Similarity Comparisons for Interactive Fine-Grained Categorization

Catherine Wah[1] Grant Van Horn[1] Steve Branson[2] Subhransu Maji[3] Pietro Perona[2] Serge Belongie[4]

[1]University of California, San Diego vision.ucsd.edu

[2]California Institute of Technology vision.caltech.edu

[3] Toyota Technological Institute at Chicago ttic.edu

[4] Cornell Tech tech.cornell.edu

Overview

Problem

- *Parts and attributes* exhibit weaknesses
  - ➢ Scalability issues; costly; reliance on experts, but experts are scarce

Proposed Solution

- Use *relative similarity comparisons* to reduce dependence on expert-derived part and attribute vocabularies

Contributions

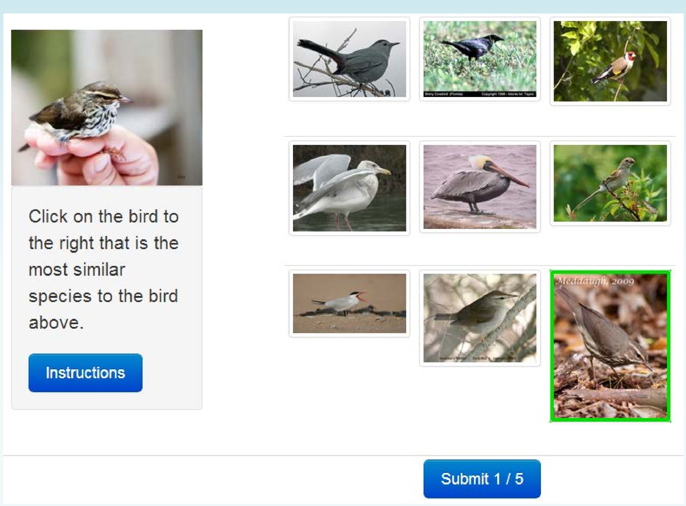
- We present an efficient, flexible, and scalable system for interactive fine-grained visual categorization
  - ➢ Based on perceptual similarity
  - ➢ Combines similarity metrics and computer vision methods in a unified framework
- Outperforms state-of-the-art relevance feedback-based and part/attribute-based approaches

Similarity Comparisons

A. Collect grid-based similarity comparisons that do not require prior expertise

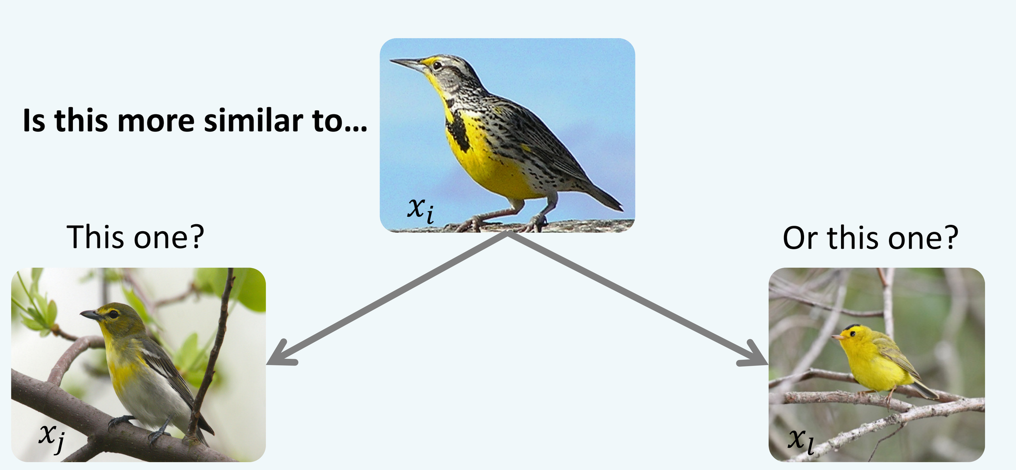
B. Broadcast grid-based comparisons to *triplet comparisons*

