scientific reports

Check for updates

OPEN

The psychophysics of human three-dimensional active visuospatial problem-solving

Markus D. Solbach✉ & John K. Tsotsos

Our understanding of how visual systems detect, analyze and interpret visual stimuli has advanced greatly. However, the visual systems of all animals do much more; they enable visual behaviours. How well the visual system performs while interacting with the visual environment and how vision is used in the real world is far from fully understood, especially in humans. It has been suggested that comparison is the most primitive of psychophysical tasks. Thus, as a probe into these active visual behaviours, we use a same-different task: Are two physical 3D objects visually the same? This task is a fundamental cognitive ability. We pose this question to human subjects who are free to move about and examine two real objects in a physical 3D space. The experimental design is such that all behaviours are directed to viewpoint change. Without any training, our participants achieved a mean accuracy of 93.82%. No learning effect was observed on accuracy after many trials, but some effect was seen for response time, number of fixations and extent of head movement. Our probe task, even though easily executed at high-performance levels, uncovered a surprising variety of complex strategies for viewpoint control, suggesting that solutions were developed dynamically and deployed in a seemingly directed hypothesize-and-test manner tailored to the specific task. Subjects need not acquire task-specific knowledge; instead, they formulate effective solutions right from the outset, and as they engage in a series of attempts, those solutions progressively refine, becoming more efficient without compromising accuracy.

Human visual ability feels so effortless that it is literally taken for granted. However, any intuitive introspection into this ability rarely reveals its profound nature. On the contrary, the tendency has been to prefer simpler descriptions in an Occam's Razor (or parsimony) sense, and although these have helped move our understanding along, they are no longer as useful. A main focus has been on how well the eyes and visual system can detect a target stimulus, i.e. Visual Function[1]. This has been studied extensively over many decades, in several literatures (in computer vision, in ophthalmology, in visual psychophysics). The visual systems of all animals do much more than simply detect stimuli; they also enable locomotion, seek food, detect threats, and more. In humans, there is much more than supporting basic survival. We use our visual abilities to create, exploit, admire, destroy, and manipulate our 3D physical world. How well we perform while interacting with the visual environment and how vision is used in everyday activities has been termed Functional Vision[1]. In contrast with visual function, this has not been as well examined, in part due to its inherent experimental difficulty and to the fact that it seems far more complex an activity to define. Research of this kind has recently become possible in humans[2–5] and a variety of animals[6–8]. We focus on how viewpoints might be determined during a purely visual task and what problem-solving strategies those sequences of viewpoints might reveal.

Although it seems easy to enumerate the many kinds of behaviours humans perform in their visual world, it is far less easy to know how to best probe the nature of such behaviours in a general sense. A classic example is seen in the well-known experiments of Sheppard and Metzler in 1971[9] where they showed subjects, seated in front of a video monitor, 2D projections of unknown 3D geometric objects (Fig. 1A). Subjects were instructed to determine if the two objects were the same or different. Note how self-occlusion is a natural characteristic of their objects. Their results showed that subjects seemed to mentally rotate objects in order to test potential geometric correspondence. This work was seminal in this area but seemed difficult to generalize if viewing real 3D objects in an unconstrained manner. Nevertheless, their experiment provided the inspiration for many important investigations as well as for what we present here. In a more foundational manner, the act of comparison has been suggested as the most primitive psychophysical task[10]; efforts to discover the deep nature of real-world visual comparison behaviour may have an impact on the full spectrum of human visual behaviour.

Department of Electrical Engineering and Computer Science, York University, Toronto, ON M3J 1P3, Canada. ✉email: solbach@yorku.ca

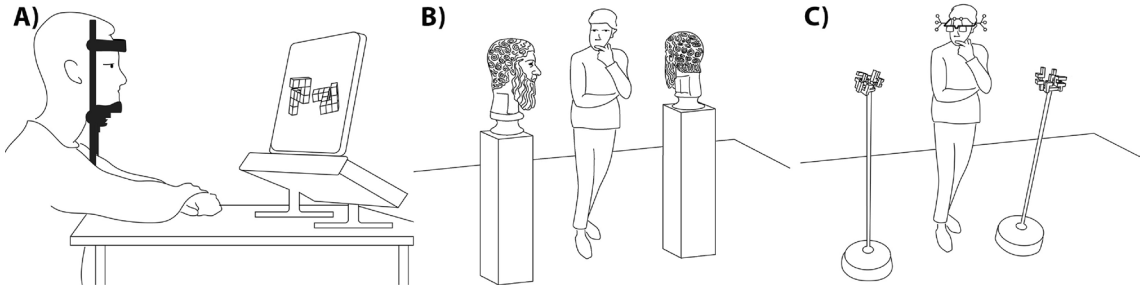

Figure 1. Our experiment as it evolved, beginning with Sheppard and Metzler's inspiration, the reality of visuospatial problem-solving "in the wild", and our actual setup. (**A**) The study of human vision often involves a two-dimensional probe, i.e., visual function. As illustrated here, the subject sits at a desk with head stabilized in a forehead and chin mount while performing an experiment involving images on a screen. (**B**) Real-world visuospatial problem-solving, i.e. functional vision, however, is distinctively different. Here we show someone in a sculpture gallery wondering if these two marble heads are the same. Humans do not passively receive stimuli; rather, they choose what to look at and how. We move our head and body in a three-dimensional world and view objects from directions and positions that are most suited to our viewing purpose. It should be clear that answering our sculpture query here might require more than one glance. It is we who decide what to look at, not an experimenter. (**C**) Our experimental setup is as shown. A subject wears a special, wireless headset and is shown two objects mounted on posts at certain 3D orientations. The subject is asked to determine, without touching, if the two objects are the same (in all aspects) or different and is completely free and untethered to move anywhere they choose. Gaze and view are precisely recorded (see Methods and Materials).

Our interests have foundation in the basic questions of active vision as laid out by Gilchrist and Findlay[11]:

(a) How is the decision made when to terminate one fixation and move the gaze?
(b) How is the decision made where to direct the gaze in order to take the next sample?
(c) What information is taken in during a fixation?
(d) How is the information from one fixation integrated with that from the previous and subsequent fixations?

It is important to emphasize that the mere pattern of fixations is not our only interest; we seek to discover whether there might be causal connections between successive fixations. What could the next gaze attempt to improve from what was learned with the previous? Before one begins to answer these, one needs data to enable an understanding of human fixation sequences in a broad sense. Experiments that address a purely visual task and cover large portions of the eye and head movement space are not common, even though the overall literature on active fixation is large.

