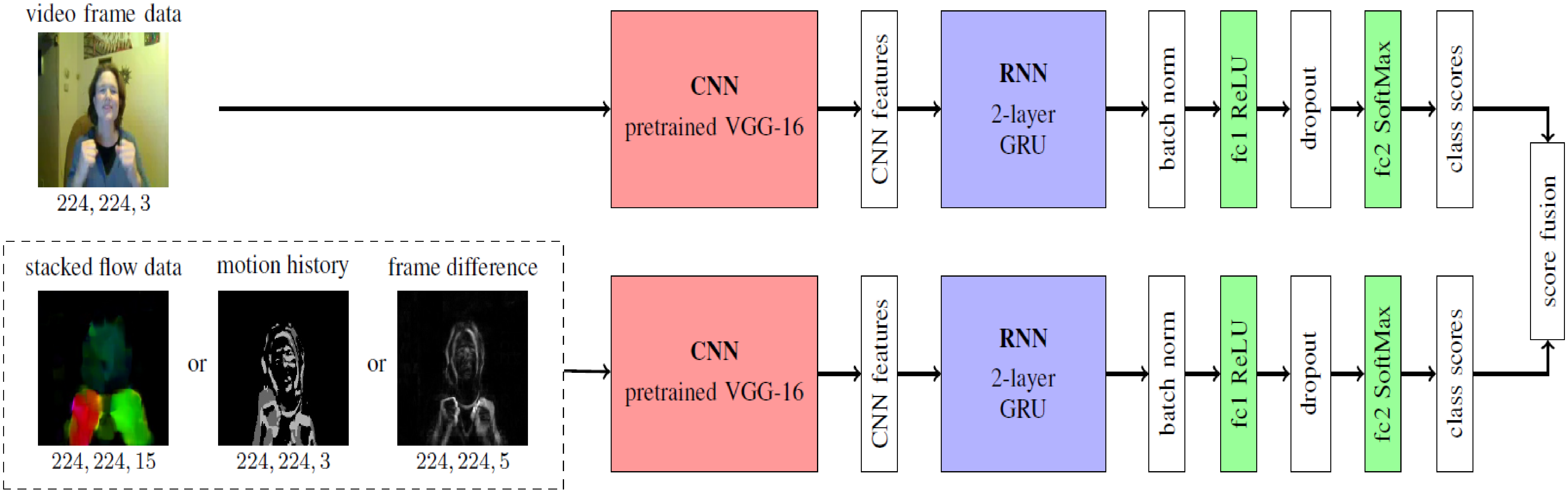


Sign Language Detection – proposed approach

- **Automated extraction of features** using a **Convolutional Neural Network (CNN)**
- Combining both **visual features** and **motion features**
- Use of a **Recurrent Neural Network (RNN)** to handle the dynamic temporal patterns present in sign languages

Proposed architecture

- Two-stream approach (Simonyan, 2014)

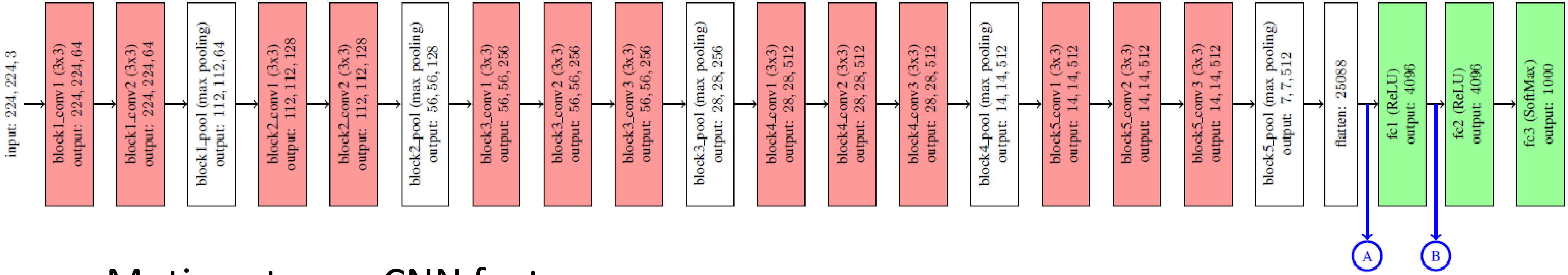


Proposed approach

- Motion stream:
  - Performance vs. computational efficiency
- Investigated:
  - Optical Flow
  - Motion History Images (MHI)
  - Multi-frame differencing

CNN features

- CNN streams
  - Pre-trained VGG-16 (Simonyan 2014)
  - CNN features:
    - Ⓐ 7 x 7 x 512 = 25088 feature map from ‘block5_conv3’ layer
    - Ⓑ 4096 feature map from ‘fc1’ layer

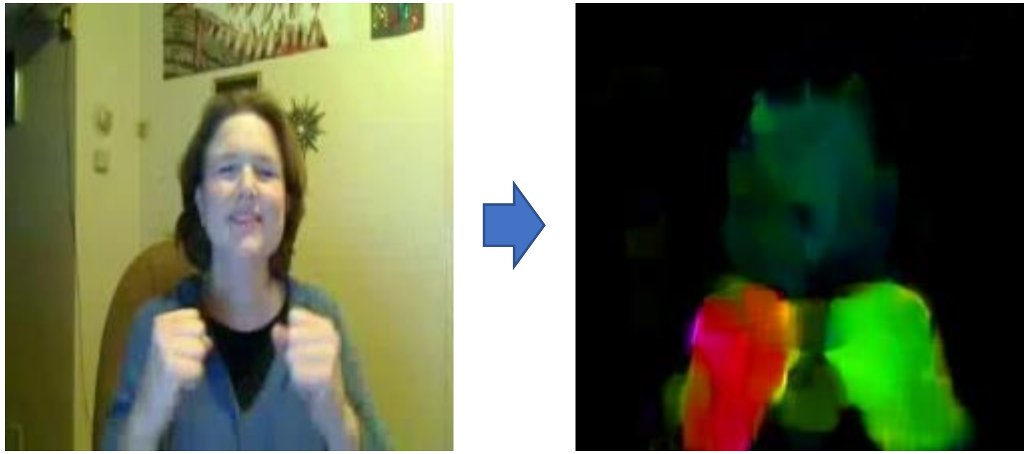


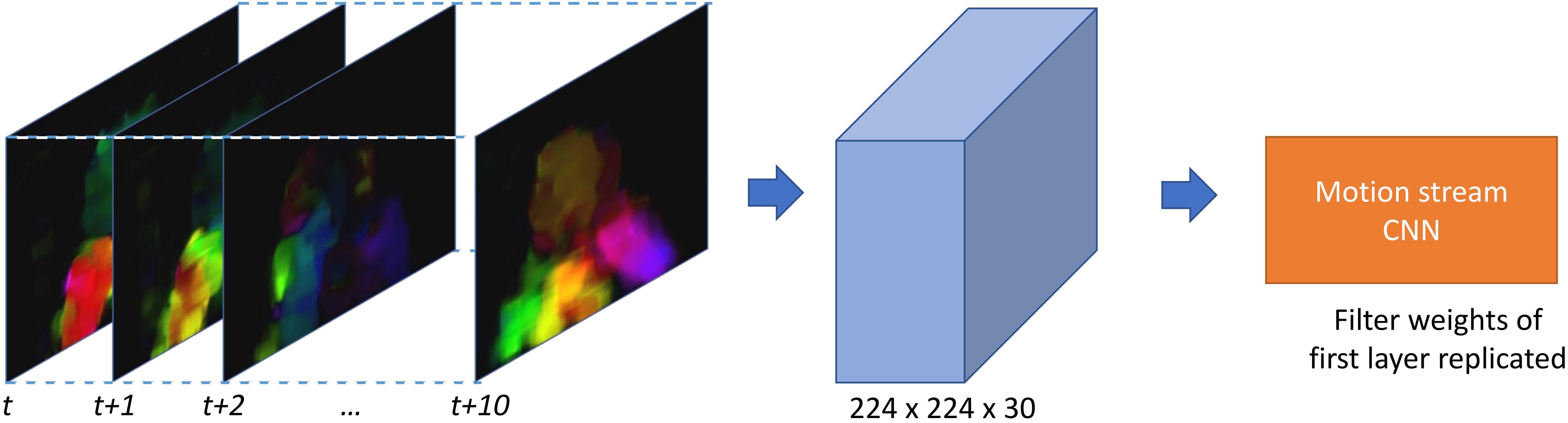
input: 224, 224, 3
block1_conv1 (3x3) output: 224, 224, 64
block1_conv2 (3x3) output: 224, 224, 64
block1_pool (max pooling) output: 112, 112, 64
block2_conv1 (3x3) output: 112, 112, 128
block2_conv2 (3x3) output: 112, 112, 128
block2_pool (max pooling) output: 56, 56, 128
block3_conv1 (3x3) output: 56, 56, 256
block3_conv2 (3x3) output: 56, 56, 256
block3_conv3 (3x3) output: 56, 56, 256
block3_pool (max pooling) output: 28, 28, 256
block4_conv1 (3x3) output: 28, 28, 512
block4_conv2 (3x3) output: 28, 28, 512
block4_conv3 (3x3) output: 28, 28, 512
block4_pool (max pooling) output: 14, 14, 512
block5_conv1 (3x3) output: 14, 14, 512
block5_conv2 (3x3) output: 14, 14, 512
block5_conv3 (3x3) output: 14, 14, 512
block5_pool (max pooling) output: 7, 7, 512
flatten: 25088
fc1 (ReLU) output: 4096
fc2 (ReLU) output: 4096
fc3 (SoftMax) output: 1000

- Motion stream CNN features:
  - We use transfer learning from a distant task (unrelated data) vs. Training from scratch (Yosinski et al., 2014)
- No fine-tuning of VGG-16 layers

Motion data

- Optical flow
  - Dense optical flow (Farnebäck’s algorithm)
  - Encoded as RGB
    - Flow vector magnitude → luminance channel
    - Flow vector angle → chrominance channels



