Active Observer Visual Problem-Solving Methods are Dynamically Hypothesized, Deployed and Tested

Markus D. Solbach

John K. Tsotsos

Department of Electrical Engineering and Computer Science

York University, Canada

November 18th, 2021

Tsotsos Lab
Active and Attentive Vision

ADVANCES in COGNITIVE SYSTEMS 2021
Ninth Annual Conference

YORK UNIVERSITÉ UNIVERSITY
Centre for Vision Research

Overview

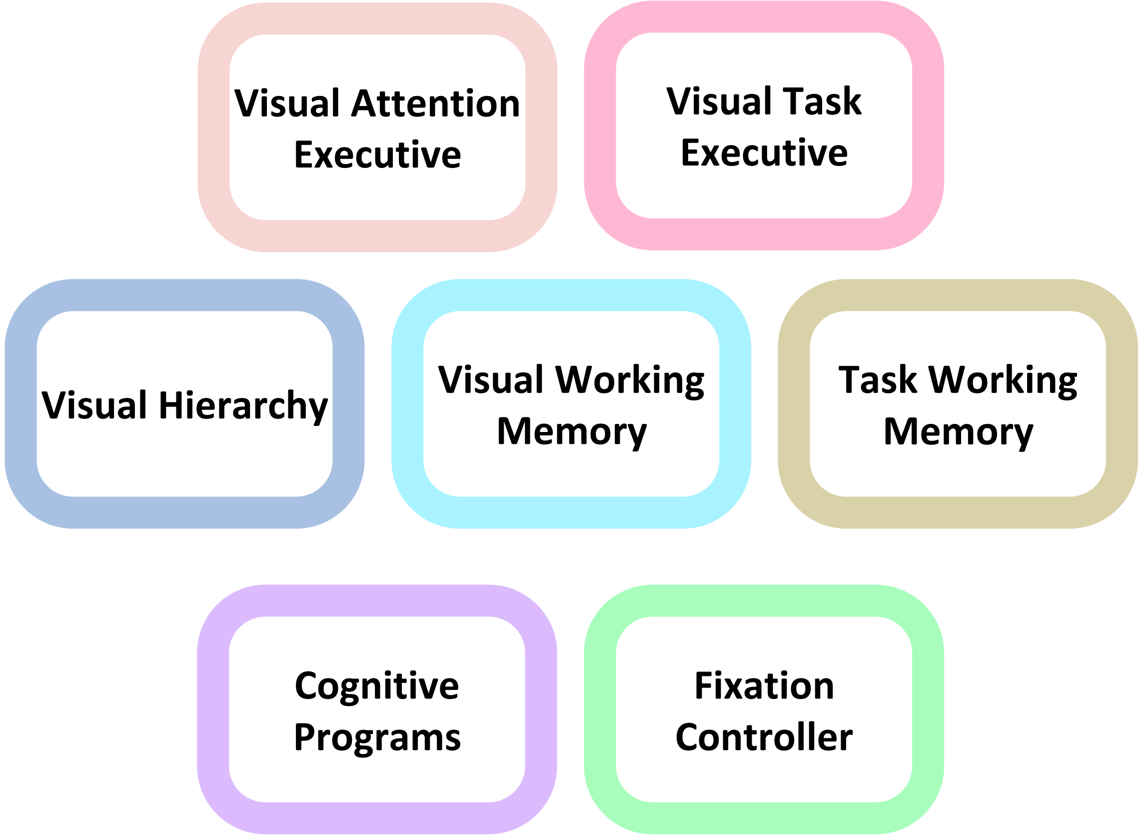
- Background
- Experimental Set Up
- Some Results
- Conclusion

Overview

- Background
- Experimental Set Up
- Some Results
- Conclusion

STAR

Selective
Tuning
Attentive
Reference



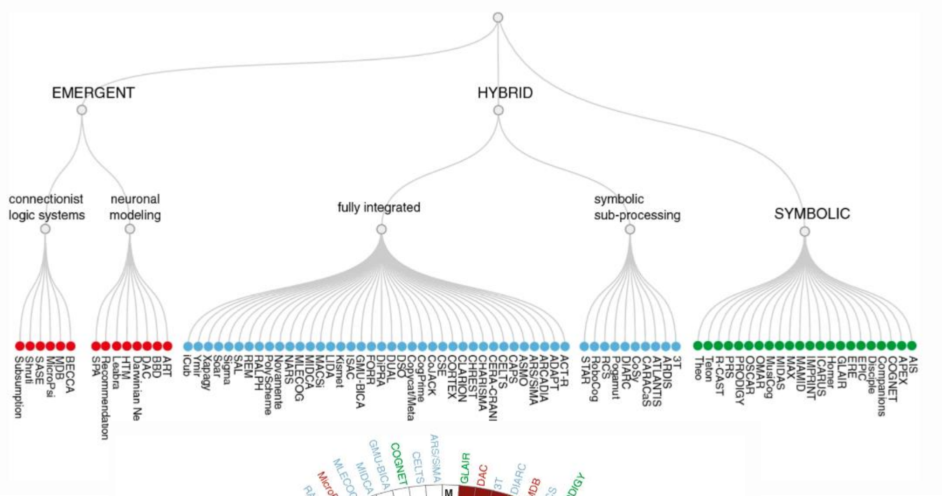
What roles does attention play in a behaving visual agent?

Tsotsos, J. K., & Kruijne, W. (2014). Cognitive programs: software for attention's executive. *Frontiers in Psychology*, 5.

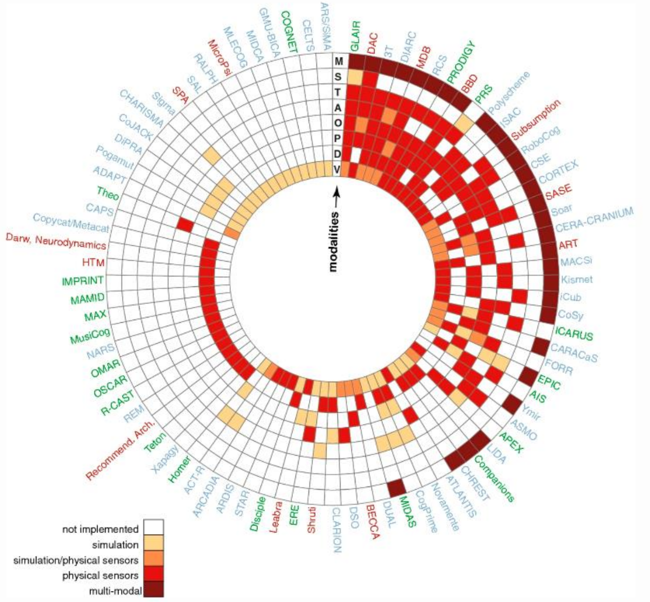
Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*(1), 17-94.

