Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

Model of Multiple Identity Tracking (MOMIT) 2.0: Resolving the serial vs. parallel controversy in tracking

Jie Li[a,*], Lauri Oksama[b], Jukka Hyönä[c]

[a] School of Psychology, Beijing Sport University, China
[b] National Defense University, Finland
[c] Department of Psychology, University of Turku, Finland

ARTICLE INFO

Keywords:
Multiple-identity tracking
Eye movements
Gaze-contingent display change
Visual attention
Working memory

ABSTRACT

The present study investigated whether during tracking of multiple moving objects with distinct identities only one identity is tracked at each moment (serial tracking) or whether multiple identities can be tracked simultaneously (parallel tracking). By adopting the gaze-contingent display change technique, we manipulated in real time the presence/absence of object identities during tracking. The data on performance accuracy revealed a serial tracking pattern for facial images and a parallel pattern for color discs: when tracking faces, the presence/absence of only the currently foveated identity impacted the performance, whereas when tracking colors, the presence of multiple identities across the visual field led to improved tracking performance. This pattern is consistent with the identifiability of the different types of objects in the visual field. The eye movements during MIT showed a bias towards visiting and dwelling on individual targets when facial identities were present and towards visiting the blank areas between targets when color identities were present. Nevertheless, the eye visits were predominately on individual targets regardless of the type of objects and the presence of object identities. The eye visits to targets were beneficial for target tracking, particularly in face tracking. We propose the Model of Multiple Identity Tracking (MOMIT) 2.0 which accounts for the results and reconcile the serial vs. parallel controversy. The model suggests that observers cooperatively use attention, eye movements, perception, and working memory for dynamic tracking. Tracking appears more serial when high-resolution information needs to be sampled and maintained for discriminating the targets, whereas it appears more parallel when low-resolution information is sufficient.

1. Introduction

When we look around, we may see objects moving around us, such as vehicles, animals, or people. In order to maintain situation awareness of our visual environment so as to interact with it or adequately respond to moving objects, we need to keep track of the object identities, knowing what each object is and where it moves to. In basketball, the player needs to track the individual players of her/his own team, for example to know where the team's excellent 3-point scorer is currently located. Similarly, a car driver approaching a busy intersection needs to track the whereabouts and movement trajectories of other vehicles and pedestrians in order to decide his/her own move. We have coined it the multiple identity tracking (MIT) task (Oksama & Hyönä, 2004). It significantly differs from the more studied multiple object tracking (MOT; Pylyshyn & Storm, 1988), where the to-be-tracked targets are identical, so only their location needs to be tracked. This process of multiple identity tracking is crucial for our everyday life, and hence the underlying mechanism of the tracking process has attracted increased research interests in recent years (Botterill, Allen, & McGeorge, 2011; Cohen, Pinto, Howe, & Horowitz, 2011; Horowitz et al., 2007; Li, Oksama, & Hyönä, 2018a, 2018b; Nummenmaa, Oksama, Glerean, & Hyönä, 2017; Oksama & Hyönä, 2004, 2008, 2016; Papenmeier, Meyerhoff, Jahn, & Huff, 2014).

1.1. The serial vs. parallel controversy in tracking

Oksama and Hyönä (2008) have proposed the Model of Multiple Identity Tracking (MOMIT) to account for the process of MIT. This model suggests that people track the identity-location bindings of multiple objects in a serial manner. There is just one attentional focus, which is tightly linked with eye movements (Corbetta et al., 1998; Deubel & Schneider, 1996). At each moment, the foveal attention is

* Corresponding author at: Room 309, Sport Science Building, Beijing Sport University, No. 48 Xinxi Road, Haidian District, Beijing 100084, China.
E-mail addresses: lijie.psy@gmail.com (J. Li), loksama@utu.fi (L. Oksama), hyona@utu.fi (J. Hyönä).

https://doi.org/10.1016/j.cognition.2018.10.016
Received 12 February 2018; Received in revised form 3 October 2018; Accepted 23 October 2018
Available online 29 October 2018

directed to one target for processing the identity information of the object and refreshing the binding between the object identity and its current location. The foveal attention switches between target objects serially so as to keep the identity-location binding of each object updated. Only the currently attended object is effectively tracked at each moment.

In contrast, the parallel tracking model suggests that people are able to track multiple objects simultaneously in a parallel manner. The models assume that people are monitoring multiple objects simultaneously as they are moving via a small number of visual indexes or multifocal attention (Cavanagh & Alvarez, 2005; Howe & Ferguson, 2015; Pylyshyn, 1989). Each visual index or attentional focus follows a target as it moves about, so that its identity information and changing location can be encoded. The information of each object is integrated into an object file (Kahneman, Treisman, & Gibbs, 1992). Hence, the visual system can track multiple object files simultaneously and retain awareness of the identity and location information of each target object.

The dispute regarding the issue of serial vs. parallel tracking not only concerns the mechanism of MIT, but also taps into the fundamental characteristics of human attention: whether we have genuine situational awareness of an enriched visual environment, or whether we are actually aware of only a single object that is in the focus of attention. Though the parallel tracking hypothesis may better match with people's intuitions assuming that we are monitoring all the objects in front of our eyes all the time, a large body of research in visual cognition suggests that such conception may be an illusion (O'Regan, 1992). In MIT studies, both models have received support (Howe & Ferguson, 2015; Oksama & Hyönä, 2008, 2016), and it is difficult to discriminate the two models simply based on observers' tracking accuracy. In fact, it is possible that MIT is carried out by conjoined systems that involve both serial and parallel components, in accordance with the two-stage theories of general visual processing: a parallel visual analysis of basic features of multiple elements in the scene and a serial scrutiny of a few elements (Buetti, Cronin, Madison, Wang, & Lleras, 2016; Wolfe & Horowitz, 2017). Even though MOMIT is usually referred to as a serial model as it emphasizes the serial attention shifts between the targets, it also includes a parallel component, as it assumes that location information about multiple moving objects is processed in parallel in the peripheral vision and multiple representations containing the objects' identity-location bindings are maintained in episodic memory (Oksama & Hyönä, 2008).

1.2. The present study

In the present study, we aimed to examine the issue of serial vs. parallel mechanism in MIT by combining MIT with eye movements. Eye movements and visual attention are tightly linked (e.g., (Hoffman, 1998; Peterson, Kramer, & Irwin, 2004). People generally attend to where they are fixating at (Deubel & Schneider, 1996; Hoffman & Subramaniam, 1995; Li, Oksama, & Hyönä, 2018a, 2018b; Meyerhoff, Schwan, & Huff, 2018), though attention can be covertly shifted to other locations without moving the eyes (Belopolsky & Theeuwes, 2009; Van Ettinger-Veenstra et al., 2009). We applied the gaze-contingent display change technique to MIT. The technique offers a powerful tool for examining the processing and utilization of visual information in various tasks. Previously this technique has been used mostly in reading research to investigate participants' attentional processing of words (Rayner, 1975, 1998). With this paradigm, it is possible to manipulate in real time what is presented on the screen contingent on where the observer is looking at from moment to moment (Loschky, McConkie, Yang, & Miller, 2005; McConkie, 1997; Perry & Geisler, 2002). Here, we introduced the paradigm to dynamic tracking in order to manipulate the presence/absence of object identities during tracking in real time according to where the observer is fixating at. The general logic is that if observers only track the identity of the foveated object at each moment, as proposed by the serial model, presenting only the identity of the foveated object would result in an equally good tracking performance as when all the target identities are present. Conversely, masking the identity of the foveated object while fully presenting the objects in the periphery would result in an equally poor performance as when no target identity information is present during tracking.

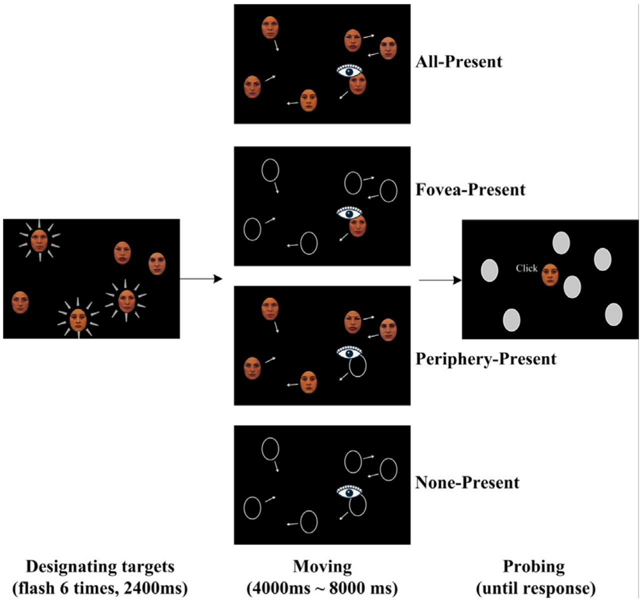

Fig. 1. Trial sequence of the gaze-contingent display change experiments. The figure depicts a condition where 3 out of 6 faces are designated as the targets at the beginning of the trial. Then all the faces started moving. The four images in the middle column illustrate the four conditions of presentation mode during tracking. The eye icon indicates the gaze position at the current moment. At the end, the objects stopped moving and they were covered by masks; the participants were required to click where each of the target faces finally stopped.

Specifically, we set up four conditions for the present experiments (see the middle column in Fig. 1): (1) all object identities present (All-Present), (2) only the foveated object identity present (Fovea-Present), (3) only the object identities in the periphery present (Periphery-Present), and (4) none of the object identities present during tracking (None-Present). The All-Present condition is identical to the setting in a typical MIT experiment, in which the identities of all objects are present all the time as they move about, regardless of the observers' gaze position. The Fovea-Present and Periphery-Present conditions involved gaze-contingent display changes. In the Fovea-Present condition, the target identity was presented whenever it was foveated by the observer, whereas the identities of all the other objects in the periphery were not presented but were replaced with empty placeholders moving around on the screen. The Periphery-Present condition was opposite to the Fovea-Present condition: the identity of an object disappeared whenever it was foveated, leaving only an empty placeholder in its place, while the identities of all objects in the periphery were presented. In the None-Present condition, none of the identities was present during the tracking stage, leaving only empty placeholders moving around the screen. This condition was first used by Pylyshyn (2004) for exploring identity tracking (see also Cohen et al., 2011).

The tracking performance in the Fovea-Present and Periphery-Present conditions was compared with that of the All-Present and None-Present conditions to investigate the dynamic tracking of identity information. If only the currently foveated object identity is tracked at each moment, only the presence/absence of the foveated object identity would affect the tracking performance, whereas the presence/absence of other target identities would not matter. Thus, when only the identity of the foveated object is presented, tracking performance would be equal to when all the identities of all the objects are presented all the time (i.e., Fovea-Present = All-Present). Moreover, tracking performance may be substantially impaired when the identity of the currently foveated object is never presented (Periphery-Present < All-Present), perhaps approximating the performance level when no identity of any object is presented (Periphery-Present = None-Present). Alternatively, if observers can track all or most of the targets simultaneously, what matters for the tracking performance is the number of identities present at each moment. In short, according to the serial hypothesis, the performance accuracy in the four conditions will be All-Present = Fovea-Present > Periphery-Present = None-Present, whereas the parallel hypothesis will predict the following pattern: All-Present > Periphery-Present > Fovea-Present > None-Present.

2. Experiment 1: Tracking multiple faces

2.1. Method

2.1.1. Participants

Forty-three participants from University of Turku, Finland took part in the experiment in exchange of course credit. The sample size was predetermined, so that the data collection was terminated when there were 20 participants in each condition. This sample size is comparable to our recent study which yielded robust effects when comparing MIT against MOT and the eyes fixed condition against free viewing (Nummenmaa et al., 2017). It also exceeds the sample size used in earlier MIT studies (e.g., Oksama & Hyönä, 2016). All participants had normal or corrected-to-normal vision, and all provided informed consent. The participants were divided into two groups so that one group tracked 3 targets and the other group tracked 4 targets. The two set sizes were chosen to be within the capacity limits of most participants (Oksama & Hyönä, 2004). By manipulating set size, we were also able to examine whether the manner of tracking remains consistent as the set size varies within the capacity limits. The data of 4 participants were excluded due to calibration problems (2 participants), program failure (1 participant), or chance-level performance (1 participant). The data of the remaining 39 participants (12 males, 27 females) were included in the analyses: 20 in tracking 3 targets and 19 in tracking 4 targets. The mean age of the participants was 22.3 years.

