Query-by-Example Image Retrieval using Visual Dependency Representations

Desmond Elliott, Victor Lavrenko, Frank Keller

University of Edinburgh

August 24, 2014

Query-by-Example Image Retrieval

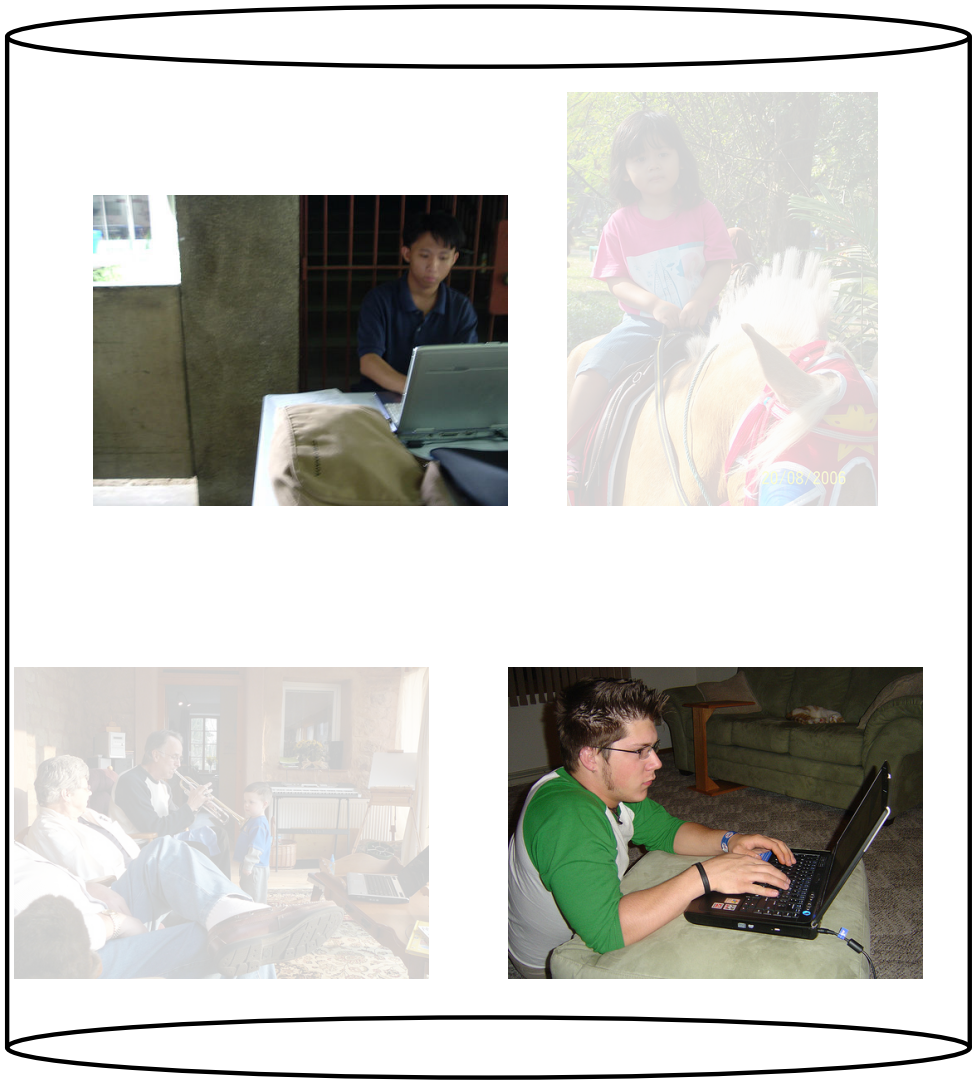
Query-by-Example Image Retrieval

In this talk, similar means same action

Query-by-Example Image Retrieval

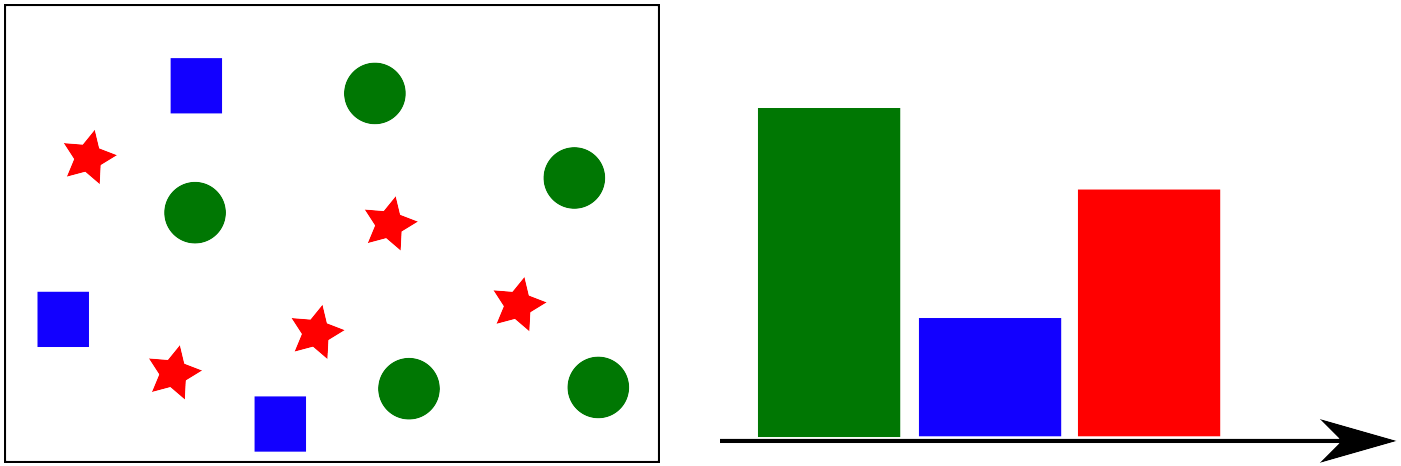
Query-by-Example Image Retrieval

Query-by-Example Image Retrieval



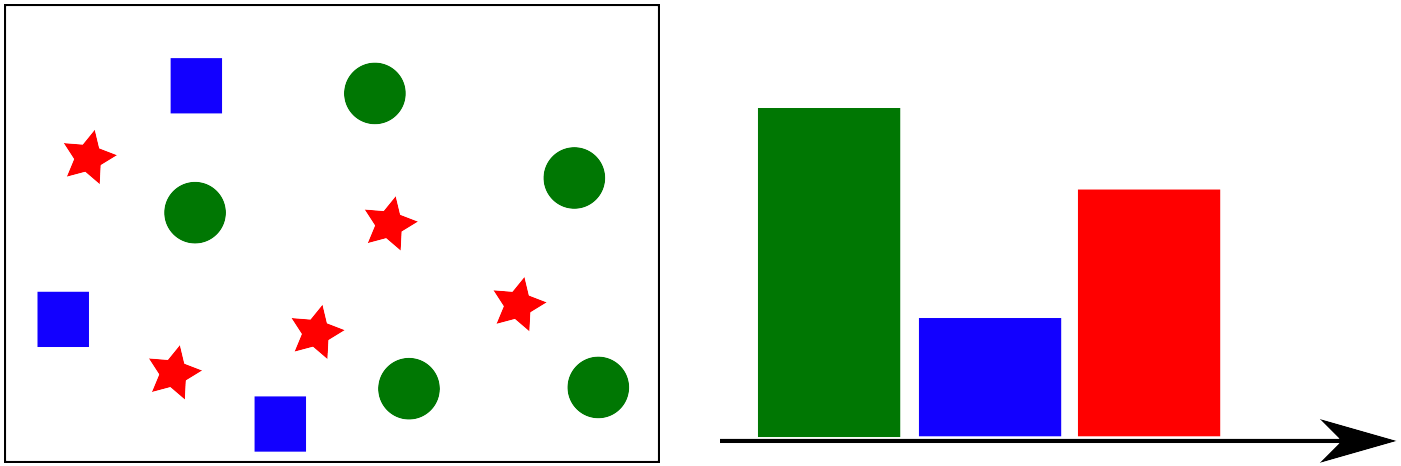
Typical Approaches to Image Retrieval

- Represent images as automatically extracted bag-of-visual-words (*visterms*)
  - SIFT, HoG, etc...



Typical Approaches to Image Retrieval

- Represent images as automatically extracted bag-of-visual-words (*visterms*)
  - SIFT, HoG, etc...



- Large heterogenous data sets
  - Corel 5K (5K images)
  - CIFAR-10 (60K images)
  - TinyImages (100K images)
  - ...

This Talk

- Represent images as annotated regions
  - Tighter connection to language than a *visterm*

↕ Girl ↕ Laptop ↕ Bed

This Talk

- Represent images as annotated regions
  - Tighter connection to language than a *visterm*

↕ Girl ↕ Laptop ↕ Bed

- Smaller data set: 341 images depicting **actions**
  - Explore the effect of action types on accuracy

