Visual Perception and Reasoning: Integrating Cognitive Programs, Working Memory, Attention Control and Visual Processing (FA9550-18-1-0054)

PI: John K. Tsotsos (York University)

AFOSR Program Review:
Computational Cognition and Machine Intelligence Program
(October 6-8, 2020, Arlington, VA)

Visual Perception and Reasoning (Tsotsos)

Research Objectives:

- How do humans reason about visual stimuli?
- How do we solve tasks that require recognizing and relating different elements of our visual field in order to do everyday tasks like make a sandwich, solve a puzzle, drive a car, walk the streets, or describe a picture?
- We wish to learn how to build flexible, adaptive active artificial agents that perceive and behave

Technical Approach:

- Data-informed, scientific method, strategy – need to understand problem nature
- Roots in 'complexity level analysis' (Tsotsos 1990)
- Core elements include Selective Tuning Theory for Visual Attention and Ullman's Visual Routines
- Evaluation based on:
  1) replication of existing experimental results;
  2) new experiments based on model predictions;
  3) new spin-off applications.

Key Scientific Contributions:

The year's contributions include:

- Novel experimental facility for human studies of active observation in 3D visuospatial tasks
- First results on 'same-different' task show a complex range of human strategies
- Demonstrated how early salience-based selection is not part of the visual categorization solution in humans
- New theoretical arguments regarding the computational nature and necessity of an attentional control process

DoD Benefits:

- DoD seeks to develop techniques and systems that are robust, scalable, and capable of learning and acting with varying levels of autonomy, and as components of networked sensors, knowledge bases, autonomous agents, and human teams.
- Active control for dynamic adaptation in perception-behavior systems

List of Project Goals

There are 8 project threads:

	Thread
I.	Selective Tuning Attentive Reference model (STAR)
II.	Selective Tuning (ST) model of attention
III.	Visual Hierarchy
IV.	Cognitive Programs (CP)
V.	Task Compilation into CPs
VI.	Working Memory
VII.	Executive Control: Attention
VIII.	Executive Control: Task

Seek to test: General purpose intelligent systems are not constructed as a large collection of uni-taskers. Although we seem to be able to build any uni-tasker we can think of (chess, GO, Starcraft...), we still need to discover how to build one system that can do all of it. Our claim is that general purpose intelligence is due to a single system that is tuned and configured differently for each required situation, and performs differently for each situation, from near immediate and perfect responses to extremely slow and error-prone responses. The key to this approach is a deep understanding of the nature of intelligent behavior.

Progress Towards Goals (or New Goals)

2019 Presentation Highlights:		2020 Presentation Highlights:
Human-equivalent performance for fixation control I		
First demonstration of attentive priming in a CNN framework II, III		
New Border Ownership Theory III	➡	New Border Ownership Theory III
New Colour Hue Theory III	➡	New Colour Hue Theory III
Elephant in the Room III		
Several new discoveries from Selective Tuning attention model II	➡	Several new discoveries from Selective Tuning attention model II
Refutation of Early Selection Theory for Vision I	➡	Refutation of Early Selection Theory for Vision I ⬅
Detailed Cognitive Program abstraction levels and demonstrated qualitative performance against 4 classic attention experiments I		
New insights open novel approach to learning executive functions VII, VIII	➡	New insights from novel approach to learning executive functions VII, VIII
		Novel 3D Active Observer Experiments IV ⬅
		New look at computational foundations for attentional control VII ⬅

Tsotsos, J. K., Kotseruba, I., & Wloka, C. (2019). Rapid visual categorization is not guided by early salience-based selection. *PLoS One*, *14*(10), e0224306.

Tsotsos, J. K., Kotseruba, I., & Wloka, C. (2019). Correction: Rapid visual categorization is not guided by early salience-based selection, *PLoS One* 14 (12), e0226429-e0226429

But please use arXiv:1901.04908

Question : is early salience-based selection part of rapid categorization in vision ?
(Broadbent's 1958 Early Selection Theory)

1. In Bylinskii et al. 2015 we considered 142 saliency models. Since then there are at least as many more. The construct seems thoroughly studied by classical, information-theoretic and deep learning methods. Validation mostly against human fixation heat maps.

2. Test them on the Potter and Thorpe image sets since they form an important part of the foundation for modern computer vision (single feedforward pass suffices for categorization).

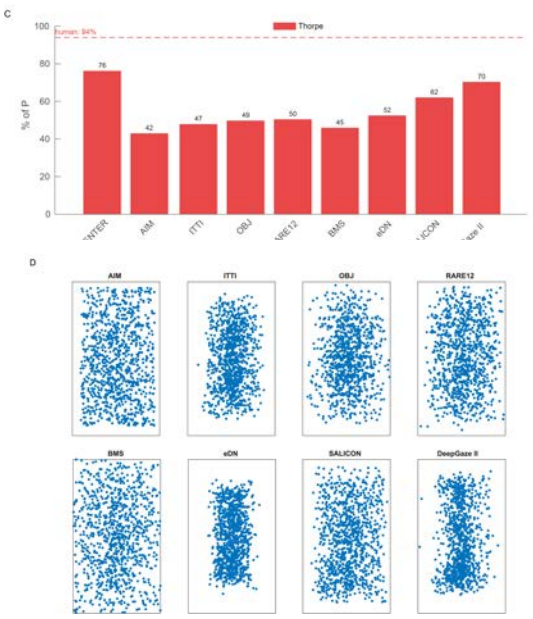
3. If early, data-driven, salience-based selection is present in human vision, do any of these methods deliver the correct fixation as a first selection? We chose a number of the best performing ones of all types.

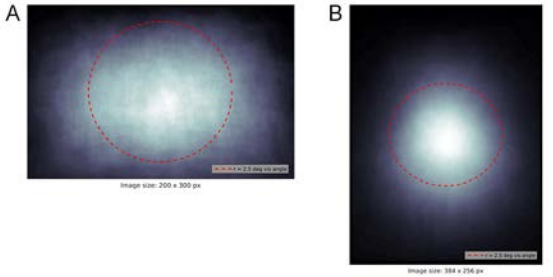
- If at least one algorithm succeeds in matching human performance, this supports the possibility of early salience-based selection in humans and its inspiration for machine vision methods.
- If not, where exactly are the algorithm fixations for these image sets? Is there anything about these images or these algorithms that lead to the failure?