- Wanted to be certain our question had novelty – confirmed
- 86 architectures and 700 application systems using those architectures included

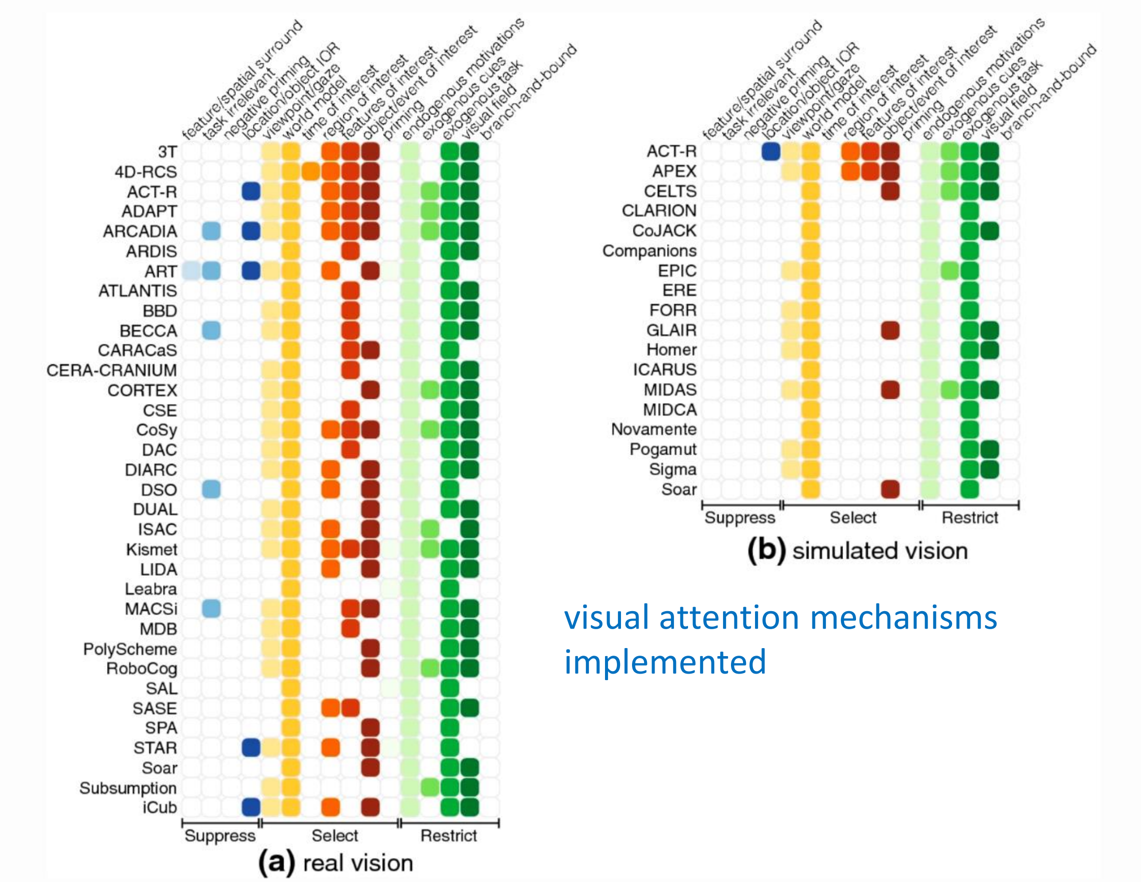
taxonomy of architectures



modalities represented across architectures



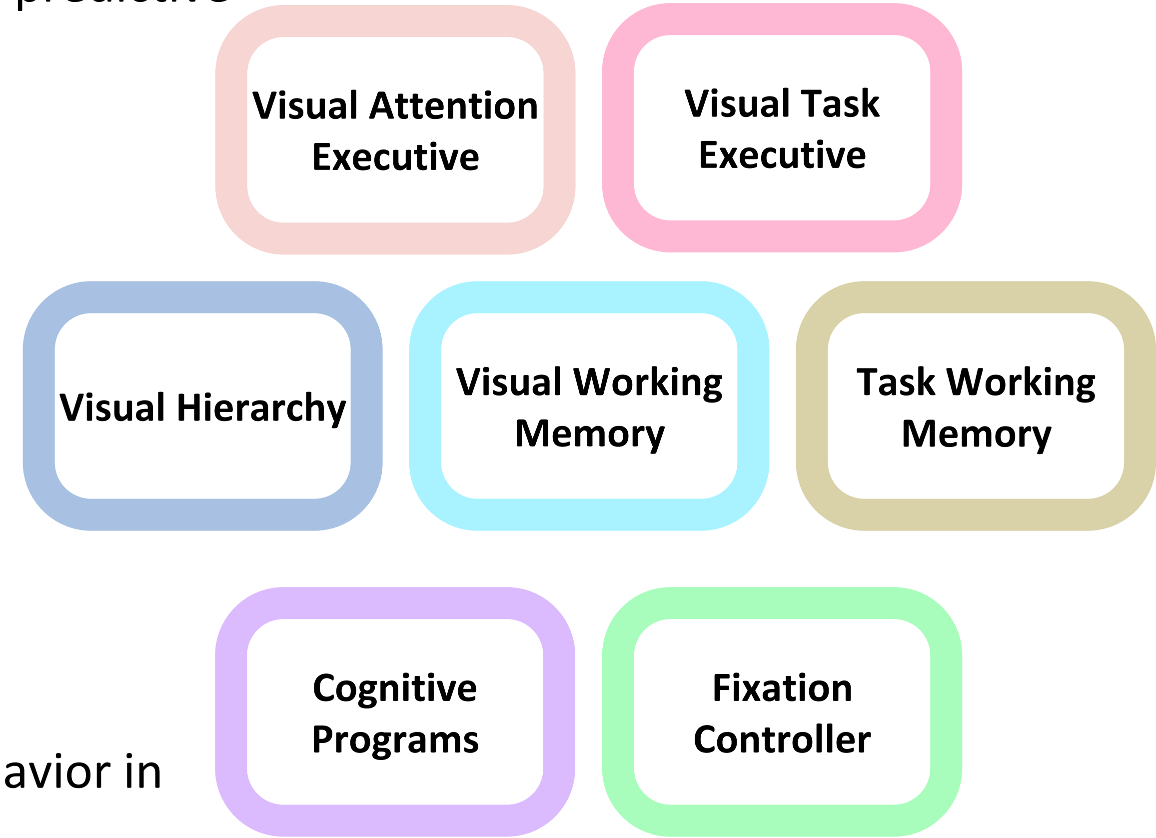
M – multi-modal
S – smell
T – touch
A – audition
O – other
P – proprioception
D – symbolic inout
V – vision



visual attention mechanisms implemented

Elements of STAR

- Elements of STAR in place included:
- Selective Tuning model of visual attention with strong predictive evidence
    - Tsotsos (2011) MIT Presss
- Attentive visual hierarchy prototypes
    - Biparva & Tsotsos (2019) ICCV-W
    - Rosenfeld et al. (2018) CVPR-W
- Human-equivalent fixation control
    - Tsotsos et al. (2016) *J. Eye Movement Research*
    - Wloka et al. (2018) CVPR
- Representation plan : Cognitive Programs
    - Tsotsos & Kruijne (2014) *Frontiers in Psychology*
- Needed a task that requires complex active visual behavior in order to understand the scope and nature of attentional/executive control – past explorations into such human behavior seem minimal



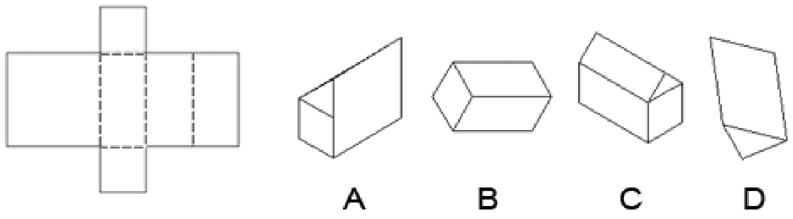
Our Broad Span of Visuospatial Abilities