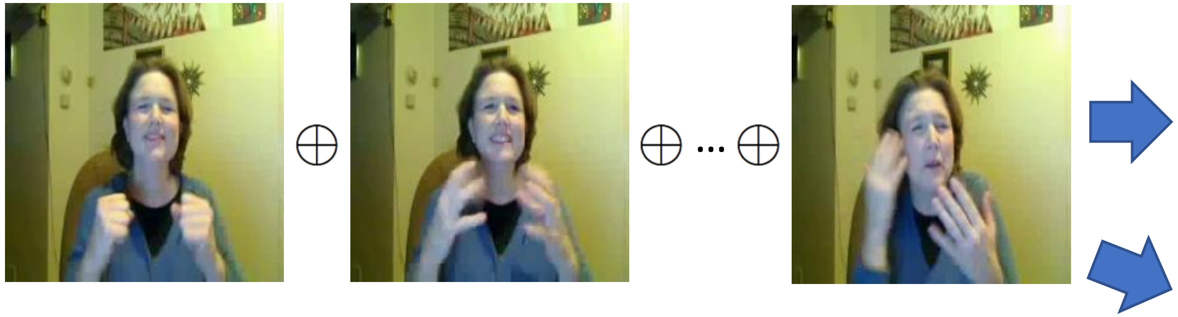


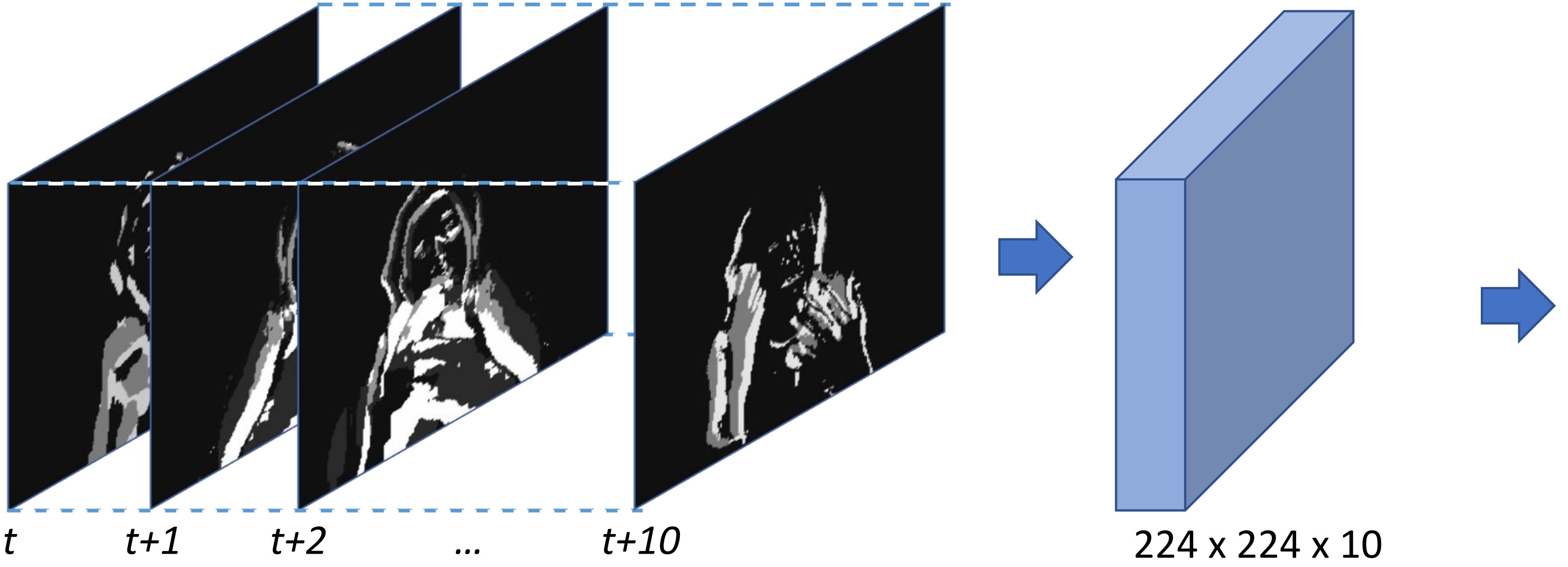
Motion stream CNN

Filter weights of first layer replicated

Motion data

- Multi-frame differencing
- Motion History Images
  - 5 frame temporal window





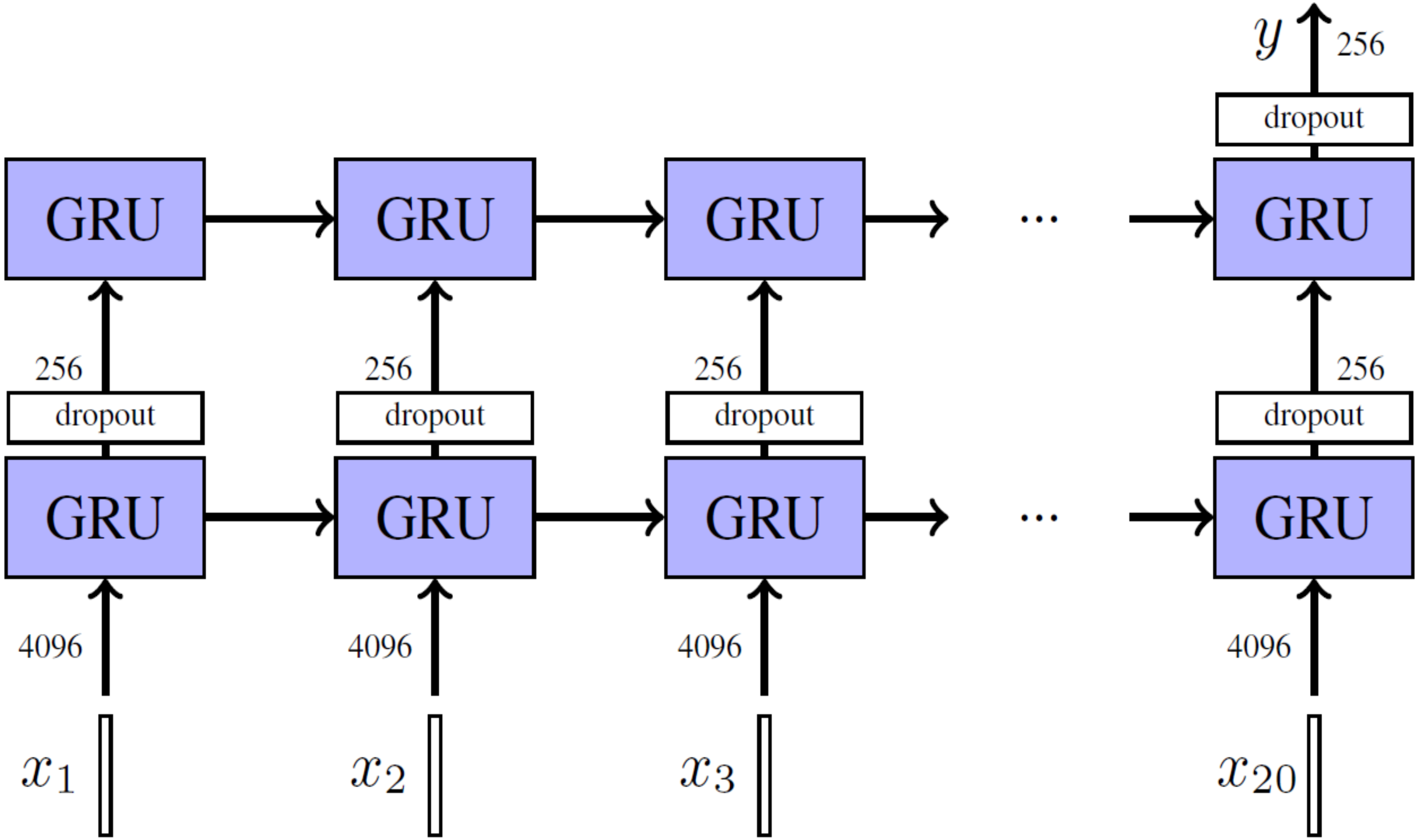
Motion stream CNN

Filter weights of first layer averaged to 1-channel, then replicated 10 times

RNN

- Various RNN options: LSTMs and GRUs
- Stacked RNNs

- 2-layer GRU
- 256 hidden units
- 20 timesteps
    - (2.5 seconds with a 5Hz sampling rate)



RNN training

- Stratified partitioning of the dataset
- video frames from a single video appear in only one partition
- 5 fold cross-validation

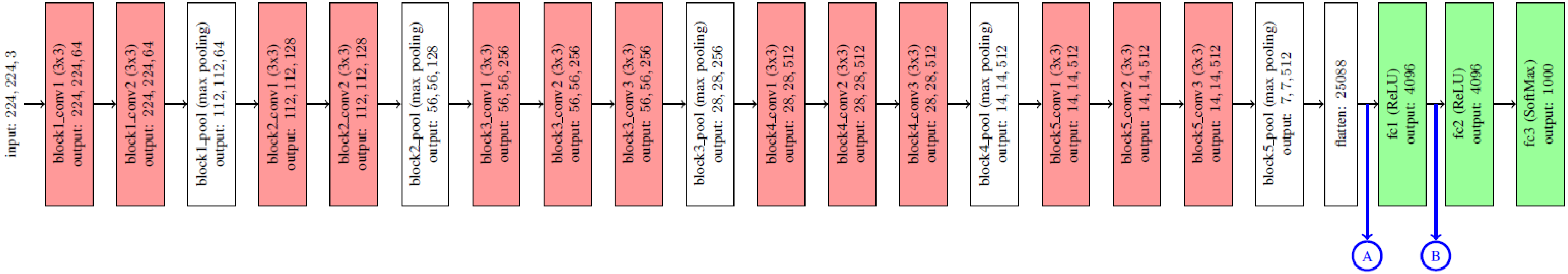
- Mini-batch stochastic gradient descent (SGD)
- Adam optimizer
- Training for 500 epochs, with early stopping (validation cross-entropy loss)

- Training strategy:
    - Initial mini-batch size of 32, learning rate of 0.001
    - Reduce learning rate when validation loss stops improving for the current combination of mini-batch size and learning rate
    - Increase mini-batch size when no more change in validation loss is observed for the given mini-batch size despite the changes to the learning rate