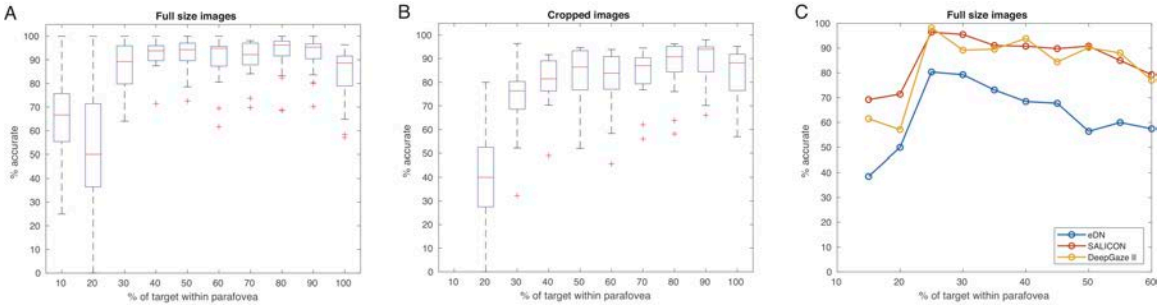
4. Do humans really need early selection for this task?

Summary of Results

1. Test of saliency algorithms on Potter/Thorpe images:
   i. no algorithm reaches human performance on first fixation
   ii. CENTER algorithm performs almost as well with significantly less cost
   iii. no algorithm provides assistance on target-absent images
   iv. both Potter and Thorpe image sets have center bias
   v. it could be that the right saliency algorithm is yet to be discovered; but we have hundreds of attempts so far
2. Human categorization with parafoveal images:
   i. human subjects perform almost as well given full or parafoveal images
   ii. algorithms require at least 30% of target within a central parafovea to perform well so system fixation must already be at the right location







Implications

- Early selection seems to not play a role in rapid categorization for humans; in fact, early selection would seem to be mis-leading rather than helpful
- Structure of original experiments implicitly biases results to centered targets and center gaze
  - *fast feedforward categorization applies only for centred targets*
- Modern computer vision models that employ early selection cannot justify this as mimicking human vision
  - *use of Thorpe's results to motivate CNN's is not valid*
- The role of data-driven saliency computation for humans may be limited to gaze selection (but that's important enough)

NOTE: Last year's progress report included

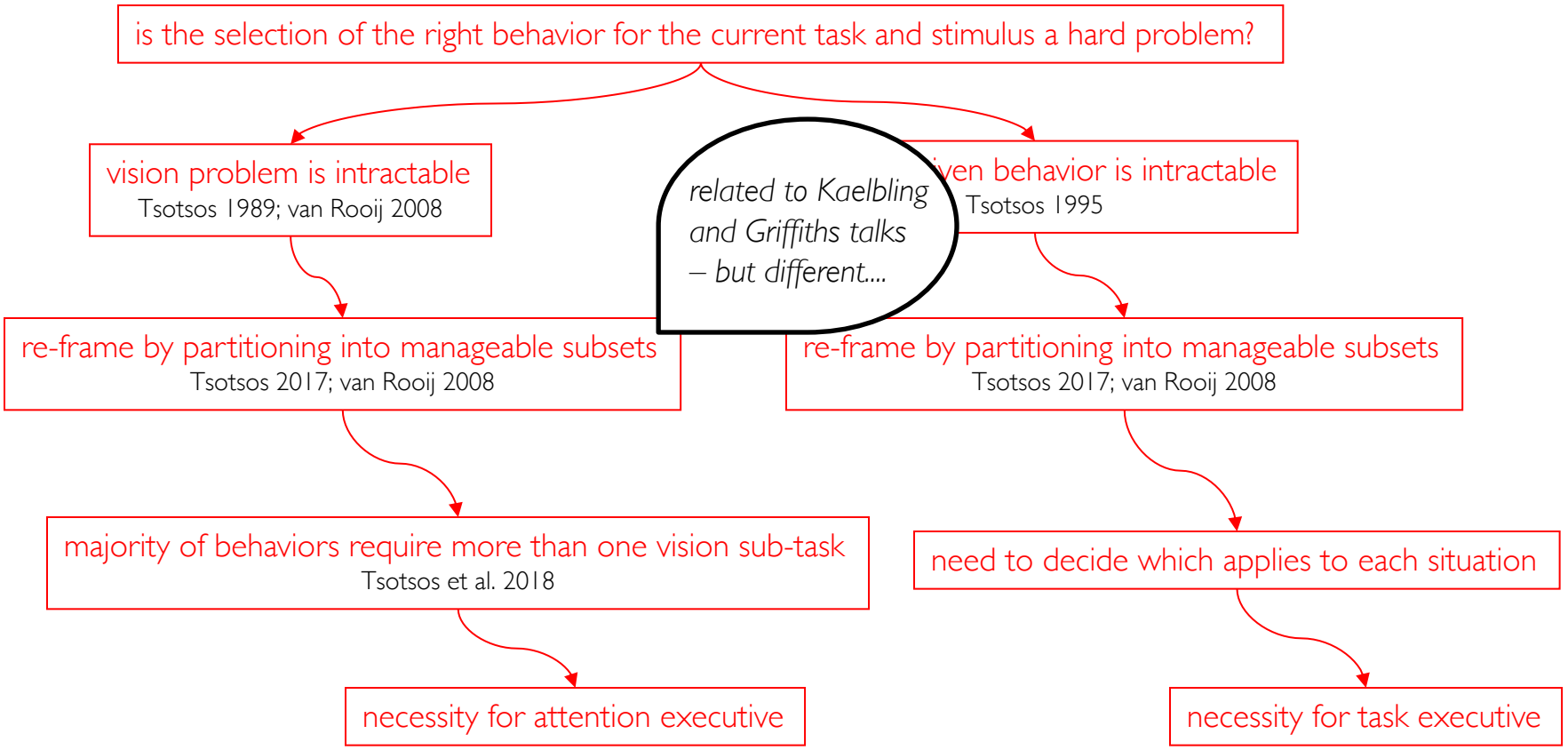
Wloka, C., Kotseruba, I, Tsotsos, J.K., Active Fixation Control to Predict Saccade Sequences, CVPR 2018.

Tsotsos, J.K., Abid, O., Kotseruba, I., Solbach, M. On the Control of Attentional Processes (submitted)

PART 1:

Humans excel at a seemingly endless variety of visuospatial tasks. How does that happen?