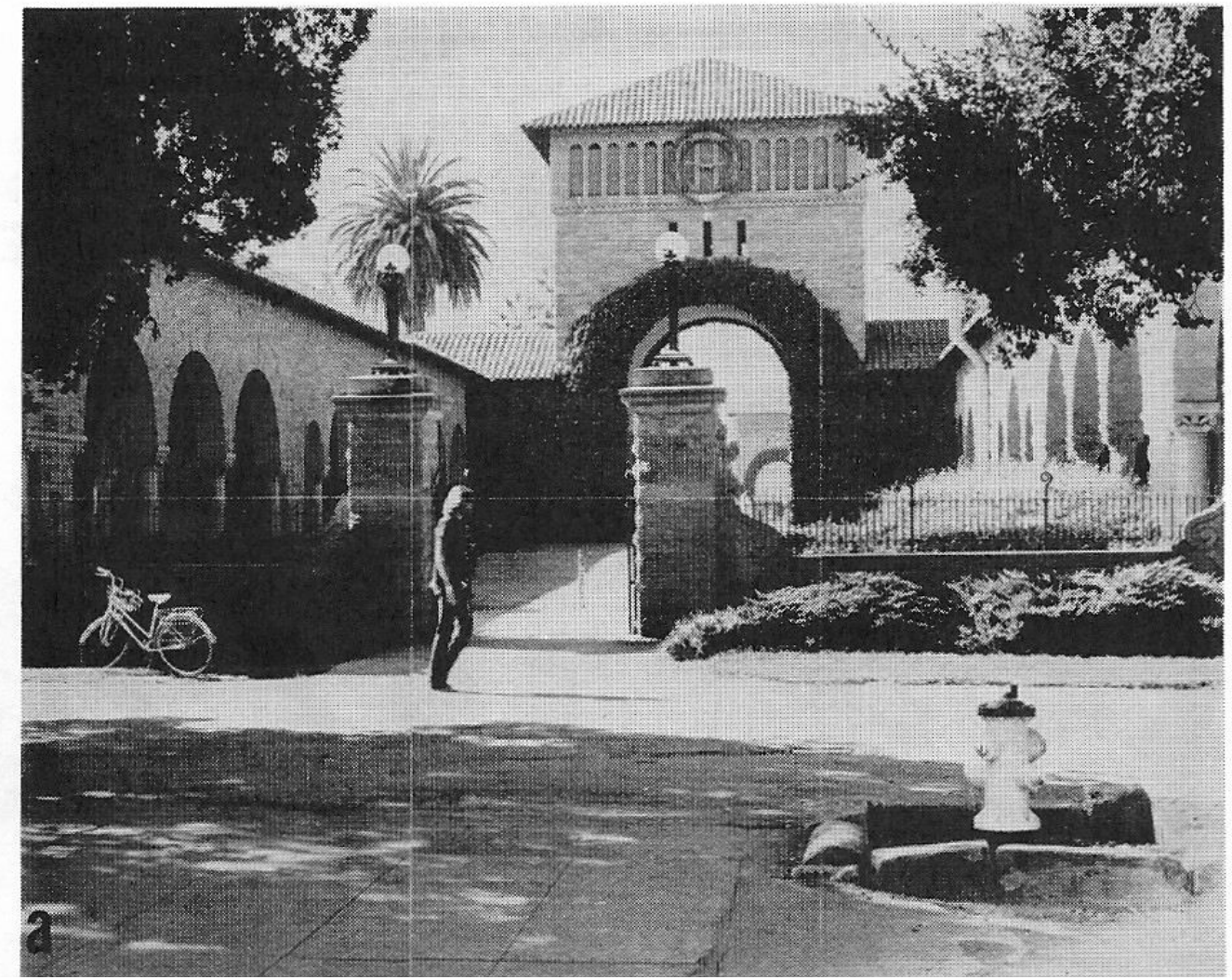
This Talk

- Represent images as annotated regions
  - Tighter connection to language than a *visterm*

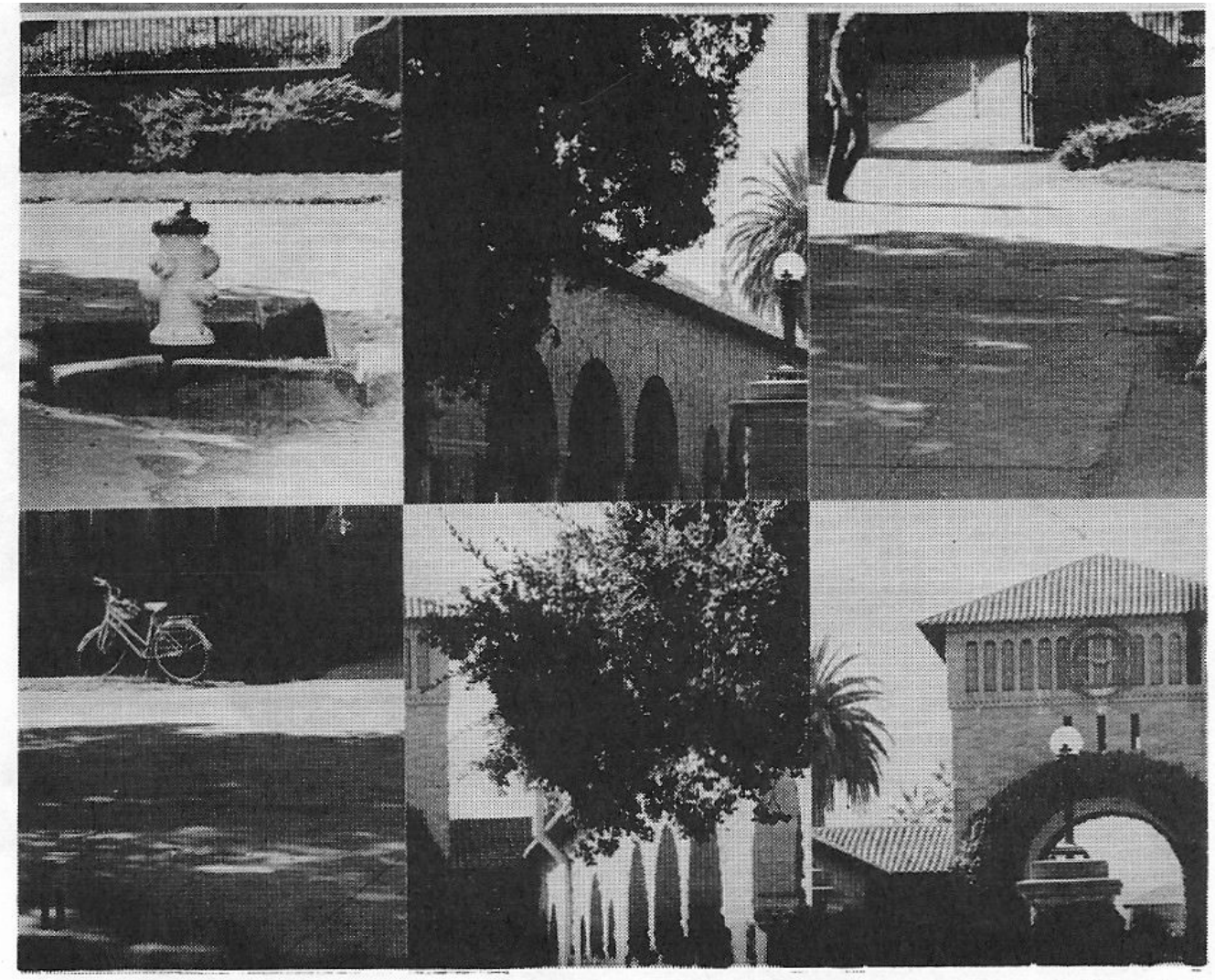
↕ Girl ↕ Laptop ↕ Bed

- Smaller data set: 341 images depicting **actions**
  - Explore the effect of action types on accuracy
- Focus on encoding the **spatial** relationships between regions

Humans benefit from consistent spatial relationships Biederman (1972)



Humans benefit from consistent spatial relationships Biederman (1972)

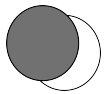


Visual Dependency Representation (Elliott and Keller, 2013)

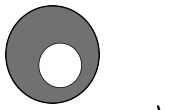
- Novel structured representation over image regions
  - Captures salient region-region relationships
  - Guided by the written description of the image
- Proved useful for describing actions in Elliott and Keller (2013)
- Inspired by dependency-syntax of language (Tesnière, 1953)
  - Tokens: image regions
  - Grammar: spatial relationships

Visual Dependency Representation (Elliott and Keller, 2013)

- Novel structured representation over image regions
  - Captures salient region-region relationships
  - Guided by the written description of the image
- Proved useful for describing actions in Elliott and Keller (2013)
- Inspired by dependency-syntax of language (Tesnière, 1953)
  - Tokens: image regions
  - Grammar: spatial relationships



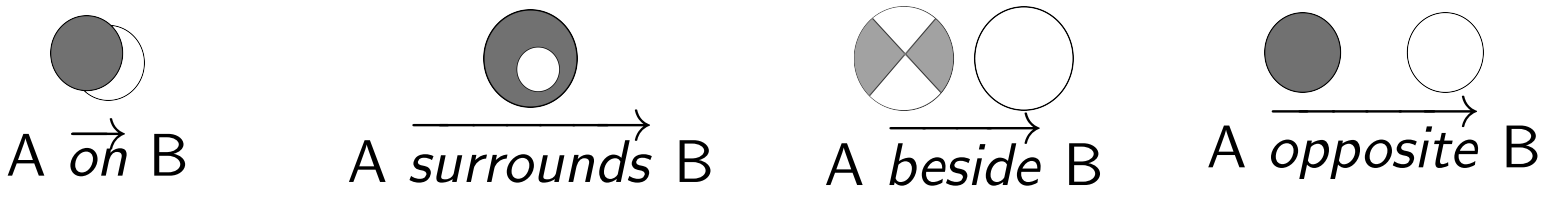
A $\overrightarrow{on}$ B



A $\overrightarrow{surrounds}$ B

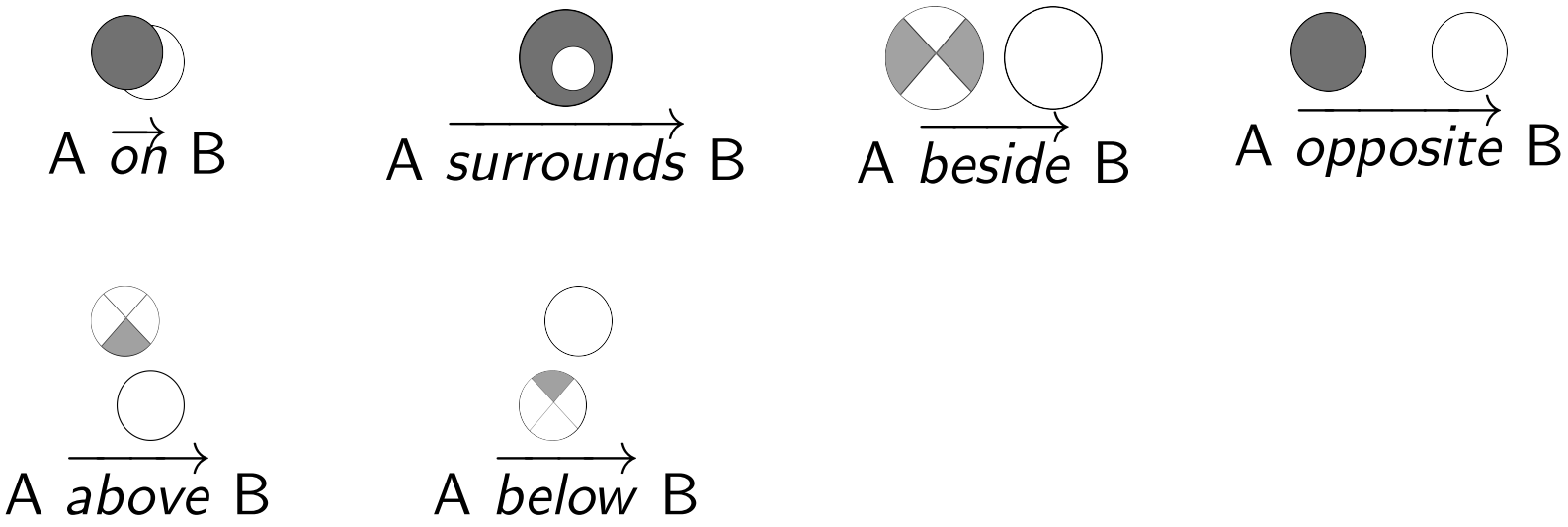
Visual Dependency Representation (Elliott and Keller, 2013)