2.1.2. Stimuli and apparatus

Faces were used as stimuli, as they are common objects to track in daily life. Eighteen digitized color photographs of faces were selected from the Karolinska Directed Emotional Faces corpus (KDEF; Lundqvist, Esteves, & Öhman, 1999). The photos were of 18 Caucasian amateur actors (9 females, 9 males), posing neutral expression and gazing directly at the viewer. Non-facial areas in each photo were removed by applying an ellipsoidal mask.

The stimuli were presented on a 21 in. CRT monitor with a screen resolution of 1024 × 768 pixels and a 120-Hz refresh rate. Participants were seated 70 cm from the monitor. The display subtended an area of 24.3° × 17.0°. The object images were shown on a black background. Each face image subtended a visual angle of 1.7° × 2.3° at the 70-cm viewing distance. The experiment was programmed in Matlab using the Psychophysics Toolbox routines (Brainard, 1997). The objects moved in random directions at a velocity of 4.5°/s. They bounced off each other when the center-to-center distance was less than 2.5° and bounced off the edges of the display when the center-to-edge distance was less than 2.5°. Eye movements of the participants were recorded with a desktop-mounted Eyelink 1000 (SR Research, Canada) system. Sampling frequency was 1000 Hz. A forehead and chin rest was used to stabilize the participants' head position.

2.1.3. Procedure and design

Prior to the experiment and after each break, the eye-tracker was calibrated using a nine-point calibration grid extending over the entire computer screen. Before each trial, a drift correction was performed by presenting a dot in the middle of the screen.

For one group of 20 participants, there were 6 distinct faces of the same gender in each trial, 3 of them were targets while the other 3 were distractors; for the other group of 19 participants, there were 8 faces (4 targets + 4 distractors). At the beginning of each trial, all the faces were presented at random locations on the screen. White circles flashed on and off around a subset (3 or 4) of them for 2.4 s, indicating that these were the targets to be tracked. Then all faces started moving for a period between 4 s and 8 s. As soon as the motion stopped, all the faces were occluded by grey ovals. Finally, the target faces appeared at the center of the screen one by one. Participants were required to report the final location of each probed target by clicking the grey oval covering that location (Fig. 1). The presentation of the object identities was manipulated during the tracking stage by gaze-contingent display changes using four presentation conditions: All-Present, Fovea-Present, Periphery-Present, and None-Present. The presentation conditions are graphically depicted in Fig. 1. An object was considered foveated when the distance between the observer's current gaze position and the center of the object image was smaller than 2.5°.

The experimental design was 2 (set size) × 4 (presentation mode), with the set size being a between-participants variable and the presentation mode being a within-participants variable. There were 24 trials in each condition of presentation mode. The 96 trials were randomly mixed. Twenty-four motion trajectory files were generated before the experiment and stored offline. Each trajectory file was used once in each of the four conditions, ensuring that the motion sequences were identical across conditions. Another eight trajectory files were generated and stored for the use in the practice trials. Before the experiment, the participants were familiarized with the procedure and the experimental task, and then performed eight practice trials, two in each condition of presentation mode.

2.2. Results

The dependent variable in the experiment is the identity tracking accuracy, which measures the participants' performance of tracking the identity-location binding of each target. In each trial, all the 3 or 4 distinct targets were probed one by one, and participants made 3 or 4 responses to judge where each of the target had finally moved to. A response was considered correct when the participant clicked the location of the probed target. We first calculated the accuracy for each trial. For instance, if a participant made 2 correct responses when tracking 3 targets, the accuracy in the trial was 0.667 (i.e., 2/3). Then the identity accuracy in each experimental condition was calculated by averaging the accuracies across the trials of each condition.

A 2 (set size: 3, 4) × 4 (presentation mode: All-Present, Fovea-Present, Periphery-Present, None-Present) repeated-measures ANOVA was conducted on the identity tracking accuracy, with presentation mode as the within-participants variable and set size as the between-participants variable. In all the analyses in the present study, a Greenhous-Geisser correction was employed when the assumption of sphericity was violated, and a Bonferroni correction was employed for pair-wise comparisons. The original degrees of freedom are reported when corrections were employed. When the Bonferroni correction was employed, the Bonferroni-corrected *p*-value is reported, which equals to the uncorrected *p*-value multiplied by the number of comparisons made, with the restriction that the corrected *p*-value should be no larger than 1.

The ANOVA yielded a significant main effect of set size, $F(1, 37) = 26.121$, $p < .001$, $\eta_p^2 = 0.414$, showing that accuracy was higher for tracking 3 faces than 4 faces ($M = 77.9\%$, $SD = 15.2\%$ vs. $M = 53.1\%$, $SD = 17.5\%$). There was also a significant main effect of presentation mode, $F(3, 111) = 14.119$, $p < .001$, $\eta_p^2 = 0.276$, while the interaction between presentation mode and set size was not significant, $F(3, 111) = 1.315$, $p = .273$, $\eta_p^2 = 0.034$. Planned pair-wise

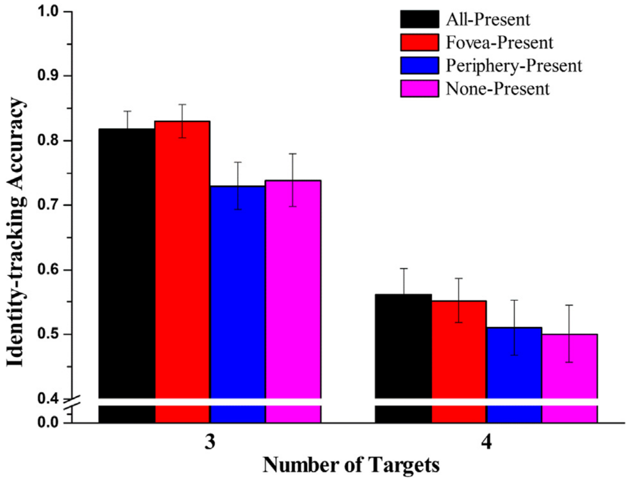

Fig. 2. Identity tracking accuracy for facial images. The error bars represent ± 1 S.E.

comparisons showed that the accuracy in the Fovea-Present condition was similar to that in the All-Present condition ($M = 69.5\%$, $SD = 19.3\%$ vs. $M = 69.3\%$, $SD = 19.7\%$, $p = 1.000$, Cohen's $d = 0.020$), and the accuracy in the Periphery-Present condition was identical to that in the None-Present condition ($M = 62.3\%$, $SD = 20.5\%$ vs. $M = 62.3\%$, $SD = 22.0\%$), $p = 1.000$, Cohen's $d = 0.001$), while the accuracies in the Fovea-Present and All-Present conditions were significantly higher than those in the Periphery-Present and None-Present conditions ($ps < .002$, Cohen's $ds > 0.641$). Thus, the pattern was All-Present = Fovea-Present > Periphery-Present = None-Present (see Fig. 2). The pattern of results was similar for set size 3 and 4, whereas the overall performance was much worse with set size 4 than 3 (tracking 4 faces appears to be difficult).

The pattern of results is completely consistent with the serial tracking model, indicating that at each moment the observers were only extracting useful information from the face that was currently foveated. Thus, presenting the identity of the foveated face led to an equal performance of presenting all the facial identities, while masking the foveated facial identity led to as poor performance as when no identity information was shown, even though all the facial identities in the periphery were visually present.

3. Testing the identifiability of different types of objects

Experiment 1 provided clear evidence for serial tracking during MIT. In contrast, some previous studies have provided evidence for parallel rather than serial tracking (Howe & Ferguson, 2015). One notable difference between the experiments is the type of objects to be tracked. While faces were used in Experiment 1, Howe and Ferguson (2015) used simple color discs. Probably faces can only be identified within the small region corresponding to the foveal vision, while colors can be easily identified across broad regions in the visual field, both in the foveal and peripheral vision. If so, faces are tracked one at a time, while multiple colors may be tracked in parallel.

Hence, we conducted an experiment to measure the identifiability of facial images and color discs in the visual field by requiring participants to identify objects presented at various distances from the fovea. Additionally, we also measured the identifiability of line drawings of daily objects (see Experiment 3 for the tracking results), which have been frequently used in previous research on dynamic tracking and visual cognition (Elmore, Passaro, & Wright, 2013; Oksama & Hyönä, 2008). The assumption is that facial configurations can be identified at the fovea but not much so in the periphery, while colors can be identified both in the foveal and the peripheral vision, and the identifiability of the line drawings would lie somewhere between the faces and colors.

A group of 18 participants (2 males, 16 females; age mean = 22.6 years) were recruited from University of Turku. None of them participated in the tracking experiments of the present study. The experiment adopted a 3 (object type: faces, colors, line drawings) × 3 (distance-to-fovea: 2.5°, 5.0°, or 7.5°) within-participant design. The same 18 facial images used in Experiment 1 were presented as the targets and probes. Nine different colors were used: red, green, blue, yellow, cyan, magenta, brown, silver, and white. Eighteen line drawings of objects (e.g., shoe, lobster, rooster, watch) were selected from a standardized set of black-and-white line drawings (Snodgrass & Vanderwart, 1980). The size of the facial images was the same as that in the tracking experiment (i.e., 1.7° × 2.3° at the 70-cm viewing distance). The diameter of each color disc subtended a visual angle of 2.0°, and each image of line drawings subtended a visual area of 2.0° × 2.0°, same as that in the tracking experiments below (Experiment 2 and 3).

At the beginning of each trial, an object was presented at the center of the screen. Participants pressed the space bar when finished looking at the object, which triggered the disappearance of the central object. Subsequently, a cross was presented at the center, and a probe object was presented for 150 ms at a randomly selected location 2.5°, 5.0°, or 7.5° away from the center. The 2.5° distance was considered to be near the foveal vision, while the 5.0°, and 7.5° distance tested peripheral vision. One MIT study showed that the majority of fixations landed no further than 4° from individual targets (Li, Oksama, & Hyönä, 2018a, 2018b). Thus, the eccentricity of 5° should cover the targets observers are visiting. The eccentricity of 7.5° corresponds to a circular area of 15° diameter. It is close to the size of display area in our MIT task (24.3° × 17.0°), so that it encompasses the majority of the targets during tracking. The probe was identical to the target in half of the trials, while it was a different object of the same type in the other half of trials. With faces, the probe and the target were always of the same gender. This is consistent with the setting in the tracking experiment (i.e., Experiment 1). The probe was then replaced by a pattern mask that covered the location for another 150 ms, after which it disappeared, leaving only the cross at the center. Participants were required to keep fixating at the center throughout the procedure and judge whether or not the probed object was the same as the target object presented at the screen center. The pattern masks were mosaic images of each type of object and of the same shape as the objects. That is, in the color disc condition the mask was a disc containing multiple colors, in the line drawing condition a drawing containing random lines, and in the face condition an oval of a scrambled face averaged from all the facial images.

Each participant performed three experimental blocks separately for identifying the three types of objects (i.e., facial images, color discs, line drawings), with the order of the blocks balanced between participants. The three distance-to-fovea conditions were randomly mixed within each block, with 24 trials in each condition, resulting in 72 experimental trials in each block. Each object was presented an equal number of times, while its appearance in each distance-to-fovea condition was randomized. Participants performed 24 practice trials prior to each block. In each practice trial, the target image and the distance-to-fovea value were randomly selected.