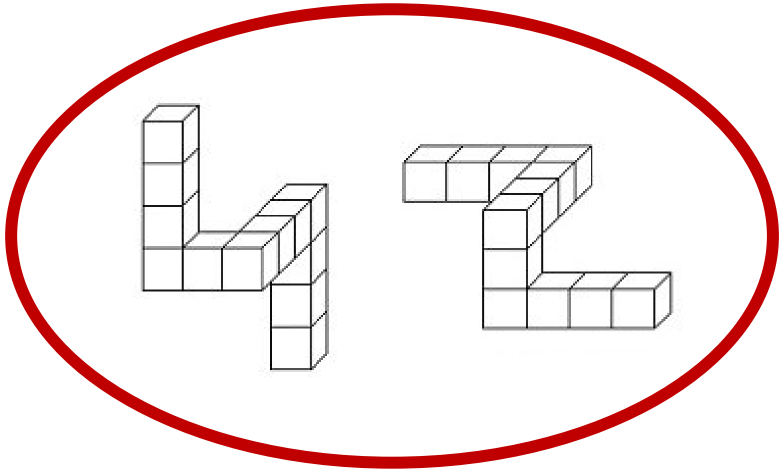
From Carroll 1993:

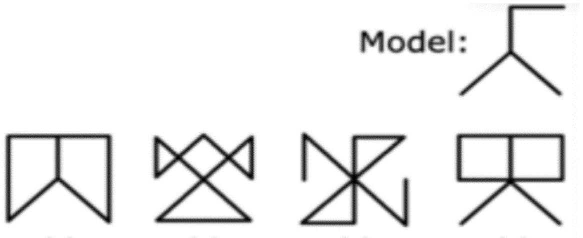
- **Spatial Visualization:** processes of apprehending, encoding, and mentally manipulating spatial forms (paper folding or spatial relations).

- **Speeded Rotation:** requires mental transformations but also involves manipulations (usually planar rotations) of two-dimensional objects and speed is emphasized (card rotation and the flag test, requiring a same-different judgment for each rotated pattern).

- **Visuospatial Perceptual Speed:** speed or efficiency of perceptual judgments (Identical Pictures Test - quickly identify which of five alternative patterns is identical to a model pattern; Hidden Patterns Test: quickly decide whether a simple target pattern is present in a more complex pattern).







Are Two Objects the Same or Different?

- This is an everyday task:
  - Often, we design objects to be easily discriminable, say by colour or size or pattern, but this is not always the case.
  - Consider a task where you are given a part during an assembly task and need to go to a bin of parts in order to find another one of the same (e.g. assembly of IKEA furniture).
  - LEGO requires one to perform such tasks many times while constructing a block configuration, either copying from a plan, mimicking an existing one or building from one's imagination

- We push this to the extreme in order to discover its characteristics and limitations

- The Problem: What is the sequence of actions to correctly determine if two objects are the same?

➢Equal interest for Human Behaviour as well as Robot Behaviour

954 I. van Rooij/Cognitive Science 32 (2008)

Table 2
A sample of computational-level theories whose combinatorial search spaces are potentially super-polynomial in $|i|$

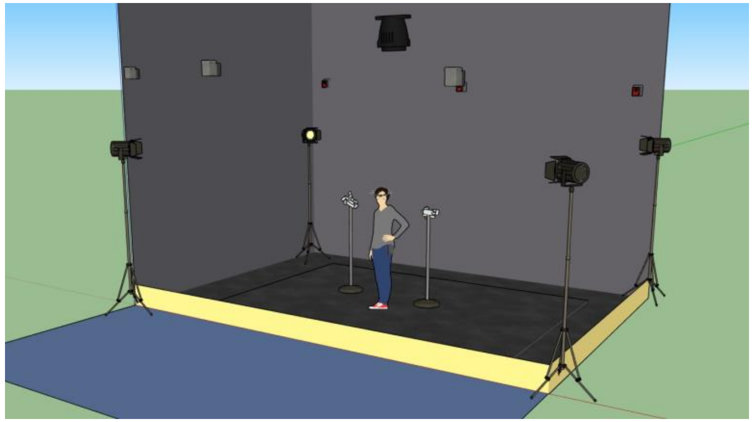
Cognitive Domain	Computational-Level Theory (ψ_T)	References
Categorization	*Input:* A set of objects, P, and a (dis)similarity value for each $(p, q) \in P \times P$. *Output:* A partition of P into categories such that within-category similarity and between-category dissimilarity is maximum.	Pothos and Chater (2001, 2002); Rosch (1973); Rosch and Mervis (1975)
Similarity	*Input:* Two objects x and y and a set of transformation rules T. *Output:* The length of the shortest sequence of transformation rules from T that, when applied to x, yields y.	Chater and Vitányi (2003a, 2003b); Hahn, Chater, and Richardson, (2003)
Coherence	*Input:* A set of propositions, P, positive and negative constraints, $C \subseteq P \times P$. *Output:* A truth assignment $T(P)$ satisfying a maximum number of constraints.	Millgram (2000); Thagard (2000); Thagard and Verbeurgt (1998); van Rooij (2003)
Gestalt perception	*Input:* A string s and a decoding function $f : C \to S$ mapping codes to strings. *Output:* A code $c \in C$ such that $f(c) = s$ and the length of c is minimum.	van der Helm (2004); van der Helm and Leeuwenberg (1986, 1996)
Visual search	*Input:* A target T, a visual display D, and two numbers x and y. *Output:* A subset $S \subseteq D$ such that the number of (mis)matching elements in S and T is at least x (at most y).	Kube (1991); Tsotsos (1990); van Rooij (2003).
Defeasible reasoning	*Input:* A knowledge base K and a set of default rules R. *Output:* All propositions $p_1, p_2, \ldots, p_k$ derivable from K using R, such that $p_1, p_2, \ldots, p_k$ and K are consistent.	Oaksford and Chater (1993, 1998); Reiter (1980).
Bayesian inference	*Input:* A knowledge base K and a set of competing hypotheses H. *Output:* A hypothesis $h \in H$ that maximizes the conditional probability $P(h\|K)$.	Chater, Tenenbaum, and Yuille (2006); Cooper (1990); Roth (1996)
Decision making	*Input:* A set of choice alternatives P and a value function $u : S \to N$ (where $S \subseteq P$ and N is a set of numbers). *Output:* A subset $S \subseteq P$ such that $u(S)$ is maximum.	Fishburn and LaValle (1993, 1996); van Rooij, Stege, and Kadlec (2005)
Language processing	*Input:* Surface form s, lexicon D, lexical-surface form relation mechanism M. *Output:* Set of lexical forms U generated by D from which M can create s.	Barton, Berwick, and Ristad (1987); Ristad (1990, 1993); Wareham (1996, 1999, 2001)
Planning	*Input:* An initial state s, a goal state g, and a collection of operators O. *Output:* A sequence of operators that when applied to s produces g.	Bylander (1994); Joseph and Plantinga (1985); Newell and Simon (1988a, 1988b)
Network harmony	*Input:* A harmonic (e.g., Hopfield) neural network. *Output:* An activation pattern that maximizes harmony.	Jagota (1997); Rumelhart et al. (1986); Smolensky and Legendre (2006)
Network learning	*Input:* A neural network N and function f. *Output:* A weight assignment to the connections in N such that N computes f.	Judd (1990); Parberry (1994, 1997)