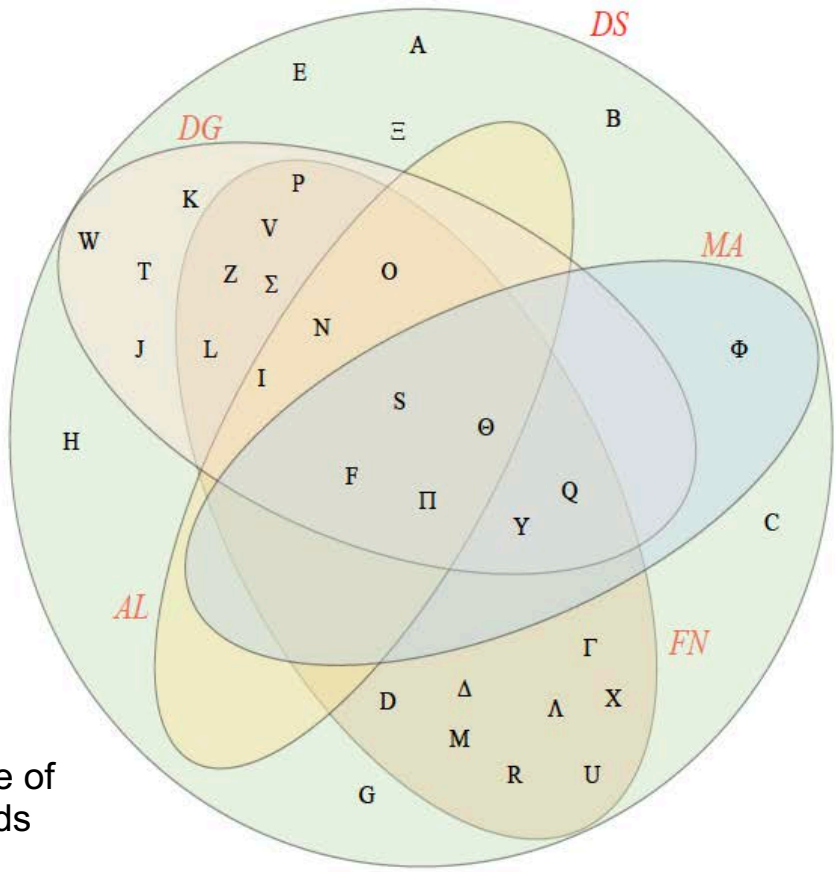
We take a 'first principles' computational approach to its understanding that is complementary to the previous data-driven or experimental approaches.



Representative Elements in an Executive Control Taxonomy

Model Classes:

DG - Directed Graph
DS - Descriptive
FN - Functional
MA - Mathematical
AL - Algorithmic



ID	Reference
A	Kahneman 1970
B	Allport 1993
C	Posner & Dehaene 1994
D	Singer & Gray 1995
E	Egeth & Yantis 1997
F	Dehaene, Kerszberg & Changeux 1998
G	Schall & Bichot 1998
H	Yantis 1998
I	Carpenter 2000
J	Hopfinger, Buonocore & Mangun 2000
K	Corbetta & Shulman 2002
L	Fuster 2002
M	Giesbrecht, Woldorff, Song & Mangun 2003
N	Roelfsema, Khayat & Spekreijse 2003
O	Roelfsema 2005
P	Cole, Schneider 2007
Q	Gold & Shadlen 2007
R	Rossi et al. 2009
S	Zylberberg, Slezak, Roelfsema, Dehaene & Sigman 2010
T	Baluch & Itti 2011
U	Niendam, Laird, Ray, Dean, Glahn & Carter 2012
V	Cole, Reynolds, Power, Repovs, Anticevic & Braver 2013
W	Gilbert & Li 2013
X	Miller & Buschman 2013
Y	Womelsdorf & Everling 2015
Z	Desrochers, Burk, Badre, & Sheinberg 2016
Γ	Mansouri, Egner, Buckley 2016
Δ	Van Rullen 2016
Λ	Hanks & Summerfield 2017
Θ	Laird, Lebiere & Rosenbloom 2017
Π	Abid 2018
Σ	Batista-Brito, Zagha, Ratliff & Vinck 2018
Φ	Tang & Bassett 2018
Ξ	Moyal & Edelman 2019

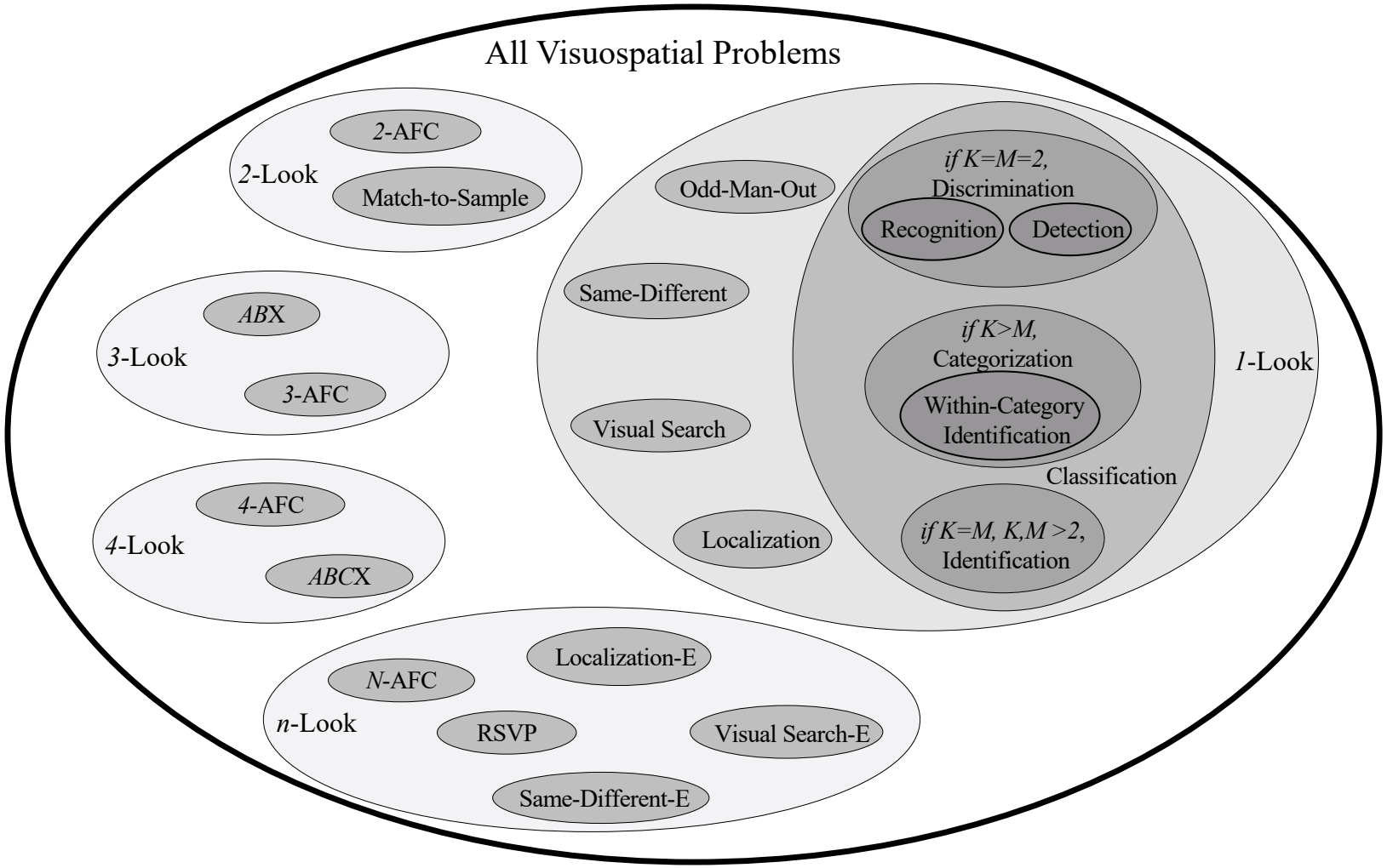
Key Conclusions:

1. attentional functions are of very many different kinds

2. the importance of flexible composition of elements to achieve solutions for dynamic task presentation

3. there is a decomposition of function that is productively used by top-down control

4. brain regions flexibly shift their functional connectivity patterns with multiple brain networks across a wide variety of tasks.



