Multiclass Recognition and Part Localization with Humans in the Loop

Heath Vinicombe

Department of Computer Science
The University of Texas at Austin

30th November 2012

Outline

- Motivation
- System Overview
- Features
- Probabilistic Model
- Prediction
- Results
- Conclusions

Motivation

- Humans vs. Computers

Easy for humans but Harder for Computers

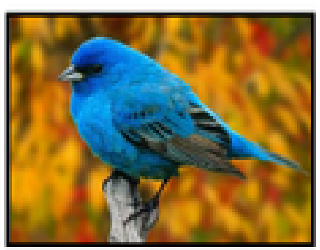
Motivation

- Leveraging abilities of Humans and Computers

Difficult for Humans and Computers

Finch? Bunting?...

Easy for Humans



Yellow Belly? Blue Belly? ...

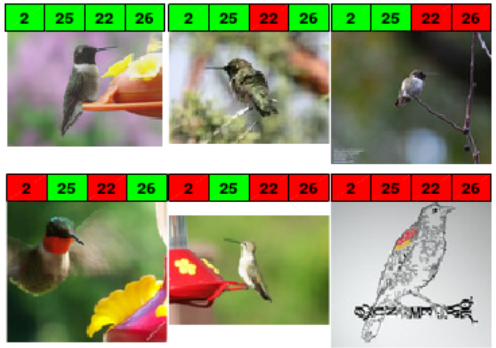
Easy for Computers

$$
\begin{aligned}
w(\ell, \delta t|\ell') &= \frac{1}{\sqrt{2\pi\sigma^2\delta t}} \exp\left[-\frac{(\ell-\ell'-a\delta t)^2}{2\sigma^2\delta t}\right] \\
&+ \frac{1}{\sqrt{2\pi\sigma^2\delta t}} \exp\left[\frac{a(\ell-\ell')}{\sigma^2} - \frac{a^2\delta t}{2\sigma^2}\right] \\
&\times \left\{\exp\left[-\frac{(\ell+\ell')^2}{2\sigma^2\delta t}\right] + \exp\left[-\frac{(2L-\ell-\ell')^2}{2\sigma^2\delta t}\right]\right\} \\
&- \frac{a}{\sigma^2} \exp\left[\frac{2\ell a}{\sigma^2}\right] \operatorname{erfc}\left[\frac{\ell+\ell'+a\delta t}{\sqrt{2\sigma^2\delta t}}\right] + \frac{a}{\sigma^2} \\
&\times \exp\left[-\frac{2(L-\ell)a}{\sigma^2}\right] \operatorname{erfc}\left[\frac{2L-\ell-\ell'-a\delta t}{\sqrt{2\sigma^2\delta t}}\right].
\end{aligned}
$$

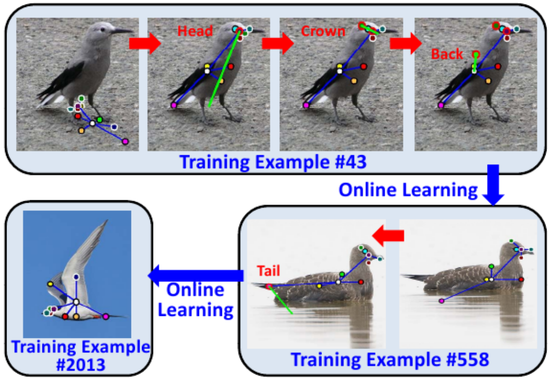
Visipedia

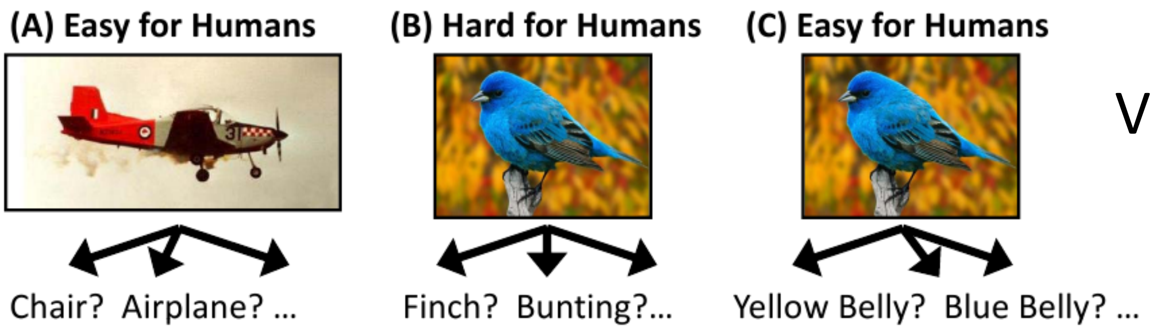
- http://www.vision.caltech.edu/visipedia/
- Visual encyclopedia of images

