Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers

by Abhinav Gupta & Larry S. Davis
presented by Arvie Frydenlund

Paper information

- ECCV 2008
- Slides at http://www.cs.cmu.edu/ ∼abhinavg/
- http://www.cs.cmu.edu/%7Eabhinavg/eccv2008.ppt

Objectives of the paper

Task:

- Auto-annotation of image regions to labels

Methods:

- Two models learned
- Training model
  - Learns classifiers for nouns and relationships at the same time
  - Learns priors on possible relationships for pairs of nouns
- Inference model given the above classifiers and priors

Issues:

- Dataset is weakly labeled
- Not all labels are used all the time in the dataset

Weakly labeled data

President Obama debates **Mitt Romney**, while the **audience** sits in the background. (while the audience sits *behind* the debaters)

Co-occurrence Ambiguities

Only have images of cars that include a street

A **man** beside a **car** on the **street** in front of a **fence**.

Noun relationships

- On(Car, Street)
- P(red labeling) $>$ P(blue labeling)

Prepositions and comparative adjective

Most common prepositions:

- above, across, after, against, along, at, behind, below, beneath, beside, between, beyond, by, down, during, in, inside, into, near, off, on onto, out, outside, over
- since, till, after, before, from, past, to, around, though, thoughout
- for, except, about, like, of

Comparative adjective:

- larger, smaller, taller, heavier, faster

Relationships Actually Used

Used 19 in total

- above, behind, beside, more textured, brighter, in, greener, larger, left, near, far, from, ontopof, more blue, right, similar, smaller, taller, shorter

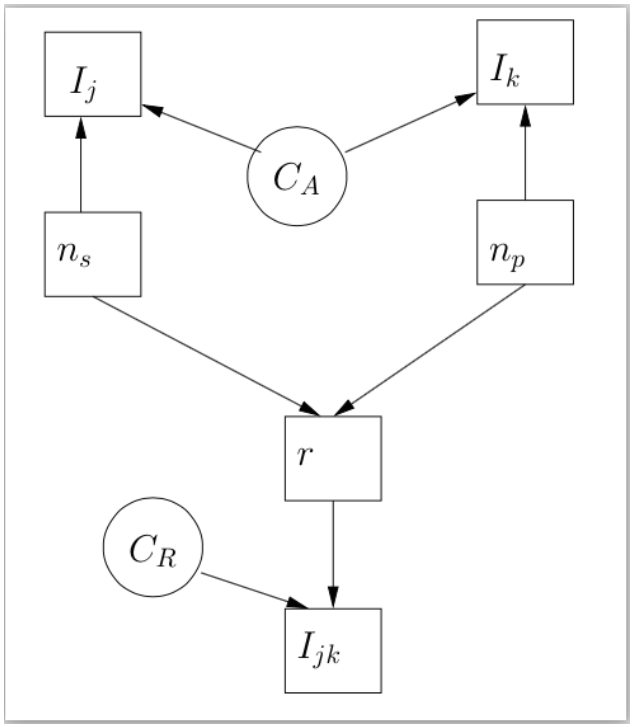
Images and regions

- Each image is pre-segmented and (weakly) annotated by a set of nouns and relations between the nouns
- Regions are represented by a feature vector based on:
  - Appearance (RGB, Intensity)
  - Shape (Convexity, Moments)
- Models for nouns are based on features of the regions
- Relationships models are based in differential features:
  - Difference of average intensity
  - Difference of location

Egg-Chicken

- Learning models for the nouns and relationships requires assigning labels
- Assigning labels requires some model for nouns and relationships
- Solution is to use EM:
  - E: compute noun annotation assignments to labels given old parameters
  - M: compute new parameters given the the E-step assignments
- Classifiers are initialized by previous automated-annotation methods i.e. Duygulu *et al.*, Object recognition as machine translation, EECV (2002)