Results

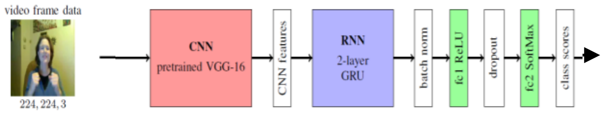
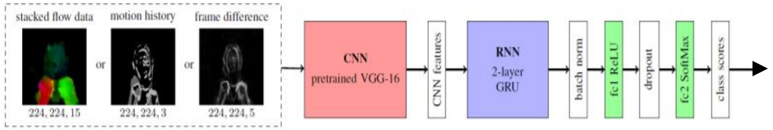
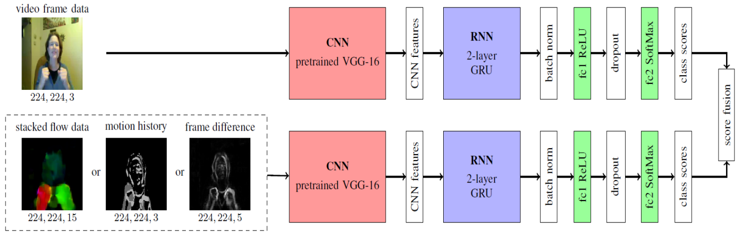
- Evaluation of different feature maps from the CNN network

CNN layer	Feature size	Loss (validation set) ↓
VGG-16 block5_conv3 Ⓐ	25088	0.6681
VGG-16 fc1 Ⓑ	4096	**0.5037**



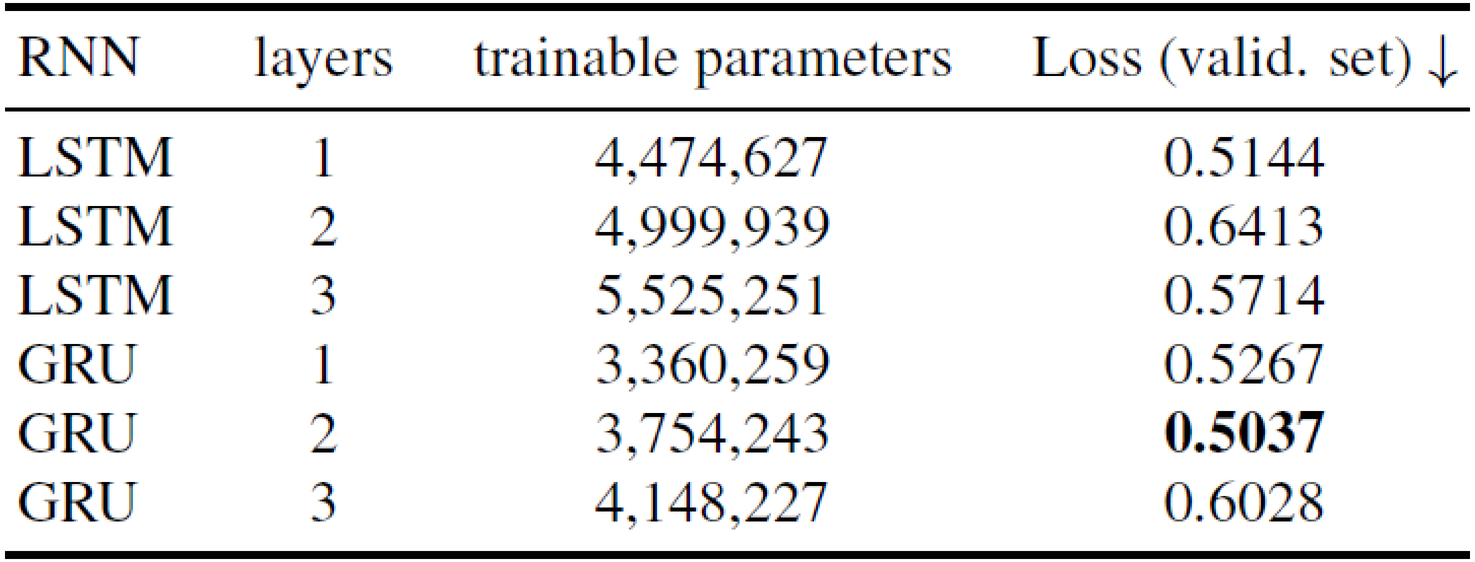
Results

- Evaluation of the individual performance of the different streams, and when fusing both the motion and RGB streams together

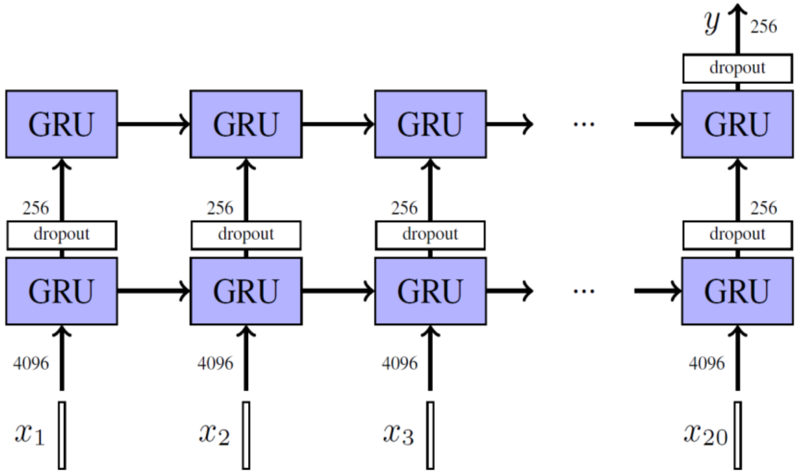
Modality	Loss ↓	Accuracy ↑	Time (ms) ↓	
RGB stream only	0.5128	85.01%	–	
optical flow only	**0.5387**	83.12%	57.8	
MHI only	0.5445	83.67%	17.1	
multi-frame diff. only	0.5738	84.08%	**9.7**	
RGB stream + optical flow stream	–	**87.67%**	–	
RGB stream + MHI stream	–	87.60%	–	
RGB stream + frame diff. stream	–	85.61%	–	

Results

- Evaluation of different RNN architectures

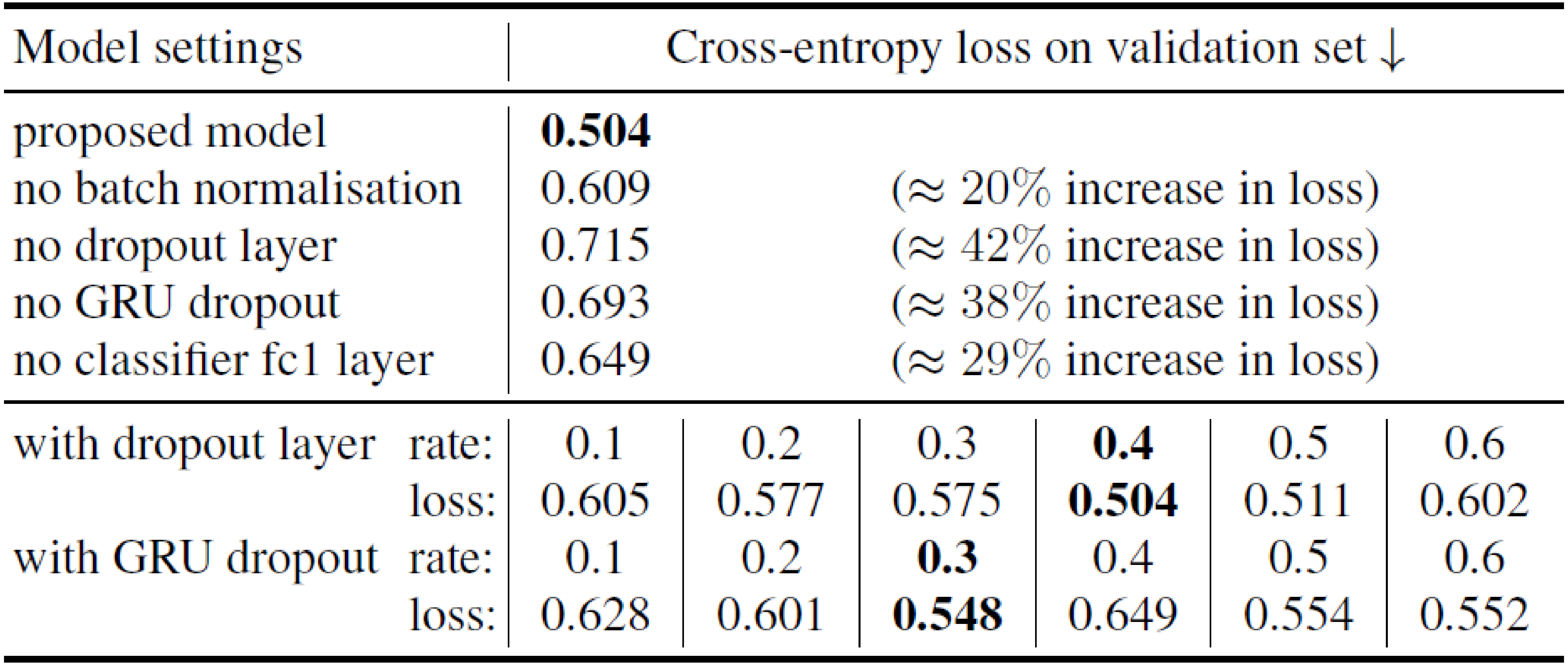


RNN	layers	trainable parameters	Loss (valid. set) ↓
LSTM	1	4,474,627	0.5144
LSTM	2	4,999,939	0.6413
LSTM	3	5,525,251	0.5714
GRU	1	3,360,259	0.5267
GRU	2	3,754,243	**0.5037**
GRU	3	4,148,227	0.6028



Results

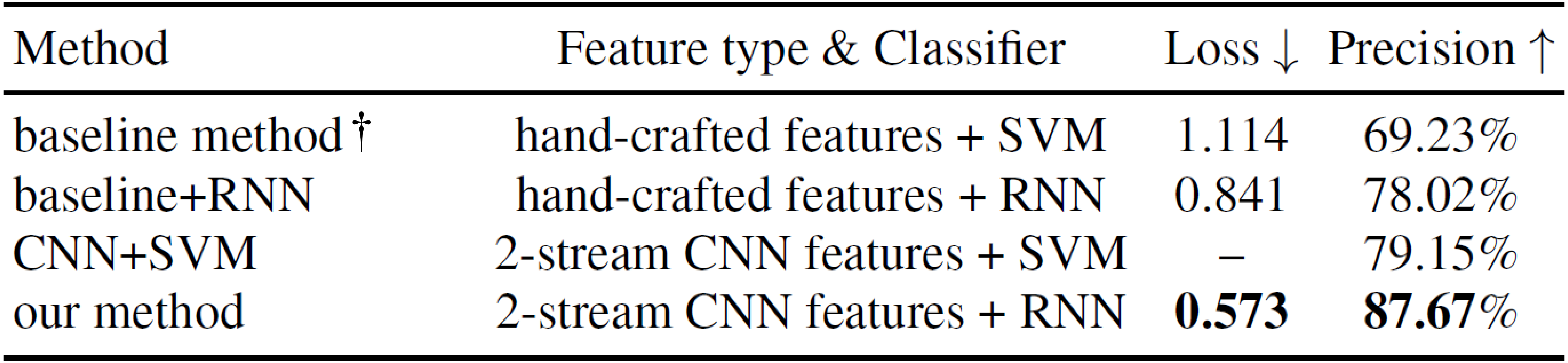
- Ablation studies on the proposed RNN network



