Retrieving Target Gestures Toward Speech Driven Animation with Meaningful Behaviors

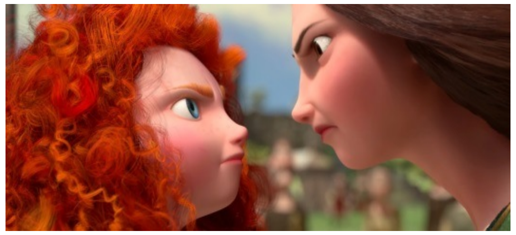
NAJMEH SADOUGHI AND CARLOS BUSSO

Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas
Erik Jonsson School of Engineering and Computer Science

Nov. 11th, 2015

Motivation

- Creating naturalistic nonverbal behaviors is important for conversation agents (CAs)
  - Animations
  - Entertainment
  - Virtual reality
- More than 90% human gestures occur while speaking
- Complex relationship between gestures and speech
  - Cross modality interplay
  - Synchronization

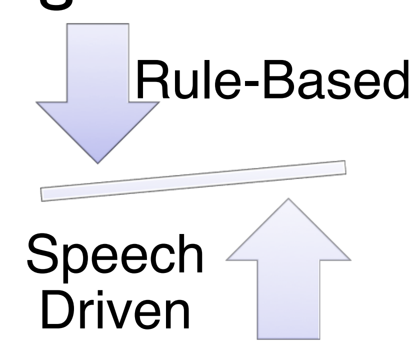


maxresdefault.jpg

ICT-USC

Previous studies on co-verbal gesture synthesis

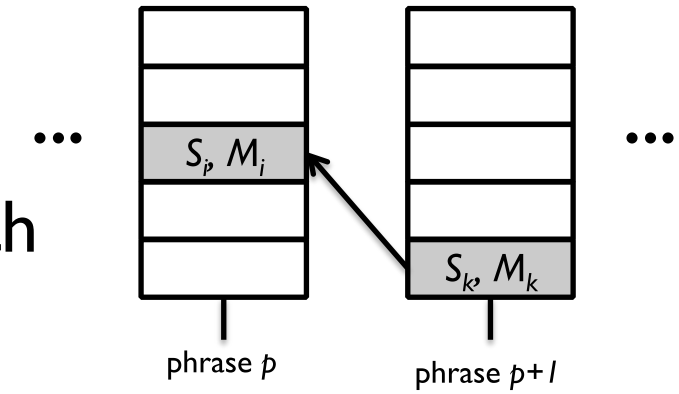
- Rule based frameworks [Cassell et al., 1994; S. Kopp 2006]
  - \+ Define rules based on the semantics
  - \- Synchronization is challenging
  - \- The variation is limited
- Speech prosody driven systems [Levine et al., 2010; Busso et al. 2007]
  - \+ Learn movements and their synchronization from recordings
  - \+ Capture the variation in the data
  - \- Disregard the context



- Combination of data driven and rule based methods [Stone et al. 2004, Marsella et al. 2013, and Sadoughi et al. 2014]
  - \+ Utilizing the advantages and overcoming the disadvantages

Previous studies using both approaches

- Stone et al., [2004]
  - Search for combination of speech and motion units with similar meaning to speech and planned behaviors

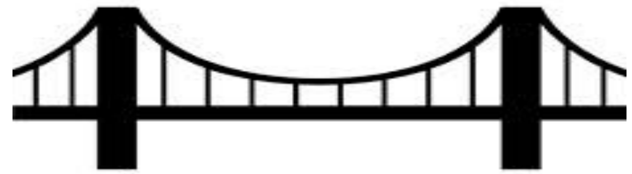


- Marsella et al., [2013]
  - Create appropriate gestures depending on the communicative goal of the utterance
  - Use speech prosody features to capture the stress and emotional state of the speaker

- Sadoughi et al., [2014]
  - Constrain a speech driven animation model based on semantic labels (e.g., Question and Affirmation)

Our Vision

Rule-based systems



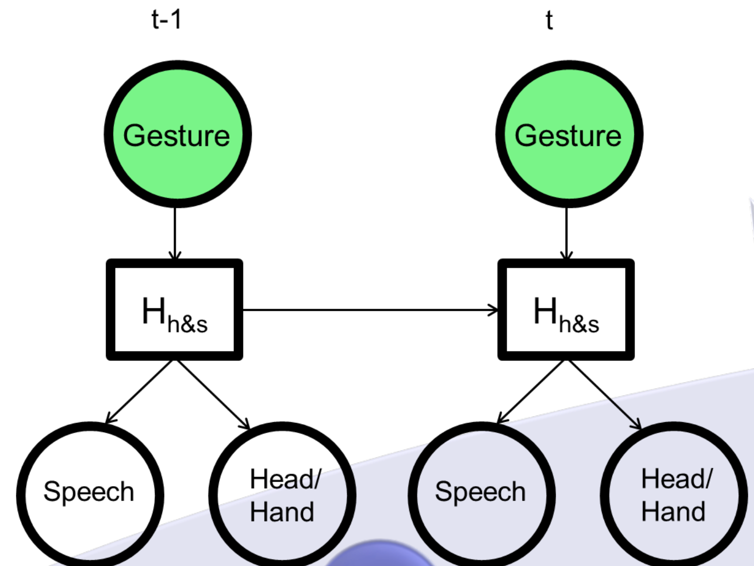
Data-driven systems

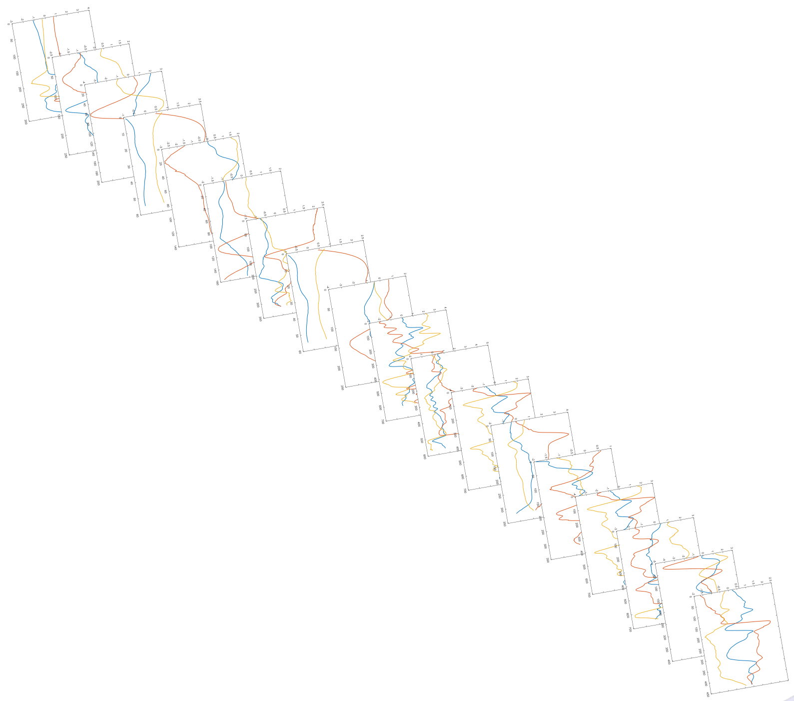
- Creating a bridge between rule based systems and data driven framework
- SAIBA framework [Kopp et al., 2006]:

