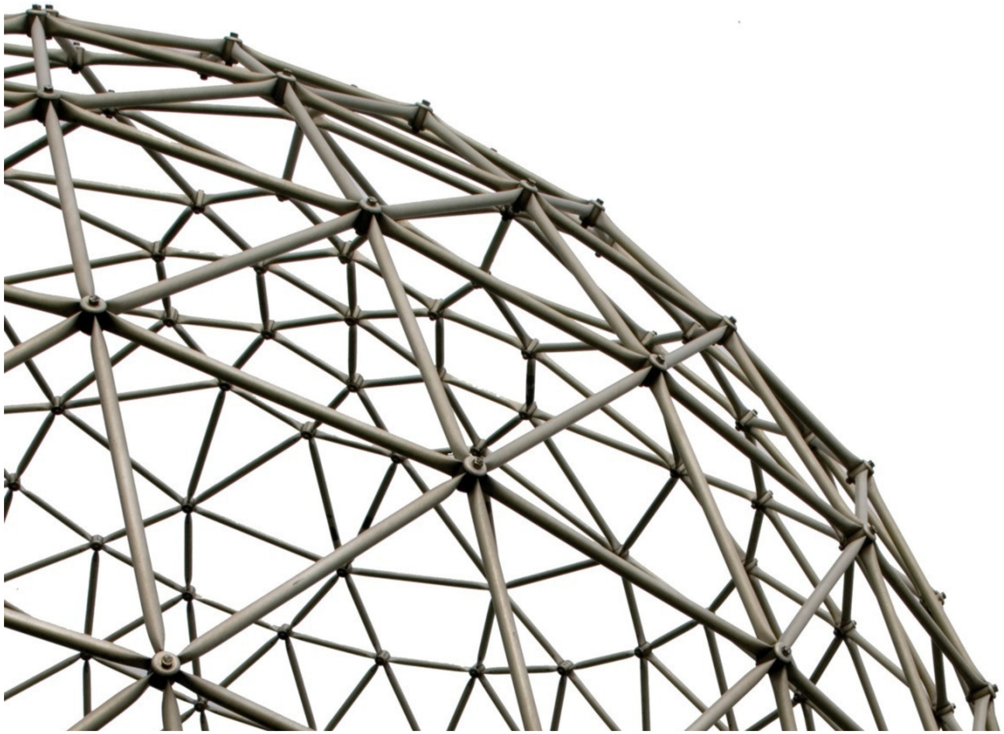
From Activity to Language:

Learning to recognise the meaning of motion

Centre for Vision, Speech and Signal Processing



Prof Rich Bowden
20 June 2011

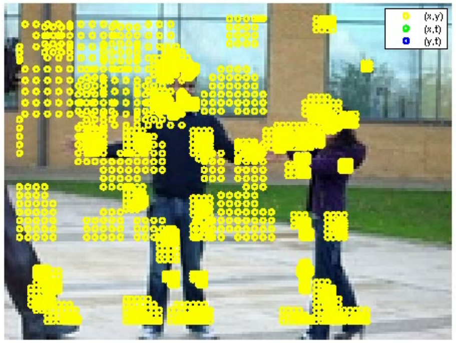
Overview

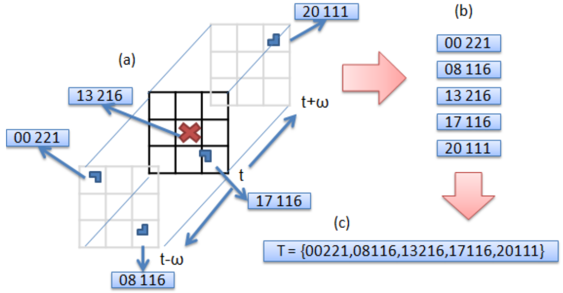
- Talk is about recognising spatio temporal patterns
- Activity Recognition
  - Holistic features
  - Weakly supervised learning
- Sign Language Recognition
  - Using weak supervision
  - Using linguistics
  - EU Project Dicta-Sign
- Facial Feature tracking
  - Lip motion
  - Non manual features

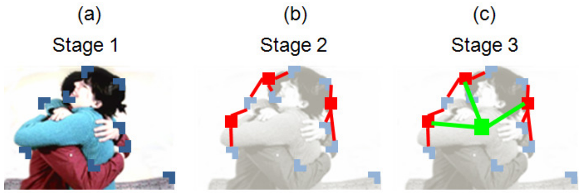
Activity Recognition

Action/Activity Recognition

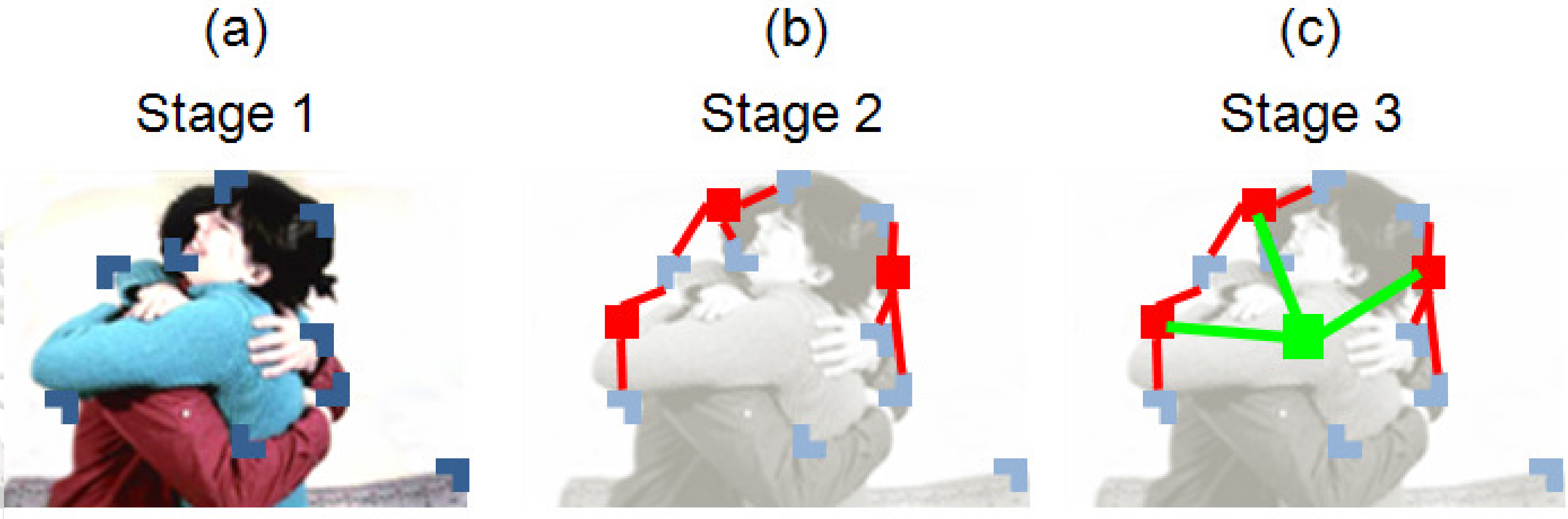
- Densely detect corners
  - (x,y), (x,t), (y,t)
  - Provides both spatial and temporal information
- Spatially encode local neighbourhood
  - Quantise corner types
  - Encode local spatio-temporal relationship
- Apply data mining
  - Find frequently reoccurring feature combinations using the association rule mining e.g Apriori algorithm
- Repeat process hierarchically





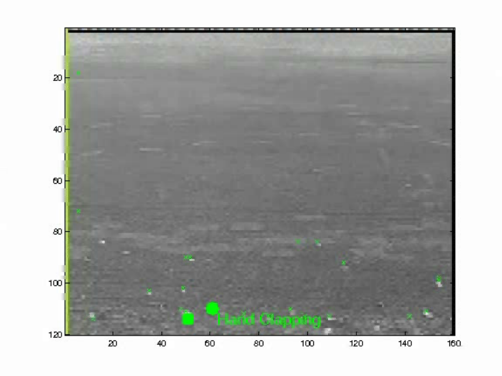


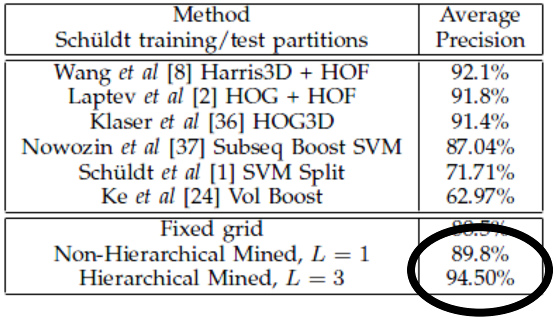
Action/Activity Recognition



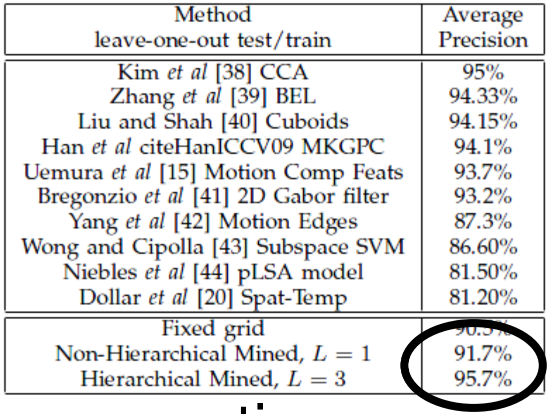
KTH Action Recognition

- Classifier is pixel based frame wise voting scheme
- KTH Dataset 94.5%(95.7%) 24fps





Method Schüldt training/test partitions	Average Precision
Wang *et al* [8] Harris3D + HOF	92.1%
Laptev *et al* [2] HOG + HOF	91.8%
Klaser *et al* [36] HOG3D	91.4%
Nowozin *et al* [37] Subseq Boost SVM	87.04%
Schüldt *et al* [1] SVM Split	71.71%
Ke *et al* [24] Vol Boost	62.97%
Fixed grid	88.5%
Non-Hierarchical Mined, $L = 1$	89.8%
Hierarchical Mined, $L = 3$	94.50%



Method leave-one-out test/train	Average Precision
Kim *et al* [38] CCA	95%
Zhang *et al* [39] BEL	94.33%
Liu and Shah [40] Cuboids	94.15%
Han *et al* citeHanICCV09 MKGPC	94.1%
Uemura *et al* [15] Motion Comp Feats	93.7%
Bregonzio *et al* [41] 2D Gabor filter	93.2%
Yang *et al* [42] Motion Edges	87.3%
Wong and Cipolla [43] Subspace SVM	86.60%
Niebles *et al* [44] pLSA model	81.50%
Dollar *et al* [20] Spat-Temp	81.20%
Fixed grid	90.3%
Non-Hierarchical Mined, $L = 1$	91.7%
Hierarchical Mined, $L = 3$	95.7%

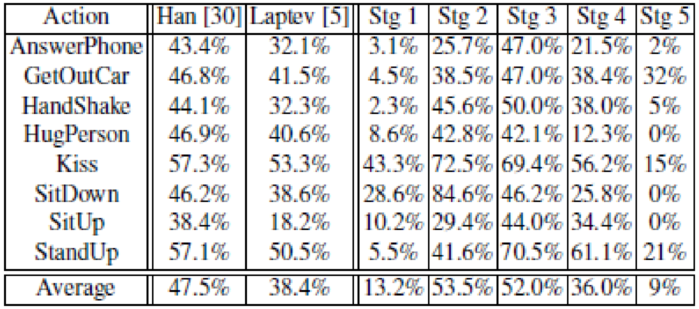
- Multi-KTH: Multiple People and Camera motion panning, zoom

	Clap	Wave	Box	Jog	Walk	Avg
Uemura *et al*	76%	81%	58%	51%	61%	65.4%
US	***69%***	***77%***	***75%***	***85%***	***70%***	***75.2%***

Gilbert, Illingworth, Bowden, Action Recognition Using Mined Hierarchical Compound Features, IEEE TPAMI, May 2011 (vol. 33 no. 5), pp. 883-897

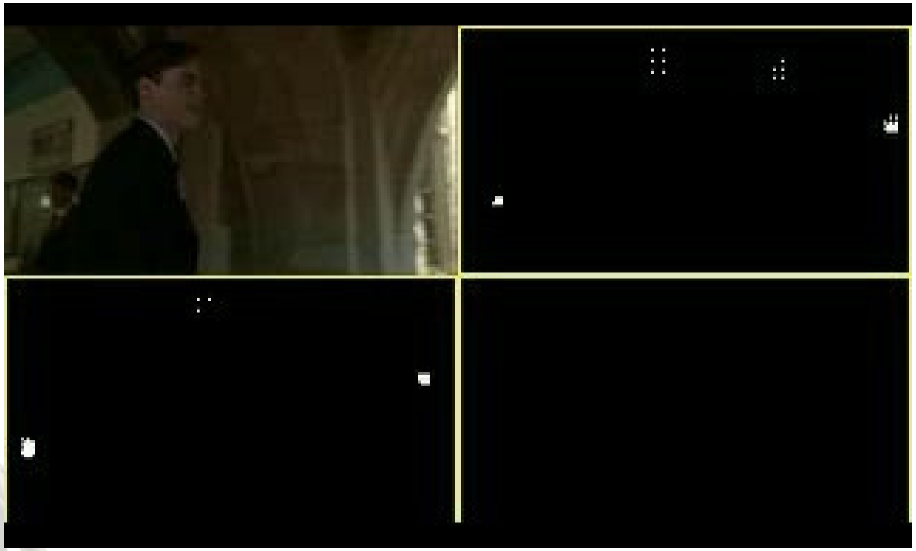
Hollywood Action Recognition

- More recent and realistic dataset
- A number of actions within Hollywood movies