Overview

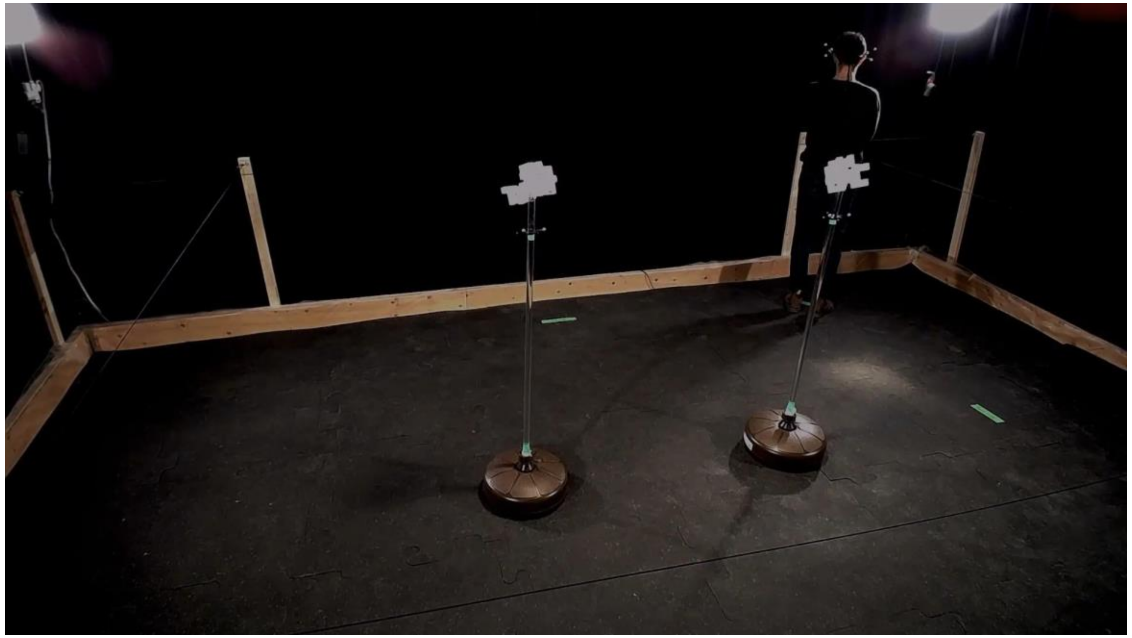
- Background
- Experimental Set Up
- Some Results
- Conclusion

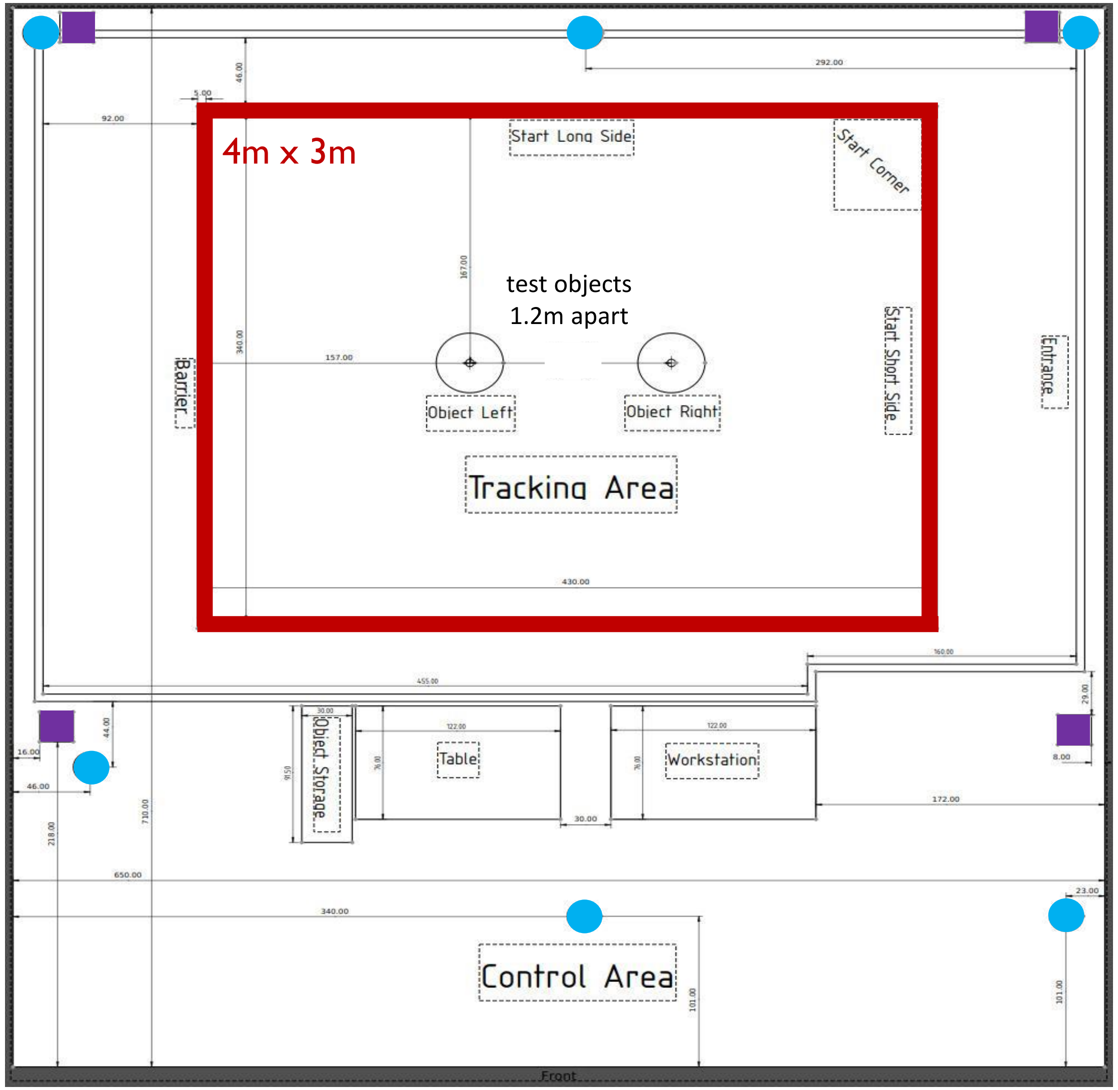
PESAO:Psychophysical Experimental Setup for Active Observers

Solbach, M.D. & Tsotsos, J.K. (2021). Tracking Active Observers in 3D Visuo-Cognitive Tasks, Proc. ACM Symposium on Eye Tracking Research and Applications, (1-3)



motion tracking camera
light source

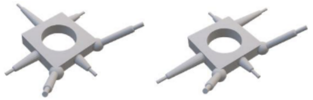




PESAO

Motion Tracking OptiTrack Robotics Package with six Flex 13 cameras on 10ft camera stands,

Gaze Tracking Tobii Pro Glasses 2 with TobiiPro Lab software and prescription lens package

Object Tracking Custom 

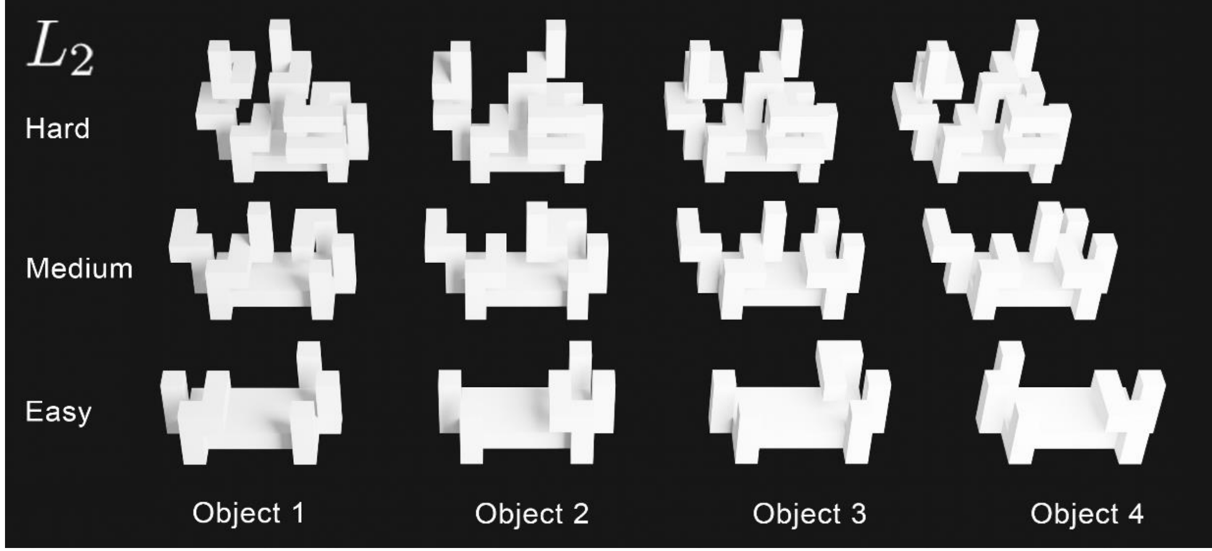
Custom Gear OptiTrack M4 markers on custom frame

Custom Software PESAOlib: control and execute experiments,
record data with accurate to microsecond-level timestamps, synchronize, analyze, display

Lighting Five 660 LED Video light-panels from Neewer one in each corner and one above
Colour temperatures from 3200 – 5600K and lumen of up to 7300 Lux/m.
Light level sensor: Yocto-Light-V2 by Yoctopuce up to 65,000 lux.