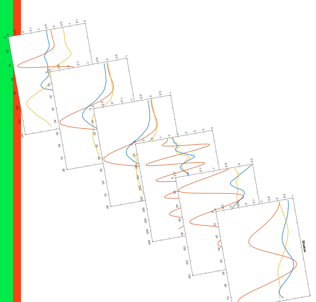
Intent Planning → Behavior Planning → Behavior Realization

- Considering the target gesture for synthesis is known
  - Synthesizing behaviors that are **timely aligned and coordinated with speech**
  - Synthesizing behaviors that **convey the right meaning**

Objective of This Study



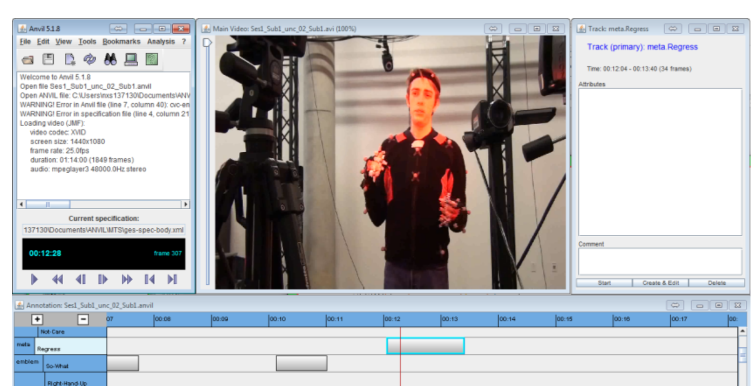




Annotating few samples of a prototypical gesture

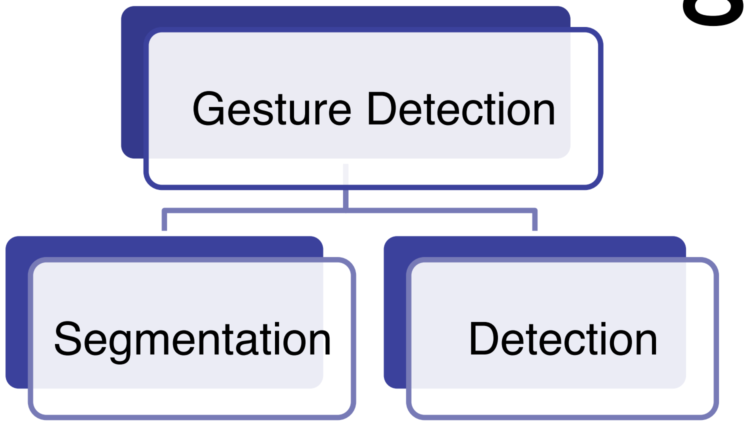
Retrieving similar gestures to the examples

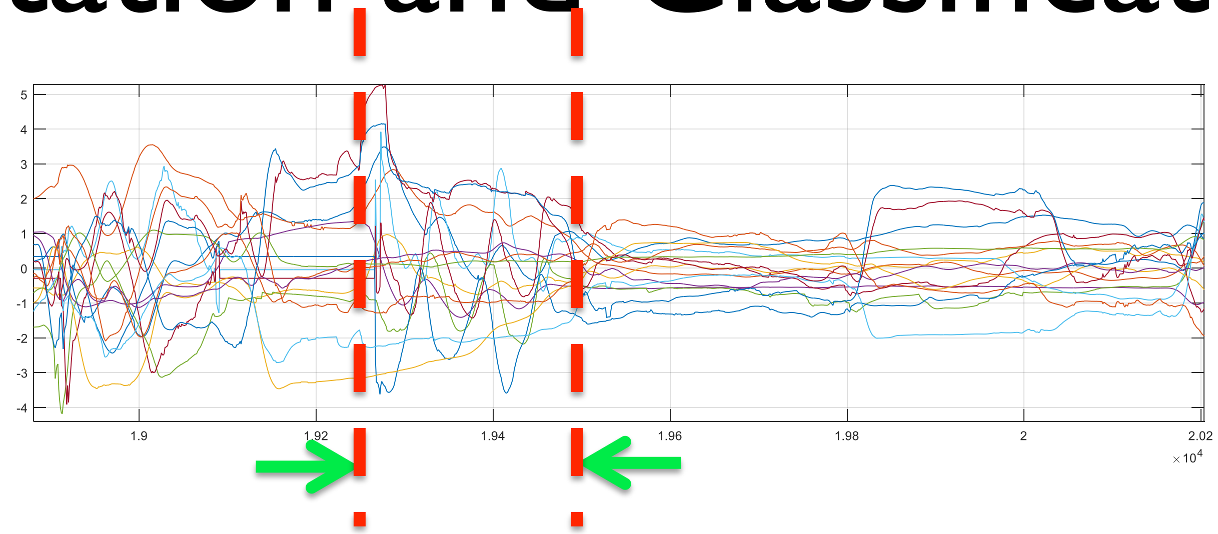
Training the Behavior Realization model



Goal: Retrieve examples of prototypical gestures

Gesture Segmentation and Classification





- Kovar et al. [2004]
  - Find gestures similar to a target gesture using DTW and use retrieved samples to expand the training samples
- Joshi et al. [2015]
  - Train a random forest model using video and depth map of the joints
  - They use a multi-scale window sliding for new data (forward search).
- Zhou et al. [2013]
  - Hierarchical aligned cluster analysis (HACA) to dynamically segment and cluster motion capture data into movement primitives

