Gesture
in Automatic Discourse Processing

Jacob Eisenstein

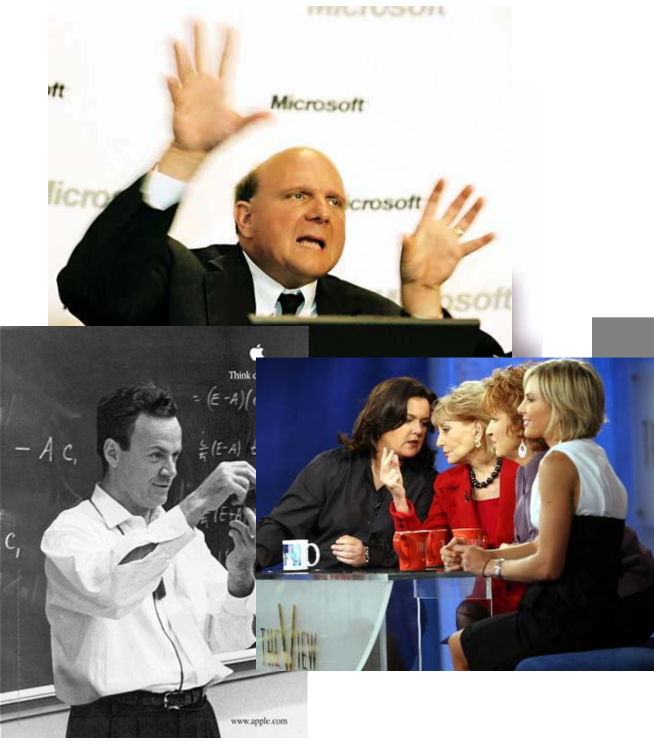
Supervised by

Regina Barzilay

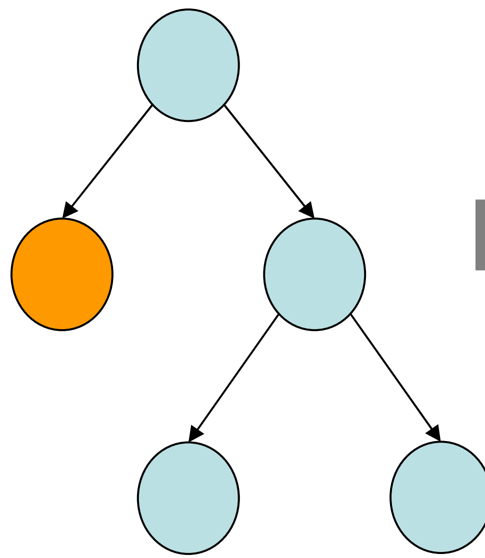
Randall Davis

natural language processing

natural language



representation

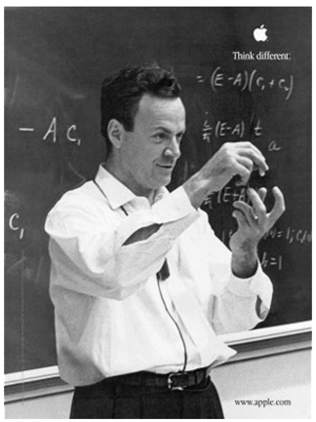


applications

$

Speech is accompanied by visual communication.

Especially: hand gesture

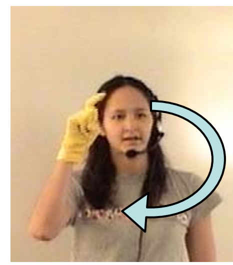


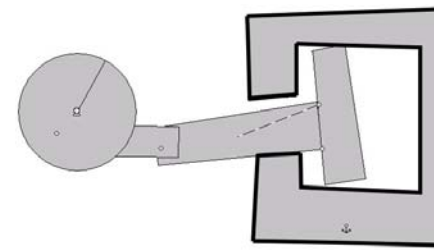
$\neq$

gesture: why should we care?

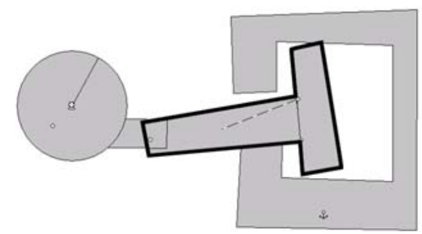
Gestural form reflects the underlying meaning.

"Think of the block letter C"

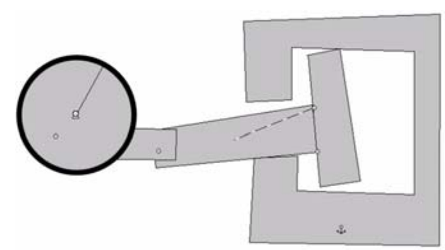




"Then there's a T-shaped thing."



"There's a wheel over here."

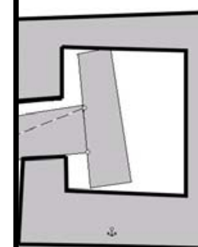


gesture: why should we care?

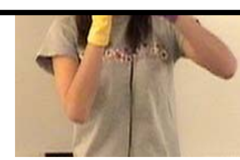
Gesture can be crucial to understanding.

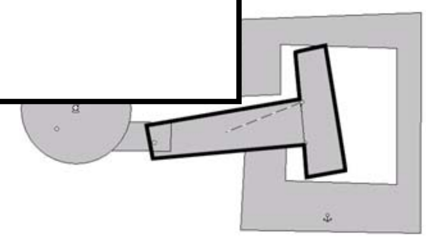
“Think of th
comes li

Can we use gestures for automatic natural language processing?

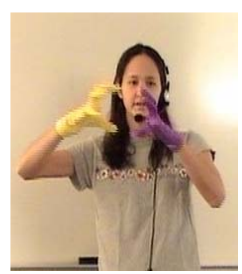


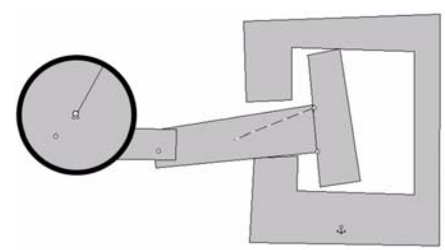
“Then there’s
thing…if t
like this.”





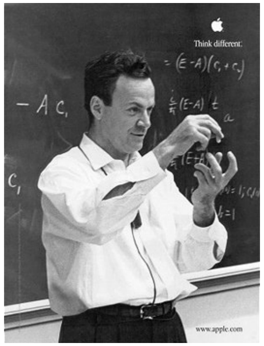
“There’s a wheel over here.”





challenges for gesture in nlp

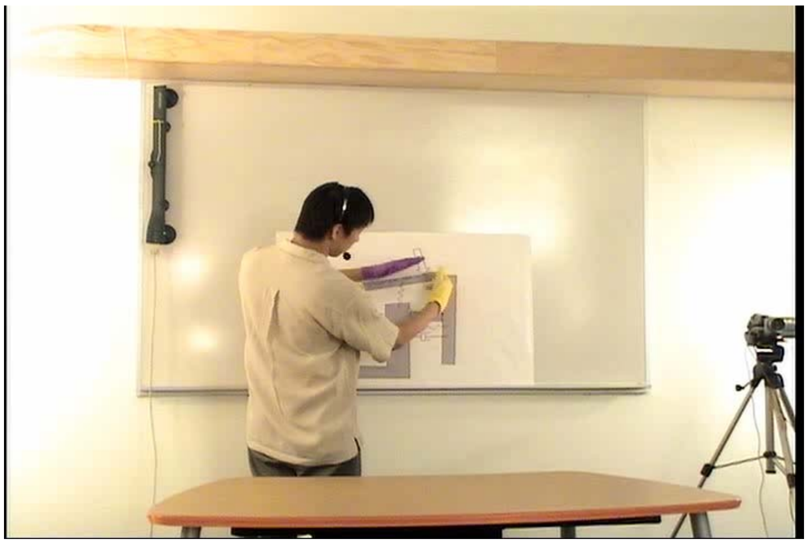
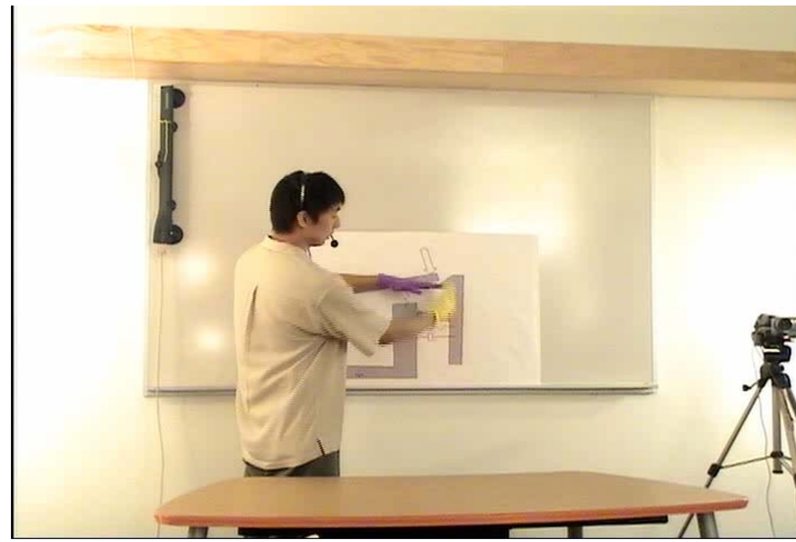
- Gesture interpretation depends on linguistic context.



- No adequate representation of individual gestures
- Raw signal → discourse analysis too difficult

patterns of gestures

- *Problem 1:* cospeech gestures are generally unstructured.

“**This thing** cli d clicks back…”

patterns in gesture predict patterns in language

Without “re they can still tell us something.

learning gestural representations

- *Problem 2:* representation of gestural form
- We want to compute gestural patterns, such as similarity.

Learn about gesture in the context of linguistic tasks.

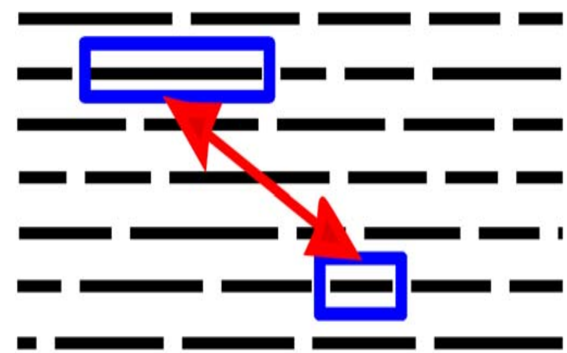
Linguistic annotation, not gestural annotation

contributions

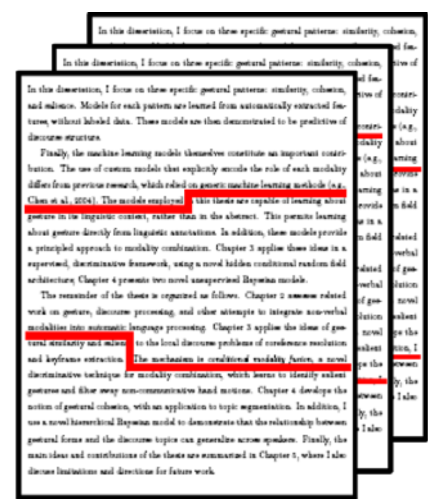
- Gesture improves discourse interpretation.
- Methods
  - Gesture patterns, not gesture recognition!
  - Key gestural properties: similarity, cohesion, and salience
  - Structured models for combining gesture, speech, and meaning

outline

- Local discourse structure



- Global discourse structure



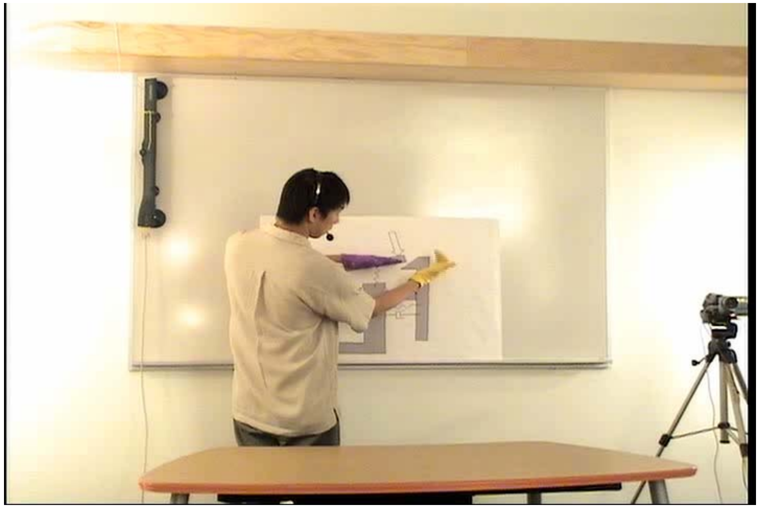
noun phrase coreference

"As this bar comes all the way down, this thing clicks back.

