Social Interactions: A First-Person Perspective.

A. Fathi, J. Hodgins, J. Rehg
Presented by Jacob Menashe

November 16, 2012

Social Interaction Detection

Objective: Detect social interactions from video footage.

Social Interaction Detection

Objective: Detect social interactions from video footage.

- Consider faces and attention

Social Interaction Detection

Objective: Detect social interactions from video footage.

- Consider faces and attention
- Account for temporal context

Social Interaction Detection

Objective: Detect social interactions from video footage.

- Consider faces and attention
- Account for temporal context
- Analyze first-person movements cues

Video Example

Red	Dialogue
Yellow	Walking Dialogue
Green	Discussion
Light Blue	Walking Discussion
Dark Blue	Monologue
None	Background

Link

Features

Features are constructed based on first- and third-person information.

Features

Features are constructed based on first- and third-person information.

1. Dense optical flow (first-person movement).

Features

Features are constructed based on first- and third-person information.

1. Dense optical flow (first-person movement).
2. Face locations (relative to first person)

Features

Features are constructed based on first- and third-person information.

1. Dense optical flow (first-person movement).
2. Face locations (relative to first person)
3. Attention and Roles. For each person x:

Features

Features are constructed based on first- and third-person information.

1. Dense optical flow (first-person movement).
2. Face locations (relative to first person)
3. Attention and Roles. For each person x:
    - Faces looking at x

Features

Features are constructed based on first- and third-person information.

1. Dense optical flow (first-person movement).
2. Face locations (relative to first person)
3. Attention and Roles. For each person x:
   - Faces looking at x
   - Whether first person looks at x

Features

Features are constructed based on first- and third-person information.

1. Dense optical flow (first-person movement).
2. Face locations (relative to first person)
3. Attention and Roles. For each person x:
   - Faces looking at x
   - Whether first person looks at x
   - Mutual attention between x and first person

Features

Features are constructed based on first- and third-person information.

1. Dense optical flow (first-person movement).
2. Face locations (relative to first person)
3. Attention and Roles. For each person x:
   - Faces looking at x
   - Whether first person looks at x
   - Mutual attention between x and first person
   - Number of faces looking at where x is looking

Feature Example

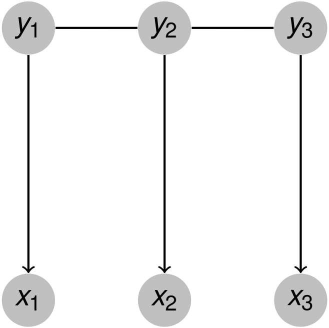
Conditional Random Fields

CRFs are described in Lafferty et al. [2001].

Conditional Random Fields

CRFs are described in Lafferty et al. [2001].

- Observations and labels form a Markov chain.
- Nodes pend on neighbors.

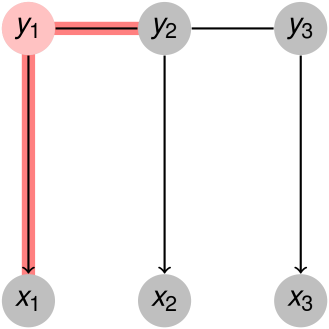


Conditional Random Fields

CRFs are described in Lafferty et al. [2001].

- Observations and labels form a Markov chain.
- Nodes pend on neighbors.

$p(y_1 | x_1, y_2)$

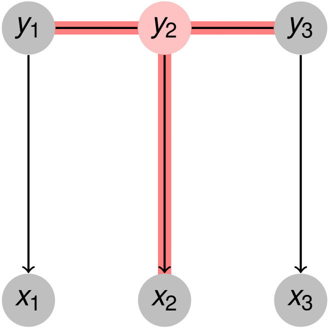


Conditional Random Fields

CRFs are described in Lafferty et al. [2001].

- Observations and labels form a Markov chain.
- Nodes pend on neighbors.

$$p(y_2|y_1, y_3, x_2)$$

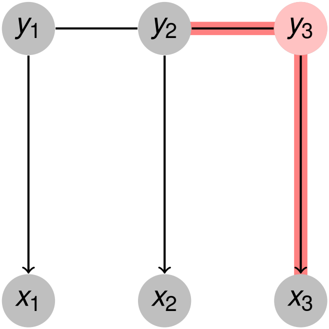


Conditional Random Fields

CRFs are described in Lafferty et al. [2001].