Online Crowdsourcing



Scalable Structure Learning and Annotation

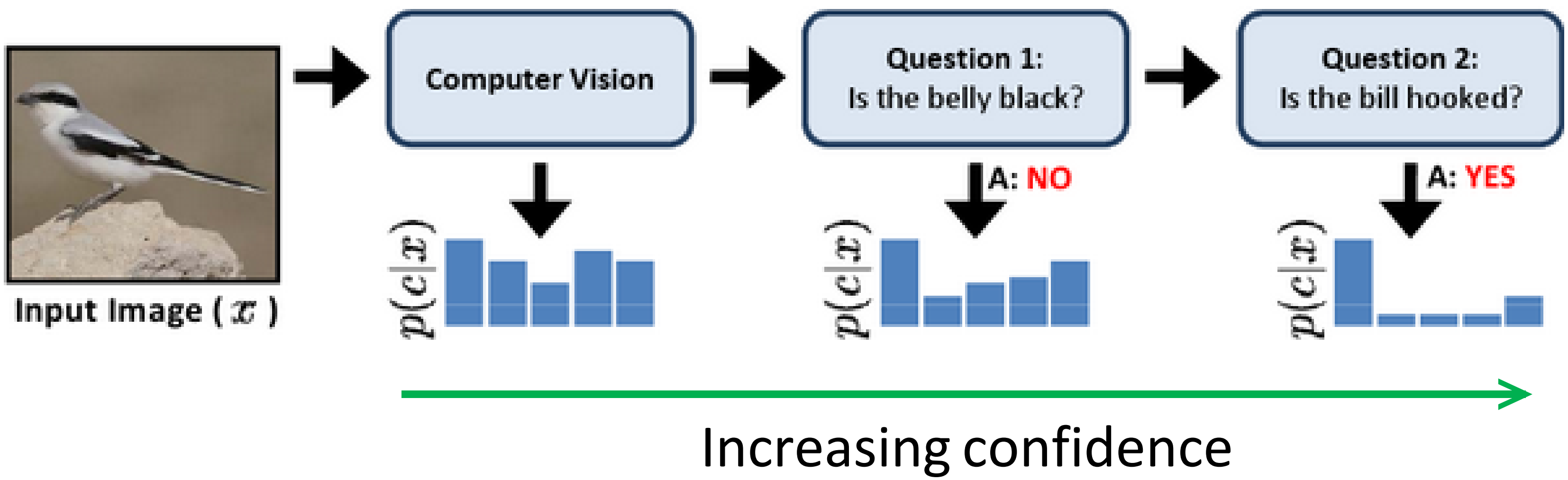




Visual Recognition with Humans in the Loop

System Overview

- Features: Attributes and Parts
- Initial probabilities from Computer Vision
- Answers to questions used to update p(c|x)



Outline

- Motivation
- System Overview
- Features
- Probabilistic Model
- Prediction
- Results
- Conclusions

Features - Attributes

- Binary vector of length 312
- Attribute vector **a**^{c} is property of **class**
- p(**a**^{c} | x) is property of **image**

Class: Big Bird

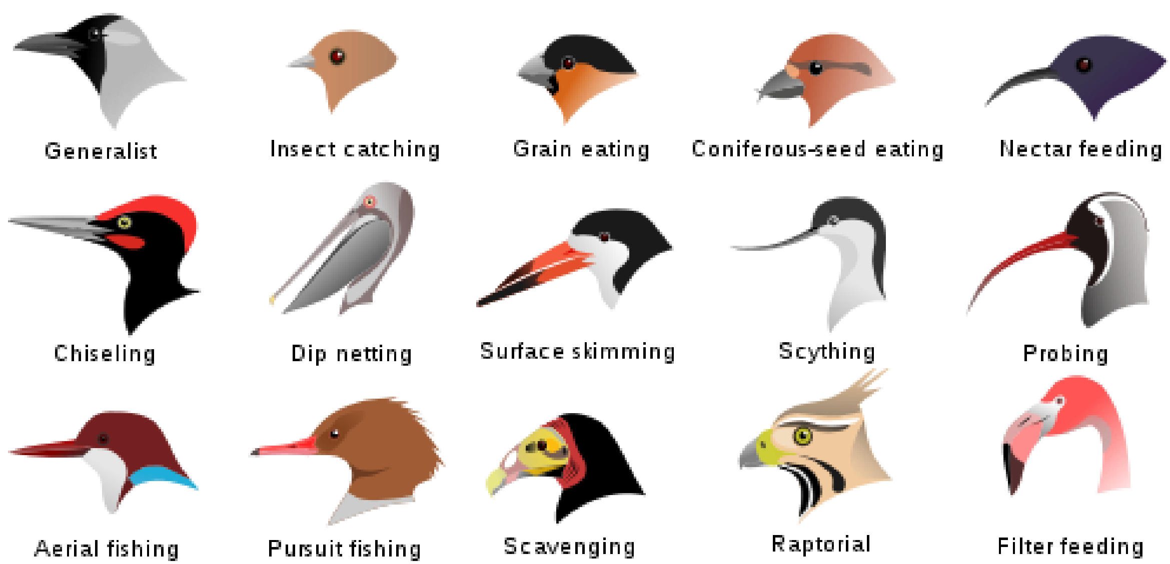
$$\boldsymbol{a}^c = [0\ 1\ 0\ 1\ 1\ 0\ 0\ 0 \ldots]$$

