Incrementally Biasing Visual Search Using Natural Language Input

Evan Krause, Rehj Cantrell, Ekaterina Potapova, Michael Zillich, Matthias Scheutz

The scenario

Search-and-rescue after a tornado

The scenario

Search-and-rescue after a tornado hits my room.

The scenario

Search-and-rescue after a tornado hits my room.

- A robot is helping me clean up.

The scenario

Search-and-rescue after a tornado hits my room.

- A robot is helping me clean up.
- But there are hundreds of objects in this room.

The scenario

Search-and-rescue after a tornado hits my room.

- A robot is helping me clean up.
- But there are hundreds of objects in this room.
- Visual search is a notoriously hard problem.

The scenario

Search-and-rescue after a tornado hits my room.

There's a book and it's red...

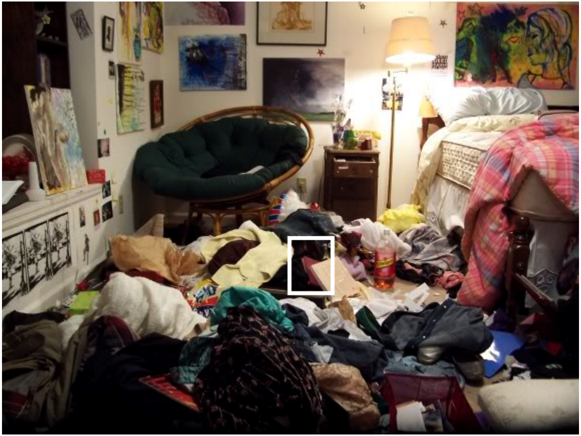
The scenario

Search-and-rescue after a tornado hits my room.

There's a book and it's red...

The scenario

Search-and-rescue after a tornado hits my room.



There's a book and it's red...

Given a natural language input with at least one visual referent, we want to:

Given a natural language input with at least one visual referent, we want to:

- derive from natural language top-down constraints on vision;

Given a natural language input with at least one visual referent, we want to:

- derive from natural language top-down constraints on vision;
- allow vision to use easy tasks to help with hard tasks.

Previous Research

- Recent work on constraining vision with top-down cues has not drawn these cues from natural language (c.f., Choi et al. [CBL+04], Frintrop et al. [FBR05], Navalpakkam et al. [NI06]).

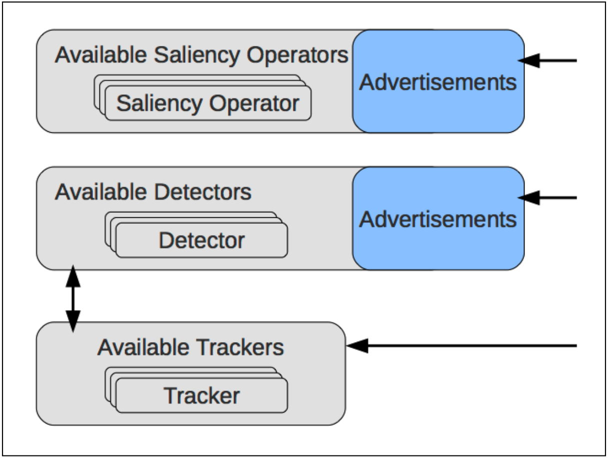
Previous Research

- Recent work on constraining vision with top-down cues has not drawn these cues from natural language (c.f., Choi et al. [CBL+04], Frintrop et al. [FBR05], Navalpakkam et al. [NI06]).
- More recent work Bergström et al. [BBK11] and Johnson-Roberson et al. [JRBS+11] use dialogue to bias object segmentation; however they explicitly require bidirectional interaction in order to refine segmentation.

Previous Research

- Recent work on constraining vision with top-down cues has not drawn these cues from natural language (c.f., Choi et al. [CBL+04], Frintrop et al. [FBR05], Navalpakkam et al. [NI06]).
- More recent work Bergström et al. [BBK11] and Johnson-Roberson et al. [JRBS+11] use dialogue to bias object segmentation; however they explicitly require bidirectional interaction in order to refine segmentation.
- We want to collect attentional cues from natural language to reduce the complexity of vision tasks.