Basic Workstation Windows 10 (Motive motion capture software)
Intel i7-7700k, Ryzen 7 2700x or comparable
> 8GB RAM ; > 128GB SSD storage ; NVIDIA Quadro, > 2GB VRAM

Objects



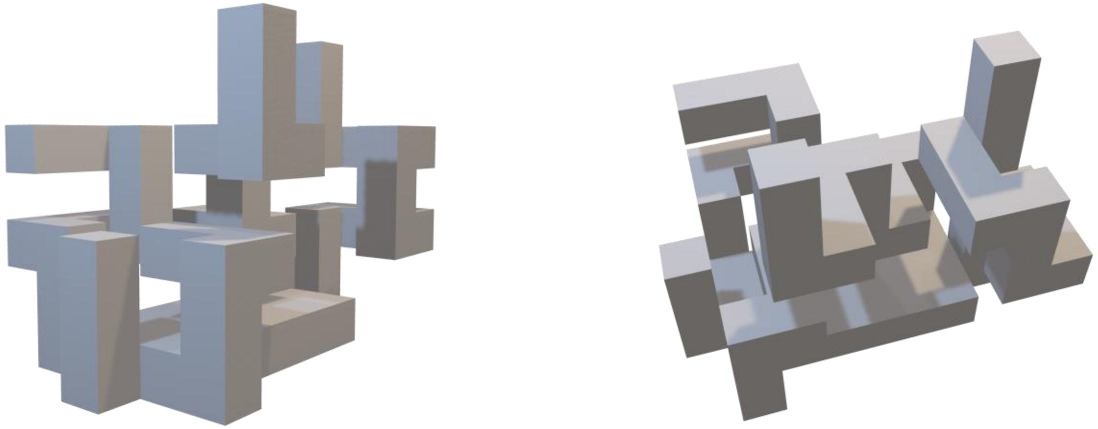
- Shepard and Metzler objects are used as an inspiration
  - Made from geometrical blocks
  - Assembled flush

- TEOS
  - All objects are made up by a base and a number of cuboids
  - Always starting with a base
  - Five connection points
  - Known complexity ($compl = n + 1$)
  - Common coordinate system

  - Also:
    - No intersection of elements
    - No continuation of direction

  - L_2 TEOS data set

The same or different?



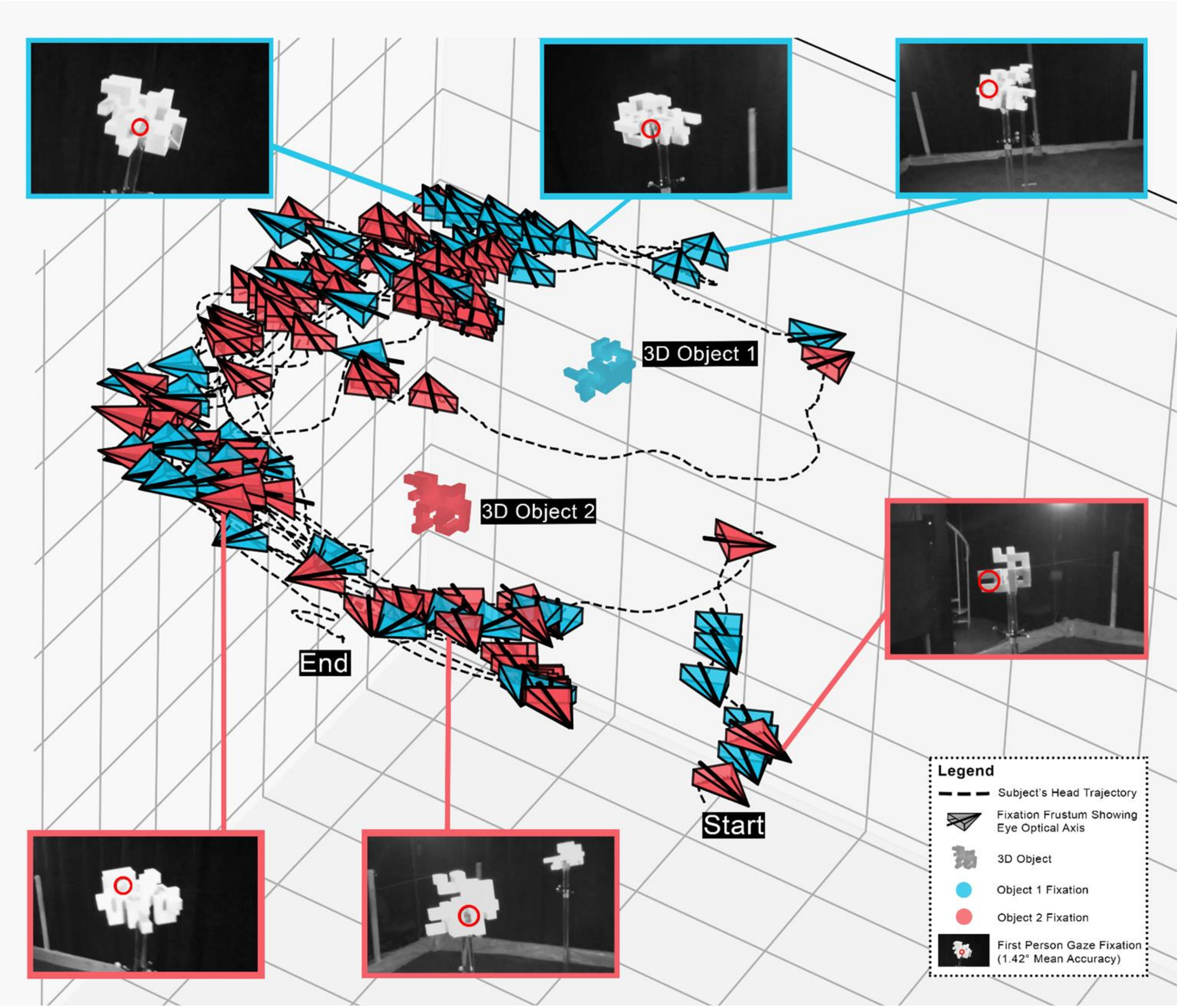
They are different!

Solbach, Markus D., Tsotsos, John K. "Blocks World Revisited: The Effect of Self-Occlusion on Classification by Convolutional Neural Networks" Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2021.