Is Yellow

Voted for Romney

Features – Example Attributes

- has_crown_color::yellow
- has_bill_shape::hooked
- has_head_pattern::striped
- has_size::very large (32 - 72 in)



Features - Parts

- 13 body parts
- 12 aspects

x position y position

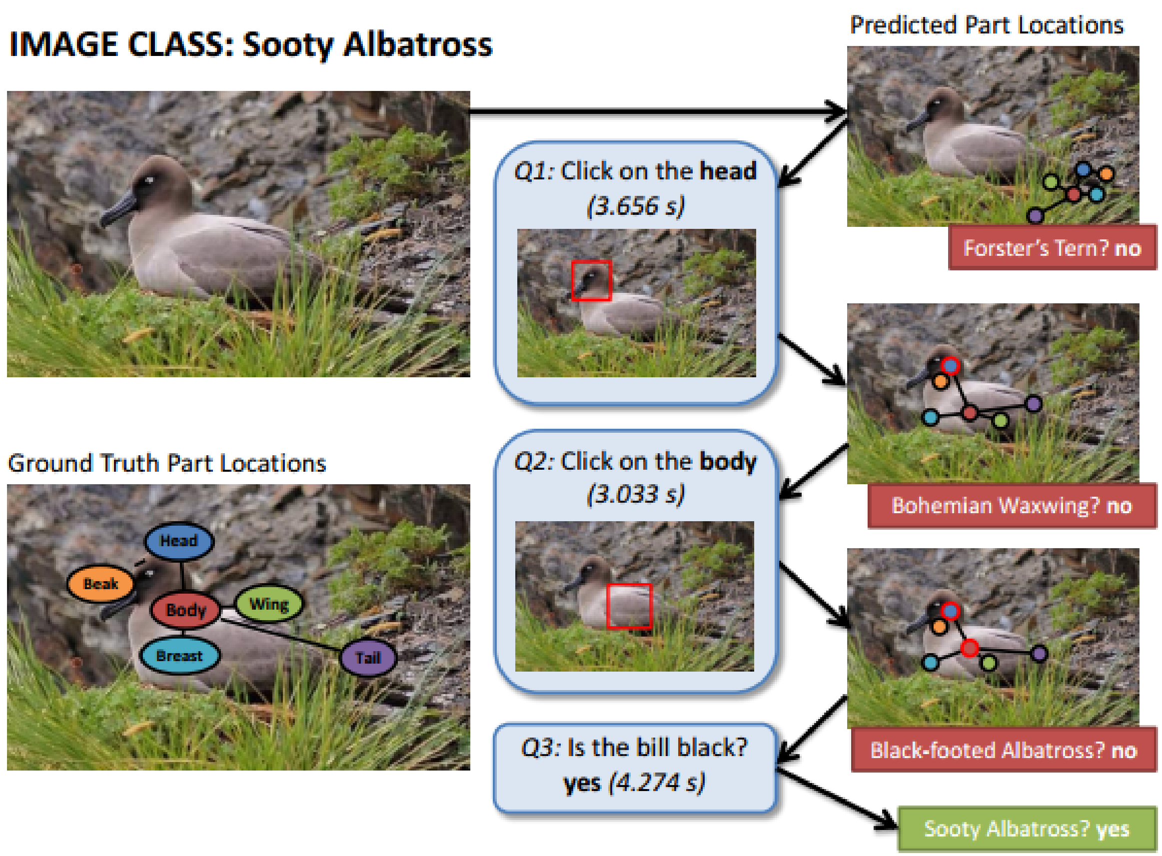
$$\theta_p = \{x_p, y_p, s_p, v_p\}$$

aspect binary visibility

$$\Theta = \{\theta_1, \dots \theta_P ,\}$$

Features – User Questions

- Attribute queries
- Part location queries



Outline

- Motivation
- System Overview
- Features
- Probabilistic Model
- Prediction
- Results
- Conclusions

