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
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 the block letter C. It comes like this.”

“Then there’s a T-shaped thing...if this is a T, rotate it like this.”

“There’s a wheel over here.”

Can we use gestures for automatic natural language processing?
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 clicks back…" "…and clicks back…"

Without "recognizing" these gestures, they can still tell us something.

patterns in gesture predict patterns in language
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?

conditional modality fusion

Gesture features are only included for salient gestures.

Salience is learned jointly with coreference.
conditional modality fusion

\[ p(y|x; w) = \sum_h p(y, h|x; w) \]

\[ = \frac{\sum_h \exp\{\psi(y, h, x; w)\}}{\sum_{y', h} \exp\{\psi(y', h, x; w)\}} \]

- \( y \) – coreference label
- \( h \) – gesture salience
- \( x \) – observed features
- \( w \) – learned weights
- \( \psi \) – potential function

potential function

\[ \psi(y, h, x; w) = \psi(y, x_v) + \psi(h, x_h) + \psi(y, x_g, h) \]

\[ \psi(y, x_v) = yw_v^T x_v \]

\[ \psi(h, x_h) = h_1 w_h^T x_{h_1} + h_2 w_h^T x_{h_2} \]

\[ \psi(y, x_g, h) = \begin{cases} yw_g^T x_g, & h_1 = h_2 = 1 \\ 0, & \text{otherwise.} \end{cases} \]
learning salience

\[
\text{sign}(y) \neq \text{sign}(w_g^T x_g) \rightarrow y w_g^T x_g < 0
\]

\[
\text{sign}(y) = \text{sign}(w_g^T x_g) \rightarrow y w_g^T x_g > 0
\]

\[
\psi(y, h, x; w) = \psi(y, x_v) + \psi(h, x_h) + \psi(y, x_g, h)
\]

\[
\psi(y, x_g, h) = \begin{cases} 
  y w_g^T x_g, & h_1 = h_2 = 1 \\
  0, & \text{otherwise.}
\end{cases}
\]
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)

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal only</td>
<td>.7945</td>
</tr>
<tr>
<td>Gesture only</td>
<td>.6732</td>
</tr>
</tbody>
</table>
## results

Evaluation in Area Under ROC Curve (AUC)

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal + All Gestures</td>
<td>.8109</td>
</tr>
<tr>
<td>Verbal only</td>
<td>.7945</td>
</tr>
<tr>
<td>Gesture only</td>
<td>.6732</td>
</tr>
</tbody>
</table>

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

Evaluation in Area Under ROC Curve (AUC)

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal + Salient Gestures</td>
<td>.8226</td>
<td>+ 1.2%</td>
</tr>
<tr>
<td>Verbal + All Gestures</td>
<td>.8109</td>
<td>+ 1.6%</td>
</tr>
<tr>
<td>Verbal only</td>
<td>.7945</td>
<td></td>
</tr>
<tr>
<td>Gesture only</td>
<td>.6732</td>
<td></td>
</tr>
</tbody>
</table>

- 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?


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…”

topic segmentation

High-level task: divide text into coherent segments

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…”
Hypothesis: gestural cohesion communicates discourse structure.

extracting gestural codewords

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


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

1. raw video
2. interest point detection
3. compute visual descriptors
4. dimensionality reduction (pca)
5. clustering (k-means)
6. codeword lexicon

N interest points

descriptor

latent description

pca

K

5
extracting gestural codewords

1. raw video
2. interest point detection
3. compute visual descriptors
4. dimensionality reduction (PCA)
5. clustering (k-means)
6. codeword lexicon

