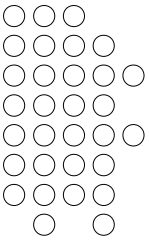
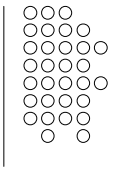
Grounded Semantic Composition for Visual Scenes

Peter Gorniak and Deb Roy
MIT Media Lab

Presentation by David DeVault



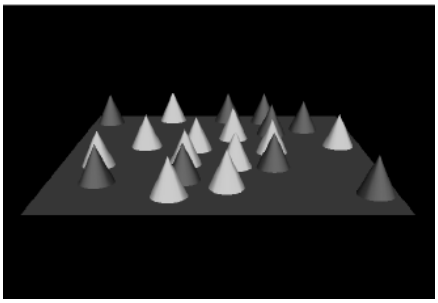
Outline



- Bird's eye view of the paper
- The BISHOP spatial description task
- Descriptive strategies
- The BISHOP language understanding system
- Evaluation
- Discussion

Domain of this work

Understanding referring expressions in visual scenes like:

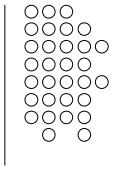


E.g. “the far back purple cone
that’s behind a row of green ones”

What they did

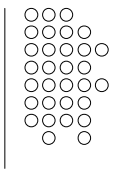
- Collected descriptions from 6 subjects
- Analyzed them for grammar → visual feature patterns
- Hand-built a grammar, a parser, and a semantic interpreter that composes visual features to determine referents

How they evaluated it



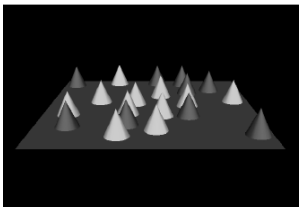
- Collected descriptions from 3 new subjects
- Used model to correctly identify 59% of speaker referents

Outline



- Bird's eye view of the paper
- The BISHOP spatial description task
- Descriptive strategies
- The BISHOP language understanding system
- Evaluation
- Discussion

The BISHOP spatial description task



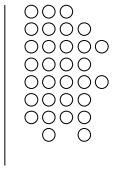
- The BISHOP task
  - Parameters
    - N <= 30 objects
    - random positions
    - 1/2 green, 1/2 purple
  - Designed to elicit spatial descriptions
  - Just reference understanding
    - No generation
    - No models of dialogue, collaboration, clarification, or agreement

Spatial description task

- Note about the BISHOP task:
  - *Only some objects are easily described:*

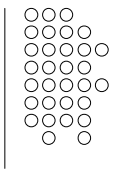
    "in the centre there are a bunch of green cones, four of them, um, actually there are more than four, but, ah, there's one that's in the centre pretty much of the pile of them up to the it's at the top, ahm, how can you say this... or the seventh cone from the right side"
    (followed by the listener counting cones by pointing at the screen).
  - So they allow their subjects to choose objects that he or she felt to be concisely yet not trivially describable.

Outline



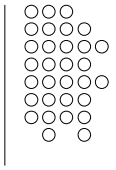
- Bird's eye view of the paper
- The BISHOP spatial description task
- Descriptive strategies
- The BISHOP language understanding system
- Evaluation
- Discussion

Experimental setup



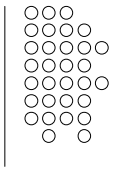
- Two subjects at a time: describer & listener
- Sit back-to-back, facing displays of same scene
- Scene starts with 30 objects
- While (objects remain)
  - Describer selects target object
  - While (target not identified)
    - Describer utters new description of target
    - Listener clicks presumed target
  - Target is removed from scene

Data collection



- "Development data": 6 subjects, 268 spoken descriptions
  - Used segmentation algorithm based on pause structure to reassemble (fused utterance, correct selection) pairs
- Transcribed speech recording verbatim including speech errors (false starts, etc.)
- "Test data": 3 subjects, 179 spoken descriptions

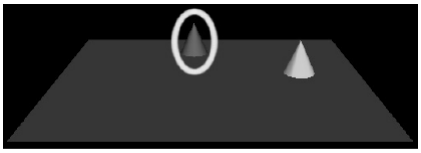
Analysis of development data



- The combination of a visual feature and corresponding linguistic device is referred to as a *descriptive strategy*.
  - E.g. "green" might be associated with a probability distribution function defined over a color space.
- Manually catalogue descriptive strategies

Color

- 96% of utterances employ “green” or “purple”
- Usually as an adjective:



“the purple cone”

- Color adjectives ***always*** immediately precede the noun they modify:
  - “the left purple one” – COMMON!
  - “the purple left one” – NEVER!