Generative training model



- C_A and C_R are classifiers (models) for the noun assignments and relationships
- I_j and I_k are region features for regions j and k. I_{jk} are the differential features.
- n_s and n_p are two nouns.
- r is a relationship.
- $L(\theta) = (C_A, C_R)$

Fig. 2 from A. Gupta & L.S. Davis

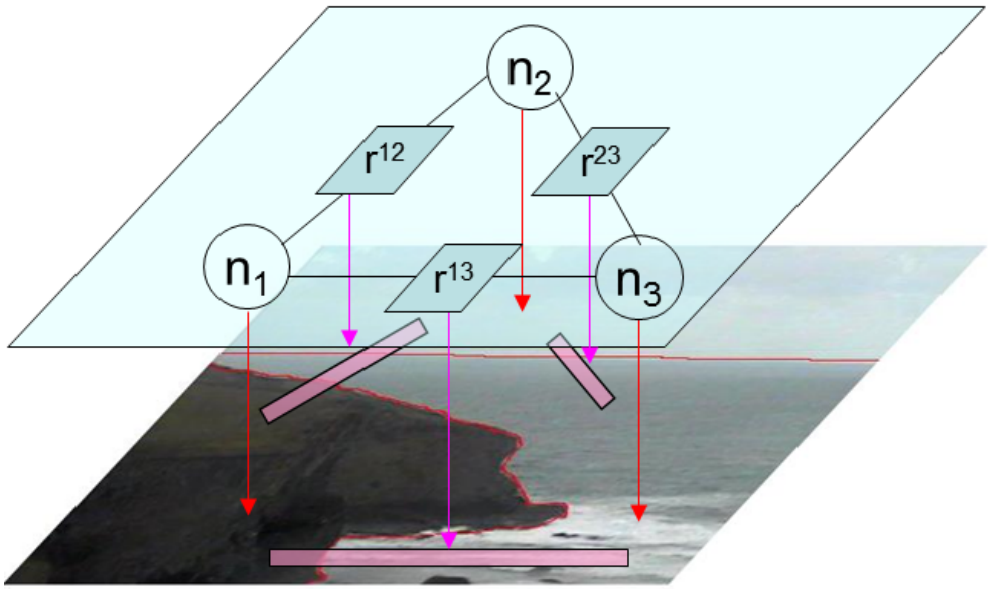
Training

- Too expensive to evaluate $L(\theta)$ directly
- Use EM to estimate $L(\theta)$, with assignments as hidden values.
- Assume predicates are independent given image and assignment
- Obviously wrong, since most predicates preclude others
- Can't be 'on top of' and 'beside'

Training relationships modelled

- C_A, noun model, is implemented as a nearest neighbour based likelihood model
- C_R, relationship mode, is implemented as a decision stump based likelihood model
- Most relationships are modelled correctly
- A few were not
  - **In**: 'Not captured by colour, shape, and location'(?)
  - **on-top-of**
  - **taller** due to poor segmentation algorithm

Inference model



- Given trained C_A and C_R from the above model
- Find $P(n_1, n_2, ...|I_1, I_2, ..., C_A, C_R)$
- Each region represented by a noun node
- Edges between nodes are weighted by the likelihood obtained by differential features