Action	Han [30]	Laptev [5]	Stg 1	Stg 2	Stg 3	Stg 4	Stg 5
AnswerPhone	43.4%	32.1%	3.1%	25.7%	47.0%	21.5%	2%
GetOutCar	46.8%	41.5%	4.5%	38.5%	47.0%	38.4%	32%
HandShake	44.1%	32.3%	2.3%	45.6%	50.0%	38.0%	5%
HugPerson	46.9%	40.6%	8.6%	42.8%	42.1%	12.3%	0%
Kiss	57.3%	53.3%	43.3%	72.5%	69.4%	56.2%	15%
SitDown	46.2%	38.6%	28.6%	84.6%	46.2%	25.8%	0%
SitUp	38.4%	18.2%	10.2%	29.4%	44.0%	34.4%	0%
StandUp	57.1%	50.5%	5.5%	41.6%	70.5%	61.1%	21%
Average	47.5%	38.4%	13.2%	53.5%	52.0%	36.0%	9%

- Hollywood
  - 57%@6 fps
  - No context
- Hollywood2
  - 51%
  - No context



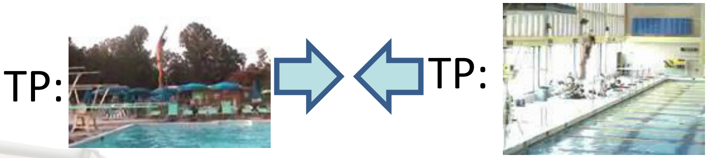
Video Mining and Grouping

- Iteratively Cluster image and video
  - Efficient and intuitive
- The user selects media that semantically belongs to the same class
  - uses machine learning to “pull” this and other related content together.
  - Minimal training period and no hand labelled training groundtruth
  - Uses two text based mining techniques for efficiency with large datasets
    - Min Hash
    - A Priori

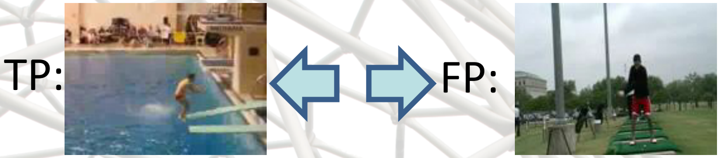
Gilbert, Bowden, iGroup : Weakly supervised image and video grouping, ICCV2011

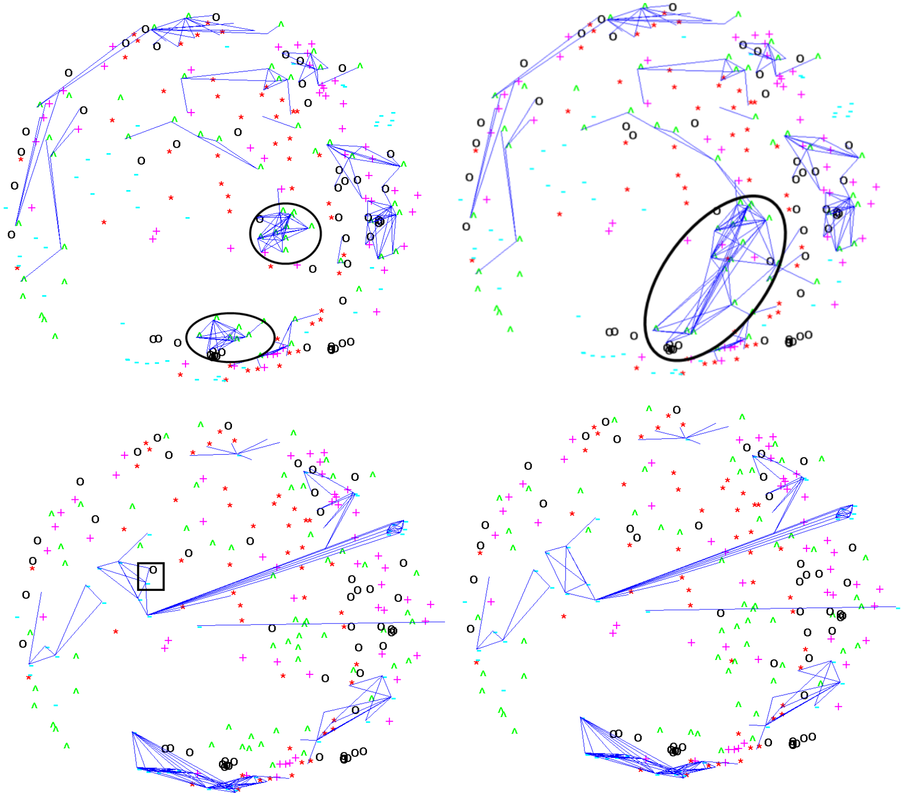
Results – YouTube dataset

- User generated dataset,
  - 1200 videos, 35 secs per iteration
- Pull true pos media together



- Push false positive media apart





- Over 15 iterations of pulling and pushing the media, accuracy of correct group label increases from 60.4% to 81.7%

Sign Recognition

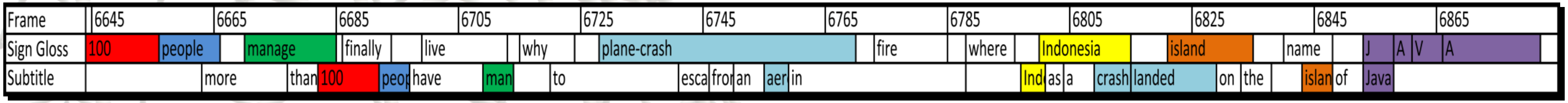
Sign Language Recognition

- Sign Language consists of
  - Hand motion
  - Finger spelling
  - Non Manual Features
  - Complex linguistic constructs that have no parallel in speech

- The problem with Sign is lack of large corpuses of labelled training data

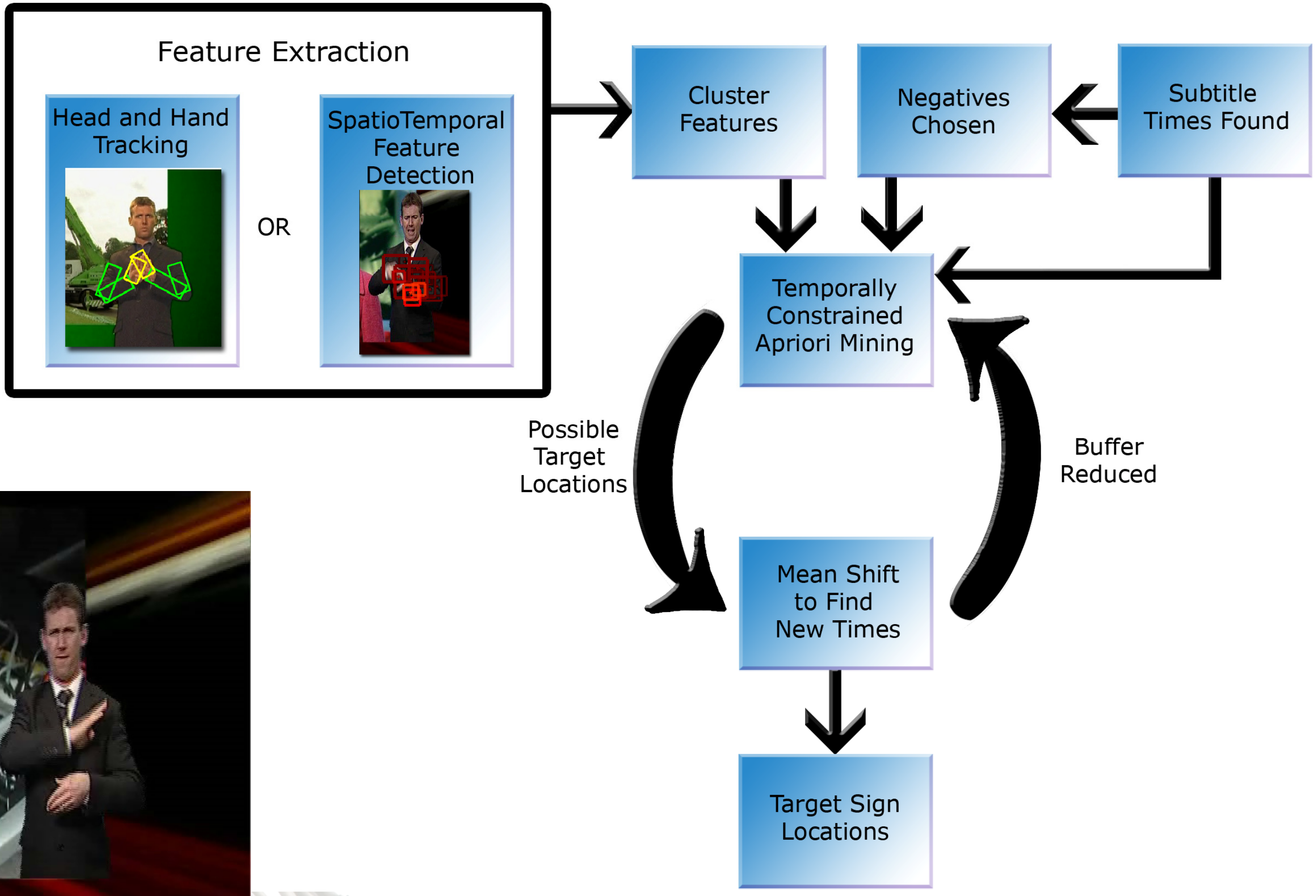
Sign Language

- Labelling large data sets is time consuming and requires expertise.
- Vast amount of sign data is broadcast daily on the BBC.
- BBC data arrives with its own weak label in the form of a subtitle.
- Can we learn what a sign looks like using the subtitle data?
  - Yes… But it's not as easy as it sounds!



Frame	6645	6665	6685	6705	6725	6745	6765	6785	6805	6825	6845	6865									
Sign Gloss	100	people	manage	finally	live	why	plane-crash	fire	where	Indonesia	island	name	J	A	V	A					
Subtitle	more	than	100	peop	have	man	to	esca	fro	an	aer	in	Ind	as	a	crash	landed	on	the	islan	of	Java

Mining Signs



Feature Extraction

Head and Hand Tracking

OR

SpatioTemporal Feature Detection

Cluster Features

Negatives Chosen

Subtitle Times Found

Temporally Constrained Apriori Mining

Possible Target Locations

Buffer Reduced

Mean Shift to Find New Times

Target Sign Locations

Mined results for the signs Army and Obese

Cooper H M, Bowden R, Learning Signs from Subtitles: A Weakly Supervised Approach to Sign Language Recognition.CVPR09. pp2568-2574.