K is determined by the set of all possible images
M is determined by the set of object categories of interest
Macmillan , N. A. , & Creelman , C. D. (2005).
Detection theory: A user ' s guide . New York : Lawrence Erlbaum Associates .

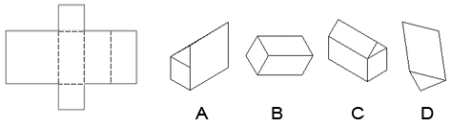
N-Look - N is the number of distinct test images

Human Visuospatial Abilities

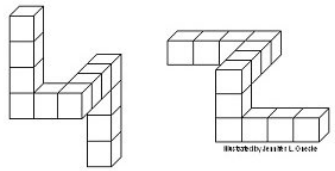
• Vision is more than categorization and attention is more than a peak in a saliency map

• Consider the full breadth of human visuospatial ability (Carroll 1993), for example:

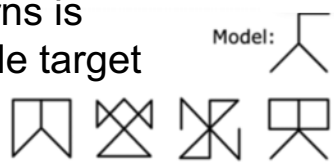
Spatial Visualization: processes of apprehending, encoding, and mentally manipulating spatial forms (paper folding or spatial relations).



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).



• Almost always studied with passive observation of a 2D display

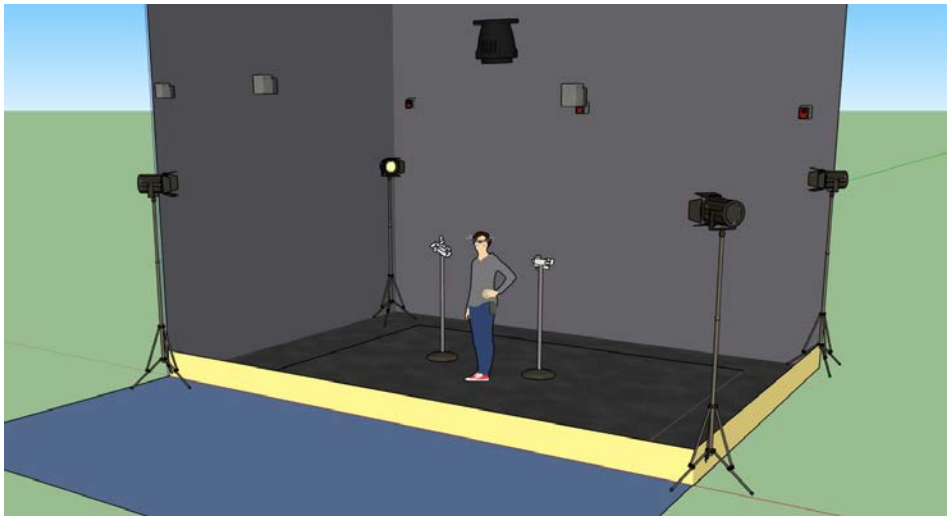
But humans are Active Observers in a 3D world

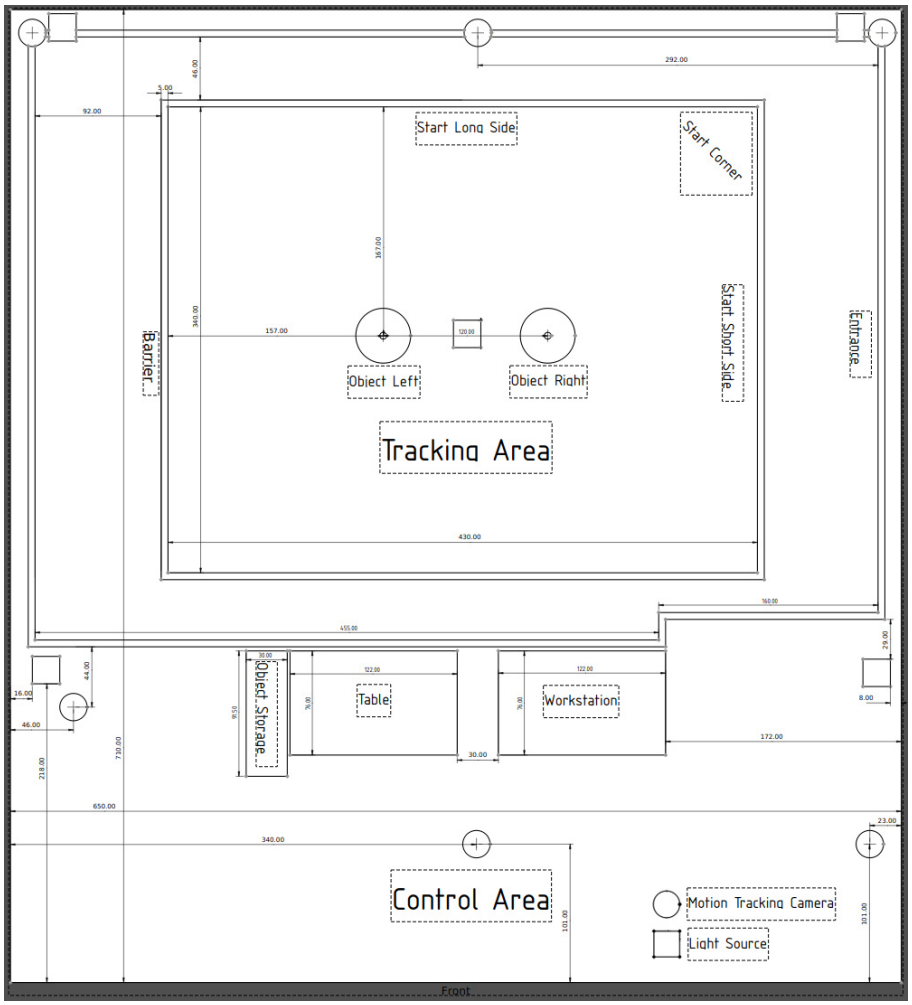
How to study this behavior?