The results were consistent with our assumptions (Fig. 3a). A 3 (object type: color, drawing, face) × 3 (distance-to-fovea: 2.5°, 5.0°, 7.5°) repeated-measures ANOVA on the identification accuracy yielded significant main effects of object type and distance-to-fovea, and a significant Object Type × Distance-to-Fovea interaction ($F(2, 34) = 72.546$, $p < .001$, $\eta_p^2 = 0.810$; $F(2, 34) = 50.002$, $p < .001$, $\eta_p^2 = 0.746$; $F(4, 68) = 11.443$, $p < .001$, $\eta_p^2 = 0.402$, respectively). The results indicate that the identifiability of the three types of objects differed from each other, and that their identifiability varied differently as a function of distance to fovea. The identification of colors was close to ceiling (around 90%) both in the foveal and peripheral vision ($M = 91.0\%$, $SD = 5.9\%$; $M = 87.5\%$, $SD = 6.7\%$; $M = 90.7\%$, $SD = 4.6\%$; for 2.5°, 5.0°, 7.5°, respectively). In contrast, the

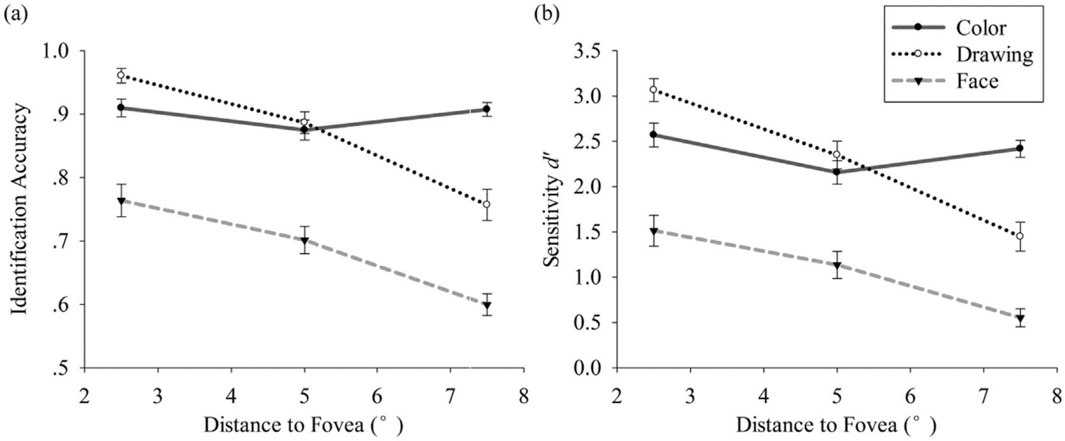

Fig. 3. The identification accuracy (a) and sensitivity *d′* (b) for colors, drawings, and faces at different distances from the fovea. The error bars represent ± 1 S.E.

identification of the drawings and faces decreased gradually as the objects were presented further away from the fovea. For the drawings, the identification was close to ceiling near the fovea and decreased in the peripheral vision but remained at a fairly accurate level, whereas the identification of the faces was fairly accurate near the fovea and dropped to poor levels in the peripheral vision.

In addition, we adopted the signal detection theory to compute the sensitivity for each type of object at each distance. The pattern of results was similar to that of the accuracy (see Fig. 3b). A 3 (object type: color, drawing, face) × 3 (distance-to-fovea: 2.5°, 5.0°, 7.5°) repeated-measures ANOVA on the sensitivity measure (*d′*) yielded significant main effects of object type and distance-to-fovea, and a significant Object Type × Distance-to-Fovea interaction ($F(2, 34) = 49.544$, $p < .001$, $\eta_p^2 = 0.745$; $F(2, 34) = 44.097$, $p < .001$, $\eta_p^2 = 0.722$; $F(4, 68) = 11.415$, $p < .001$, $\eta_p^2 = 0.402$, respectively). Simple effect analyses showed that the sensitivity for color discs was at a high level at each distance ($M = 2.57$, $SD = 0.56$, $M = 2.16$, $SD = 0.55$, $M = 2.42$, $SD = 0.40$, for 2.5°, 5.0°, 7.5°, respectively), and there was neither a significant difference between 2.5° and 7.5° ($p = 1.000$), nor between 5.0° and 7.5° ($p = .165$), albeit a small but significant drop from 2.5° to 5.0° ($p = .021$). The sensitivity for line drawings gradually dropped from the fovea to periphery ($M = 3.07$, $SD = 0.54$; $M = 2.35$, $SD = 0.65$; $M = 1.45$, $SD = 0.67$), with the differences being significant between each pair ($ps < .001$). The sensitivity for facial images was low at all distances ($M = 1.51$, $SD = 0.72$; $M = 1.14$, $SD = 0.64$; $M = 0.55$, $SD = 0.42$). The difference was significant between 2.5° and 7.5° ($p = .001$), between 5.0° and 7.5° ($p = .035$), while not significant between 2.5° and 5.0° ($p = .143$). Overall, the pattern of sensitivity resembles that of the accuracy, indicating that the results were not confounded by response bias.

The results demonstrate that color discs can be identified accurately in both foveal and peripheral vision, while the facial images can be identified well in foveal vision but are nearly unidentifiable in peripheral vision. The line drawings lie somewhere between the color discs and the facial images in terms of identifiability.[1]

4. Experiment 2: Tracking multiple color discs

The identification accuracy of objects may be related to the manner they are tracked. In Experiment 1, the tracking of facial images, which can be identified in foveal vision but are nearly unidentifiable in the periphery, showed a serial pattern. In Experiment 2, we examined whether tracking color discs that can be accurately identified in both foveal and peripheral vision would show evidence for more parallel tracking, as observed by Howe and Ferguson (2015).

4.1. Method

Experiment 2 was identical to Experiment 1, except that color discs were used as target objects instead of facial images. The nine color discs used in the identifiability test were used in the tracking experiment. Each color disc appeared in one of the nine colors: red, green, blue, yellow, cyan, magenta, brown, silver, and white. The diameter of the color disc subtended 2.0° at a 70-cm viewing distance. Forty-two participants from University of Turku, Finland took part in the experiment in exchange of course credit. None of them participated in the other experiments of the present study. All participants had normal or corrected-to-normal vision, and all provided informed consent. The data of 2 participants were excluded due to calibration problems (1 participant) or program failure (1 participant). The data of the remaining 40 participants (12 males, 28 females) were included in analyses: 21 in tracking 3 color discs and 19 in tracking 4 color discs. The mean age of the participants was 24.4 years.

4.2. Results

A 2 (set size: 3, 4) × 4 (presentation mode: All-Present, Fovea-Present, Periphery-Present, None-Present) repeated-measures ANOVA yielded significant main effects of presentation mode and set size ($F(3, 114) = 21.551$, $p < .001$, $\eta_p^2 = 0.362$; $F(1, 38) = 8.602$, $p = .006$, $\eta_p^2 = 0.185$), as well as a significant Set Size × Presentation Mode interaction, $F(3, 114) = 10.927$, $p < .001$, $\eta_p^2 = 0.223$ (see Fig. 4).

We then separately examined the effect of presentation mode for set size 3 and 4. When tracking 3 color discs, the accuracies were around 90% in all conditions ($M = 89.6\%$, $SD = 7.2\%$, $M = 90.9\%$, $SD = 6.4\%$, $M = 89.0\%$, $SD = 7.9\%$, $M = 86.0\%$, $SD = 7.6\%$, for the All-, Fovea-, Periphery-, None-Present condition, respectively), i.e., nearly close to ceiling. Planned pair-wise comparisons showed that the differences between the conditions were not significant ($ps > .385$), except that the difference between Fovea-Present and None-Present conditions approached significance ($p = .067$). The near-ceiling effect indicates that tracking 3 colors is well within the observers' tracking capacity.

When tracking 4 color discs, the pattern was All-Present = Periphery-Present > Fovea-Present > None-Present ($M = 86.3\%$, $SD = 9.5\%$, $M = 84.5\%$, $SD = 12.6\%$, $M = 77.6\%$,

[1] In this task, participants need to process target objects and probes of the same type and then judge whether they are identical or different. Strictly speaking, it is not sure to what level the participants process the objects. Nevertheless, considering that discriminating similar objects are closely intertwined with identifying the objects (Palmeri & Gauthier, 2004), and it is the case both in this task and the MIT task, we use the terms identify and identifiability throughout the paper.

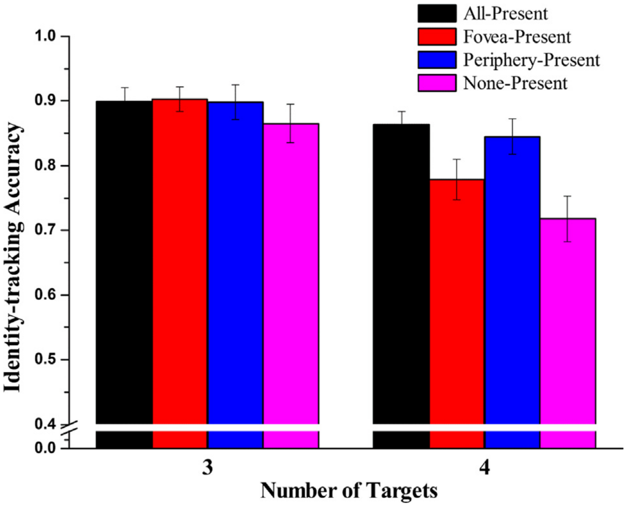

Fig. 4. Identity tracking accuracy for color discs. The error bars represent ± 1 S.E.

SD = 14.5%, M = 71.7%, SD = 16.3%, respectively). Planned pair-wise comparisons showed that the accuracies in the All-Present condition and Periphery-Present condition were similar to each other (p = 1.000, Cohen's d = 0.288), and both were higher than that in the Fovea-Present and None-Present conditions (ps < .002, Cohen's ds > 0.820). The accuracy in the Fovea-Present condition was higher than that in the None-Present condition (p = .019, Cohen's d = 0.699).

The pattern of results is consistent with parallel tracking. That is, what matters for tracking is how many color identities are presented in the visual field at each moment, no matter whether or not they were foveally attended. The tracking performance improved as the number of visually available color identities increased (0, 1, and 3 in the None-Present, Fovea-Present, Periphery-Present conditions, respectively), though seeing 4 colors (3 in the periphery + 1 at the fovea) did not lead to a significantly better performance than seeing 3 colors (i.e., All-Present = Periphery-Present). Possibly the benefit for having all identities visually present becomes less robust, as the number of targets increases.

5. Experiment 3: Tracking multiple line drawings

Experiment 1 and 2 yielded a distinct pattern of results, suggesting that the manner of tracking may differ for different types of objects that vary in identifiability. Facial images are tracked in a serial manner, while color discs are tracked in parallel. A feasible inference is that tracking objects that lie in between faces and colors in terms of identifiability may exhibit a pattern between purely serial and purely parallel. In Experiment 3, we examined this possibility by using line drawings as target objects in the tracking task.

5.1. Method

Experiment 3 was identical to Experiment 1, except that line drawings were used as stimuli. The eighteen line drawings used in the identifiability test were used in the experiment. Each image subtended a visual angle of 2.0° × 2.0°. Forty participants (12 males, 28 females) from University of Turku, Finland took part in the experiment in exchange of course credit. None of them participated in the other experiments of the present study. All participants had normal or corrected-to-normal vision, and all provided informed consent. Twenty participants tracked 3 line drawings, while the other 20 tracked 4 drawings. The mean age of the participants was 26.4 years.

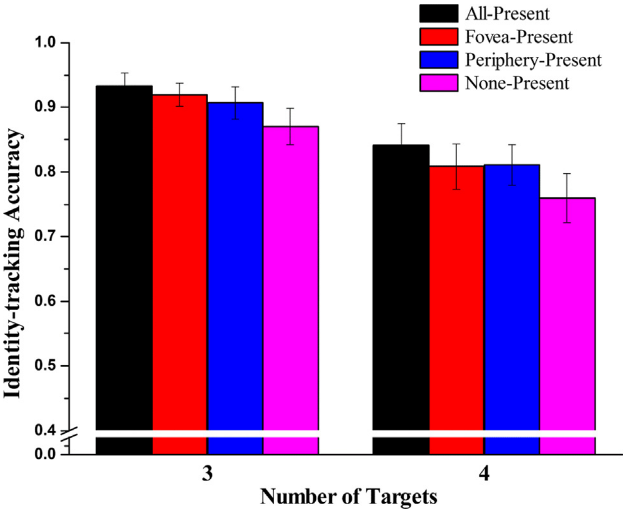

Fig. 5. Identity tracking accuracy for line drawings. The error bars represent ± 1 S.E.

5.2. Results

A 2 (set size: 3, 4) × 4 (presentation mode: All-Present, Fovea-Present, Periphery-Present, None-Present) repeated-measures ANOVA was conducted. The result yielded a significant main effect of set size, $F(1, 38) = 8.356$, $p = .006$, $\eta_p^2 = 0.180$, showing that the performance was higher for tracking 3 than 4 drawings (M = 90.7%, SD = 7.0% vs. M = 80.5%, SD = 15.5%). The main effect of presentation mode was also significant, $F(3, 114) = 15.451$, $p < .001$, $\eta_p^2 = 0.289$, while the Set Size × Presentation Mode interaction was not significant, $F(3, 114) = 0.434$, $p = .729$, $\eta_p^2 = 0.011$ (see Fig. 5).