In everyday human behaviour, and when solving 3D physical visuospatial tasks (e.g., filling a dishwasher, re-arranging furniture, searching for a specific tool, etc.), we typically move our eyes as well as our head and body to acquire visual information needed for a solution. Specifically, our head may exhibit 3 kinds of movement: dorsal or ventral flexion, tipping the head forward or backward about 60° in each direction; about 40° in left and right lateral bending or pivoting; and, left or right rotation, or panning, of about 70° either way. These motions can also be described as combinations of movements along six dimensions of a head-centred coordinate system—translation in x, y and z (elevation) coordinates (partially due to head and partially due to body movement) and rotations ϕ (pitch—flexion), θ (roll—bending), ψ (yaw—panning). Within the head, the eyes also move: abduction-adduction of about 40° (pitch - looking away or towards the nose); supraduction-infraduction of about 40° (yaw—looking up or down); and, a few degrees of dextrocycloduction-levocycloduction (roll around the eye's optical axis towards or away from the nose)—there are several general sources[11–14]. The movements of the two eyes are usually executed in a coupled manner so that there is a single fixation location in the 3D world.

Each of these dimensions of movement has limits to be sure, some defined by the context of the task or location while others are defined by the range of possible movements allowed by our physical bodies (as suggested above), as well as limits on how they might be performed in combination. The exact extent of all of these motions is documented in many sources (given earlier) and under many conditions. The point here is to highlight that these motions are part of normal visual behaviour and to show the space of behaviours that we are interested in probing.

Consider this space of possible eye-head movement (the nine degrees of freedom listed above) against the kinds of experimental setups common in vision and experiments related to ours. Experiments where subjects are seated, have head restraints and are instructed to maintain fixation on a point within a fixed distance display permit no movement of the kind described above. If subjects are seated, have head restraints, but may freely fixate to stimuli on display or respond by eye movements, a small range of movements may be investigated (e.g., Posner classic cueing experiments or classic visual search). Remove the head restraint, and the space of movement increases (e.g., in typical reading experiments). Using 3D objects or scenes projected onto a 2D display does not change this (as in[15], or block copying[16]). Moving away from 2D displays and using physical 3D scenes permits a greater range of movement even if seated because the eyes need not fixate on a single depth plane (physical block copying[17], or driving[18]). Allow the subject to move their body (e.g., making a cup of tea[19] or making a sandwich[2]) and the range along the variables described above increases to its fullest.

Past work with physical 3D objects/scenes with goals most strongly related to ours includes making a cup of tea[19], making a sandwich[2], physical block copying[17] and several studies involving locomotion[3–5,20]. All of these studies include the goal of uncovering what strategies subjects employ in performing non-trivial visuospatial tasks, as in our work. However, all of them also deal with tasks that inherently involve hand, body or foot movements as components of the task beyond simply a final response action, and these are of central interest to those experimenters. To be sure, important dimensions of active vision and eye-body coordination are uncovered: the role of binocular impairment for walking[3,20]; the nature of optic flow computations during locomotion[4,5]; extensions to image saliency methods[21,22]; and, the impact of gaze allocation on change blindness performance[23], just to give a few examples. In the most recent of these the main concern is how locomotion affects retinal input patterns[5]. They show how these patterns are shaped by gaze and behaviour and the relationship to motion sensitivity and receptive field properties as they vary across the visual field. The direction of travel determines the sequence of fixations, as they note.

There are at least two other variations on this general scenario. Some experimenters explore active vision by determining the viewpoint and scene in advance that a subject sees while the subjects remain stationary (e.g.,[24]). We wish to probe how the subject determines the sequence of gazes to acquire and acts to achieve them. Moving visuospatial experiments to virtual reality (as in[22,23]) can provide access to setups that are difficult or hazardous to do in-person; however, in our case, our setup is neither difficult nor hazardous. It simply needed some novel engineering to marry existing technologies[25]. VR, moreover, is, by definition, a simulation, and all simulations necessarily include abstractions away from reality. Our brains can fill in those gaps or ignore them, as we do often when watching movies, so these seem, to the viewer, to not matter. However, many researchers have found that they do matter, particularly when it comes to impairments or deficits due to length of viewing and thus duration of experimental sessions[26–28].

Our motivating question is how do humans decide if two objects are the same or different (as Sheppard and Metzler asked) within a physical 3D environment where they are free to explore those objects? Where do they look? How do they look? How do they move? Such questions are central to this study. Figure 1B shows an example of one common such setting within a sculpture gallery: are these two sculptures the same? This may be easy, or it may be difficult to determine. What is clear is that a single image of the objects generally will not suffice. Suppose we consider 3D versions of Sheppard and Metzler's unknown geometric objects. We could construct a range of such objects of varying physical complexity and show them to subjects. They would be free to examine without touching them and free to move around the objects, just like in the sculpture gallery. This requires a means to precisely record exactly what subjects are looking at and from what viewing position while deciding the answer to the simple question: are two objects the same or different? Figure 1C gives a sketch of our experiment showing a human subject wearing the recording device.

With this question in mind, we have designed a novel set of stimuli with known geometric complexity. We also employ a novel experimental setup that allows for precise, synchronized tracking of head motion with 6 degrees of freedom and eye gaze while subjects are entirely untethered to allow natural task execution. The tracking is accomplished with purely visual data without the need for other on-person devices that require frequent calibrations.

We have collected detailed data on humans performing a visuospatial task in hundreds of experiments and present an in-depth analysis with respect to various cumulative performance metrics such as the number of fixations, response time, amount of movement, and learning effect. We are not restricted to summary statistics of single actions such as these, but we also examine sequences of actions that reveal a fascinating array of problem-solving strategies employed differently by each subject and for each test case.

We also challenge the current intuitive views that perhaps such a visual ability can be captured by a large dataset of trials that feed a sophisticated learning method or that clever use of visual saliency maps might be the key, or that novel policies for a reinforcement learning approach might suffice. Our expectation is that this, and related non-trivial tasks, need something different. Our experimental goal is to reveal what human solutions to such 3D visuospatial tasks actually involve.

To analyze the effect of stimulus complexity, our set of objects consists of 12 objects divided into three complexity levels each of which can be presented at an arbitrary 3D pose. How each object appears (the image it projects to the subject) changes, sometimes radically, with change in viewing position. As a result, we include not only the object poses as experimental variables but also the subjects' initial viewing position. To keep the experiment tractable we investigate three object pose orientations and three viewer starting positions. Every subject performed 18 trials with randomly selected experimental configurations to test if a learning effect exists.

