The Language of Vision

Patrick Cavanagh

2021

Presented by: Saaketh Medepalli (1/18/23)

Agenda

- Motivation
- Language for Vision
- Components of Visual Language
- Visual Grammar
- Vision is the Original Language
- Conclusions
- Piazza Discussions

Motivation

- Studies show that language may have first developed close to ~1 million years ago (Uomini & Meyer, 2013)
- Even with liberal estimates, *not enough* time to reflect radical changes in brain structure/function to support language

How did language develop so quickly?

Motivation

- Studies show that language may have first developed close to ~1 million years ago (Uomini & Meyer, 2013)
- Even with liberal estimates, *not enough* time to reflect radical changes in brain structure/function to support language

How did language develop so quickly?

"In the beginning was the grammar of vision – in the end came the word"

- Richard Gregory

Why Vision?

- Visual processing uses ~30% of cortex in humans
- Visual information not only received, but sent to other centers of the brain (language, motor, etc.)
  - E.g. Describing a waterfall at a park, Jumping over an obstacle
- How is this information organized and sent?

Why Vision?

- Visual processing uses ~30% of cortex in humans
- Visual information not only received, but sent to other centers of the brain (language, motor, etc.)
  - E.g. Describing a waterfall at a park, Jumping over an obstacle
- How is this information organized and sent?
  - Visual information → requires other centers to have their own visual systems

Why Vision?

- Visual processing uses ~30% of cortex in humans
- Visual information not only received, but sent to other centers of the brain (language, motor, etc.)
  - E.g. Describing a waterfall at a park, Jumping over an obstacle
- How is this information organized and sent?
  - Visual information → requires other centers to have their own visual systems
  - Packaged as a language that can be decoded → Plausible…

Why Language?

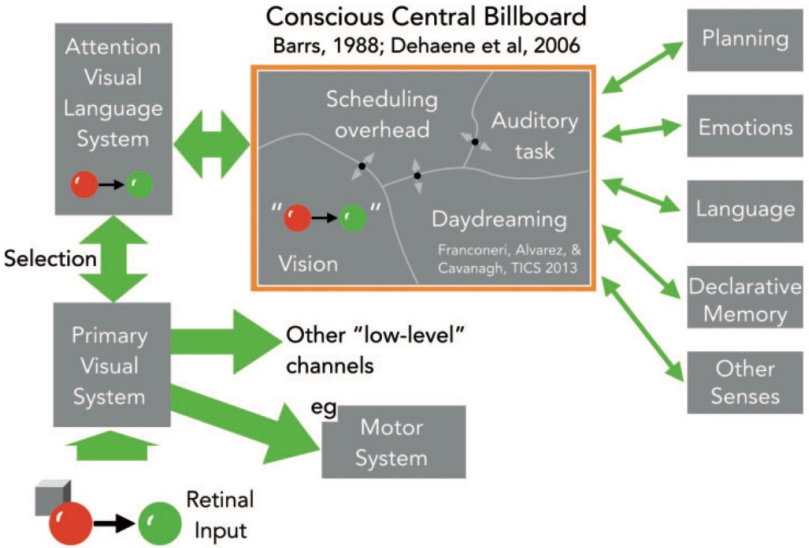
- Information can be sent using unordered labels
  - Suppose 2000 labels learned → Only 2000 messages possible
  - Divide into 2 classes → $(2000/2)^2 = 10^6$ messages possible!
- Dividing into classes more efficient *but* format/grammar needed to specify classes

Why Language?

- Information can be sent using unordered labels
  - Suppose 2000 labels learned → Only 2000 messages possible
  - Divide into 2 classes → $(2000/2)^2 = 10^6$ messages possible!
- Dividing into classes more efficient *but* format/grammar needed to specify classes
- ***Enter Language!***

Language for Vision

- Some broadband connections (e.g., motor)
- Remaining are sent from attention-selecting & packaging
- Billboard → conscious visual percepts

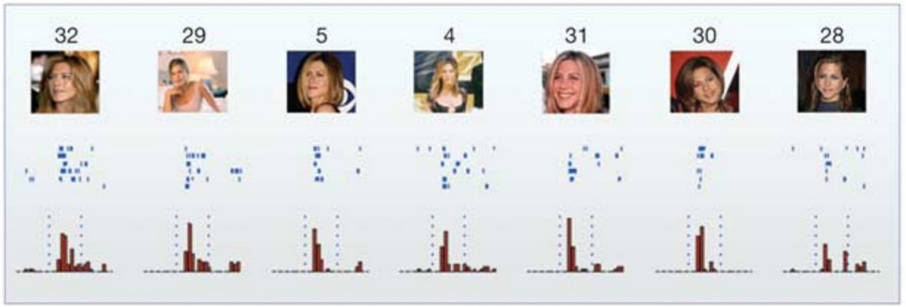


Components of Visual Language

- Visual and spoken languages describe world → *likely* similar components including:
  - Nouns ⟺ Objects
  - Verbs ⟺ Actions
  - Prepositions ⟺ Spatiotemporal Relations
- Four key elements:
  1. Compositionality
  2. Arbitrariness
  3. Displacement
  4. Recursion