Vision


Available Saliency Operators
Saliency Operator
Advertisements
Available Detectors
Detector
Advertisements
Available Trackers
Tracker

Saliency

- color - distance between the color of a point and a known value for a color term such as “red”

Saliency

- color - distance between the color of a point and a known value for a color term such as “red”
- height - distance between a point and a supporting surface below

Saliency

- color - distance between the color of a point and a known value for a color term such as “red”
- height - distance between a point and a supporting surface below
- location - saliency decreases in the form of a Gaussian from the selected image border or image center

Measuring the Cost of Visual Search

Different types of saliency operators have different computational costs.

$$ \tag{1} $$

Measuring the Cost of Visual Search

Different types of saliency operators have different computational costs.

$$C(S) \tag{1}$$

- $C(S)$: the cost of a search

Measuring the Cost of Visual Search

Different types of saliency operators have different computational costs.

$$C(S) = c_p \tag{1}$$

- $C(S)$: the cost of a search
- c_p: cost of checking one point; depends only on saliency type

Measuring the Cost of Visual Search

Different types of saliency operators have different computational costs.

$$C(S) = c_p * n_p \tag{1}$$

- $C(S)$: the cost of a search
- c_p: cost of checking one point; depends only on saliency type
- n_p: number of points

Measuring the Cost of Visual Search

Different types of saliency operators have different computational costs.

$$C(S) = c_p * n_p + c_f \tag{1}$$

- $C(S)$: the cost of a search
- c_p: cost of checking one point; depends only on saliency type
- n_p: number of points
- c_f: fixed costs; depends only on saliency type

Measuring the Cost of Visual Search

Different types of saliency operators have different computational costs.

$$C(S) = c_p * n_p + c_f \tag{1}$$

- $C(S)$: the cost of a search
- c_p: cost of checking one point; depends only on saliency type
- n_p: number of points
- c_f: fixed costs; depends only on saliency type

Incremental Vision

Incremental Vision

Incremental Vision

Incremental Vision

- Non-incremental

Incremental Vision

- Non-incremental
- Incremental

Incremental Vision

- Non-incremental
- Incremental

Processing Order

- Parallel

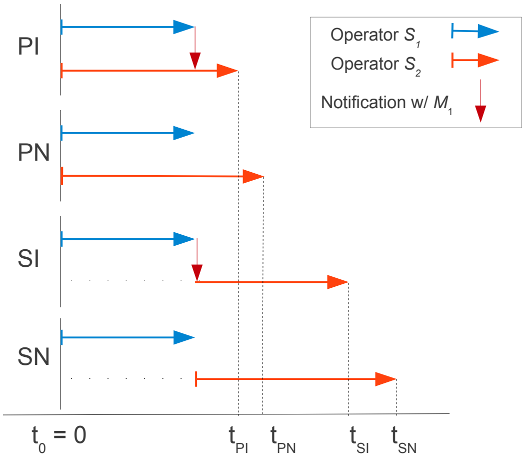
Processing Order

- Parallel
- Sequential

Processing Order

- Parallel
- Sequential
- Staggered

Four saliency operator configurations and their relative times to completion



Attention

- A mechanism that describes object detection modes

Attention

- A mechanism that describes object detection modes
- Checks all objects for saliency in some order until the entire space has been searched

Attention

- A mechanism that describes object detection modes
- Checks all objects for saliency in some order until the entire space has been searched
- With attention: Begins with objects containing the most salient points

Coordinating a visual search using spoken input

1. We begin a visual search whenever we spot a noun phrase (signaled by a determiner, adjective or noun).

Coordinating a visual search using spoken input

1. We begin a visual search whenever we spot a noun phrase (signaled by a determiner, adjective or noun).
2. Once a visual search has begun, we send all descriptors to vision immediately.

Coordinating a visual search using spoken input

1. We begin a visual search whenever we spot a noun phrase (signaled by a determiner, adjective or noun).
2. Once a visual search has begun, we send all descriptors to vision immediately.
3. We end the visual search when we find some word that cannot be part of it.