- Novel structured representation over image regions
  - Captures salient region-region relationships
  - Guided by the written description of the image
- Proved useful for describing actions in Elliott and Keller (2013)
- Inspired by dependency-syntax of language (Tesnière, 1953)
  - Tokens: image regions
  - Grammar: spatial relationships



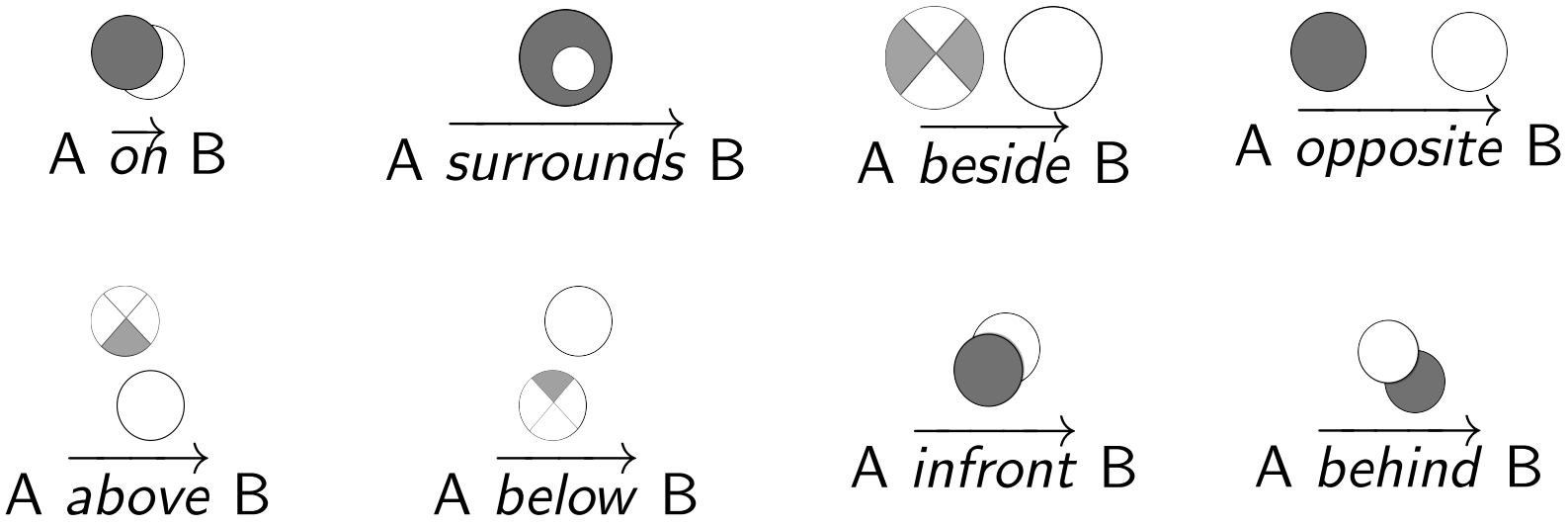
Visual Dependency Representation (Elliott and Keller, 2013)

- Novel structured representation over image regions
  - Captures salient region-region relationships
  - Guided by the written description of the image
- Proved useful for describing actions in Elliott and Keller (2013)
- Inspired by dependency-syntax of language (Tesnière, 1953)
  - Tokens: image regions
  - Grammar: spatial relationships



Visual Dependency Representation (Elliott and Keller, 2013)

- Novel structured representation over image regions
  - Captures salient region-region relationships
  - Guided by the written description of the image
- Proved useful for describing actions in Elliott and Keller (2013)
- Inspired by dependency-syntax of language (Tesnière, 1953)
  - Tokens: image regions
  - Grammar: spatial relationships



Gold Standard Example

"A girl is using a laptop. She is sitting on a bed."

Girl

Laptop

Bed

Gold Standard Example

"A girl is using a laptop. She is sitting on a bed."

ROOT Girl Laptop Bed

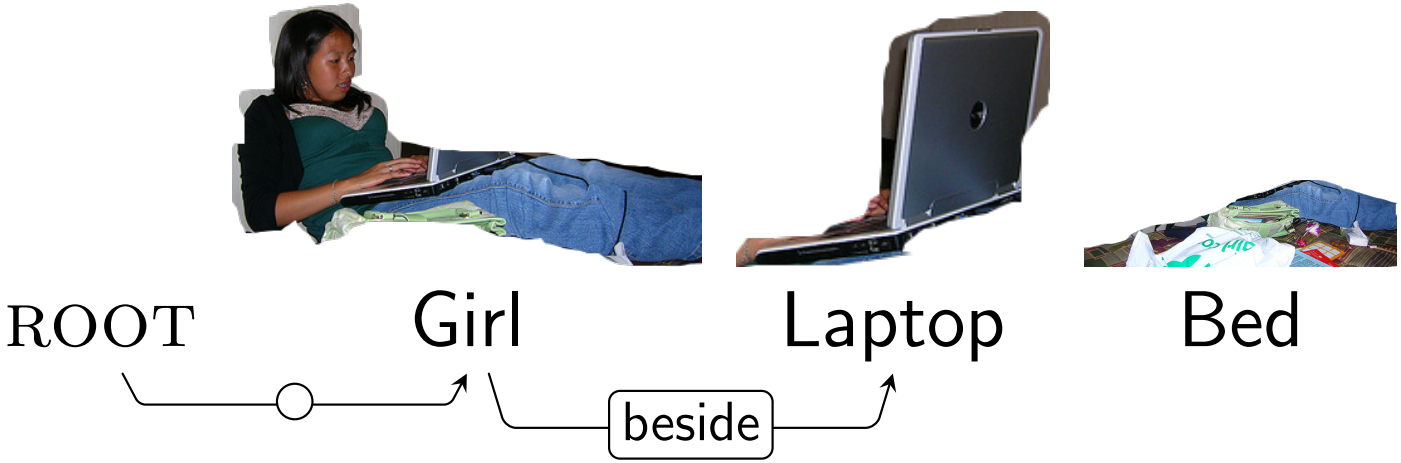
Gold Standard Example

"A girl is using a laptop. She is sitting on a bed."

ROOT Girl Laptop Bed

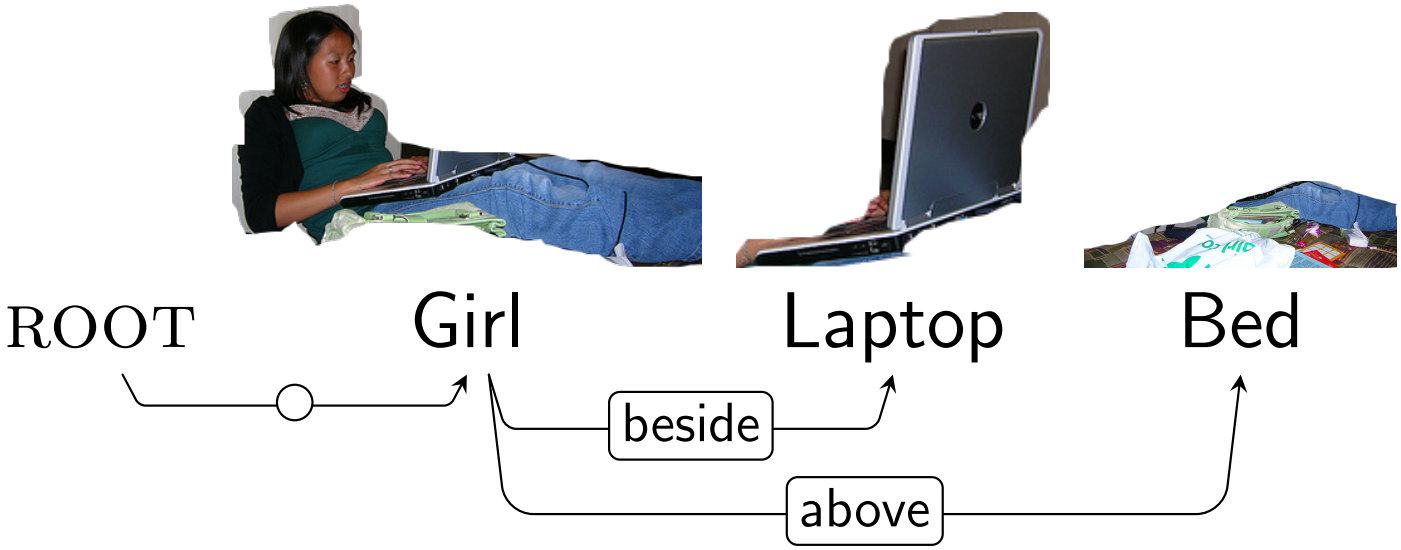
Gold Standard Example

"A girl is using a laptop. She is sitting on a bed."



Gold Standard Example

"A girl is using a laptop. She is sitting on a bed."