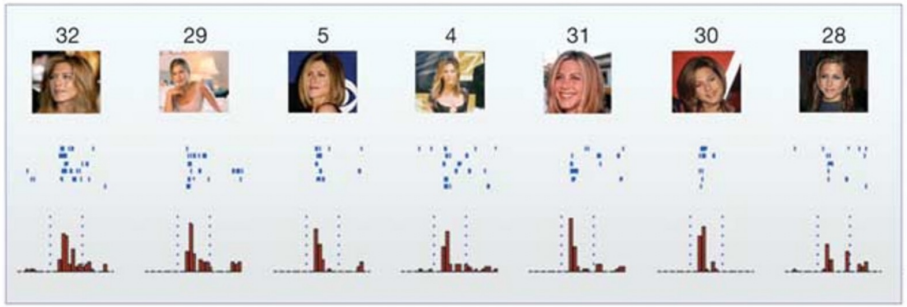
Visual Nouns = *Objects*

- Output of ventral object recognition area (population code)
- From Quiroga et al. (2005), single unit in hippocampus fires for different pictures of Jennifer Aniston



Visual Nouns = *Objects*

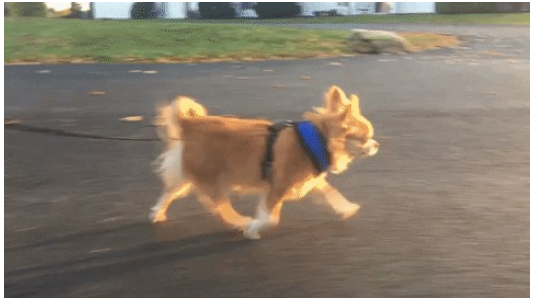
- Output of ventral object recognition area (population code)
- From Quiroga et al. (2005), single unit in hippocampus fires for different pictures of Jennifer Aniston



$\Longrightarrow$ Arbitrariness ✓

Visual Verbs = *Actions*

- Familiar actions = verbs of vision (e.g., walking)



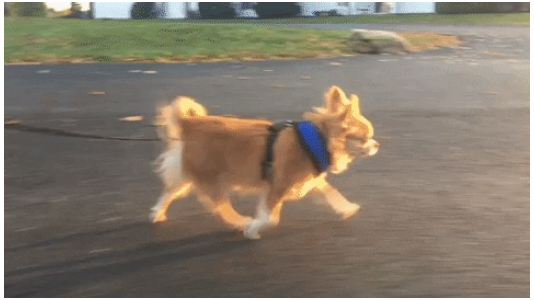
Dog walking

Man walking

- Common motion patterns giving characteristic label

Visual Verbs = *Actions*

- Familiar actions = verbs of vision (e.g., walking)



Dog walking

Man walking

- Common motion patterns giving characteristic label

$\Longrightarrow$ Compositionality ✓

Past Tense and Future Tense

- Like spoken language, visual verbs also have tenses
- Past tense: record of the cause of the current state
  - Visual vs. Cognitive Inference?

Visual Inference

Cognitive Inference

Past Tense and Future Tense

- Like spoken language, visual verbs also have tenses
- Future tense: immediate prediction of what is to happen
  - Visual vs. Cognitive inference less clear

Woman about to fall in the pool

Visual Prepositions = *Spatiotemporal Relations*

- *Behind* of special significance: addresses occlusion handling

*Attention also necessary on both objects of comparison

Visual Prepositions = *Spatiotemporal Relations*

- *Behind* of special significance: addresses occlusion handling

$\Longrightarrow$ Displacement ✓

*Attention also necessary on both objects of comparison

Causality

- To complete a sentence: subject, verb, *and* object (in that order)
  - Who did what to whom?
  - Critical component to a package sent out to other centers
- Studies have shown that some levels of causality worked out directly in visual system
  - E.g., Series of causal events (collisions) entirely retinotopic

Recursion

- Last element to language system
- Occurs when one description embedded in another

Recursion

- Last element to language system
- Occurs when one description embedded in another

$\Longrightarrow$ Recursion ✓

What about the Grammar?

- Grammar: how components are structured together
- So far, grammars for machine vision stop at descriptions of static images
- Rules of grammar left for future work
  - *Hint: ungrammatical* images

Break in Syntax?

What about the Grammar?

- Grammar: how components are structured together
- So far, grammars for machine vision stop at descriptions of static images
- Rules of grammar left for future work
  - Careful!