Pair-wise comparison between the presentation modes showed that the overall pattern was All-Present ≥ Fovea-Present = Periphery-Present > None-Present (M = 88.7%, SD = 12.1%, M = 86.4%, SD = 12.5%, M = 85.9%, SD = 11.8%, M = 81.5%, SD = 14.8%, respectively). The accuracy in the All-Present condition was statistically marginally higher than that in the Fovea-Present condition (p = .060, Cohen's d = 0.427), and significantly higher than that in the Periphery-Present (p = .030, Cohen's d = 0.477) and the None-Present condition (p < .001, Cohen's d = 0.945). The accuracies in the Fovea-Present condition and the Periphery-Present condition were similar to each other (p = 1.000, Cohen's d = 0.081). The accuracy in the None-Present condition was significantly lower than that in all the other conditions (ps < .005, Cohen's ds > 0.574).

The pattern of results is not completely serial, since seeing line drawings in the periphery resulted in a better tracking performance than seeing no identity (Periphery-Present > None-Present). The pattern is not completely parallel either, since seeing multiple drawings in the periphery did not lead to better performance than seeing only one object at the fovea (Periphery-Present = Fovea-Present). Instead, the pattern appears like a combination of parallel and serial tracking. The target identities both in the foveal as well as in the peripheral vision are processed, yet the identities in the periphery are processed less precisely than the foveated one. Thus, the target identities presented in the periphery are helpful for tracking, but not as helpful as those at the fovea. The pattern of results was similar for set size 3 and 4, although the overall performance was lower with set size 4 than 3.

Taken together, the data on performance accuracy obtained in Experiments 1–3 indicate that multiple identity tracking is modulated by the relative success in identifying objects in the peripheral vision. For objects that need to be identified by foveal vision (e.g., faces), tracking appears serial. On the other hand, for objects that can be identified perfectly by both the foveal and peripheral vision (e.g., colors), tracking appears parallel. For objects that can be identified in the peripheral vision but to a lesser degree than that in the foveal vision (e.g., line drawings), tracking shows a manner between purely serial and purely parallel. Moreover, an increase in target set-size led to poorer tracking performance overall, but the manner of tracking remained the same, except that the performance of tracking three color discs was at ceiling under all presentation conditions. In addition, it is noteworthy that for all types of objects, when no identities were present during tracking, tracking was poorer than in the other conditions but still at a substantially high level. This result suggests that tracking does not rely solely on the visual information sampled during tracking, but also on the representations maintained in short-term memory.

6. Eye movements during tracking of different types of objects

In order to further understand the mechanism of MIT, we analyzed the eye movement data during tracking. Eye movements are closely associated with visual sampling and shift of attention and working memory resources (Bays & Husain, 2008; Shao et al., 2010), and can shed light on parallel vs. serial processing (Young & Hulleman, 2013). In the following analyses, we firstly examined how eye movements are affected by the presence of different types of object identities in the visual field (Section 6.1), then examined the parallel vs. serial patterns of the eye movements (Section 6.2), followed by the exploration of the relationship between eye movements and working memory during tracking (Sections 6.3 and 6.4).

Raw eye movement data (i.e., x and y coordinates sampled at each time point) were used for the analyses. We computed the distance from the sampled eye gaze position to the center of each object on the screen, so as to find the closest object to the current gaze position. If the distance between the closest object and the gaze position was shorter than 2.5°, the gaze was considered to be on this object; otherwise, the gaze was considered to be on blank area between the moving objects. A visit to an object or the blank area was counted as valid when the eyes had continuously dwelled on it for more than 80 ms.

6.1. *The influence of identity presentation on eye movements*

In order to gain a deeper understanding of how object identities are processed during MIT, we examined the eye behavior in the different identity presentation conditions. Identifying faces requires high resolution visual information, while low resolution information is sufficient for identifying color discs. In human visual system, high resolution information can be sampled only via the small central region corresponding to the fovea, while low resolution information can be sampled in the broad visual field in the periphery (Anstis, 1974; Land & Tatler, 2009). Thus, we predicted that when faces as opposed to color discs are presented in the periphery, participants may be more likely to move the eyes to the objects in order to sample high resolution information. Moreover, once the facial identities are foveated, the eye gaze may dwell there for a relatively long time for sampling the visual information.

We calculated the frequency and dwell time of the eye visits to targets for each type of objects in each presentation mode. We then conducted 3 (object type: color disc, drawing, face) × 2 (set size: 3, 4) × 4 (presentation mode: All-, Fovea-, Periphery-, None-Present) ANOVAs on the frequency of visits and dwell time, with presentation mode as the within-participants variable and object type and set size as the between-participants variables. Overall, as shown in Fig. 6, the patterns of eye movements differ between different types of objects. The ANOVAs yielded significant Object Type × Presentation Mode interactions (see Tables 1 and 2 for details), indicating that the presence/absence of different types of object identities influences the eye movements differently.

To further examine the interaction, we conducted separate one-way (object type: color disc, drawing, face) ANOVAs for each presentation mode. When no identities were shown (i.e., the None-Present condition), there was no significant difference between color discs, drawings, and faces in the frequency of target visits, $F(2,116) = 0.936$, $p = .395$, $\eta_p^2 = 0.016$, or in the dwell time on targets, $F(2,116) = 1.364$, $p = .260$, $\eta_p^2 = 0.023$.

When the identities were present in the periphery (i.e., the Periphery-Present condition), participants made more visits to targets when the targets were drawings or faces than color discs ($M = 12.1$, $SD = 3.9$, and $M = 11.9$, $SD = 3.9$, vs. $M = 9.0$, $SD = 3.4$; $ps < .003$), while there was no difference between drawings and faces ($p = 1.000$). The dwell time on targets was significantly longer when tracking color discs than drawings ($M = 593$ ms, $SD = 345$, vs. $M = 441$ ms, $SD = 169$; $p = .023$), while there was no difference between drawings and faces ($M = 441$ ms, $SD = 169$, vs. $M = 495$ ms, $SD = 198$; $p = 1.000$) or between color discs and faces ($p = .258$). The results indicate that when the identities of faces and line drawings are present in the periphery, in comparison with the color discs, observers are more likely to terminate the current fixation and move the eyes to other peripheral targets.

When identities were present at the fovea (i.e., Fovea-Present), participants made fewer visits to targets when the targets were faces in comparison with color discs ($M = 6.8$, $SD = 2.8$, vs. $M = 9.0$, $SD = 3.3$; $p = .008$), and marginally significantly ($M = 6.8$, $SD = 2.8$, vs. $M = 8.5$, $SD = 3.3$; $p = .058$) fewer visits to faces in comparison drawings, while there was no difference between color discs and drawings ($p = 1.000$). The dwell time on targets was significantly longer when the targets were faces rather than color discs or drawings ($M = 943$ ms, $SD = 412$, vs. $M = 564$ ms, $SD = 302$, and $M = 604$ ms, $SD = 245$; $ps = .001$), while there was no difference between color discs and drawings ($p = 1.000$). The results indicate that when facial identities are present at the fovea, observers are inclined to dwell longer on the foveated targets and make fewer eye movements, in comparison with the color discs and line drawings.

In addition, the ANOVA results showed that eye movements were also affected by set size. There was a significant Set Size × Presentation Mode interaction in the ANOVA in the frequency of visits to targets (see Table 1). Thus, we separately conducted one-way ANOVAs for each set size with the presentation mode as the within-participants variable. The results showed that the effect of presentation mode was significant for both set size 3 and 4 ($ps < .001$). The interaction was due to the frequency of target visits being greater for set-size 3 in the Periphery-Present condition than in the All-Present condition ($p = .003$), and similar in the Fovea-Present and the None-Present conditions ($p = 1.000$) (i.e., Periphery > All > Fovea = None; $M = 10.8$, $SD = 4.1$, $M = 9.9$, $SD = 3.1$, $M = 7.9$, $SD = 3.3$, $M = 8.0$, $SD = 3.1$, respectively), whereas for set size 4 the frequency was similar in the Periphery-Present and All-Present conditions ($p = 1.000$), and greater in the Fovea-Present condition than in the None-Present condition ($p = .008$) (i.e., Periphery = All > Fovea > None; $M = 11.1$, $SD = 3.9$, $M = 10.9$, $SD = 3.3$, $M = 8.4$, $SD = 3.3$, $M = 7.8$, $SD = 2.8$, respectively). The results suggest that when there are more targets to track, the observers are likely to make more eye movements to targets, particularly when the target identity is visually available while foveated.

The ANOVA on the dwell time yielded a significant main effect of set size and a significant Object Type × Set Size interaction (see Table 2). Simple effect analyses showed that when tracking colors and faces the dwell time decreased as the set size increased from 3 to 4 ($M = 698$ ms, $SD = 387$, vs. $M = 429$ ms, $SD = 92$; $p < .001$ for colors; $M = 761$ ms, $SD = 334$, vs. $M = 607$ ms, $SD = 369$; $p = .039$ for faces), yet when tracking drawings the dwell time did not differ between set size 3 and 4 ($M = 523$ ms, $SD = 184$, vs. $M = 540$ ms, $SD = 249$; $p = .810$). It is not clear why there was no difference for drawings, since typically the dwell time on targets decreases as a function of set-size (Oksama & Hyönä, 2016).

Overall, the results are consistent with our predictions, showing that when faces and drawings were present in the periphery, participants were more likely to terminate the current fixation and move the eyes to

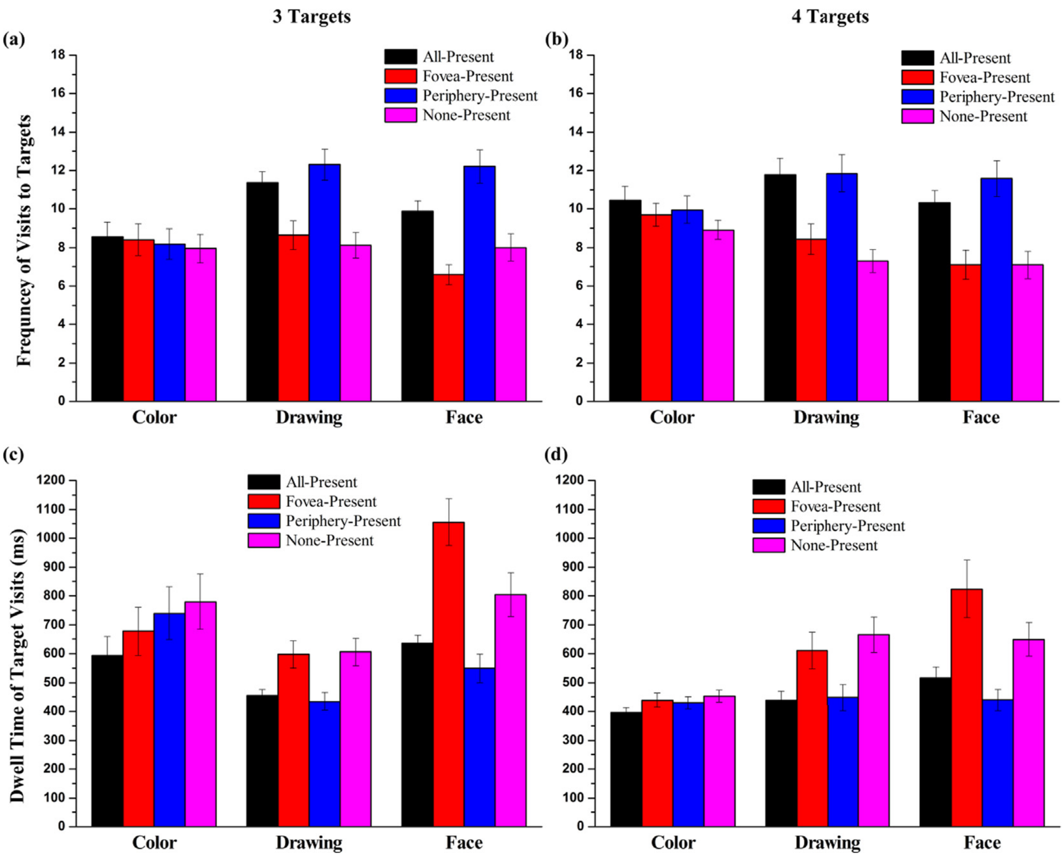

Fig. 6. Frequency of visits to targets (a, b) and dwell time of target visits (c, d) when tracking 3 or 4 targets of each object type in each presentation mode.