Novel Active Vision Experimental Setup

Markus D. Solbach, John K. Tsotsos , PESAO: Psychophysical Experimental Setup for Active Observers , arXiv:2009.09933, Sept. 2020

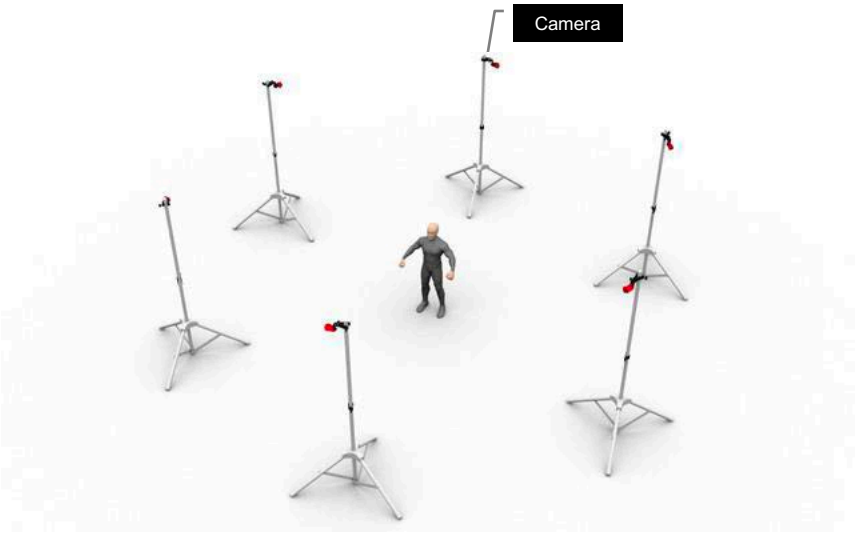
- 6 x 7.1m controlled environment
- Constrained 4.3 x 3.4m tracking area



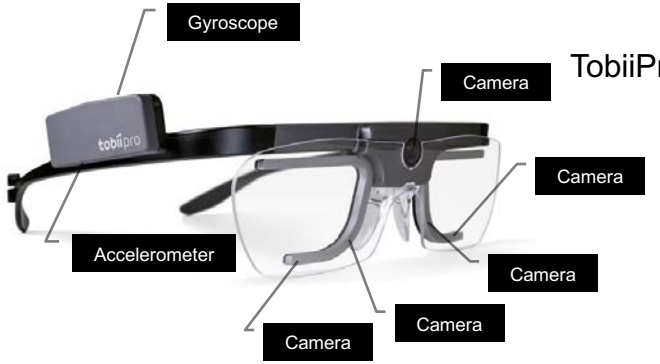


Need integrated, high precision, controlled gaze tracking, head tracking

PESAO Elements



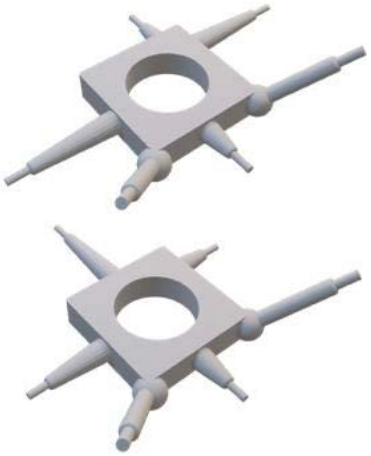
OptiTrack – 6 camera system



TobiiPro Glasses 2

Custom Tracking Attachment

Custom
Object Tracker



Five 660 LED Video light-panels from Neewer (3200 – 5600K and lumen of up to 7300 Lux/m.), one in each corner and one ceiling light.

Integrated software and interface:
http://data.nvision2.eecs.yorku.ca/PESAO/

error: eye gaze - ~2° @ 310cm
head position - <1.5mm

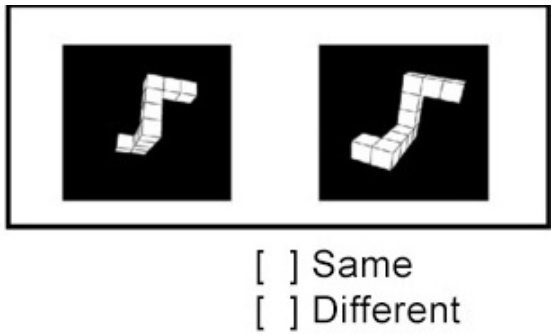
Hardware	Parameter	Value Range
Eye-Tracking	Tracking Frequency	<50 / <100 Hz
	IMU Frequency	<50 / <100 Hz
	Gaze Tracking	<50 / <100 Hz
	Eye movement type	<50 / <100 Hz
	First-Person Camera	<25 FPS
Motion-Tracking	Tracking Frequency	<120 Hz
Light	Light Intensity	<7300 Lux/m
	Colour Temperature	3200-5600 K
Camera	Frequency	<30 Hz

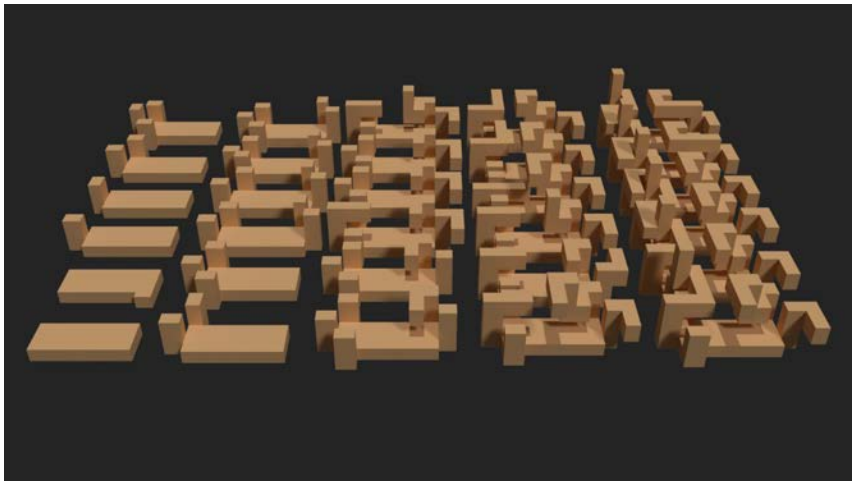
Table 1 PESAO Hardware Specifications.