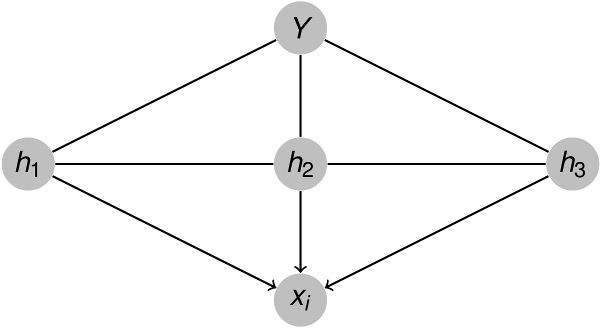
- Observations and labels form a Markov chain.
- Nodes pend on neighbors.

$p(y_3|y_2, x_3)$



Hidden Conditional Random Fields

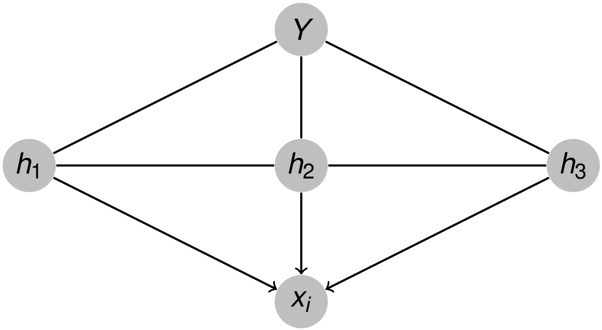
A micro view of the HCRF model as described in Quattoni et al. [2007].



Hidden Conditional Random Fields

A micro view of the HCRF model as described in Quattoni et al. [2007].

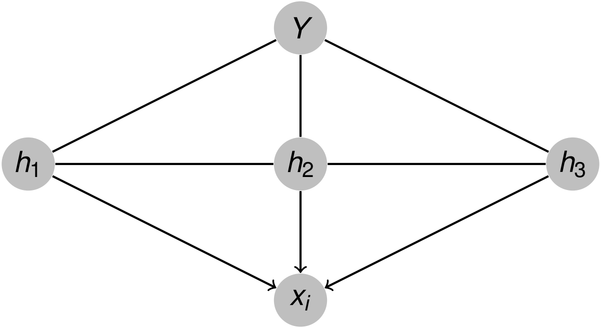
- Y is a label for the whole sequence.



Hidden Conditional Random Fields

A micro view of the HCRF model as described in Quattoni et al. [2007].

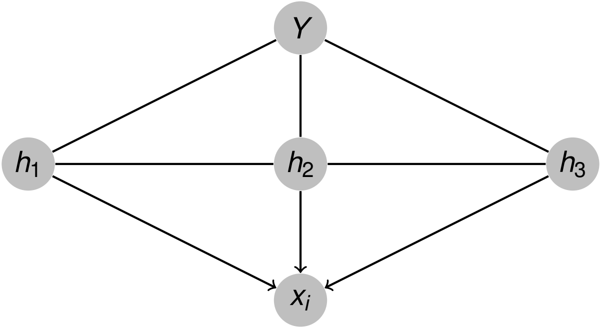
- Y is a label for the whole sequence.
- x_i is a single observation in the sequence.



Hidden Conditional Random Fields

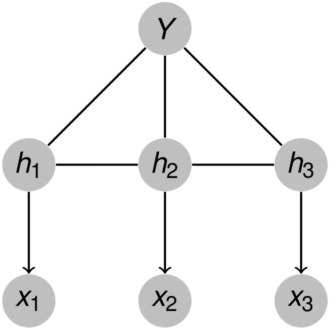
A micro view of the HCRF model as described in Quattoni et al. [2007].

- Y is a label for the whole sequence.
- x_i is a single observation in the sequence.
- Each h_i is a possible hidden state.



Hidden Conditional Random Fields (cont.)

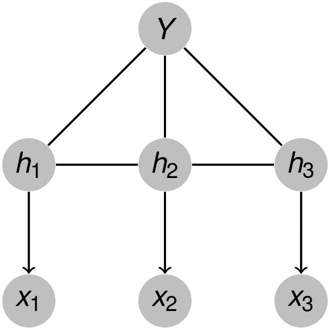
A macro view of the HCRF model as described in Quattoni et al. [2007].



Hidden Conditional Random Fields (cont.)

A macro view of the HCRF model as described in Quattoni et al. [2007].

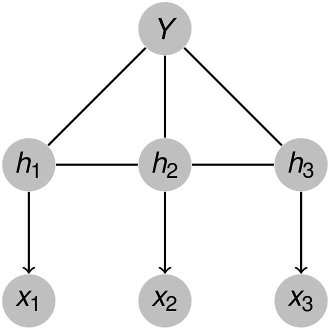
- Y is a label for the whole sequence.



Hidden Conditional Random Fields (cont.)

A macro view of the HCRF model as described in Quattoni et al. [2007].