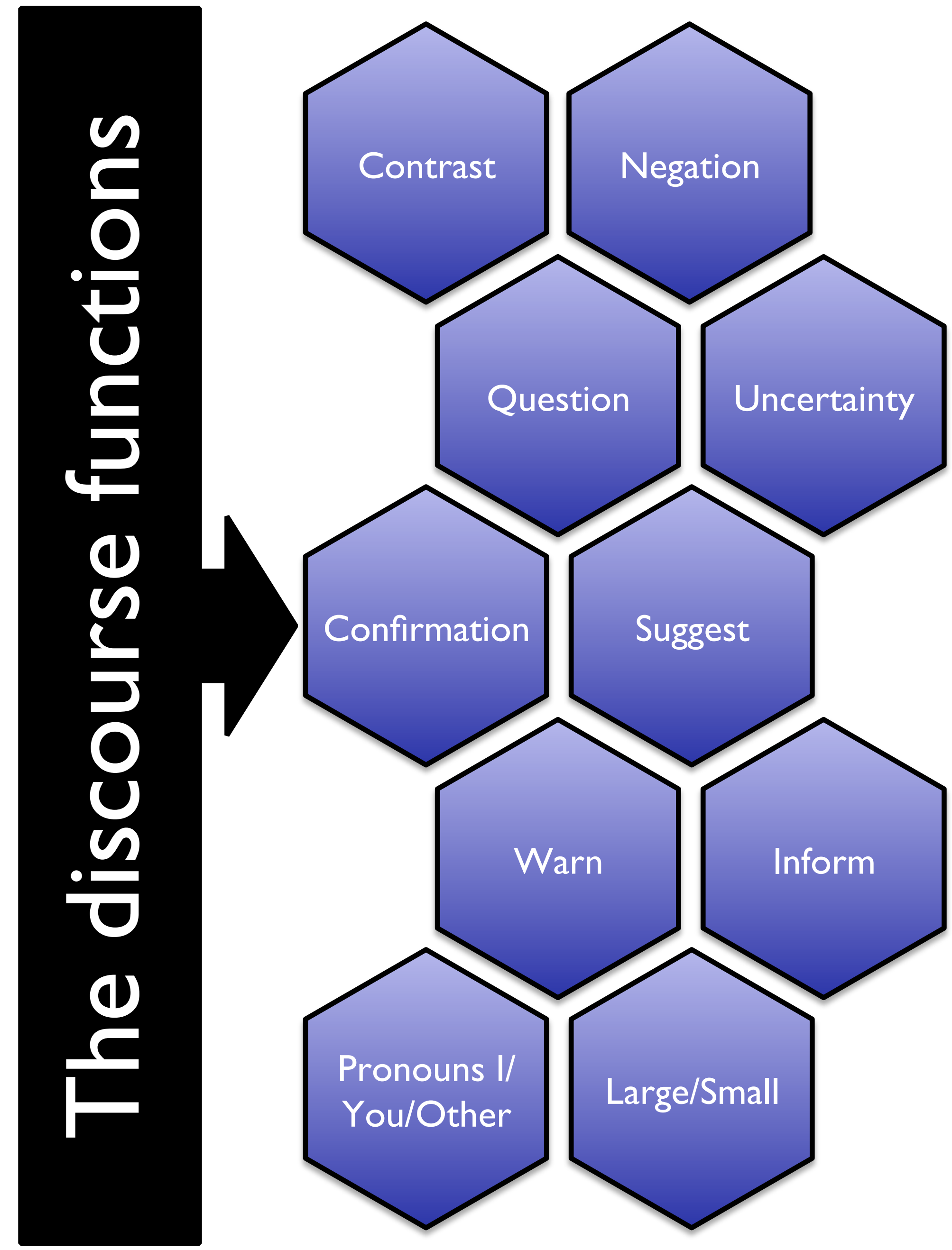
MSP-AVATAR Corpus

- Multimodal database comprising:
  - Motion capture data
  - Video camera
  - Speech recordings

- Four dyadic interaction between actors
  - We motion captured one of the actors
- Database rich in terms of discourse functions

Discourse Functions in MSP-AVATAR corpus

- Discourse functions that elicit specific gestural behaviors
- Selection guided by previous studies
  - Poggi et al [2005]
  - Marsella et al. [2013]
- 2-5 scenarios per discourse function
- We used the recordings from one of the actors (66 mins)



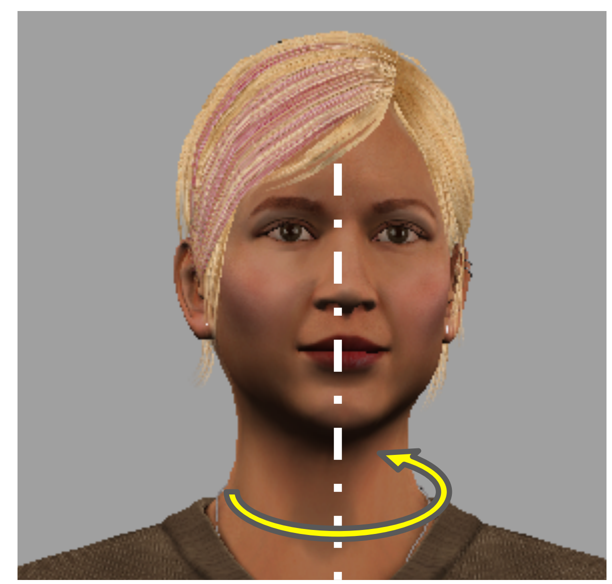
Prototypical Behaviors

So-What

To-Fro

Regress

Nods

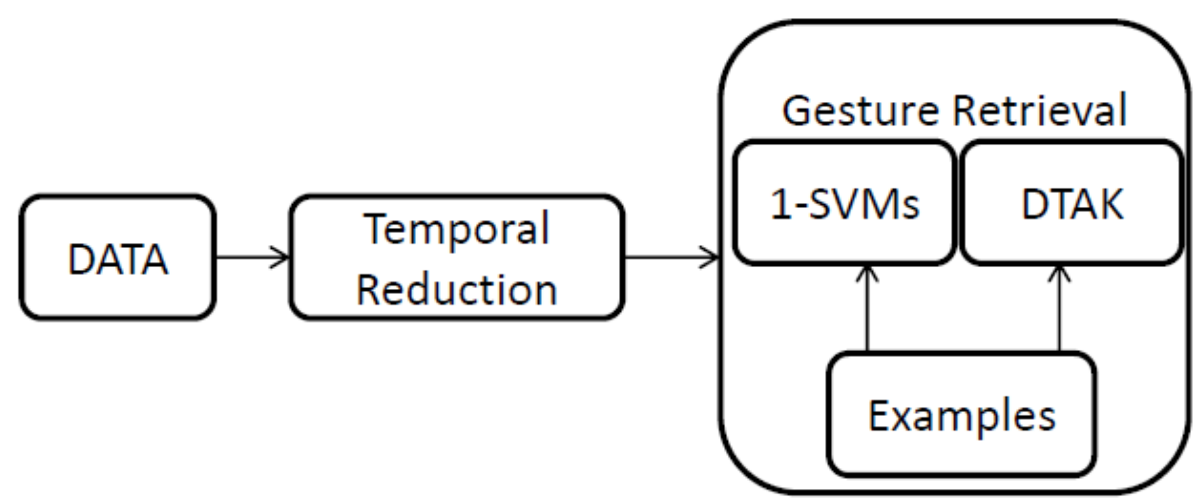


Shakes

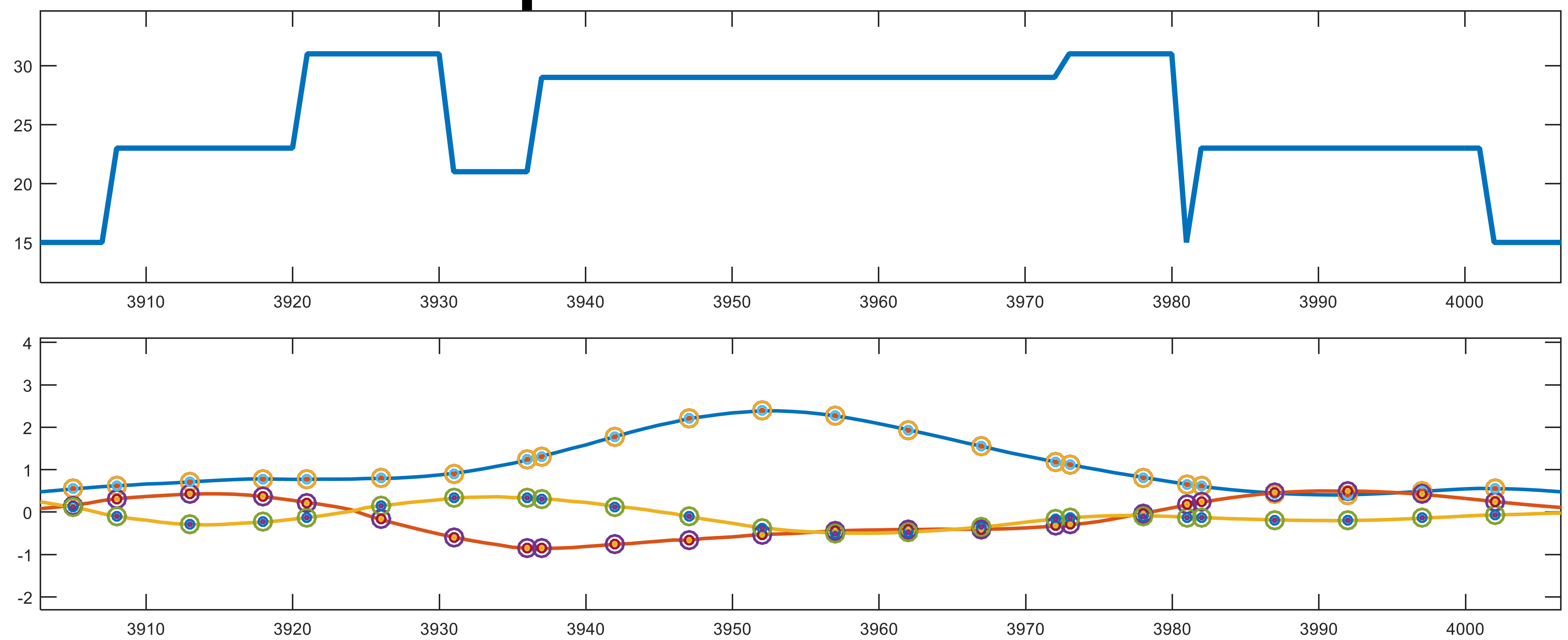
	So-What	To-Fro	Regress	Nods	Shakes
$Samples_{train}$	14	27	26	24	27
$Samples_{test\&developing}$	21	29	73	138	115