Incremental Processing of Linguistic Input

can you see

(vision waiting)

Incremental Processing of Linguistic Input

can you see a
(search started)

Incremental Processing of Linguistic Input

can you see a tall

(constraint: tall)

Incremental Processing of Linguistic Input

can you see a tall red
(constraint: red)

Incremental Processing of Linguistic Input

can you see a tall red object

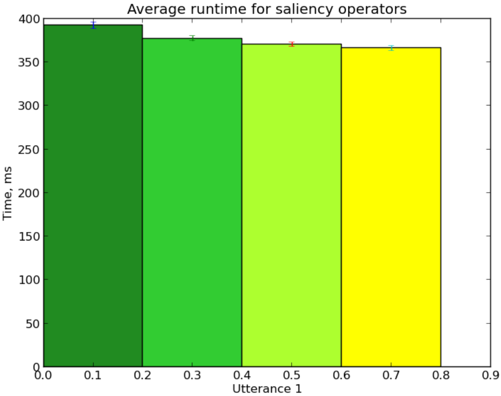
Incremental Processing of Linguistic Input

can you see a tall red object at left

(constraint: left)

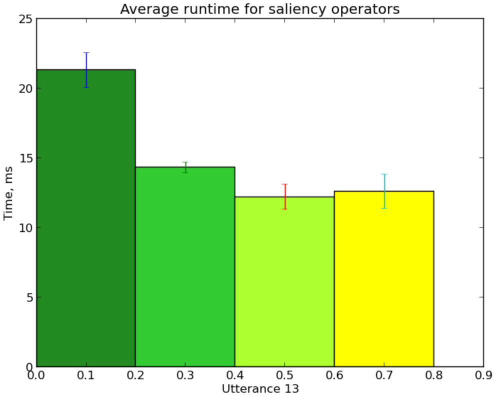
(results returned)

Experimental results.



"The short green top object"
Left to Right: SN, SI, PN, PI

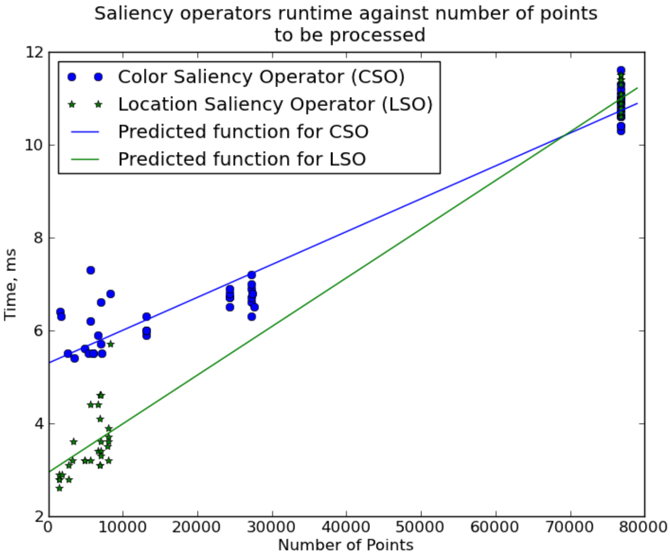
Experimental results.



"The red right object"

Left to Right: SN, SI, PN, PI

Experimental results.



Color and location saliency operator runtime vs. n_p.

Experimental Design

- Seven scenes each contained 11 objects for each of which a uniquely-identifying verbal description containing one to four descriptive or locational constraints (e.g., *red*, *tall*, *right*, *front*) was formulated.