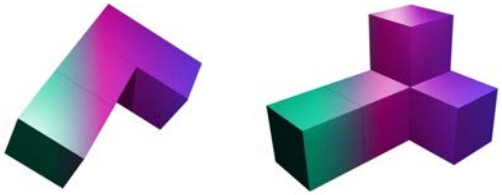
3D Same-Different Task w/Active Observer

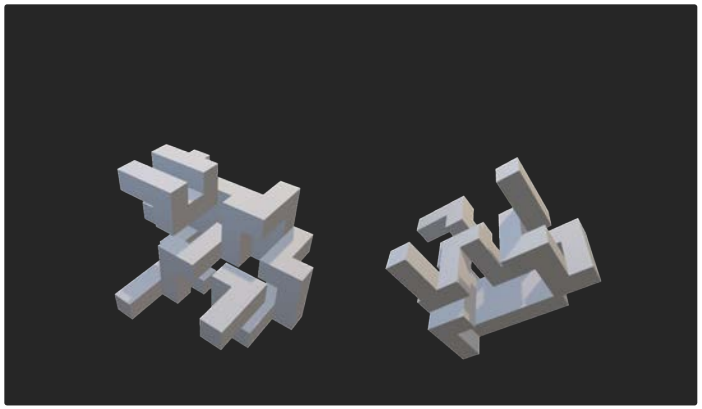
Real 3D Objects

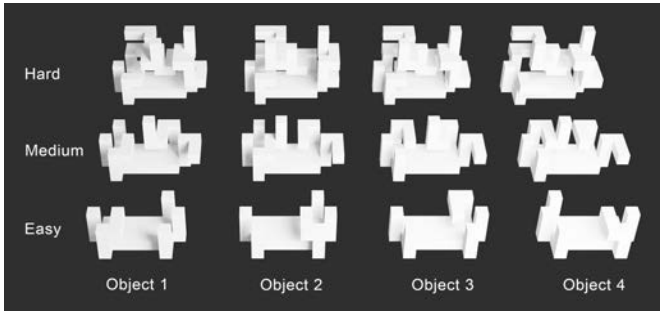
- Inspired by *Shepard and Metzler (1971)*
  - Rich psychophysical literature
- Different Complexity Levels
- Common Coordinate System
- Self-Occlusion







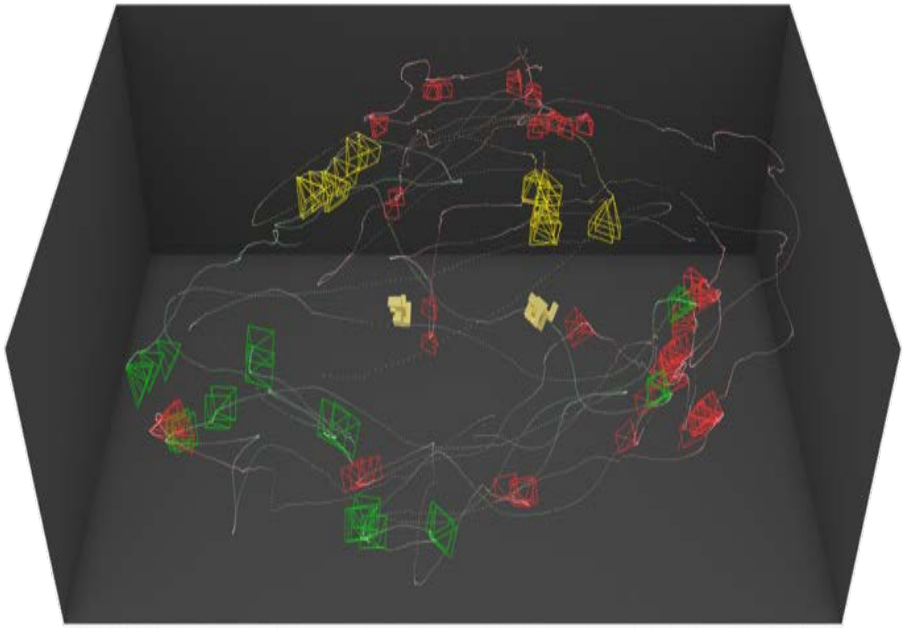




So Far:

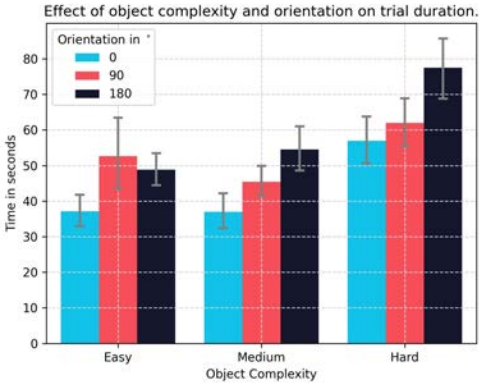
- 66 Subjects, ~23min/subj. , ~11M tracking points (Head Pose alone), 25h of video
- Configuration space sampled 22 times, 1188 trials to date

  (3 complexity levels x 2 possible responses x 3 rotations x 3 starting positions = 54)

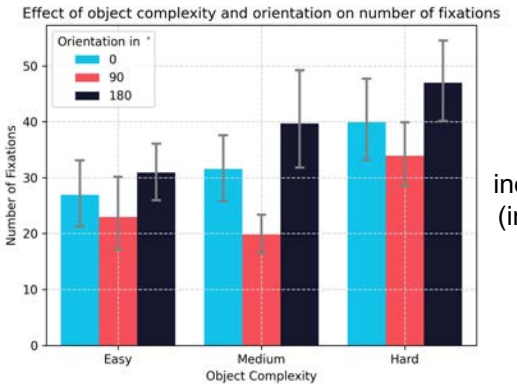


Some Preliminary Results

- **Questionnaire & Recorded Data**
  Strategy depends on Object complexity, Object orientation, Initial viewpoint