Gesture Retrieval Framework Overview

- Temporal reduction
  - The data is captured by 120 fps, and may have redundant information
- Gesture segmentation
  - Gestures can happen with arbitrary durations
- Gesture detection
  - Binary decision per segment



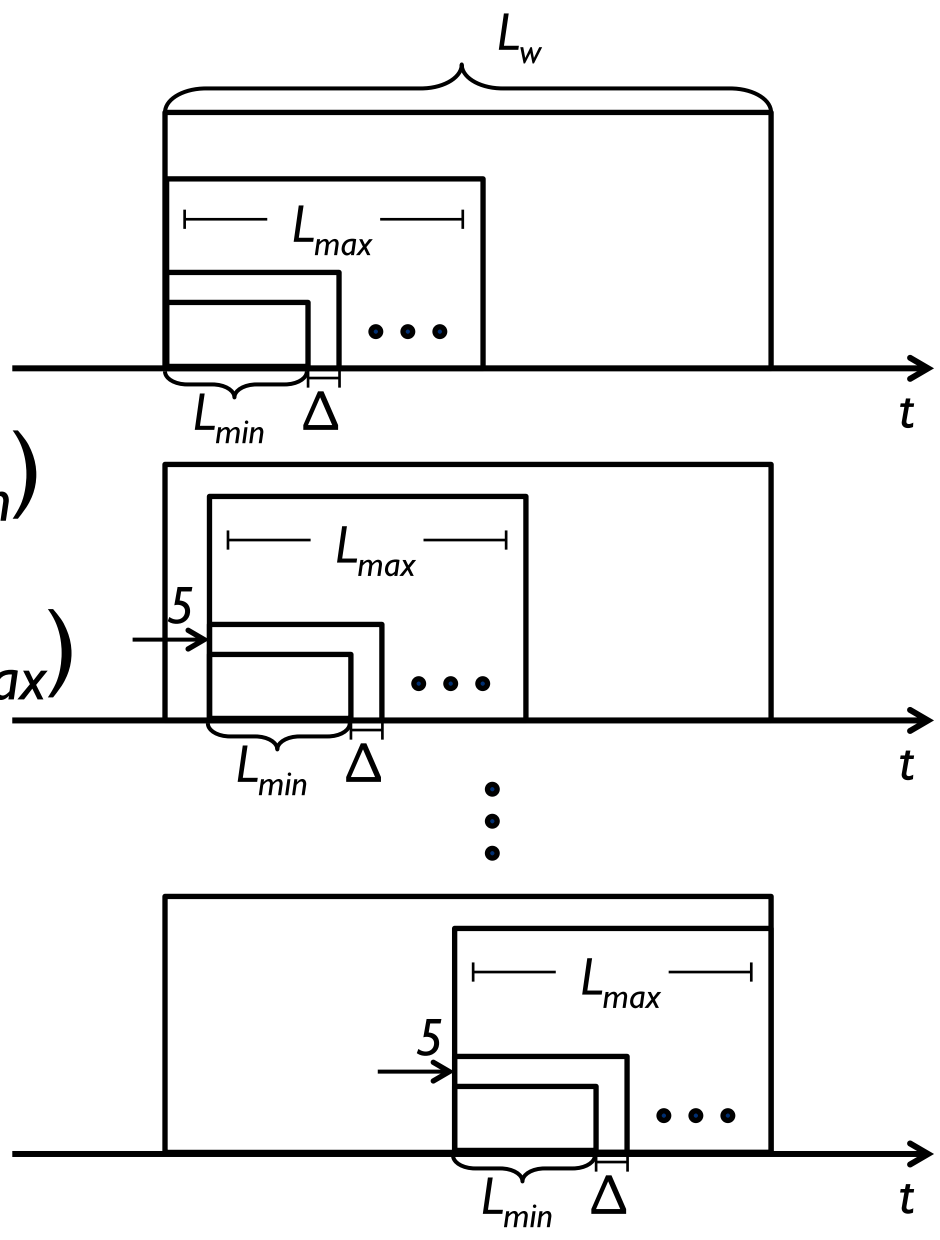
Temporal Reduction



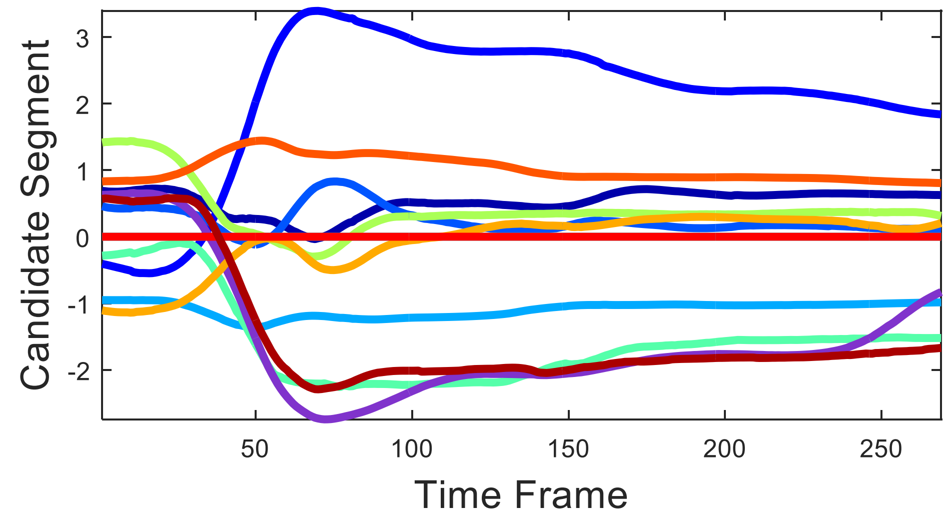
- Reduce the complexity of the system
  - Inspired by Zhou et al. [2013]
- Non-uniform downsampling
  - Based on Linde-Buzo-Gray vector quantization (LBG-VQ)
  - Discard consecutive frames up to 5 frames if they are in the same cluster

