BİL722

A Sentence is Worth a Thousand Pixels

Sanja Fidler, et al.

Pınar Küllü

N12247681

Goal

- to reason jointly about the scene type, objects, their location and spatial extent in an image, while exploiting textual information in the form of complex sentential image descriptions generated by humans.

- Being able to extract semantic information from text does not entirely solve the image parsing problem.
- Solution: a holistic model for semantic parsing which employs text and image information.

- A CRF model which employs complex sentential image descriptions to jointly reason about multiple scene recognition tasks.
- parse the sentences and extract objects and their relationships, and incorporate those into the model, both via potentials as well as by re-ranking the candidate bounding boxes.

Automatic Text Extraction

- extract part of speech tags (POS) of all sentences
- parse syntactically the sentences and obtain a parse tree
- Given the POS, parse trees and type dependencies, extract
  - whether an object class was mentioned,
  - its cardinality,
  - the relationships between the different objects, e.g., object A is near or on top of object B.

Automatic Text Extraction Presence of a Class

- To detect the presence/absence of a class
  - Extract nouns from the POS
  - Match nouns to the object classes
    - Not only to the name of the class, also to its synonyms and plural forms

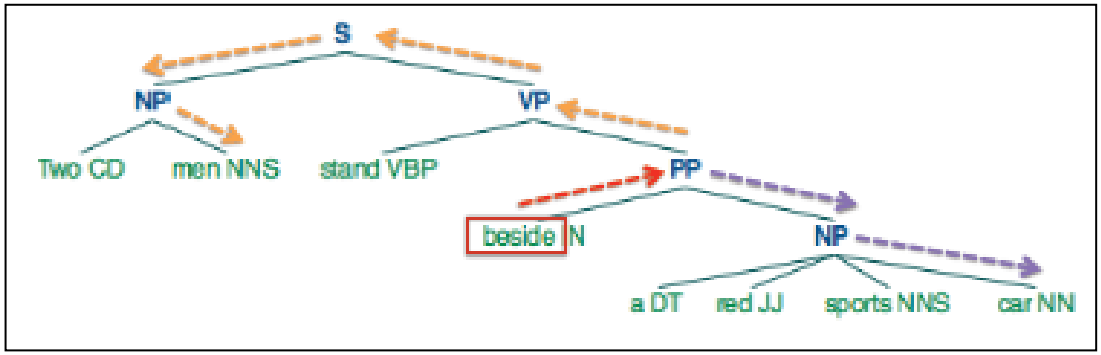
Automatic Text Extraction
Object Cardinality

- Can appear in the sentence in two different forms:
  - Explicitly mentioned.
    - "**two** children playing on the grass."
  - Implicit.
    - Parse entire sentence from left to right
    - Increase count by 1 or 2 with each mention of a class.
    - Lower bound on the cardinality can be extracted.

Automatic Text Extraction
Object Relations

- Extract prepositions and the objects they modify.
- For each preposition, use parse tree to locate the objects modified by the preposition.
  - ($object_1$; prep; $object_2$)
- To compute $object_2$
  - search for NPs on the right side of the preposition by traversing the tree.
  - return the nouns in the NP which are synonyms of object classes.
- To compute $object_1$
  - move up the tree until hitting S or NP
  - return the nouns.