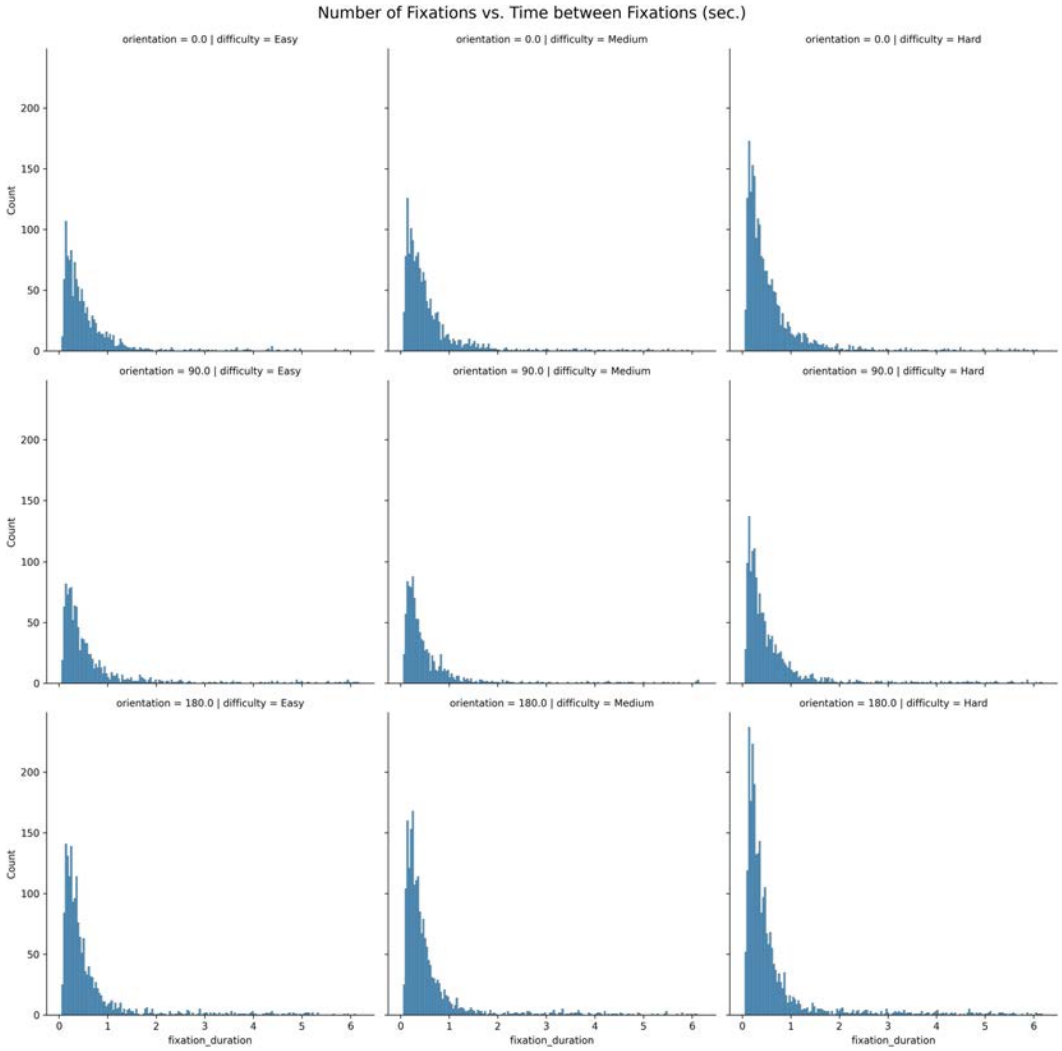


increasing
target complexity
⇩
increasing response
time

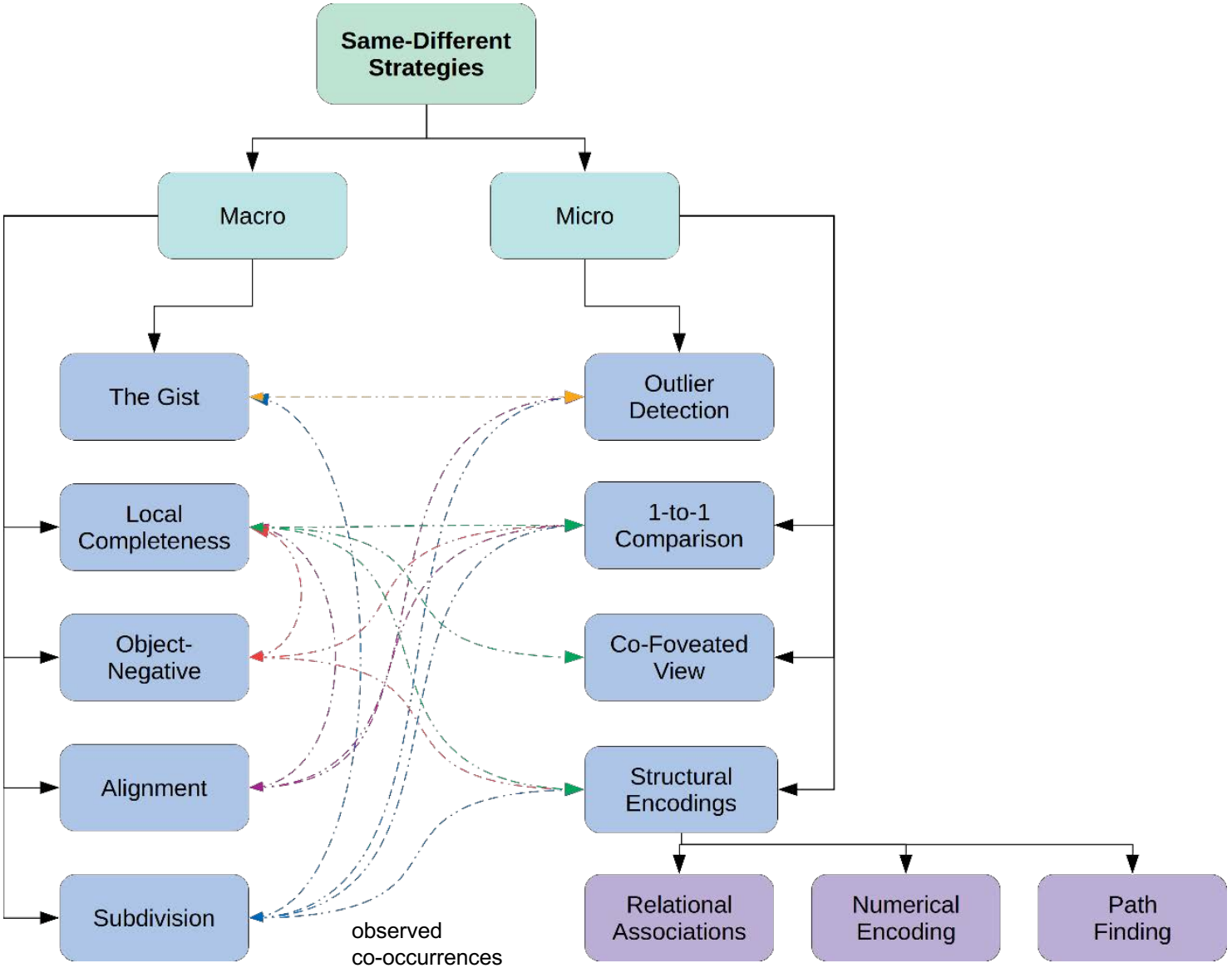


increasing
target complexity
⇩
increasing # fixations
(interesting 90° data)

many fixations with over
half-second dwell time



More Preliminary Results



Why Do these Experiments Matter?

Deep Learning and Active Perception are inherently incompatible

"Active vision breaks deep learning" - Yann LeCun, AFOSR Workshop on the Future of Machine Learning, Arlington, VA, May 17-19, 2017

Active Perception adds an inductive component to data acquisition.
It is essentially the way to complete inductive reasoning actions.

Inductive reasoning takes specific information (premises) and makes a broader generalization (conclusion) that is considered probable. The only way to know is to test the conclusion; a passive sensing strategy could only do this by accident.

Passive sensing thus impedes the use of any form of inductive reasoning
(e.g., inductive generalization, Bayesian inference, analogical reasoning, prediction)

Our experiments attempt to discover exactly what humans hypothesize while performing a complex visuo-spatial task and how they go about testing their hypotheses

Why Do these Experiments Matter?....cont'd

- There are limits to

> "This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves."
>
> - C. Anderson,The End of Theory: Will the Data Deluge Makes the Scientific Method Obsolete? (WIRED 2008)

Our experiments tests this. The domain space is far too large to think it can be sensibly sampled for training sets.