Table 1
ANOVA results for the frequency of visits to targets.

	df	*MS*	*F*	*p*	η_p^2
Main effects					
Object Type	2	44.480	1.163	.316	0.020
Set Size	1	15.224	0.398	.529	0.004
Presentation Mode	3	362.845	198.837	< .001	0.638
Two-way interactions					
Object Type × Set Size	2	38.087	0.996	.373	0.017
Object Type × Presentation Mode	6	71.682	39.281	< .001	0.410
Set Size × Presentation Mode	3	8.999	4.931	.005	0.042
Three-way interaction					
Object Type × Set Size × Presentation Mode	6	1.533	0.840	.520	0.015

Note: *MS* = Mean Square.

Table 2
ANOVA results of dwell time on targets.

	df	*MS*	*F*	*p*	η_p^2
Main effects					
Object Type	2	1,016,666	4.751	.010	0.078
Set Size	1	2,168,130	10.133	.002	0.082
Presentation Mode	3	1,693,180	68.021	< .001	0.376
Two-way interactions					
Object Type × Set Size	2	827,726	3.868	.024	0.064
Object Type × Presentation Mode	6	554,452	22.274	< .001	0.283
Set Size × Presentation Mode	3	11,512	0.462	.647	0.004
Three-way interactions					
Object Type × Set Size × Presentation Mode	6	46,319	1.861	.112	0.032

Note: *MS* = Mean Square.

a peripheral target; when faces were present at the fovea, participants were more likely to dwell longer on the target. In addition, as the number of targets increased from 3 to 4, participants typically made more frequent but shorter visits to the targets.

Furthermore, we quantified the participants' inclination of moving the eyes to the peripheral targets for visually sampling the identity information by calculating the difference between the Periphery-Present and None-Present condition. When the identities were drawings and faces, the difference was positive in the frequency of visits to targets (around 4.3) and negative in the dwell time on targets (around −200 ms). When the identities were colors, the differences were close to 0 in both measures (see left panels in Fig. 7a and b), and significantly differed from that when the identities were drawings and faces ($M = 0.6$, $SD = 1.2$, vs. $M = 4.4$, $SD = 2.3$, and $M = 4.3$, $SD = 2.0$; $ps < .001$, in frequency; $M = -31$ ms, $SD = 82$, vs. $M = -194$ ms, $SD = 166$, and $M = -233$ ms, $SD = 174$; $ps < .001$, in dwell time), while there were no significant differences between the drawings and faces ($ps > .320$). The results indicate that the observers are much more inclined to disengage from the currently foveated target and move the eyes to the peripheral targets when drawings and faces are present in the periphery, whereas such inclination is fairly weak when colors are present in the periphery. This is likely due to the fact (see Fig. 3) that the drawings and faces can be identified by the foveal vision but not much so in the peripheral visual field, while colors can be accurately identified both in the foveal and the peripheral visual field.

Analogously, by calculating the difference between the Fovea-Present condition and the None-Present condition (right panels in Fig. 7a and b), we quantified the participants' inclination of continually dwelling on the current foveated identity. This difference revealed a significant increase in dwell time ($M = 215$ ms, $SD = 288$; $p < .001$) along with a significant decrease in the frequency of target visits for tracking faces ($M = -0.8$, $SD = 1.7$; $p = .012$); conversely, for tracking

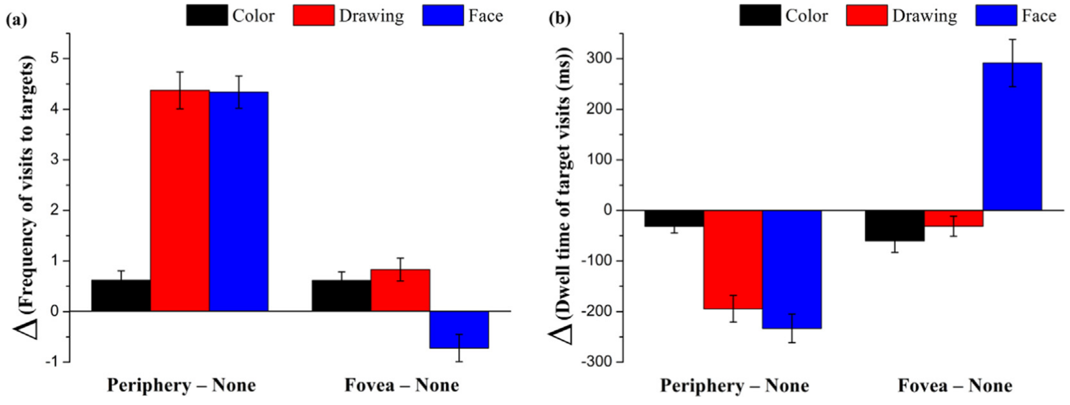

Fig. 7. The difference in (a) the frequency of visits to targets and (b) the dwell time of target visits between the Periphery-Present condition and the None-Present condition, and between the Fovea-Present condition and the None-Present condition.

Table 3
Number of visited targets, and frequency of visits to targets, blank area, and distractors in the All-Present condition. Standard deviations are in parentheses.

Object Type	Set Size	Number of Visited Targets	Frequency of Visits			Percentage of Total Dwell Time (%)		
			Target	Blank	Distractor	Target	Blank	Distractor
Color disc	3	2.6 (0.4)	8.6 (3.5)	3.5 (1.5)	0.7 (0.3)	75.4 (14.4)	19.8 (11.6)	4.8 (3.2)
	4	3.3 (0.4)	10.5 (3.1)	4.2 (1.7)	1.8 (0.5)	70.7 (13.4)	19.8 (10.4)	9.5 (3.6)
Line drawing	3	2.9 (0.2)	11.4 (2.6)	1.9 (1.6)	0.4 (0.2)	88.6 (12.0)	9.1 (10.5)	2.4 (1.8)
	4	3.6 (0.7)	11.8 (3.8)	2.0 (1.8)	1.4 (0.7)	81.1 (18.0)	11.2 (15.5)	7.7 (4.1)
Facial image	3	2.9 (0.1)	9.9 (2.4)	0.9 (0.6)	0.5 (0.3)	93.0 (5.3)	4.3 (4.8)	2.6 (1.5)
	4	3.6 (0.6)	10.3 (2.8)	1.3 (1.1)	1.4 (0.4)	84.1 (1.3)	7.4 (12.1)	8.5 (3.0)

colors, there was a significant decrease in dwell time ($M = -60$ ms, $SD = 143$; $p = .011$) along with a significant increase in frequency ($M = 0.6$, $SD = 1.1$; $p = .001$); for tracking drawings, there was no significant change in dwell time ($M = -30$ ms, $SD = 125$; $p = .123$) but a significant increase in frequency ($M = 0.8$, $SD = 1.4$; $p = .001$). The results indicate that the participants make fewer eye movements and dwelled for extra time on targets to extract visual information of the faces but not that of color discs or drawings.

In sum, the results in this section showed that eye visits to individual targets increased in frequency and dwell time when facial identities rather than color discs were present. The results are consistent with the low identifiability of faces in the peripheral vision and the serial tracking of faces revealed by the performance accuracy, as well as with the evidence from performance accuracy supporting parallel tracking of colors that can be identified perfectly both in the foveal and peripheral vision.

6.2. Serial vs. parallel tracking manifested in eye visits to individual targets and blank area between targets

In this section, we report the overall pattern of eye movements for tracking different types of objects: whether eye movements demonstrate a predominately serial pattern (i.e., switching from one target to another) when tracking faces, and a predominately parallel pattern (i.e., dwelling on blank areas between targets) when tracking colors. During a fixation the observer may carry out parallel processing of objects around the fixation, whereas shifting the fixation from one object to another is coupled with serial shifting of attention (Li, Oksama, & Hyönä, 2018a, 2018b; Motter & Holsapple, 2007; Young & Hulleman, 2013). When looking at static scenes, observers are predisposed to gaze at the centroid of the objects when trying to monitor multiple objects (Fluharty, Jentzsch, Spitschan, & Vishwanath, 2016), while the gaze lands more frequently on individual objects as visual search turns from parallel to serial (Williams, Reingold, Moscovitch, & Behrmann, 1997; Young & Hulleman, 2013). Previous studies on dynamic tracking showed that in MOT tasks observers predominately gaze at the blank areas between targets for tracking the target locations (Fehd & Seiffert, 2010), whereas in MIT tasks observers predominately gaze at individual targets for tracking the identities of line drawings (Oksama & Hyönä, 2016). Therefore, we hypothesized that in the present study the eye movement pattern would also be serial when tracking the faces and line drawings, while it might switch to parallel when the MIT task becomes parallel, as is the case with color discs.

We calculated the number of targets that were visited during tracking, and computed the frequency of eye visits to targets and blank areas between targets, as well as the percentage of dwell time spent on targets and blank areas. In the analyses in this section, the eye movement data from the All-Present condition (i.e., standard MIT) was used.

The results showed that for all types of objects, the majority of the targets was visited by the eyes, the majority of the eye visits was directed to targets, and the majority of the time was spent on visiting targets (see Table 3). The blank areas were visited less frequently and for less time than the individual targets, and the distractors were seldom visited. Even when tracking color discs, 2.6 out of 3, and 3.3 out of 4 targets were visited, over 2/3 of the visits were directed to targets, and over 3/4 of the time was spent on targets. This is inconsistent with our prediction, indicating that the eye movement pattern during MIT is generally serial for all types of objects.

Despite the finding that the eye movement pattern was found to be generally serial for all types of objects, subsequent analyses showed that the eye movement pattern was relatively parallel when tracking colors as opposed to drawings and/or faces, as evidenced by fewer targets visited, fewer visits to targets and more frequent visits to blank areas. This is in line with our hypothesis derived from the results obtained via the gaze-contingent display change paradigm. The results are reported below in more detail.

We conducted a 3 (object type: color, drawing, face) × 2 (set size: 3, 4) ANOVA on the number of visited targets. The results showed significant main effects of object type ($F(2, 113) = 4.887$, $p = .009$, $\eta_p^2 = 0.080$) and set size ($F(1, 113) = 64.090$, $p < .001$, $\eta_p^2 = 0.362$), while the interaction was not significant, $F(2, 113) = 0.183$, $p = .833$, $\eta_p^2 = 0.003$. Planned pair-wise comparisons showed that more targets were visited when set size was 4 rather than 3 ($M = 3.5$, $SD = 0.6$, vs. $M = 2.8$, $SD = 0.3$; $p < .001$). More importantly, compared with tracking colors, more targets were visited when tracking drawings ($M = 3.2$, $SD = 0.6$, vs. $M = 3.0$, $SD = 0.6$; $p = .036$) and when tracking faces ($M = 3.3$, $SD = 0.5$, vs. $M = 3.0$, $SD = 0.6$; $p = .016$), while there was no significant difference between tracking drawings and faces ($p = 1.000$).

A 3 (object type: color, drawing, face) × 2 (set size: 3, 4) ANOVA on the frequency of visits to targets yielded a significant main effect of object type, $F(2, 113) = 4.628$, $p = .012$, $\eta_p^2 = 0.076$, while the main effect of set size and the interaction between object type and set size were not significant ($F(1, 113) = 2.580$, $p = .111$, $\eta_p^2 = 0.022$; $F(2, 113) = 0.745$, $p = .477$, $\eta_p^2 = 0.013$, respectively). Planned pair-wise comparisons showed that the frequency of visits to targets was smaller when tracking colors than tracking drawings ($M = 9.5$, $SD = 3.4$, vs. $M = 11.6$, $SD = 3.2$; $p = .011$), whereas there was no significant difference between tracking colors and tracking faces ($M = 9.5$, $SD = 3.4$, vs. $M = 10.1$, $SD = 2.6$; $p = 1.000$) or between tracking drawings and tracking faces ($M = 11.6$, $SD = 3.2$, vs. $M = 10.1$, $SD = 2.6$; $p = .116$).

A 3 (object type: color, drawing, face) × 2 (set size: 3, 4) ANOVA on the frequency of visits to blank areas yielded a significant main effect of object type, $F(2, 113) = 36.711$, $p < .001$, $\eta_p^2 = 0.394$, while the main effect of set size and the interaction between object type and set size were not significant ($F(1, 113) = 2.396$, $p = .124$, $\eta_p^2 = 0.021$; $F(2, 113) = 0.421$, $p = .658$, $\eta_p^2 = 0.007$, respectively). Planned pair-wise comparisons showed that there were significantly more visits to blank area when tracking color discs than tracking line drawings or faces ($M = 3.8$, $SD = 1.6$, vs. $M = 2.0$, $SD = 1.7$, and $M = 1.1$, $SD = 0.9$; $ps < .001$), and when tracking drawings than faces ($M = 2.0$, $SD = 1.7$, vs. $M = 1.1$, $SD = 0.9$; $p = .029$).