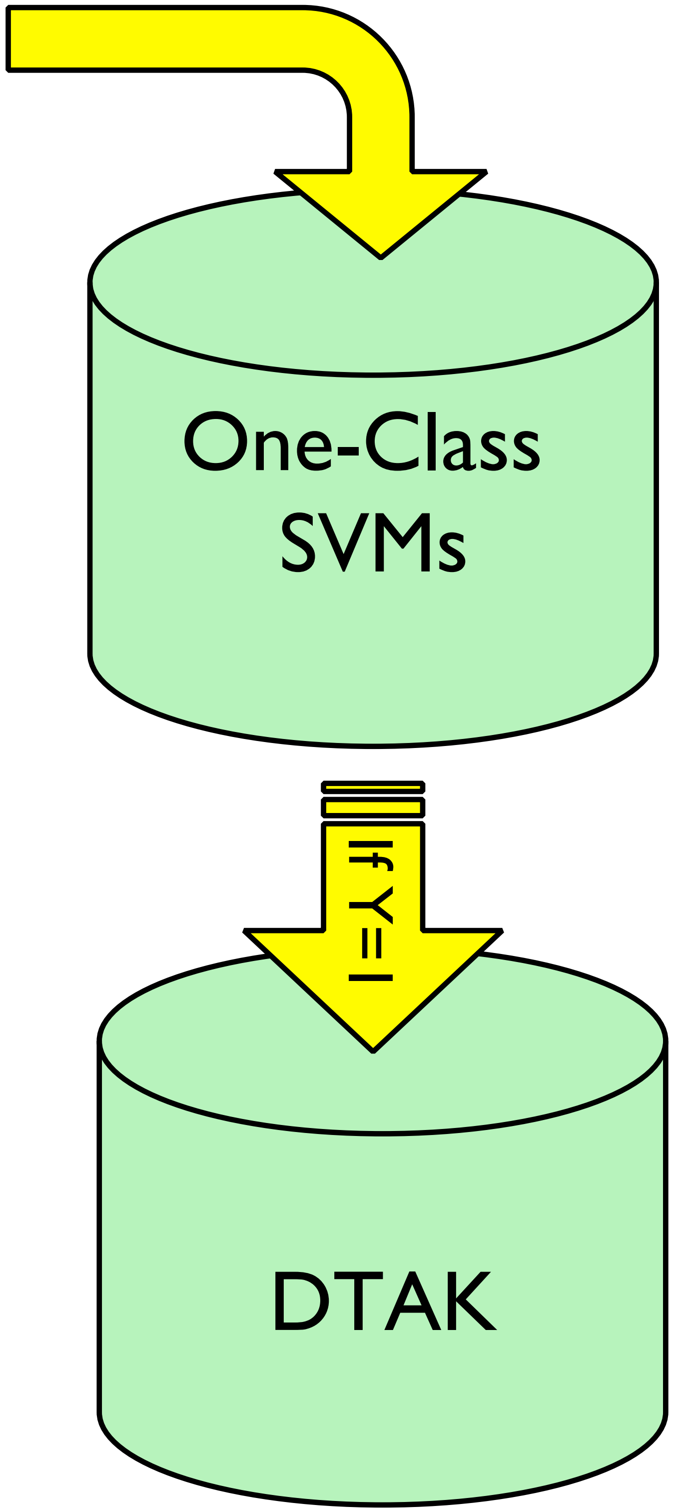
Gesture Segmentation

- Window size (L_w)
- Minimum length of search segment (L_{min})
- Maximum length of search segment (L_{max})
- Increment frames between iterations
  - $\Delta = (L_{max} - L_{min})/30$
- One winner per window



Gesture Detection

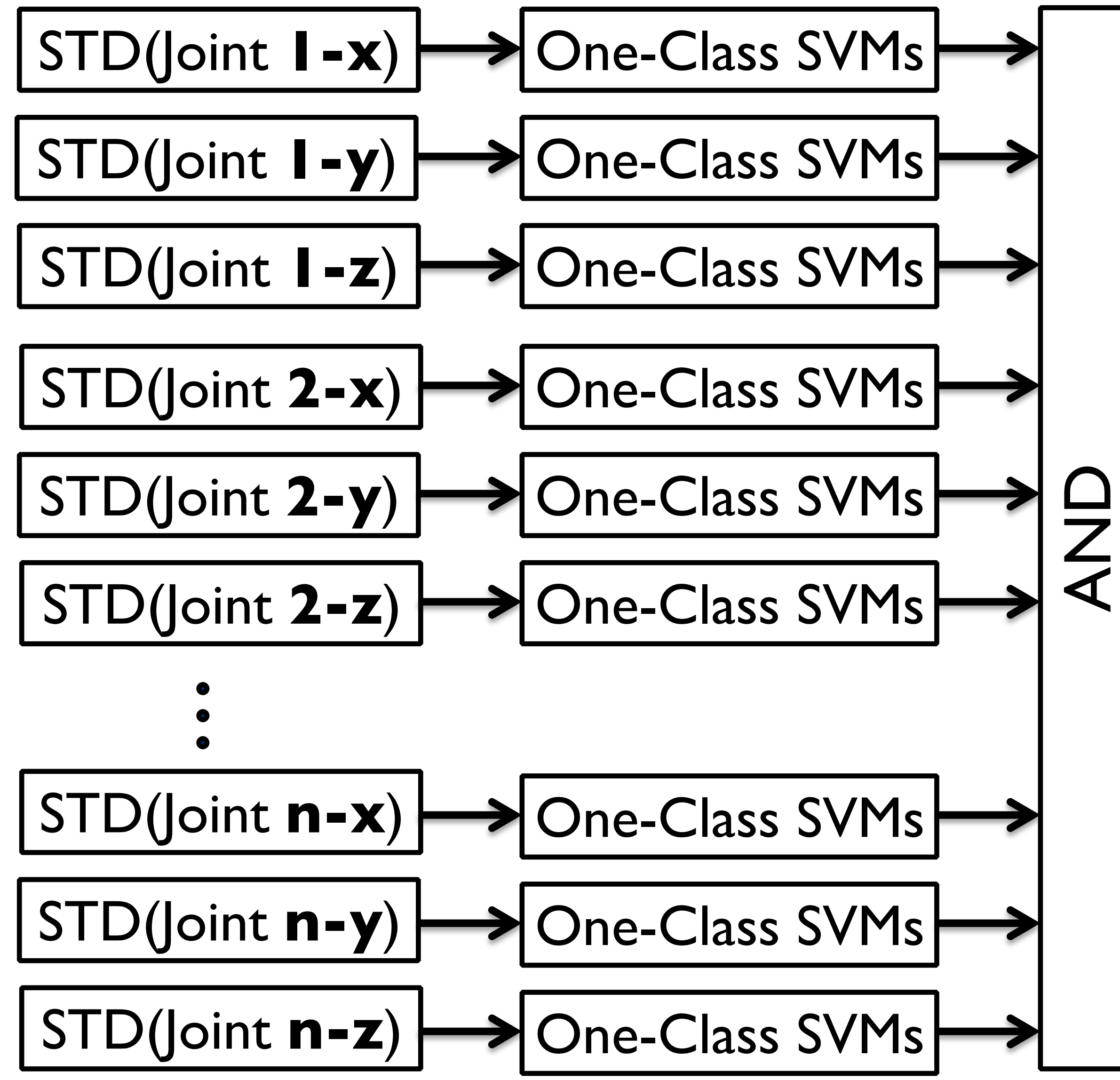




- One-class SVMs
  - Efficiently reduce the number of candidates
- Dynamic time alignment kernel (DTAK)
  - To increase precision