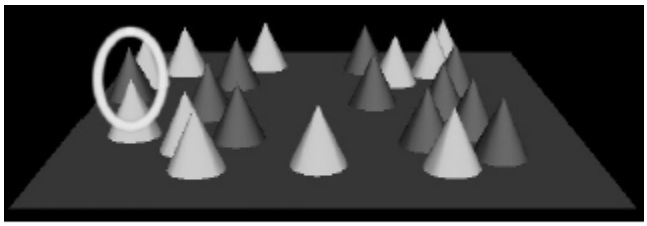
Color

- 7% of utterances employ “green” or “purple” as a noun or elliptically omit the noun
  - E.g. “the leftmost purple”

“the purple cone”

Spatial regions and extrema

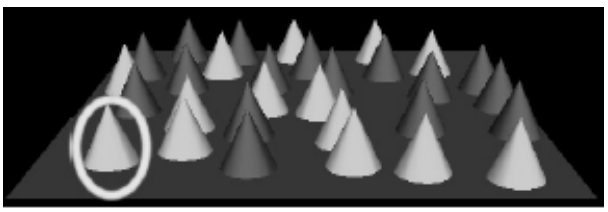
- 72% of utterances used single spatial extrema



"the purple one on the left side"

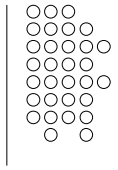
Spatial regions and extrema

- 20% of utterances used spatial regions

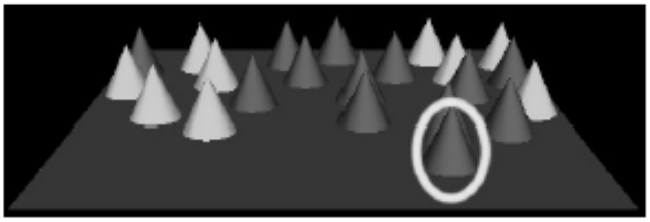


"the green cone at the left bottom"

Spatial regions and extrema

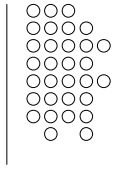


- 28% of utterances used multiple extrema or regions

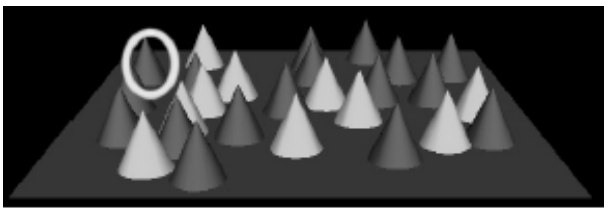


"the lowest purple on the right hand side"

Grouping



- 12% of utterances exhibited grouping



"there's three on the left side; the one in the furthest back"

Spatial relations

- 6% of utterances used spatial relations

"the green cone below the green cone"

"there's a purple cone that's it's all the way on the left hand side but it's it's below another purple"

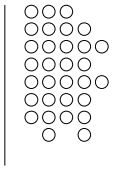
Anaphora

- 4% of utterances contained anaphora

"the closest purple one on the far left side"

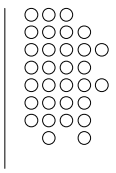
"the green one right behind that one"

Outline



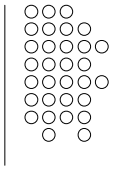
- Bird's eye view of the paper
- The BISHOP spatial description task
- Descriptive strategies
- The BISHOP language understanding system
- Evaluation
- Discussion

Synthetic vision



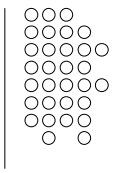
- Treat 2D perspective projection of 3D scene as pseudo camera image
- Segment image into individual objects
- Use segmentation compute object *visual features*:
  - average RGB color
  - center of mass
  - inter-object distances
  - etc.