Experiment

Experiment

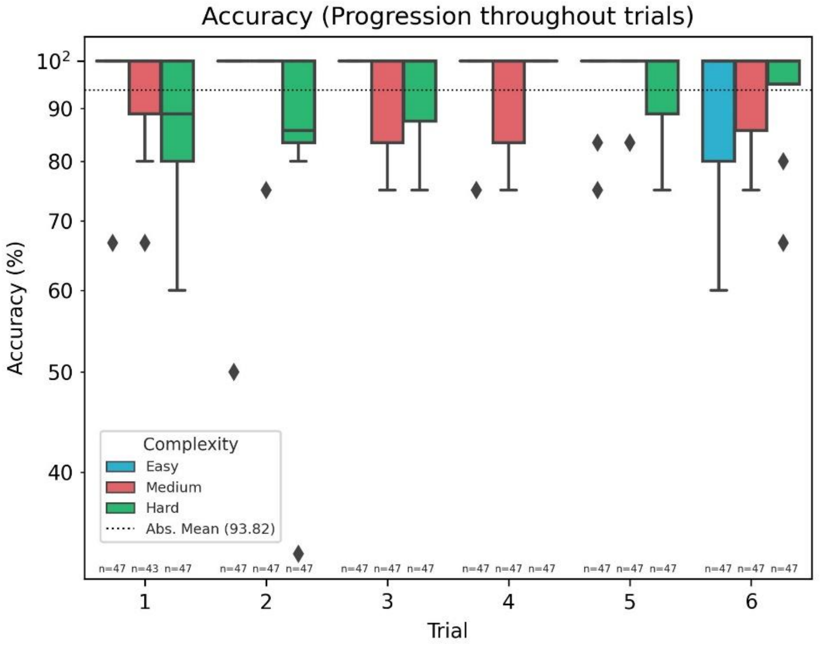
- 47 subjects recorded
- 846 trials
- About 80,000 fixations; 40,000sec of video
- Constant lighting; shadows minimized
- Subjects interviewed afterwards



Overview

- Background
- Experimental Set Up
- Some Results
- Conclusion

Characteristics across Trials (1)

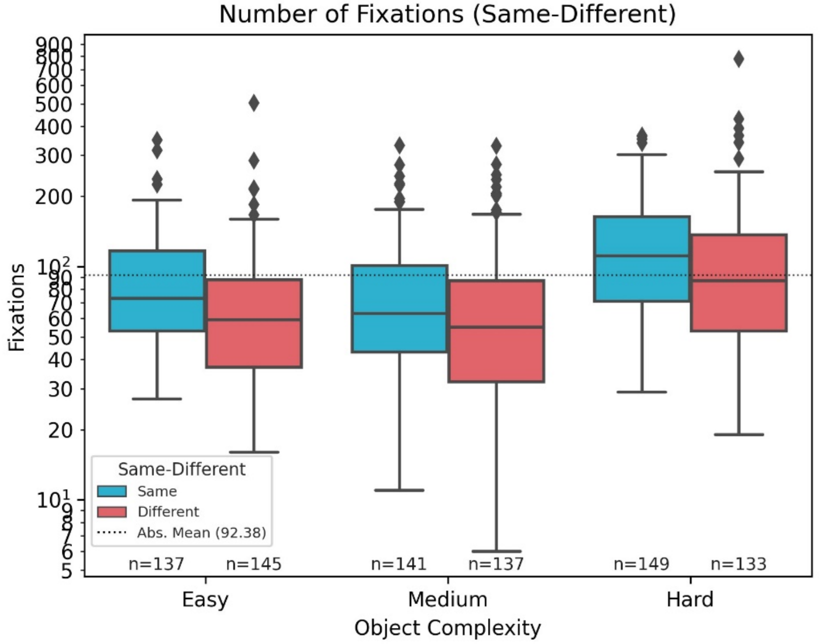


- People are very good at this task. The range of response times from simplest to most difficult cases ranged from 4 - 298 sec. and accuracy from 80% to 100%.

- There is a great deal of data acquisition occurring during all trials with the range of eye movements (and thus separate fixations and separate images processed) from 6 to 800 fixations.

- Error responses take more time and require more fixations

- Subjects did up to 18 trials each; no learning effect observed

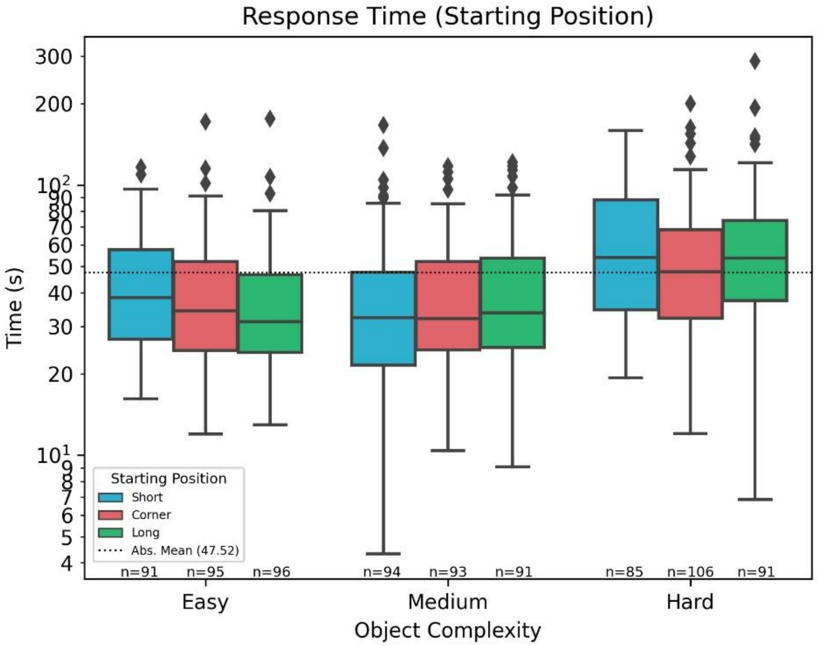


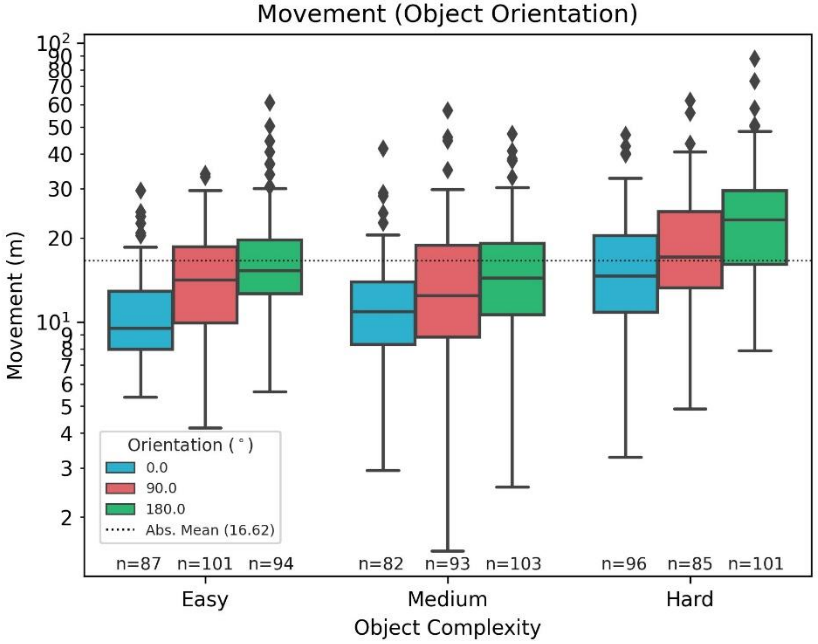
Characteristics across Trials (2)

- Overall average of 93 fixations during an average of 48 seconds; with 300ms per fixation change, this leaves over 20 seconds for 'thinking' (reasoning, planning, decision-making, working memory).

- The absolute mean of head movement was 16.62m and no trial was less than 1m.