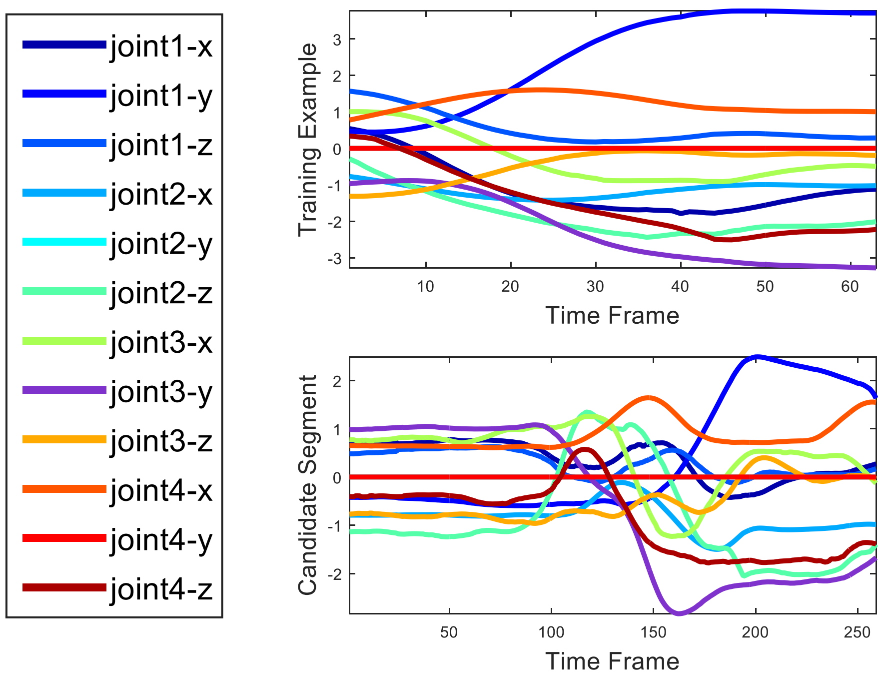
One-Class SVMs

- Only positive samples
- Limited number of training instances
  - Train separately for different features
  - Fuse the classifiers using the AND operator
- Feature selection by cross-validation
  - Sort features according to accuracy
  - Remove one by one to get accuracy>0.85



DTAK by Zhou et al. [2013]

- DTAK finds similarity between two segments regardless of their length in term of a kernel (Gaussian)



$$K_{i,j} = \exp\left(- \frac{\left\| x_i - y_j \right\|^2}{2\sigma^2} \right)$$

$$\tau(X,Y) = \frac{u_{l_x, l_y}}{l_x + l_y}, u_{i,j} = \max \begin{cases} u_{i-1,j} + K_{i,j} \\ u_{i-1,j-1} + 2K_{i,j} \\ u_{i,j-1} + K_{i,j} \end{cases}$$

- Final score: the median of the similarity measure to the training examples

- Find a threshold by maximizing the F-score on the developing set

Evaluation of Retrieved Gestures

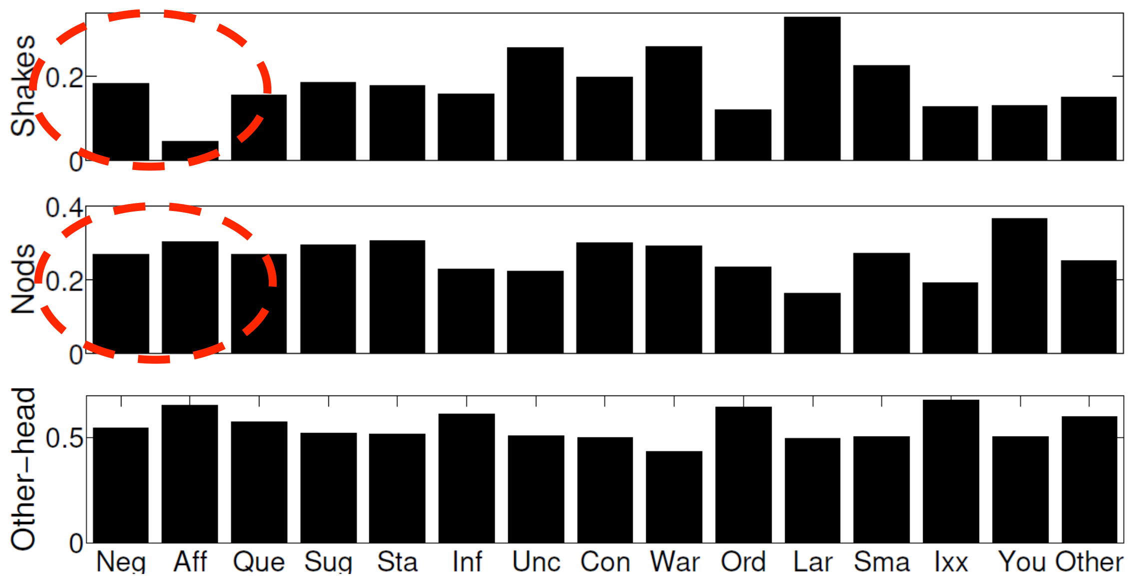
- Precision in head gestures > 0.85
- Precision in hand gestures > 0.59
- Head vs. hand gestures:
  - Less complex

Gesture	19 Sessions
	Precision [%]
Head Shake	**91.32**
Head Nod	**85.04**
To-Fro	**59.52**
So-What	**76.68**
Regress	**71.77**

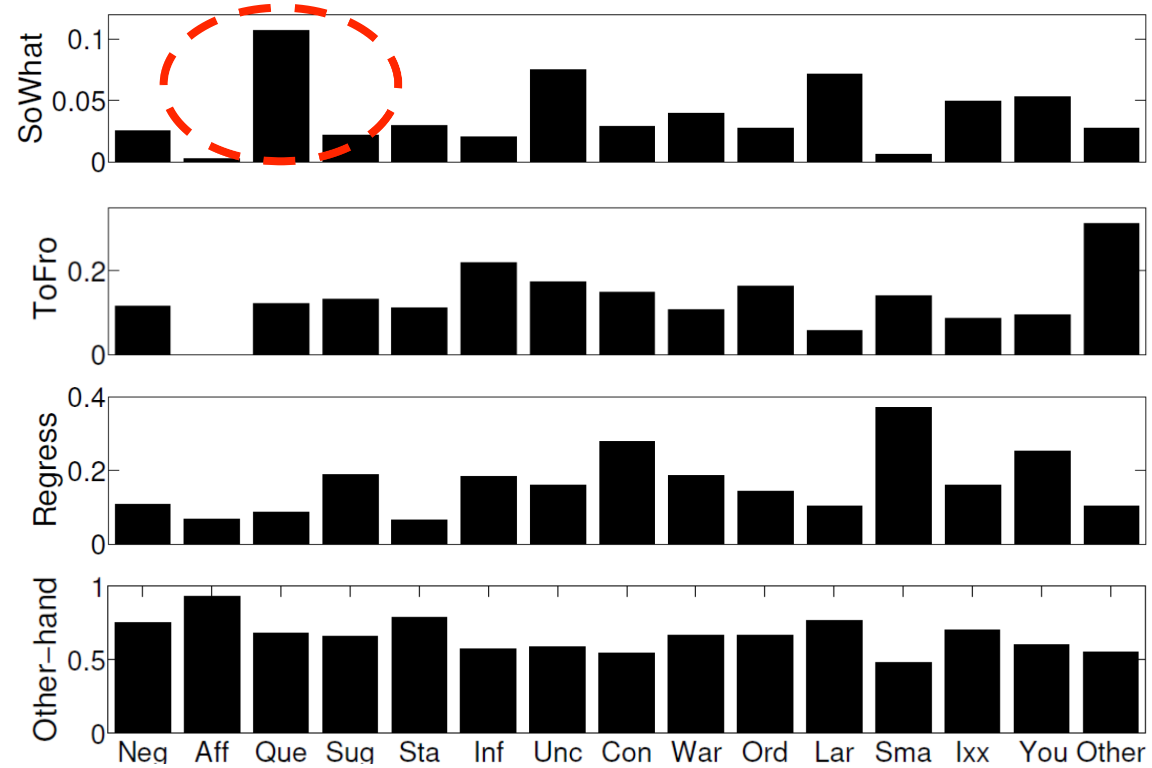
Gesture	Test & Developing Sessions	
	Precision [%]	Recall [%]
Head Shake	**95.65**	**42.31**
Head Nod	**87.10**	**61.36**
To-Fro	**67.86**	**67.86**
So-What	**76.92**	**47.62**
Regress	**78.85**	**57.75**