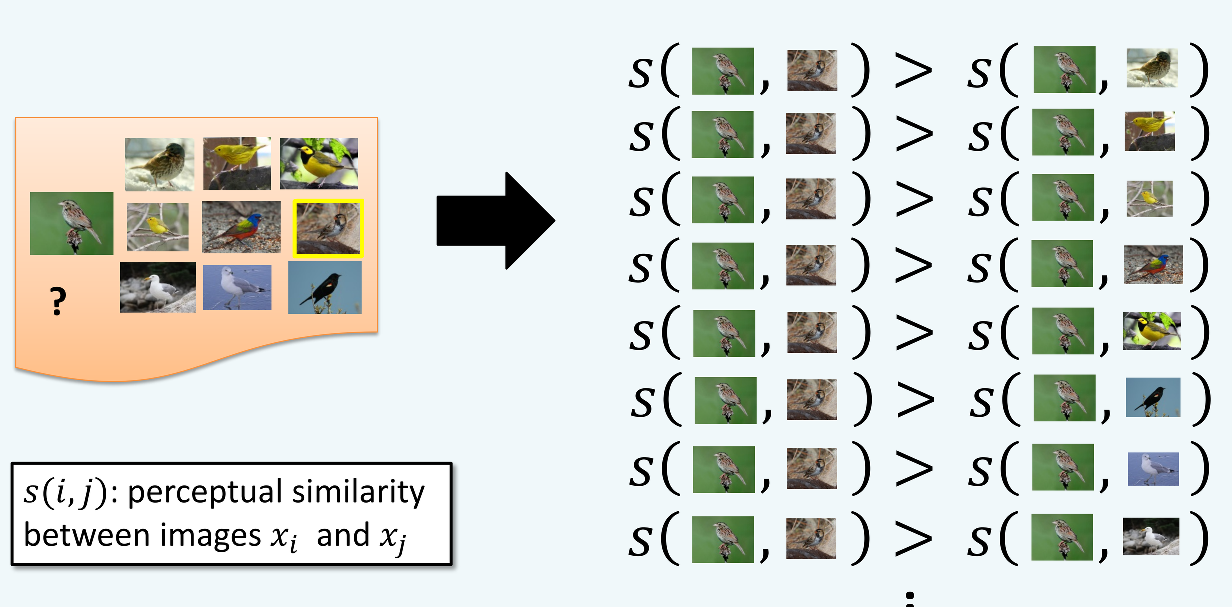
A



B

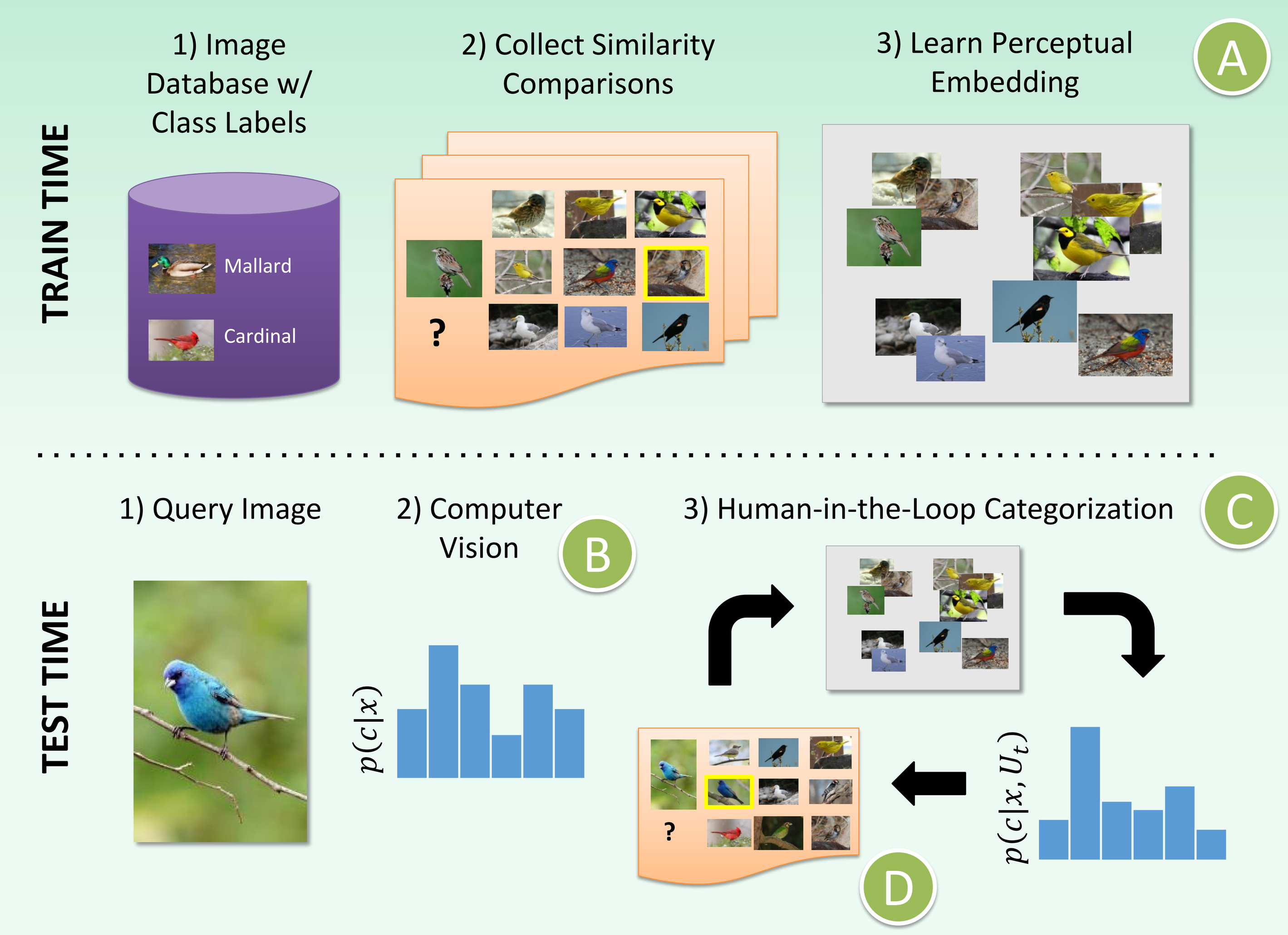
$$\mathcal{T} = \{(i,j,l) | x_i \text{ more similar to } x_j \text{ than } x_l\}$$