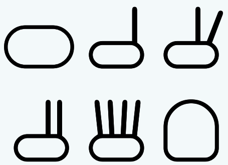
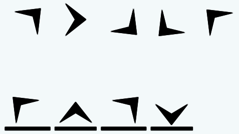
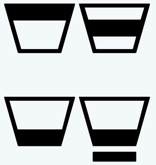
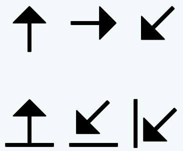
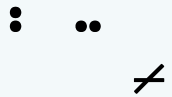
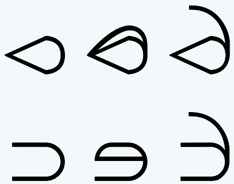
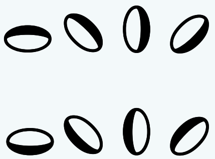
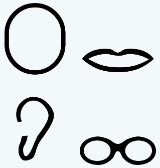
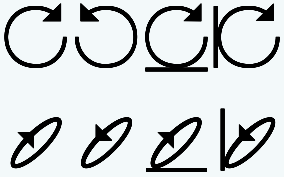
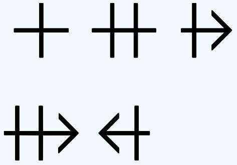
Sign Language Recognition

- New project with Zisserman (Oxford) and Everingham (Leeds)
  - Learning to Recognise Dynamic Visual Content from Broadcast Footage

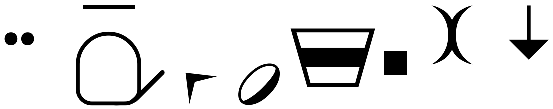
- Currently working on the project Dicta-Sign
- Parallel corpora across 4 sign languages
- Automated tools for annotation using HamNoSys
- Web2.0 tools for the Deaf Community
  - Demonstration: Sign Wiki

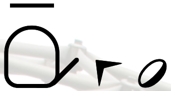
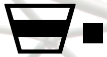
HamNoSys

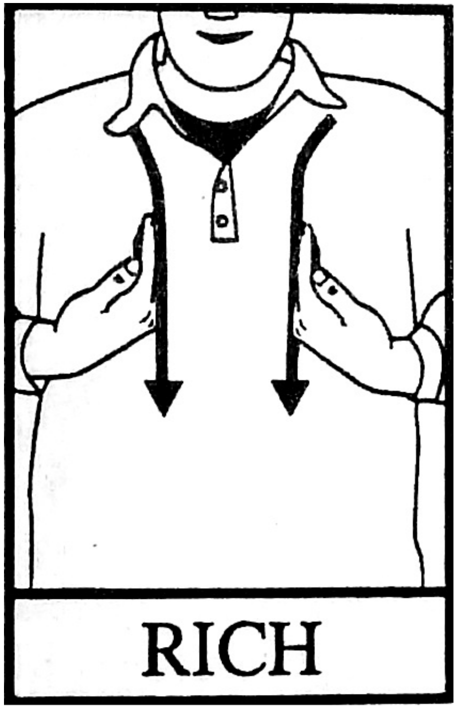
- Linguistic documentation of sign data
- Pictorial representation of phonemes
  - e.g:

Handshape	Orientation	Location	Movement	Constructs
Open	Finger	Torso	Straight	Symmetry
				
Closed	Palm	Head	Circle/Ellipse	Repetition
				

HamNoSys Example

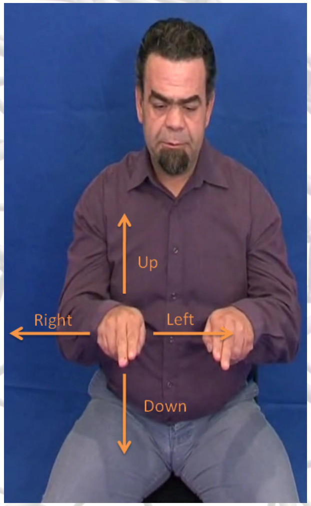


- ¨ left - right mirror
-  hand shape/orientation
-  Right side of torso
-)(contact with torso
- ↓ downwards motion

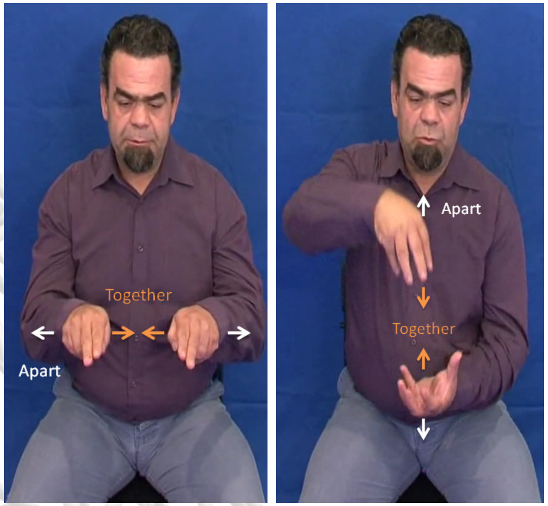


Motion Features

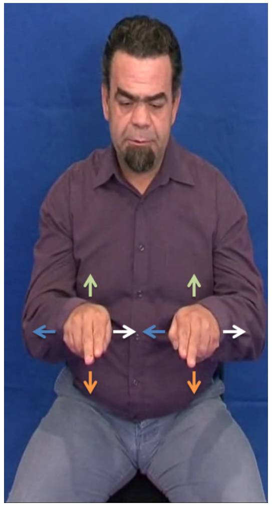
- Automated tools help for annotation
- Useful in recognition as they generalise
- Features follow subset of HamNoSys
  - Location
  - Motion
  - Handshape



Direction



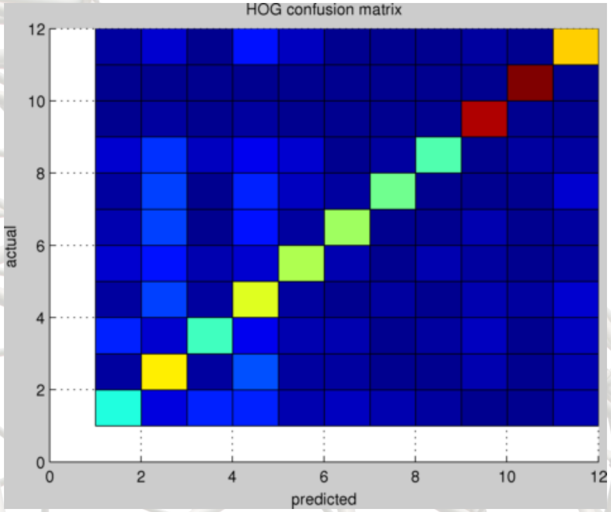
Relative together/apart

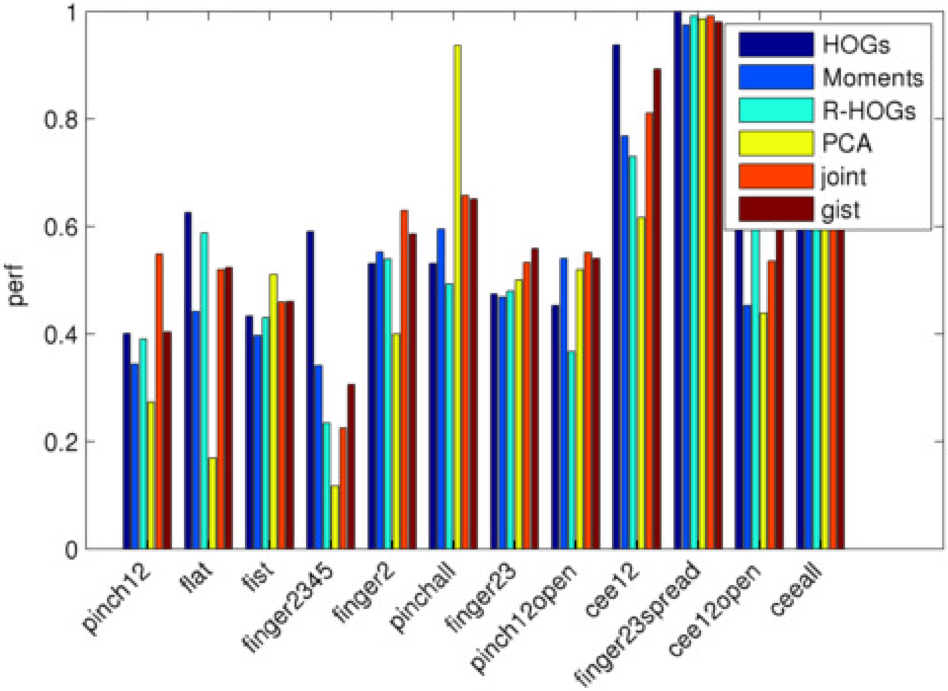


Synchronous motion

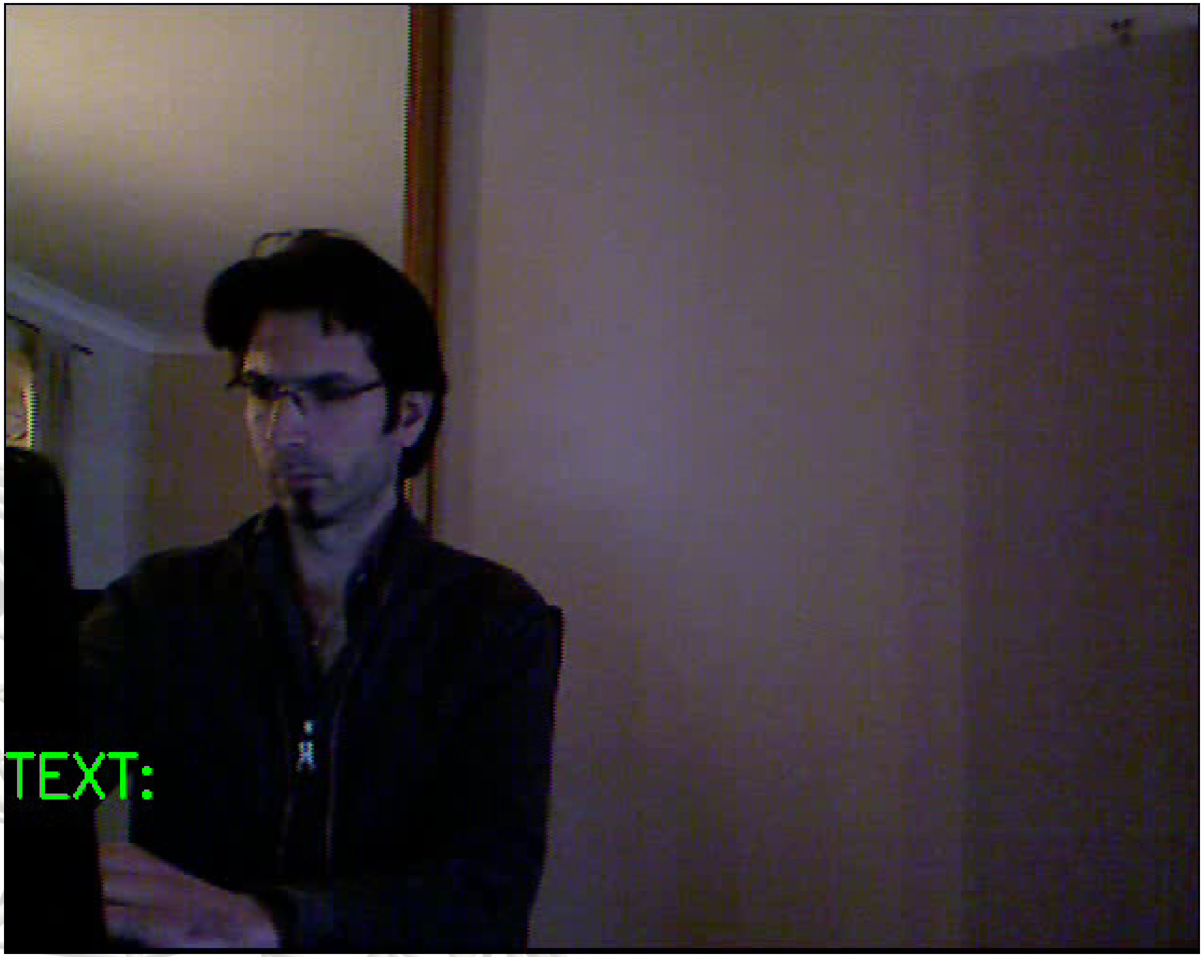
Mapping Hands to HamNoSys

- Align PDTS with HamNoSys
  - Identify which hand shapes are likely in which frame
  - Extract features for that frame e.g. HOG, GIST, Sobel, moments
- RDF, multiclass classifier





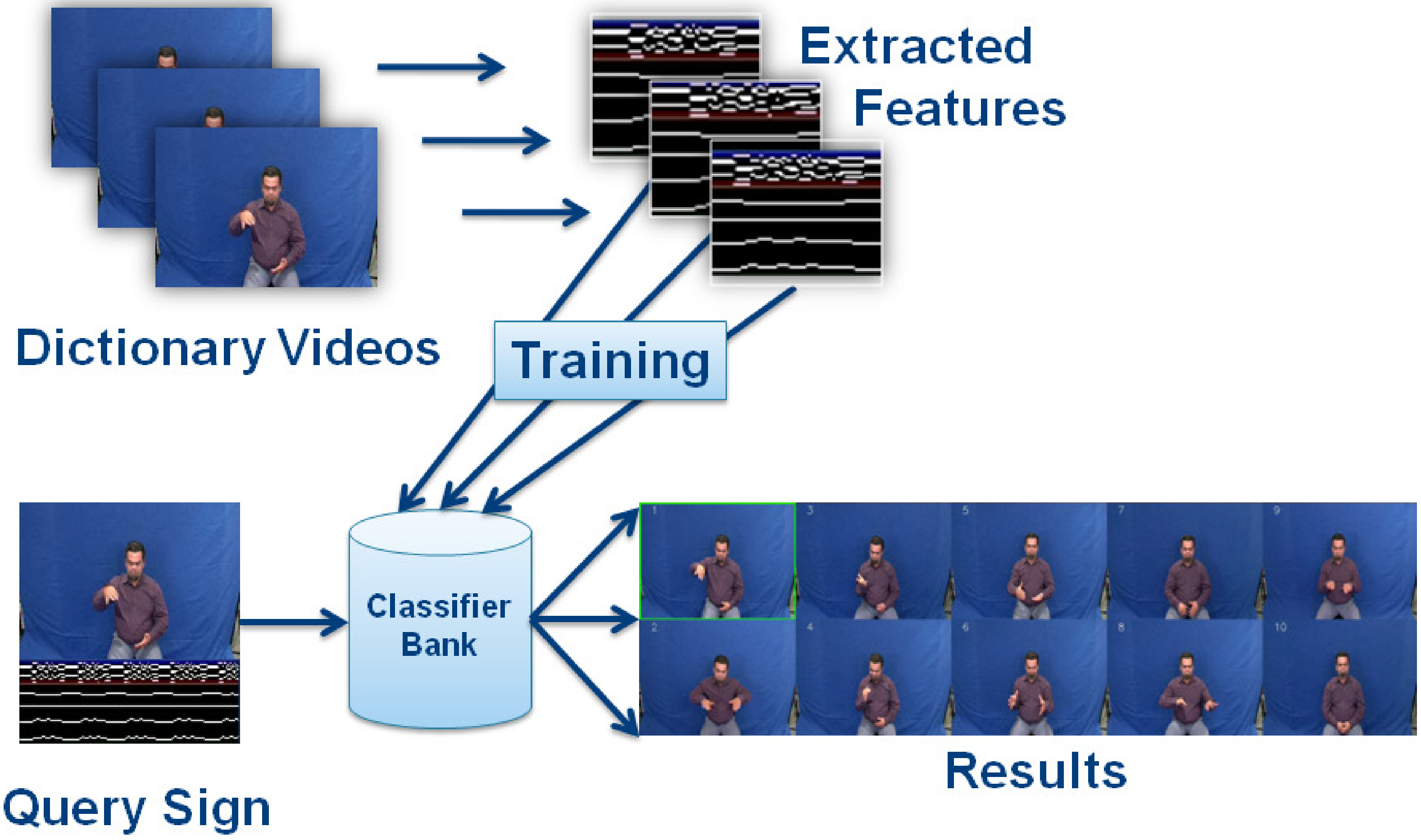
Handshape demonstrator



Motion Features

- Features are not mutually exclusive and can fire in combination.