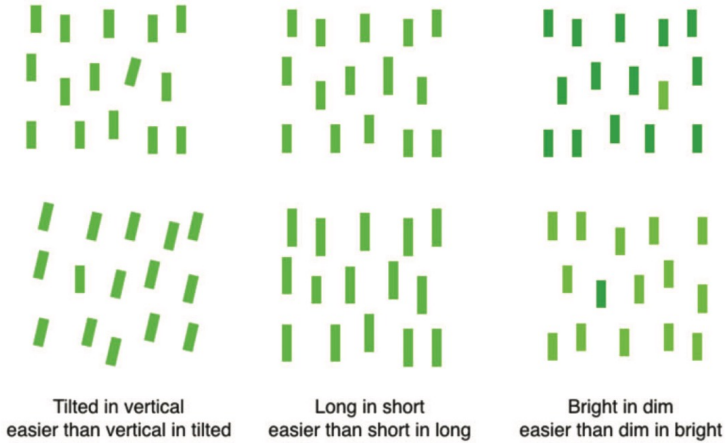
Cognitive inference breach, not visual!

Is Vision the Original Language?

- How does vision acquire grammar?
  - Regularities in visual input → classes of entities (nouns, verbs and prepositions of vision)
- Possible that vision created the foundation for acquisition of grammar for spoken, solving evolution issue!
- Likely too simplistic…

Is Vision the Original Language?

- Through evolution, vision likely created a template for acquiring *any* language
  - I.e., mechanism developed for humans to extract regularities in visual input, and eventually speech input, for vision and language



Tilt denotes **tilted** or vertical
Length denotes **long** or short
Brightness denotes **bright** or dim

Conclusions

- Vision seen as only reception (Pure Vision)
- Language of vision corrects this view to *include production* of compressed packages of information (Interactive Vision)

3 main claims:

1. Attention helps format & send out descriptions
2. Message packaged by language = contents of visual perception
3. Language of vision requires grammar

Piazza Discussion #1

Vision = 'ur-language' *(@26_f2)*

- General agreement that vision seems to provide a template for acquiring new languages
  - Nicaraguan Sign Language: Interesting example supporting this hypothesis, where children invented their own sign language *without adult supervision*
- Do other species' visual language resemble ours?

Piazza Discussion #1 (from Class)

Vision = 'ur-language' *(@26_f2)*

- The idea that vision is the origin of development of language seems to be a tenuous claim
  - E.g., Things like paintings developed very recently (~4000 BC or so), which directly interferes with the inclusion of *recursion* as an element
- Much more likely that a template for 'perceptual language' is present but perhaps not one from vision

Piazza Discussion #2

Language of Neural Networks *(@26_f5)*

- Neural network, modeled loosely after the brain, formulates its own visual language of sorts to perform tasks
  - DETR uses a convolutional backbone to collect a set of features (description) which are then used by transformer encoder-decoder for object detection (language processing?)
- Since many DL backbones perform compression (optimized by some loss), could the most efficient encoding resemble language as described in the article?

Piazza Discussion #2 (from Class/Piazza)

Language of Neural Networks *(@26_f5)*

- Neural networks seem to generate their own grammar and can achieve incredible feats in language
  - This doesn't seem to be mentioned or discussed until the very end, or at least put into context with the rest of the paper
- With similar and inputs and cost functions, NNs perhaps learn representations necessarily alike those of the brain
  - Rajiv Govindjee: "I would not be surprised if the most efficient way to encode the whole input space into a useful (latent) representation for common output queries resembles language in terms of the structures described (reusable relations, recursion)"
  - Personal thoughts: I think this goes back to whether the brain performs something like backpropagation (if not, above conclusion doesn't hold)

Piazza Discussion #3

Visual grammar influenced by spoken language? *(@26_f8)*

- Just as how visual input influences spoken language, perhaps abstractions learned from spoken language also influence visual processing?
- Possible that humans develop *perceptual* grammar from all sensory inputs → used for all internal communication
- Does the learned grammar depend on external rewards (outside of regularities)?

Piazza Discussion #3 (from Piazza)

Visual grammar influenced by spoken language? *(@26_f8)*

- Goes back to the first Discussion Question
  - Perhaps the original grammar does not come from vision but is rather a general template that is influenced by multiple senses
- From Prof. Kuipers: Highly unlikely that language influenced foundational parts of vision given that vision develops well into the first year of the babies' life while language learning occurs later

Citations

Cavanagh, P. (2021). The Language of Vision*. Perception, 50(3), 195–215. https://doi.org/10.1177/0301006621991491

Quiroga, R., Reddy, L., Kreiman, G. *et al.* Invariant visual representation by single neurons in the human brain. *Nature* **435**, 1102–1107 (2005). https://doi.org/10.1038/nature03687

Uomini NT, Meyer GF (2013) Shared Brain Lateralization Patterns in Language and Acheulean Stone Tool Production: A Functional Transcranial Doppler Ultrasound Study. PLoS ONE 8(8): e72693. https://doi.org/10.1371/journal.pone.0072693