Lexical entries



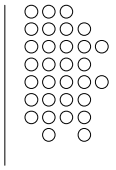
- Lexical entries are marked up with semantic features
  - whether entry refers
  - its *semantic composer*
    - function describing compositional behavior of entry
    - Inputs and outputs are *concepts* = ranked sets of objects or groups
  - whether arguments are on left or right in syntax
  - etc.

Example lexical entry



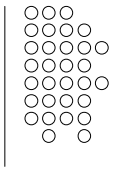
```
<LIST NAME="above">
  <MAP>
    <BOOL NAME="Ambiguous" VALUE="true"/>
      <INT NAME="Arity" VALUE="0"/>
      <MAP NAME="Composer">
        <INT NAME="SpatialFeatureIndex" VALUE="0"/>
        <STRING NAME="Type" VALUE="SpatialSemanticComposer"/>
      </MAP>
      <BOOL NAME="FixedArity" VALUE="true"/>
      <INT NAME="LeftArity" VALUE="1"/>
      <STRING NAME="POS" VALUE="P"/>
      <STRING NAME="ReferenceType" VALUE="none"/>
      <INT NAME="RightArity" VALUE="1"/>
  </MAP>
</LIST>
```

Parsing



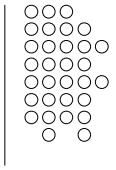
- Use a CFG
- Bottom-up chart parse
- Do composition as new constituents are built
- Partial parse extends constituents to span unknown words
- Longest constituent wins

Semantic composers



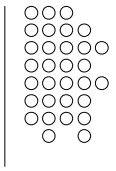
- Composers act on incoming objects
- Produce a set of objects with attached referent strengths (a “concept”)
- Composition is delayed when
  - arguments do not refer
    - E.g. composers for “the left green” are delayed until some referring head noun arrives
  - any argument is unavailable

Color composer



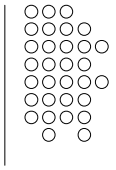
- Used for “green” and “purple”
- Learn a model
  - Collect set of labelled green and purple cones from synthetic vision
  - Construct 3D Gaussian distribution over average RGB values
- Output value of the *pdf* as “reference strength”
- Drop objects not exceeding some threshold

Spatial extrema/region composers



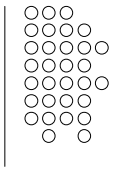
- Used for
  - “bottom”, “leftmost” - minima
  - “top”, “rightmost” - maxima
  - “middle” - region
- Use hand-built exponential decay functions to determine “reference strength”
- Drop objects not exceeding some threshold

Grouping composer



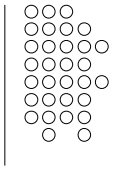
- Used for “group”, “cones”, etc.
- Find all groups in the scene
- Assign “reference strength” = average distance between inter-group objects
- Output ranked set of groups
- “of” can split apart groups
  - “the leftmost one of the three green ones”

Spatial relation composers



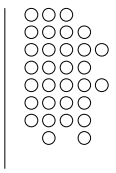
- Uses hand-built attentional vector sum (AVS) calculation
- Returns target objects with “reference strength” a function of target and possible landmark objects:
  - Angle of the vector connecting centers of mass w.r.t. reference vector
  - Angle of vector connecting closest points w.r.t. reference vector
  - Relative spatial dimensions
- Drop objects not exceeding some threshold

Anaphoric composers



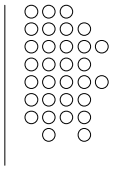
- Used for
  - “that” in “left of that one”
  - “previous” in “left of the previous one”
- Returns the last object removed
- Marks it so further composition uses last visual scene