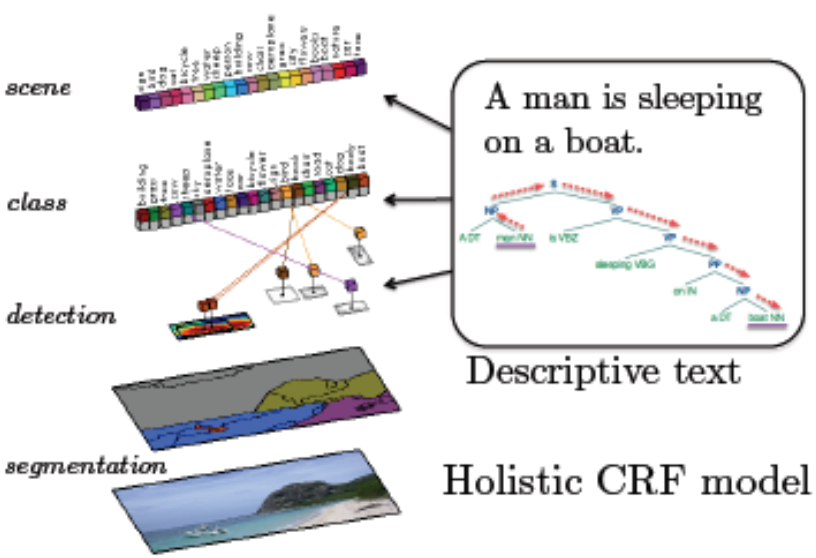
Automatic Text Extraction Object Relations



- To compute $object_2$
  - search for NPs on the right side of the preposition by traversing the tree.
  - return the nouns in the NP which are synonyms of object classes.
- To compute $object_1$
  - move up the tree until hitting S or NP
  - return the nouns.

Holistic Scene Understanding

- CRF model
  - variables representing the class labels of image segments at two levels in a segmentation hierarchy
  - binary variables indicating the correctness of candidate object detections
  - binary variables encode the presence/absence of a class in the scene.



Holistic Scene Understanding

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{b}, \mathbf{s}) = \frac{1}{Z} \prod_{type} \prod_{\alpha} \psi_{\alpha}^{type}(\mathbf{a}_{\alpha})$$

- $x_i \in \{1,\ldots,C\}$
  - the class label of the i-th segment in the lower level of the hierarchy
- $y_j \in \{1,\ldots,C\}$
  - the class label of the j-th segment of the second level of the hierarchy
- $b_l \in \{0,1\}$
  - a candidate detection, taking value 0 when the detection is a false detection
- $z_k \in \{0,1\}$
  - variable which takes value 1 if class k is present in the image
- ψ_{α}^{type} encodes potential functions over set of variables.

Segmentation Potentials

- **Unary Segmentation potential**: the unary potential for each region at segment and super-segment level by averaging the pixel potentials inside each region:

$$\phi_s(x_i|I) \quad \phi_s(y_j|I)$$

- **Segment-SuperSegment compatibility**: To encode compatibility between the two levels of the hierarchy.

$$\phi_{i,j}(x_i, y_j) = \begin{cases} -\gamma & \text{if } x_i \neq y_j \\ 0 & \text{otherwise.} \end{cases}$$

Class Presence Potentials

- **Class Presence form Text**: two types of unary potentials, depending on whether a class was mentioned or not in the text.

$$\phi^{class}_{ment}(z_i|T) = \begin{cases} \bar{C}ard(i) & \text{if } z_i = 1 \text{ and class } i \text{ mentioned} \\ 0 & \text{otherwise.} \end{cases}$$

$$\phi^{class}_{notment}(z_i|T) = \begin{cases} 1 & \text{if } z_i = 0 \text{ and class } i \textbf{ not} \text{ mentioned} \\ 0 & \text{otherwise.} \end{cases}$$

- **Class Presense Statistics from Images**: the unary potential, depending on the presence or absence of class z_i in image.

Class Presence Potentials

- **Class-Segment compatibility**: This potential ensures the compatibility of classes that are inferred to be present in the scene and the classes that are chosen at the segment level.

$$\phi_{j,k}(y_j, z_k) = \begin{cases} -\eta & \text{if } y_j = k \wedge z_k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Object Detection Potentials

- **Object Candidate Score**:
  - reduce thresholds
  - upper bound the number of candidate objects to be 3 per class.
  - use the boxes that pass the DPM thresholds, unless the object class is specifically mentioned in text.
  - for each box, compute a feature vector composed of
    - the original detector's score,
    - the average cardinality for that class extracted from text,
    - size relative to the image size.
  - train a SVM classifier with these features.