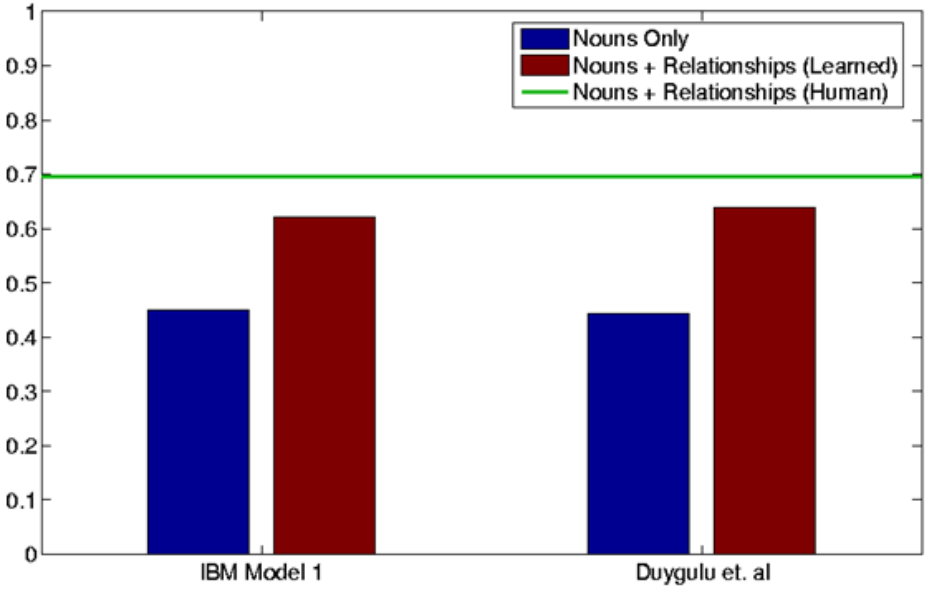
Fig. 3 from A. Gupta & L.S. Davis

Experimental setup

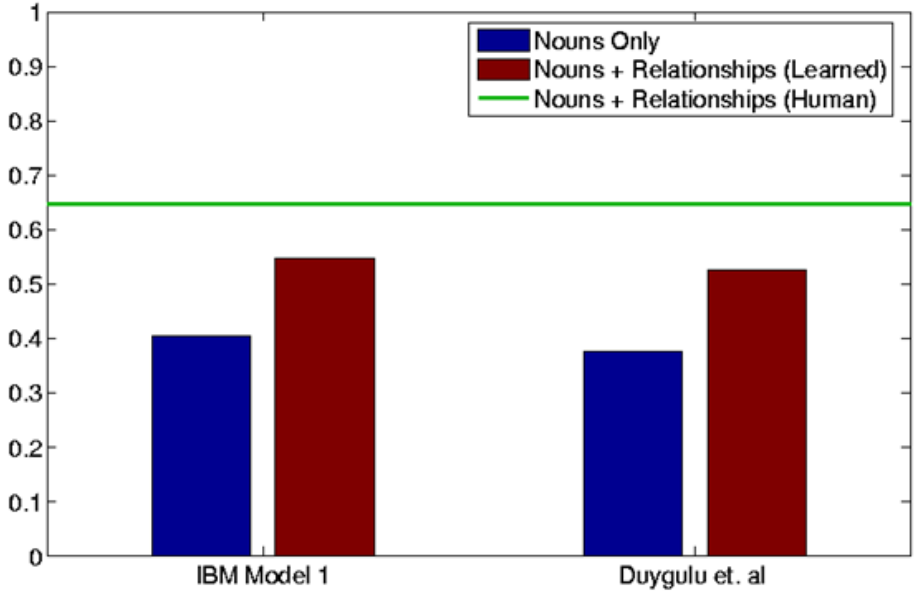
- Corel5K dataset
- 850 training images, tagged with nouns and manually labeled relationships
- Vocabulary size 173 nouns, 19 relationships
- Same segmentation and feature vectors as Duygulu *et al.*, Object recognition as machine translation, EECV (2002)
- Training model test set 150 images (from training set)
- Inference model test set 100 images (given that those images have the same vocabulary)

Training model evaluation

- Use two metrics:
  - Range semantics: counts number of correctly labeled words, while treating each label with the same weight
  - Frequency counts: counts number of correctly labeled regions, which weights more frequent words heigher
- Compared to simple IBM1 (MT model, 1993) and Duygulu *et al.*, MT model