Model settings		Cross-entropy loss on validation set ↓					
proposed model		**0.504**					
no batch normalisation		0.609		(≈ 20% increase in loss)			
no dropout layer		0.715		(≈ 42% increase in loss)			
no GRU dropout		0.693		(≈ 38% increase in loss)			
no classifier fc1 layer		0.649		(≈ 29% increase in loss)			
with dropout layer	rate:	0.1	0.2	0.3	**0.4**	0.5	0.6
	loss:	0.605	0.577	0.575	**0.504**	0.511	0.602
with GRU dropout	rate:	0.1	0.2	**0.3**	0.4	0.5	0.6
	loss:	0.628	0.601	**0.548**	0.649	0.554	0.552

Results

- Comparison with the state-of-the-art in sign language detection
  - ≈ 18% improvement over baseline
  - ≈ 9% improvement when using an RNN versus SVM

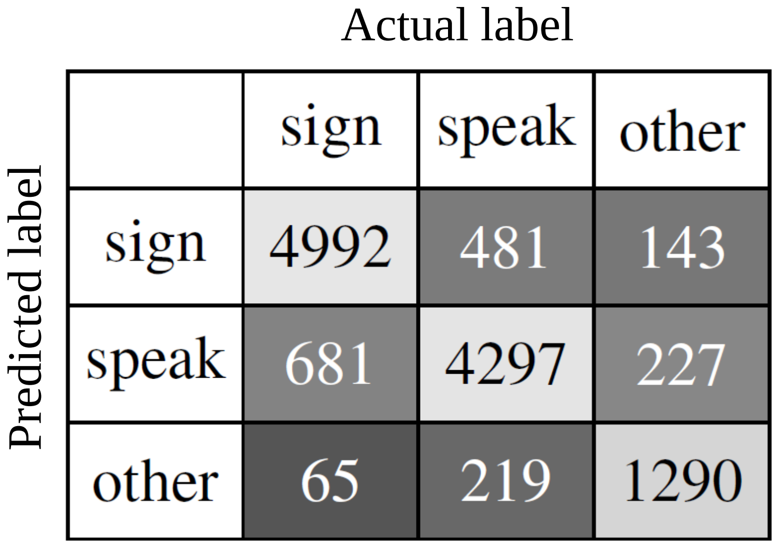


Method	Feature type & Classifier	Loss ↓	Precision ↑
baseline method †	hand-crafted features + SVM	1.114	69.23%
baseline+RNN	hand-crafted features + RNN	0.841	78.02%
CNN+SVM	2-stream CNN features + SVM	–	79.15%
our method	2-stream CNN features + RNN	**0.573**	**87.67**%

† Shipman et al. (JCDL 2015, ACM SigAccess 2017)

Results

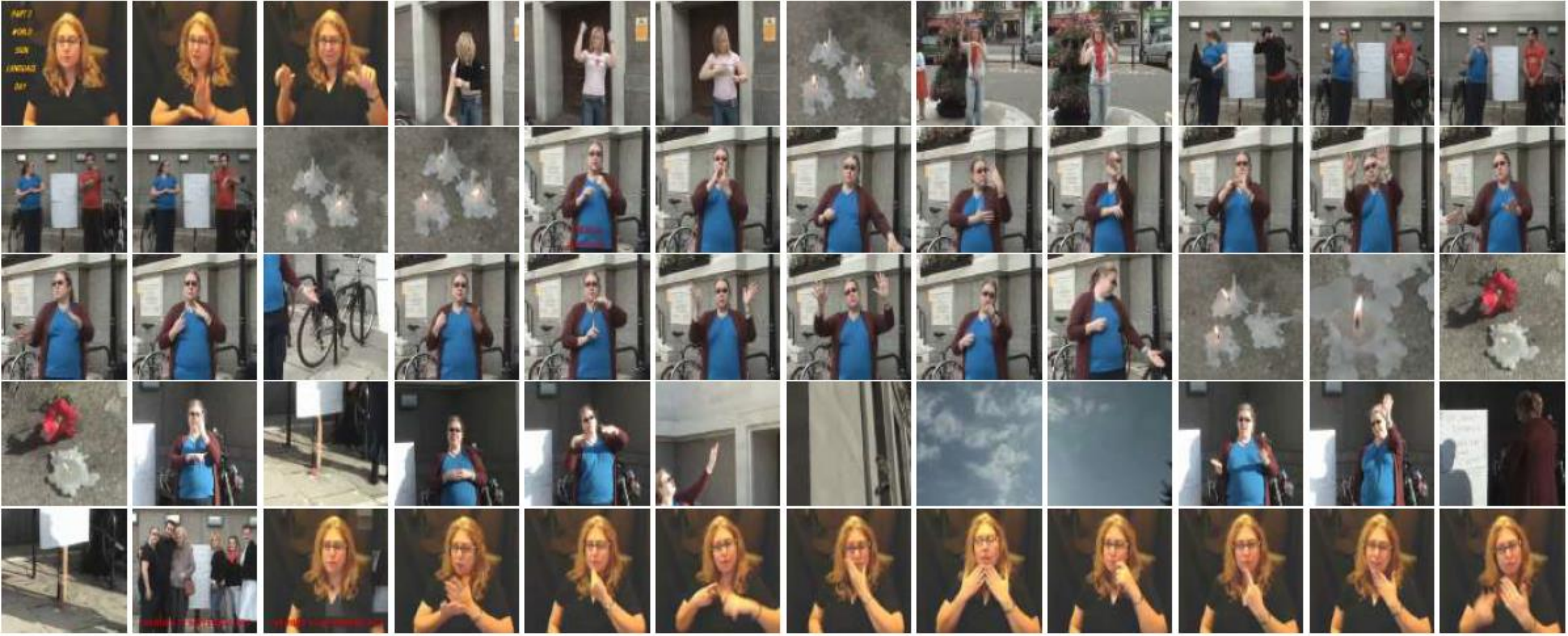
- Confusion matrix



Actual label

Predicted label		sign	speak	other
	sign	4992	481	143
	speak	681	4297	227
	other	65	219	1290

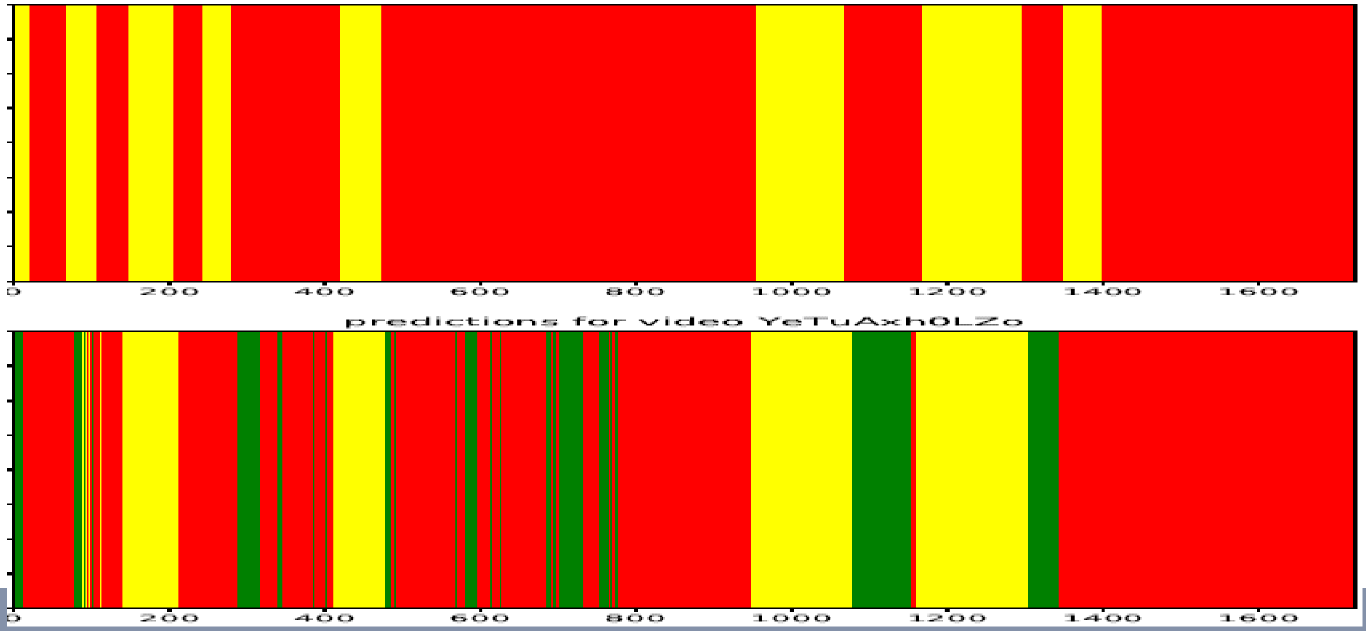
Results

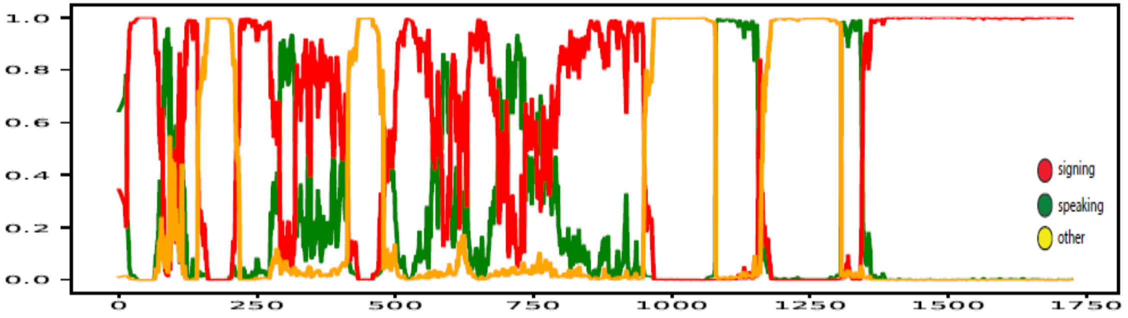


Video “YeTuAxh0LZo”