Dictionary Overview



Results

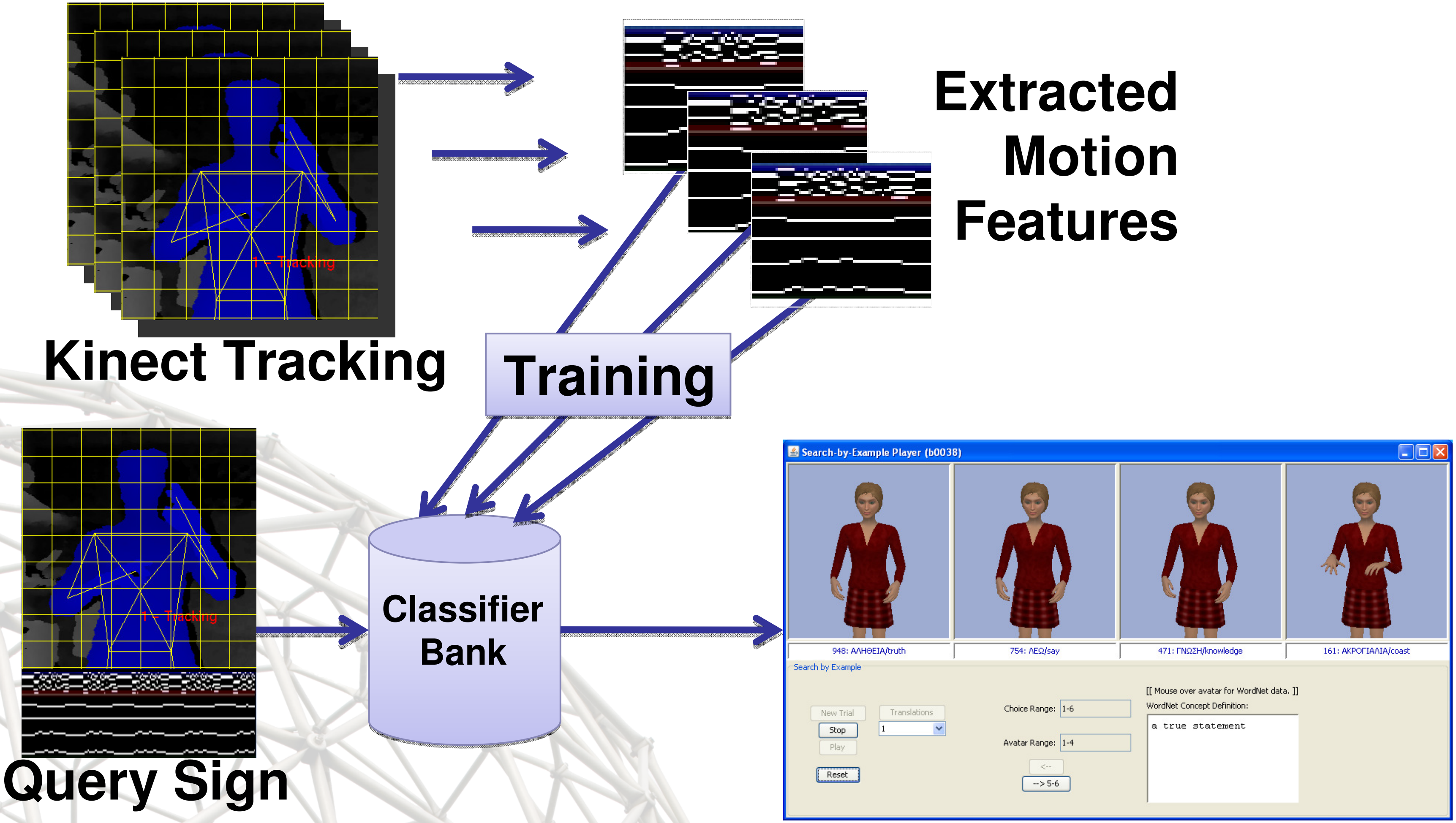
- 984 isolated signs, single signer, 5 rep
- Using feature types individually or in pairs

Results Returned	Motion	Location	Handshape	Motion + Handshape	Motion + Location	Location + Handshape
1	25.1%	60.5%	3.4%	36.0%	66.5%	66.2%
10	48.7%	82.2%	17.3%	60.7%	82.7%	86.9%

- Using all types of features in combination

Results Returned	1st Order Transitions	2nd Order Transitions	WTA Handshape + 2nd Order	WTA Handshape + 1st Order
1	68.4%	71.4%	54.0%	52.7%
10	85.3%	85.9%	59.9%	59.1%

Live Demo



Search-by-Example Player (b0038)

948: ΑΛΗΘΕΙΑ/truth

754: ΛΕΩ/say

471: ΓΝΩΣΗ/knowledge

161: ΑΚΡΟΓΙΑΛΙΑ/coast

Search by Example

Translations: 1

Choice Range: 1-6

Avatar Range: 1-4

[[Mouse over avatar for WordNet data.]]

WordNet Concept Definition:

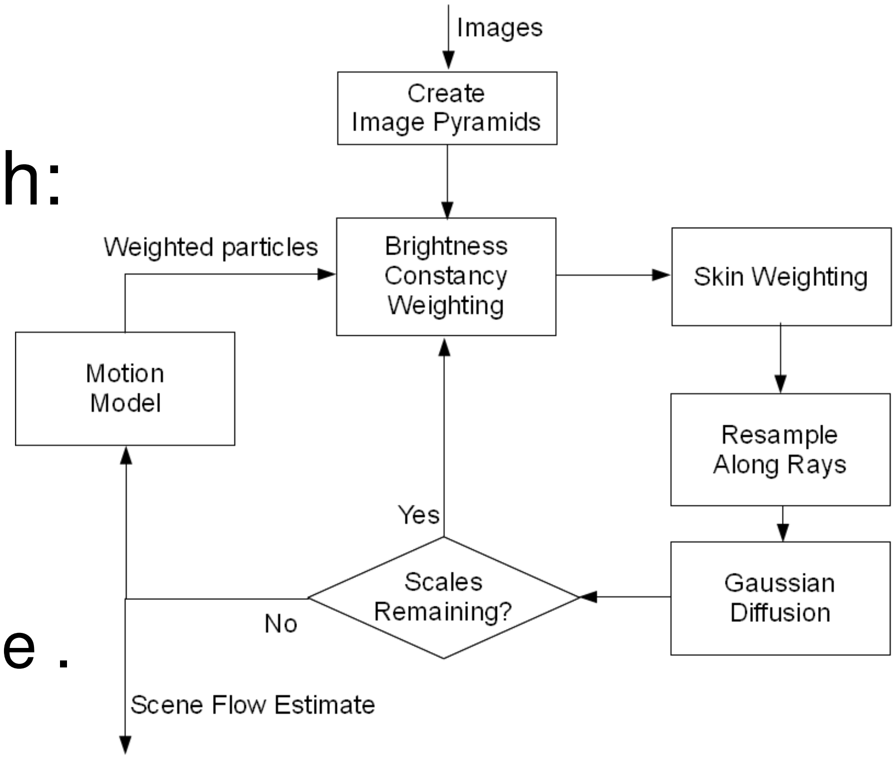
```
a true statement
```

Kinect Demo

Moving to 3D features

Scene Particle approach

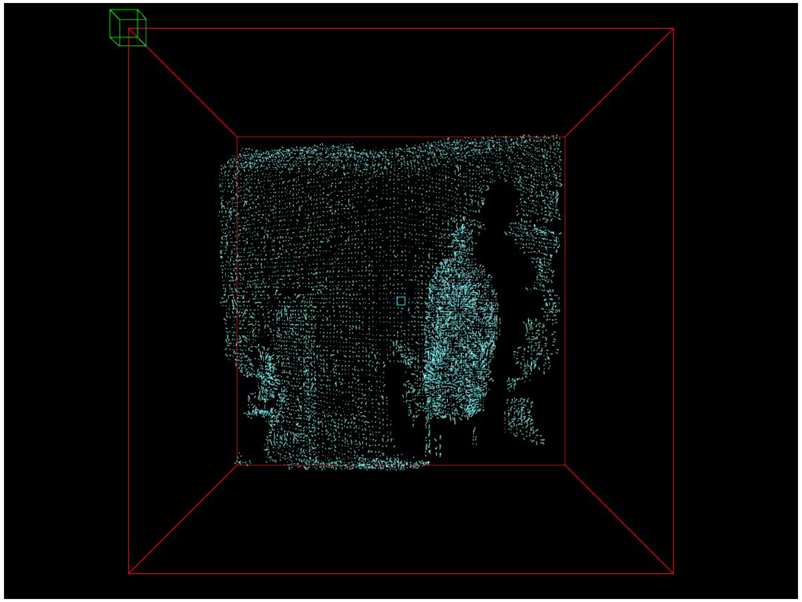
- Scene Particle approach:
  - Particle Filter inspired.
  - Multiple hypotheses.
  - No smoothing artifacts.
  - Easily parallelisable.
  - Kinect: 10 secs per frame .
  - Multi-view: 2 mins per frame.



Hadfield, Bowden. Kinecting the dots: Particle Based Scene Flow from depth sensors, ICCV2011

Scene Particles

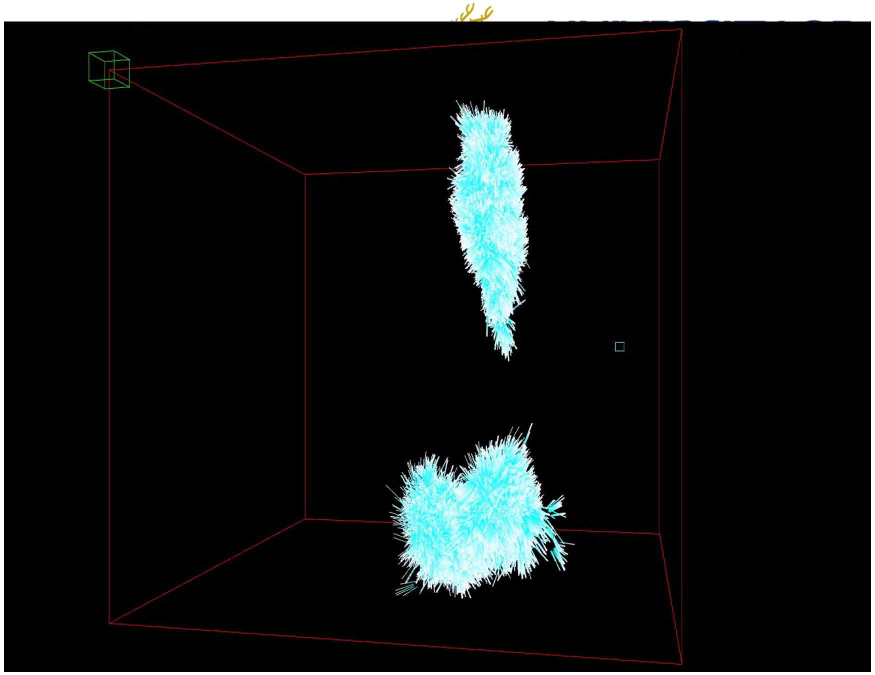
- Middlebury stereo dataset:
- Structure 20x better.
- Motion mag. 5x better.