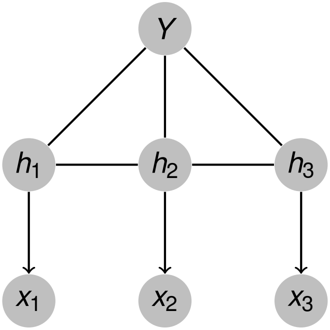
- Y is a label for the whole sequence.
- Each x_i is a single observation in the sequence.



Hidden Conditional Random Fields (cont.)

A macro view of the HCRF model as described in Quattoni et al. [2007].

- Y is a label for the whole sequence.
- Each x_i is a single observation in the sequence.
- Each h_i is the hidden state label assigned to x_i.

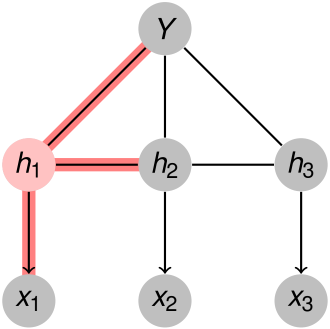


Hidden Conditional Random Fields (cont.)

A macro view of the HCRF model as described in Quattoni et al. [2007].

- Y is a label for the whole sequence.
- Each x_i is a single observation in the sequence.
- Each h_i is the hidden state label assigned to x_i.

$$p(h_1 | Y, h_2, x_1)$$

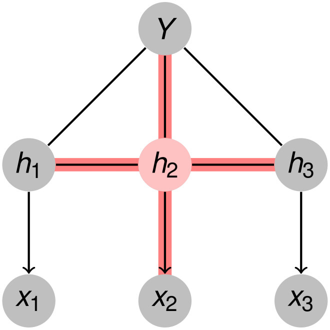


Hidden Conditional Random Fields (cont.)

A macro view of the HCRF model as described in Quattoni et al. [2007].

- Y is a label for the whole sequence.
- Each x_i is a single observation in the sequence.
- Each h_i is the hidden state label assigned to x_i.

$$p(h_2 | Y, h_1, h_3, x_2)$$

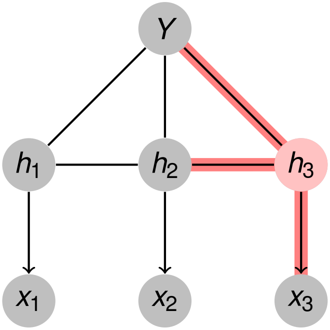


Hidden Conditional Random Fields (cont.)

A macro view of the HCRF model as described in Quattoni et al. [2007].

- Y is a label for the whole sequence.
- Each x_i is a single observation in the sequence.
- Each h_i is the hidden state label assigned to x_i.

$$p(h_3 | Y, h_2, x_3)$$

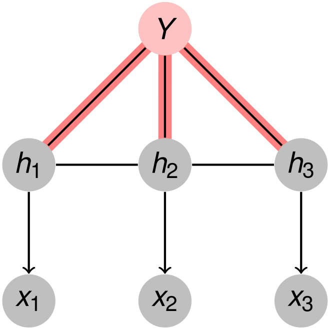


Hidden Conditional Random Fields (cont.)

A macro view of the HCRF model as described in Quattoni et al. [2007].

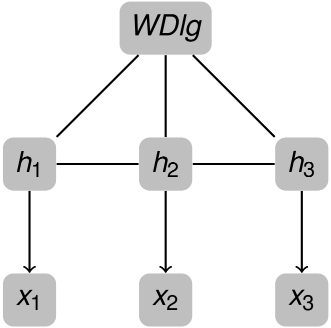
- Y is a label for the whole sequence.
- Each x_i is a single observation in the sequence.
- Each h_i is the hidden state label assigned to x_i.

$$p(Y|\{h_i\}) = p(Y|\{x_i\})$$



HCRF Example

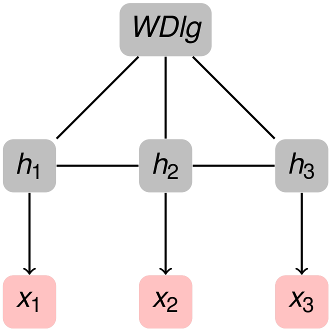
Suppose we want to find the likelihood of “walking dialogue” (*WDlg*) vs “walking discussion” (*WDisc*).



HCRF Example

Suppose we want to find the likelihood of “walking dialogue” (*WDlg*) vs “walking discussion” (*WDisc*).

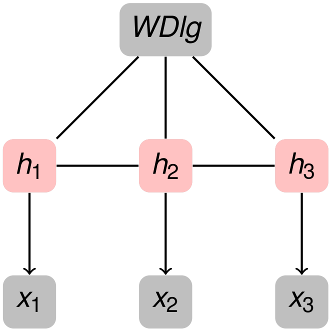
- Each x_i is now a feature extracted from video frames.