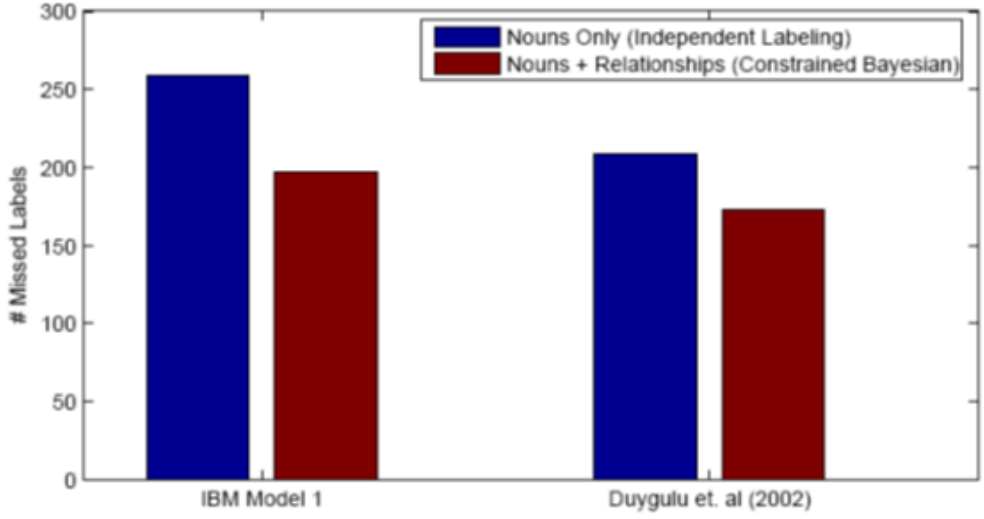
(a) Frequency Correct



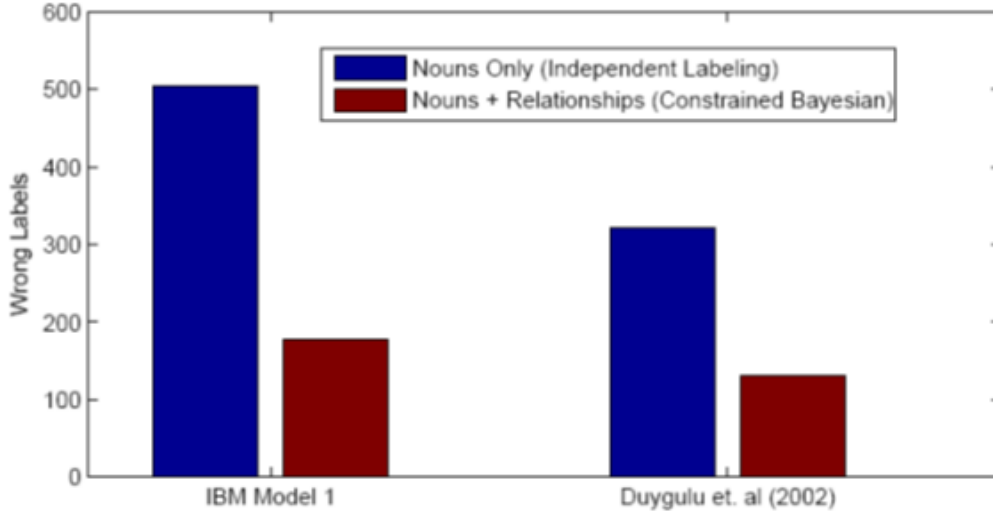
(b) Semantic Range

Inference model evaluation

- Annotating unseen images
- Doesn't use Corel annotations due to missing labels
- 24% and 17% reduction in missed labels
- 63% and 59% reduction false labels