Object Detection Potentials

$$\phi_{cls}^{BBox}(b_l|I,T) = \begin{cases} \sigma(r_l) & \text{if } b_l = 1 \wedge c_l = cls \\ 0 & \text{otherwise.} \end{cases}$$

- r_l=object scores obtained from classifier.
- c_l=detector's class
- $\sigma(x)=1/(1+\exp(-1.5x))$=logistic function

Object Detection Potentials

- **Cardinality Potential**: A potential on the b_l variables to exploit the cardinality estimated from text.

$$\phi^{BBox}_{card-1}(\mathbf{b}^i|T) = \begin{cases} -\zeta_1 & \text{if } \bar{C}ard(i) \geq 1 \text{ and } \sum_j b^i_j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\phi^{BBox}_{card-2}(\mathbf{b}^i|T) = \begin{cases} -\zeta_2 & \text{if } \bar{C}ard(i) \geq 2 \text{ and } \sum_j b^i_j = 1 \\ 0 & \text{otherwise.} \end{cases}$$

  - b^i = all detections of class i.

Object Detection Potentials

- **Using Prepositions**:
  - extract prepositions from text and use them to score pairs of boxes.
    - rel = (cls1 ; prep; cls2).
  - train a SVM classifier for each preposition that uses features defined over pairs of bounding boxes of the referred object classes.
    - distance and signed distance between the closest left/right or top/bottom sides,
    - amount of overlap between the boxes,
    - scores of the two boxes.
  - compute the new score for each box using the preposition classifier on the pairwise features

$$\hat{r}_i = \max_{j,prep} score(r_{i,j,prep})$$

Object Detection Potentials

- **Using Prepositions**:

$$\phi_{prep}(b_\ell|T) = \begin{cases} \hat{r}_i & \text{if } b_l = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Object Detection Potentials

- **Class-detection compatibility**: allows the bounding box to be active only when the class label of that object is also declared as present in the scene.

$$\phi_{l,k}^{BClass}(b_l, z_k) = \begin{cases} -\alpha & \text{if } z_k = 0 \wedge c_l = k \wedge b_l = 1 \\ 0 & \text{otherwise.} \end{cases}$$

- **Shape prior**: The mixture components of a part-based model typically reflect the pose and shape of the object.

$$\phi_{cls}^{sh}(x_i, b_l | I)$$

Scene Potentials

- **Text Scene Potential**:
  - extract a vocabulary of words appearing in the full text corpus.
  - train an SVM classifier over the bag-of-words (textual, not visual).

$$\phi^{Scene}(s = u|T) = \sigma(t_u)$$

  - t_u = the classifier score for scene class u
  - σ = logistic function

Scene Potentials

- **Scene-class compatibility**:

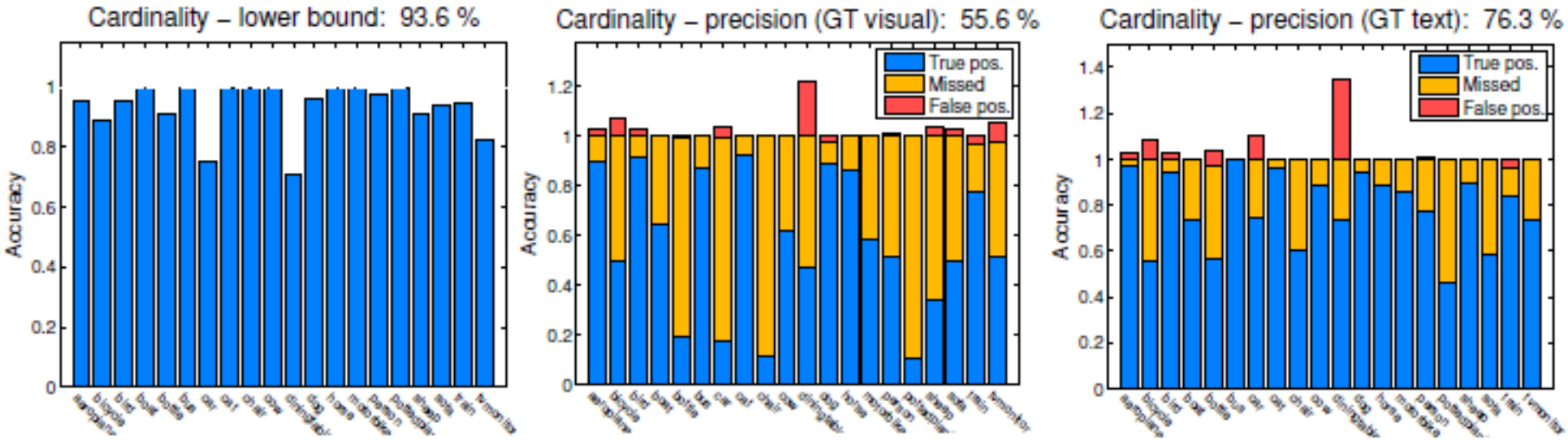
$$\phi^{SC}(s, z_k) = \begin{cases} f_{s,z_k} & \text{if } z_k = 1 \wedge f_{s,z_k} > 0 \\ -\tau & \text{if } z_k = 1 \wedge f_{s,z_k} = 0 \\ 0 & \text{otherwise.} \end{cases}$$