Probability Model

- $p(c|U^t, x) = \frac{p(\boldsymbol{a}^c, U^t|x)}{\sum_c p(\boldsymbol{a}^c, U^t|x)}$

- $p(\boldsymbol{a}^c, U^t|x) = \int_\Theta p(\boldsymbol{a}^c, U^t, \Theta|x) d\Theta$

- $p(\boldsymbol{a}^c, U^t, \Theta|x) = \underbrace{p(\boldsymbol{a}^c|\Theta, x)}_{\text{Attributes detector}} \underbrace{p(\Theta|x)}_{\text{Parts detector}} \underbrace{p(U^t|\boldsymbol{a}^c, \Theta, x)}_{\text{User's answers to questions}}$

Attribute Detection

- Linear classifier for each $a^c_i \in \boldsymbol{a}^c$
- SIFT and RGB quantized to 128 codewords
- Independence assumption

- $p(\boldsymbol{a}^c|\Theta, x) = \prod_{a^c_i \in \boldsymbol{a}^c} \underbrace{p(a^c_i|\theta_{part(a_i)}, x)}_{\text{output of linear classifier}}$

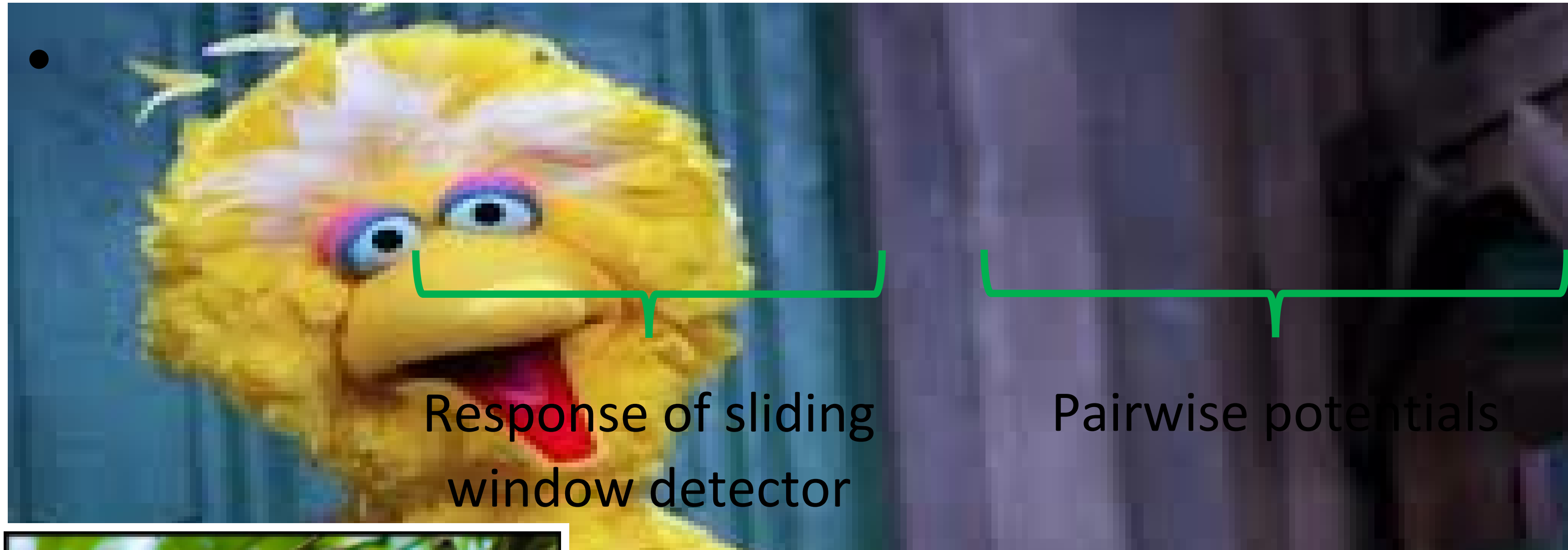
Single Attribute

Full Attribute Vector

Discussion

- Why such a simple choice of attribute detector?
- Is the independence assumption in calculating $\boldsymbol{a}^c$ appropriate?