Surprisingly, humans have virtually no difficulty with this task, even for hard cases. The accuracy ranged from 80-100% across all configurations. Much data acquisition occurs with a minimum of 6 and a maximum of 800 eye fixations. Interestingly, no statistical change was observed in accuracy throughout the trials. However, a learning effect was seen for the number of fixations on the objects, response time, and head movement. The sequence of actions we observed strongly suggests that human problem-solving strategies are dynamically determined and deployed in a seemingly directed hypothesize-and-test manner tailored to the particular task instance at hand. Subjects do not need to learn the task; they develop good solutions from the start, and over the set of trials, those solutions become smoother or more efficient while maintaining accuracy.

Materials and methods

Stimuli and task

The task, including the stimuli, is illustrated in Fig. 1C and is designed to be a two-alternative forced choice. Subjects were allowed to move within a constrained area of about 3.4 m by 4.3 m, presented with two static three-dimensional stimuli mounted on acrylic posts. The task was to determine whether the two stimuli were

the same or different. Sameness in our experiment is defined as geometric congruence—all stimuli share the same colour and surface texture.

The stimuli are part of a three-dimensional physical objects set called *TEOS*[29]. The objects are inspired by the stimuli of Sheppard and Metzler. *TEOS* objects are all three-dimensional and have a known geometric complexity. Furthermore, a shared common-coordinate system allows quantifying the orientational difference of two objects. An illustration of the objects is shown in Fig. 2A. The set contains twelve objects split equally into three different complexity levels, which is defined by the number of blocks used to build an object. The level of object complexity C will be indicated by the subscript, such as C_e, C_m, and C_h for easy, medium, and hard, respectively. *TEOS* objects are 3D printable and templates are available. The objects are roughly $12cm \times 14cm \times 18cm$ in size. Movements of the subject were not restricted, and no time constraints were given. However, a definite answer (same, different, i.e. 2AFC) must be given to end the trial. Each subject performed 18 trials evenly split among complexity levels.

We also investigated different starting positions as they determine the initial observation of the objects. Figure 3A illustrates a top view of the experimental space marking the three positions from which a trial can start: equidistant from both objects (P_l), in line with both objects (P_s) and oblique to both (P_c).

Furthermore, we looked at the effect of object orientation difference. We limited the large space of possibilities to three values of orientation difference, 0°, 90°, and 180°. For all trials, we have used the same three poses as illustrated in Fig. 3B. Furthermore, all experimental variables were selected randomly for each trial, including sameness, complexity, starting position, and object orientation.

The experimental environment conditions are designed to limit the number of visual features that could potentially distract the participants: dark flooring, blackout curtains and directed lights pointing at the stimuli. The participant starts facing away from the stimuli (facing the curtain) and awaits a starting signal to turn around and start the trial. Once the subject is ready to answer, the subject is instructed to face away from the objects and provide their response. Before the experiment, all subjects are briefed with a written script to better control the amount of prior knowledge about the task. Further information about the conditions is presented in[30].

Lastly, after the experiment, subjects completed a questionnaire about their approach to solving this task (e.g. "what was your strategy for approaching the task", "did you notice any changes in your approach throughout the trials", "which instances were more challenging than others and why").

Data acquisition and analysis

We created a novel active vision experimental facility (named PESAO—Psychophysical Experimental Setup for Active Observers[25]) which formed the basis for all data acquisition and analysis. Its primary components are:

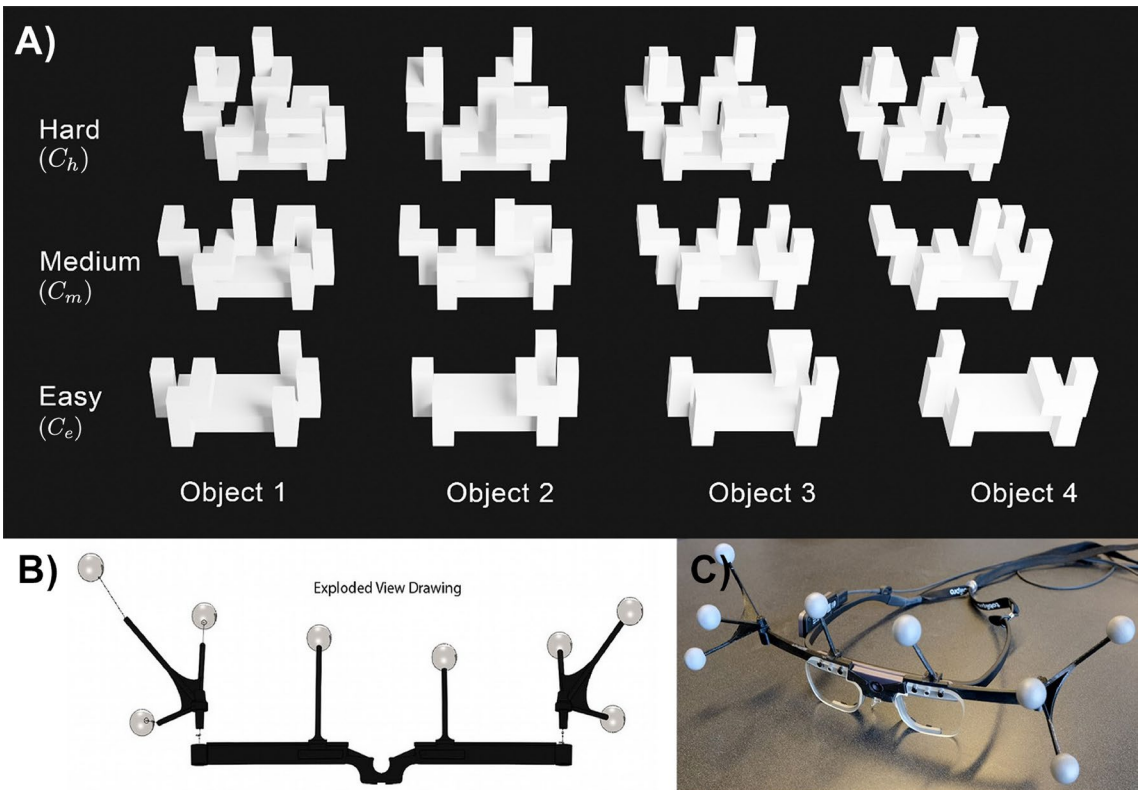

Figure 2. (**A**) Illustration of *TEOS* objects used as stimuli. The set is split into three different complexity levels. Complexity is defined as the number of blocks used to build an object. (**B**) Expanded view drawing of the custom clip-on tracking equipment. It uses 8 rotationally variant positioned tracking markers to avoid ambiguities. (**C**) Photograph of the assembled eye tracking glasses with tracking equipment.