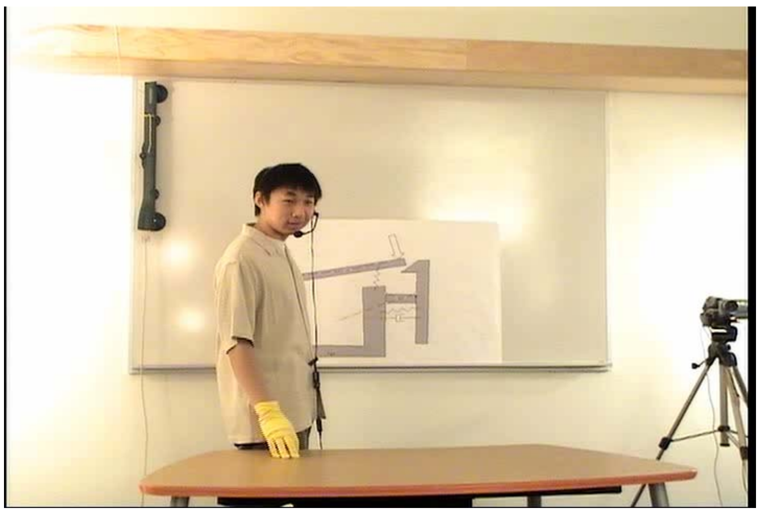
…

That happens three times during the video, so this comes down, it clicks over. And then I think the video resets or something, but it's restored back to this state, and then it comes down again, this thing goes out and clicks back."

noun phrase coreference

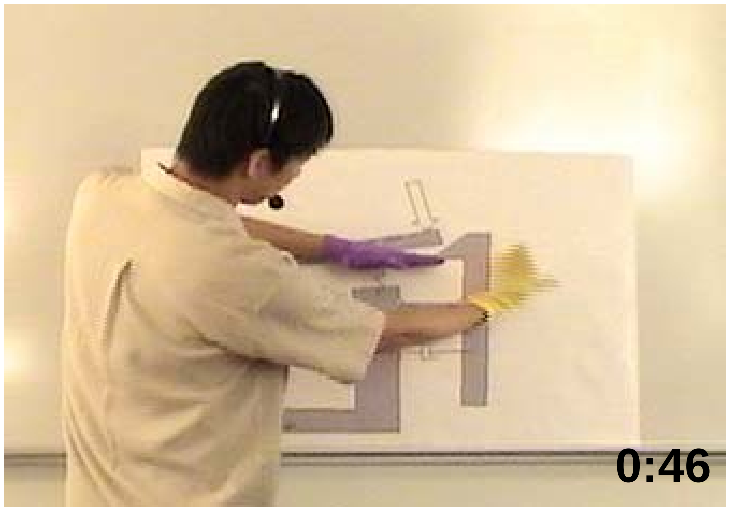


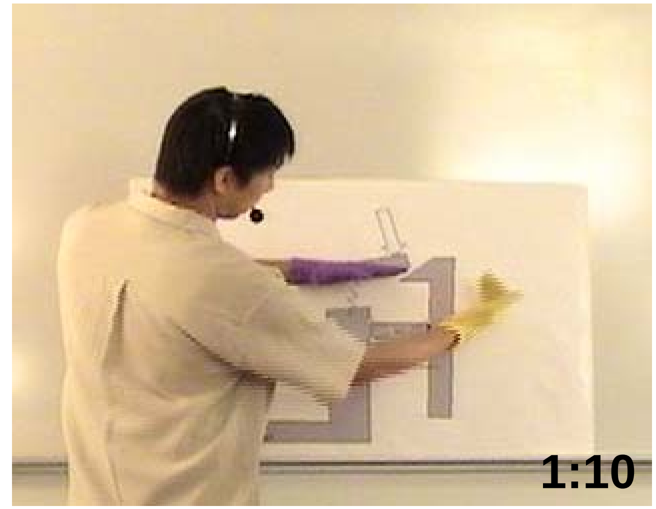
"As this bar comes all the way down, this thing clicks back…



That happens three times during the video, so this comes down, it clicks over. And then I think the video resets or something, but it's restored back to this state, and then it comes down again, this thing goes out and clicks back."

an example



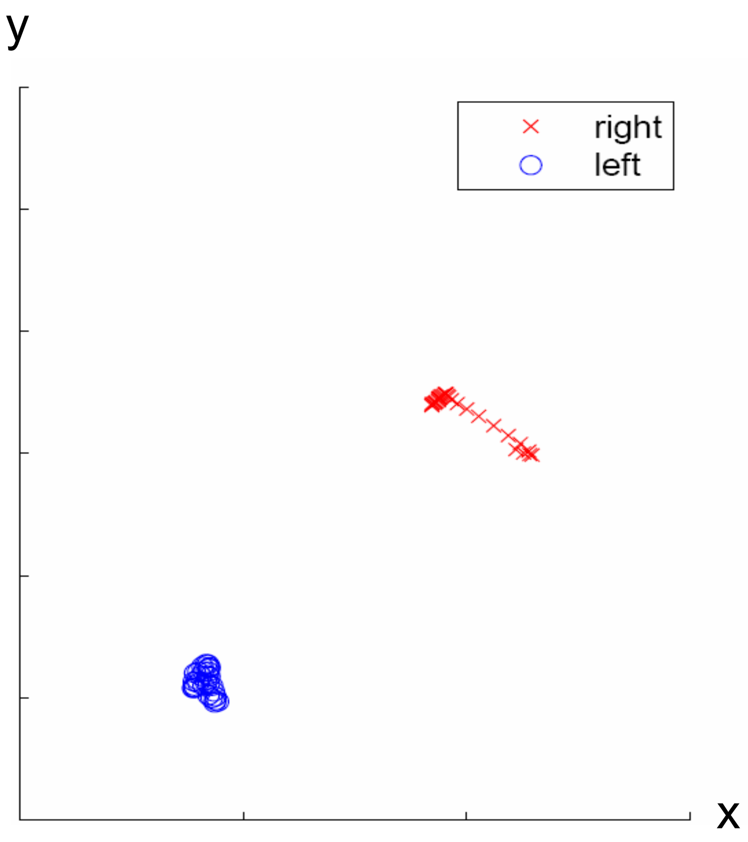


"**This thing** clicks back."

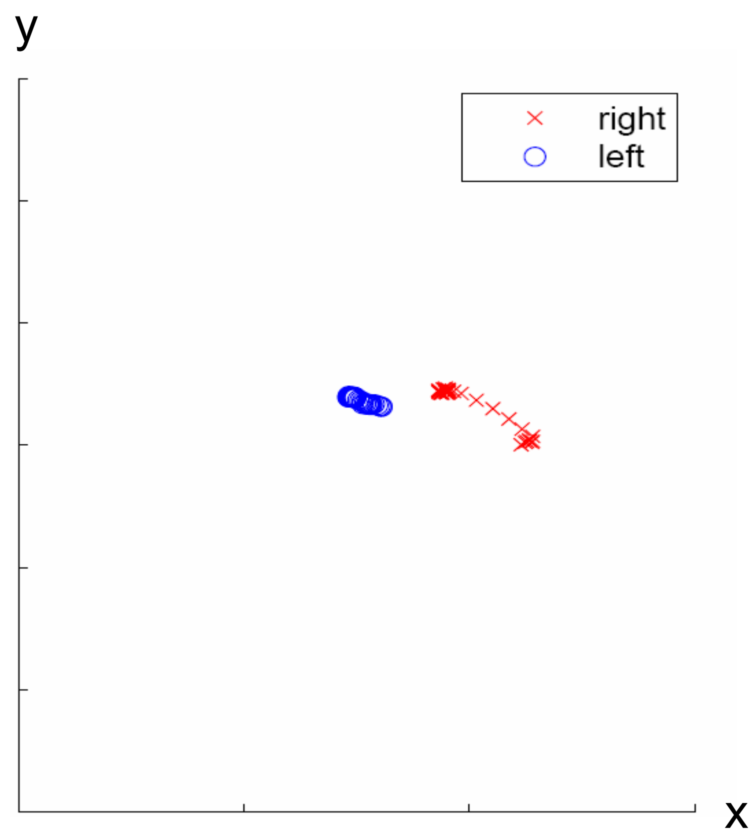
…

"**it** clicks over…"

hand trajectories

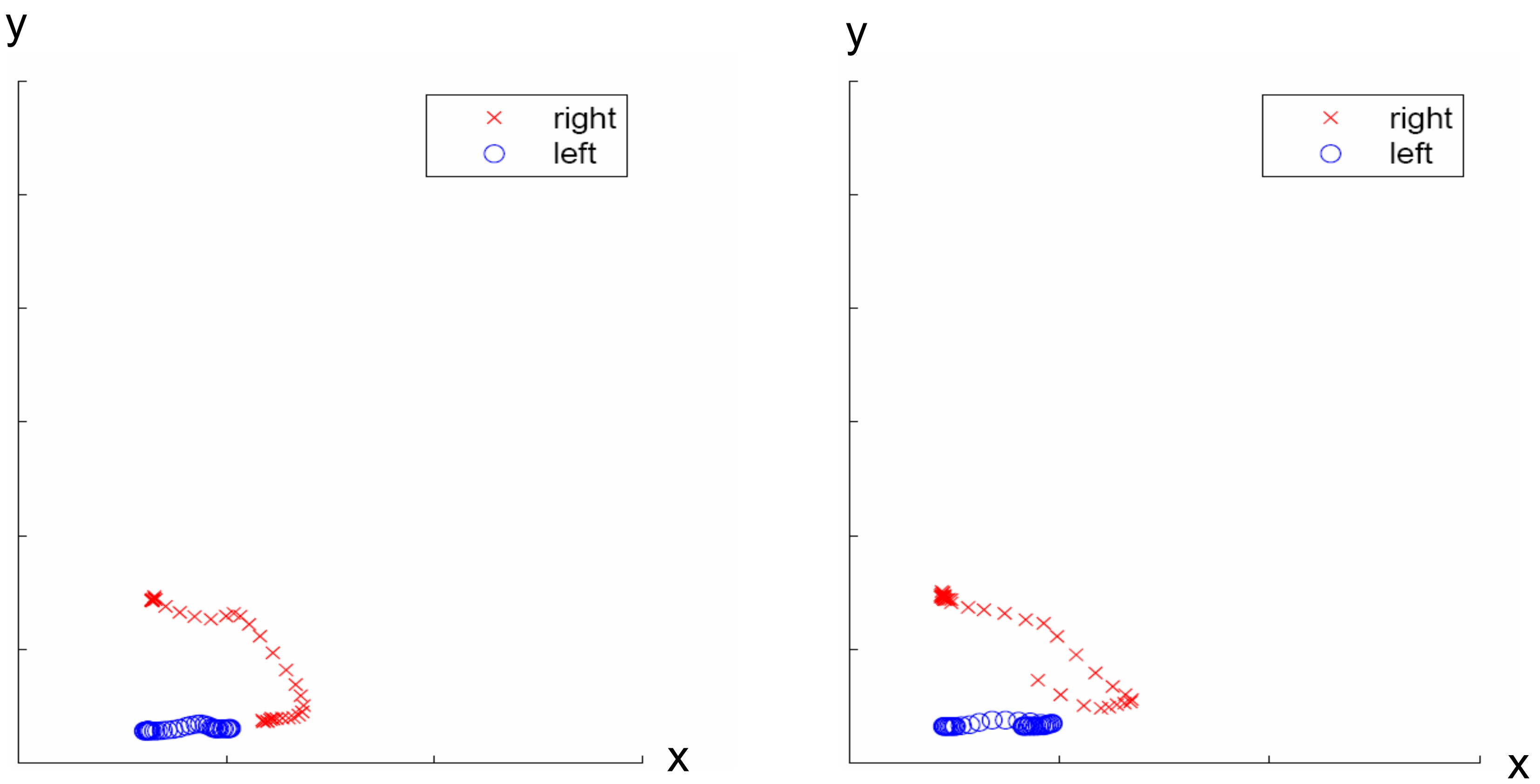


“This thing”