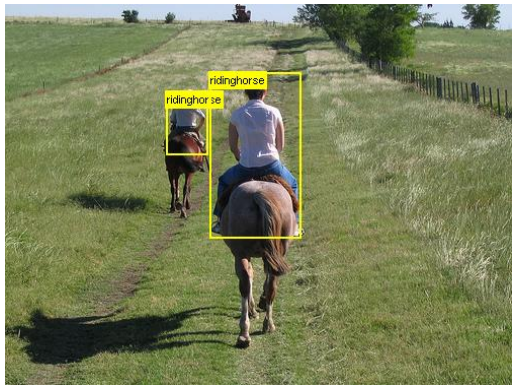
Data

341 images
from PASCAL VOC
Action Recognition
gold action labels

Data: 341 Images

- 10 types of actions

Reading



Ride horse

Phoning

Ride bike

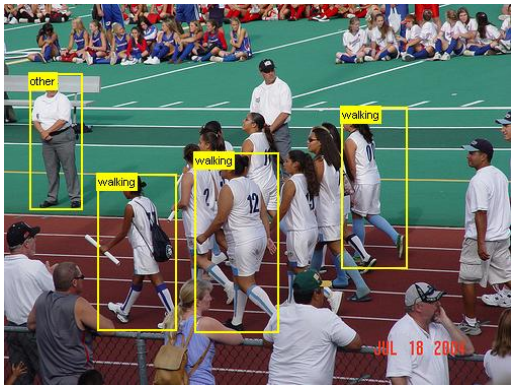
Play instrument

Data: 341 Images

- 10 types of actions

Jumping

Running



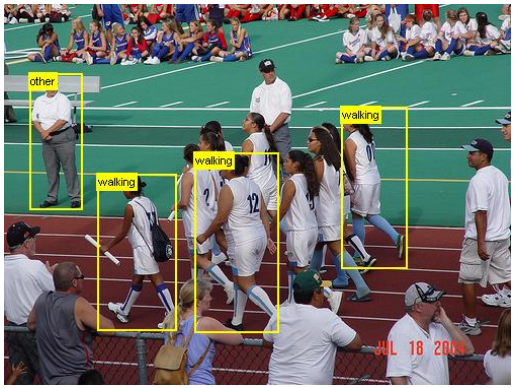
Walking

Data: 341 Images

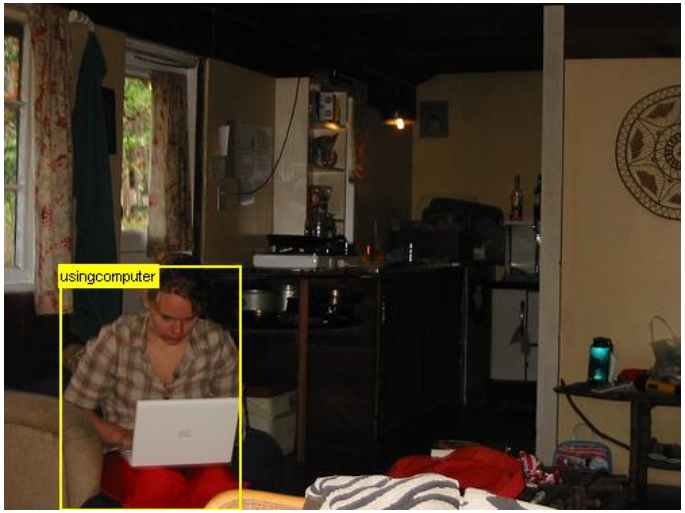
- 10 types of actions

Jumping

Running



Walking



Use computer

Take photo

Data

D1 D2 D3

341 images
from PASCAL VOC
Action Recognition
with action labels

3 descriptions/image

Data: 1,023 Descriptions from Mechanical Turk

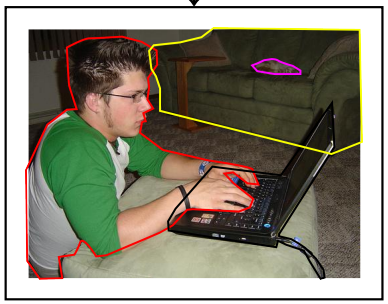
1. A teenage girl is using a laptop. She is sitting on a bed.
2. A girl is using a laptop. There is a lamp beside her.
3. A girl is using a computer. There is a picture behind her.

Data

341 images
from PASCAL VOC
Action Recognition
with action labels

D1 D2 D3

3 descriptions/image



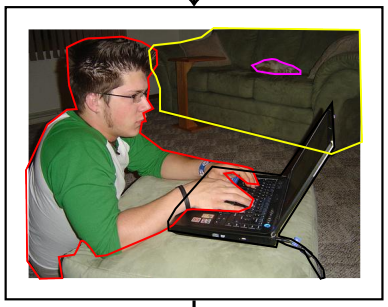
Objects for 341 images

Data

341 images
from PASCAL VOC
Action Recognition
with action labels

D1 D2 D3

3 descriptions/image



Objects for 341 images

VDR 1 VDR 2 VDR 3

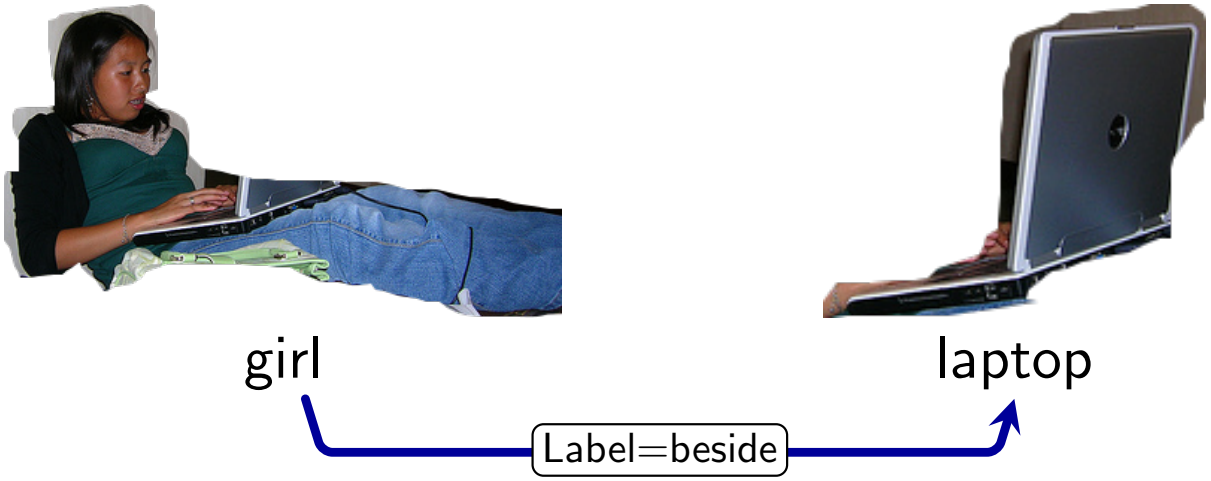
1,023 VDRs

Automatic VDR Prediction

- Framed as a dependency parsing task
  - MaltParser (Nivre et al., 2004) seems unsuitable because it is incremental
- Construct a complete graph between all regions using MSTParser (McDonald et al., 2005)
  - Remove all features that encode the linear order of the input
  - Extract features from the image regions

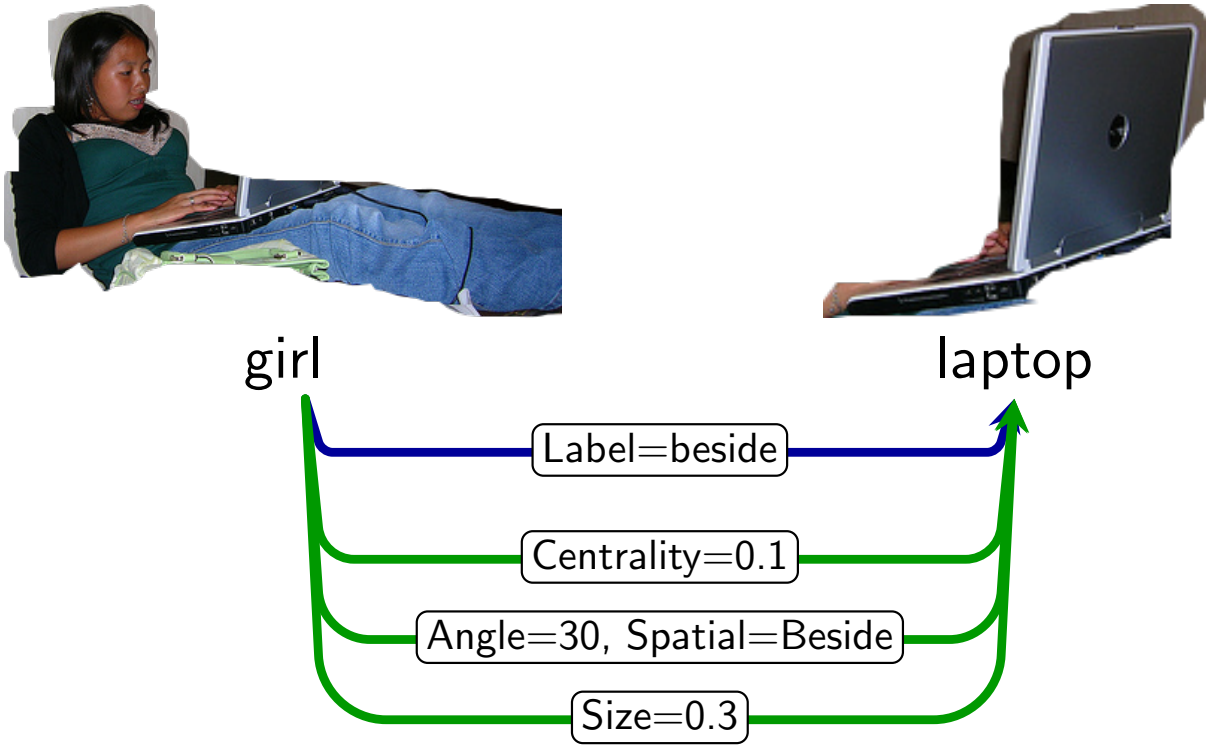
Automatic VDR Prediction

- Framed as a dependency parsing task
  - MaltParser (Nivre et al., 2004) seems unsuitable because it is incremental
- Construct a complete graph between all regions using MSTParser (McDonald et al., 2005)
  - Remove all features that encode the linear order of the input
  - Extract features from the image regions