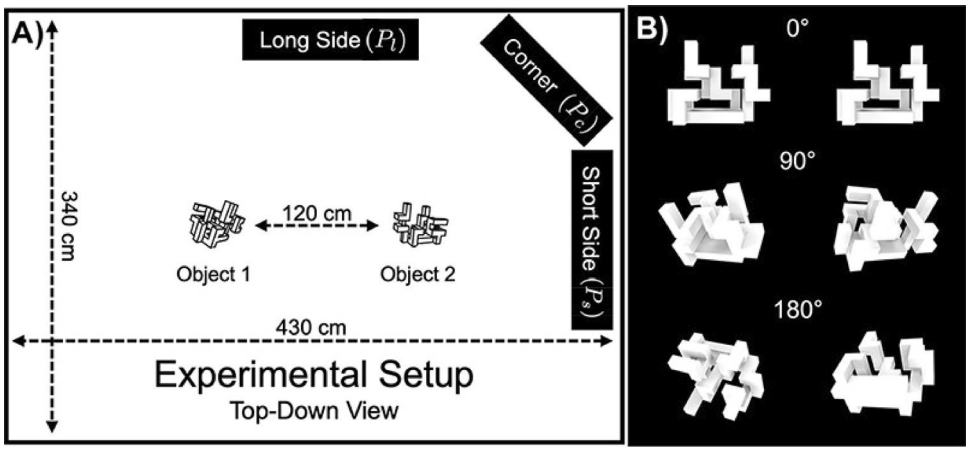

Figure 3. (**A**) Top-down illustration of the experimental setup. It shows dimensions, as well as the three different starting positions investigated. (**B**) We have investigated three orientational differences between the stimuli. 0° (top) means that there is no rotational difference between both object poses—the object rotations are aligned. 90° and 180° means that the poses have a rotational difference of 90° and 180°, respectively. For all trials, we have used the same poses—as shown in this illustration.

A pair of eye tracking glasses by Tobii (Pro Glasses 2) to capture the gaze direction, a motion tracking system by OptiTrack for head tracking, 1st and 3rd person video and homogenous lighting set up. Specialized software synchronized and aligned data streams from each source at microsecond precision[25]. To track the position and orientation of the stimuli, we developed motion-tracking markers that were attached to the stand of the objects. The subject's head motion was tracked using a custom tracking body attached to the eye-tracking glasses with 6 degrees of freedom (3 translation + 3 rotation). Figure 2B shows the custom clip-on equipment and Fig. 2C shows a photo of the assembly with the glasses. The tracking frequency for objects and head motion was 120 Hz, and for the eyes 50 Hz. The accuracy for the motion tracking system was ≈ 0.2 mm (RMSE) in 97% of the capture volume. The gaze was tracked with 1.42° mean accuracy. A fixation is defined as a sequence of raw gaze points with an estimated velocity below $30°/s$. Lastly, our statistical significance analysis is performed using GLMM (generalized linear mixed model) and n-way repeated-measures ANOVA. The performance accuracy is close to 100% for many participants; hence, to avoid a ceiling effect, we use GLMM here. All other independent variables and their interactions are analyzed using n-way repeated-measures ANOVA. For comparison, we also report the results of one-way repeated-measures ANOVA with adjusted p values in the supplementary Table 1.

Subjects

47 participants took part in our experiment. The average age was 23.4 years, ranging from 19 to 52 years of age. All subjects had normal, or corrected-to-normal vision, granted informed consent and were paid for participation. The experiment was approved by the Office of Research Ethics at York University (Certificates #2020-137 and #2020-217).

Results

In total, we conducted 846 trials. Each subject performed 18 trials, which in total sampled each configuration of experimental variables more than 15 times. We recorded about 80,000 fixations with over 4,500,000. head poses and 11 h of footage of 1st and 3rd person video each. A visualization of a trial is shown in Fig. 4, visualized using the graphical functions of *PESAO*. The subject required 61 fixations, moved a total of 18.75 m of head movement to complete the task, and answered correctly (same). We visualize three kinds of fixations; red, blue and grey for fixations on object 1, object 2, and environment, respectively. For the analyses involving fixations, we utilize the fixations on objects 1 and 2.

We next present observations on accuracy, number of fixations, response time, movement, and fixation patterns.

Accuracy

Throughout all configurations, participants achieved an absolute mean accuracy of 93.83%, $\sigma = 3.9\%$ (Figure S1, Supporting Information (SI)). The best-performing configuration was with stimuli of C_e, starting position P_l and a difference in object orientations of 0.0°. Not a single trial of this configuration was answered incorrectly, regardless of the object sameness. Object complexity plays a relevant role in how well participants performed this task. Objects of C_e complexity yield an average accuracy of 96.1%, C_m objects with 94.18%, and C_h with 91.2%. However, the complexity does not have a significant main effect ($p = 0.427$).

The sameness of objects has a significant main effect on accuracy ($p = 0.032$, see Figure S1 c, SI). If the objects are the same, in general, a higher accuracy is seen (94.3%) than for different pairings (91.6%).

We were also interested in investigating the effect of the starting position (Figure S1 a, SI). While for the C_e, the best mean performance was observed from P_c, for C_m, the best performance was seen starting from P_s, and