Sample video frames (every 100)

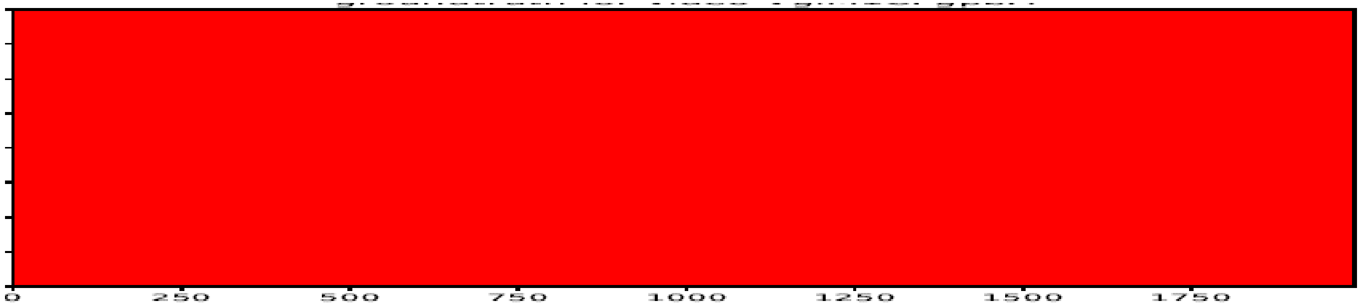
Groundtruth + prediction strips (signing, speaking, other)



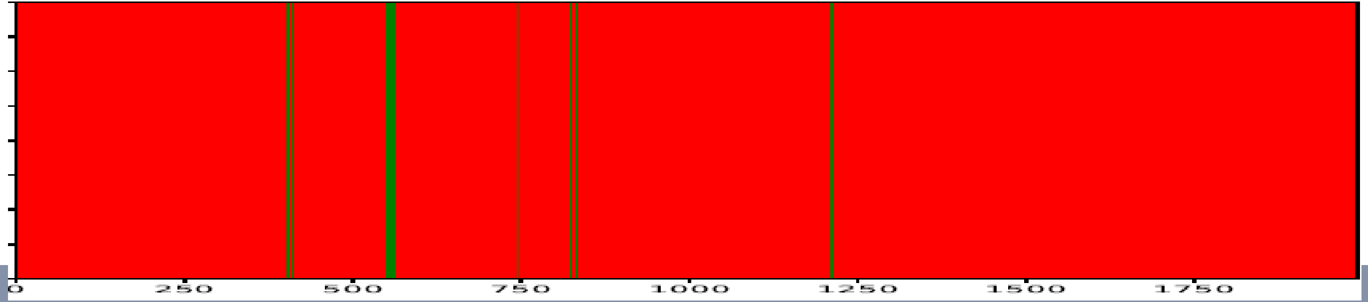


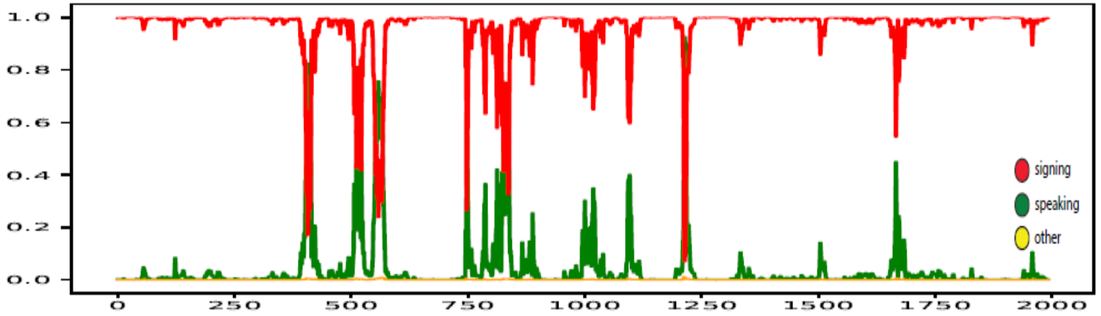
Results

Video “VgiM1CFgpbA”



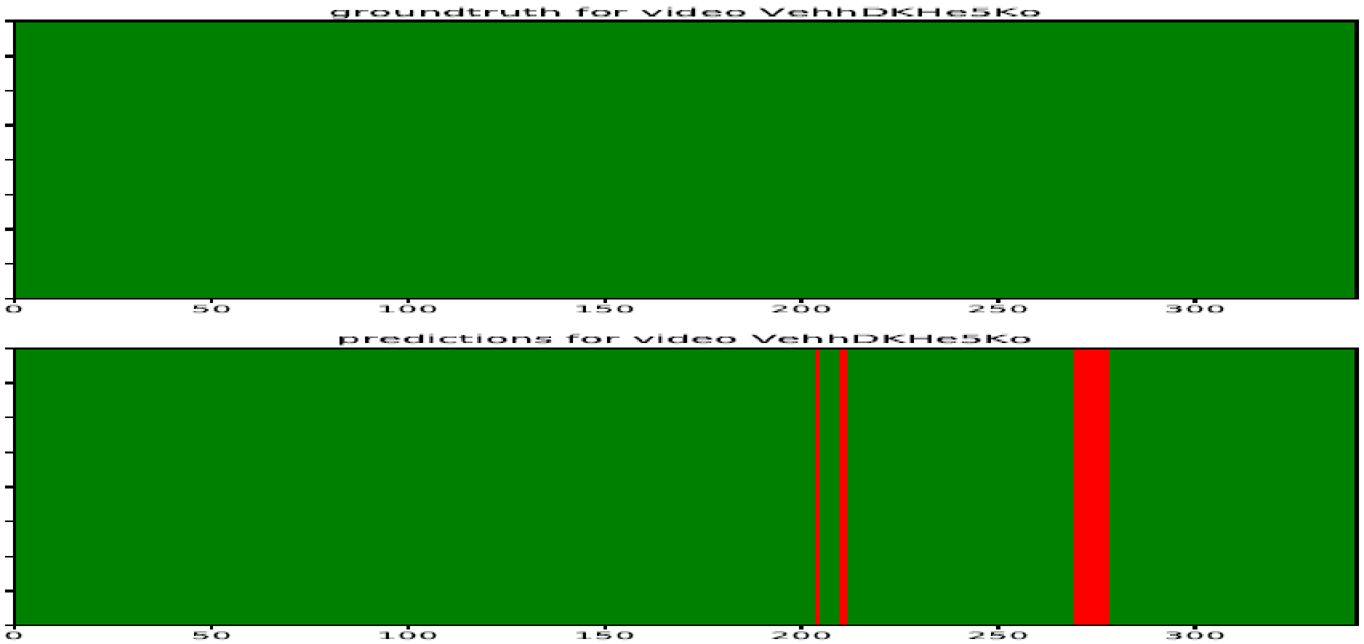
predictions for video VgiM1CFgpbA



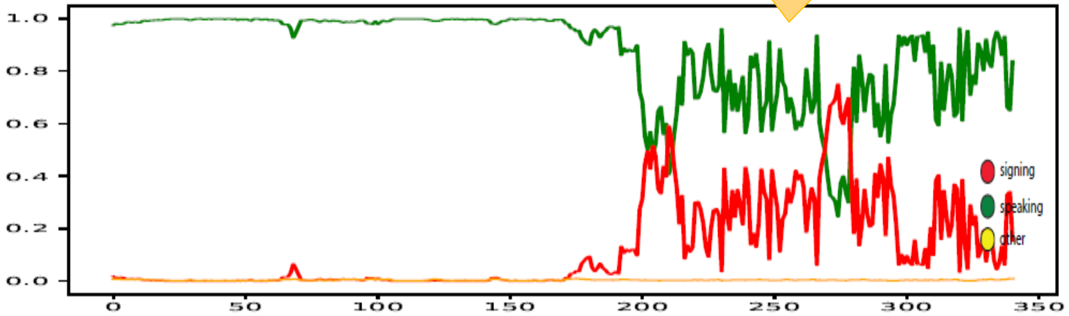


Results

Video “VehhDKHe5Ko”