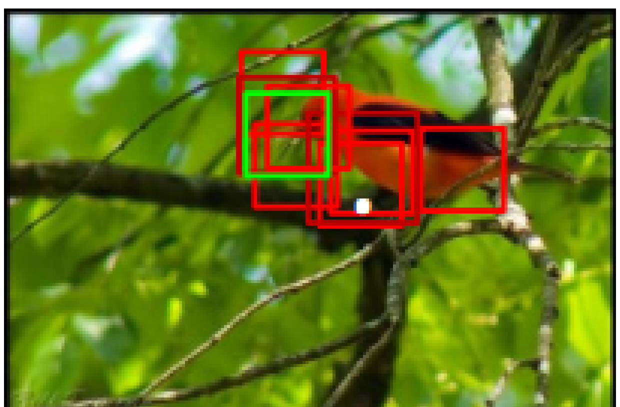
Part Detection



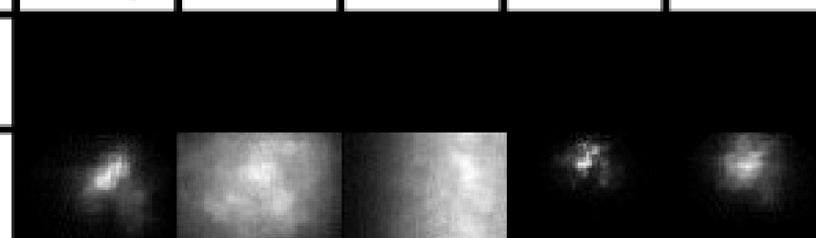
-

Response of sliding window detector

Pairwise potentials



Pose	body	breast	tail	head	throat
Facing_ right					
Facing_ left					



Discussion

- In this case are the pairwise potential terms useful or not?

User Model

- Models likelihood of user's answers based on current hypothesis

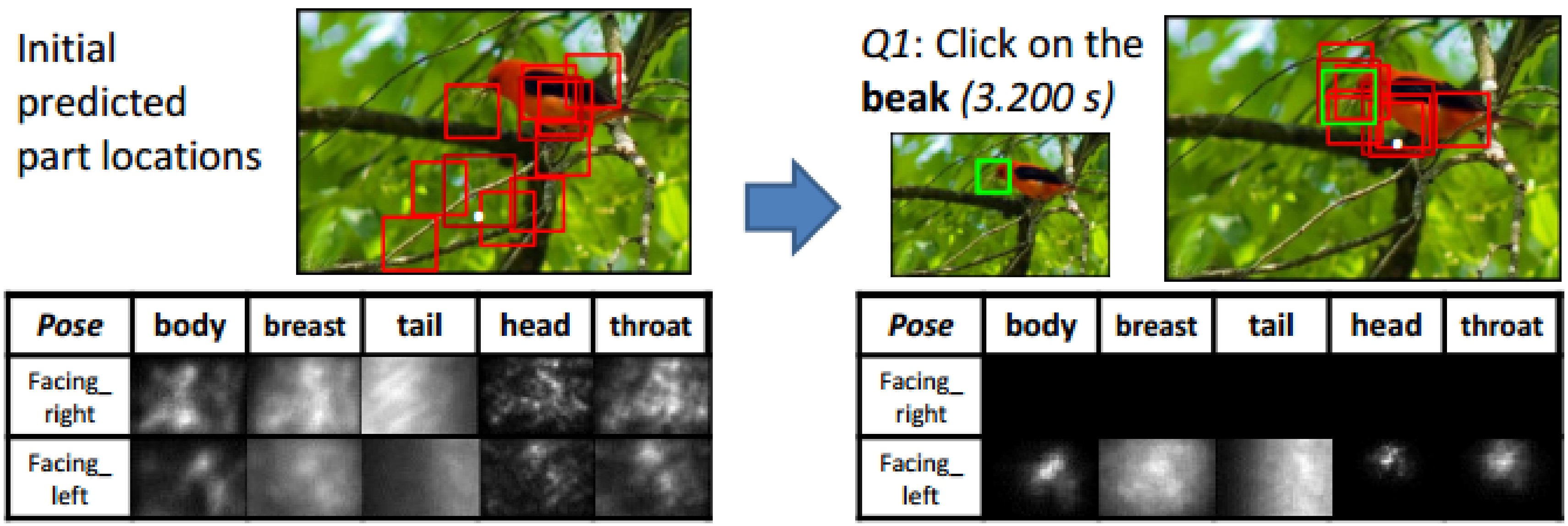
$$p(U^t|\boldsymbol{a}^c, \Theta, x) = \underbrace{\prod_{p \in U^t_{\Theta}} p(\tilde{\theta}_p|\theta_p)}_{\substack{\text{Part Locations:} \\ \text{Normal distribution}}} \underbrace{\prod_{\tilde{a}_i \in U^t_a} p(\tilde{a}_i|a^c{}_i)}_{\substack{\text{Attribute Values:} \\ \text{Binomial distribution}}}$$