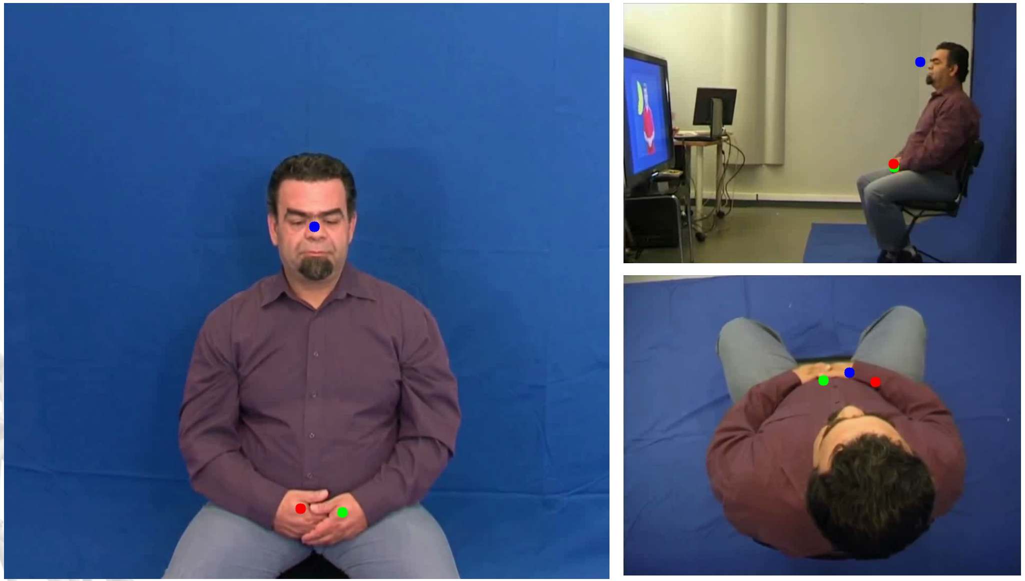


Approach	Structure	Op. Flow	Z Flow	AAE	
Scene Particle	**0.31**	**0.16**	**0.00**	3.43	
Basha 2010	6.22	1.32	0.01	**0.12**	
Huguet 2007	5.55	5.79	8.24	0.69	

3D Tracking

- Scene Particle system.
- Adaptive skin model.
- 6D (x+dx) clustering.
- 3D trajectories.





Kinect Data Set

- 20 Signs
  - Randomly chosen GSL
  - Some similar motions (e.g. April and Athens)
- 6 people ~7 repetitions per sign
- OpenNI / NITE skeleton data
- Extracted HamNoSys motion and location features
- Motion Features same as 2D case plus the Z plane motions.

3D Kinect Results

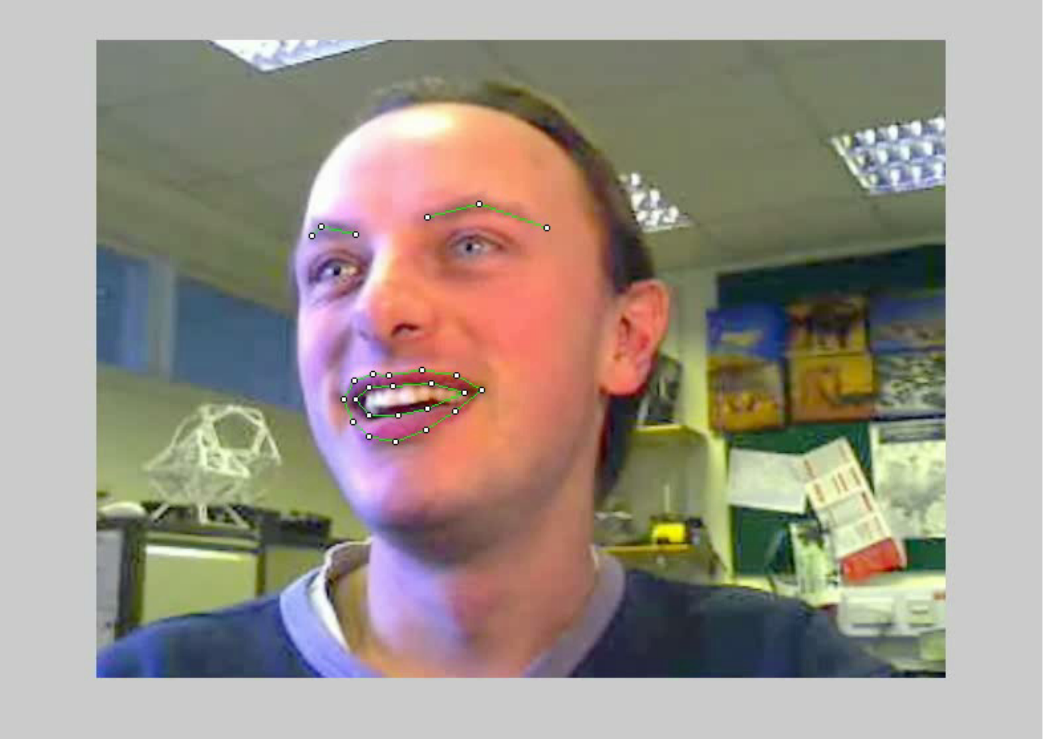
- User Independent (5 subject train,1 test)
- All Users (leave one out method)

Test Subject	Markov Chain		Sequential Patterns	
	Top 1	Top 4	Top 1	Top 4
B	56%	80%	72%	91%
E	61%	79%	80%	98%
H	30%	45%	67%	89%
N	55%	86%	77%	95%
S	58%	75%	78%	98%
J	63%	83%	80%	98%
Average	**54%**	**75%**	**76%**	**95%**
All	**79%**	**92%**	**92%**	**99.9%**

Facial Feature Tracking

Facial Feature Tracking

- Primarily built for lip reading
- Flocks of Linear Predictors
  - provide fast accurate regresser functions for tracking
  - generic, can track any object or feature
  - accurate tracking of any facial feature
  - allows accurate pose estimation



Ong, Bowden, Robust Facial Feature Tracking Using Shape-Constrained Multi-Resolution Selected Linear Predictors, IEEE TPAMI, accepted, to appear

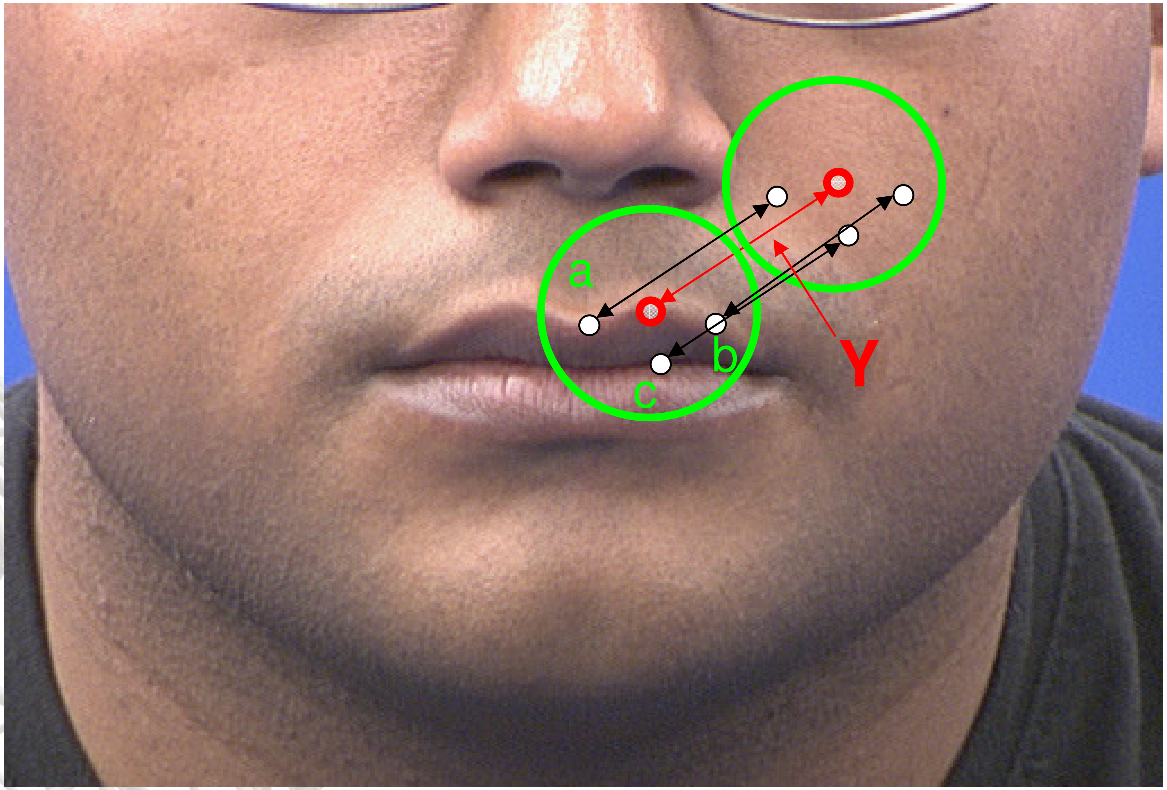
Linear Predictors

(Marchand et al 1999, Jurie & Dhome 2002, Matas et al 2006)

- Reference Point + Support Pixels (a,b,c)
- Linear mapping (H) from support pixel intensity difference to translation vector

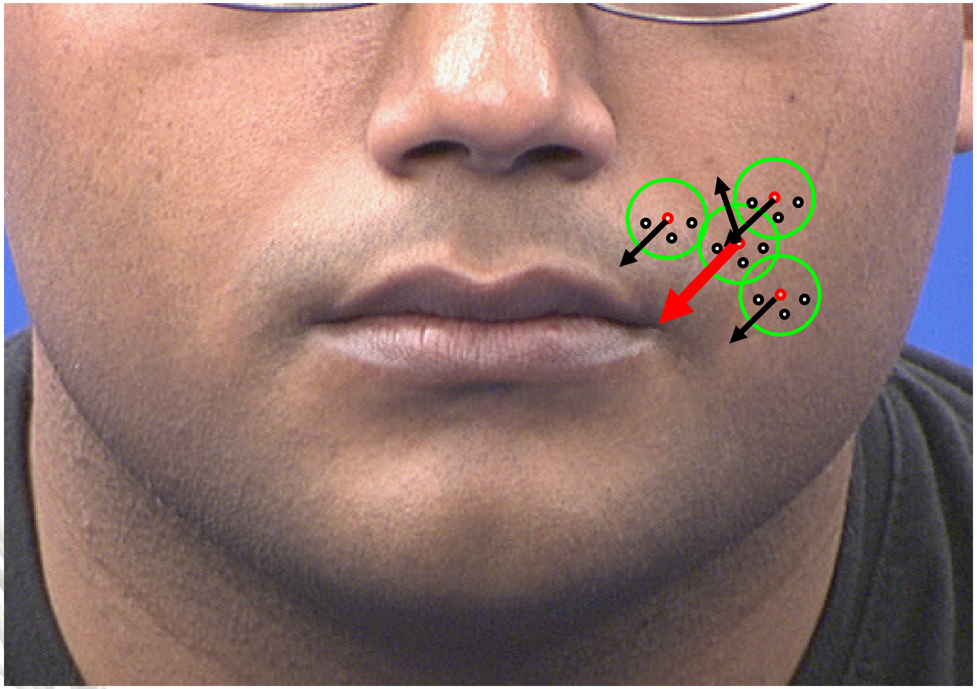
$$\delta P = [\ Ia - I'a,\ Ib - I'b,\ Ic - I'c\ ]$$

$$Y = H\delta P$$



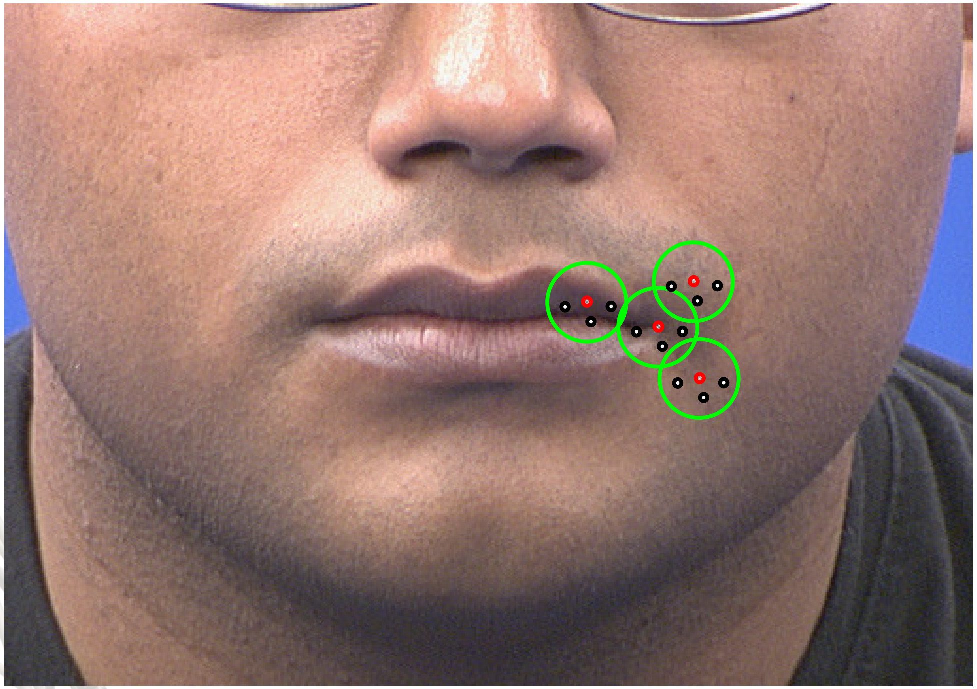
Linear Predictors

- Linear Predictor “Bunches”
  - Single LPs are not stable enough for tracking image features
  - Use a set (“bunch”) of LPs instead
  - Final prediction = consensus of the most common predicted translation