A 3 (object type: color, drawing, face) × 2 (set size: 3, 4) ANOVA on the percentage of dwell time on targets yielded significant main effects of object type and set size ($F(2, 113) = 14.559$, $p < .001$, $\eta_p^2 = 0.205$; $F(1, 113) = 8.261$, $p = .005$, $\eta_p^2 = 0.068$, respectively), while the interaction was not significant, $F(2, 113) = 0.254$, $p = .776$, $\eta_p^2 = 0.004$. Pair-wise comparisons showed a smaller percentage of dwell time on targets when tracking colors than tracking drawings and faces ($M = 73.2\%$, $SD = 14.0\%$, vs. $M = 84.8\%$, $SD = 15.6\%$, and $M = 88.7\%$, $SD = 11.0\%$; $ps < .001$), while there was no difference between drawings and faces ($p = .646$). Moreover, the percentage of dwell time on targets decreased as the set-size increased ($M = 85.7\%$, $SD = 13.5\%$, vs. $M = 78.6\%$, $SD = 16.0\%$).

A 3 (object type: color, drawing, face) × 2 (set size: 3, 4) ANOVA on the percentage of dwell time on blank areas yielded a significant main effect of object type, $F(2, 113) = 15.933$, $p < .001$, $\eta_p^2 = 0.220$, while neither the main effect of set size nor the interaction between object type and set size was significant ($F(1, 113) = 0.713$, $p = .400$, $\eta_p^2 = 0.006$; $F(2, 113) = 0.193$, $p = .824$, $\eta_p^2 = 0.003$, respectively). Pair-wise comparisons showed that more time was spent on fixating blank areas when tracking colors than drawings and faces ($M = 19.8\%$, $SD = 10.9\%$, vs. $M = 10.2\%$, $SD = 13.1\%$, and $M = 5.8\%$, $SD = 9.1\%$; $ps < .001$), while there was no difference between drawings and faces ($p = .279$).

Taken together, the analyses in this section showed that the eye movement pattern was relatively parallel when tracking colors, yet it was predominantly serial regardless of the object type with the majority of the visits paid to individual targets and the majority of the time spent on individual targets. This would not be the case if eye visits to targets were employed only when high resolution information needs to be sampled for target identification. Rather, there seems to be an inherent serial process in MIT. Considering that working memory is actively involved in MIT (Oksama & Hyönä, 2004, 2008), and eye movements are intertwined with attention and working memory (Ferreira, Apel, & Henderson, 2008; Theeuwes, Belopolsky, & Olivers, 2009; Van der Stigchel & Hollingworth, 2018), probably eye visits to targets are also employed to facilitate the temporary maintenance of target representations by refreshing each in turn. This assumption is tested in the following explorative analyses.

6.3. Eye movements when no target identities were present during tracking (i.e., the None-Present condition)

In this section, we report explorative analyses of eye movements when no object identity was present in the visual field (i.e., the None-Present condition). The observed eye behavior in this condition presumably reflects refreshing of target representations in short-term memory. The results showed that even in the absence of identity information, participants still frequently visited with their eyes the empty target locations and also for long time (see Table 4). Overall, 65.0%, 65.3%, 65.1% of the visits, and 75.5%, 74.4%, 75.2% of the dwell time was on target locations for tracking colors, line drawings, and faces,

Table 4
Number of visited targets, and frequency of visits to targets, blank area, and distractors in the None-Present condition. Standard deviations are in parentheses.

Object Type	Set Size	Number of Visited Targets	Frequency of Visits			Percentage of Dwell Time (%)		
			Target	Blank	Distractor	Target	Blank	Distractor
Color disc	3	2.5 (0.4)	8.0 (3.4)	2.8 (1.1)	0.5 (0.2)	81.3 (9.4)	15.8 (8.7)	2.9 (1.3)
	4	3.2 (0.5)	8.9 (2.1)	4.0 (1.3)	1.9 (0.7)	69.2 (10.8)	19.8 (8.1)	11.0 (3.9)
Line drawing	3	2.6 (0.4)	8.1 (3.0)	3.2 (1.6)	0.5 (0.2)	78.3 (15.1)	18.7 (13.5)	3.0 (2.0)
	4	2.9 (0.7)	7.3 (2.7)	3.0 (1.6)	1.5 (0.5)	70.6 (16.0)	18.9 (14.6)	10.5 (3.6)
Facial image	3	2.6 (0.4)	8.0 (3.2)	2.5 (0.9)	0.6 (0.3)	83.3 (8.6)	13.2 (8.1)	3.6 (1.5)
	4	2.8 (0.7)	7.1 (3.1)	3.4 (1.2)	1.7 (0.6)	66.6 (14.4)	20.8 (13.1)	12.5 (3.4)

respectively. This phenomenon is reminiscent of the “looking at nothing” phenomenon observed in previous studies (Altmann, 2004; Ferreira et al., 2008; Richardson & Spivey, 2000). The phenomenon demonstrates that looking at the empty location of previously occupied by an object can activate the object representation and benefit the memory performance for it (e.g., Johansson & Johansson, 2014). Analogously, the frequent eye visits to the empty target locations during MIT may serve to refresh the target representations, which is a serial process regardless of the content of the representations. Hence, the eyes switch between targets by default, resulting in a predominately serial pattern of eye movements regardless of the type of objects and the visual presence of object identities. Yet, the default cycle may be modulated by the degree to which sampling of visual information required for successful performance, becoming more serial when high resolution information is required and relatively more parallel when low resolution is sufficient for identity tracking.

6.4. The relationship between target visits and tracking performance

In this section, we report data on the relationship between eye visits to targets and their tracking performance. The analyses are exploratory in nature. Assuming that target visits facilitate the sampling of high resolution information from the targets and the refreshing of the target representations, the visits should be beneficial for the tracking performance. We tested this prediction by examining whether the recently visited targets are tracked better than those that have not been visited for some time.

For the analysis, we extracted the sequence of object visits during tracking in each trial. The sequence is a series of numbers listing the visited objects in the order they were fixated from the beginning to the end of the tracking stage. For example, when tracking 4 targets in a total of 8 moving objects, we may get a sequence of visits such as “3 4 1 2 1 2 4 5 3 1”, in which objects 1, 2, 3, 4 are targets while the other objects are distractors. For each trial, we calculated the number of intervening visits to other objects after each target was visited for the last time. In the example sequence mentioned above, the number of intervening object visits for target 1, 2, 3, 4 are 0, 4, 1, 3, respectively (i.e., 0 means that it was the last target fixated prior to movement termination).

We used logistic regressions to estimate how the response accuracy varies with the number of intervening object visits for each type of objects under each presentation and set size condition. The results are listed in Table 5. The effect of set size was significant for all the objects in all the presentation modes, showing that the response accuracy was lower for tracking 4 than 3 targets. The effect of the number of intervening object visits was significant for faces in all the presentation modes; for line drawings, the effect was significant in the All-Present and the Periphery-Present conditions, marginally significant in the Fovea-Present condition, and non-significant in the None-Present condition; for colors, the effect was significant only in the Fovea-Present condition (see Table 5 and Fig. 8).

The results show that eye visits to targets benefit their tracking (see also Li, Oksama, & Hyönä, 2018a, 2018b). The benefit was particularly prominent when the targets were faces that require high resolution information to be identified. This is probably due to the sampling of the high-resolution information being enhanced at the fovea and due to the maintenance of high-resolution information in working memory relying particularly on timely refreshing (Li, 2016).

In sum, the analyses in Section 6 demonstrated a bias towards visiting and dwelling on individual targets when facial identities are present, and towards visiting the blank areas between targets when color identities are present. Yet, the eye visits were predominately directed to the individual targets regardless of the object type and the presence of the object identity. The eye visits to targets were beneficial for the tracking performance, particularly when the targets were faces. The results suggest that eye movements during MIT serve to refresh the target representations in working memory in a serial manner, as well as to sample visual information both in a serial (for high-resolution information) and parallel (for low-resolution information) manner.

7. General discussion

The present study addressed the serial vs. parallel controversy in MIT by adopting the gaze-contingent display change paradigm. The result for tracking multiple faces showed that at each moment only the currently foveated facial identity affects tracking performance, supporting serial tracking (Experiment 1). Considering that our facial stimuli were nearly unidentifiable in peripheral vision, we then used color discs that can be identified accurately in foveal and peripheral vision. The results showed that the simultaneous presence of multiple color identities led to higher tracking performance (Experiment 2), supporting parallel tracking in MIT. When tracking objects depicted as line drawings, which can be identified in the peripheral vision to some but a lower degree than in the foveal vision, the results showed a pattern lying between purely serial and purely parallel tracking (Experiment 3). Taken together, the performance accuracy results indicate that the manner of tracking multiple objects varies in the serial-parallel continuum according to the identifiability of the objects.

In line with the performance accuracy results, also the eye movement results showed a more serial pattern when tracking faces and drawings and a relatively parallel pattern when tracking colors. That is, when tracking faces and drawings, the eyes were more inclined to visit each target serially, whereas in tracking multiple moving color discs an increased number visits to the blank areas between targets were

Table 5
The effect of intervening visits and set size for each object type in each presentation condition.

	Color				Drawing				Face			
	B	SE	Z	*p*	B	SE	Z	*p*	B	SE	Z	*p*
All-Present												
Intervening Visits	−0.017	0.021	−0.797	.425	−0.096	0.022	−4.370	< .001	−0.090	0.017	−5.326	< .001
Set Size	−0.254	0.119	−2.132	.033	−0.974	0.127	−7.643	< .001	−1.202	0.087	−13.883	< .001
Fovea-Present												
Intervening Visits	−0.039	0.019	−2.057	.040	−0.044	0.023	−1.889	.059	−0.180	0.021	−8.752	< .001
Set Size	−0.964	0.118	−8.162	< .001	−0.951	0.125	−7.635	< .001	−1.179	0.096	−12.252	< .001
None-Present												
Intervening Visits	−0.018	0.017	−1.077	.282	−0.026	0.019	−1.358	.175	−0.039	0.016	−2.511	.012
Set Size	−0.919	0.103	−8.933	< .001	−0.810	0.107	−7.592	< .001	−0.971	0.086	−11.273	< .001
Periphery-Present												
Intervening Visits	−0.003	0.023	−0.137	.891	−0.125	0.016	−7.636	< .001	−0.047	0.013	−3.710	< .001
Set Size	−0.377	0.120	−3.146	.002	−0.751	0.114	−6.561	< .001	−0.875	0.080	−10.880	< .001

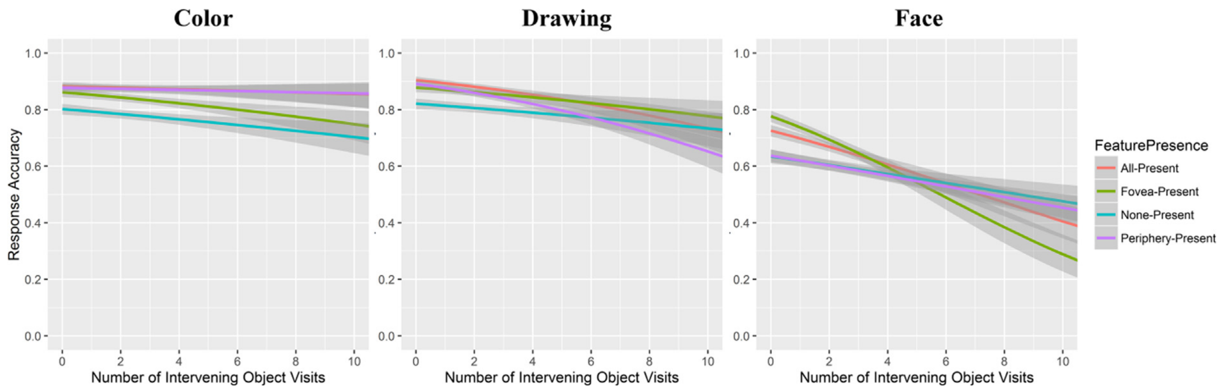

Fig. 8. The regression lines show how response accuracy varies with the number of intervening object visits for each type of objects under each presentation mode.

observed. The results indicate that eye movements serve the purpose of efficiently sampling visual information during tracking. When high-resolution information is needed for identifying a target, the foveal vision is directed to it. On the other hand, when low-resolution information is sufficient for discriminating the targets, the eyes tend to land on blank areas to sample information from multiple targets.