Post-parse filtering



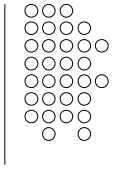
- Extract longest referring constituent(s) from the chart
  - If referent unambiguous, select it
  - If referent is unambiguous group, select best matching object
  - If referent is ambiguous group, select random object

Outline



- Bird's eye view of the paper
- The BISHOP spatial description task
- Descriptive strategies
- The BISHOP language understanding system
- Evaluation
- Discussion

Evaluation

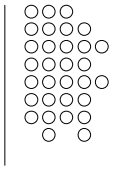


Utterance Set	Accuracy - Development	Accuracy - Testing
All	76.5%	58.7%
All except 'Other'	83.2%	68.8%
All except 'Other' and 'Errors' (clean)	86.7%	72.5%

Table 2: Overall Results

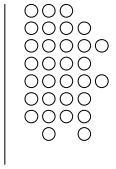
- All except 'Other' : excludes ignored descriptive strategies
- All except 'Other' and 'Errors' : also excludes segmentation errors (<1%) and speech errors

Sources of error



- Speech segmenter and utterance reassembler produced a few errors (< 1% of utterances)
- Producing an accurate covering grammar difficult
  - though helped by loose parsing, they say
- Errors in treatment of descriptive strategies

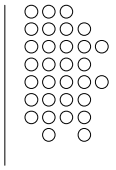
Errors by descriptive strategy



Utterance Set	Accuracy - Development	Accuracy - Test
Spatial Extrema	86.8% (132/152)	77.4% (72/93)
Combined Spatial Extrema	87.5% (49/56)	75.0% (27/36)
Grouping	34.8% (8/23)	38.5% (5/13)
Spatial Relations	64.3% (9/14)	40.0% (8/20)
Anaphora	100% (6/6)	75.0% (3/4)

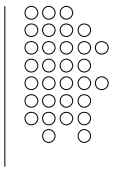
Table 3: Detailed Results

Example error 1



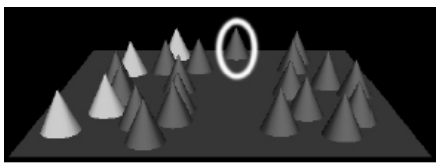
- “the leftmost one in the front”
  - “leftmost” is relative to front cones, but BISHOP treats it as absolute

Example error 2



- Bad semantics for “middle”

“the purple one right in the middle”



“the purple one in the middle”

Example error 3

- Insufficiently rich grouping strategy

"the cone on the right in the pair of cones"

"the purple cone on at the front of the row of three purple cones"

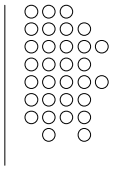
Example error 4

- Insufficiently rich coverage of anaphora

"the next cone in the row"

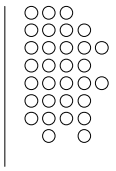
"the last cone in the row"

Some sample descriptions



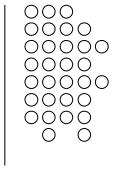
- the green cone in the middle
- the purple cone behind it
- the purple cone all the way to the left
- the purple cone in the corner on the right
- the green cone in the front
- the green cone in the back next to the purple cone
- the purple cone in the middle front
- the purple cone in the middle
- the frontmost purple cone
- the green cone in the corner
- the most obstructed green cone
- the purple cone hidden in the back
- the purple cone on the right in the rear
- the green cone in the front
- the solitary green cone
- the purple cone on at the front of the row of three purple cones
- the next cone in the row
- the last cone in the row
- the cone on the right in the pair of cones

Outline



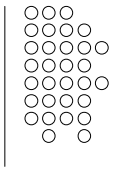
- Bird's eye view of the paper
- The BISHOP spatial description task
- Descriptive strategies
- The BISHOP language understanding system
- Evaluation
- Discussion

Grounded Semantic Composition



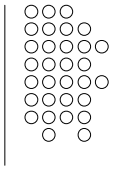
- "We use the term grounded semantic composition to highlight that both the semantics of individual words and the word composition process itself are visually-grounded."
  - Note *grounded* here means perceptually grounded.
- "In our model, each lexical entry's meaning is grounded through an association to a visual model."