  - f_{s,z_k} = the probability of occurence of class z_k for scene type s

Experimental Evaluation

- **Dataset**:
- UIUC dataset
  - 1000 images form PASCAL VOC 2008.
  - 600 images for training, 400 images for testing.
  - 3 sentences on average for each images.
- VOC'10 trainval images are used to train DPM detectors.

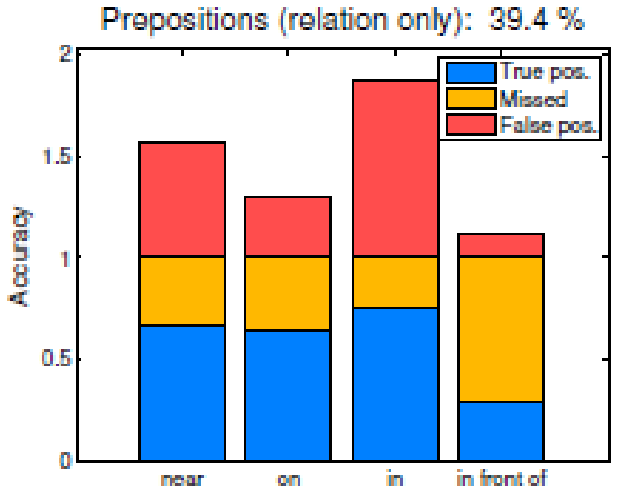
Experimental Evaluation
Extracting information from text

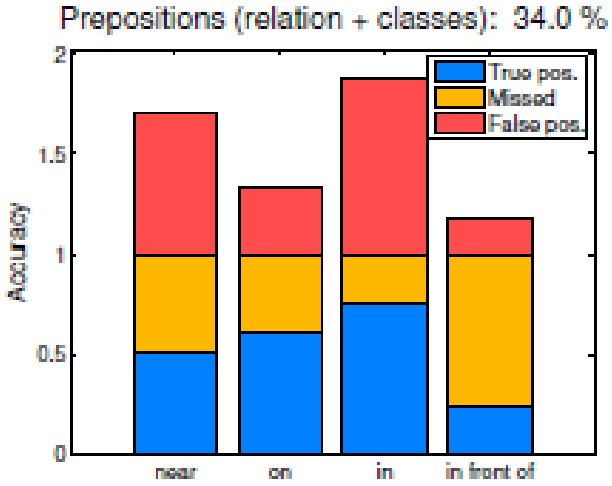


Text extraction accuracies: Cardinality

- (a) Reliability – the number of times the image contains at least as many objects as the predicted cardinality,
- (b) visual information – the number of true, missed and extra predictions of the objects from text, compared to the visual GT,
- (c) textGT – the number of true, missed and extra predictions of the object from text, compared to the textGT.