“it”

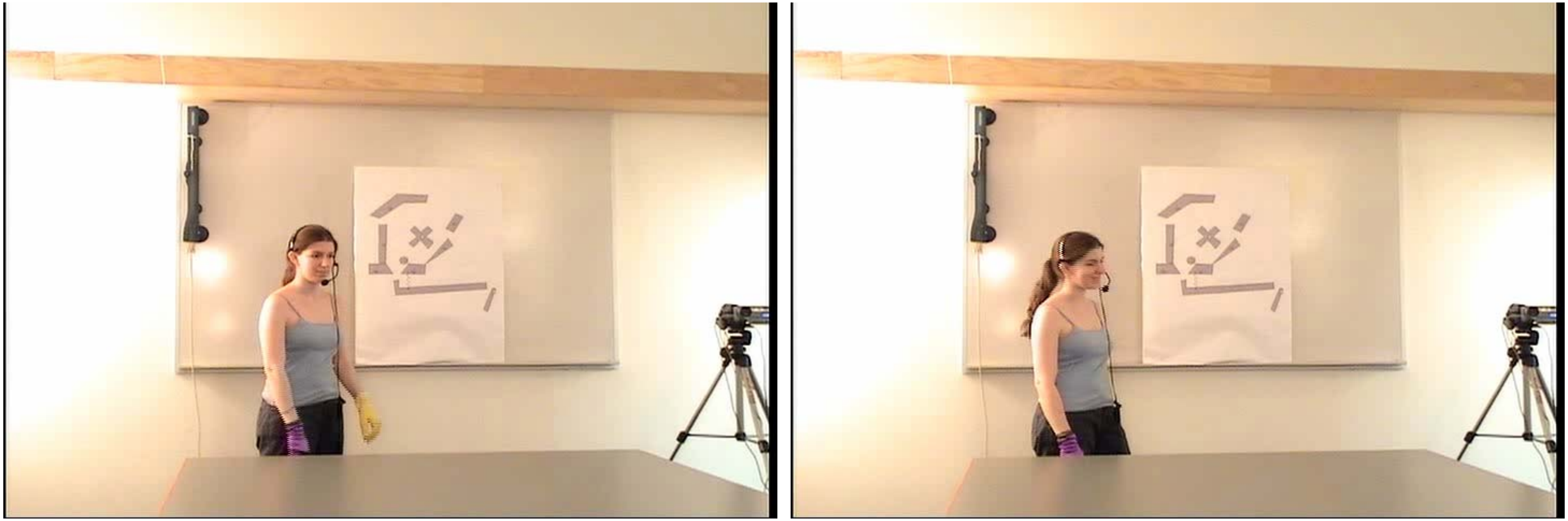
another example



another example

~~"Similar gestures imply similar content"~~

Similar *meaningful* gestures imply similar content.



gestural salience

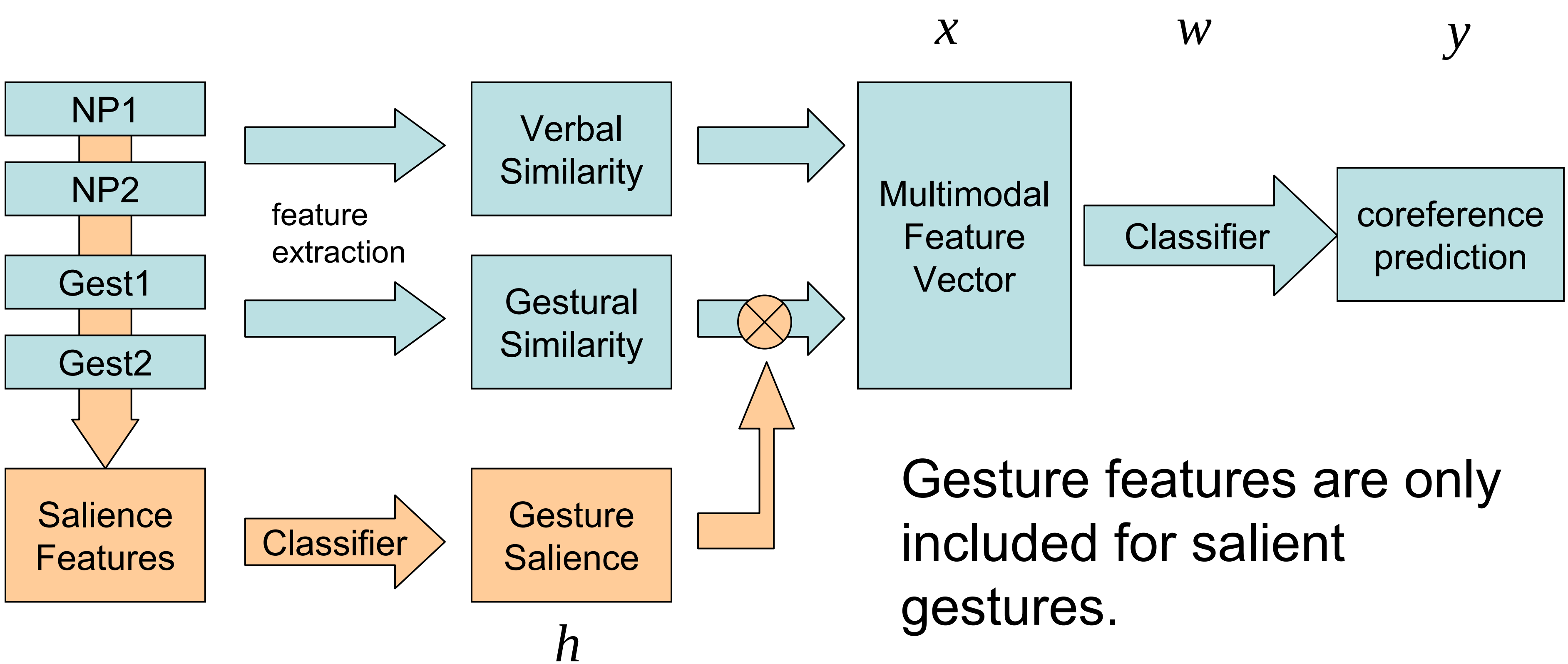
- Viewers distinguish communicative gestures from other hand movements consistently and robustly (Kendon 1978).

Can we train a computer to do the same thing?

Can we do it without labeled data?

Kendon, "Differential Perception and attentional frame: two problems for investigation." Semiotica, 24 (1978): 305-315.

conditional modality fusion



Gesture features are only included for salient gestures.

Salience is learned jointly with coreference.

conditional modality fusion

$$p(y|\mathbf{x};\mathbf{w}) = \sum_{\mathbf{h}} p(y,\mathbf{h}|\mathbf{x};\mathbf{w})$$

$$= \frac{\sum_{\mathbf{h}} exp\{\psi(y,\mathbf{h},\mathbf{x};\mathbf{w})\}}{\sum_{y',\mathbf{h}} exp\{\psi(y',\mathbf{h},\mathbf{x};\mathbf{w})\}}$$

- y – coreference label
- $\mathbf{h}$ – gesture salience
- $\mathbf{x}$ – observed features
- $\mathbf{w}$ – learned weights
- ψ – potential function

Quattoni, Collins and Darrell, “Conditional Random Fields for Object Recognition.” NIPS 2004.

potential function

$$\psi(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) = \overset{\text{verbal similarity}}{\boxed{\psi(y, \mathbf{x}_v)}} + \overset{\text{gesture salience}}{\boxed{\psi(\mathbf{h}, \mathbf{x}_h)}} + \overset{\text{gesture similarity}}{\boxed{\psi(y, \mathbf{x}_g, \mathbf{h})}}$$

$$\psi(y, \mathbf{x}_v) = y\mathbf{w}_v^T \mathbf{x}_v$$

$$\psi(\mathbf{h}, \mathbf{x}_h) = h_1 \mathbf{w}_h^T \mathbf{x}_{h_1} + h_2 \mathbf{w}_h^T \mathbf{x}_{h_2}$$

$$\psi(y, \mathbf{x}_g, \mathbf{h}) = \begin{cases} y\mathbf{w}_g^T \mathbf{x}_g, & h_1 = h_2 = 1 \\ 0, & \text{otherwise.} \end{cases}$$

learning salience

$$\text{sign}(y) \neq \text{sign}(\mathbf{w}_g^T \mathbf{x}_g) \rightarrow y\mathbf{w}_g^T \mathbf{x}_g < 0$$

$$\text{sign}(y) = \text{sign}(\mathbf{w}_g^T \mathbf{x}_g) \rightarrow y\mathbf{w}_g^T \mathbf{x}_g > 0$$

$$\psi(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) = \psi(y, \mathbf{x}_v) + \psi(\mathbf{h}, \mathbf{x}_h) + \psi(y, \mathbf{x}_g, \mathbf{h})$$

$$\psi(y, \mathbf{x}_g, \mathbf{h}) = \begin{cases} \boxed{y\mathbf{w}_g^T \mathbf{x}_g, \quad h_1 = h_2 = 1} \\ \boxed{0, \qquad\qquad \text{otherwise.}} \end{cases}$$

dataset

- Spoken, spontaneous explanations of the behavior of mechanical devices
- Visual aids are provided
- 16 videos, nine speakers
- Data processing
  - Automatic detection of hand position and velocities in an articulated model
  - Manual transcriptions of speech

results

Evaluation in Area Under ROC Curve (AUC)

Model	AUC
Verbal only	.7945
Gesture only	.6732

results

Evaluation in Area Under ROC Curve (AUC)

Model	AUC
Verbal + All Gestures	.8109
Verbal only	.7945
Gesture only	.6732

+ 1.6%

multimodal beats verbal only: $t(15) = 4.45$, $p < .01$

results

Evaluation in Area Under ROC Curve (AUC)

Model	AUC
Verbal + Salient Gestures	.8226
Verbal + All Gestures	.8109
Verbal only	.7945
Gesture only	.6732

+ 1.2%

+ 1.6%

multimodal beats verbal only: $t(15) = 4.45$, $p < .01$
hidden variable beats flat model: $t(15) = 3.73$, $p < .01$

contribution of gesture features increases by a relative 73%

is it really salience?