HCRF Example

Suppose we want to find the likelihood of “walking dialogue” (*WDlg*) vs “walking discussion” (*WDisc*).

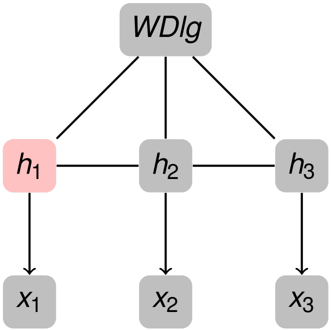
- Each x_i is now a feature extracted from video frames.
- Each h_i is determined from training:



HCRF Example

Suppose we want to find the likelihood of “walking dialogue” (*WDlg*) vs “walking discussion” (*WDisc*).

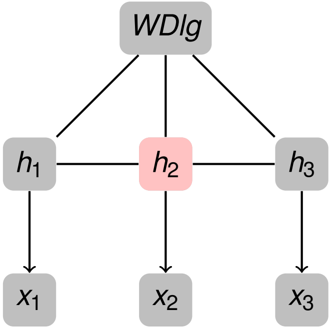
- Each x_i is now a feature extracted from video frames.
- Each h_i is determined from training:
  - h_1: John wants to hear about my weekend.



HCRF Example

Suppose we want to find the likelihood of “walking dialogue” (*WDlg*) vs “walking discussion” (*WDisc*).

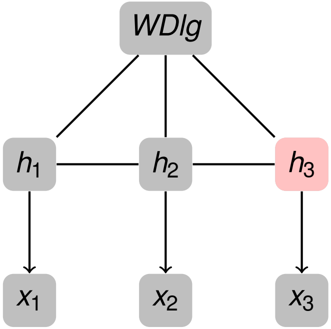
- Each x_i is now a feature extracted from video frames.
- Each h_i is determined from training:
  - h_2: I’m feeling talkative.



HCRF Example

Suppose we want to find the likelihood of “walking dialogue” (*WDlg*) vs “walking discussion” (*WDisc*).

- Each x_i is now a feature extracted from video frames.
- Each h_i is determined from training:
  - h_3: Mary wants to listen to her iPod.

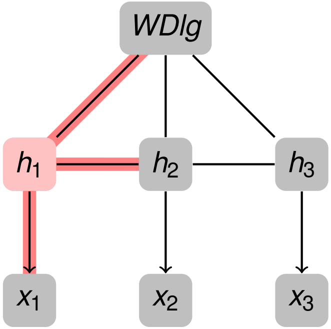


HCRF Example

Suppose we want to find the likelihood of “walking dialogue” (*WDlg*) vs “walking discussion” (*WDisc*).

- Each x_i is now a feature extracted from video frames.
- Each h_i is determined from training:
  - h_1: John wants to hear about my weekend.

$$p(h_1 \mid Y, h_2, x_1)$$

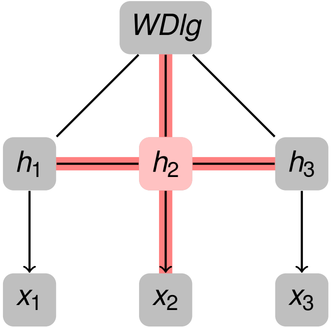


HCRF Example

Suppose we want to find the likelihood of “walking dialogue” (*WDlg*) vs “walking discussion” (*WDisc*).

- Each x_i is now a feature extracted from video frames.
- Each h_i is determined from training:
  - h_2: I’m feeling talkative.

$$p(h_2 | Y, h_1, h_3, x_2)$$

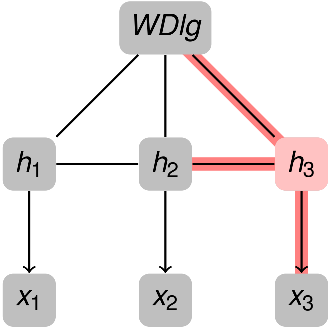


HCRF Example

Suppose we want to find the likelihood of "walking dialogue" (*WDlg*) vs "walking discussion" (*WDisc*).

- Each x_i is now a feature extracted from video frames.
- Each h_i is determined from training:
  - h_3: Mary wants to listen to her iPod.

$$p(h_3 | Y, h_2, x_3)$$

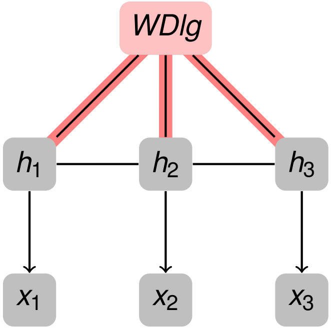


HCRF Example

Suppose we want to find the likelihood of “walking dialogue” (*WDlg*) vs “walking discussion” (*WDisc*).