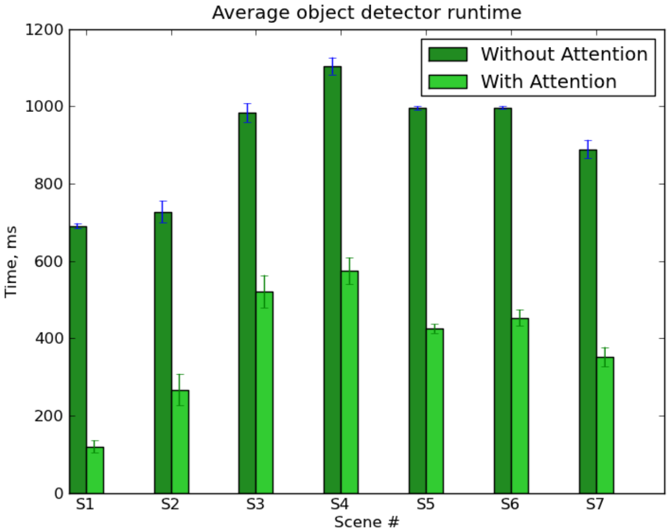
Experimental Design

- Seven scenes each contained 11 objects for each of which a uniquely-identifying verbal description containing one to four descriptive or locational constraints (e.g., *red*, *tall*, *right*, *front*) was formulated.
- This method resulted in 77 separate description/scene pairs.

Experimental Design

- Seven scenes each contained 11 objects for each of which a uniquely-identifying verbal description containing one to four descriptive or locational constraints (e.g., *red*, *tall*, *right*, *front*) was formulated.
- This method resulted in 77 separate description/scene pairs.
- As each scene was viewed, each associated description was presented incrementally (as in spoken natural language) 10 times.

Experimental results.



Average object detector runtime across scenes.

Future Work

- Work on human subjects (Chiu and Spivey [CS12]) has shown that in some cases parallel processing is faster than serial, but in other cases the reverse is true; additional experiments are needed to show whether our model accounts for this as well.

Future Work

- Work on human subjects (Chiu and Spivey [CS12]) has shown that in some cases parallel processing is faster than serial, but in other cases the reverse is true; additional experiments are needed to show whether our model accounts for this as well.
- A threshold value is used to decide whether a detected object meets the description. Future work will employ probabilistic models of object properties (such as the incrementally learned KDE based representations of Skocaj et al. [SJK+10]).

Future Work

- Work on human subjects (Chiu and Spivey [CS12]) has shown that in some cases parallel processing is faster than serial, but in other cases the reverse is true; additional experiments are needed to show whether our model accounts for this as well.
- A threshold value is used to decide whether a detected object meets the description. Future work will employ probabilistic models of object properties (such as the incrementally learned KDE based representations of Skocaj et al. [SJK+10]).
- Development is also needed to fuse confidence measures from natural language with such probabilistic measures from the vision system.

Future Work

- Work on human subjects (Chiu and Spivey [CS12]) has shown that in some cases parallel processing is faster than serial, but in other cases the reverse is true; additional experiments are needed to show whether our model accounts for this as well.
- A threshold value is used to decide whether a detected object meets the description. Future work will employ probabilistic models of object properties (such as the incrementally learned KDE based representations of Skocaj et al. [SJK+10]).
- Development is also needed to fuse confidence measures from natural language with such probabilistic measures from the vision system.
- The reverse direction, using visually-acquired information to constrain natural language interpretation, needs to be explored.

Acknowledgements

This work was in part funded by ONR grant N00014-11-1-0493 and the Austrian Science Fund (FWF) under project TRP 139-N23 InSitu.

References

N. Bergstrom, M. Bjorkman, and D. Kragic, *Generating Object Hypotheses in Natural Scenes through Human-Robot Interaction*, IROS, 2011, pp. 827–833.

Sang Choi, Sang Ban, Minho Lee, Jang Shin, Dae Seo, and Hyun Yang, *Biologically motivated trainable selective attention model using adaptive resonance theory network*, Biologically inspired approaches to advanced IT, Springer Berlin / Heidelberg, 2004, pp. 456–471.

Eric M. Chiu and Michael J. Spivey, *The role of preview and incremental delivery on visual search*, Proceedings of the 34th Annual Conference of the Cognitive Science Society, 2012, pp. 216–221.

Simone Frintrop, Gerriet Backer, and Erich Rome, *Selecting what is important: Training visual attention*, Proc. of the 28th Annual German Conf. on AI (KI'05), 2005.

M. Johnson-Roberson, J. Bohg, G. Skantze, J. Gustavson, R. Carlsson, and D. Kragic, *Enhanced Visual Scene Understanding through Human-Robot Dialog*, IROS, 2011, pp. 3342–3348.

Thanks for your attention!
...Questions?