$s(i,j)$: perceptual similarity between images x_i and x_j

Approach



x Query image	t Time step	$\mathbf{z}$ True location of x in
c Class	U_t User responses at t	perceptual space

INTERACTIVE CATEGORIZATION (C) (D)

- Compute per-class probabilities as:

$$p(c,|x,U_t) \propto p(c,U_t|x) = \int_{\mathbf{z}} p(c,\mathbf{z},U_t|x)d\mathbf{z}$$

where

$$w^t = p(c,\mathbf{z},U_t|x) = p(U_t|\,c,\mathbf{z},x)\,p(c,\mathbf{z}|x)$$

Efficient computation

- Approximate per-class probabilities as:

$$p(c,|x,U_t) \approx \frac{\sum_{k,c_k=c} w_k^t}{\sum_k w_k^t}$$

i.e. sum of weights of examples of class c where k enumerates training examples

- Weight w_k represents how likely $\mathbf{z}_k$ is true location $\mathbf{z}$:

$$w_k^t = p(c_k,\mathbf{z}_k,U_t|x) = p(U_t|\,c_k,\mathbf{z}_k,x)\,p(c_k,\mathbf{z}_k|x)$$

such that

$$w_k^{t+1} = p(u_{t+1}|\mathbf{z}_k)w_k^t = \frac{\phi(S_{ik})}{\sum_{j\in D}\phi(S_{jk})}w_k^t$$

Efficient update rule:

1. Initialize weights $w_k^0 = p(c_k,\mathbf{z}_k|x)$
2. Update weights w_k^{t+1} when user answers a similarity question
3. Update per-class probabilities

Learning a Metric (A)

- Given set of triplet comparisons $\mathcal{T}$, learn embedding $\mathbf{Z}$ of N training images with *stochastic triplet embedding* [van der Maaten & Weinberger 2012]
- From $\mathbf{Z}$, generate similarity matrix $S \in N \times N$

Computer Vision (B)

- Easy to map off-the-shelf CV algorithms into framework, *e.g.*, multiclass classification scores

$$p(c,\mathbf{z}|x) \propto p(c|x)$$

Incorporating Users (D)

- D is grid of images for each question Incorporate independent user response as:

$$p(u|\mathbf{z}) = \frac{\phi(s(\mathbf{z},\mathbf{z}_i))}{\sum_{j\in D}\phi(s(\mathbf{z},\mathbf{z}_j))}$$