- A clear trend between amount of head movement and orientation; 0° least amount, 90° increase of 2-5m, and 180° additional increase of 1-5m.

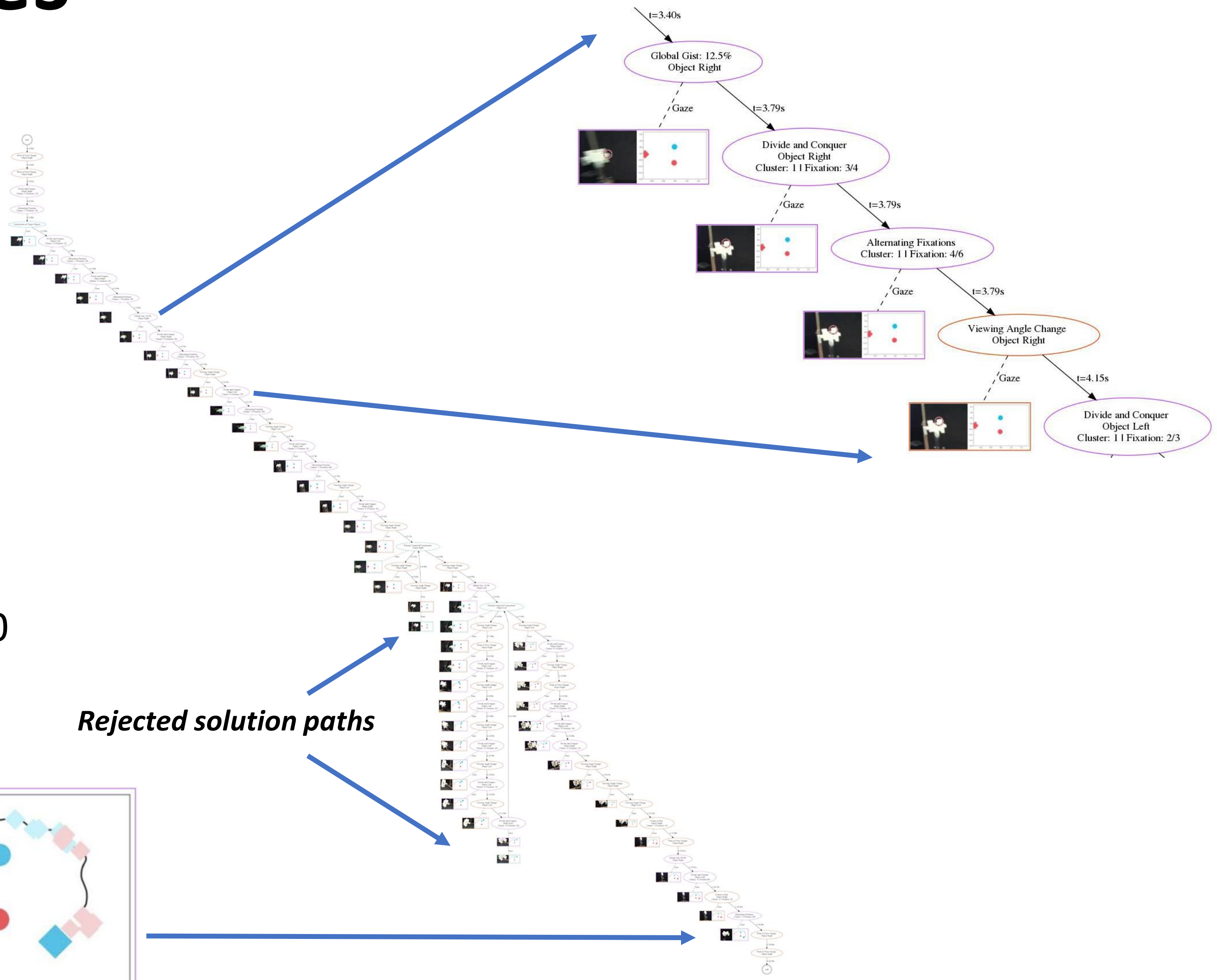


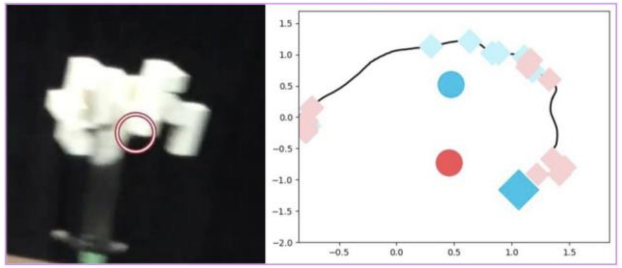


Action Sequences

- An example sequence of actions
  - Complexity: Easy
  - Orientation: 90°
  - Sameness: Same
  - Start: Long

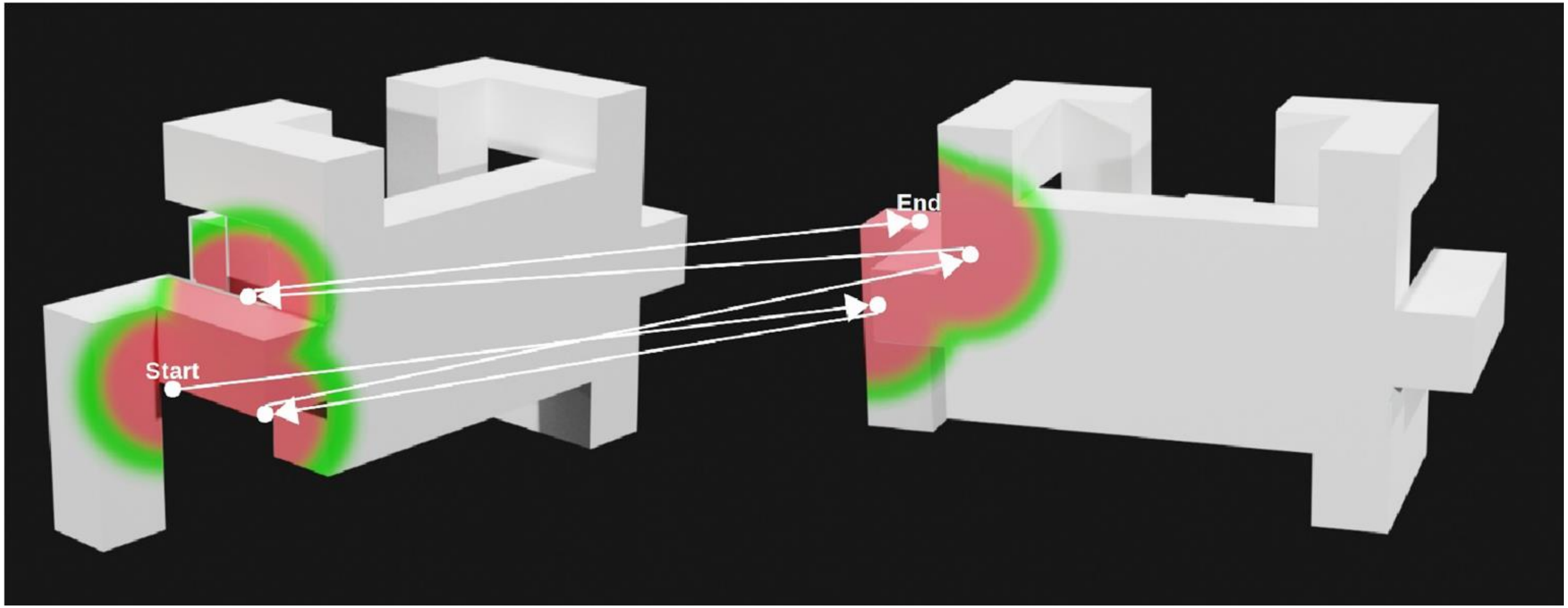
- No direct path to a solution
- Almost always several trials and error components
- Generally, complex graphs, with up to 800 nodes