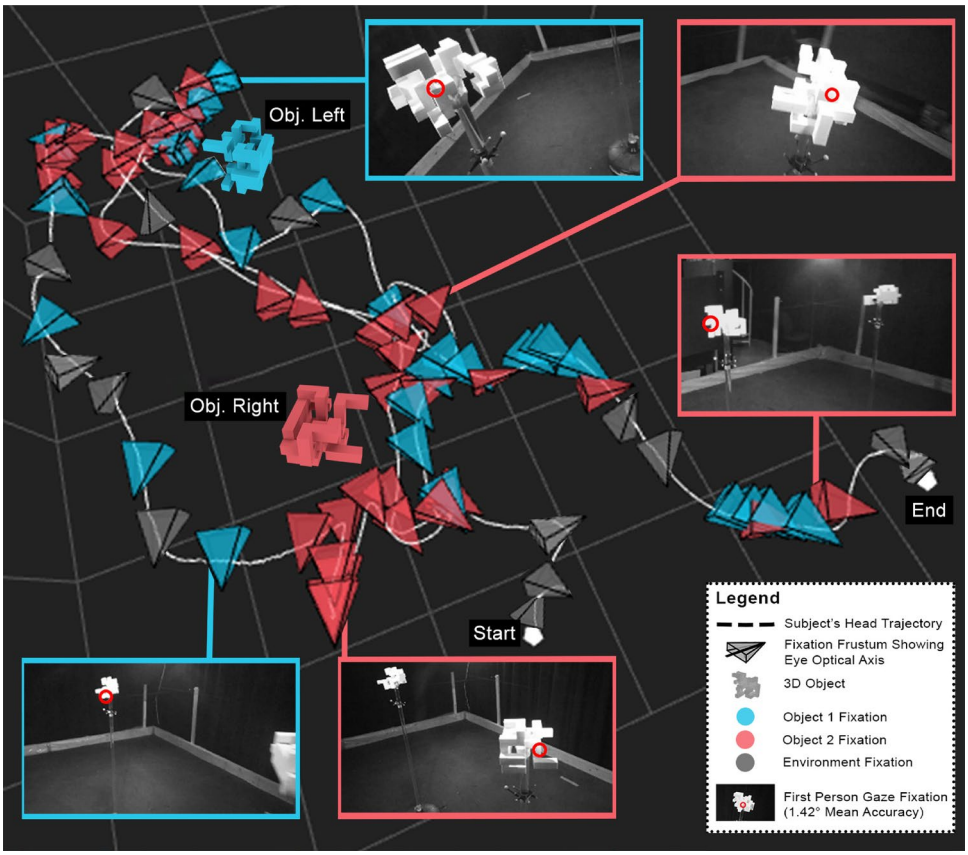

Figure 4. A visualization of the recorded data from PESAO. The subject's movement is plotted as a dashed line in white, and fixations are illustrated as colour-coded frustum—red, blue and grey for fixations on object 1, object 2, and environment (i.e., not on an object, likely to understand spatial layout for locomotion) , respectively. Selected fixation frusta are annotated with snapshots of the subject's first-person view and the gaze at a particular fixation (red circle). In this example, the objects are the same, of complexity level C_h, they differ in pose by 180°, and the subject started from position P_s. This visualization was created using *PESAO*[25].

finally, for C_h, starting from P_l achieved the highest accuracy. While the worst performances varied between P_s and P_c starting position, P_l did seem to result, generally, in higher accuracy. However, no significant effect of the starting position with respect to accuracy was observed ($p = 0.182$).

An investigation of the object orientation with regard to accuracy yields the following observations (Figure S1 b, SI). For the C_e case, there is a clear gradient of accuracy following the increase of orientation difference. Notably, trials of C_e objects with orientation 0° had an accuracy of 100%. However, for the other complexity levels, a different pattern can be identified; 90° was most accurately identified with 94.82% and 90% for C_m and C_h, respectively. 0° and 180° ranked second and third. In fact, the object orientation does have a significant effect on the accuracy ($p = 0.041$).

Every subject performed 18 trials, and no target object configurations were repeated. Nevertheless, we expected that some improvement in accuracy would begin to appear. Surprisingly, this was not the case; no significant learning effect was observed ($p = 0.539$), see Figure S1 d, SI.

Number of fixations

A substantial amount of data acquisition occurs during a solution to this task, as subjects used a minimum of 6 different eye fixations while averaging 92.38 across all trials. The object complexity plays a role in how many fixations are required to solve this task. C_e objects required about 76.56 fixations, C_m 79.53 fixations, and C_h 121.06 fixations on average. The effect of object complexity is statistically significant ($F_{2,92} = 39.8, p < 0.0001$), see Figure S2 e, SI.

The evaluation of sameness against the number of fixations revealed two major insights (Figure S2 c, SI). Firstly, the same pairings always required significantly more fixations than different pairings ($F_{1,46} = 9.12, p = 0.0026$). Secondly, the same pairings needed at least 10, in some cases up to 20, fixations on average more. Furthermore, error responses resulted in significantly more fixations than correct answers ($F_{1,46} = 9.762, p < 0.003$).

C_m and C_h cases, starting from P_s resulted in the most fixations on average, followed by starting from the P_c and P_l (Figure S2 a, SI). The starting position, similar to the accuracy of answering correctly, does not have any effect with respect to the number of fixations ($F_{2,92} = 0.34, p = 0.07$).

In terms of object orientation, (Figure S2 b), SI), orientations 0° and 90° are similar, varying only a few fixations for the median and upper and lower quartile. In terms of absolute values, a few trials of C_h and orientation of 0° required about 800 fixations. Notably, these trials started from P_l. In summary, larger orientation differences required significantly more fixations regardless of object complexity ($F_{2,92} = 8.98, p = 0.00013$).

It is interesting that a significant learning effect with respect to the number of fixations is observed ($F_{5,230} = 3.239, p = 0.0075$). This means that participants require fewer fixations (Figure S2 d, SI), hence solving the task more efficiently but not more accurately, as the trials progress.

Lastly, we found no significant interactions between these variables ($p > 0.05$).

Response time

The response time is the time elapsed from the first fixation of the trial to the time when the subject provided the answer. On average, the response time was 47.52 s ($\sigma = 30.39$). Among all trials, the shortest response was for an C_e level, starting from P_s, with 180° orientational difference and only taking 4.2 s. The longest response time was recorded for a C_h level, starting from P_l and required 298 s.

The complexity of the stimuli affects the response time gradually for C_e (on average 40.03 s), and C_m (on average 42.01 s) cases, and distinctively for C_h (on average 60.53 s)—increasing object complexity also means a significant increase in response time ($F_{2,92} = 48.3, p < 0.0001$), see Figure S3 e, SI. Furthermore, the response time is approximately a linearly increasing function of the angular difference of the objects; C_e ($n = 7, \Delta_t = 5.72$ s), C_m ($n = 10, \Delta_t = 4.2$ s), and C_h ($n = 18, \Delta_t = 3.36$ s), where n is the number of elements used to create the object and Δ_t is the normalized response time with respect to a single element (Figure S3 b, SI). This relates well to Sheppard and Metzler conclusions, but in our 3D active setting.

Similarly to the number of fixations required, the sameness of the stimuli has a distinct effect on the response time ($F_{1,46} = 19.1, p < 0.0001$), see Figure S3 c, SI—same cases take significantly longer than different ones.

Consistent with other measures, the response time is not significantly affected by the starting position ($F_{2,92} = 1.49, p = 0.225$), see Figure S3 a, SI. The object orientation, however, does affect response time ($F_{2,92} = 12.95, p < 0.0001$). In general, a smaller orientation difference also means a quicker response time—0° was answered the quickest, followed by 90°, and 180°, see Figure S3 b, SI.