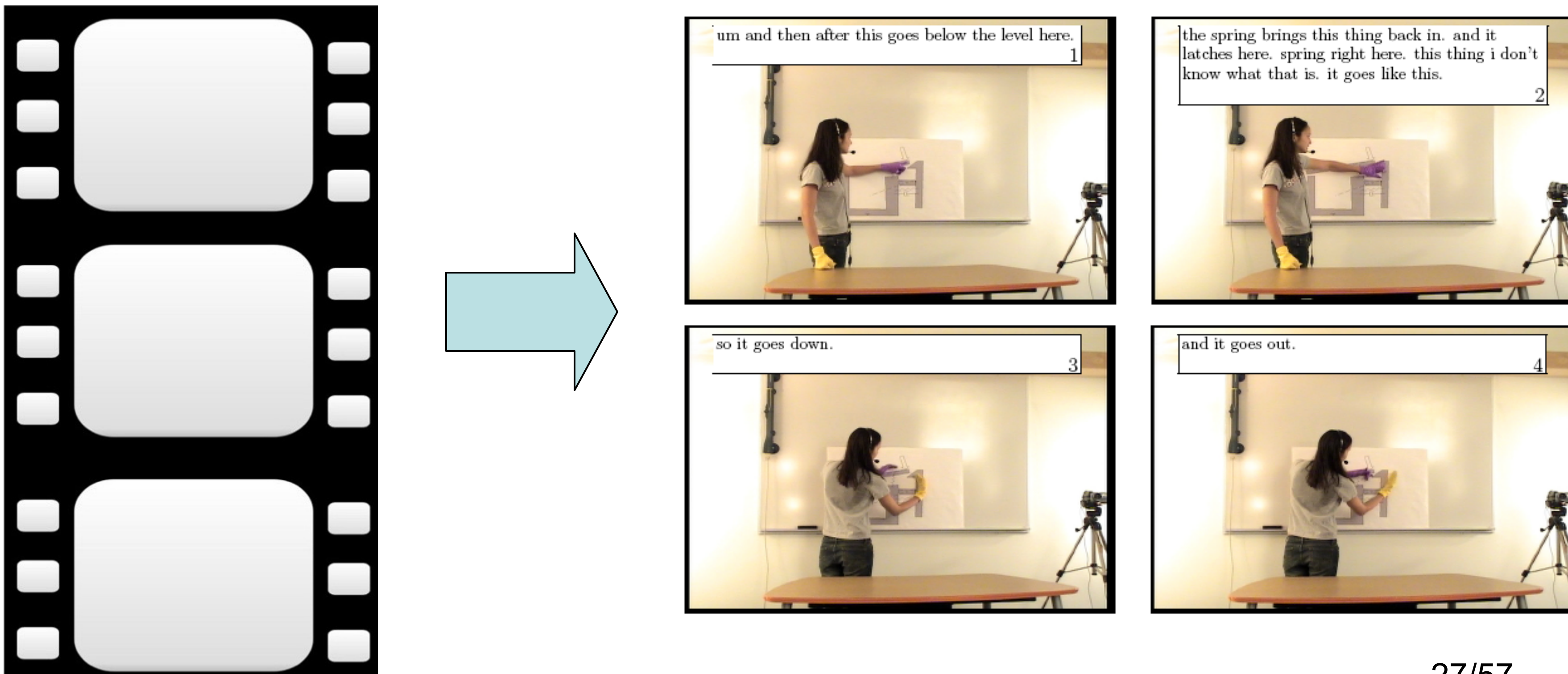
- Coreference performance improves
- But is it really learning gesture salience?
  - Do the system's estimates of salience agree with human perception?

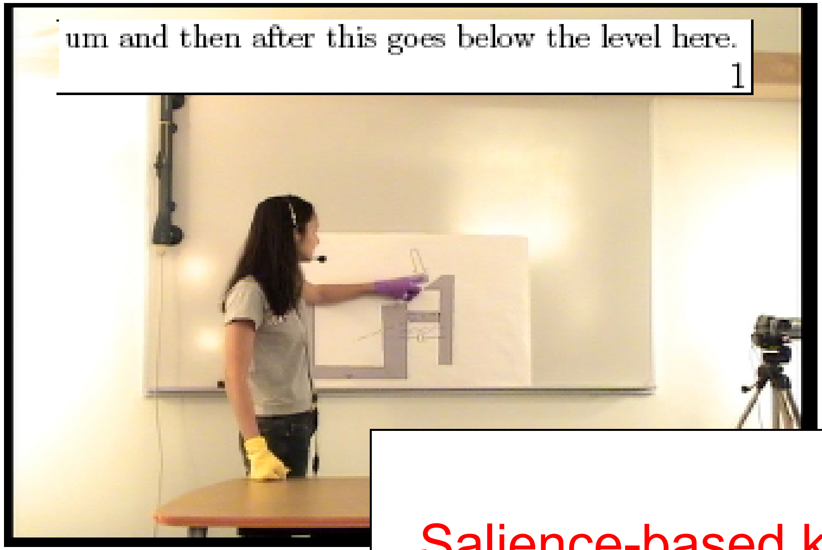
Eisenstein, Barzilay and Davis, "Turning Lectures into Comic Books with Linguistically Salient Gestures," AAAI 2007.

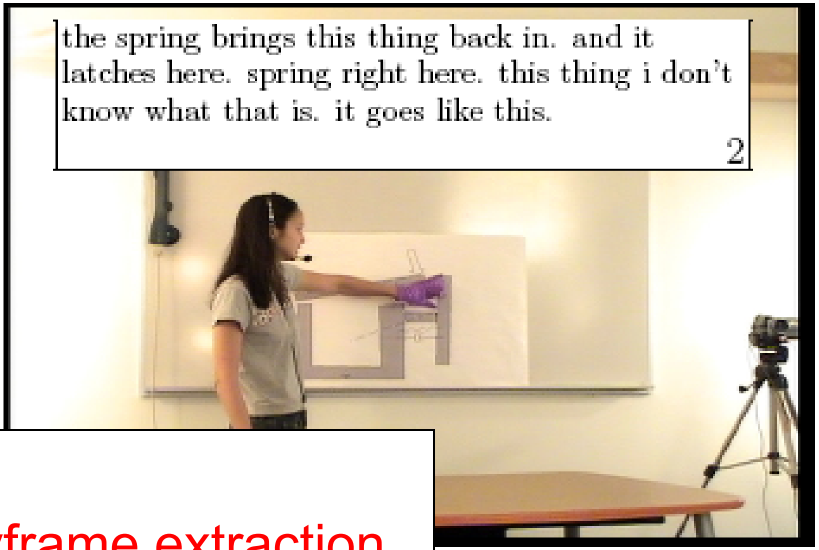
Eisenstein, Barzilay and Davis, "Modeling Gesture Salience as a Hidden Variable for Coreference Resolution and Keyframe Extraction." JAIR, 2008.

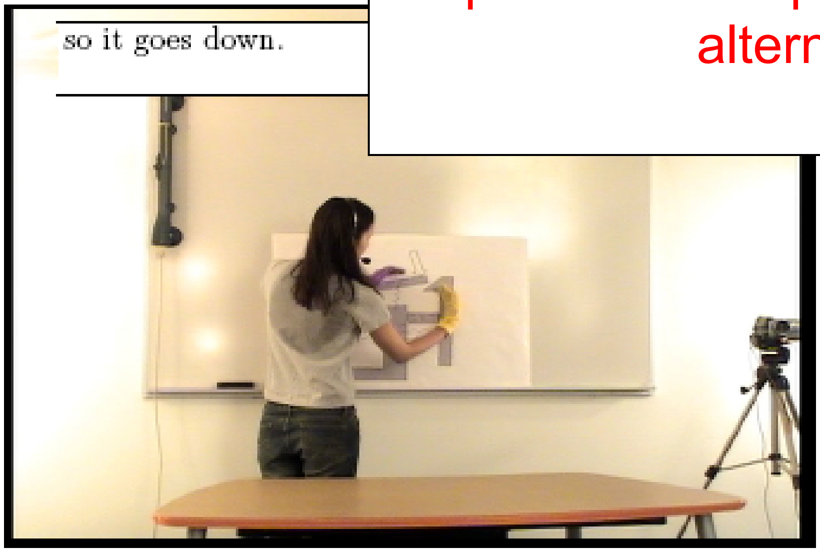
keyframe summarization

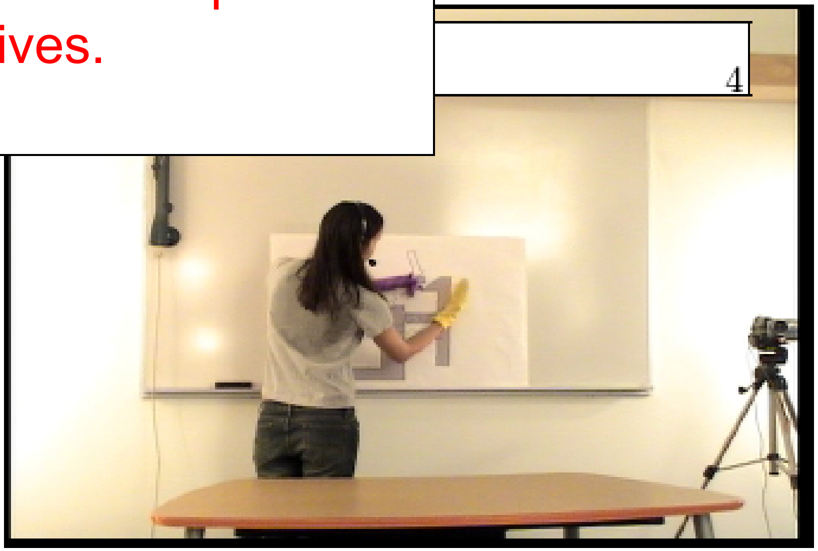
- Application: keyframe summaries showing salient gestures.







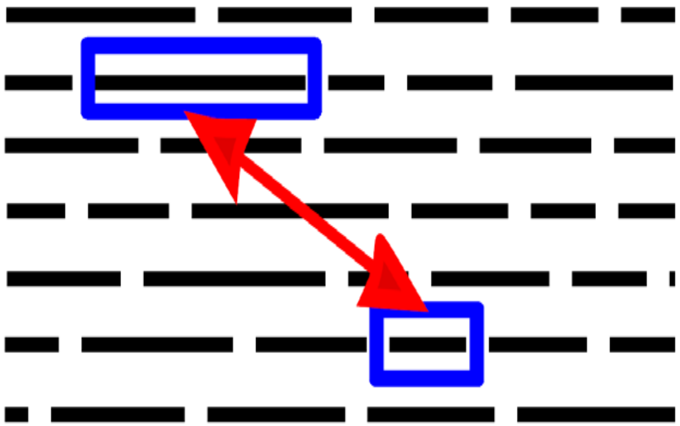




Salience-based keyframe extraction outperforms competitive unsupervised alternatives.

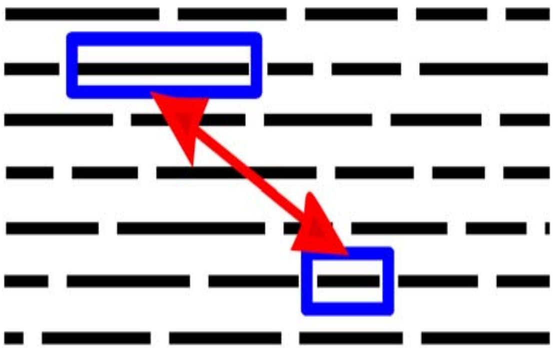
local discourse: summary

- Gesture similarity predicts NP coreference.
- Salience substantially increases the usefulness of gesture features.
- Conditional modality fusion learns salience as a hidden variable.