- N interest points
- descriptor
- latent description
- cluster index

K PCA

1 15 2 1 2 11

39/57
extracting gestural codewords

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

sentence number → codeword index

sentence number → word
Bayesian segmentation

\[ \hat{S} = \arg\max_S p(S, x, y|\theta_0) \]

\[ S = \{ z, \theta \}, \quad \theta_i^{(v)} = E[\theta|x_i, \theta_0^{(v)}] \]

\[ p(S, x, y|\theta_0) = p(x, y|S, \theta_0)p(S|\theta_0) \]

\[ = \prod_{i}^{K} p(x_i|\theta_i^{(v)})p(y_i|\theta_i^{(g)})p(\theta_i^{(v)}|\theta_0^{(v)})p(\theta_i^{(g)}|\theta_0^{(g)}) \]

- segmentation
- words
- gesture codewords
- priors
- segmentation points
- language models
- multinomials
- Dirichlet priors
High likelihood segmentations have **compact** language models.
priors control modality weight

\[ \theta_0 = .1 \]

\[ \theta_0 = 1 \]

\[ \theta_0 = 10 \]

speech dominates

speech ignored
evaluation setup

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Pk</th>
<th>WindowDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>.473</td>
<td>.526</td>
</tr>
<tr>
<td>Equal-width</td>
<td>.508</td>
<td>.515</td>
</tr>
</tbody>
</table>
## results

<table>
<thead>
<tr>
<th>Model</th>
<th>Pk</th>
<th>WindowDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>.473</td>
<td>.526</td>
</tr>
<tr>
<td>Equal-width</td>
<td>.508</td>
<td>.515</td>
</tr>
<tr>
<td>Gesture only</td>
<td>.460</td>
<td>.489</td>
</tr>
</tbody>
</table>
## results

<table>
<thead>
<tr>
<th>Model</th>
<th>Pk</th>
<th>WindowDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>.473</td>
<td>.526</td>
</tr>
<tr>
<td>Equal-width</td>
<td>.508</td>
<td>.515</td>
</tr>
<tr>
<td>Gesture only</td>
<td>.460</td>
<td>.489</td>
</tr>
<tr>
<td>ASR only</td>
<td>.458</td>
<td>.472</td>
</tr>
</tbody>
</table>
results

<table>
<thead>
<tr>
<th>Model</th>
<th>Pk</th>
<th>WindowDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>.473</td>
<td>.526</td>
</tr>
<tr>
<td>Equal-width</td>
<td>.508</td>
<td>.515</td>
</tr>
<tr>
<td>Gesture only</td>
<td>.460</td>
<td>.489</td>
</tr>
<tr>
<td>ASR only</td>
<td>.458</td>
<td>.472</td>
</tr>
<tr>
<td>ASR + gesture</td>
<td>.388</td>
<td>.401</td>
</tr>
</tbody>
</table>
# results

<table>
<thead>
<tr>
<th>Model</th>
<th>Pk</th>
<th>WindowDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.473</td>
<td>0.526</td>
</tr>
<tr>
<td>Equal-width</td>
<td>0.508</td>
<td>0.515</td>
</tr>
<tr>
<td>Gesture only</td>
<td>0.460</td>
<td>0.489</td>
</tr>
<tr>
<td>ASR only</td>
<td>0.458</td>
<td>0.472</td>
</tr>
<tr>
<td>ASR + gesture</td>
<td>0.388</td>
<td>0.401</td>
</tr>
<tr>
<td>Transcript only</td>
<td>0.370</td>
<td>0.384</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Pk</th>
<th>WindowDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>.473</td>
<td>.526</td>
</tr>
<tr>
<td>Equal-width</td>
<td>.508</td>
<td>.515</td>
</tr>
<tr>
<td>Gesture only</td>
<td>.460</td>
<td>.489</td>
</tr>
<tr>
<td>ASR only</td>
<td>.458</td>
<td>.472</td>
</tr>
<tr>
<td>ASR + gesture</td>
<td>.388</td>
<td>.401</td>
</tr>
<tr>
<td>Transcript only</td>
<td>.370</td>
<td>.384</td>
</tr>
<tr>
<td>Transcript + gesture</td>
<td>.321</td>
<td>.336</td>
</tr>
</tbody>
</table>
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.

global discourse: summary

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


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.