Analysis of Gestures vs. Discourse Functions

HEAD



HAND



- The histograms of the discourse functions vs. behaviors
  - Different gestures appear with different frequencies across different discourse functions
  - Shakes happen in Negation more than in Affirmation
  - Nods happen in Affirmation more than in Negation
  - So-What happens more in Question than other discourse functions

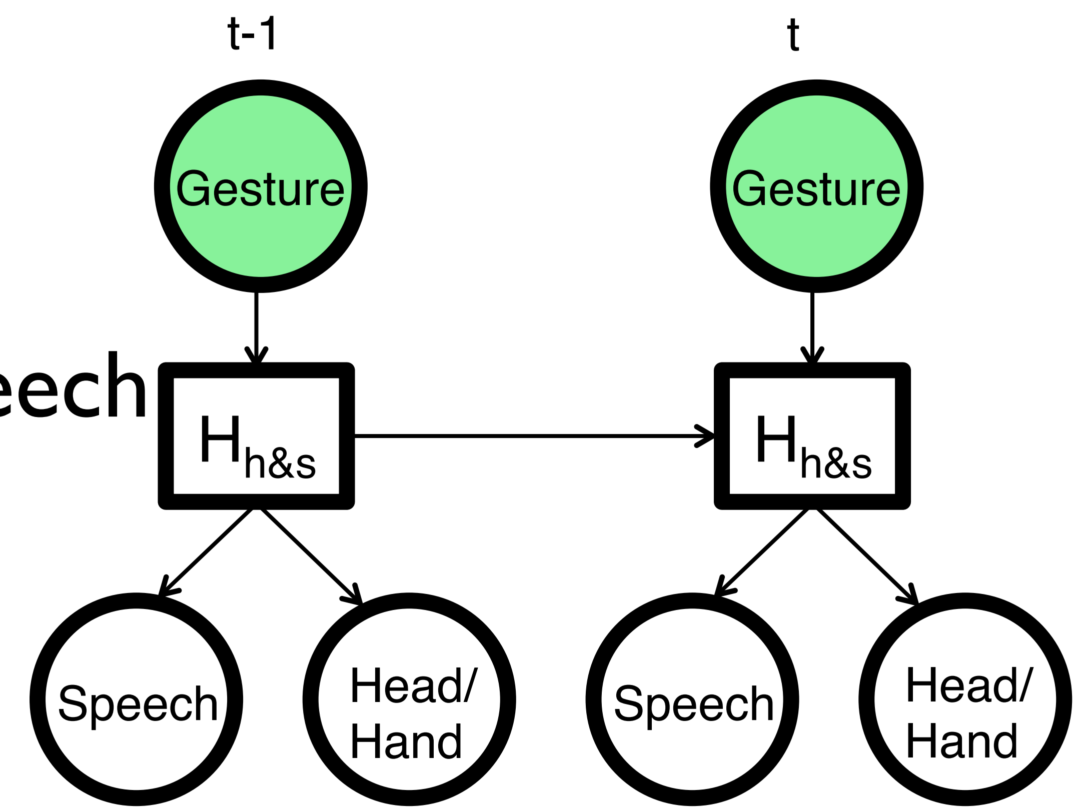
Modeling the gestures

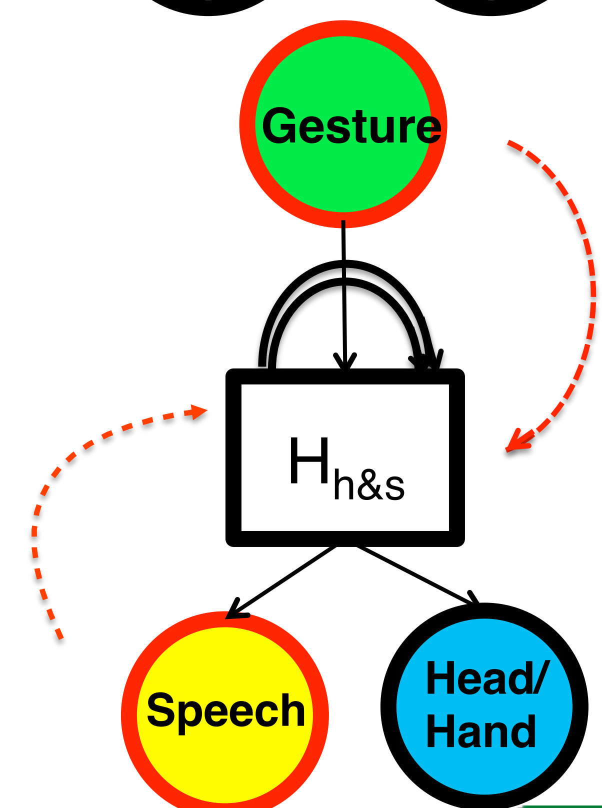
- Gesture retrieval → more samples to train the models
- Assumptions
  - Target gesture is known
  - Speech prosody features are known
- How to model the gesture?
  - Speech driven models
  - Training: **speech prosody features**, **motion capture data**, and **prototypical gesture**
  - Testing (synthesis): **speech prosody features**, and **prototypical gesture**

Gesture	#Retrieved
Head Shake	287
Head Nod	535
To-Fro	223
So-What	114
Regress	262

Speech driven animation

- Dynamic Bayesian Network
- Shared hidden variable between speech and head/hand
- Constrained on gestures
- Add the constraint node as parent of the hidden state:
  - More robust to unbalanced data
  - Learns separately:
    - Prior probabilities of the gestures
    - The affect of gestures on transition matrices