Subjects seem to develop more efficient strategies with increasing trials completed. Starting at about 47 s (Median) at the first trial, the response time drops to about 34 s (Median) for trials two to four and drops further to 29 s Median at trials five and six. For C_h cases, a drop from the first trial (70 s Median) to the second trial (about 50 s Median) can be seen (Figure S3 d, SI). Overall, looking at the impact of progressing trials and their response time, a significant effect is noticed ($F_{5,230} = 1.69, p = 0.039$).

Finally, no significant interaction was observed between these variables ($p > 0.05$).

Movement

The mean extent of head movement was 16.62 m. The amount of head movement slightly increased from complexity cases of C_e to C_m but increased more distinctly for C_h cases—the object complexity significantly affects the amount of head movement ($F_{2,92} = 51.0, p < 0.0001$, Figure S4 d, SI).

Aligned with the number of fixations, response time and accuracy, the amount of movement is greater for the same object pairings across all complexity levels ($F_{1,46} = 32.4, p < 0.0001$), see Figure S4 c, SI. For cases where the objects are different, the increased upper and lower quartiles indicate a greater variability across subjects concerning viewing strategies.

We found no relationship between amount of head movement and starting position ($F_{2,92} = 57.4, p = 0.563$, Figure S4 a, SI). However, there exists a significant effect on the correctness of the answer and head movement (Figure S4 e, SI). Error responses were accompanied by significantly more head movement ($F_{1,46} = 11.4, p < 0.0001$).

A clear trend ($F_{2,92} = 37.0, p < 0.0001$) can be observed between the amount of head movement and the amount of orientational difference (see Figure S4 b, SI); at 0° the least amount of movement was required, at 90° an increase of 2–5 m on average is recorded, and at 180° an additional increase of 1–5 m across all complexity classes is recorded.

A significant reduction in head movement is noticable over the course of the trials ($F_{5,230} = 1.95, p = 0.0119$). This means that participants show a learning effect in the sense that they execute a strategy with less head moment as the experiment progresses. For C_m and C_h cases, a trend is not visible directly, but for the C_e cases, it is (Figure S4 d, SI). C_h cases start off at the first trial with just above 20 m and drop to the absolute mean value of 16.62 m and stay steady, marginally falling below and exceeding it repetitively; similarly, for the C_m case, where no learning trend can be observed. However, the C_e case, while noticing a slight up-trend for the second trial, consecutively decreases from about 16 m down to about 10 m, which is the equivalent of an improvement of 37.5%.

Lastly, in Figure S5, SI, we plot normalized measured variables (accuracy, amount of fixations, response time, and movement) against trial number. It is easily seen that every measurement decreases over the trials except for accuracy.

The multi-factor analysis did not show any significant interactions for these variables ($p > 0.05$).

Fixation patterns

Lastly, we looked at fixation patterns. We considered the ratio of fixations landing on each object respectively and fixation groupings, specifically fixations falling on the same object before shifting away.

To evaluate the fixation ratio, we looked at the total number of fixations for either object. The object with the most fixations is considered the primary object, and the object with fewer or the same number of fixations is the

secondary object. On average, the primary object accounted for 59.53% of fixations, and the secondary object for 40.47%, considering object fixations only.

None of the experimental variables have a significant effect on the fixation ratio: object complexity ($F_{2,92} = 0.858, p = 0.424$), starting position ($F_{2,92} = 0.94, p = 0.39$), object orientation ($F_{2,92} = 1.37, p = 0.254$), object sameness ($F_{2,46} = 1.41, p = 0.234$), trial progression ($F_{5,230} = 1.03, p = 0.395$), and correctness of answer ($F_{1,46} = 0.082, p = 0.777$) (Figure S6, SI).

However, for some configurations, a difference in fixation ratio is noticeable. For instance, the difference in the average number of fixations with respect to object orientation decreases with increasing orientational difference (Figure S6 c, SI).

Next, we looked at fixation groupings, which means the number of fixations on one object before changing focus to the other object (Figure S7, SI). On average, 18.7% ($\sigma = 10.01\%$) are fixations that change focus between each object every time—here, we call them single fixations. Couple fixations is defined as two fixations on one object before changing the object. Triple fixations means three fixations, and so on. It follows that these groupings include no environment fixations. The remaining 81.3% of fixations are divided as follows: couple fixations 18.43% ($\sigma = 11.48\%$), triple fixations 12.26% ($\sigma = 8.51\%$), quadruple fixations 8.57% ($\sigma = 7.41\%$), quintuple fixations 6.71% ($\sigma = 7.20\%$), sextuple fixations 4.87% ($\sigma = 7.47\%$), septuple fixations 3.74% ($\sigma = 5.94\%$), octuple fixations 3.23% ($\sigma = 6.13\%$), and higher groupings (> 8) 18.41% ($\sigma = 17.90\%$).

The object complexity has a significant effect on single ($F_{2,92} = 8.53, p = 0.0003$), couple ($F_{2,92} = 9.35, p = 0.0002$), triple ($F_{2,92} = 4.34, p = 0.015$), quintuple ($F_{2,92} = 6.428, p = 0.002$), octuple ($F_{2,92} = 5.10, p = 0.007$) and higher ($F_{2,92} = 15.56, p < 0.0001$) fixation groupings (Figure S7 a, SI). Notably, for single, couple, and triple, the probability of occurrence significantly decreases with increasing object complexity. While for octuple and higher groupings, the opposite is true—their occurrence increases with increasing object complexity.

While object complexity has a significant effect on grouping patterns, the starting position does not. No investigated fixation group sizes showed to be significantly affected by the starting position: single ($F_{2,92} = 0.433, p = 0.648$), couple ($F_{2,92} = 0.036, p = 0.962$), triple ($F_{2,92} = 0.804, p = 0.447$), quadruple ($F_{2,92} = 1.91, p = 0.149$), quintuple ($F_{2,92} = 0.629, p = 0.533$), sextuple ($F_{2,92} = 0.187, p = 0.829$), septuple ($F_{2,92} = 1.37, p = 0.255$), octuple ($F_{2,92} = 1.06, p = 0.348$), higher ($F_{2,92} = 1.3, p = 0.272$).

However, the object orientation has a significant effect on single ($F_{2,92} = 14.5, p < 0.0001$) and couple ($F_{2,92} = 25.6, p < 0.0001$) groupings as this is the dominant method for object orientation 0° and decreases steadily with increasing object orientation. Similar to object complexity, larger fixation groupings are affected by object orientation as well. Specifically, septuple ($F_{2,92} = 2.67, p = 0.07$) and higher ($F_{2,92} = 11.3, p < 0.0001$) groupings occur more frequently with increasing object orientation (Figure S7 c, SI).