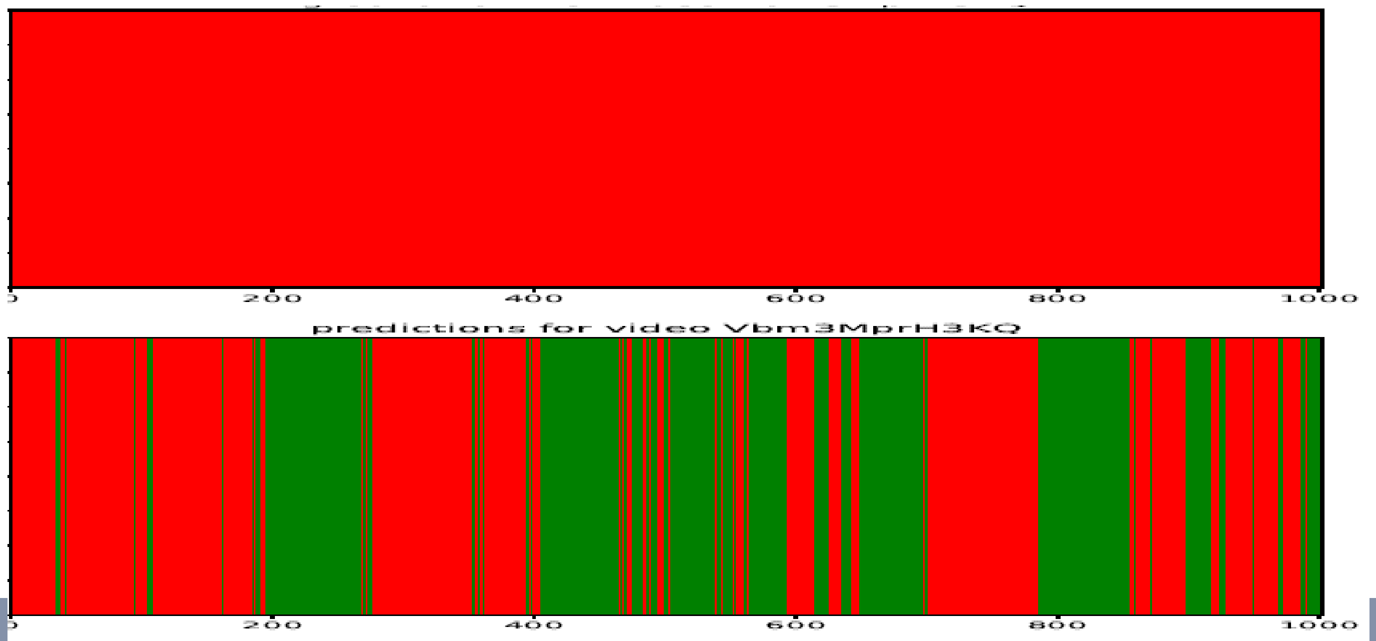


Second part of this video contains hand motion, clapping and singing

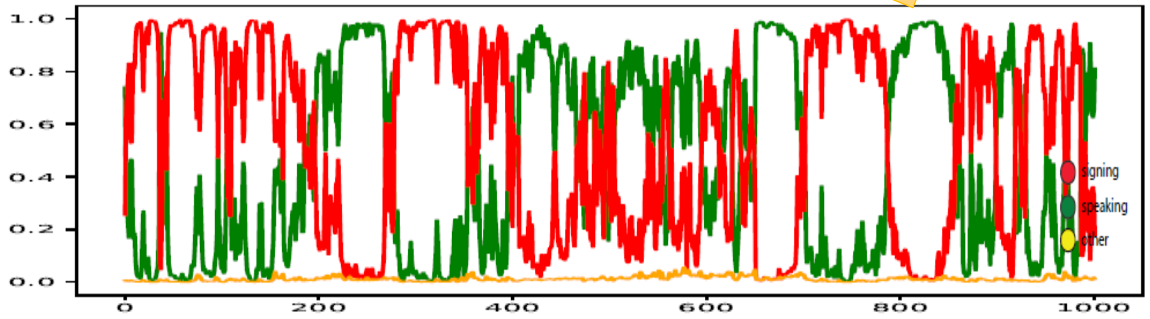


Results

Video “Vbm3MprH3KQ”

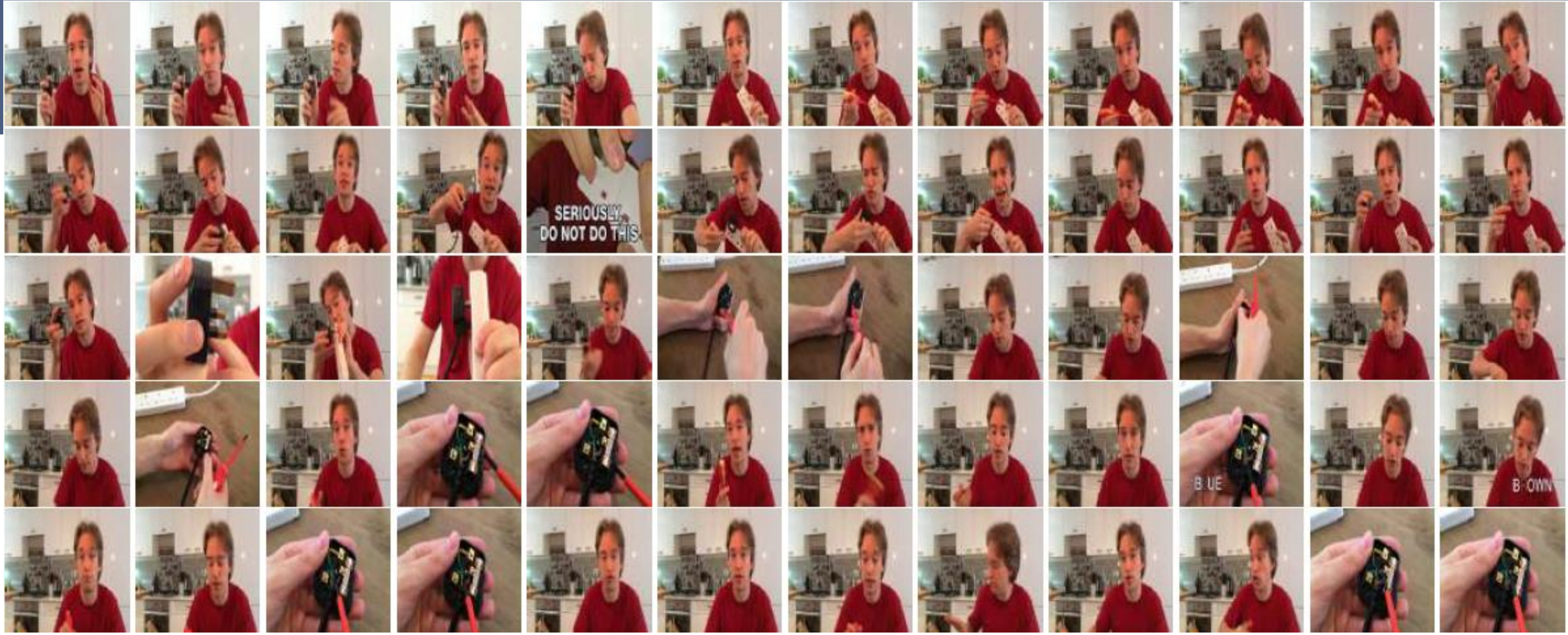


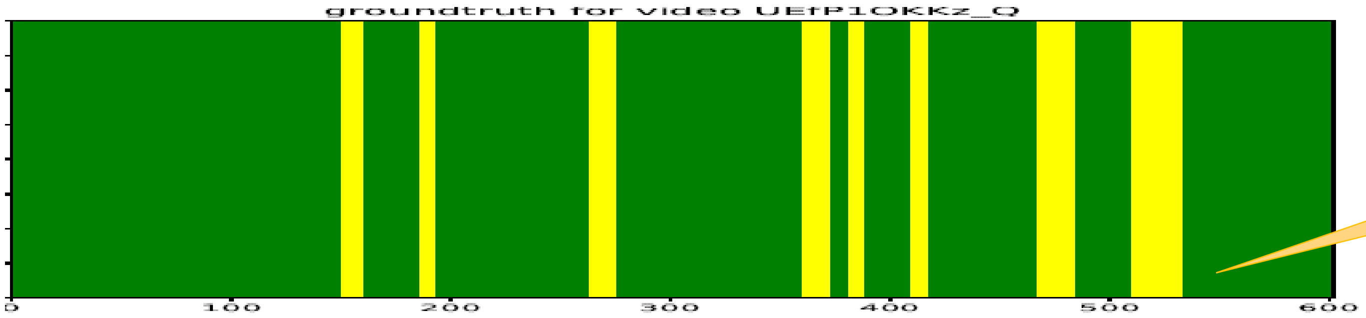
Two signers around a table. Several segments mislabeled as speech

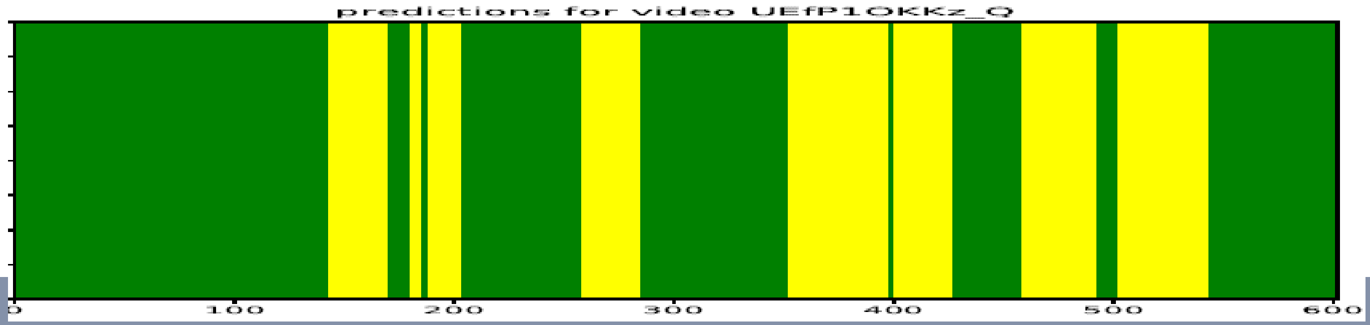


Results

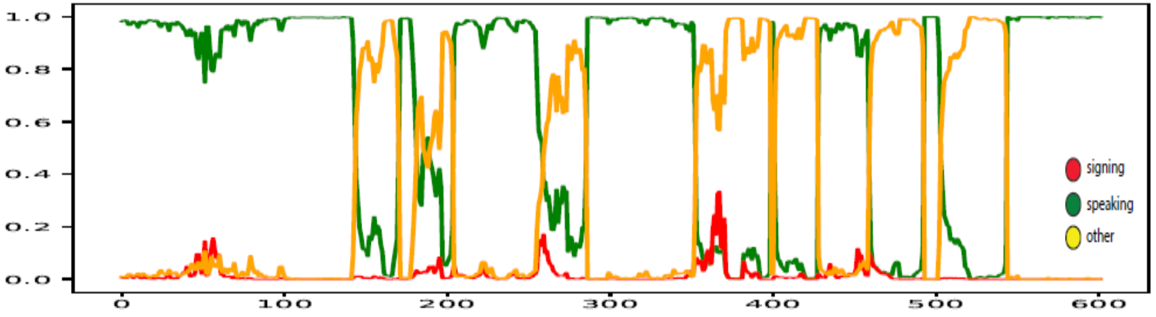
Video “UEfP1OKKz_Q”