- Each x_i is now a feature extracted from video frames.
- Each h_i is determined from training:
  - h_1: John wants to hear about my weekend.
  - h_2: I’m feeling talkative.
  - h_3: Mary wants to listen to her iPod.

$$p(WDlg|\{h_i\}) = p(WDlg|\{x_i\})$$

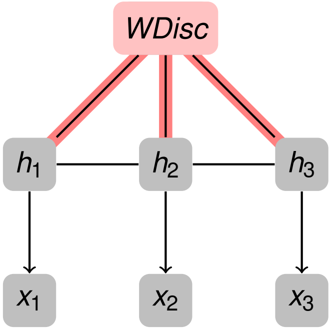


HCRF Example

Suppose we want to find the likelihood of “walking dialogue” (*WDlg*) vs “walking discussion” (*WDisc*).

- Each x_i is now a feature extracted from video frames.
- Each h_i is determined from training:
  - h_1: John wants to hear about my weekend.
  - h_2: I’m feeling talkative.
  - h_3: Mary wants to listen to her iPod.

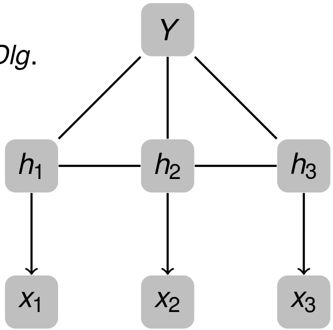
$$p(WDisc|\{h_i\}) = p(WDisc|\{x_i\})$$



HCRF Example

Suppose we want to find the likelihood of “walking dialogue” (*WDlg*) vs “walking discussion” (*WDisc*).

- Each x_i is now a feature extracted from video frames.
- Each h_i is determined from training:
  - h_1: John wants to hear about my weekend.
  - h_2: I’m feeling talkative.
  - h_3: Mary wants to listen to her iPod.
  - If $p(WDlg) > p(WDisc)$, assign $Y = WDlg$.



Experiment Outline

The following experiments are presented:

Experiment Outline

The following experiments are presented:

- Video Processing

Experiment Outline

The following experiments are presented:

- Video Processing
- Caltech image dataset

Experiment Outline

The following experiments are presented:

- Video Processing
- Caltech image dataset
- Adjusted parameters:

Experiment Outline

The following experiments are presented:

- Video Processing
- Caltech image dataset
- Adjusted parameters:
  - Iterations

Experiment Outline

The following experiments are presented:

- Video Processing
- Caltech image dataset
- Adjusted parameters:
  - Iterations
  - Hidden States

Experiment Outline

The following experiments are presented:

- Video Processing
- Caltech image dataset
- Adjusted parameters:
  - Iterations
  - Hidden States
  - Optimization Function

Experiment Outline

The following experiments are presented:

- Video Processing
- Caltech image dataset
- Adjusted parameters:
    - Iterations
    - Hidden States
    - Optimization Function
    - Clusters

Experiment Outline

The following experiments are presented:

- Video Processing
- Caltech image dataset
- Adjusted parameters:
  - Iterations
  - Hidden States
  - Optimization Function
  - Clusters
- Compared with linear SVM baseline

Experiment 1: Video Processing

Mine	Theirs
40 training intervals	4,000 training intervals
40 testing intervals	[unspecified]
Dialogue vs Discussion	One vs. All
All Features	Location First-Person Motion Attention All Features

Experiment 1: Video Processing

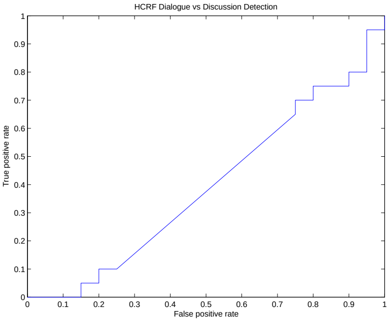
Mine	Theirs
40 training intervals	4,000 training intervals
40 testing intervals	[unspecified]
Dialogue vs Discussion	One vs. All
All Features	Location First-Person Motion Attention All Features

~42 hours = 11,340 intervals

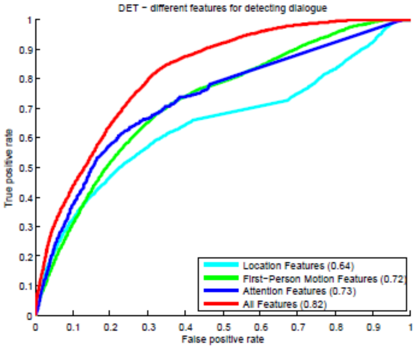
11,340 intervals @ 24 hours per 20 intervals $>$ 18 months

Experiment 1: Video Processing (cont.)