Experimental Evaluation
Extracting information from text





Text extraction accuracies: prepositions

Experimental Evaluation
Scene Classifier

- 15 scene types
  - dining area, room, furniture, potted-plant, cat, dog, city, motorbike, bicycle, field/farm, sheep, sky, airport, train, sea
- Text-based scene classifier achieved 76% classification accuracy. (The best visual scene classifier achieved only 40%.)

Experimental Evaluation
Holistic parsing using text

	back.	aerop.	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	pplant	sheep	sofa	train	monitor	averg.
Textonboost (unary) [24]	77.8	14.1	3.4	0.7	11.3	3.3	25.5	30.9	10.3	0.7	13.2	10.8	5.2	15.1	31.8	41.0	0.0	3.7	2.4	17.1	33.7	16.8
Holistic Scene Understanding [29]	77.3	25.6	12.9	14.2	19.2	31.0	34.6	38.6	16.1	7.4	11.9	9.0	13.9	25.4	31.7	38.1	11.2	18.8	6.2	23.6	34.4	23.9
[29] num boxes from text	77.8	26.7	14.3	11.5	18.6	30.8	34.4	37.9	17.2	5.7	19.0	7.3	12.4	27.3	36.5	37.1	11.6	9.4	6.2	25.7	43.8	24.3
ours	76.9	31.3	29.7	37.3	27.7	29.5	52.1	40.0	38.0	6.6	55.9	25.2	33.2	38.2	44.3	42.5	15.2	32.0	20.2	40.7	48.4	36.4

Comparison to the state-of-the-art that utilizes only image information in the UIUC sentence dataset. By leveraging text information our approach improves 12.5% AP. Note that this dataset contains only 600 PASCAL VOC 2008 images for training, and thus is significantly a more difficult task than recent VOC challenges which have up to 10K training images.

Experimental Evaluation
Parsing with Oracle

	back.	aerop.	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	pplant	sheep	sofa	train	monitor	averg.
Oracle Z - noneg	76.8	36.7	28.3	34.8	21.9	30.9	56.1	47.6	36.8	10.0	58.2	28.8	33.4	54.8	42.6	41.8	15.1	28.1	16.3	35.7	48.7	37.3
Oracle Z - neg	76.8	31.2	28.2	34.7	21.7	31.1	56.0	50.3	36.7	10.5	57.4	29.5	34.4	55.1	42.5	41.9	16.9	32.2	18.8	35.7	49.2	37.7
ours	76.9	31.3	29.7	37.3	27.7	29.5	52.1	40.0	38.0	6.6	55.9	25.2	33.2	38.2	44.3	42.5	15.2	32.0	20.2	40.7	48.4	36.4

Comparison to oracle Z

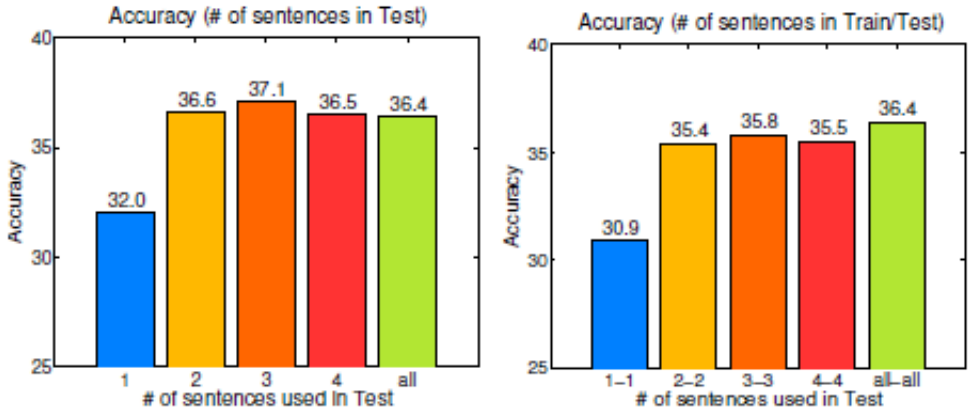
- “GT noneg”: At least as many boxes as dictated by the cardinality is encouraged to be on in the image.
- “GT-neg”: The boxes for classes with card. 0 is suppressed.