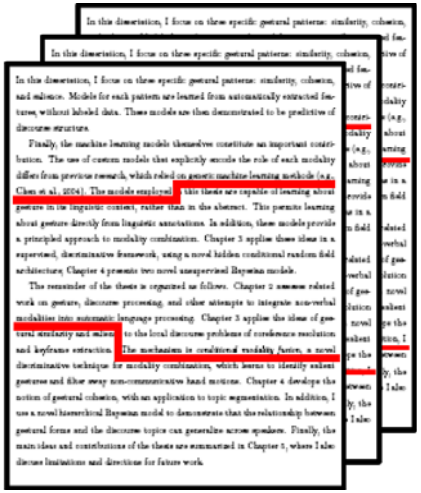


outline

- Local discourse structure
  - **Task:** noun phrase references
  - **Gesture:** similarity and salience
  - **Learning:** transfer from linguistic annotations
  - **Visual features**: hand tracking



- Global discourse structure
  - **Task:** topic segmentation
  - **Gesture:** cohesion
  - **Learning:** unsupervised
  - **Visual features:** interest points



topic segmentation

- High-level task: divide text into coherent segments

1: “Ok, so there’s this -- like if you think of the block letter c. It comes like this, right?”
2: “OK, backward C”
1: “Well I’m drawing it the right way. Just draw it as a C. And – but it comes in at the top and bottom.”
1: “OK, and then there’s a T-shaped thing such that the… if this is a t, rotate it like this. And this part is inside the C. And this part is in the opening, and it’s connected.”
1: “And then to the t, there’s this other short piece that’s connected. That’s can rotate around an axis a little bit but not too much.”
2: “Where is it connected?”
1. “To the… to here.”
2. “So it’s like a little flap”
1. “No it’s like a… it’s a stub. It’s like the length of the t. It’s a bar, connecting bar.”
1. “And that bar is connected to this wheel. So there’s a wheel over here. And it’s connected at a specific point on the wheel…”

Eisenstein, Barzilay and Davis, “Gestural Cohesion for Discourse Segmentation,” ACL 2008.

topic segmentation

High-level task: divide text into coherent segments

Eisenstein, Barzilay and Davis, "Gestural Cohesion for Discourse Segmentation," ACL 2008.

Topic: the backward C

1: "Ok, so there's this -- like if you think of the block letter c. It comes like this, right?"

2: "OK, backward C"

1: "Well I'm drawing it the right way. Just draw it as a C. And – but it comes in at the top and bottom."

Topic: the T

1: "OK, and then there's a T-shaped thing such that the… if this is a t, rotate it like this. And this part is inside the C. And this part is in the opening, and it's connected."

1: "And then to the t, there's this other short piece that's connected. That's can rotate around an axis a little bit but not too much."

2: "Where is it connected?"

1. "To the… to here."

2. "So it's like a little flap"

1. "No it's like a… it's a stub. It's like the length of the t. It's a bar, connecting bar."

Topic: the wheel

1. "And that bar is connected to this wheel. So there's a wheel over here. And it's connected at a specific point on the wheel…"

topics and gestural form

Topic: the backward C

1: “Ok, so there’s this -- like if you think of the block letter c. It comes like this, right?”

2: “OK, backward C”

1: “Well I’m drawing it the right way. Just draw it as a C. And – but it comes in at the top and bottom.”

Topic: the T

1: “OK, and then there’s a T-shaped thing such that the… if this is a t, rotate it like this. And this part is inside the C. And this part is in the opening, and it’s connected.”

1: “And then to the t, there’s this other short piece that’s connected. That’s can rotate around an axis a little bit but not too much.”

2: “Where is it connected?”

1. “To the… to here.”

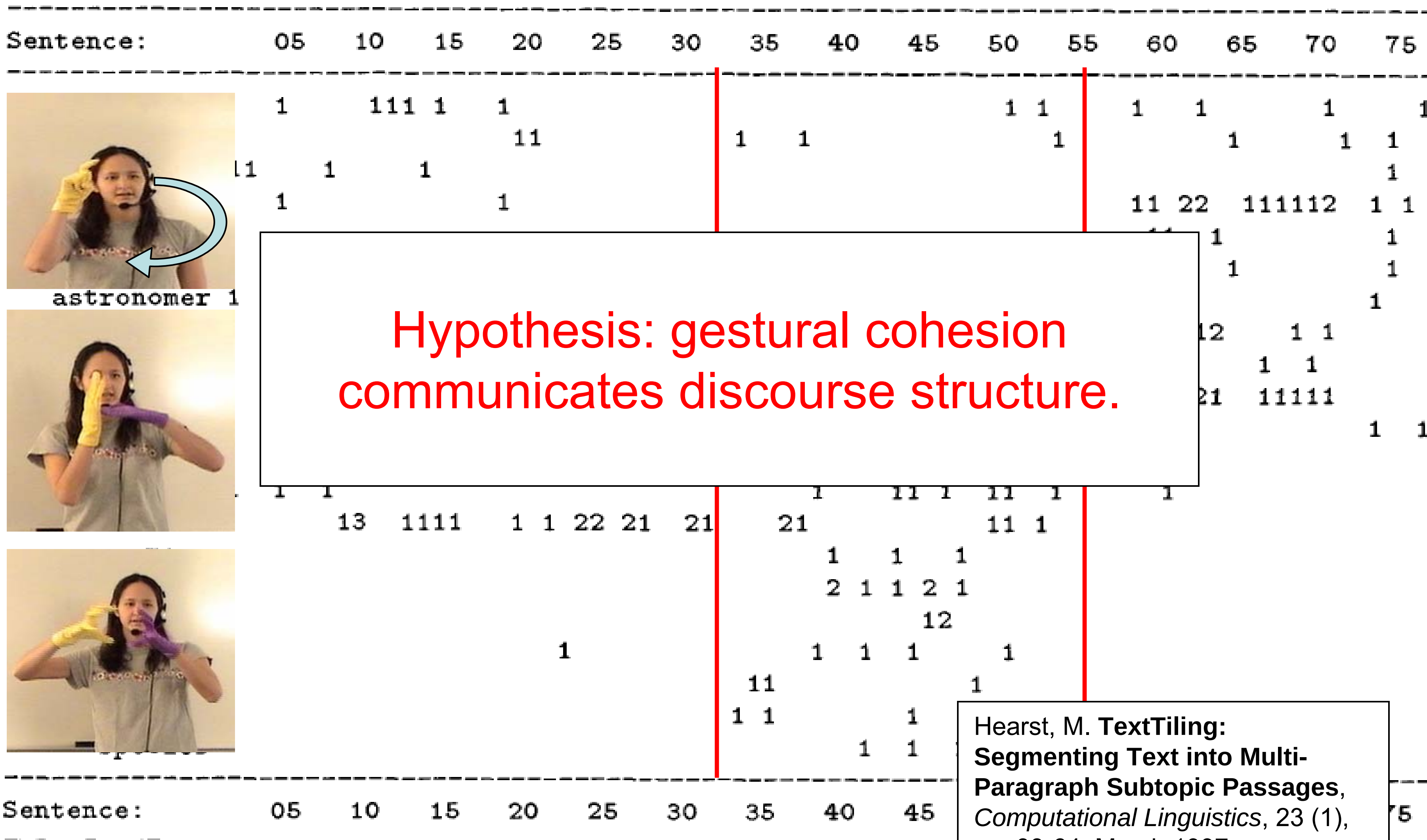
2. “So it’s like a little flap”

1. “No it’s like a… it’s a stub. It’s like the length of the t. It’s a bar, connecting bar.”

Topic: the wheel

1. “And that bar is connected to this wheel. So there’s a wheel over here. And it’s connected at a specific point on the wheel…”

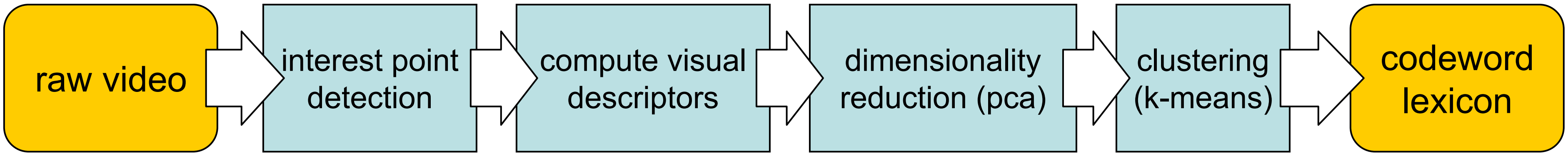
segmentation by cohesion



Hypothesis: gestural cohesion communicates discourse structure.

Hearst, M. **TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages**, *Computational Linguistics*, 23 (1), pp. 33-64, March 1997.

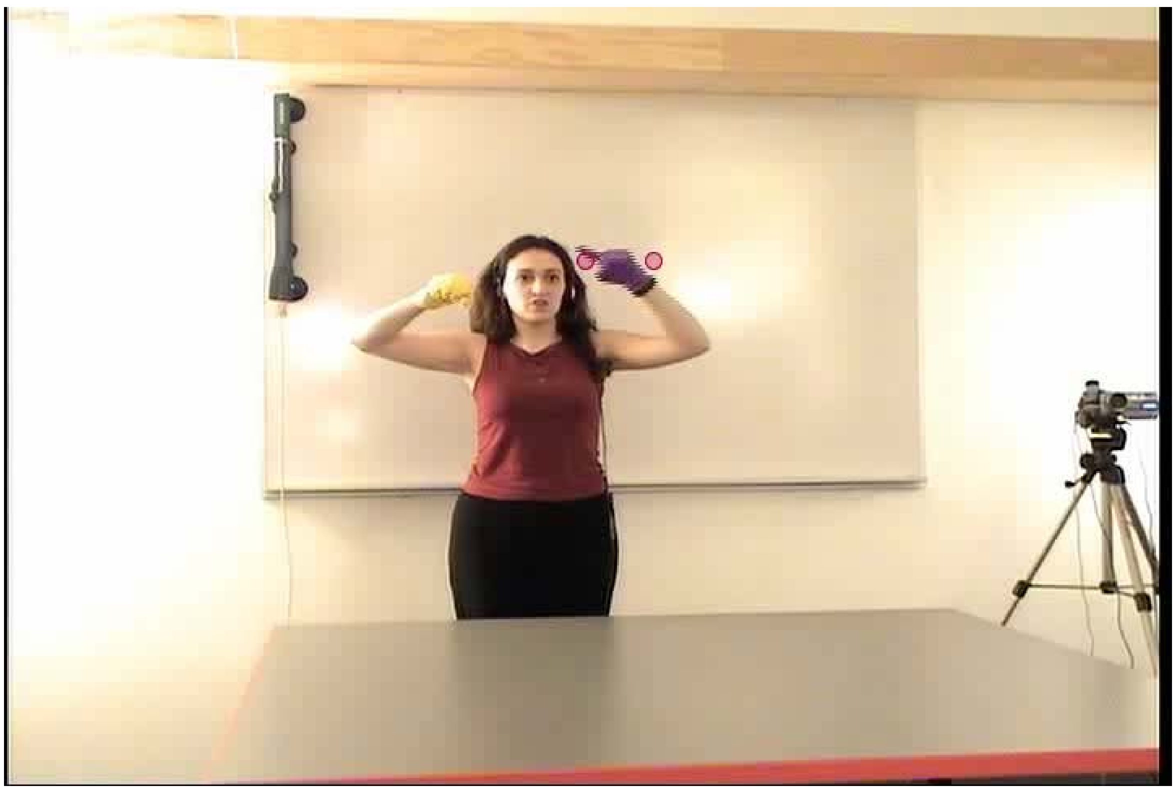
extracting gestural codewords



```
----------------------------------------------------------------------------------------------------------------
Sentence:       05   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95
----------------------------------------------------------------------------------------------------------------
14       form   1     111 1   1                               1 1     1   1       1     1     1    1
 8  scientist                  11            1   1               1          1      1  1
 5      space11    1     1                                                              1
25       star   1             1                                       11 22  111112  1 1  1   11 1111     1
 5     binary                                                           11  1          1                    1
 4    trinary                                                            1   1         1                    1
 8 astronomer1                1                                         1 1           1   1    1  1
 7      orbit  1                    1                                      12    1 1
 6       pull                         2     1 1                               1  1
16     planet  1    1       11               1           1                 21  11111                  1     1
 7     galaxy   1                                                  1                1  11     1           1
 4      lunar           1  1     1       1
19       life1  1  1                              1    11 1  11  1      1                 1 1    1 111  1 1
27       moon       13  1111   1 1 22 21  21    21           11 1
 3       move                                      1   1   1
 7  continent                                      2 1 1 2 1
 3  shoreline                                            12
 6       time                     1               1  1   1     1                                         1
 3      water                                  11           1
 6        say                                 1 1         1       11                1
 3    species                                        1   1 1
----------------------------------------------------------------------------------------------------------------
Sentence:       05   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95
----------------------------------------------------------------------------------------------------------------
```

extracting gestural codewords

raw video → interest point detection → compute visual descriptors → dimensionality reduction (pca) → clustering (k-means) → codeword lexicon