My Results



Their Results



(a)

Experiment 2: Caltech Dataset

Experiment 2 focuses on the Caltech image dataset.

Experiment 2: Caltech Dataset

Experiment 2 focuses on the Caltech image dataset.

- Multi-class HCRF evaluated

Experiment 2: Caltech Dataset

Experiment 2 focuses on the Caltech image dataset.

- Multi-class HCRF evaluated
- Classes are evaluated in isolation.

Experiment 2: Caltech Dataset

Experiment 2 focuses on the Caltech image dataset.

- Multi-class HCRF evaluated
- Classes are evaluated in isolation.
- Temporal context is simulated with clustering

Experiment 2: Caltech Dataset

Experiment 2 focuses on the Caltech image dataset.

- Multi-class HCRF evaluated
- Classes are evaluated in isolation.
- Temporal context is simulated with clustering
- Initial parameters are based on Fathi et al. [2012]:

Experiment 2: Caltech Dataset

Experiment 2 focuses on the Caltech image dataset.

- Multi-class HCRF evaluated
- Classes are evaluated in isolation.
- Temporal context is simulated with clustering
- Initial parameters are based on Fathi et al. [2012]:
  - Hidden States: 5

Experiment 2: Caltech Dataset

Experiment 2 focuses on the Caltech image dataset.

- Multi-class HCRF evaluated
- Classes are evaluated in isolation.
- Temporal context is simulated with clustering
- Initial parameters are based on Fathi et al. [2012]:
  - Hidden States: 5
  - Window Size: 5

Experiment 2: Caltech Dataset

Experiment 2 focuses on the Caltech image dataset.

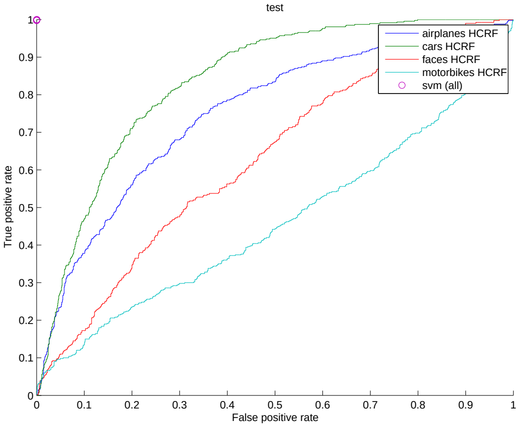
- Multi-class HCRF evaluated
- Classes are evaluated in isolation.
- Temporal context is simulated with clustering
- Initial parameters are based on Fathi et al. [2012]:
  - Hidden States: 5
  - Window Size: 5
  - Max Iterations: 100

Experiment 2: Caltech Dataset

Experiment 2 focuses on the Caltech image dataset.

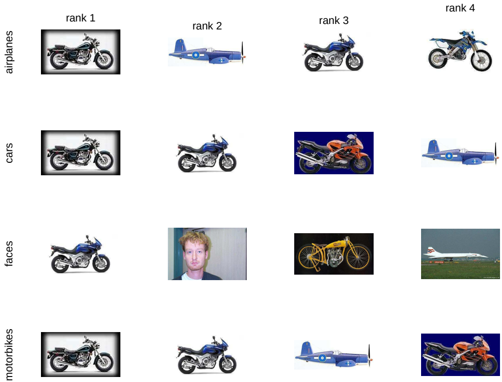
- Multi-class HCRF evaluated
- Classes are evaluated in isolation.
- Temporal context is simulated with clustering
- Initial parameters are based on Fathi et al. [2012]:
  - Hidden States: 5
  - Window Size: 5
  - Max Iterations: 100
  - Optimizer: Broyden–Fletcher-Goldfarb-Shanno (BFGS)

Exp. 2a: Initial Settings

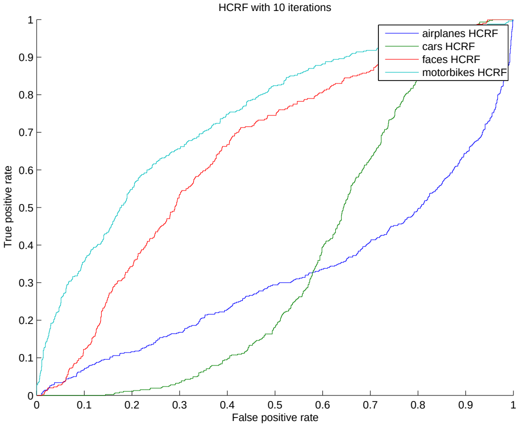


Processing: ~18 minutes, 1 MB

Exp. 2a: Initial Settings (cont.)