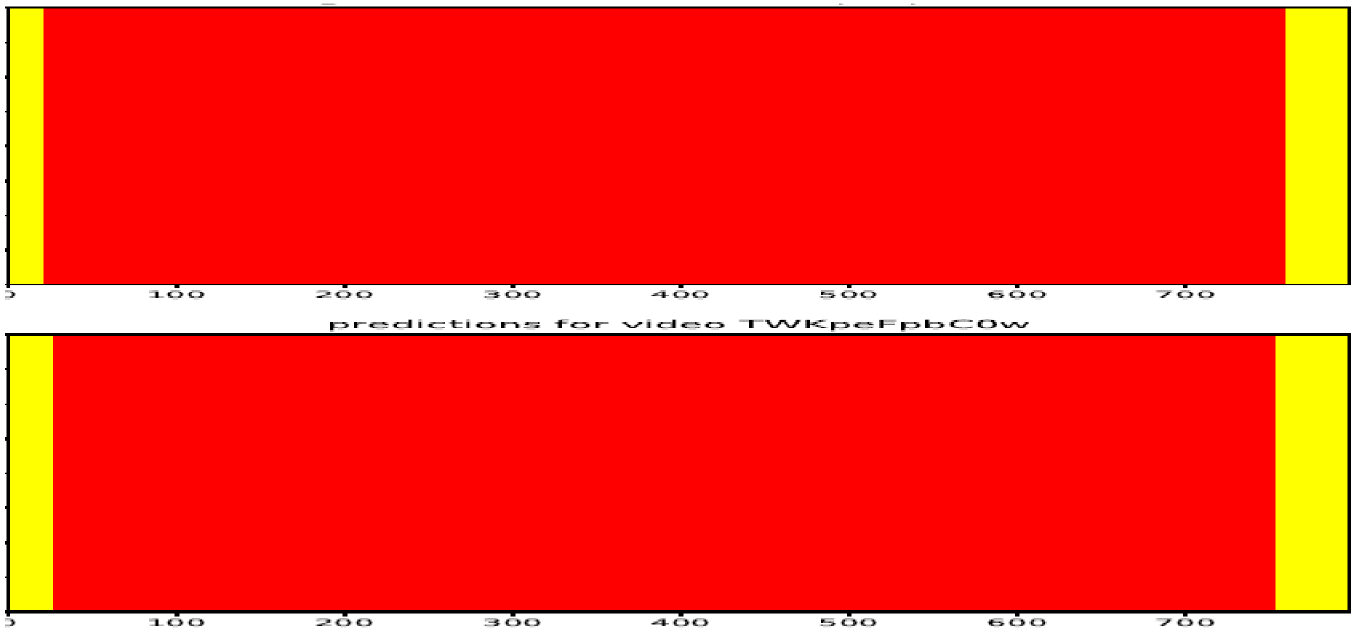


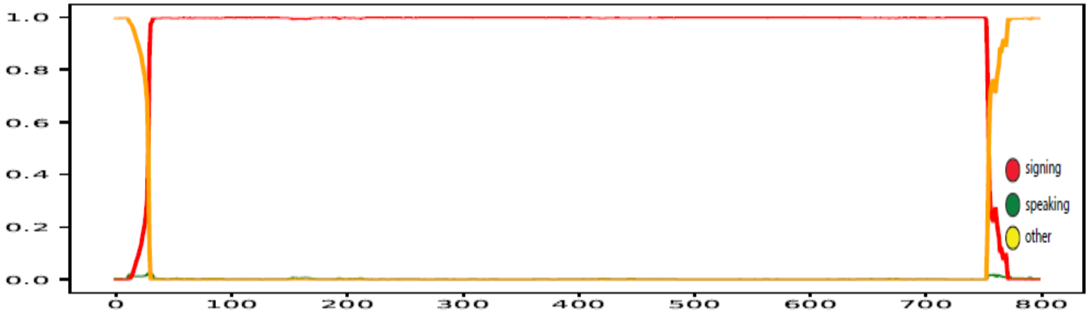
Note the boundary errors between speech and other categories



Results

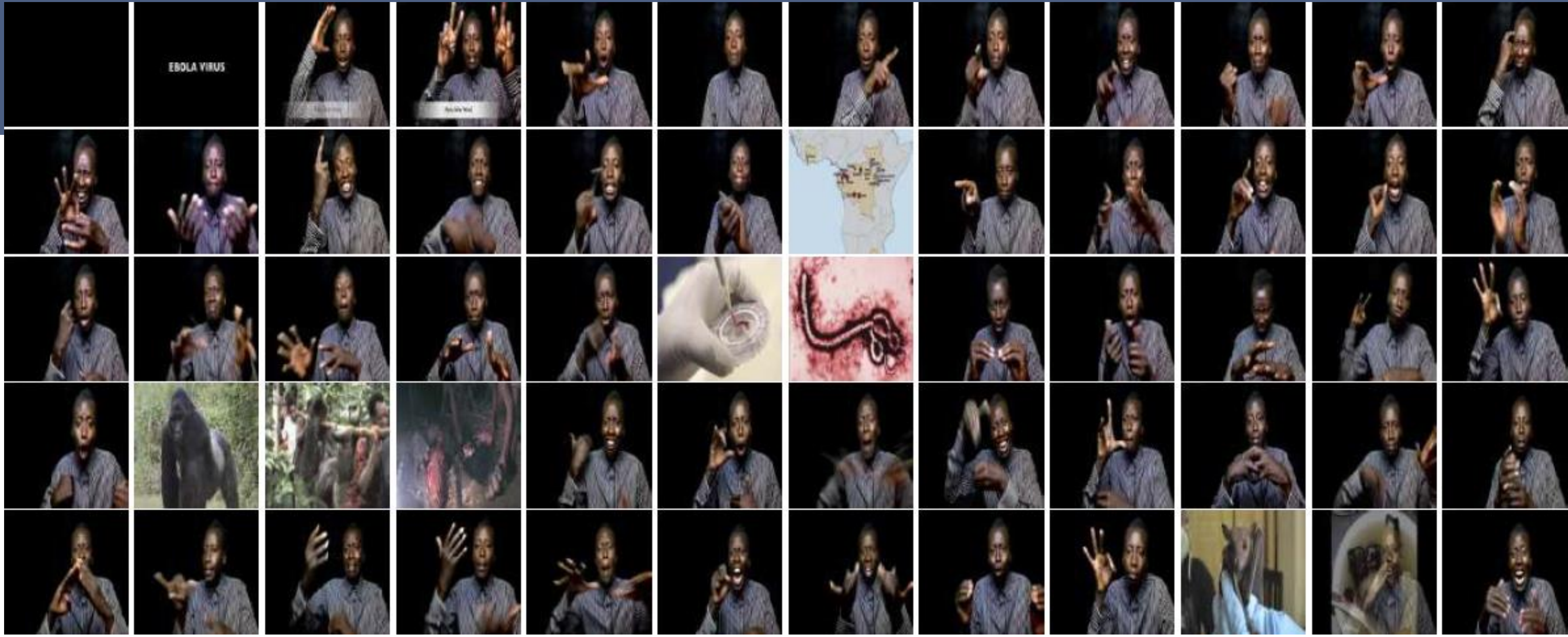
Video “TWKpeFpbC0w”

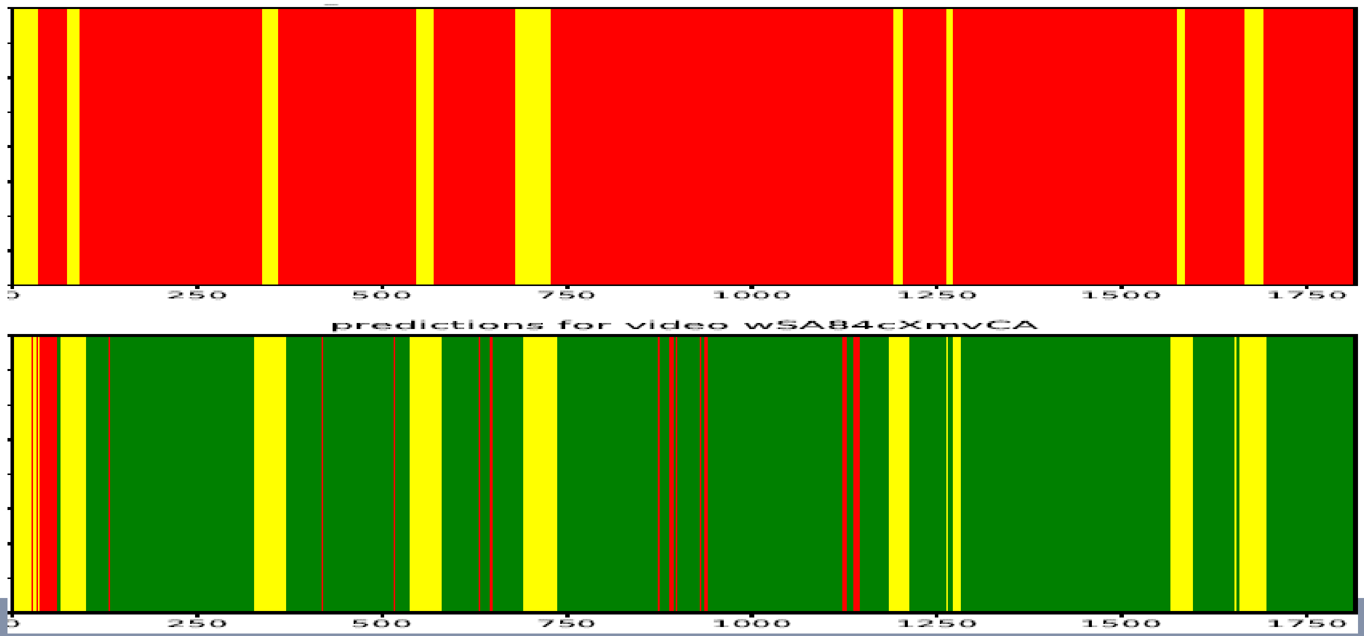




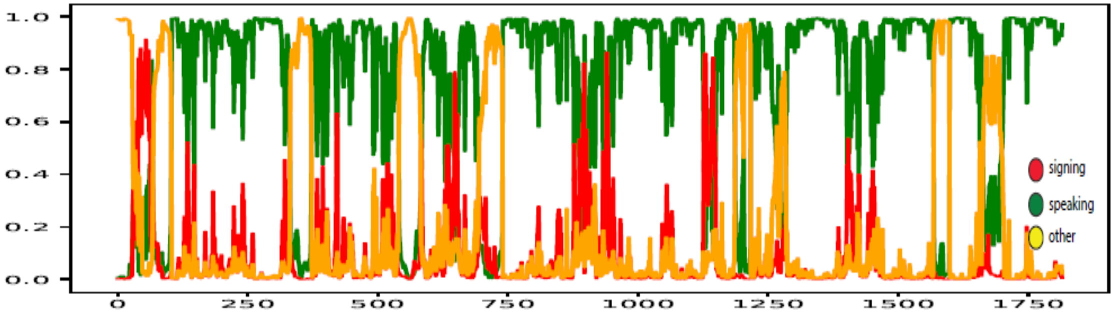
Results

Video “wSA84cXmvCA”