The sameness of the object had largely no significant effect on fixation groups (Figure S7 d, SI). Only single ($F_{2,92} = 5.02, p = 0.025$) and septuple ($F_{2,46} = 5.39, p = 0.02$) groups are more significantly used for the same objects than different ones. As trials advanced, subjects used single groupings progressively less ($F_{2,92} = 1.26, p = 0.281$). A similar trend is observed for couple, triple, quadruple and quintuple groups—none are significant, however. Larger groupings see an increase in probability as trials proceed. Notably, only a few sparse data points are recorded for octuple pairings up to trial 8. Octuple pairings are more regularly seen for trials 9–19 (Figure S7 e, SI).

The correctness of the answer significantly correlated with octuple pairs ($F_{2,92} = 10.1, p = 0.001$). Octuple fixation groups are significantly used more for error answers than correct responses (Figure S7 f, SI).

Finally, all interactions between these variables were not significant ($p > 0.05$).

The fixation groupings reveal a purposeful gaze toward the solution of 3D visuospatial problems. Further analysis and experiments are hoped to flesh out behaviours comprising human functional vision.

Discussion

The goal of this study was to examine functional vision in human subjects, specifically, how they solve a visuospatial problem in a three-dimensional physical space. We addressed this by developing a three-dimensional version of the well-known same-different task as a probe and an experimental setup allowing natural, visual problem-solving and precision recording. Such physical object comparisons seem a fundamental cognitive ability[31].

Our main results follow. No training trials were required. The range of response times from simplest to most complex cases ranges from 4 to 298 sec. and accuracy from 80 to 100%. A great deal of data acquisition occurs during all trials with the range of eye movements (separate fixations and separate images processed) from 6 to 800 fixations.

Furthermore, we showed that not only multiple fixations are required, but also multiple fixations in sequence on the same stimulus. Only about 20% of all fixations are single fixations (not preceded or followed by a fixation on the same object), and fixation groups get larger and more frequent with increasing levels of complexity and orientation. These groups seem to develop throughout the course of the trial. Simpler groupings (single, couple, triple) are replaced by more complex ones (septuple, octuple, larger) as the trial progresses. This hints that subjects use what they know to dynamically compose visuospatial strategies. In our analysis of fixation ratios, subjects did not simply observe each object with the same number of fixations. They chose one object as their primary object (59.53% of total fixations, regardless of experimental setup) and spent just about 40% of fixations on the secondary to solve this problem—brute-force approaches would have averaged a 50 : 50 ratio leading to the conclusion that subjects did not use random or uninformed search strategies. Could they be building internal models of the objects, which are then compared? This possibility needs further investigation. In Figure S8, SI, we illustrate a set of fixations observed in a trial with C_m complexity, same objects, starting from P_c, presented at

90°. This pattern looks like an observation strategy; the gaze goes back and forth between stimuli. Interestingly, the fixations land on similar-looking areas. It appears that the subject compares object features.

No statistical change was observed in accuracy with increasing trials for individual subjects. However, a change was observed in the number of fixations, response time, and amount of head movement. This is surprising, as the accuracy did not significantly change throughout the trials and shows that a subject's set of visuospatial problem-solving techniques, innate and learned through a lifetime, generalized well to this specific task. However, the decrease in the number of fixations, response time, and amount of head movement shows that participants seem to fine-tune their techniques leading to greater efficiency. This learning effect seems counterintuitive, especially when compared to modern computational attempts at active learning (however, see[32] for review and a promising change).

Our work also shows consistency with the classic version of the same-different task[9] which showed that the time required to determine if two perspective drawings portray objects of the same three-dimensional shape is a linearly increasing function of the angular difference in the portrayed orientations for the two objects.

Other three-dimensional stimuli were considered. Some examples are the well-studied greebles families[33], the *CLEVR* data set[34], or *T-LESS*[35]. All are virtual but could be physically created, for instance, using fused manufacturing modelling. Various versions of the greebles objects also exist. These objects would function well as a stimulus for this experiment as they are textureless, and a common-coordinate system could be defined easily (greebles are all structured similarly). However, different greebles appear quite different and would make the same-different task trivial. The *CLEVR* data set does use simple blocks to build the stimulus, similar to *TEOS*; however, neither a systematic measure for self-occlusion nor a common coordinate system to define the object pose exists. *T-LESS* objects do have an associated pose and also have a textureless appearance, but the objects are easily differentiable. *TEOS* combines crucial properties to discover any patterns in solving the same-different task; novel and unfamiliar, textureless appearance, known complexity, common coordinate system, and varying self-occlusion.

Humans use vision for a vast array of behaviours in the real world; visuospatial intelligence is much more than simply detecting a stimulus or recognizing an object or scene. Although past methodologies have limited these studies, new techniques are beginning to explore more complex problems. Our work adds to this, using a novel setup and probe task to examine a fundamental aspect of visual intelligence, object comparison. We do so in a manner where all subject behaviour is dedicated to the visual aspects of the task, as there is no motor action other than those that determine visual gaze and viewpoint. Where do we look? How do we look? How do we move? How do we seek out the data that enables problems to be solved? The first steps towards these answers are presented along with an experimental infrastructure appropriate for many further studies. Our data shows that we actually do a great deal of 'looking' in order to solve a problem. Ongoing work is examining fixation patterns to uncover strategies that subjects deploy, such as the local structure comparison of Figure S8, including the detection of degenerate views that lead to disambiguating next views, moving closer to enhance resolution, and more that exhibit clear causal connections among fixations. Adult humans do not need to learn where to look and exhibit an array of complex fixation patterns. We also move about as we look, most likely because we choose what we wish to see and from what vantage in order to support the task at hand.

We found that the sequence of steps strongly suggests that human problem-solving strategies are a dynamic process of formulating and testing hypotheses adapted to the specific task at hand. The task is not learned by the subjects; instead, participants generate effective solutions right from the beginning and gradually improve in efficiency while retaining their accuracy. Although the use of our eyes may seem effortless, the complexity of actions is staggering and unravelling their purpose—the 'why' behind all this looking—poses an exciting challenge.

Data availability

The data and code used to perform the analysis of this work, as well as the *TEOS* templates, are publicly available at https://gitlab.nvision.eecs.yorku.ca/solbach/activevisuospatial.

