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

Chair? Airplane? ...
Motivation

- Leveraging abilities of Humans and Computers

Difficult for Humans and Computers

Easy for Humans

Yellow Belly? Blue Belly? ...

Easy for Computers

\[
w(\ell, \delta t | \ell') = \frac{1}{\sqrt{2\pi \sigma^2 \delta t}} \exp \left[ -\frac{(\ell - \ell' - a\delta t)^2}{2\sigma^2 \delta t} \right] \\
+ \frac{1}{\sqrt{2\pi \sigma^2 \delta t}} \exp \left[ \frac{\alpha(\ell - \ell')}{\sigma^2} - \frac{\alpha^2 \delta t}{2\sigma^2} \right] \\
\times \left\{ \exp \left[ -\frac{(\ell + \ell')^2}{2\sigma^2 \delta t} \right] + \exp \left[ -\frac{(2L - \ell - \ell')^2}{2\sigma^2 \delta t} \right] \right\} \\
- \frac{a}{\sigma^2} \exp \left[ \frac{2(L - \ell) a^2}{\sigma^2} \right] \text{erfc} \left[ \frac{\ell + \ell' + a\delta t}{\sqrt{2\sigma^2 \delta t}} + \frac{a}{\sigma^2} \right] \\
\times \exp \left[ -\frac{2(L - \ell) a^2}{\sigma^2} \right] \text{erfc} \left[ \frac{2L - \ell - \ell' - a\delta t}{\sqrt{2\sigma^2 \delta t}} \right]
\]
Visipedia

- [http://www.vision.caltech.edu/visipedia/](http://www.vision.caltech.edu/visipedia/)
- Visual encyclopedia of images

**Online Crowdsourcing**

**Scalable Structure Learning and Annotation**

(A) Easy for Humans  (B) Hard for Humans  (C) Easy for Humans

Chair?  Airplane?  ...  Finch?  Bunting?  ...  Yellow Belly?  Blue Belly?  ...

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

$$a^c = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & \ldots \end{bmatrix}$$

- 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

\[ \theta_p = \{x_p, y_p, s_p, v_p\} \]

\[ \Theta = \{\theta_1, \ldots, \theta_p\} \]
Features – User Questions

- Attribute queries
- Part location queries
Outline

• Motivation
• System Overview
• Features
• Probabilistic Model
• Prediction
• Results
• Conclusions
Probability Model

\[ p(c|U^t, x) = \frac{p(a^c, U^t|x)}{\sum_c p(a^c, U^t|x)} \]

\[ p(a^c, U^t|x) = \int_{\Theta} p(a^c, U^t, \Theta|x) d\Theta \]

\[ p(a^c, U^t, \Theta|x) = p(a^c|\Theta, x)p(\Theta|x)p(U^t|a^c, \Theta, x) \]

Attributes detector
Parts detector
User’s answers to questions
Attribute Detection

• Linear classifier for each \( a^c_i \in \mathbf{a}^c \)
• SIFT and RGB quantized to 128 codewords
• Independence assumption

\[
p(\mathbf{a}^c | \Theta, x) = \prod_{a^c_i \in \mathbf{a}^c} p(a^c_i | \theta_{part(a^c_i), x})
\]

Single Attribute  Full Attribute Vector  output of linear classifier
Discussion

• Why such a simple choice of attribute detector?
• Is the independence assumption in calculating $\alpha^c$ appropriate?
Part Detection

- Response of sliding window detector
- Pairwise potentials
Discussion

• In this case are the pairwise potential terms useful or not?
User Model

• Models likelihood of user’s answers based on current hypothesis

\[ p(U^t | \alpha^c, \Theta, x) = \prod_{p \in U^t_{\Theta}} p(\tilde{\theta}_p | \theta_p) \prod_{\tilde{a}_i \in U^t_a} p(\tilde{a}_i | \alpha^c_i) \]

Part Locations: Normal distribution
Attribute Values: Binomial distribution
Outline

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

- Inference updates probabilities after each question
Inference

• We need to evaluate:

\[ \int_{\Theta} p(\alpha^c | \Theta, x)p(\Theta | x)p(U^t | \alpha^c, \Theta, x) \, d\Theta \]

For all possible combinations of:

• Classes: 200 in total
• Part Locations: ~1000’s of windows per part
• Exponential in number of parts
Inference

- Don’t bother here
- Evaluate here

Part Probability Density
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

\[
\text{Maximize } \frac{IG(q_j)}{\mathbb{E}[\text{time}(q_j)]}
\]
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 \( l \)

• Entropy:

\[
H = - \sum_{n} \]
Results

- Time to classify using IG criterion
Results

• Comparison of criterion

Information Gain criterion

Time criterion

- Binary+Click Q’s (93.990 sec)
- Binary Q’s Only (127.774 sec)
- No Comp. Vision (162.218 sec)
- Random Q’s (173.352 sec)

- Binary+Click Q’s (58.410 sec)
- Binary Q’s Only (78.823 sec)
- No Comp. Vision (110.909 sec)
- Random Q’s (165.838 sec)
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