Automatic VDR Prediction

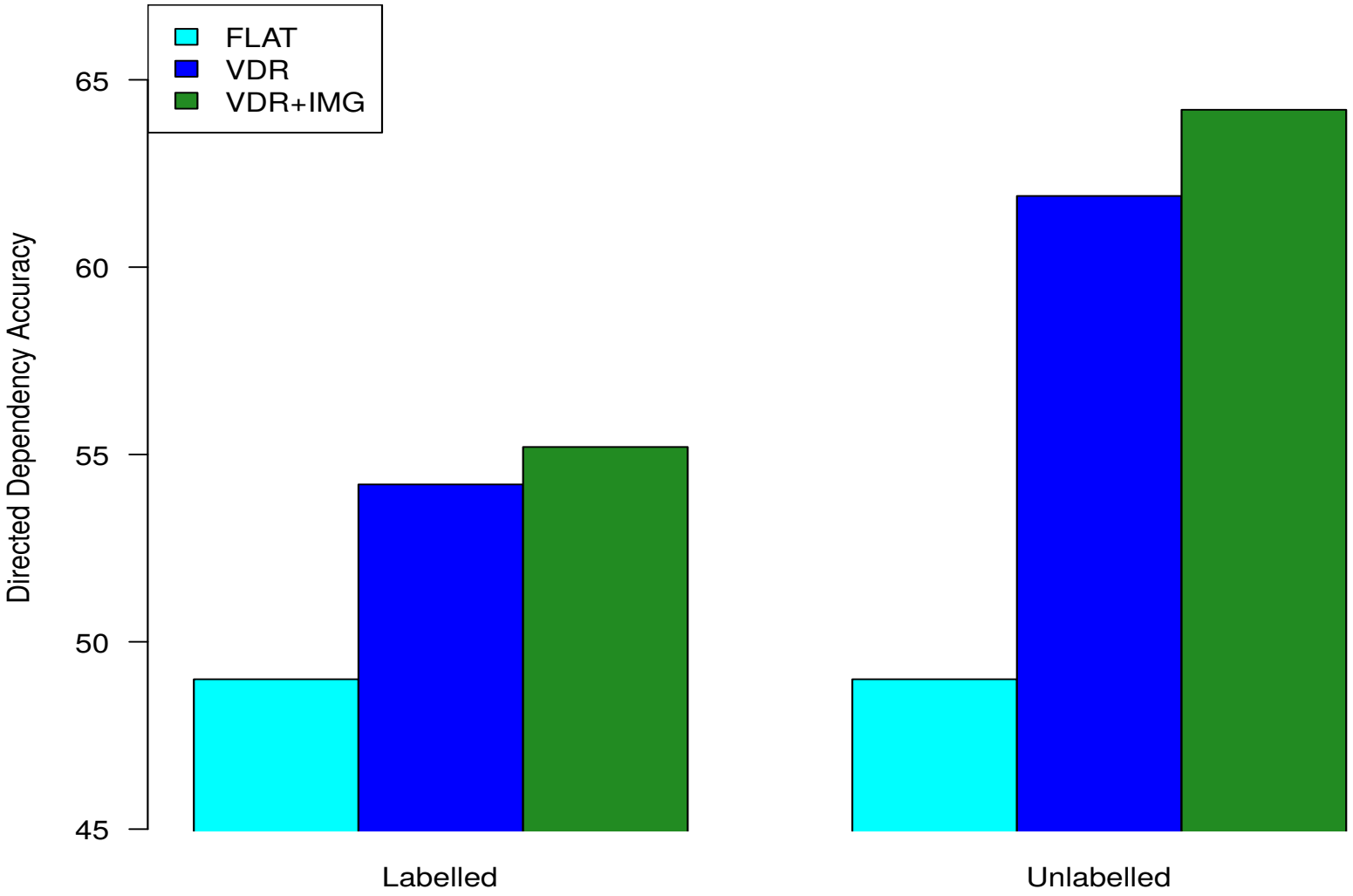
- Framed as a dependency parsing task
  - MaltParser (Nivre et al., 2004) seems unsuitable because it is incremental
- Construct a complete graph between all regions using MSTParser (McDonald et al., 2005)
  - Remove all features that encode the linear order of the input
  - Extract features from the image regions



VDR Parsing Experiment

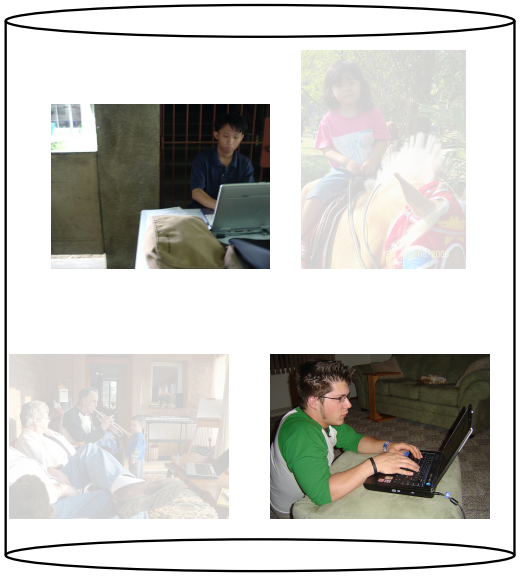
- Task
  - Predict VDR over region-annotated image
- Data
  - 1,023 VDR data set
  - 10 fold cross-validation
- Evaluation
  - Unlabelled/labelled directed attachment accuracy
- Models
  - FLAT is a bag-of-regions baseline
  - VDR uses only input features
  - VDR+IMG also uses visual features

VDR Parsing Results



Query-by-Example Image Retrieval

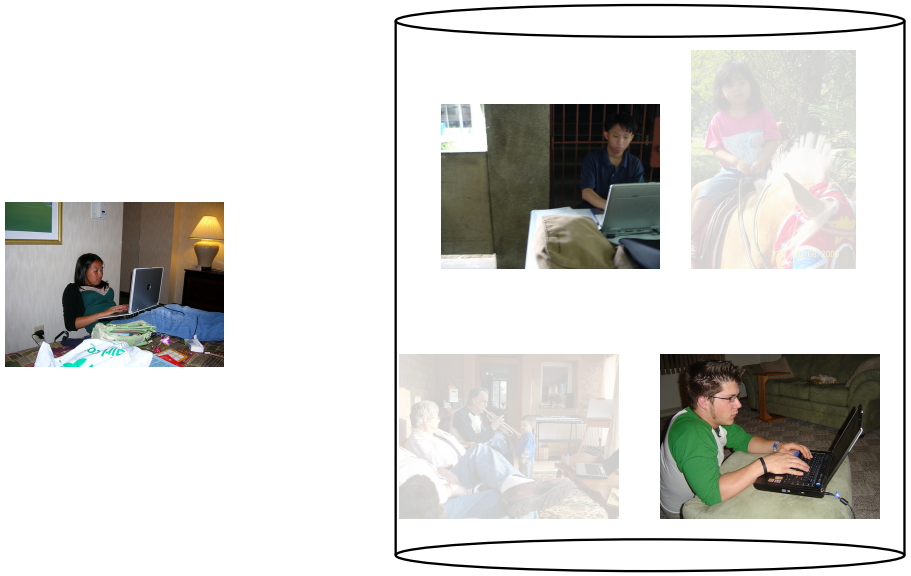
- Given a query example, find images of the same action



- Matching function: cosine with *tf-idf* weighting

Query-by-Example Image Retrieval

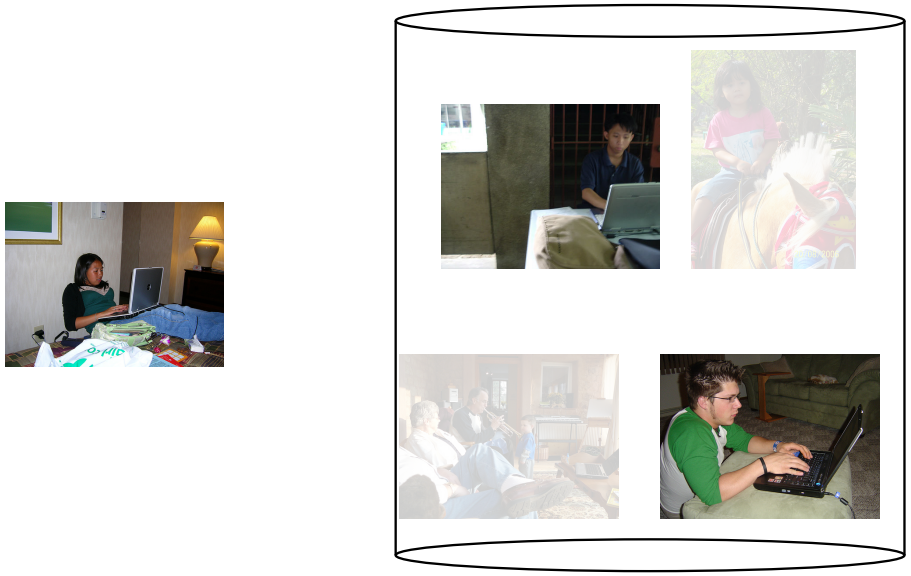
- Given a query example, find images of the same action



  - Matching function: cosine with *tf-idf* weighting
- Evaluate with Mean Average Precision and Precison@10
  - Relevance means same action annotation

Query-by-Example Image Retrieval

- Given a query example, find images of the same action



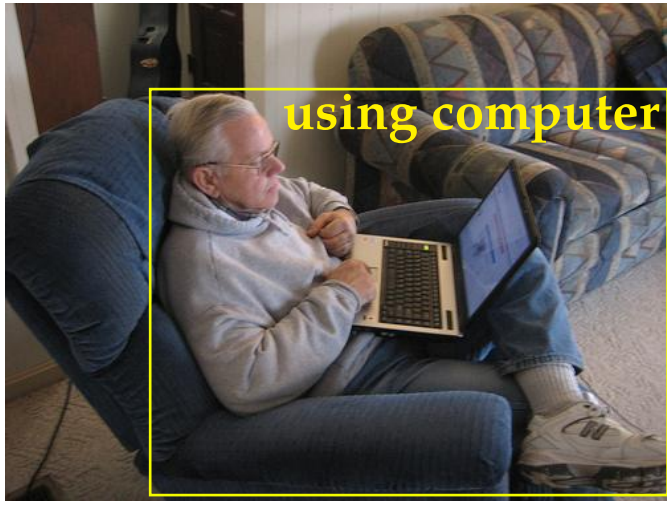
  - Matching function: cosine with *tf-idf* weighting
- Evaluate with Mean Average Precision and Precison@10
  - Relevance means same action annotation
- Models:
  - Bag-of-Regions representation
  - Visual Dependency Representation
  - Both use gold-standard object annotations