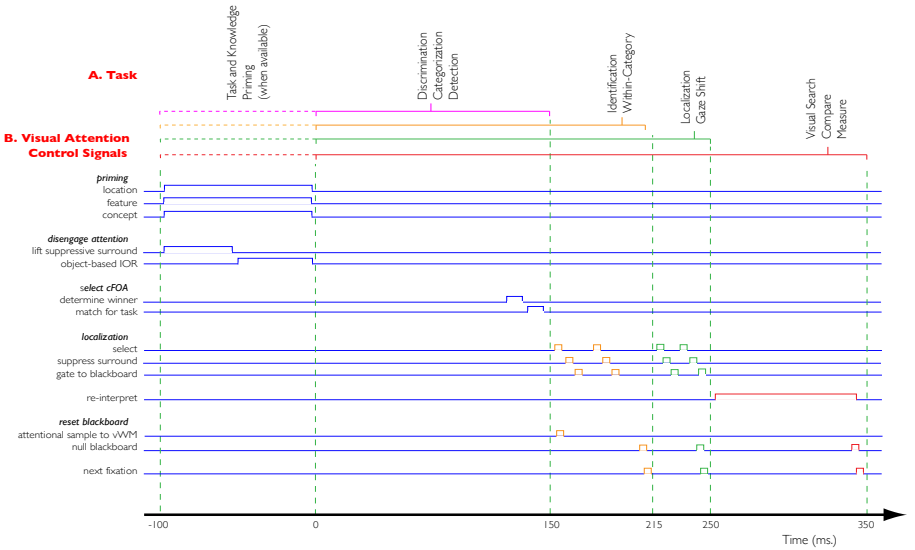
- Provide insight into different strategies (cognitive programs)
- Appears that subjects decide based on individual cases – they choose simple strategies for simple cases and more complex ones for difficult cases, just as the original reason for problem re-framing would require
- This supports our main hypothesis as the strategy for creating tractable solutions to difficult tasks

Tsotsos, J.K., Abid, O., Kotseruba, I., Solbach, M. On the Control of Attentional Processes (submitted)

PART 2:

Start with a Cognitive Program
For each visual sub-program

A1 - initiate and terminate program



A2 - monitor program progress
(objective function gives target)

A3 - modulate errant program

A4 - re-start failures (re-plan)

A5 - terminate and move to next program

The fovea, and specifically its foveola, has the highest density of cones, it should always be placed at the location of maximum interest. Let the gaze position be *(x, y)* in retinal coordinates and the centroid of target *S* be (x_S, y_S). The closer the point of gaze - or the center of the foveola - is to the target centroid, the more retinal cones will fall within the target, so a controller will seek to minimize the distance between these two points with the following objective,

$$\min_{(x,y)} \|(x_S, y_S) - (x, y)\|$$

The control signal for the fixation mechanism would be $(\delta x, \delta y)$, the change from the current (x, y).

Self-Assessment

- **Seek to test:** General purpose intelligent systems are not constructed as a large collection of uni-taskers. Although we seem to be able to build any uni-tasker we can think of (chess, GO, Starcraft...), we still need to discover how to build one system that can do all of it. Our claim is that general purpose intelligence is due to a single system that is tuned and configured differently for each required situation, and performs differently for each situation, from near immediate and perfect responses to extremely slow and error-prone responses. The key to this approach is a deep understanding of the nature of intelligent behavior.

- Our focus is on vision and limited visual behavior
- Latest experimental literature reaches similar conclusion
- Major progress on understanding nature of vision (saliency, active 3D observers, attention mechanisms, fixation control, visual hierarchy details)
- Quantitative arguments for why attention and task executive control are necessary
- Evidence for role and character of Cognitive Programs accumulating
- New learning strategy being incubated: *Learning by Composition and Exploration*

List of Publications, Awards, Honors, etc. Attributed to the Grant

Awards

1st Prize: Best Computer Vision Poster, International Conference on Predictive Vision, Toronto, June 10 - 13, 2019
Wloka, C., Kunic, T., Kotseruba, I, Tsotsos, J.K., SMILER: An Easy and Consistent Way to Compute Saliency Maps

Calden Wloka, Best Doctoral Dissertation Award, 2019, CIPPRS (Canadian Image Processing and Pattern Recognition Society)

Full Papers in Refereed Journals

Mehrani, P, Mouraviev, A., Tsotsos, J.K. (2020). Multiplicative modulations enhance diversity of hue-selective cells, *Scientific Reports* 10 (1), 1-15.

Tsotsos, J. K., Kotseruba, I., & Wloka, C. (2019). Rapid visual categorization is not guided by early salience-based selection. *PloS one*, *14*(10), e0224306.

Yoo, S. A., Tsotsos, J. K., & Fallah, M. (2019). Feed-forward visual processing suffices for coarse localization but fine-grained localization in an attention-demanding context needs feedback processing. *PloS one*, *14*(9), e0223166.

Tsotsos, J.K., Attention: The Messy Reality, *Yale Journal of Biology and Medicine*, Special Issue on Attention, March 2019

Rasouli, A., Tsotsos, J.K. (2019). Autonomous vehicles that interact with pedestrians: A survey of theory and practice, *IEEE Trans. on Intelligent Transportation Systems, p1* -19 , 15 March, DOI: 10.1109/TITS.2019.2901817

Rosenfeld, A., Tsotsos, J.K.. (2020). Incremental Learning Through Deep Adaptation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1-13, doi: 10.1109/TPAMI.2018.2884462; March, Vol 42, No3, p651-663.

Full Papers in Refereed Conference Proceedings

Kotseruba, I., Rasouli, A., Tsotsos, J.K. (2020). Do They Want to Cross? Understanding Pedestrian Intention for Behavior Prediction, IEEE Intelligent Vehicles Symposium, June 23-26, Las Vegas, NV, USA

Tsotsos, J., Kotseruba, I., Andreopoulos, A., & Wu, Y. (2019). Why Does Data-Driven Beat Theory-Driven Computer Vision?. In Proc. of the IEEE International Conference on Computer Vision Workshops, Oct. 28, 2019, Seoul, Korea.

Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J.K.. (2019). PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction, Int. Conf. on Computer Vision,, Oct. 29 - Nov 1, Seoul, Korea.

Kotseruba, I., Wloka, C., Rasouli, A., Tsotsos, J.K., (2019). Do Saliency Models Detect Odd-One-Out Targets? New Datasets and Evaluations, British Machine Vision Conference.

Raosuli, A., Kotseruba, I., Tsotsos, J.K. (2019). Pedestrian Action Anticipation using Contextual Feature Fusion in Stacked RNNs, British Machine Vision Conference.

Rosenfeld, A., Tsotsos, J.K., (2019). Intriguing Properties of Randomly Weighted Networks: Generalizing While Learning Next to Nothing, Canadian Conference on Computer and Robot Vision.