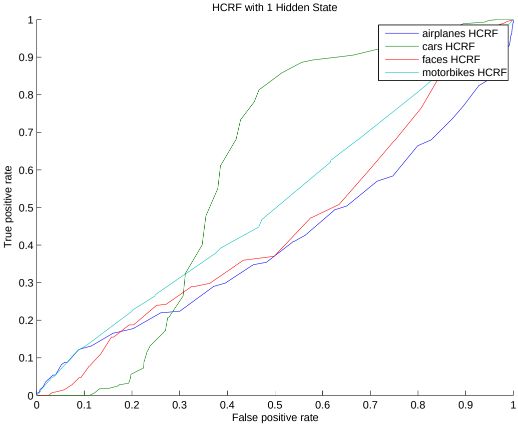


Exp. 2b: Low Iterations



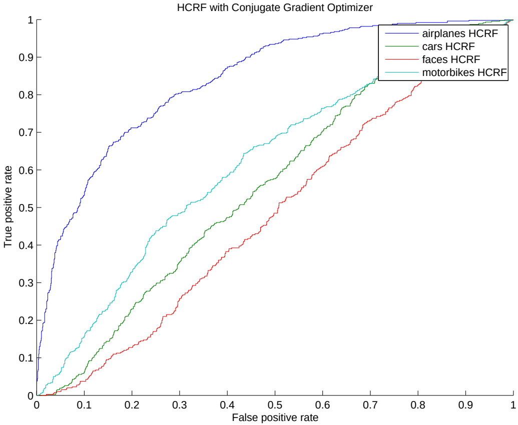
Processing: ~3 minutes, 1 MB

Exp. 2c: Low Hidden States



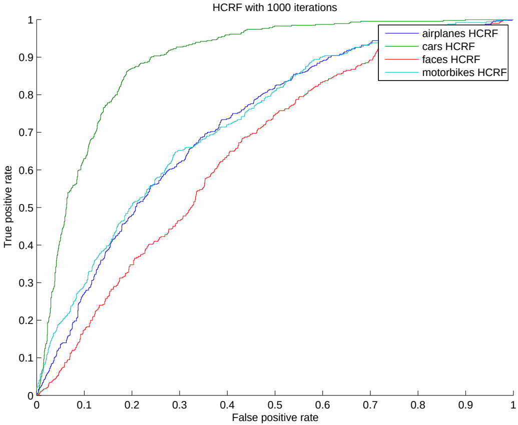
Processing: ~2 minutes, 1 MB

Exp. 2d: CG Optimizer



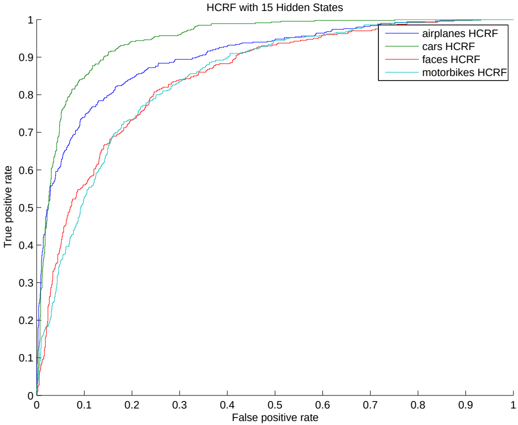
Processing: ~11 minutes, 1 MB

Exp. 2e: Increased Iterations



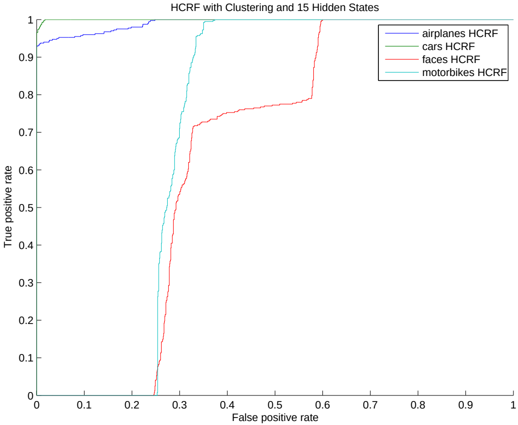
Processing: ~30 minutes, 1 MB

Exp. 2f: Increased Hidden States



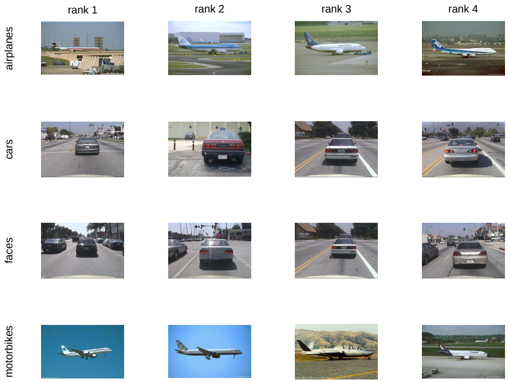
Processing: ~1 hour, 3 GB

Exp. 2g: Clustering + 15 Hidden States

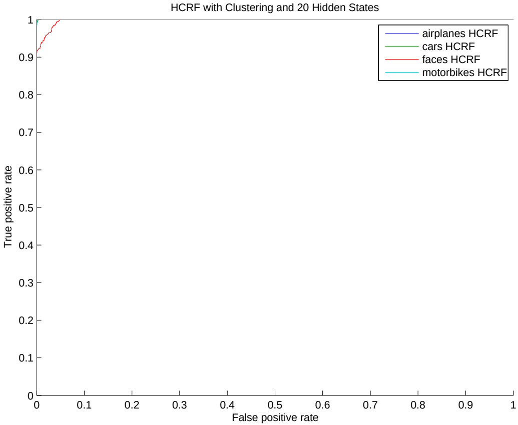


Processing: ~1 hour 10 minutes, 3 GB

Exp. 2g: Clustering + 15 Hidden States (cont.)

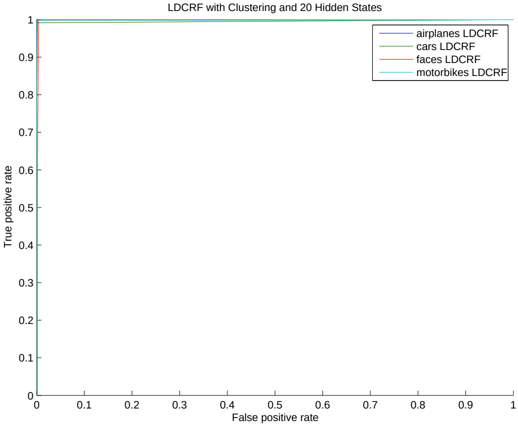


Exp. 2h: Clustering + 20 Hidden States



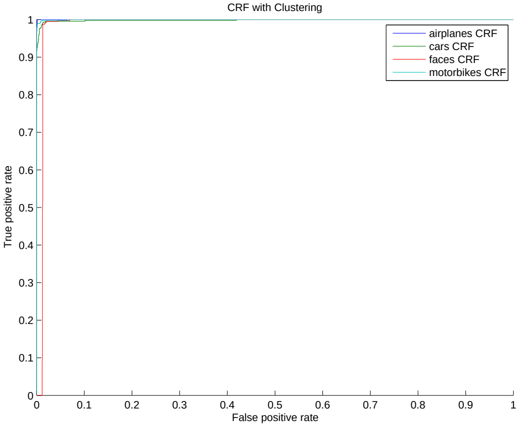
Processing: ~1 hour 40 minutes, 5 GB

Exp. 2i: LDCRF with 20 Hidden States



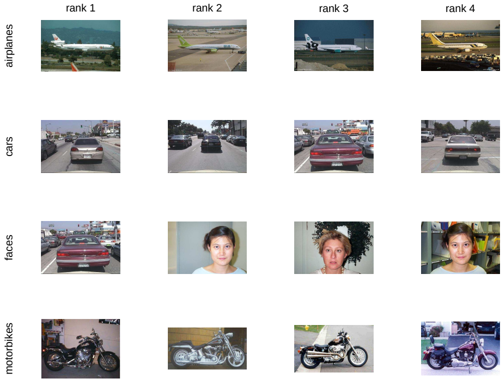
Processing: ~5 hours 20 minutes, 5 GB

Exp. 2j: CRF with Initial Parameters



Processing: ~21 seconds, 1 MB

Exp. 2j: CRF with Initial Parameters (cont.)



Overall Results

- SVM, CRF, and LDCRF perform best

Overall Results

- SVM, CRF, and LDCRF perform best
- CRF almost outperforms all with negligible memory and processing requirements

Overall Results

- SVM, CRF, and LDCRF perform best
- CRF almost outperforms all with negligible memory and processing requirements
- Hidden states increase accuracy but at significant memory cost

Conclusion

- HCRF is accurate, but has a heavy performance cost.
- May be optimal for particular domains.

References I

Alireza Fathi, Jessica K. Hodgins, and James M. Rehg. Social interactions: A first-person perspective. In *CVPR*, pages 1226–1233. IEEE, 2012. ISBN 978-1-4673-1226-4. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2012.html#FathiHR12.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL http://dl.acm.org/citation.cfm?id=645530.655813.

References II

Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell. Hidden conditional random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29 (10):1848–1852, October 2007. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1124. URL http://dx.doi.org/10.1109/TPAMI.2007.1124.