The analysis of eye behavior during MIT also demonstrated that, regardless of the type of objects and the presence or absence of object identities in the visual field, eye movements displayed a predominately serial pattern. Most of the eye visits were directed to individual targets rather than to blank areas even when tracking colors that can be easily identified by the peripheral vision. This suggests that eye visits to targets are not performed solely for sampling high-resolution information. Even when no object identities were visually present during tracking, the majority of eye visits were still paid to target locations, and the performance accuracy remained substantially high. The eye visits to targets benefited their tracking, particularly when tracking faces. This was shown by the finding that recently visited targets were tracked better than those that had not been visited for some time. The results indicate that object representations maintained in working memory support MIT, and eye movements serve to serially refresh the representations by visiting each target location in turn.

Taken together, the results indicate that MIT is not solely about deploying attention to multiple objects, either in a serial or parallel manner. Instead, MIT is a complex process fulfilled with the cooperation of attention, eye movements, perception, and working memory. The MIT performance manifests a more serial pattern when high-resolution information is required and a relatively parallel pattern when low-resolution information is sufficient for tracking. Below, we discuss the implications of the results and then propose an upgraded model of multiple identity tracking – MOMIT 2.0.

7.1. *Dissociated processing of coarse and detailed information underlies parallel vs. serial tracking*

The performance accuracy results demonstrate parallel tracking for colors and serial tracking for faces. Such a pattern is in line with the two-stage theories of visual processing, which posit a parallel visual analysis of all the elements in the scene followed by a serial scrutiny of a few elements (Buetti et al., 2016; Wolfe, 1994; Wolfe & Horowitz, 2017). The features processed in the first stage are basic features such as colors and orientations, which can be analyzed in parallel by low-level detectors (Wolfe & Horowitz, 2017). The features are encoded at each location in the visual field as coarse proto-objects (Rensink, 2000; Wolfe & Bennett, 1997). In the subsequent stage, focal attention can be guided to each of the locations to process detailed information so as to construct elaborated object-files (Xu & Chun, 2009). Previous research has investigated the distinct processing of the coarse information and detailed information mostly in the context of visual search and change detection (Gao, Gao, Li, Sun, & Shen, 2011; Hoffman, 1979; Wolfe & Horowitz, 2017). The present study extended this line of research by showing that in dynamic circumstances objects that can be identified by coarse information can be tracked in parallel, whereas objects that require detailed information to be identified are tracked serially. It reconciles the serial vs. parallel controversy evident in previous MIT studies of Oksama and Hyönä (2008, 2016) and Howe and Ferguson (2015). Color discs were used in Howe and Ferguson's (2015) study, which demonstrated parallel tracking, while line drawings and faces were used in Oksama and Hyönä's (2008, 2016) studies, which demonstrated serial tracking.

7.2. *Eye movements are employed for visual sampling*

Eye movements are employed to adjust the visual sampling of the low- and high-resolution information when tracking multiple objects. Though human observers are able to enhance resolution at a location by covertly deploying attention to it without moving the eyes (Yeshurun & Carrasco, 1999), it is more effective and natural to move the eyes along with attention (He, Cavanagh, & Intriligator, 1996; Meyerhoff et al., 2018). Visual resolution drops off precipitously from fovea to periphery. While coarse information can be sampled across a broad visual field, detailed information can be sampled only within the small region corresponding to the fovea (Anstis, 1974; Land & Tatler, 2009). In MIT, the variation in the sampling resolution translates to parallel processing for coarse information and serial processing for detailed information.

An efficient way of visual sampling would be to gaze at the area between multiple targets for sampling low-resolution information from these targets and to direct the eyes to individual targets when high-resolution information is needed. Such eye movement patterns have been found in previous studies on static scenes. These studies have shown that observers are predisposed to gaze at the centroid of multiple objects when trying to monitor multiple objects (Fluharty et al., 2016), while the gaze lands more frequently on individual objects when more detailed information is required for visual search (Williams et al., 1997; Young & Hulleman, 2013). The present study demonstrates that such eye movement behavior applies to dynamic circumstances as well. When tracking multiple moving color discs, the eyes tend to be directed to blank areas between targets, whereas in face tracking the eyes are mostly directed to individual targets.

7.3. *Eye movements are employed for refreshing memory representations*

MIT is not fulfilled solely by sampling visual information during tracking, but also by maintaining object representations in working memory (Oksama & Hyönä, 2004, 2008). Tracking accuracy in the present study was surprisingly high when no object identities were present during tracking (Cohen et al., 2011; Pylyshyn, 2004). In order to carry out the MIT task, the object representations in working memory need to be constantly refreshed to avoid decay. The refreshing not only keeps the location information of the moving objects updated, but also prevents the resolution of the identity information from declining. Without timely refreshing, the resolution may quickly decline making the object representations less distinguishable from each other.

Eye movements are closely intertwined with working memory and can be adopted to facilitate memory (Theeuwes et al., 2009; Van der Stigchel & Hollingworth, 2018). Object information perceived during a fixation is transferred to working memory, and after each saccade correspondence is established between the object representations in memory and the currently perceived objects, so as to achieve continuous experiences of the scene (Papenmeier et al., 2014; Van der Stigchel & Hollingworth, 2018). The representation of the saccade target is activated and receives biased allocation of attention and working memory resources, leading to a high-resolution representation maintained in working memory (Bays & Husain, 2008; Ferreira et al., 2008; Li, Zhou, Huang, He, & Shen, 2013; Shao et al., 2010). The present study suggests that during MIT eye movements and subsequent fixations are employed to refresh each target representation in turn, which then prevents the resolution of the active representations from declining. This is vital for tracking targets that require high-resolution information to be identified and kept distinguishable from other targets.

Above we argue that eye movements during MIT are functional both for visual sampling and for refreshing memory representations, so the observed eye behavior reflects the combination of the two functions. The default cycle is to move the eyes from target to target for refreshing each target representation in turn, regardless of the type of the objects and the presence or absence of the object identities. The default cycle may then be adjusted for visual sampling: the eyes move to peripheral targets more frequently and dwell there for longer time when sampling high-resolution information, while they move more to blank areas between targets when sampling low-resolution information from targets.

7.4. MOMIT 2.0

Based on the current findings and other recent research on MIT, we propose an upgraded model of multiple identity tracking – MOMIT 2.0 (for the original MOMIT model, see Oksama & Hyönä, 2008). The model not only addresses the apparent serial vs. parallel controversy in MIT, but also provides a general framework for explaining how people maintain situation awareness in dynamic circumstances via the cooperation of attention, eye movements, perception, and working memory.

7.4.1. The underlying assumption of MOMIT 2.0

The underlying assumption of MOMIT 2.0 is that the brain aims to maintain maximal (or good enough) situational awareness of the environment. This is particularly a challenge in dynamic circumstances, since it’s impossible for the observers to accumulate information to construct a valid representation of the scene and maintain it by long-term memoryas the situation is changing all the time. Thus, the observer cooperatively uses the perceptual system and working memory system in combination with the use of attention shifts and eye movements for sampling visual information and refreshing memory representations, so as to keep track of each moving object.

The key for successful tracking is to retain the resolution of the target representations in working memory at a proper level. The memory resolution naturally declines with time making target objects less distinguishable from each other. To maintain the memory resolution at sufficient level, two critical processes are involved: sampling visual information from the objects to elaborate their representations, and periodically refreshing the representations to prevent the resolution from declining. Visual sampling takes place in parallel for low-resolution information across the visual field, whereas the process is serial for high-resolution information. On the other hand, memory refreshing is serial regardless of the content of the object representations, as it requires focal attention and the eyes being directed to each target location in turn. Efficient sampling and refreshing are achieved by shifts of attention and the corresponding eye movements. The default eye movement cycle is to visit each target in turn for memory refreshing, but the cycle may be adjusted for visual sampling needs.

7.4.2. The process of MIT according to MOMIT 2.0

The process of MIT includes the following steps operating as a constant loop:

1. Sampling visual information during each fixation: high-resolution information from the object(s) at the fovea, low-resolution information from the objects in the periphery.
2. Linking the currently perceived object(s) with the corresponding object representations in working memory and updating the object locations in memory with the perceived locations. In case no memory representation exists for an object, a new representation is created.
3. For the foveally attended object, elaborating the representation with high-resolution visual information and actively maintaining it. For objects in the periphery, memory resolution quickly declines, and low-resolution visual information may be added to the representations in case the memory resolution drops to an even lower level.
4. Selecting the location of the next eye visit by assessing the following conditions in the order of priority.
   (a) If an object has not been foveally attended for a long time and its resolution in memory is susceptible to decline to a low level, move the eyes to that object location.
   (b) If high-resolution visual information of an object is available in the periphery and may be beneficial for tracking, move the eyes to the target.
   (c) If a blank area is sufficient for sampling low resolution information from multiple targets, move the eyes to such area.
5. Move the eyes to the selected location (and continue the cycle from step 1).

7.5. Comparison between MOMIT 1.0 and 2.0

MOMIT 2.0 resembles MOMIT 1.0 (Oksama & Hyönä, 2008) in some aspects but it also differs from the original version in several ways. The similarities between MOMIT 1.0 and 2.0 are the following.

1. Both models consider MIT a complex task carried out by a series of cognitive processes, including attention, eye movements, perception, and working memory.
2. Both models involve serial and parallel components.
3. Both models assume that serial shifts of attention and eye movements among targets are carried out during tracking.
4. Both models assume that information of multiple targets is maintained in working memory during tracking.

The differences between MOMIT 1.0 and 2.0 include:

1. MOMIT 2.0 provides a more general framework for addressing not only MIT but also situation awareness in dynamic circumstances more generally.
2. MOMIT 1.0 is based on the assumption that objects’ location information and feature information is initially processed separately and then bound together (Treisman, 1996; Ungerleider & Haxby, 1994, see also Botterill et al., 2011; Li, Zhou, Shui, & Shen, 2015). Instead, MOMIT 2.0 is based on the assumption that object files are constructed in a coarse-to-fine manner (Gao, Ding, Yang, Liang, & Shui, 2013; Schyns & Oliva, 1994; Xu & Chun, 2009): Objects’ location information and feature information is bound together to form proto-objects, while subsequent processes elaborate on the

constructed object files (Braet & Humphreys, 2009).

3. MOMIT 1.0 posits that the key to MIT is to establish and refresh the identity-location bindings of the target objects, while according to MOMIT 2.0 the key is to increase and retain the resolution of the object representations in working memory.
4. Compared with MOMIT 1.0, MOMIT 2.0 is a more parallel model. It assumes that the outputs of parallel processing are not non-indexed locations but proto-objects that contain both location and basic featural information, which can be sufficient for tracking in case no detailed information is required.
5. Eye movements serve an additional function in MOMIT 2.0 – facilitating the refreshing of object representations in working memory.

7.6. Implications of MOMIT 2.0

MOMIT 2.0 has the potential of explaining several results in existing studies on dynamic tracking. For instance, in research on MOT (i.e., tracking the locations of multiple identical objects) there has also been a controversy concerning parallel vs. serial tracking. Pylyshyn and Storm's (1988) original study proposed that tracking multiple object locations is a parallel process, and there are more recent studies supporting this notion (e.g., Howe, Cohen, Pinto, & Horowitz, 2010). In contrast, Holcombe and Chen's (2013) study suggests that during MOT a single tracking focus switches among the targets, sampling the location of only one target at a time. Possibly the discrepancy may stem from different requirements for resolution (Iordanescu, Grabowecky, & Suzuki, 2009), not necessarily concerning resolution of featural information but rather that of location information and/or temporal information. In classic MOT studies, objects move along distinct trajectories with fairly large inter-object space. In such cases, low-resolution location and temporal information is sufficient for discriminating the objects. Thus, the objects can be tracked in parallel, and observers tend to gaze at the center of multiple targets for visual sampling (e.g., Fehd & Seiffert, 2010). In contrast, when objects move along the same trajectories (e.g., Holcombe & Chen, 2013) and/or are close to each other, high resolution information is required for discriminating the objects. Thus, tracking becomes more serial, and observers are inclined to switch attention and eye gaze from one object to another (Li, Oksama, & Hyönä, 2018a, 2018b; Meyerhoff et al., 2018).