Outline

- Motivation
- System Overview
- Features
- Probabilistic Model
- Inference
- Results
- Conclusions

Inference

- Inference updates probabilities after each question



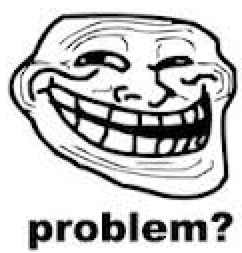
Inference

- We need to evaluate:

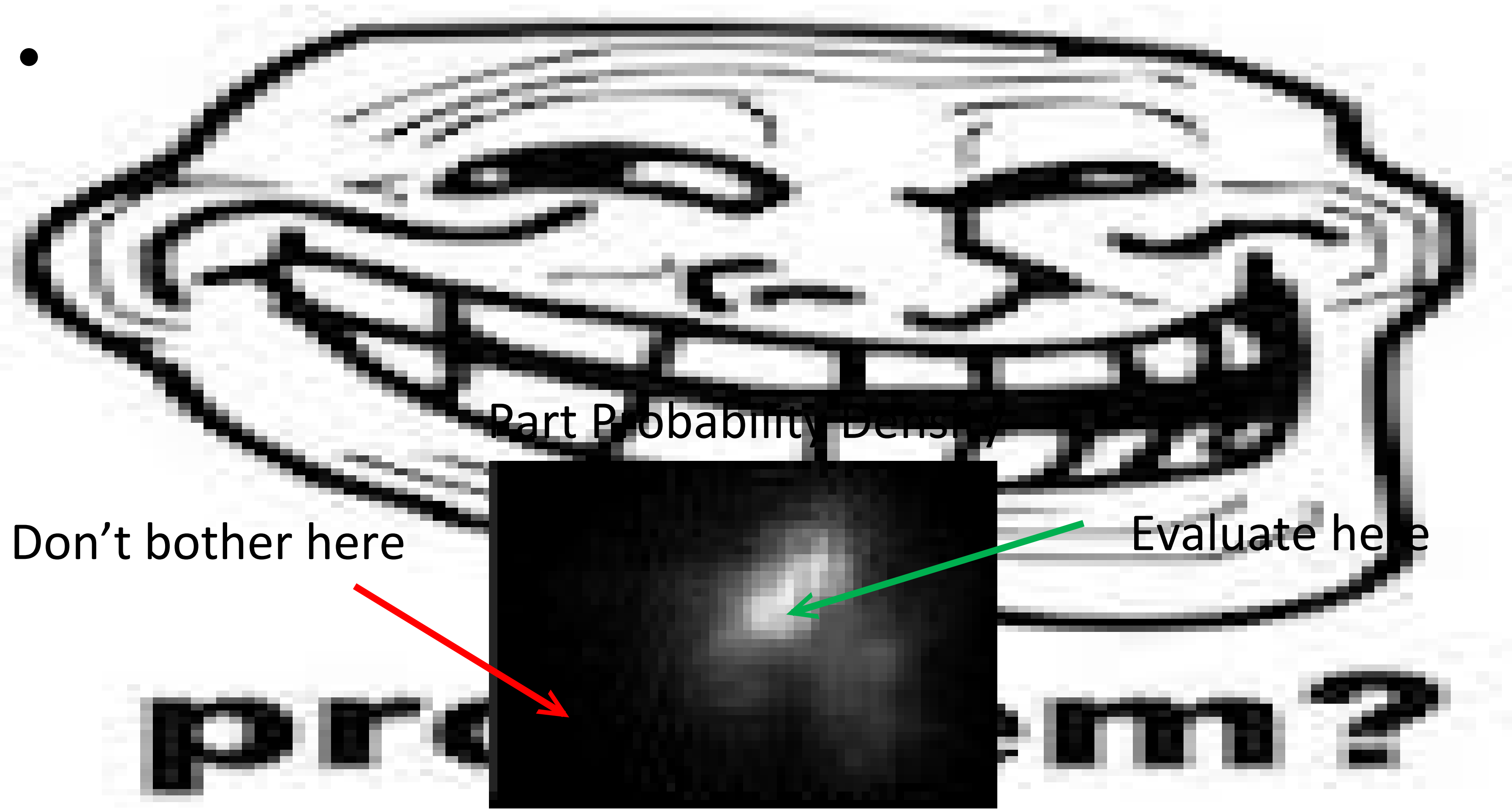
$$\int_{\Theta} p(\boldsymbol{a}^c|\Theta, x)p(\Theta|x)p(U^t|\boldsymbol{a}^c, \Theta, x)\, \mathrm{d}\Theta$$

For all possible combinations of:

- Classes: 200 in total
- Part Locations: ~1000's of windows per part
- Exponential in number of parts



Inference

- 

Choice of Questions

- Minimize user input by asking “best” questions

Two candidate classes

Bad question: Is the head white?