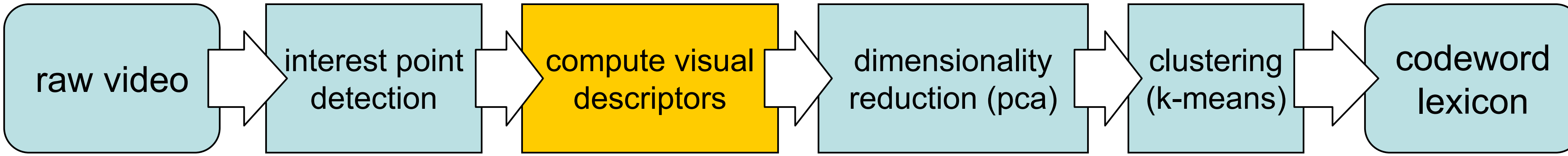


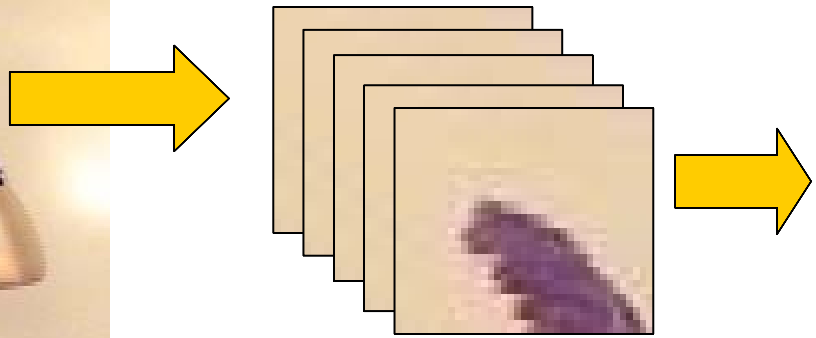
Spatio-temporal interest points give a sparse representation of motion.

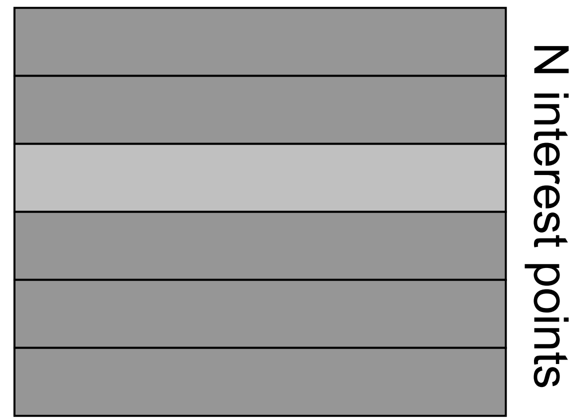
Laptev, "On Space-Time Interest Points." IJCV (64). 2005.

Dollar et al, "Behavior recognition via sparse spatio-temporal features." In ICCV VS-PETS 2005.

extracting gestural codewords



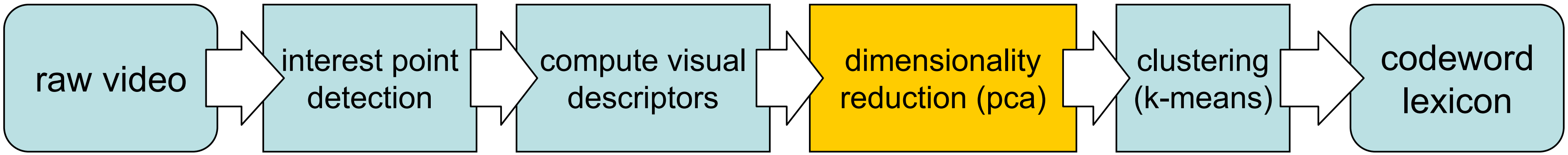


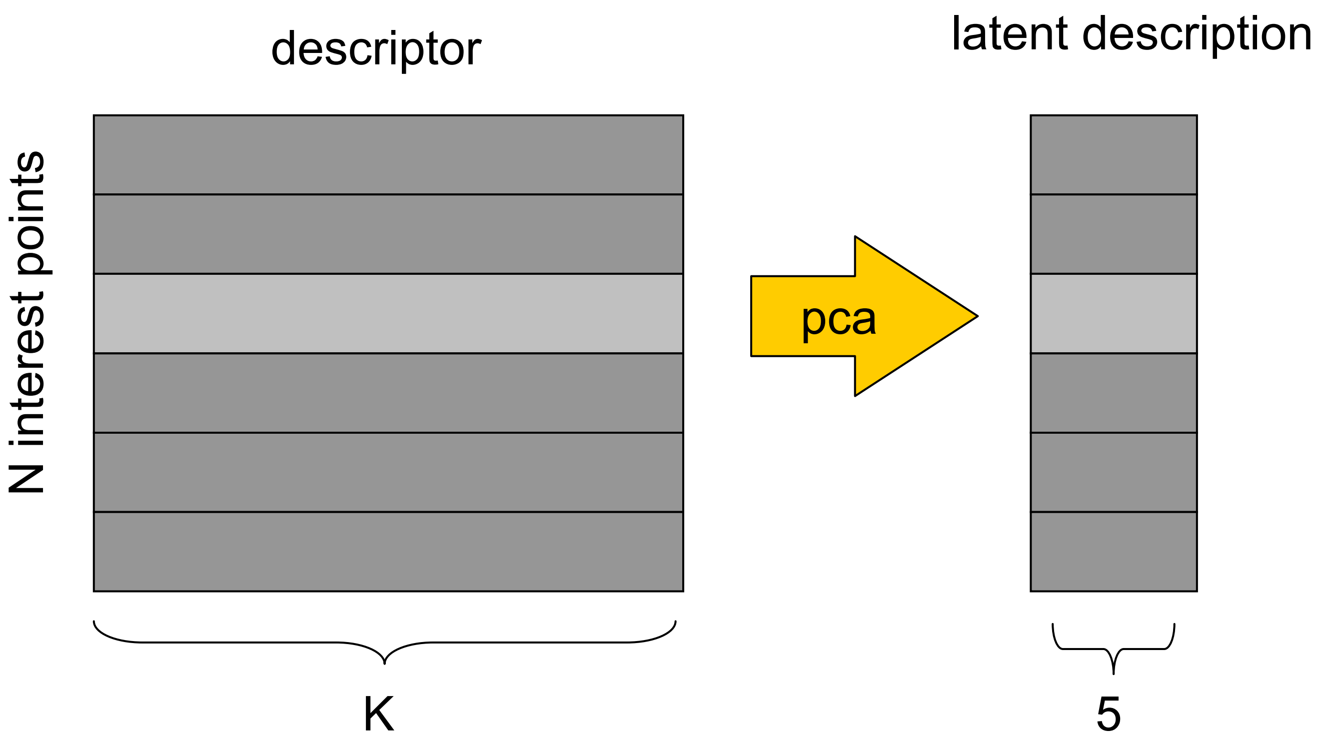


A small space-time volume is extracted at each interest point.

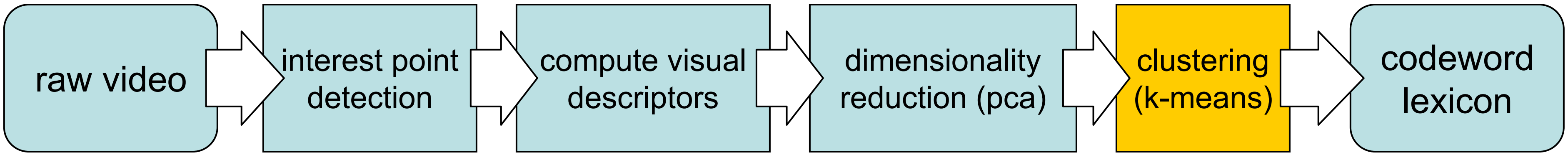
Each volume is described in terms of the brightness gradient.