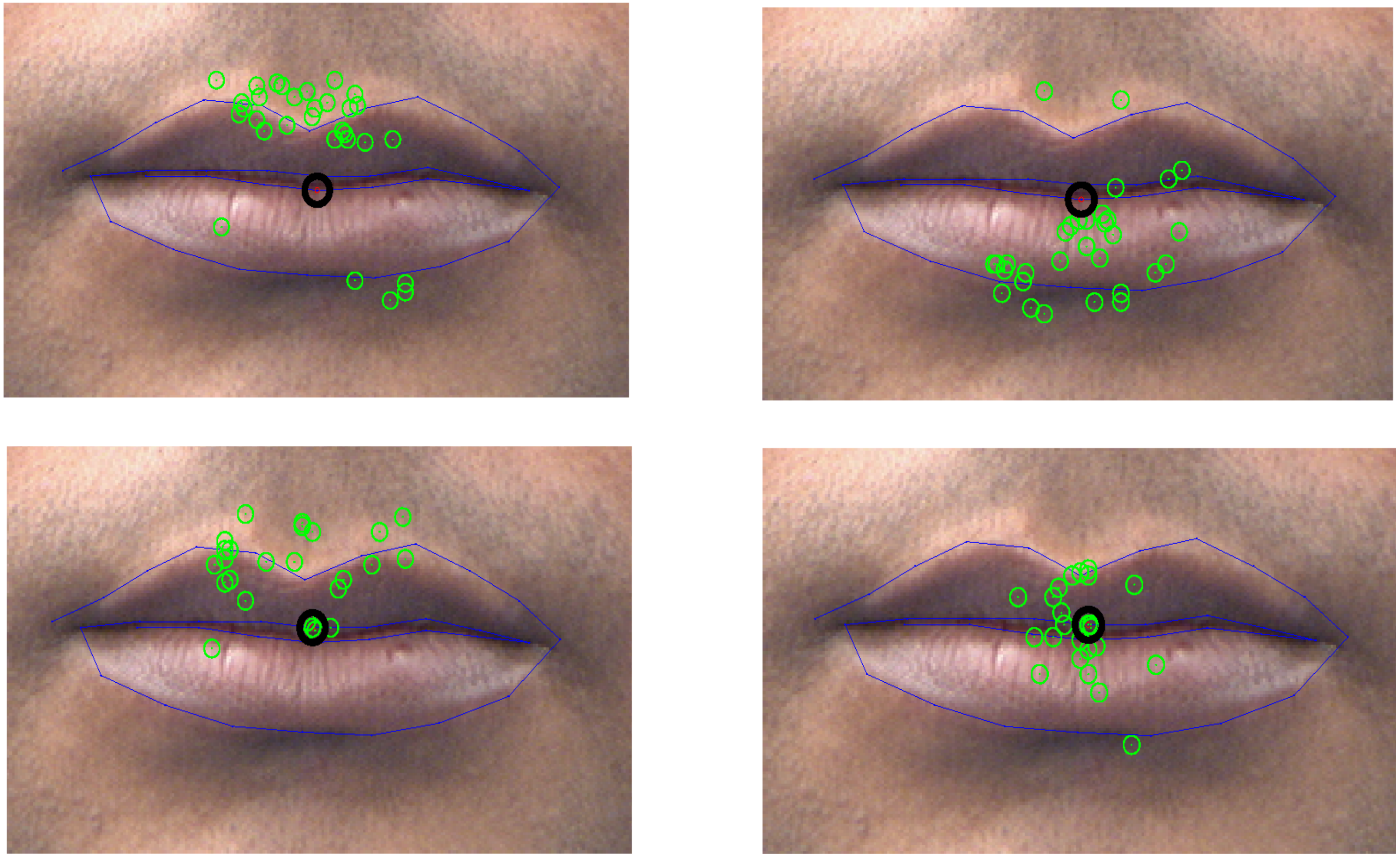


Linear Predictors

- Linear Predictor “Bunches”
  - Single LPs are not stable enough for tracking image features
  - Use a set (“bunch”) of LPs instead
  - Final prediction = consensus of the most common predicted translation