Bag-of-Regions Representation

$$cos(a, b) = \frac{a \cdot b}{||a||\ ||b||}$$

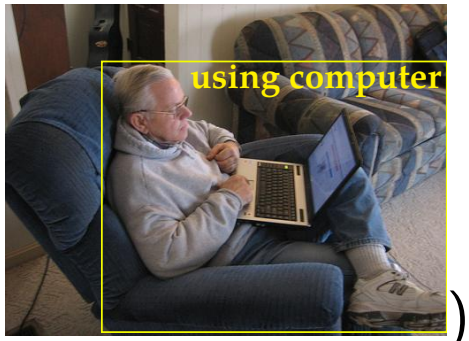
Bag-of-Regions Representation

$\cos($ 

, 

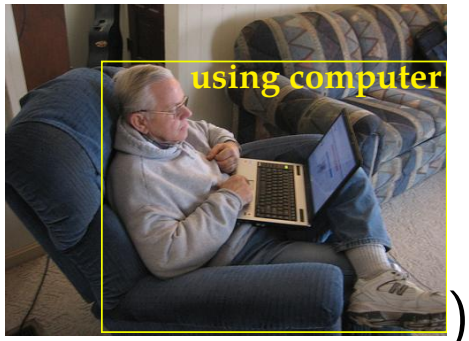
$)$

Bag-of-Regions Representation

$cos($  ,  $) =$

$$\frac{<person, laptop> \cdot <person, laptop>}{}$$

Bag-of-Regions Representation

$cos($  ,  $) =$

$$\frac{<person, laptop> \cdot <person, laptop>}{\left| person, laptop, \cdots \right| \left| person, laptop, \cdots \right|}$$

Bag-of-Regions Representation

$cos($ 

, 

$)$

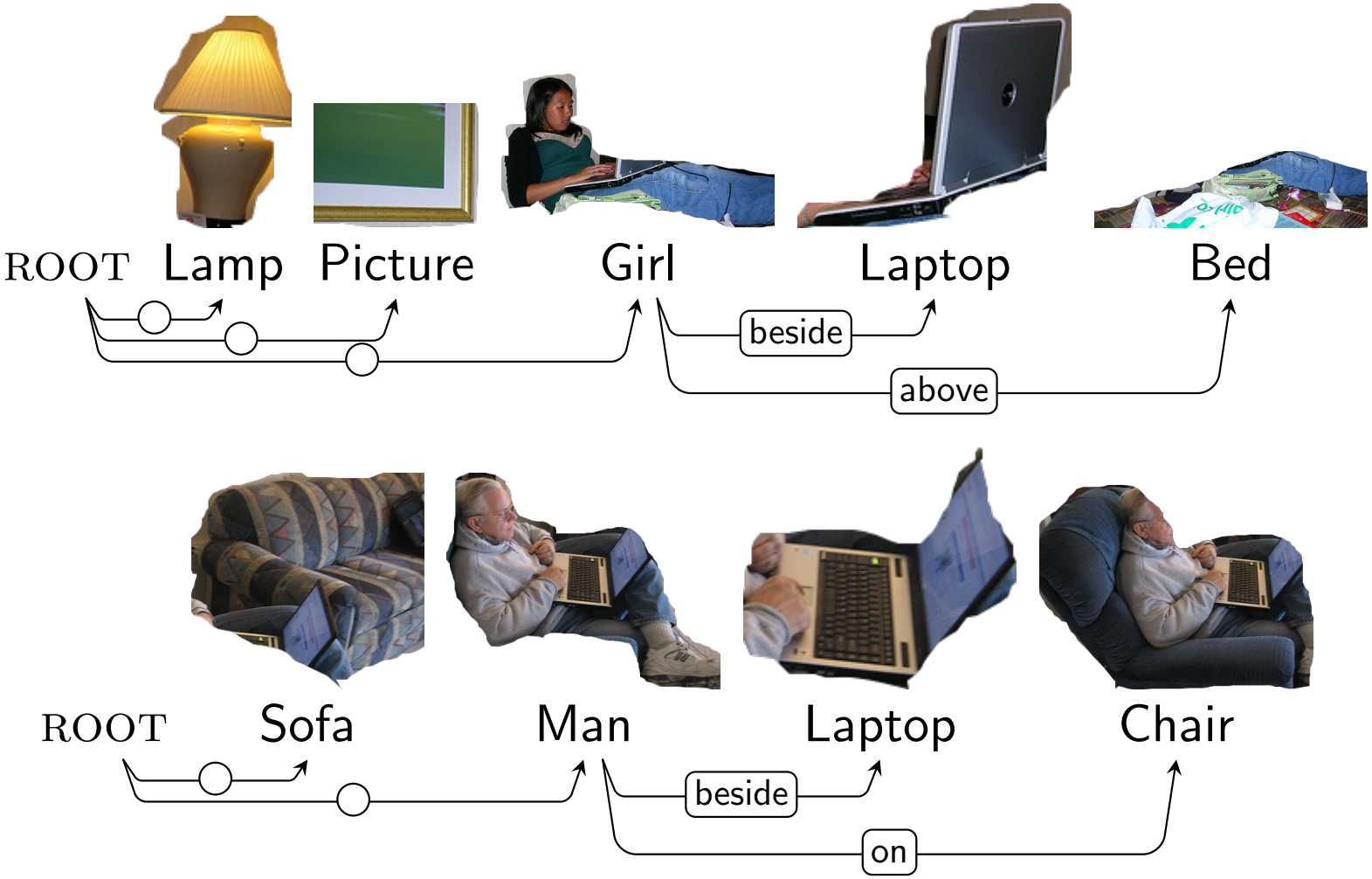
Bag-of-Regions Representation

$cos($  ,  $) =$

$$\frac{<person, laptop> \cdot <person, laptop>}{\left| person, laptop, \cdots \right| \left| person, laptop, \cdots \right|}$$

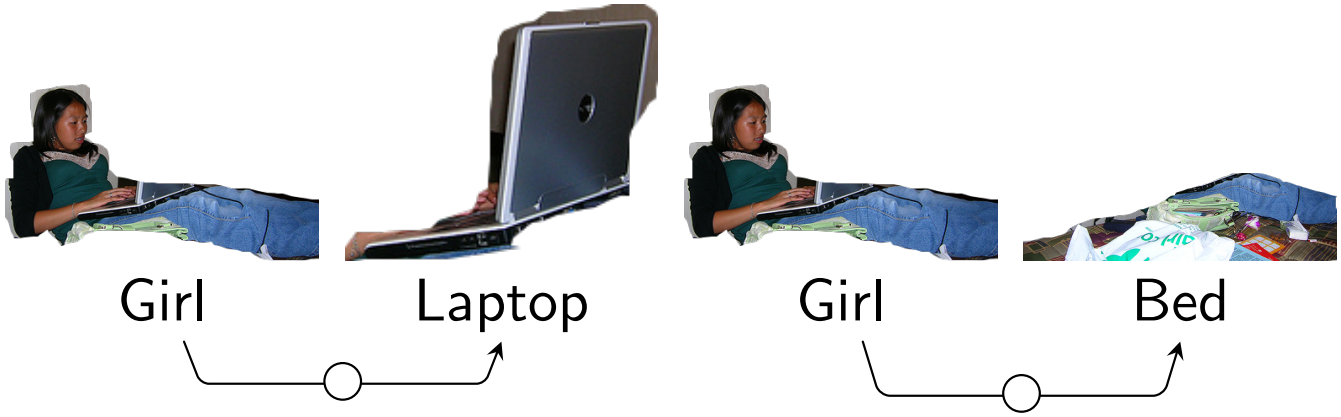
Visual Dependency Representation

- How to compare two trees?



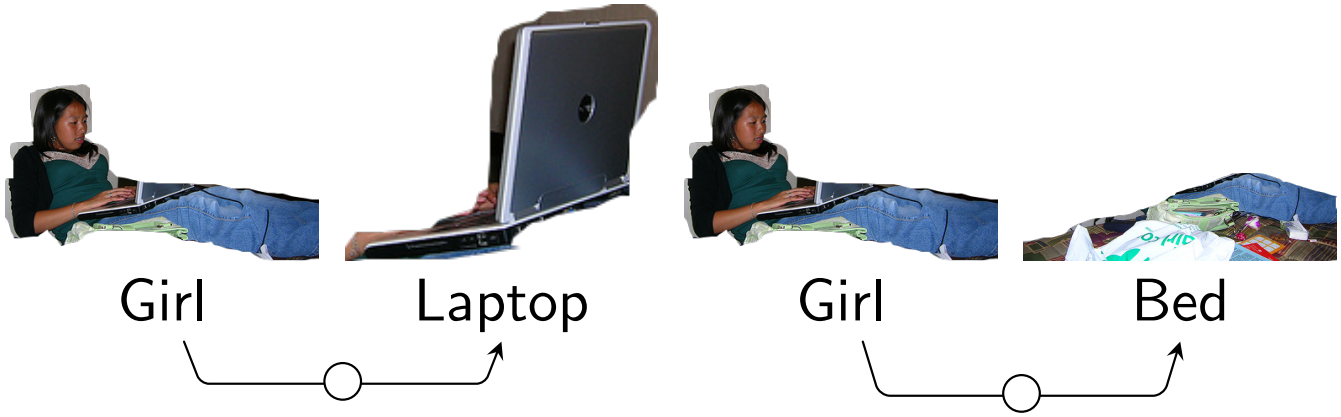
Visual Dependency Representation

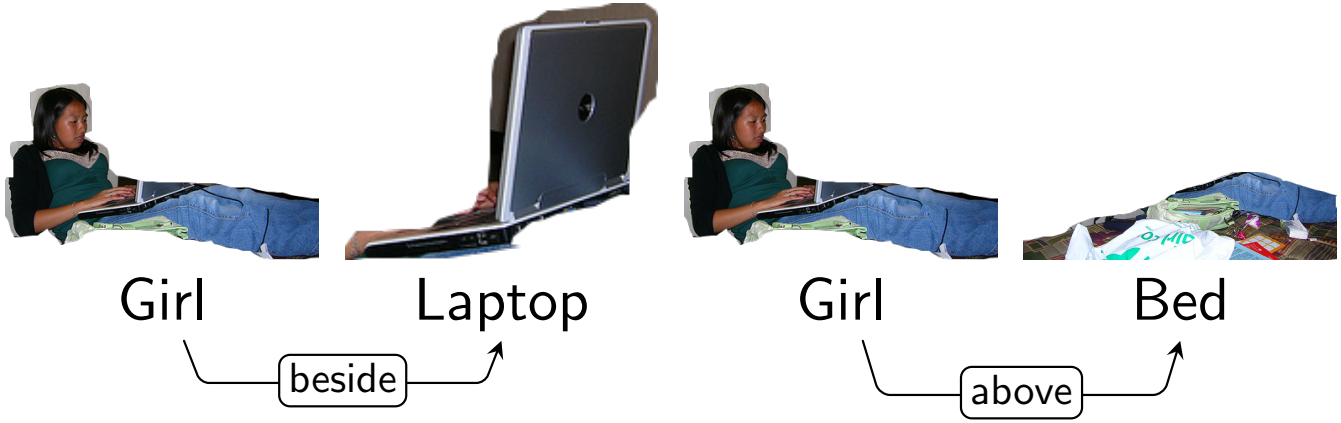
- How to compare two trees?
  - Decompose all edges into bigrams and trigrams