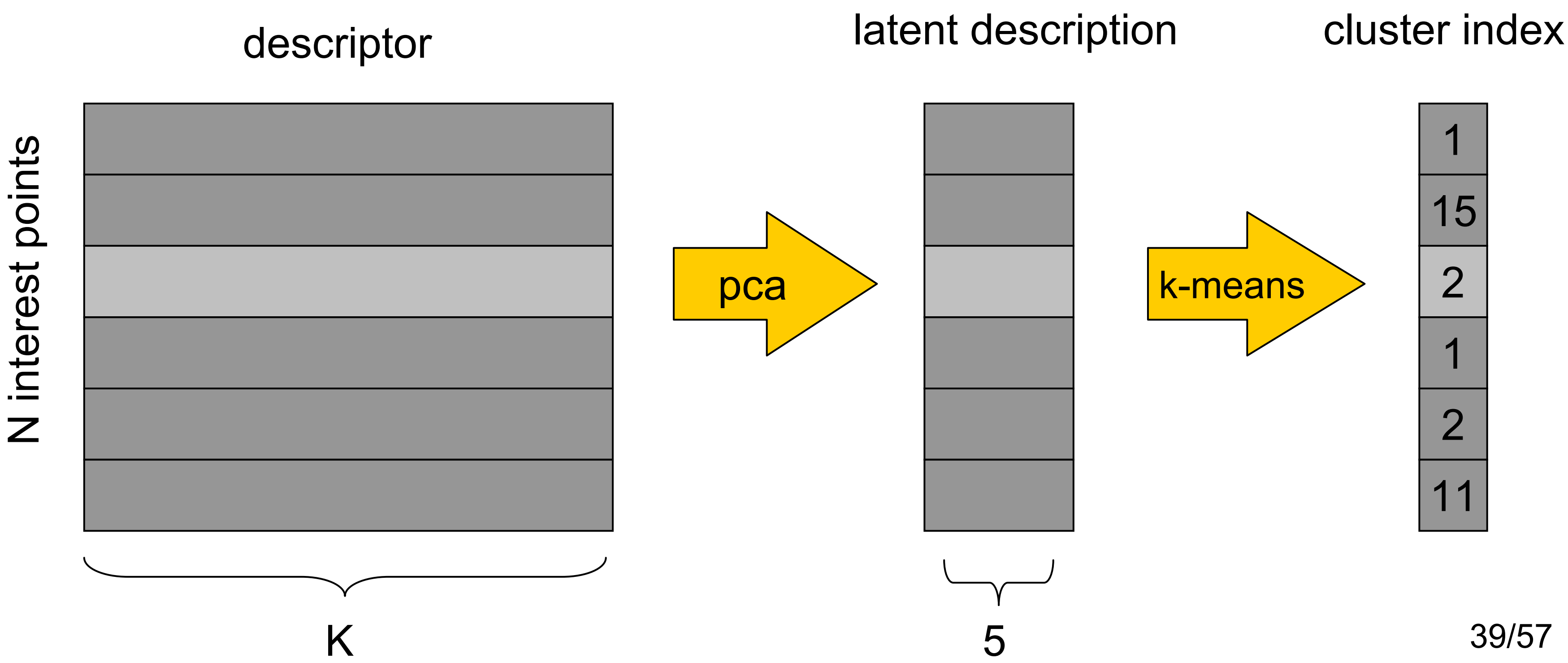
extracting gestural codewords





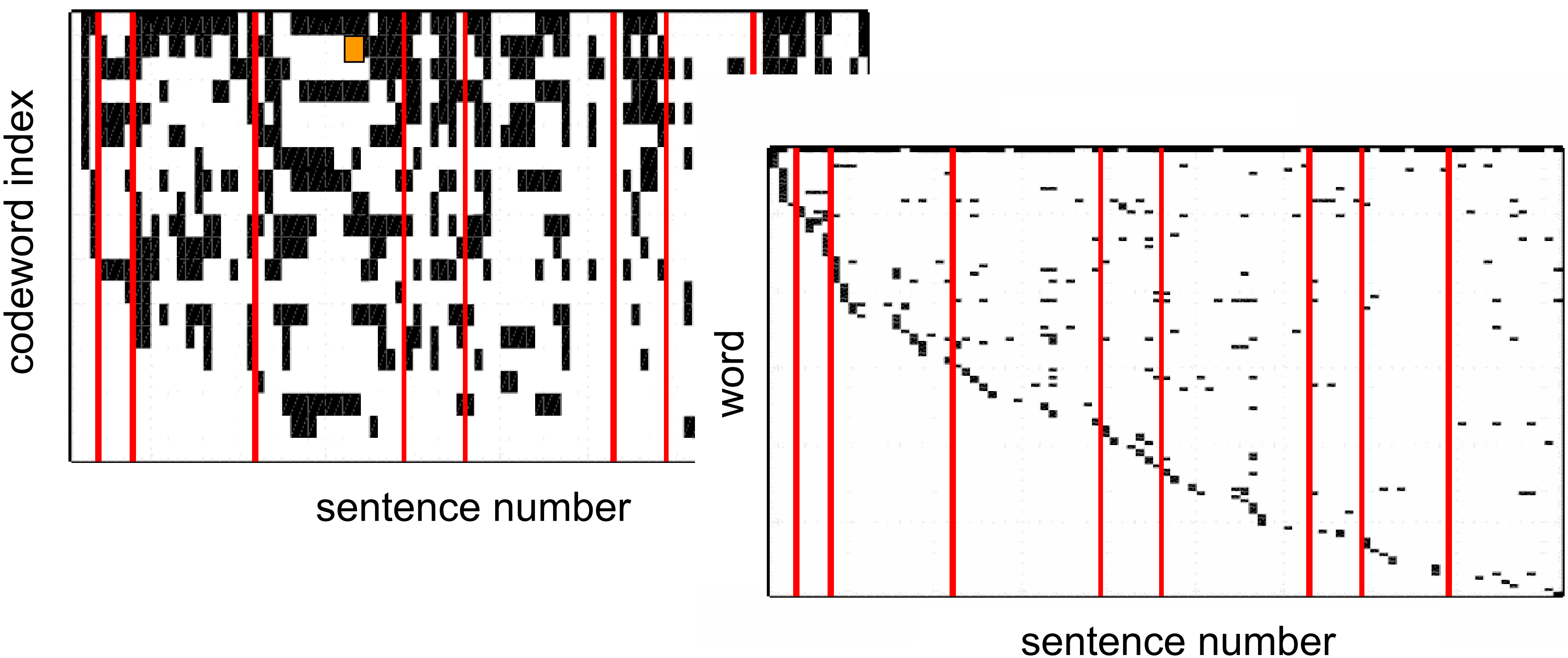
extracting gestural codewords





extracting gestural codewords

raw video → interest point detection → compute visual descriptors → dimensionality reduction (pca) → clustering (k-means) → codeword lexicon



Bayesian segmentation

$$\boxed{\hat{S}} = \mathrm{argmax}_S\, p(\boxed{S}, \boxed{\mathbf{x}}, \boxed{\mathbf{y}} | \boxed{\theta_0})$$

segmentation   words   gesture codewords   priors

$$S = \langle \boxed{\mathbf{z}}, \boxed{\theta} \rangle, \qquad \theta_i^{(v)} = E[\theta | \mathbf{x}_i, \theta_0^{(v)}]$$

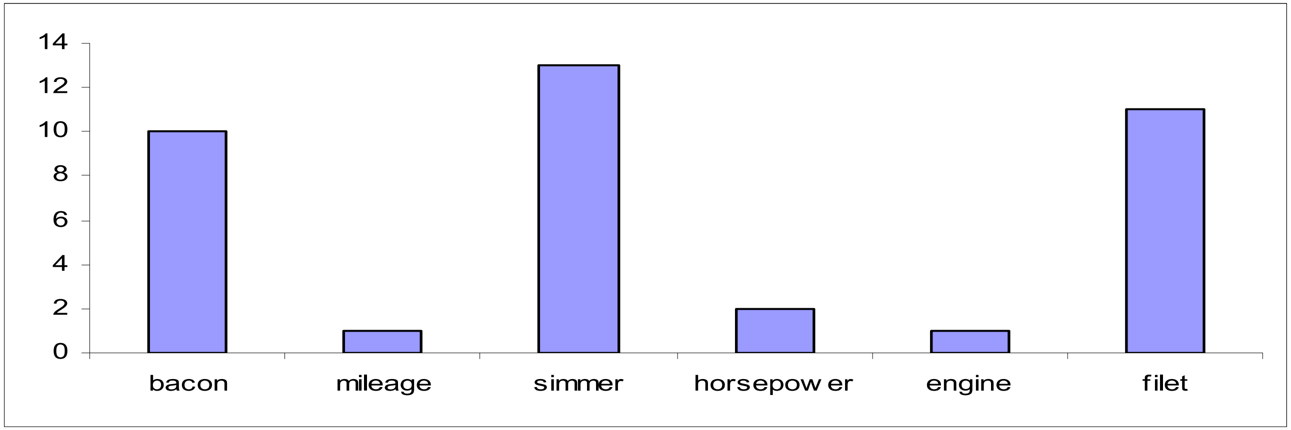
segmentation points   language models

$$p(S, \mathbf{x}, \mathbf{y} | \theta_0) = p(\mathbf{x}, \mathbf{y} | S, \theta_0) p(S | \theta_0)$$

$$= \prod_i^K \underbrace{p(\mathbf{x}_i | \theta_i^{(v)}) p(\mathbf{y}_i | \theta_i^{(g)})}_{\text{multinomials}} \underbrace{p(\theta_i^{(v)} | \theta_0^{(v)}) p(\theta_i^{(g)} | \theta_0^{(g)})}_{\text{Dirichlet priors}}$$

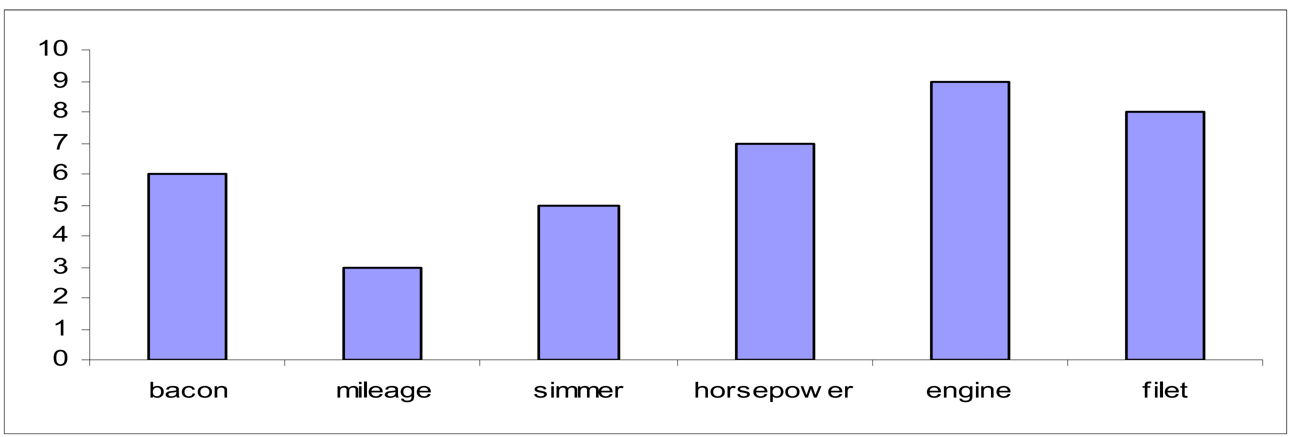
High likelihood segmentations have **compact** language models

p

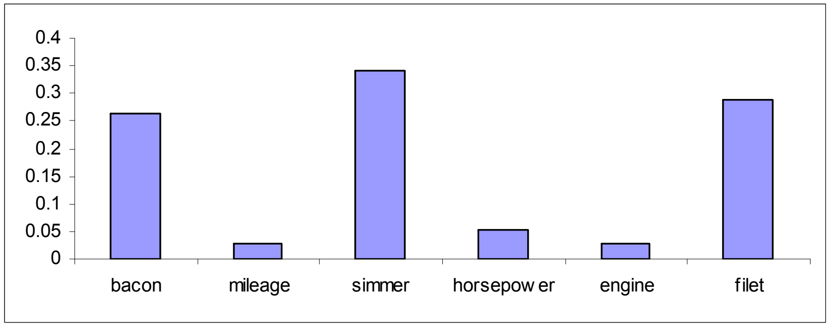


>

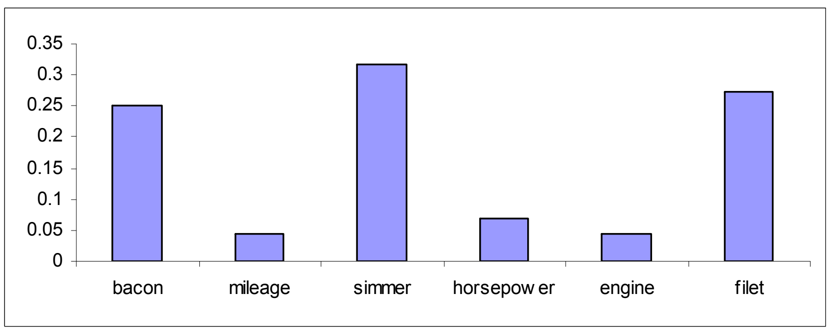
p



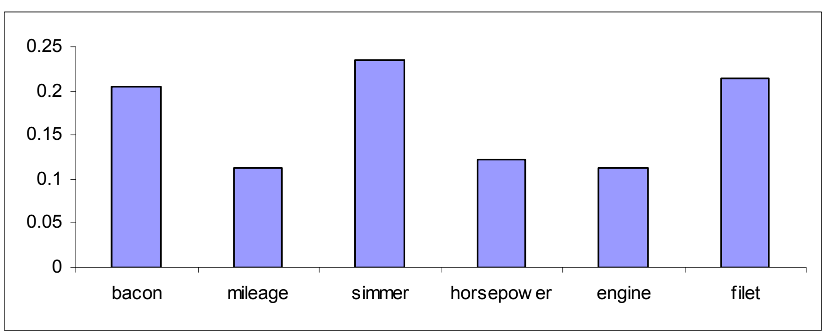
priors control modality weight



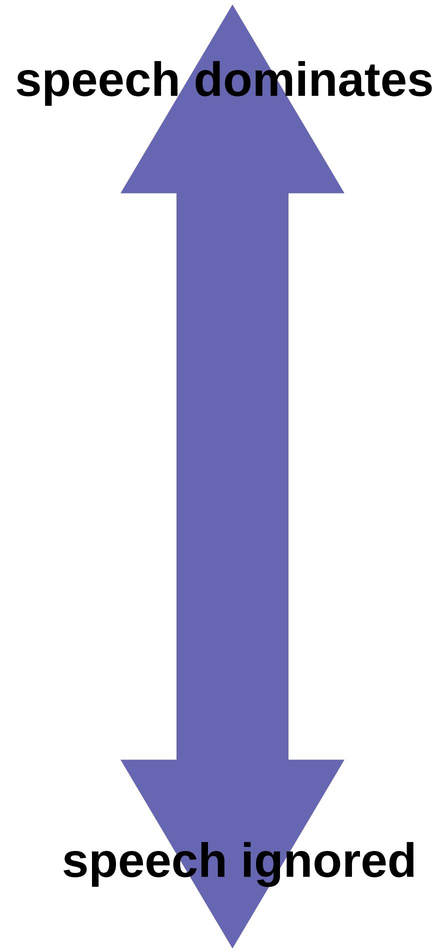
$\theta_0 = .1$



$\theta_0 = 1$



$\theta_0 = 10$



evaluation setup

- Dataset
  - 15 videos, 9 speakers
  - mechanical devices + cartoon narrative
  - no visual aids
  - transcriptions
    - manual transcript
    - automatic speech recognition (ASR)

results

Model	Pk	WindowDiff
Random	.473	.526
Equal-width	.508	.515

results

Model	Pk	WindowDiff
Random	.473	.526
Equal-width	.508	.515
Gesture only	.460	.489

results

Model	Pk	WindowDiff
Random	.473	.526
Equal-width	.508	.515
Gesture only	.460	.489
ASR only	.458	.472

results

Model	Pk	WindowDiff
Random	.473	.526
Equal-width	.508	.515
Gesture only	.460	.489
ASR only	.458	.472
ASR + gesture	***.388***	***.401***