Good question: ???

Information Gain

- Expected change in Entropy
- Entropy:

$$H = -\sum_{i=1}^{n} P(x_i) \ln P(x_i)$$

High Entropy RV: H = 1.38

Low Entropy RV: H = 0.71

Discussion

- What other factors should be taken into consideration when choosing a question?

Selection by Time

- Want to minimize time rather than number of questions required
- Expected time of questions vary

$$Maximize \quad \frac{IG(q_j)}{\mathbb{E}[time(q_j)]}$$

Outline

- Motivation
- System Overview
- Features
- Probabilistic Model
- Prediction
- Results
- Conclusions

Dataset

- Caltech-UCSD Birds 200 (CUB-200)
- 11,800 images of birds
- 200 classes
- 312 binary attributes
- 15 part labels
- Part labels obtained through MTurk

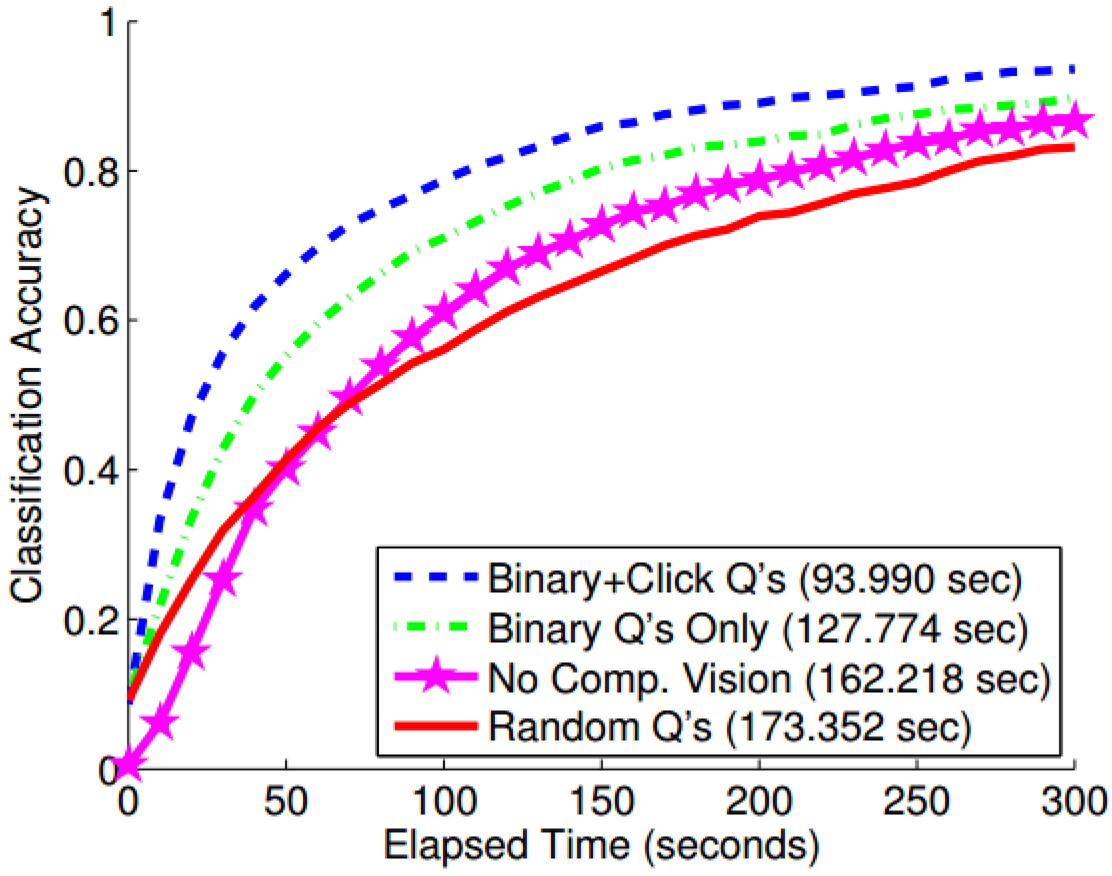
Dataset

- Expected change in
- Entropy:

$$H = - \sum^{n}$$

Results

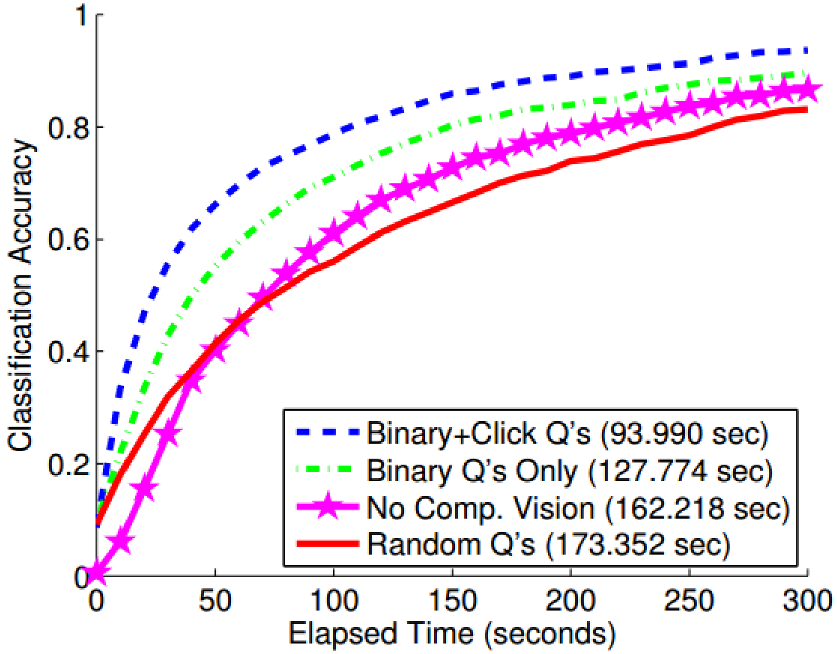
- Time to classify using IG criterion



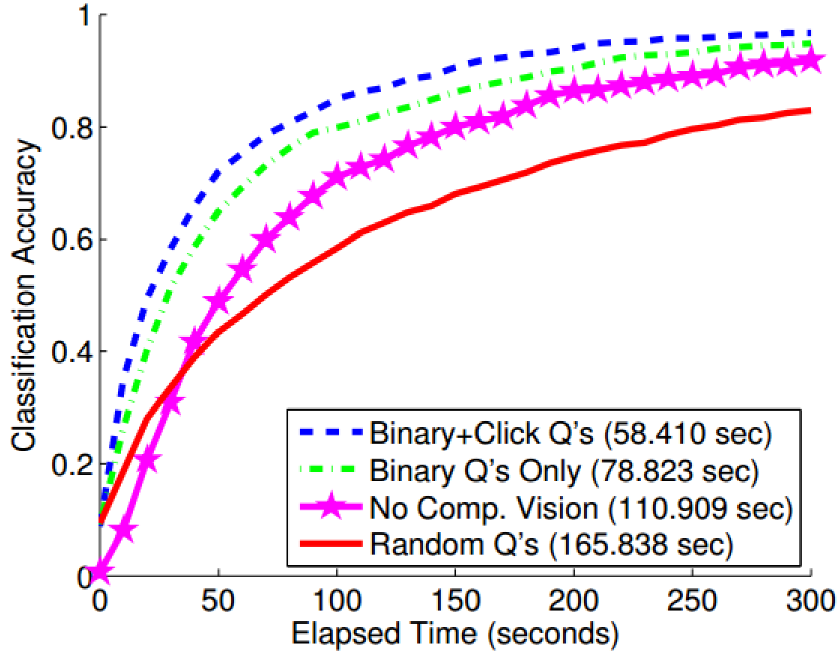
Results

- Comparison of criterion

Information Gain criterion



Time criterion



Results Analysis

- Computer Vision reduces time to classify
- Time criterion reduces time to classify
- Part localization improves performance (attribute detectors 17.3% on ground truth locations vs. 10.3% on predicted)
- Part localization questions are quicker to answer (3s vs. 7.6s)

Future Work

- Visipedia iPad App

Interactive Part Labeling

- Video

Conclusion

- Better performance by combining strengths of humans and computers
- Using two types of questions and simple computer vision, bird species are classified in ~ 60s
- Human input can “guide” computer vision algorithms to produce better results

References

- Multiclass Recognition and Part Localization with Humans in the Loop. C. Wah et al. ICCV 2011
- http://www.vision.caltech.edu/visipedia/index.html
- A Discriminatively Trained, Multiscale, Deformable Part Model, by P. Felzenszwalb, D. McAllester and D. Ramanan