Experimental Evaluation
Model Components

	back.	aerop.	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	pplant	sheep	sofa	train	monitor	**averg.**
[29] num boxes from text	77.8	26.7	14.3	11.5	18.6	30.8	34.4	37.9	17.2	5.7	19.0	7.3	12.4	27.3	36.5	37.1	11.6	9.4	6.2	25.7	43.8	24.3
text rescored det.	77.9	28.1	12.8	31.0	32.3	32.3	44.3	42.1	27.4	6.4	26.0	22.1	24.1	29.5	37.3	40.8	11.4	21.6	16.5	26.8	45.1	30.3
+ card and scene	76.9	30.2	29.3	37.4	26.2	29.5	52.0	40.5	38.0	6.8	55.4	25.5	31.9	37.4	44.3	42.4	15.2	32.1	18.4	33.3	48.4	35.8
+ prep	76.9	31.3	29.7	37.3	27.7	29.5	52.1	40.0	38.0	6.6	55.9	25.2	33.2	38.2	44.3	42.5	15.2	32.0	20.2	40.7	48.4	36.4

Performance gain when employing different amounts of text information. Our method is able to gradually increase performance.

Experimental Evaluation
Amount of Text



Segmentation accuracy as a function of the number of sentences used per image: (**a**) all sentences in training, and different number in test, (**b**) N in train and N test

- (a) by using all available sentences per image in training, but different number in test,
- (b) by varying also the number of training sentences.

Experimental Evaluation
Amount of Text

# train	1	2	3	4	all	all	all	all	all
# test	1	2	3	4	1	2	3	4	all
T-DPM	35.6	36.1	36.3	36.7	35.6	36.3	36.4	36.6	36.6
DPM	31.5								

Results: Detector's AP (%) using text-based re-scoring for different number of sentences used per image in train and test.

sent 1: "A dog herding two sheep." **sent 2:** "A sheep dog and two sheep walking in a field." **sent 3:** "Black dog herding sheep in grassy field."

sent 1: "Passengers at a station waiting to board a train pulled by a green locomotive engine." **sent 2:** "Passengers loading onto a train with a green and black steam engine." **sent 3:** "Several people waiting to board the train."

sent 1: "Cattle in a snow-covered field." **sent 2:** "Cows grazing in a snow covered field." **sent 3:** "Five cows grazing in a snow covered field." **sent 4:** "Three black cows and one brown cow stand in a snowy field."

sent 1: "A yellow and white sail boat glides between the shore and a yellow buoy." **sent 2:** "Sail boat on water with two people riding inside." **sent 3:** "Small sailboat with spinnaker passing a buoy."

sent 1: "A table is set with wine and dishes for two people." **sent 2:** "A table set for two." **sent 3:** "A wooden table is set with candles, wine, and a purple plastic bowl."

sent 1: "An old fashioned passenger bus with open windows." **sent 2:** "Bus with yellow flag sticking out window." **sent 3:** "The front of a red, blue, and yellow bus." **sent 4:** "The idle tourist bus awaits its passengers."

image Yao [29] one sent two sent three sent

sent 1: "Passengers at a station waiting to board a train pulled by a green locomotive engine." **sent 2:** "Passengers loading onto a train with a green and black steam engine." **sent 3:** "Several people waiting to board the train."

sent 1: "Black and white cows grazing in a pen.' **sent 2:** "The black and white cows pause in front of the gate." **sent 3:** "Two cows in a field grazing near a gate."

image Yao [29] one sent two sent three sent

sent 1: "Two men on a plane, the closer one with a suspicious look on his face.'." **sent 2:** "A wide-eyed blonde man sits in an airplane next to an Asian man." **sent 3:** "Up close photo of man with short blonde hair on airplane."

sent 1: "Man using computer on a table." **sent 2:** "The man sitting at a messy table and using a laptop." **sent 3:** "Young man sitting at messy table staring at laptop."

QUESTIONS?