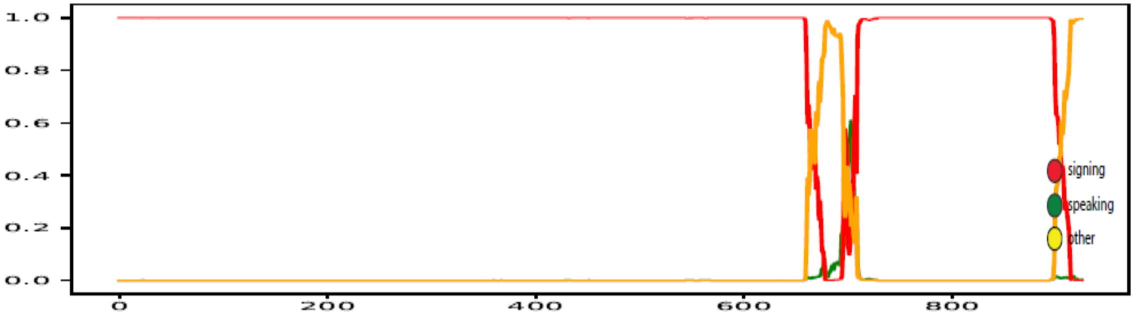
A failure case



Results

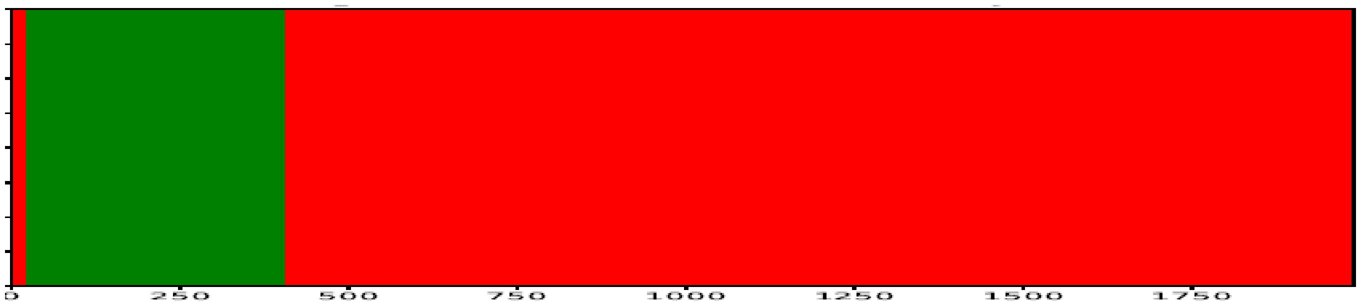
Video “sWxjJaRj1EE”

Correct detection. Signer is wearing a mask

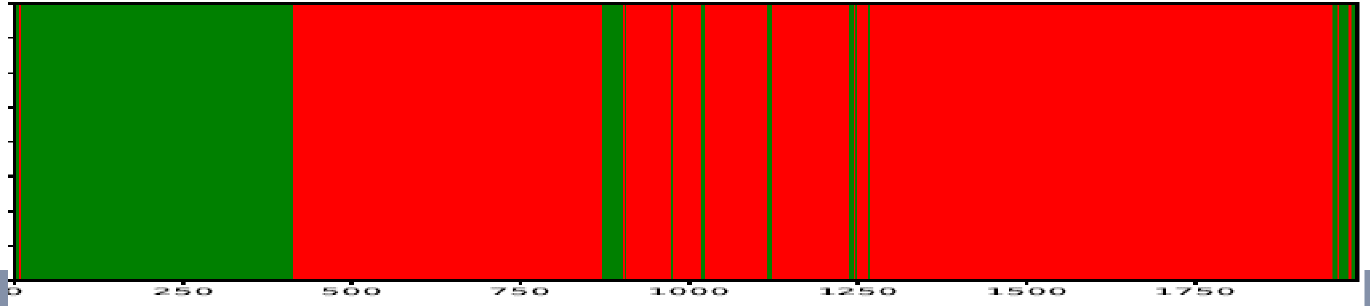


Results

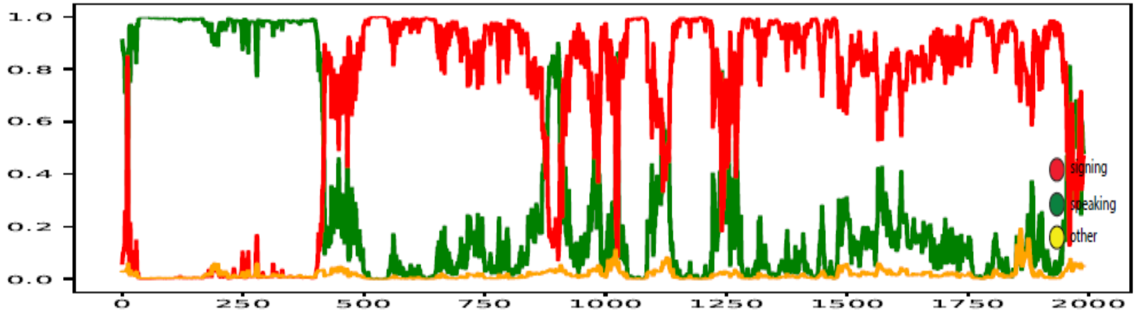
Video “HUMEcnkvhJU”



predictions for video HUMEcnkvhJU



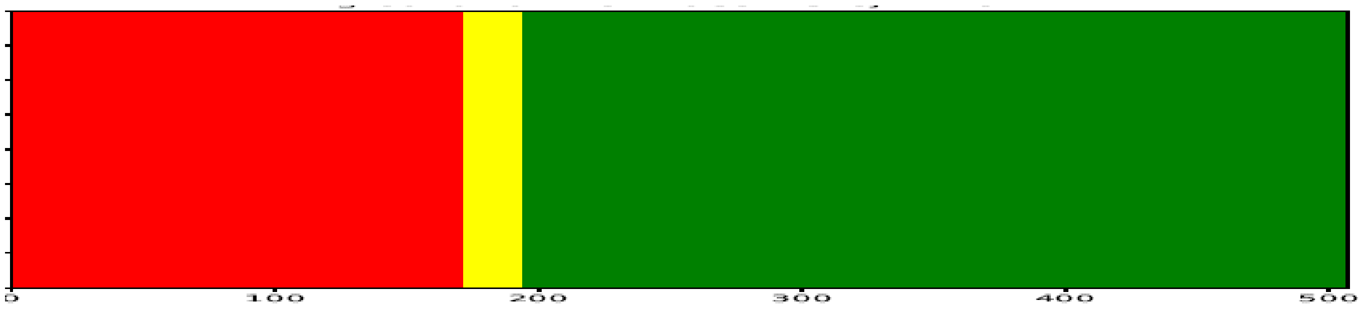
Same person first speaks, then starts signing



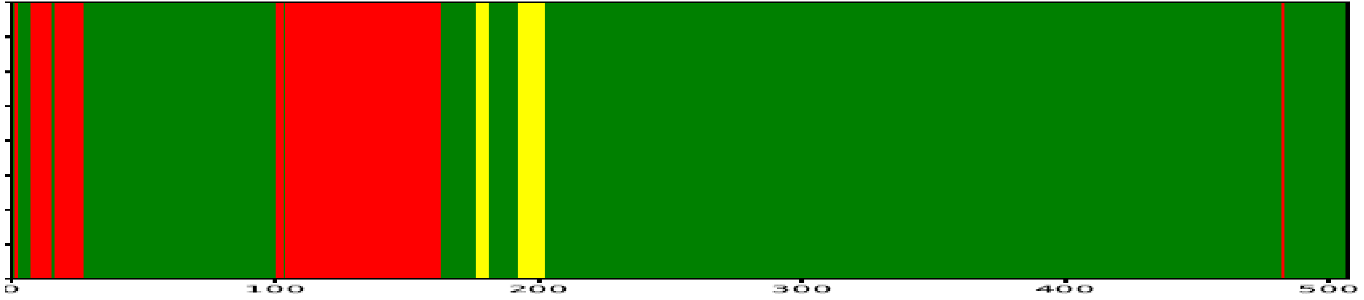
Results

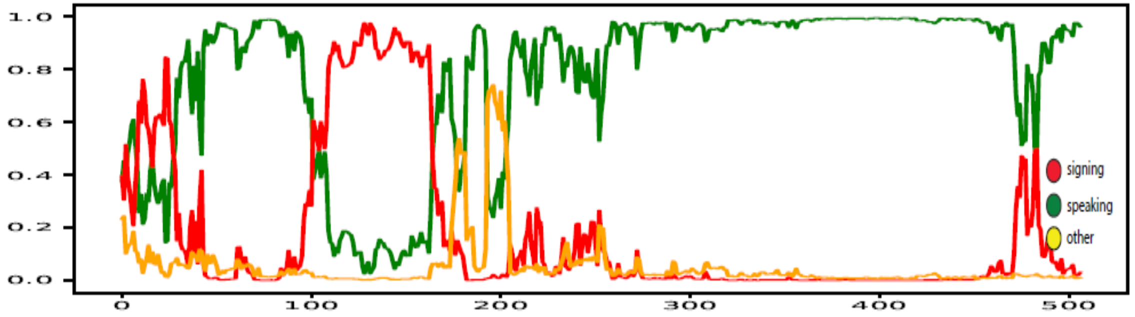
Video “wCFoJviRw9k”

Video contains a person doing both signing and speaking



predictions for video wCFoJviRw9k





Conclusion & Future Work

- A new dataset “Signing in the Wild”
- Successful sign language detection via a two-stream CNN+RNN
  - $\approx 18\%$ improvement over state-of-the-art
- Visual + Motion features are both important for signing

- Future Work:
  - Signer localisation
  - Identification of particular sign languages & sign language constructs
  - Investigating what the CNN+RNN is basing recognition on

- Thank you for your attention