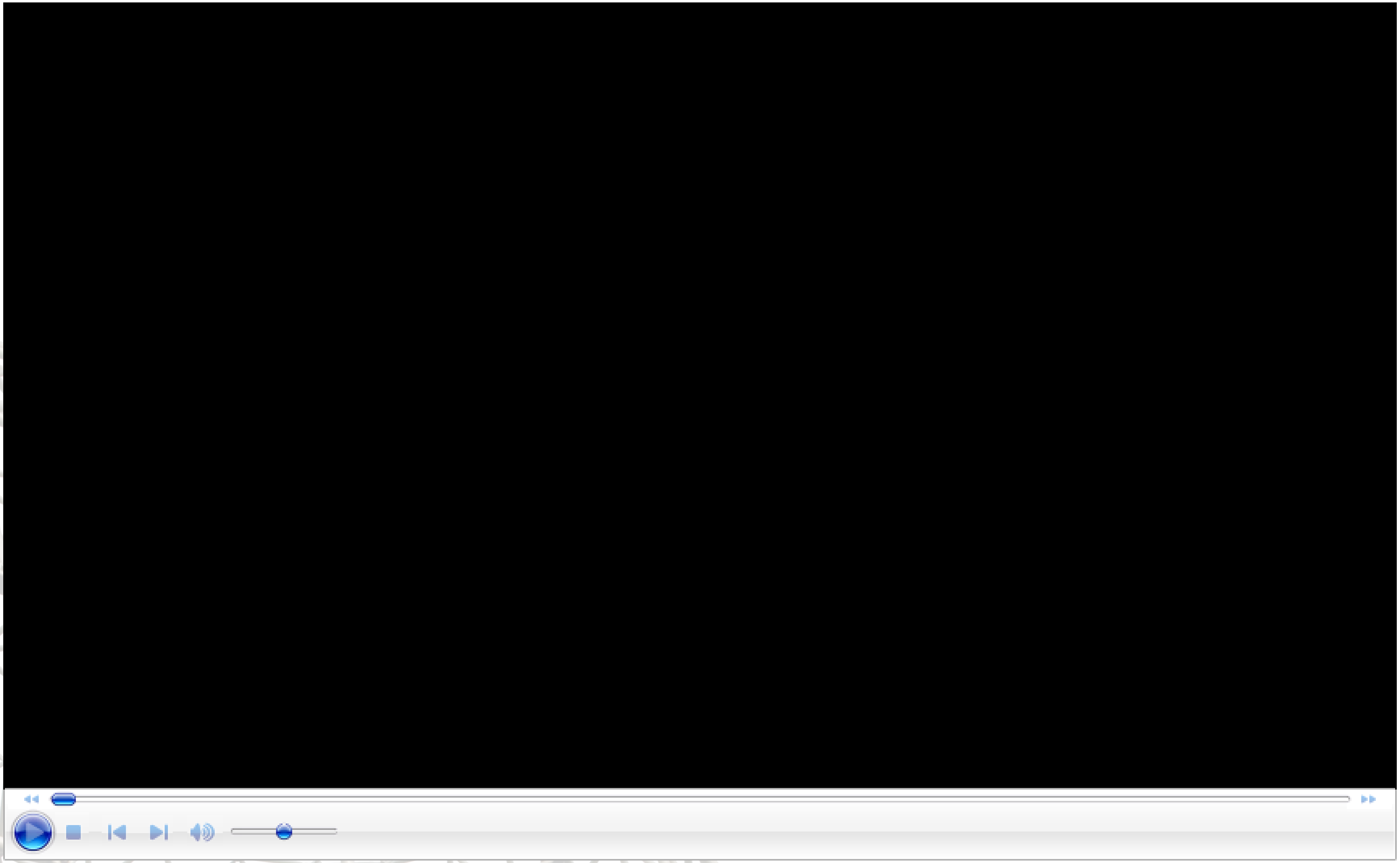
Tracking lips with Linear Predictors



X Translation

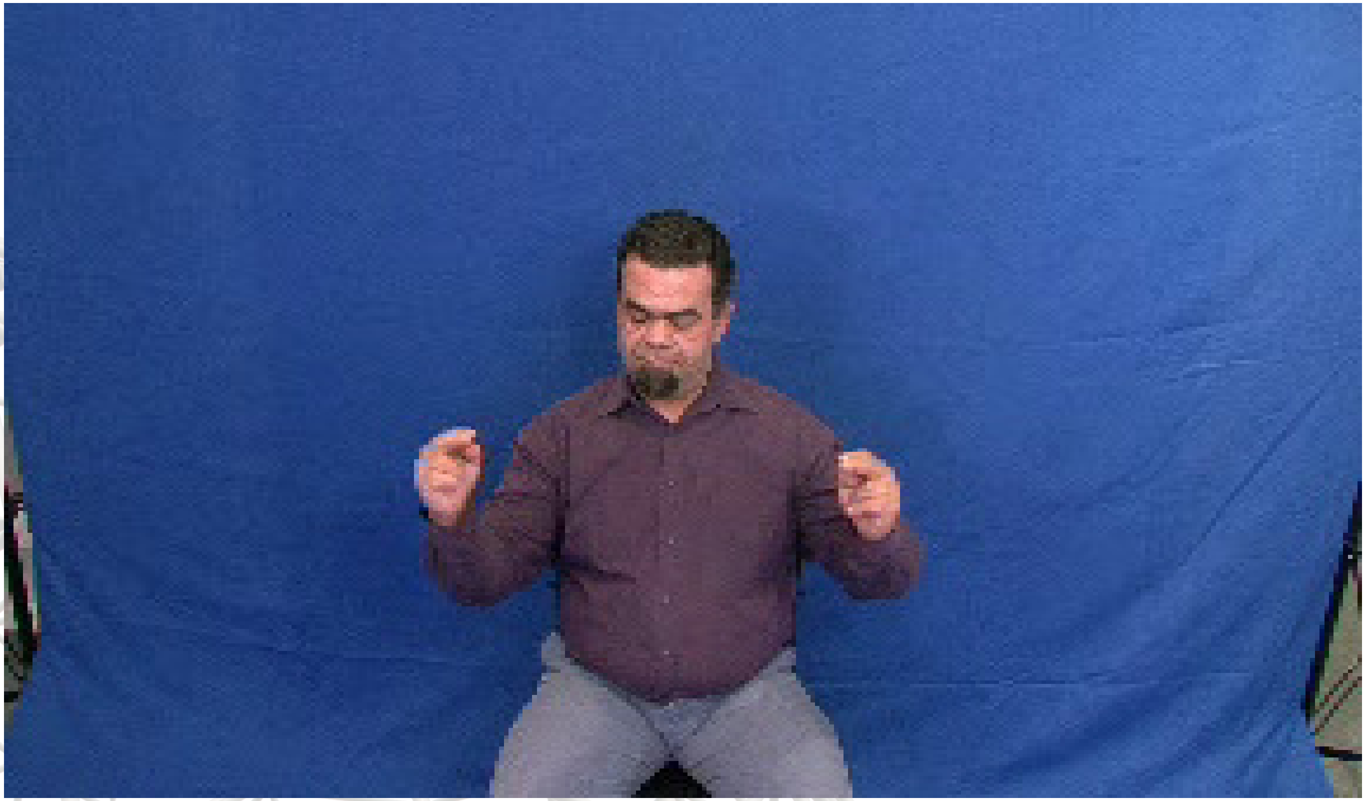
Y Translation

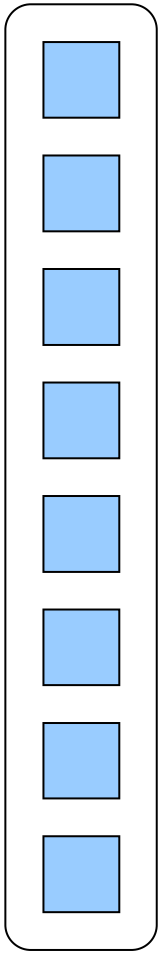
Facial Feature Tracking



Sequential Patterns

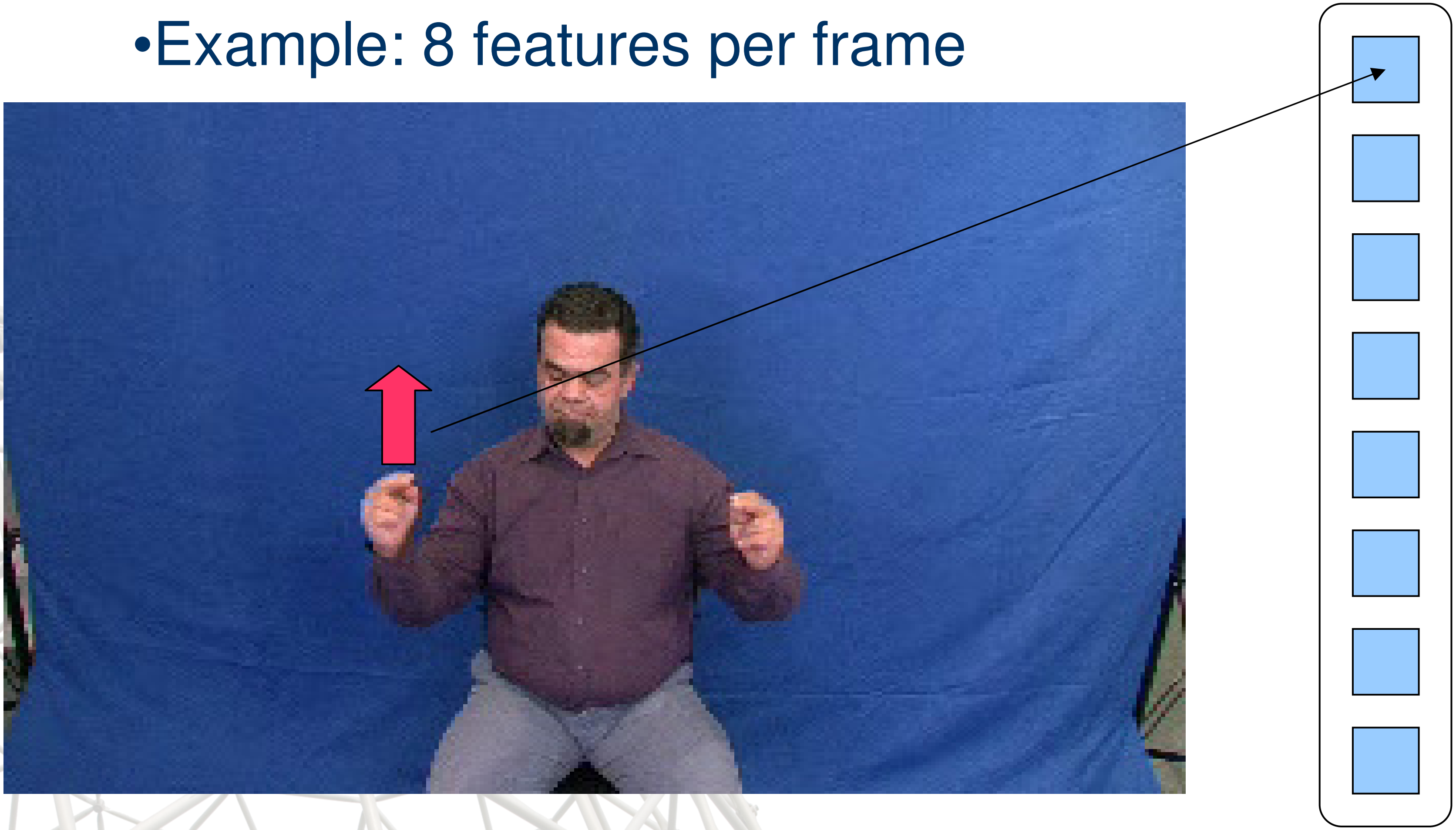
- Sequential Patterns: Sequence of feature subsets
- Example: 8 features per frame





Sequential Patterns

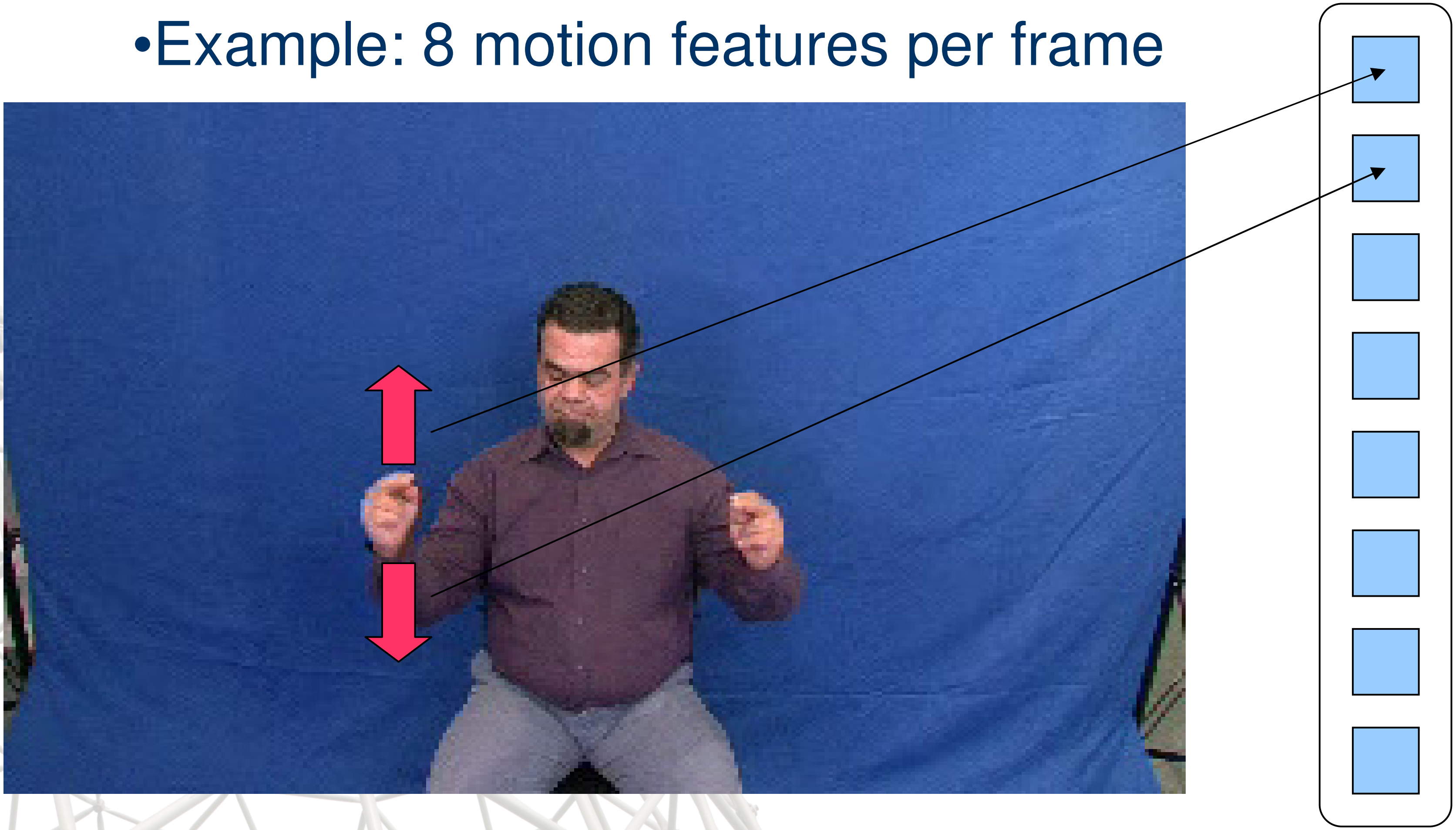
- Sequential Patterns: Sequence of feature subsets
- Example: 8 features per frame



Sequential Patterns

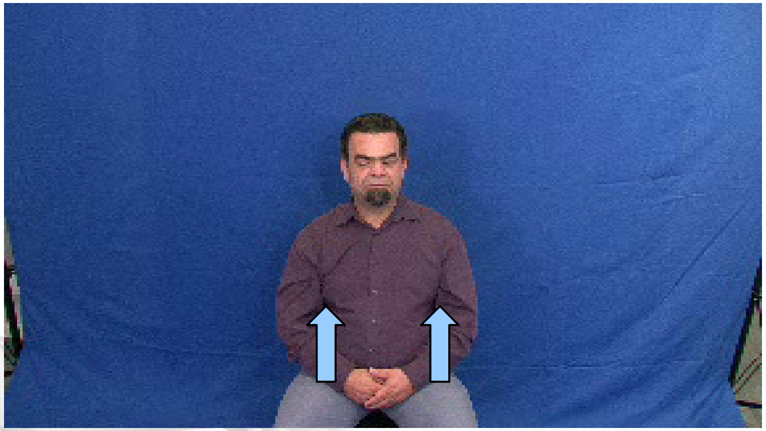
- Sequential Patterns: Sequence of feature subsets