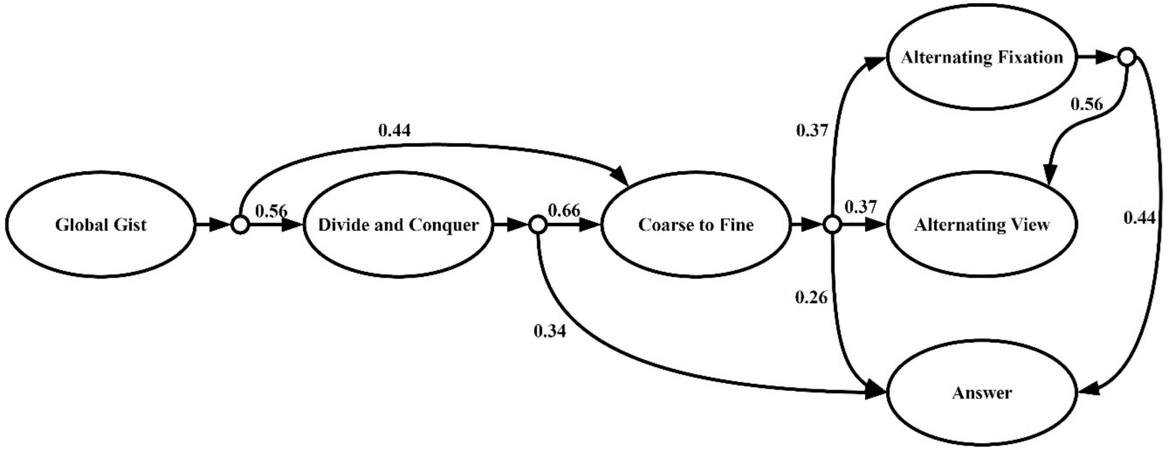
Patterns within Action Sequences (1)

- Found many patterns
  - Each has usage frequency dependent on complexity, orientation, and starting position
  - For example, subjects would move gaze back and forth between objects seemingly inspecting a single spatial region for similarity – we termed this the ***Alternating Fixation*** strategy



Patterns within Action Sequences (2)

- Over the course of a trial, subjects used several such strategies in sequence and these formed higher order patterns – they form directed graphs
- These higher order patterns were compositions of the strategies but with different frequency of occurrence depending on the experiment initial conditions – have found several to date
- They bear a remarkable similarity to the ***Cognitive Programs*** of Tsotsos & Kruijne (2014), to the Dynamic Bayes Nets of Ballard & Hayhoe (2009) and to the Visual Routines of Ullman (1984).



If target objects are the same, this Cognitive Program was used in 99.7% of trials.

Patterns within Action Sequences (3)

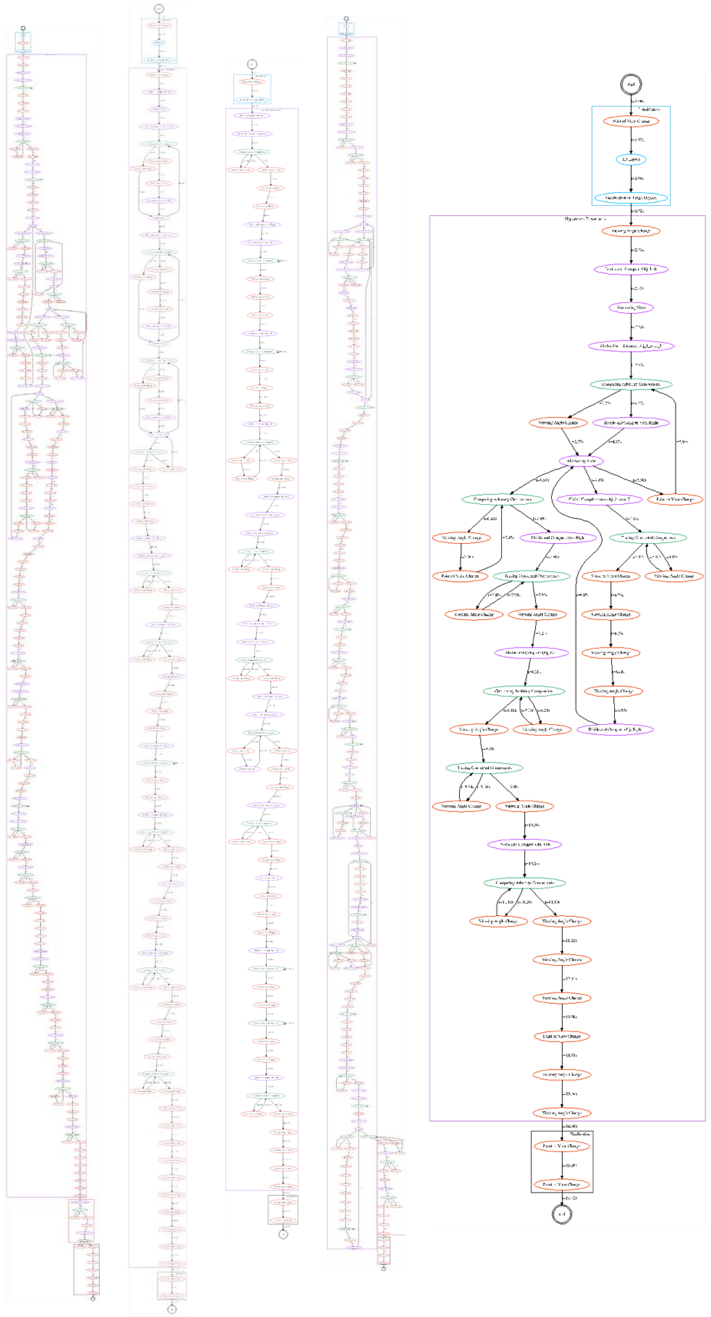
- These provide the answer to our motivating question

- Across each Cognitive Program, attention is seen to play one of several roles as specified by the attentional mechanisms of Selective Tuning:
  - Task Priming
  - Fixation change
  - Viewpoint change
  - Top-down surround suppression and localization
  - Selection

- An attentional/executive controller seems needed to choose, parameterize, sequence, initiate, monitor for success, plan remedial action, and more

Conclusions

- Motivated to understand the scope and nature of attentional/executive control for STAR for an active observer
- Lack of human experimental knowledge led us to do the experiment ourselves
- The Same-Different task for an active observer seems like an excellent testbed for systems that purport intelligent behavior
- To date, our subjects show a complex solution process that is dynamically deployed, highly accurate, composing known elements into a hypothesize and test framework until a solution achieved
- Human solutions seem to fill the criteria for Cognitive Program representations, an upgraded form of Ullman's Visual Routines
- Within these, attention has many roles and the dynamic nature of their application indicates what the nature of a controller might be, whose design is now underway
- Moving forward
    - We intend to also experiment with 3D ‘spatial relations’, 3D ‘visual search’, and to add shadowing to all the tasks, within the PESAO facility.
    - The goal is to discover the common elements of a generic visual problem-solving strategy. (Not to solve each separately.)



Active Observer Visual Problem-Solving Methods are Dynamically Hypothesized, Deployed and Tested

Markus D. Solbach

John K. Tsotsos

Department of Electrical Engineering and Computer Science

York University, Canada

November 18th, 2021

Tsotsos Lab
Active and Attentive Vision

ADVANCES in COGNITIVE SYSTEMS 2021
Ninth Annual Conference

YORK UNIVERSITÉ UNIVERSITY
Centre for Vision Research