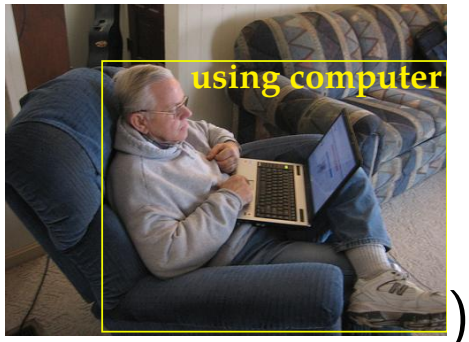
Visual Dependency Representation

- How to compare two trees?
  - Decompose all edges into bigrams and trigrams

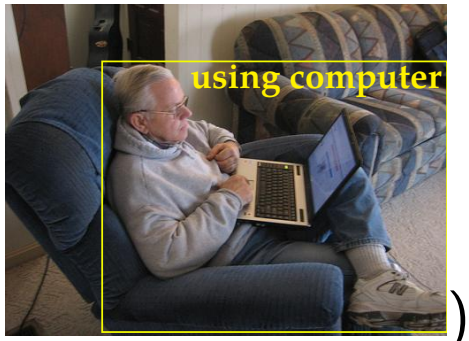




Visual Dependency Representation

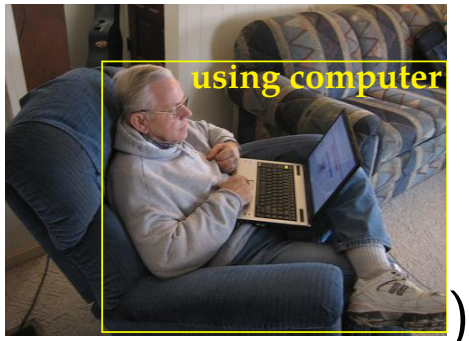
$cos($  ,  $) =$

Visual Dependency Representation

$cos($  $,$  $) =$

$$\frac{\langle \text{Girl Laptop} \rangle \cdot \langle \text{Man Laptop} \rangle}{\text{beside} \qquad \text{beside}}$$

Visual Dependency Representation

$cos($  ,  $) =$

$$\frac{\langle \text{Girl Laptop} \rangle \xrightarrow{\text{beside}} \cdot \langle \text{Man Laptop} \rangle \xrightarrow{\text{beside}}}{\left| \text{Girl Laptop} \xrightarrow{\text{beside}} , \text{Girl Bed} \xrightarrow{\text{above}} \cdots \right| \quad \left| \text{Man Laptop} \xrightarrow{\text{beside}} \cdots \right|}$$

Visual Dependency Representation

$cos($  ,  $) =$

Visual Dependency Representation

$cos($  ,  $) =$

$$\frac{<>}{}$$

Visual Dependency Representation

$cos($  ,  $) =$

$$\frac{\langle\rangle}{\left|\begin{array}{ccccc} \text{Girl} & \text{Laptop} & , & \text{Girl} & \text{Bed} \;\cdots \\ & \text{beside} & & \text{above} & \end{array}\right| \quad \left|\begin{array}{cc} \text{Man} & \text{Trumpet} \;\cdots \\ & \text{beside} \end{array}\right|}$$

Results

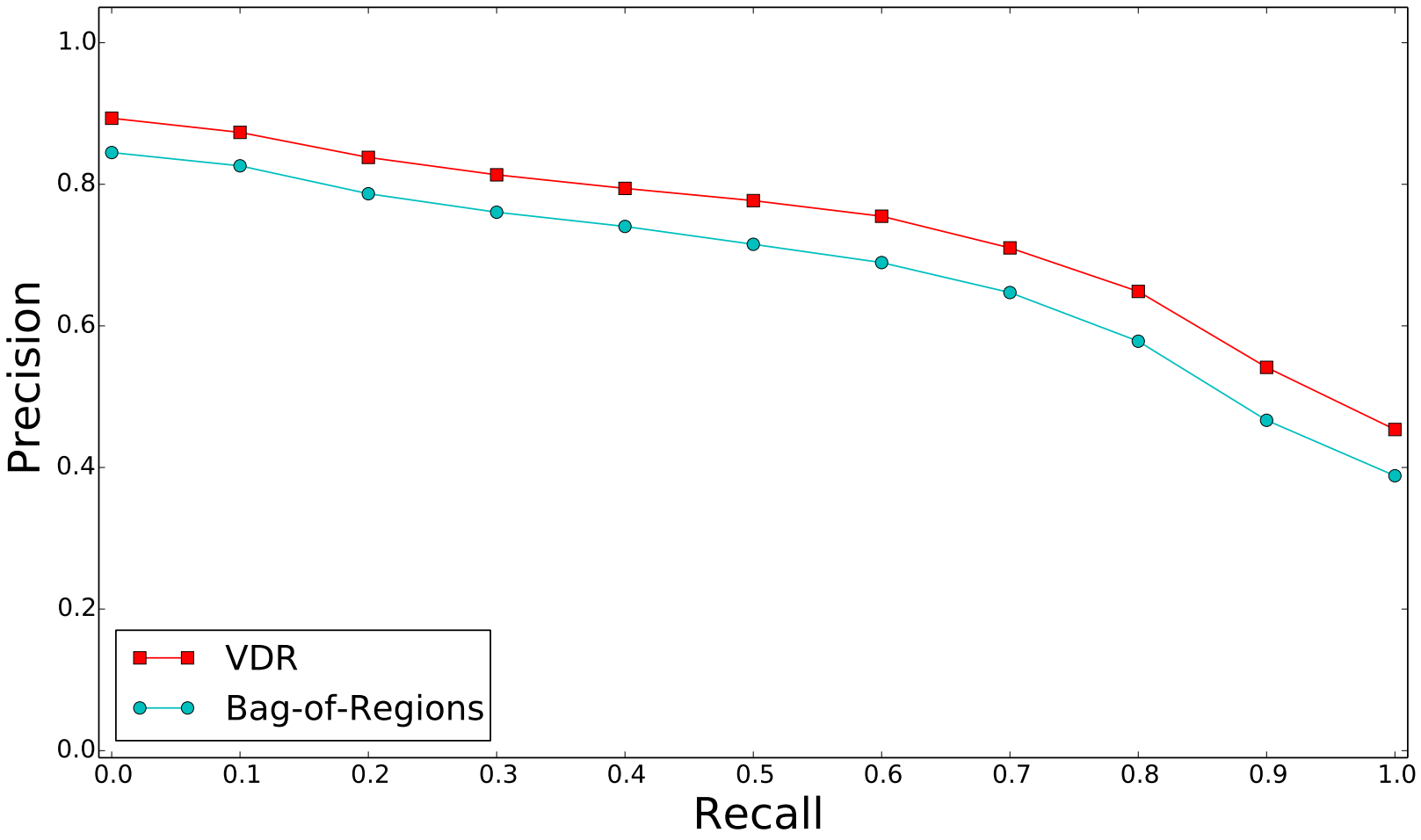
	MAP	P@10
Bag-of-Regions	0.467	0.415

Results

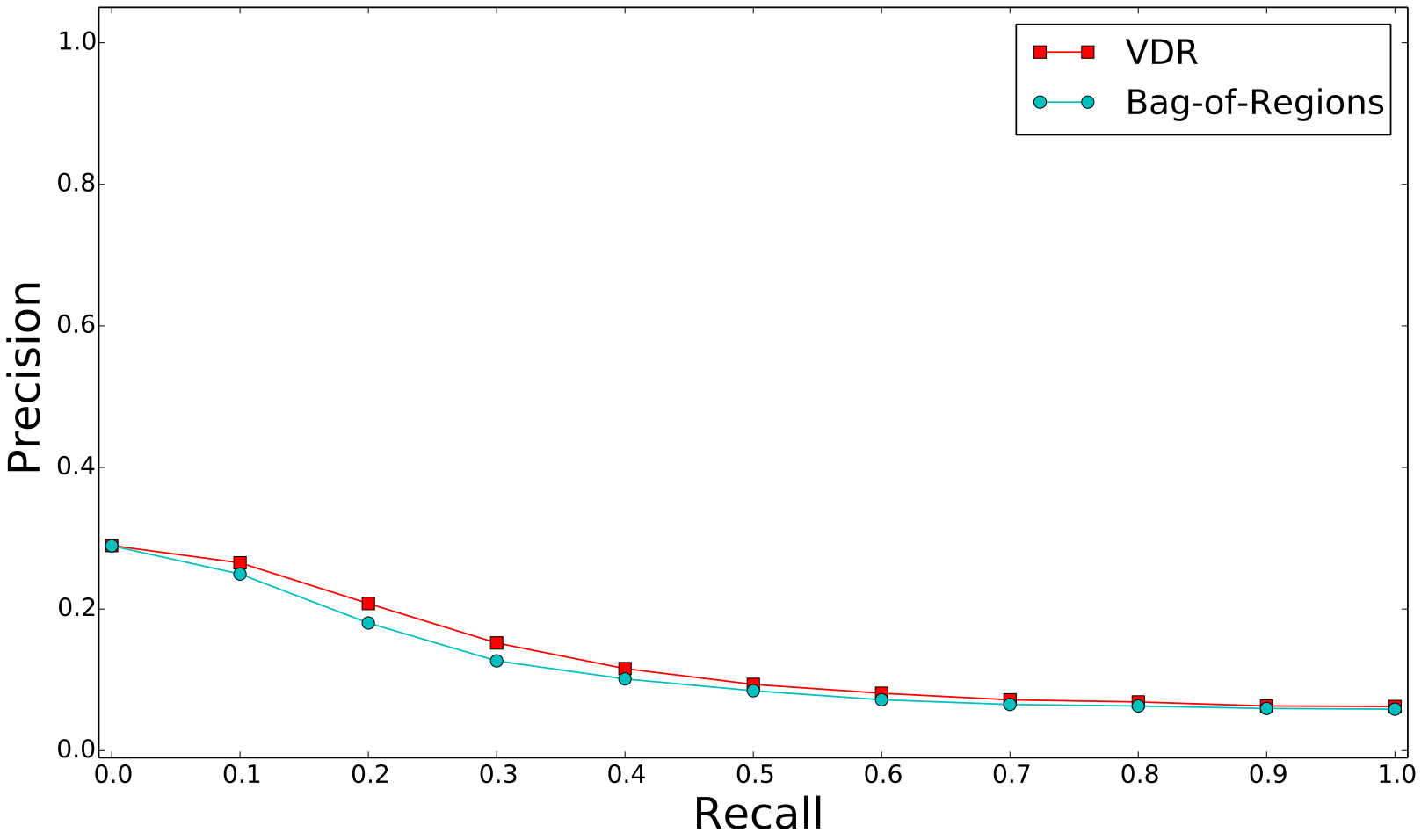
	MAP	P@10
Bag-of-Regions	0.467	0.415
VDR	$0.508^{\star}$	$0.451^{\star}$