Received: 21 June 2023; Accepted: 9 November 2023
Published online: 15 November 2023

References

1. Bennett, C. R., Bex, P. J., Bauer, C. M. & Merabet, L. B. The assessment of visual function and functional vision. *Semin. Pediatr. Neurol.* **31**, 30–40. https://doi.org/10.1016/j.spen.2019.05.006 (2019).
2. Hayhoe, M. Vision using routines: A functional account of vision. *Vis. Cogn.* **7**, 43–64 (2000).
3. Bonnen, K. *et al.* Binocular vision and the control of foot placement during walking in natural terrain. *Sci. Rep.* **11**, 20881 (2021).
4. Matthis, J. S., Muller, K. S., Bonnen, K. L. & Hayhoe, M. M. Retinal optic flow during natural locomotion. *PLoS Comput. Biol.* **18**, e1009575 (2022).
5. Muller, K. S. *et al.* Retinal motion statistics during natural locomotion. *Elife* **12**, e82410 (2023).
6. Parker, P. R., Abe, E. T., Leonard, E. S., Martins, D. M. & Niell, C. M. Joint coding of visual input and eye/head position in V1 of freely moving mice. *Neuron* **110**, 3897-3906.e5. https://doi.org/10.1016/j.neuron.2022.08.029 (2022).
7. Kadohisa, M. *et al.* Frontal and temporal coding dynamics in successive steps of complex behaviour. *Neuron*https://doi.org/10.1016/j.neuron.2022.11.004 *(2022).*
8. Martinho, A. & Kacelnik, A. Ducklings imprint on the relational concept of "same or different". *Science* **353**, 286–288. https://doi.org/10.1126/science.aaf4247 (2016) arXiv:1011.1669v3.
9. Shepard, R. N. & Metzler, J. Mental rotation of three-dimensional objects. *Science* **171**, 701–703 (1971).
10. Macmillan, N. A. & Creelman, C. D. *Detection Theory: A User's Guide* (Psychology press, 2004).
11. Findlay, J. M. & Gilchrist, I. D. *Active vision* (Oxford University Press, 2003).
12. Duke-Elder, S. System of ophthalmology. *Ocul. Motil. Strabismus* **6**, 223–228 (1973).
13. Carpenter, R. H. S. *Movements of the Eyes, 2nd Revision* (Pion Limited, 1988).
14. Liversedge, S., Gilchrist, I. & Everling, S. *The Oxford Handbook of Eye Movements* (OUP, 2011).

15. Davitt, L., Cristino, F., Wong, A. & Leek, E. C. Fixation preference for concave surface discontinuities during object recognition generalises across levels of stimulus classification. *J. Exp. Psychol. Hum. Percept. Perform.* **40**, 451–456 (2014).
16. Ballard, D. H., Hayhoe, M. M. & Pelz, J. B. Memory representations in natural tasks. *J. Cogn. Neurosci.* **7**, 66–80 (1995).
17. Pelz, J., Hayhoe, M. & Loeber, R. The coordination of eye, head, and hand movements in a natural task. *Exp. Brain Res.* **139**, 266–277 (2001).
18. Land, M. F. & Lee, D. N. Where we look when we steer. *Nature* **369**, 742–744 (1994).
19. Land, M. F., Mennie, N. & Rusted, J. Eye movements and the roles of vision in activities of daily living: Making a cup of tea. *Perception* **28**, 1311–1328 (1999).
20. Matthis, J. S., Yates, J. L. & Hayhoe, M. M. Gaze and the control of foot placement when walking in natural terrain. *Curr. Biol.* **28**, 1224–1233 (2018).
21. Tatler, B. W., Hayhoe, M. M., Land, M. F. & Ballard, D. H. Eye guidance in natural vision: Reinterpreting salience. *J. Vis.* **11**, 5 (2011).
22. Rothkopf, C. A., Ballard, D. H. & Hayhoe, M. M. Task and context determine where you look. *J. Vis.* **7**, 16 (2007).
23. Triesch, J., Ballard, D. H., Hayhoe, M. M. & Sullivan, B. T. What you see is what you need. *J. Vis.* **3**, 9 (2003).
24. Wang, R. F. & Simons, D. J. Active and passive scene recognition across views. *Cognition* **70**, 191–210 (1999).
25. Solbach, M. D. & Tsotsos, J. K. PESAO: Psychophysical experimental setup for active observers 1–20. arXiv preprint arXiv:2009.09933 (2020).
26. Mišiak, M., Fuhrmann, A. & Latoschik, M. E. The impact of reflection approximations on visual quality in virtual reality. In *ACM Symposium on Applied Perception* 1–11, Vol. 2023 (2023).
27. Mon-Williams, M. & Wann, J. P. Binocular virtual reality displays: When problems do and don't occur. *Hum. Factors* **40**, 42–49 (1998).
28. Zhao, Q. 10 scientific problems in virtual reality. *Commun. ACM* **54**, 116–118 (2011).
29. Solbach, M. D. & Tsotsos, J. K. Blocks world revisited: The effect of self-occlusion on classification by convolutional neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 3505–3514 (2021). https://doi.org/10.1109/iccvw54120.2021.00390. arXiv:2102.12911.
30. Solbach, M. D. *Active Observers in a 3D World: Human Visual Behaviours for Active Vision* (York University, 2022).
31. Carroll, J. B. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies* (Cambridge University Press, 1993).
32. Taylor, A. T., Berrueta, T. A. & Murphey, T. D. Active learning in robotics: A review of control principles. *Mechatronics* **77**, 102576. https://doi.org/10.1016/j.mechatronics.2021.102576 (2021) arXiv:2106.13697.
33. Gauthier, I. & Tarr, M. J. Becoming a"Greeble" expert: Exploring mechanisms for face recognition. *Vis. Res.* **37**, 1673–1682. https://doi.org/10.1016/S0042-6989(96)00286-6 (1997).
34. Johnson, J. *et al.* CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2901–2910. https://doi.org/10.1109/CVPR.2017.215 (2017). arXiv:1612.06890.
35. Hodaň, T. *et al.* T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. In *Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017* 880–888. https://doi.org/10.1109/WACV.2017.103 (2017). arXiv:1701.05498.

Author contributions

Conceptualization: M.D.S. and J.K.T. Methodology: M.D.S. and J.K.T. Investigation: M.D.S. Visualization: M.D.S. Supervision: J.K.T. Writing original draft: M.D.S. Review and editing: J.K.T., and M.D.S.

Funding

This material is based upon work supported by the Air Force Office of Scientific Research under award numbers FA9550-18-1-0054 and FA9550-22-1-0538 (Computational Cognition and Machine Intelligence, and Cognitive and Computational Neuroscience portfolios); the Canada Research Chairs Program (Grant Number 950-231659); Natural Sciences and Engineering Research Council of Canada (Grant Numbers RGPIN-2016-05352 and RGPIN-2022-04606).

Competing Interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-47188-4.

Correspondence and requests for materials should be addressed to M.D.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.