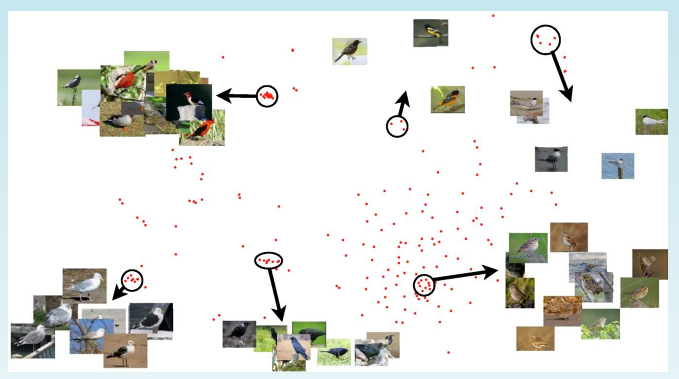
Selecting the Display (D)

- **Approximate solution**: maximizes expected information gain in terms of entropy of $p(c,\mathbf{z}_k,U_t|x)$
- Group images into equal-weight clusters [Fang & Geman 2005]
- From each cluster, select image with largest w_k^t

Results

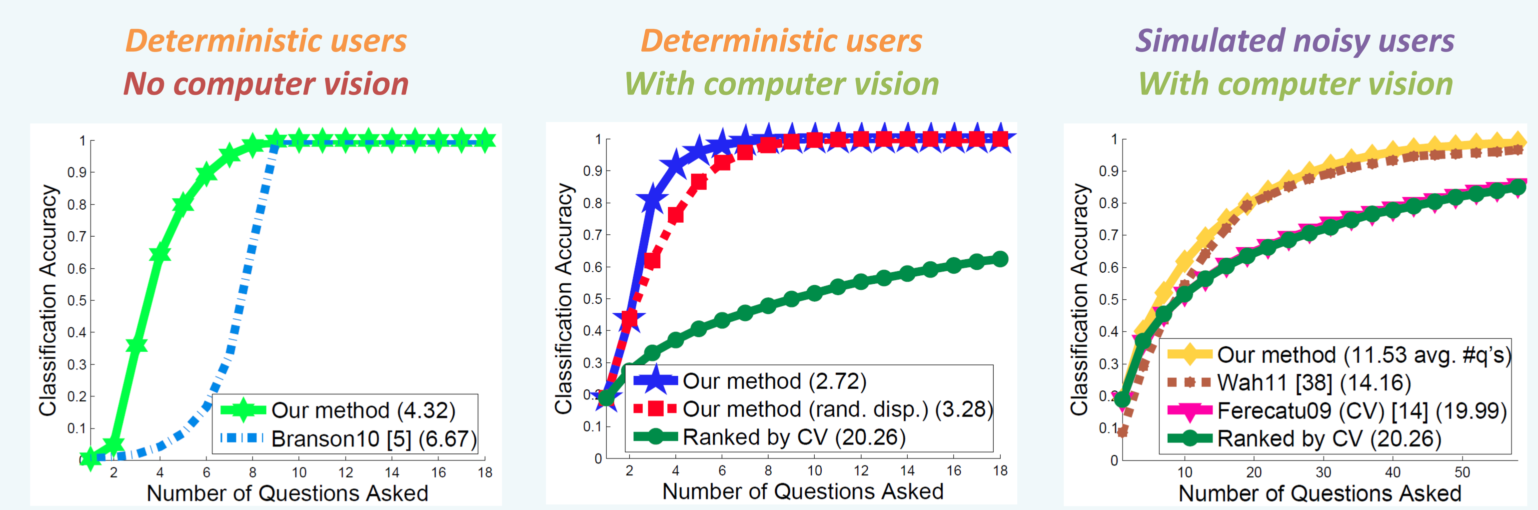
Learned Embedding

- Learn category-level embedding of $N = 200$ nodes
- Category-level embedding requires much fewer comparisons compared to at the instance-level



Interactive Categorization

- Similarity comparisons are advantageous compared to part/attribute questions
- Using computer vision reduces the burden on the user
- Intelligently selecting image displays reduces effort
- The system is robust to user noise



Multiple Metrics

- System supports multiple similarity metrics as different types of questions
- Simulate perceptual spaces using CUB-200-2011 attribute annotations

Method	Avg. #Qs
CV, Color Similarity	2.70
CV, Shape Similarity	2.67
CV, Pattern Similarity	2.67
CV, Color/Shape/Pattern Similarity	**2.64**
No CV, Color/Shape/Pattern Similarity	4.21

Qualitative Results