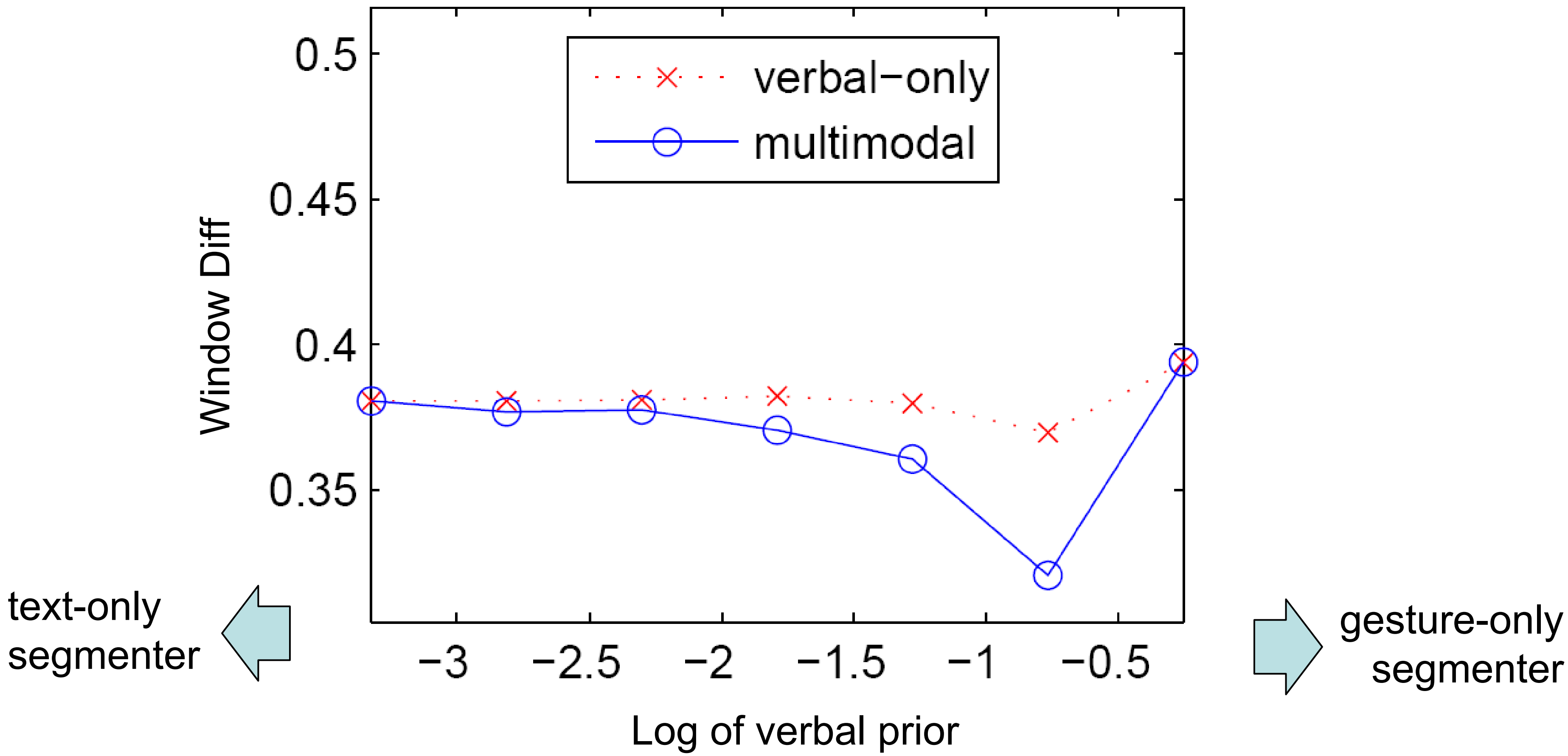
results

Model	Pk	WindowDiff
Random	.473	.526
Equal-width	.508	.515
Gesture only	.460	.489
ASR only	.458	.472
ASR + gesture	***.388***	***.401***
Transcript only	.370	.384

results

Model	**Pk**	**WindowDiff**
Random	.473	.526
Equal-width	.508	.515
Gesture only	.460	.489
ASR only	.458	.472
ASR + gesture	***.388***	***.401***
Transcript only	.370	.384
Transcript + gesture	***.321***	***.336***

priors control modality weight



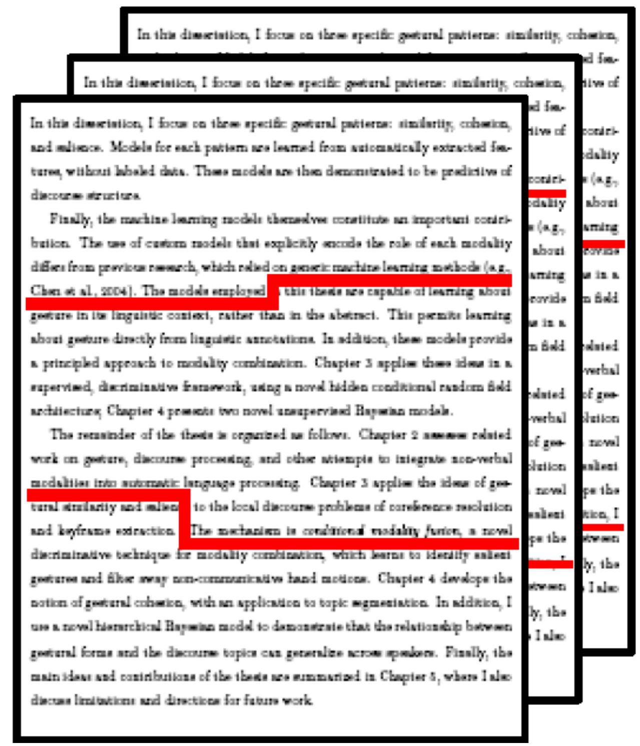
speakers and topics

- How idiosyncratic are gestures?
  - Very difficult to answer with manual annotation.
  - To what extend is the codewords distribution governed by the speaker and the topic?
- Codeword representation demonstrates consistency across speakers.

Eisenstein, Barzilay, and Davis. "Discourse Topic and Gestural Form." AAAI 2008.

global discourse: summary

- Gestural cohesion predicts segment boundaries.
- Gesture adds new information beyond lexical cohesion alone.
- Hand tracking not necessary for gestural analysis

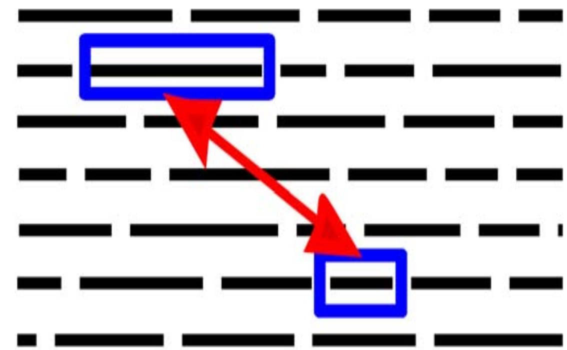

In this dissertation, I focus on three specific gestural patterns: similarity, cohesion, and salience. Models for each pattern are learned from automatically extracted features, without labeled data. These models are then demonstrated to be predictive of discourse structure.

Finally, the machine learning models themselves constitute an important contribution. The use of custom models that explicitly encode the role of each modality differs from previous research, which relied on generic machine learning methods (e.g., Chen et al., 2004). The models employed in this thesis are capable of learning about gesture in its linguistic context, rather than in the abstract. This permits learning about gesture directly from linguistic annotations. In addition, these models provide a principled approach to modality combination. Chapter 3 applies these ideas in a supervised, discriminative framework, using a novel hidden conditional random field architecture; Chapter 4 presents two novel unsupervised Bayesian models.

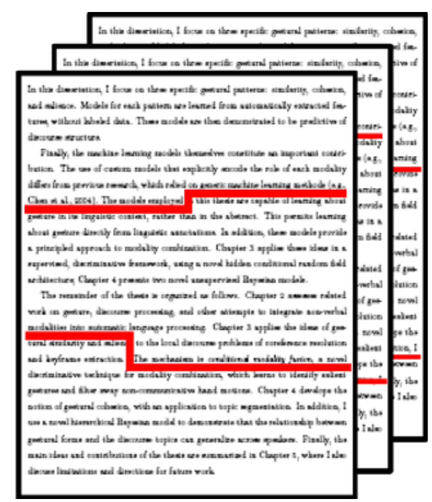
The remainder of the thesis is organized as follows. Chapter 2 assesses related work on gesture, discourse processing, and other attempts to integrate non-verbal modalities into automatic language processing. Chapter 3 applies the ideas of gestural similarity and salience to the local discourse problems of coreference resolution and keyframe extraction. The mechanism is conditional modality fusion, a novel discriminative technique for modality combination, which learns to identify salient gestures and filter away non-communicative hand motions. Chapter 4 develops the notion of gestural cohesion, with an application to topic segmentation. In addition, I use a novel hierarchical Bayesian model to demonstrate that the relationship between gestural forms and the discourse topics can generalize across speakers. Finally, the main ideas and contributions of the thesis are summarized in Chapter 5, where I also discuss limitations and directions for future work.

outline

- Local discourse structure
  - **Task:** noun phrase references
  - **Gesture:** similarity and salience
  - **Learning:** transfer from linguistic annotations
  - **Visual features**: hand tracking



- Global discourse structure
  - **Task:** topic segmentation
  - **Gesture:** cohesion
  - **Learning:** unsupervised
  - **Visual features:** interest points



prior work

- David McNeill
  - Gestural catchments
- Francis Quek *et al.*
  - Gesture patterns correlate with discourse structure.
- Lei Chen *et al.*
  - Gesture as visual punctuation

McNeill. **Hand and Mind**. University of Chicago Press, 1992.

Quek. "The Catchment Feature Model for Multimodal Language Analysis," ICCV 2003.

Chen, Harper, and Huang. "Using Maximum Entropy (ME) Model to Incorporate Gesture Cues for SU Detection." ICMI 2006.

contributions

- Gesture improves discourse interpretation.
- Methods
  - Gesture patterns, not gesture recognition!
  - Key gestural properties: similarity, cohesion, and salience
  - Structured models for combining gesture, speech, and meaning.

Thank You!

My committee: Regina Barzilay, Michael Collins, Randy Davis, and Candy Sidner.

The NLP and multimodal understanding groups, and many other good friends at MIT.

contributions

- Gesture improves discourse interpretation.
- Methods
  - Gesture patterns, not gesture recognition!
  - Key gestural properties: similarity, cohesion, and salience
  - Structured models for combining gesture, speech, and meaning.