HEAD Synthesis

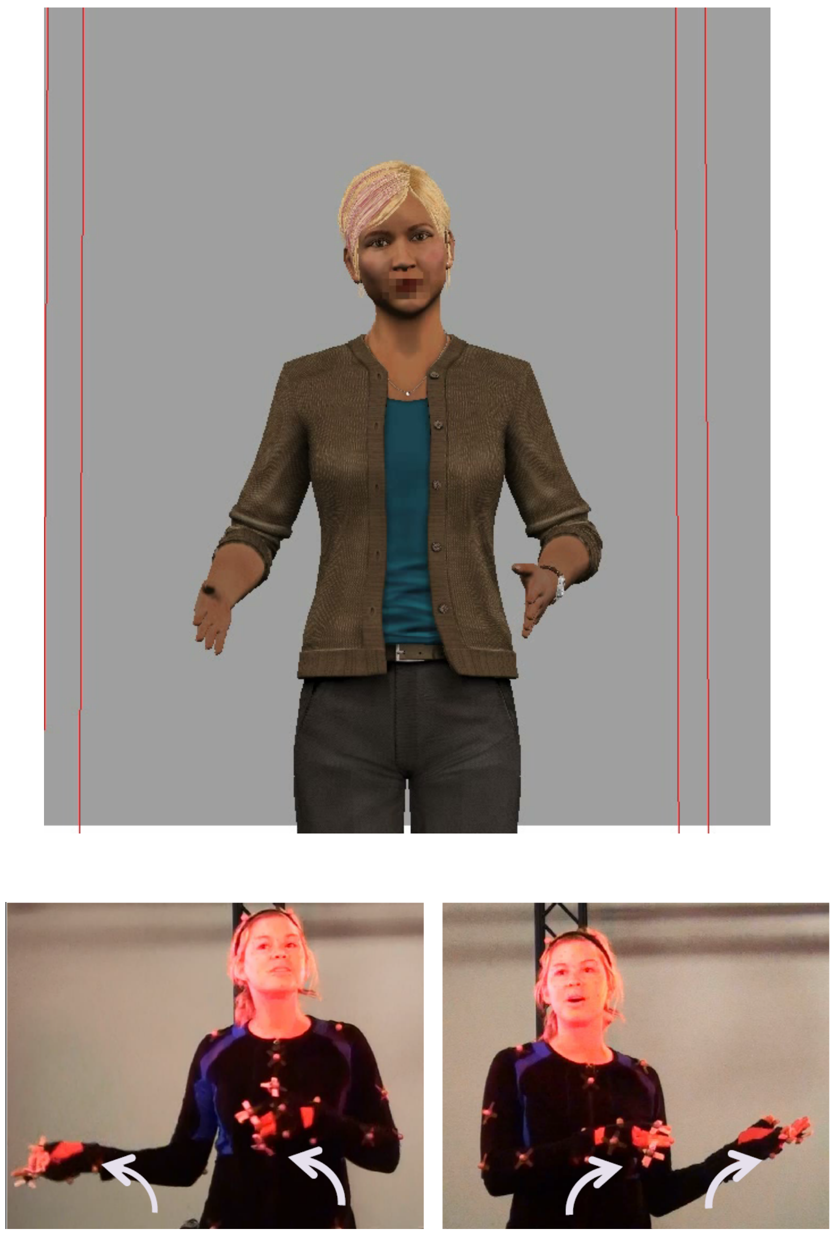
For illustration gesture is always “on”

Nods

Shakes

HAND Synthesis

For illustration gesture is always "on"



To-Fro

So-What

Regress

Conclusions

- This paper proposed a framework to automatically detect target gestures
  - Using few examples in a motion capture database
  - The advantage of this framework is its flexibility to retrieve any gesture
- The approach jointly solved the segmentation and detection of gestures
  - Multi scale windows
  - Two-step detection framework
- We used the retrieved samples to synthesize novel realizations of these gestures
  - Speech-driven animations constrained by these target behaviors

Future Work

- Explore the minimum number of examples per gesture to achieve acceptable detection rates
- Using adaptation to generalize the models to retrieve similar gestures from different subjects
    - With more data, more restrictive threshold can be considered
- Explore the effects of detection errors on the performance of the speech driven models

Multimodal Signal Processing (MSP)

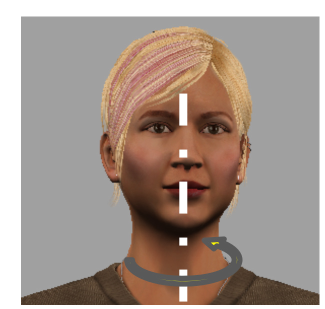
- Questions?

http://msp.utdallas.edu/

HEAD Synthesis

For illustration gesture is always "on"

Nods



Shakes

HAND Synthesis

For illustration gesture is always “on”

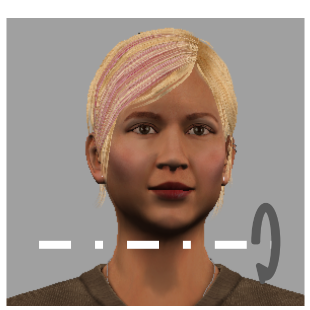
To-Fro

So-What

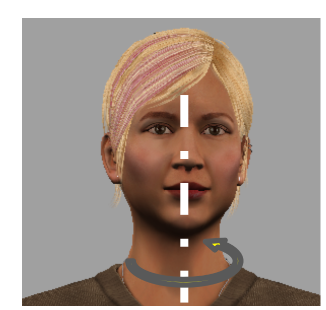
Regress

HEAD Synthesis

For illustration gesture is always "on"



Nods



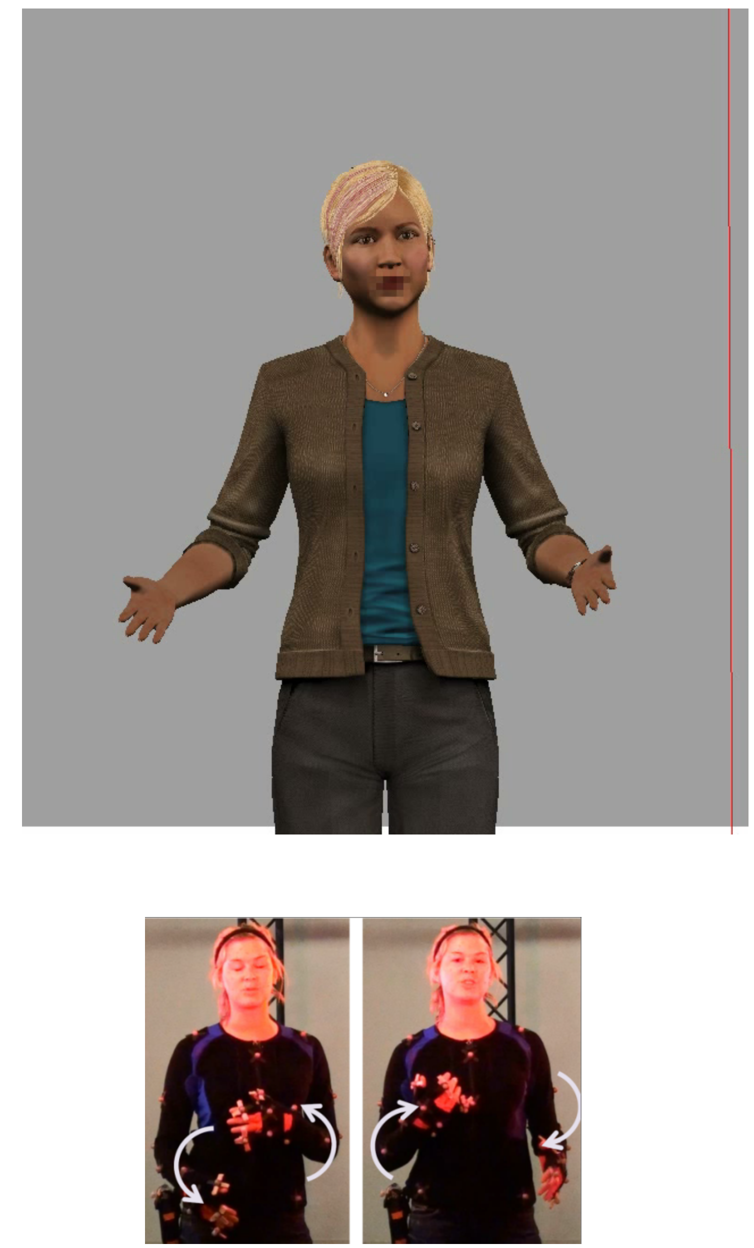
Shakes

HAND Synthesis

For illustration gesture is always “on”

To-Fro

So-What



Regress