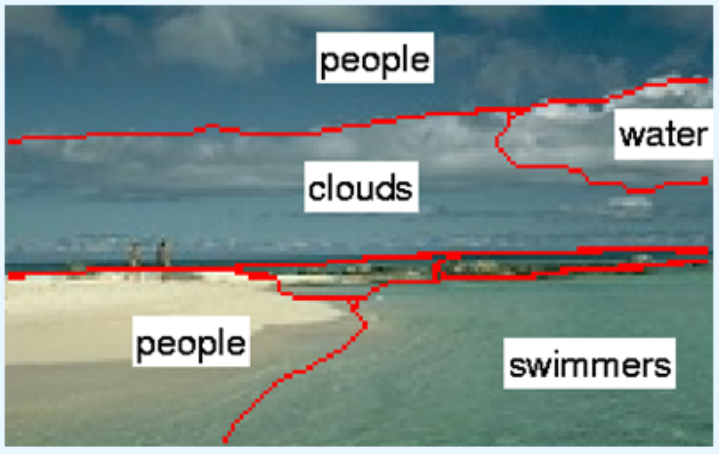


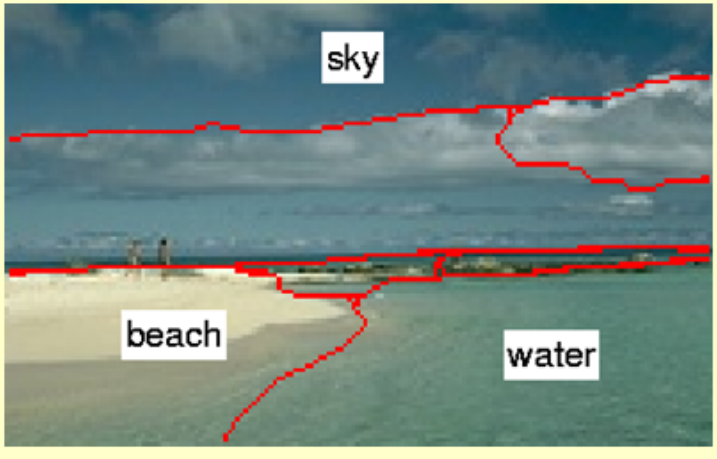
(a) Missed Labels

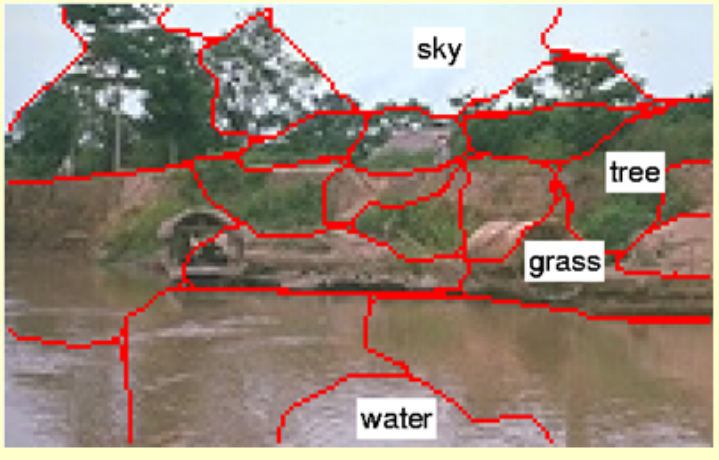


(b) False Labels

Inference model examples







Duygulu *et al.* is the top and the paper's results are the bottom

Inference model Precision-Recall

Duygulu *et al.* is [1]

	Recall		**Precision**	
	[1]	Ours	[1]	Ours
Water	0.79	0.90	0.57	0.67
Grass	0.70	1.00	0.84	0.79
Clouds	0.27	0.27	0.76	0.88
Buildings	0.25	0.42	0.68	0.80
Sun	0.57	0.57	0.77	1.00
Sky	0.60	0.93	0.98	1.00
Tree	0.66	0.75	0.7	0.75

Novelties and limitations

Achievements:

- Novel use of prepositions and comparative adjectives for automatic annotation
- Use previous annotation models for bootstrapping
- Good results

Limitations:

- Only uses two argument predicates, results in ‘greener’
- Can’t do pink flower example
- Assumes one to one relationship between nouns and image segments

Questions?

- One of the motivations was the co-occurrence problem. Wouldn't a simpler model with better training data solve this problem?
- Image caption generation to annotation stack?
- Model simplification: assuming independence of predicates?
- Scale with vocabulary and number of relationships used? 'Bluer' and 'greener' work for outdoor scenes