$\star$: significantly better than Bag-of-Regions at $p < 0.01$

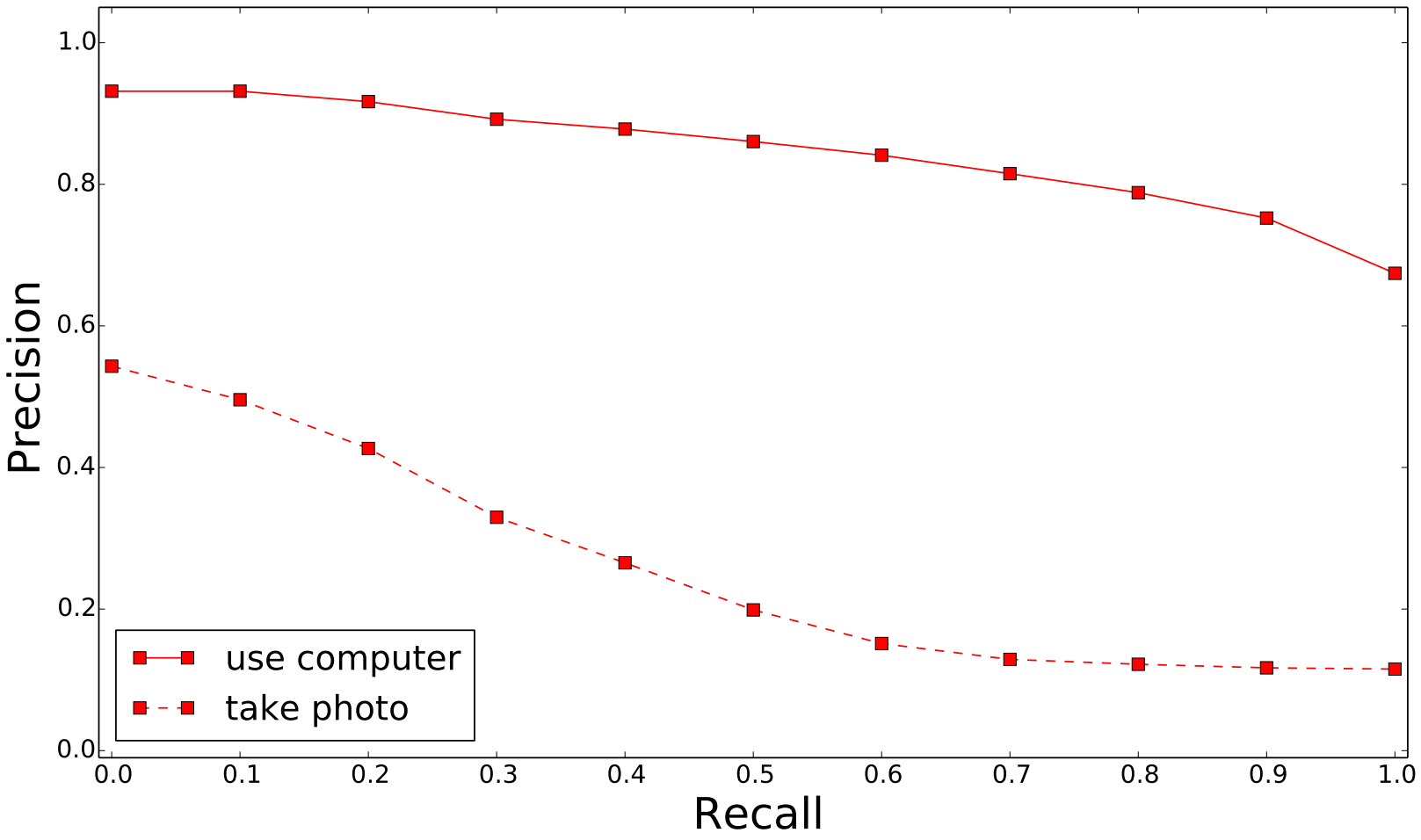
Transitive actions



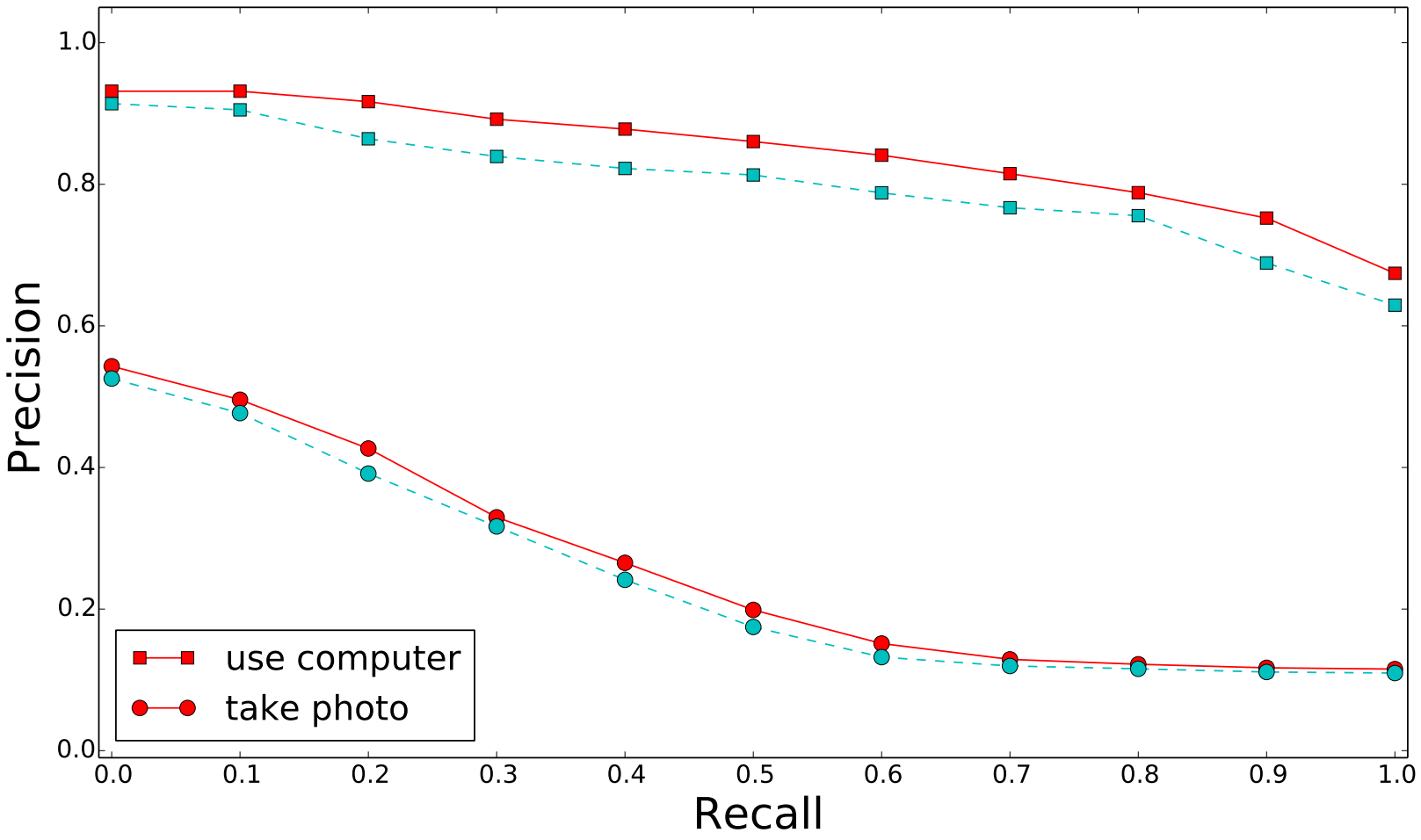
Intransitive actions



“Light” actions - use computer / take photo



“Light” actions - use computer / take photo



Conclusions

- VDR increases the accuracy of query-by-example image retrieval compared to a bag-of-regions baseline
- Improvement depends on the type of action:
  - Most pronounced for transitive verbs
  - Least pronounced when no object is required for the action
- Future work:
  - Scaling to larger data sets
  - Different matching paradigms, e.g. RankSVM
  - Explore the effect of other languages on actions

Questions?

- VDRParser: `http://github.com/elliottd/vdrparser`
- Data: `http://homepages.inf.ed.ac.uk/s0128959/dataset/`

- `http://homepages.inf.ed.ac.uk/s0128959/`
- d.elliott@ed.ac.uk // @delliott

References

Biederman, I. (1972). Perceiving real-world scenes. *Science*, 177(4043):77–80.

Elliott, D. and Keller, F. (2013). Image Description using Visual Dependency Representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Seattle, Washington, U.S.A.

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada.

Nivre, J., Hall, J., and Nilsson, J. (2004). Memory-based dependency parsing. In *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning*, pages 49–56, Boston, Massachusetts, USA.

Tesnière, L. (1953). *Esquisse d'une syntaxe structurale*. Librairie C. Klincksieck.