Grounded vs. symbolic semantic representations?



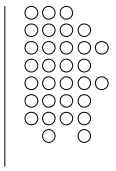
- "Symbolic formal approaches to semantics leave the details of non-linguistic influences on meaning unspecified, whereas we take the computational modeling of these influences as our primary concern."

In their words…



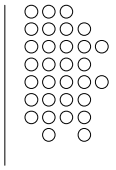
- P. 433: “Most prior systems use a declaratively stated set of semantic facts that is disconnected from perception... Our emphasis, however, is on a system that can actively ground word and utterance meanings through its own sensory system…. Schuler's [and Winograd's etc.] system requires a **human-specified** clean logical encoding of the world state, which ignores the noisy, complex and difficult-to-maintain process linking language to a sensed world. We consider this process, which we call the grounding process, one of the most important aspects of situated human-like language understanding.

But what does this come to?

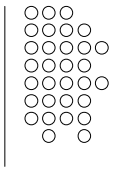


- "SAM (Brown, Buntschuh, & Wilpon, 1992) and Ubiquitous Talker (Nagao & Rekimoto, 1995) are language understanding systems that map language to objects in visual scenes. Similar to SHDRLU, the underlying representation of visual scenes is symbolic and loses much of the subtle visual information that our work, and the work cited above, focus on. Both SAM and Ubiquitous Talker incorporate a vision system, phrase parser and understanding system. The systems translate visually perceived objects into a symbolic knowledge base and map utterances into plans that operate on the knowledge base. In contrast, we are primarily concerned with understanding language referring to the objects and their relations as they appear visually.“

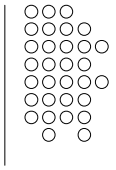
Socially grounded reference



- They need dynamic models of dialogue, including correction and clarification to really model reference correctly.
- But they have gotten pretty far without them!

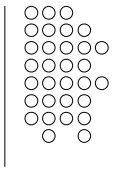


Comparison to earlier work 1



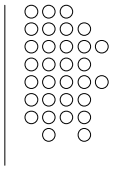
- SHRDLU: claim their system has
  - more robust / broader grammatical coverage,
  - advantages in not requiring a clean symbolic formulation
  - Cares about context: not just set theoretic logical semantics

Comparison to earlier work 2



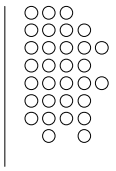
- Review of Brown-Schmidt et al. 2002
  - Nice ambiguity results
  - Agreement behavior
  - People do agreement and use discourse and visual context to disambiguate underspecified referring expressions
    - G&R eliminate dialogue so they can “computationally model the strategies our participants employ”

Comparison to earlier work 3



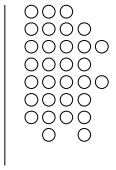
- SAM / Ubiquitous Talker

Comparison to earlier work 4: lessons about grounding?



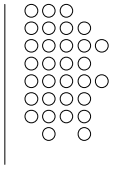
- P. 433: “Most prior systems use a declaratively stated set of semantic facts that is disconnected from perception... Our emphasis, however, is on a system that can actively ground word and utterance meanings through its own sensory system…. Schuler's [and Winograd's etc.] system requires a **human-specified** clean logical encoding of the world state, which ignores the noisy, complex and difficult-to-maintain process linking language to a sensed world. We consider this process, which we call the grounding process, one of the most important aspects of situated human-like language understanding.
- Re: SAM / Ubiquitous Talker: Even though they use a vision system, because they use Symbolic encodings of visual scenes they lose much of the subtle visual information that our work… focuses on."

Comparison to earlier work 5



- (Roy & Pentland, 2002) *multiplicative* semantic composition
- “highest” is associated with a probability distribution centered at a particular height
- DeVault & Stone (2004) employ a better computational model of vague adjectives

Remarks



- Good:
  - Tries to connect semantic representations to the perceived physical world
  - Methodology for building systems: collect data with wizard-of-oz setup, use it to design system
    - Reduces imposition of grammatical limitations through lack of foresight
  - Partial parsing