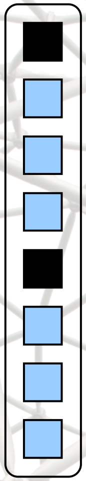
•Example: 8 motion features per frame



Sequential Patterns

- Sequential pattern example for Bridge



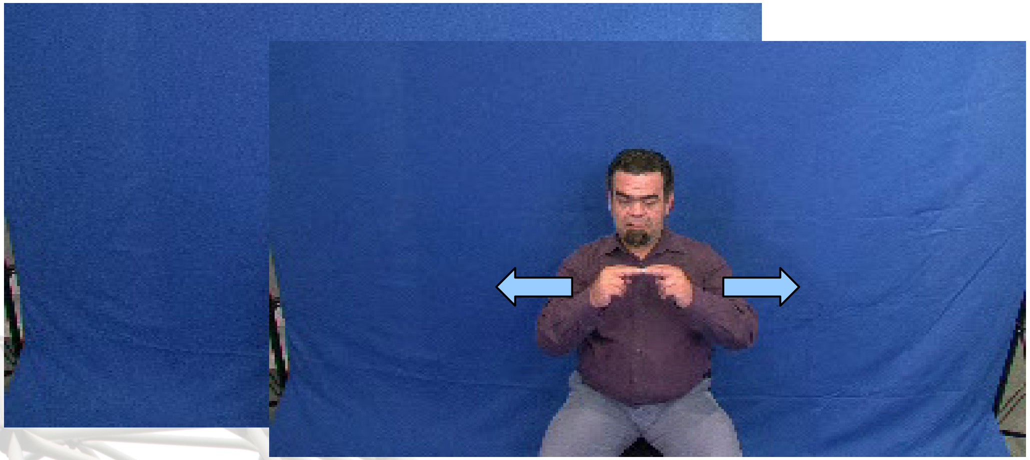


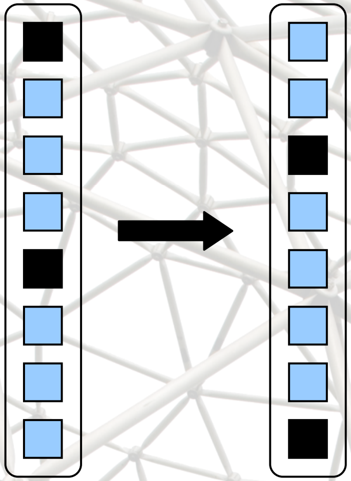
🟦 Motion not present

■ Motion present

Sequential Patterns

- Sequential pattern example for Bridge



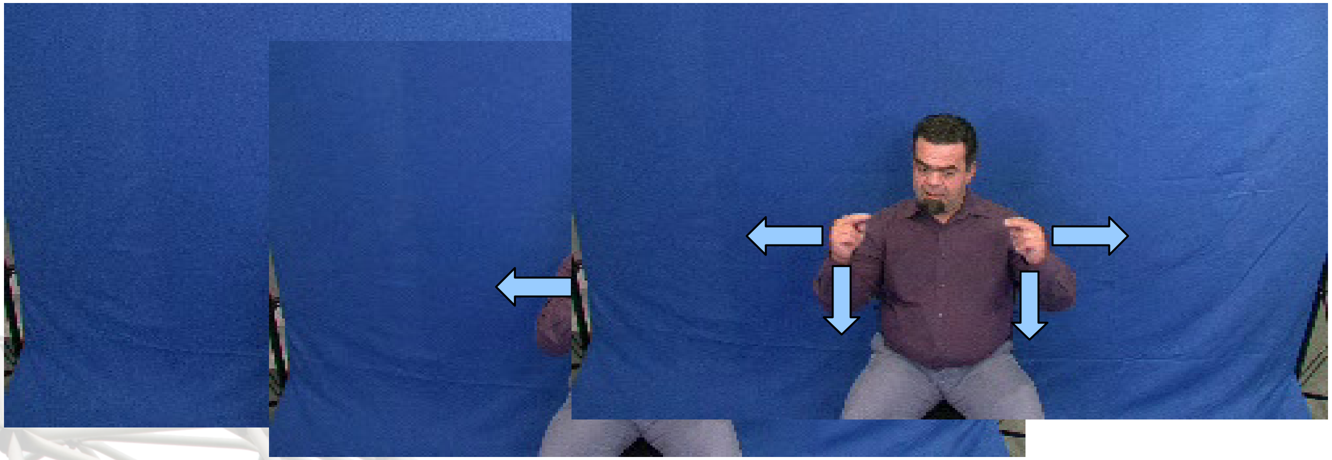


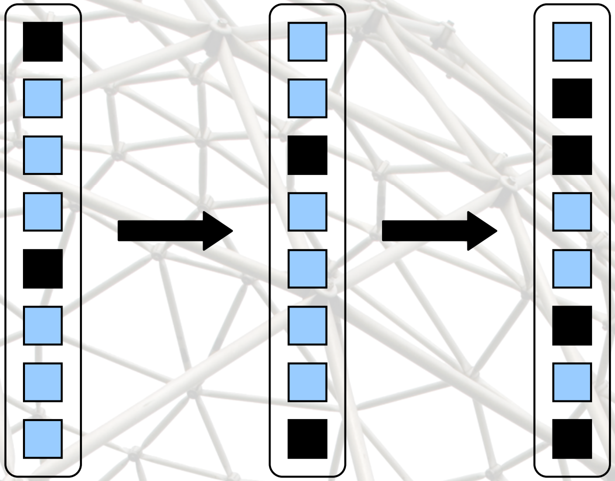
🟦 Motion not present

⬛ Motion present

Sequential Patterns

- Sequential pattern example for Bridge





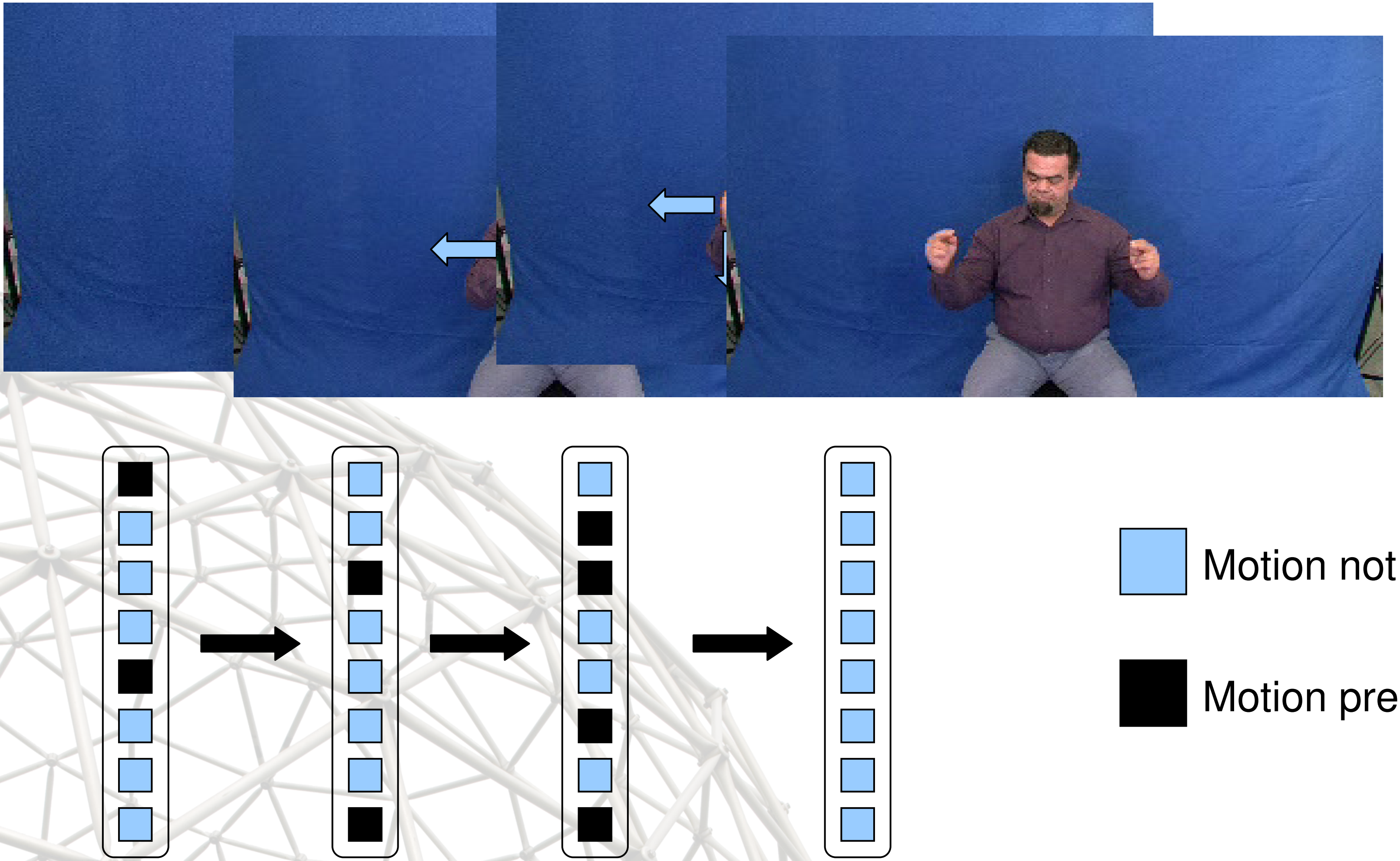
⬛	🟦	🟦
🟦	🟦	⬛
🟦	⬛	⬛
🟦	🟦	🟦
⬛	🟦	🟦
🟦	🟦	⬛
🟦	🟦	🟦
🟦	⬛	⬛

🟦 Motion not present

⬛ Motion present

Sequential Patterns

- Sequential pattern example for Bridge



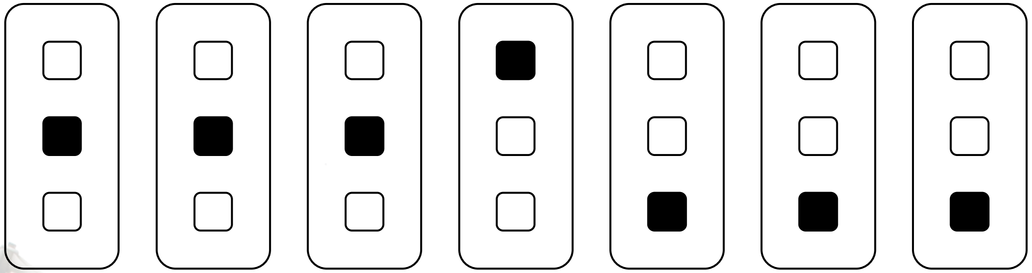
⬛	🟦	🟦	🟦
🟦	🟦	⬛	🟦
🟦	⬛	⬛	🟦
🟦	🟦	🟦	🟦
⬛	🟦	🟦	🟦
🟦	🟦	⬛	🟦
🟦	🟦	🟦	🟦
🟦	⬛	⬛	🟦

🟦 Motion not present

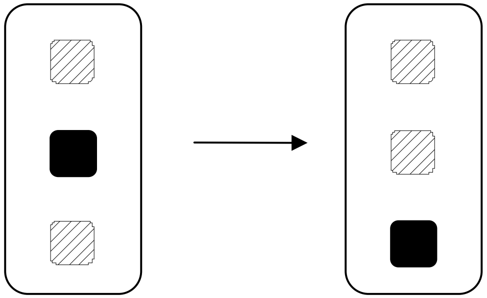
⬛ Motion present

Sequential Patterns

- Matching a sequential pattern to an input sequence:
  - Suppose we are given an input sequence of features

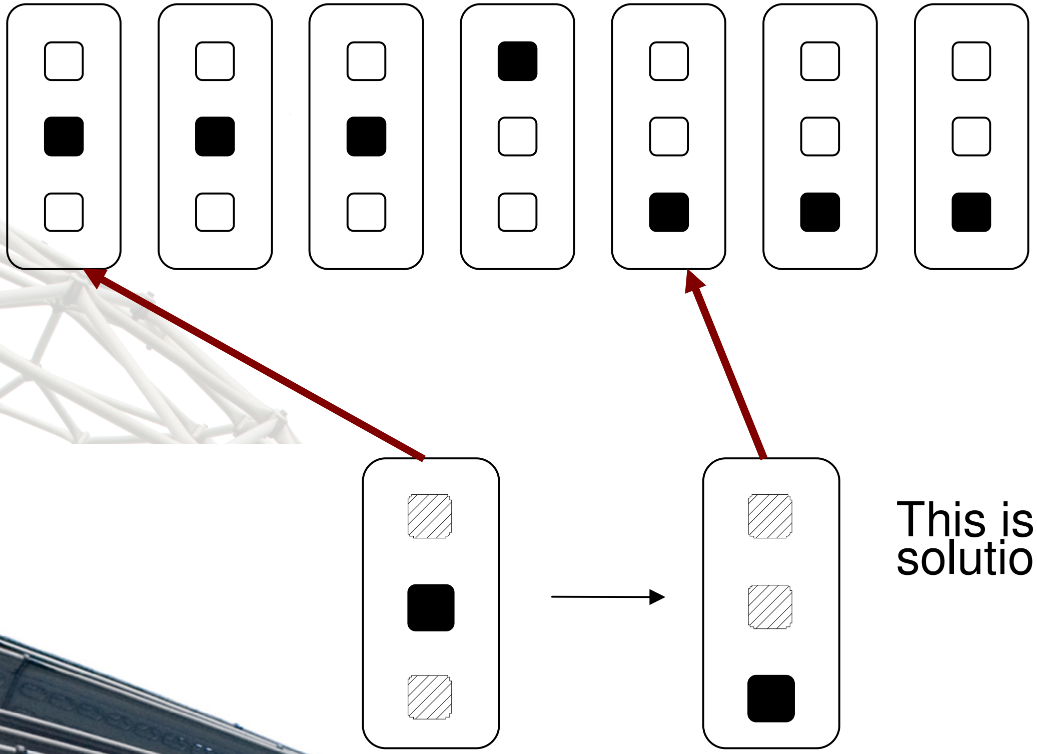


The goal is to find whether this sequence of classification results exists within the input sequence



Sequential Patterns

- Matching a sequential pattern to an input sequence:
  - There are multiple solutions to how a sequential pattern can be found in an input sequence



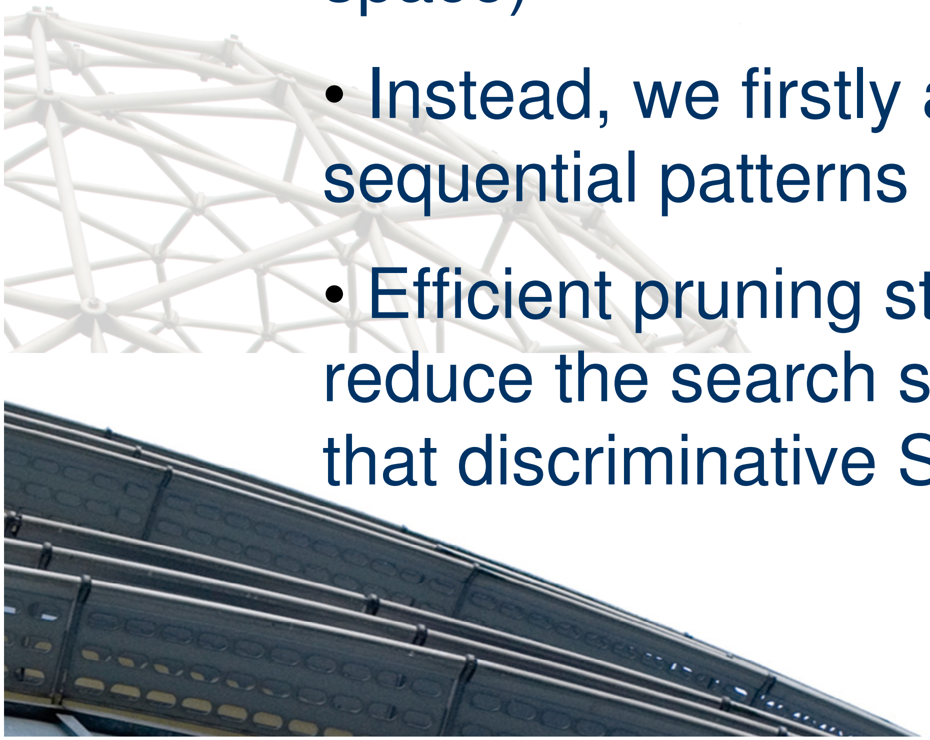
This is one possible solution

Sequential Patterns

- Pros:
  - Allows the use of different subsets of features
  - Can handle different speeds in temporal pattern
- Cons:
  - Potential sequential patterns very large: 2^ND (D = number of features)
  - Example: if we have 200 features, for sequences up to length 5, we have 2^{1000} configurations.
  - Assuming we can do 2^{64} searches in a second, we need to wait 2^{936} seconds to do 1 exhaustive search. (Longer than age of the universe).

Sequential Patterns

- Learning
  - With sequential patterns, a naive approach will be to generate all possible sequence configurations. NOT POSSIBLE (2^{ND} search space)
  - Instead, we firstly approach possible sequential patterns as a tree structure.
  - Efficient pruning strategies can then vastly reduce the search space, while guaranteeing that discriminative SPs can be found.



- Show word spotting vid

Conclusions

- Interpreting the meaning of motion is common across all these examples
- Interpreting the meaning of sign is far more complex than just recognising motion
- While approaches therefore differ to suit complexity new learning approaches which can cope with noise in training are important for all areas
- Needless to say we still need more and varied datasets to move forward and need to be careful about optimising our results over them
  - (hopefully preaching to the converted)