The notion of resolution in MOMIT 2.0 can be linked with resource theories of dynamic tracking (Alvarez & Franconeri, 2007; Holcombe, Chen, & Howe, 2014; Vul, Frank, Tenenbaum, & Alvarez, 2009) and working memory (Bays & Husain, 2008; Li, Shao, Xu, Shui, & Shen, 2013; Ma, Husain, & Bays, 2014). Allocating additional resources to an object leads to higher resolution and hence better tracking performance (Chen, Howe, & Holcombe, 2013; Srivastava & Vul, 2016). Resource allocation fluctuates dynamically with the shifts of attention and eye movements, being biased towards the foveally attended object and away from other objects (Bays & Husain, 2008; Shao et al., 2010). Thus, the more recently foveated object is maintained with higher resolution and tracked with higher accuracy.

Target set-size affects the tracking performance in two ways. Firstly, as the number of targets increases, the average amount of resources allocated to each target decreases, leading to lower resolution and lower tracking performance overall (Alvarez & Franconeri, 2007; Horowitz & Cohen, 2010; Tripathy & Barrett, 2004). Secondly, as the number of targets increases, the cycle of target visits is prolonged. Consequently, the resolution of some targets decreases due to delayed refreshing and unfavorable resource allocation resulting in poorer performance.

In the future, MOMIT 2.0 may be developed to include tracking by various type of information from multiple modalities. Current theories of dynamic tracking (including MOMIT 2.0) mostly focus on tracking by vision. Yet, to support tracking people may employ all sorts of information from multiple modalities to form combined object files (Jordan, Clark, & Mitroff, 2010). Information sampled in one modality may be transformed into another. For instance, a basketball player may track other players by touching and hearing in addition to looking; or people may recode visually seen colors in verbal form as “red”, “green”, etc. Information obtained via different modalities may be used to a different degree in tracking depending on the type of objects and other circumstances. Resources may be distributed across modalities, and increased identifiability of the objects in one modality may lower the need for resolution in other modalities. Thus, a verbal label that can readily be activated for an object may compensate for poor resolution in the visual modality. The optimal strategy would rely on the information that discriminate the objects most distinctively. As proposed by MOMIT 2.0, people may cooperatively use perception, working memory, and attention to increase and retain the resolution of the information from different modalities. How multiple modalities work interactively and how resources are distributed among them requires further research.

7.7. Summary

MOMIT 2.0 suggests that people cooperatively use attention, eye movements, perception, and working memory for dynamic tracking. Tracking appears more serial when high resolution information needs to be sampled and maintained for discriminating the targets, whereas it appears more parallel when low resolution information is sufficient.

Acknowledgements

Li was supported by the National Natural Science Foundation of China [grant number 31500887], and Li’s post-doc year in Turku, Finland, was supported by the University of Turku. Oksama was supported by Suomen Akatemia [grant number 273413]. We thank Timo Heikkilä for his assistance in the data collection.

Declarations of interest

The authors declared that there is no conflict of interest.

References

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm'. *Cognition, 93*(2), B79–B87.

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track? Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision, 7*(13) 14-1–10.

Anstis, S. M. (1974). A chart demonstrating variations in acuity with retinal position. *Vision Research, 14*(7), 589–592.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321*(5890), 851–854.

Belopolsky, A. V., & Theeuwes, J. (2009). When are attention and saccade preparation dissociated? *Psychological Science, 20*(11), 1340–1347.

Botterill, K., Allen, R., & McGeorge, P. (2011). Multiple-object tracking: The binding of spatial location and featural identity. *Experimental Psychology, 58*(3), 196–200.

Braet, W., & Humphreys, G. W. (2009). The role of reentrant processes in feature binding: Evidence from neuropsychology and TMS on late onset illusory conjunctions. *Visual Cognition, 17*(1–2), 25–47.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436.

Buetti, S., Cronin, D. A., Madison, A. M., Wang, Z., & Lleras, A. (2016). Towards a better understanding of parallel visual processing in human vision: Evidence for exhaustive analysis of visual information. *Journal of Experimental Psychology: General, 145*(6), 672–707. https://doi.org/10.1037/xge0000163.

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences, 9*(7), 349–354.

Chen, W., Howe, P. D., & Holcombe, A. O. (2013). Resource demands of object tracking and differential allocation of the resource. *Attention, Perception, & Psychophysics, 75*(4), 710–725. https://doi.org/10.3758/s13414-013-0425-1.

Cohen, M. A., Pinto, Y., Howe, P. D. L., & Horowitz, T. S. (2011). The what-where tradeoff in multiple-identity tracking. *Attention Perception & Psychophysics, 73*(5), 1422–1434.

Corbetta, M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Ollinger, J. M., Drury, H. A., ... Shulman, G. L. (1998). A common network of functional areas for attention and eye movements. *Neuron, 21*(4), 761–773. https://doi.org/10.1016/S0896-6273(00)80593-0.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*(12), 1827–1837.

Elmore, L. C., Passaro, A. D., & Wright, A. A. (2013). Change detection for the study of object and location memory. *Behavioural Processes, 93*, 25–30. https://doi.org/10.1016/j.beproc.2012.11.002.

Fehd, H. M., & Seiffert, A. E. (2010). Looking at the center of the targets helps multiple object tracking. *Journal of Vision, 10*(4) 19-1–13.

Ferreira, F., Apel, J., & Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends in Cognitive Sciences, 12*(11), 405–410.

Fluharty, M., Jentzsch, I., Spitschan, M., & Vishwanath, D. (2016). Eye fixation during multiple object attention is based on a representation of discrete spatial foci. *Scientific Reports, 6*(31832), 1–13. https://doi.org/10.1038/srep31832.

Gao, Z. F., Ding, X. W., Yang, T., Liang, J. Y., & Shui, R. D. (2013). Coarse-to-fine construction for high-resolution representation in visual working memory. *Plos One, 2*(8), e579132.

Gao, T., Gao, Z., Li, J., Sun, Z., & Shen, M. (2011). The perceptual root of object-based storage: An interactive model of perception and visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 37*(6), 1803–1823.

He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature, 383*(6598), 334–337.

Hoffman, J. E. (1979). A two-stage model of visual search. *Perception & Psychophysics, 25*(4), 319–327. https://doi.org/10.3758/bf03198811.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics, 57*(6), 787–795. https://doi.org/10.3758/BF03206794.

Hoffman, J. E. (1998). Visual attention and eye movements. In H. Pashler (Ed.). *Attention* (pp. 119–153). Hove, UK: Psychology Press ((reprinted) (1998)).

Holcombe, A. O., & Chen, W. (2013). Splitting attention reduces temporal resolution from 7 Hz for tracking one object to < 3 Hz when tracking three. *Journal of Vision, 13(1), 12*, 1–19.

Holcombe, A. O., Chen, W. Y., & Howe, P. D. L. (2014). Object tracking: Absence of long-range spatial interference supports resource theories. *Journal of Vision, 14*(6), https://doi.org/10.1167/14.6.1 1-1–21.

Horowitz, T., & Cohen, M. (2010). Direction information in multiple object tracking is limited by a graded resource. *Attention, Perception, & Psychophysics, 72*(7), 1765–1775.

Horowitz, T. S., Klieger, S. B., Fencsik, D. E., Yang, K. K., Alvarez, G. A., ... Wolfe, J. M. (2007). Tracking unique objects. *Perception & Psychophysics, 69*(2), 172–184.

Howe, P. D., Cohen, M. A., Pinto, Y., & Horowitz, T. S. (2010). Distinguishing between parallel and serial accounts of multiple object tracking. *Journal of Vision, 10*(8), https://doi.org/10.1167/10.8.11 11-1–13.

Howe, P. D. L., & Ferguson, A. (2015). The identity-location binding problem. *Cognitive Science, 39*(7), 1622–1645. https://doi.org/10.1111/cogs.12204.

Iordanescu, L., Grabowecky, M., & Suzuki, S. (2009). Demand-based dynamic distribution of attention and monitoring of velocities during multiple-object tracking. *Journal of Vision, 9*(4), 1. https://doi.org/10.1167/9.4.1.

Johansson, R., & Johansson, M. (2014). Look here, eye movements play a functional role in memory retrieval. *Psychological Science, 25*(1), 236–242. https://doi.org/10.1177/0956797613498260.

Jordan, K. E., Clark, K., & Mitroff, S. R. (2010). See an object, hear an object file: Object correspondence transcends sensory modality. *Visual Cognition, 18*(4), 492–503. https://doi.org/10.1080/13506280903338911.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24*(2), 175–219.

Land, M. F., & Tatler, B. W. (2009). *Looking and acting: Vision and action in natural behaviour.* Oxford, England: Oxford University Press.

Li, J. (2016). Dissociable loss of the representations in visual short-term memory. *The Journal of General Psychology, 143*(1), 1–15. https://doi.org/10.1080/00221309.2015.1100979.

Li, J., Zhou, Y., Huang, X., He, J., & Shen, M. (2013). *The function of looking at nothing: Eye movement to an absent object benefits the memory for it. Paper presented at the 17th European Conference on Eye Movements, Lund, Sweden*.

Li, J., Oksama, L., & Hyönä, J. (2018a). Close coupling between eye movements and serial attentional refreshing during multiple-identity tracking. *Journal of Cognitive Psychology, 30*(5–6), 609–626. https://doi.org/10.1080/20445911.2018.1476517.

Li, J., Oksama, L., & Hyönä, J. (2018b). Close coupling between eye movements and serial attentional refreshing during multiple-identity tracking. *Journal of Cognitive Psychology*, 1–18. https://doi.org/10.1080/20445911.2018.1476517.

Li, J., Shao, N., Xu, H., Shui, R., & Shen, M. (2013). Does visual working memory work as a few fixed slots? *The Quarterly Journal of Experimental Psychology, 66*(11), 2103–2117.

Li, J., Zhou, Y., Shui, R., & Shen, M. (2015). Visual working memory for dynamic objects: Impaired binding between object feature and location. *Visual Cognition, 23*(3), 357–378. https://doi.org/10.1080/13506285.2014.1001010.

Loschky, L., McConkie, G., Yang, J., & Miller, M. (2005). The limits of visual resolution in natural scene viewing. *Visual Cognition, 12*(6), 1057–1092. https://doi.org/10.1080/13506280444000652.

Lundqvist, D., Esteves, F., & Öhman, A. (1999). The face of wrath: Critical features for conveying facial threat. *Cognition & Emotion, 13*(6), 691–711. https://doi.org/10.1080/026999399379041.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17*(3), 347–356. https://doi.org/10.1038/nn.3655.

McConkie, G. W. (1997). Eye movement contingent display control: Personal reflections and comments. *Scientific Studies of Reading, 1*(4), 303–316. https://doi.org/10.1207/s1532799xssr0104_1.

Meyerhoff, H. S., Schwan, S., & Huff, M. (2018). Oculomotion mediates attentional guidance toward temporarily close objects. *Visual Cognition, 26*(3), 166–178. https://doi.org/10.1080/13506285.2017.1399950.

Motter, B. C., & Holsapple, J. (2007). Saccades and covert shifts of attention during active visual search: Spatial distributions, memory, and items per fixation. *Vision Research, 47*(10), 1261–1281. https://doi.org/10.1016/j.visres.2007.02.006.

Nummenmaa, L., Oksama, L., Glerean, E., & Hyönä, J. (2017). Cortical circuit for binding object identity and location during multiple-object tracking. *Cerebral Cortex, 27*(1), 162–172. https://doi.org/10.1093/cercor/bhw380.

Oksama, L., & Hyönä, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach. *Visual Cognition, 11*(5), 631–671.

Oksama, L., & Hyönä, J. (2008). Dynamic binding of identity and location information: A serial model of multiple identity tracking. *Cognitive Psychology, 56*(4), 237–283.

Oksama, L., & Hyönä, J. (2016). Position tracking and identity tracking are separate systems: Evidence from eye movements. *Cognition, 146*, 393–409. https://doi.org/10.1016/j.cognition.2015.10.016.

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology, 46*(3), 461–488.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience, 5*(4), 291–303. https://doi.org/10.1038/nrn1364.

Papenmeier, F., Meyerhoff, H. S., Jahn, G., & Huff, M. (2014). Tracking by location and features: Object correspondence across spatiotemporal discontinuities during multiple object tracking. *Journal of Experimental Psychology: Human Perception and Performance, 40*(1), 159–171. https://doi.org/10.1037/a0033117.

Perry, J. S., & Geisler, W. S. (2002). Gaze-contingent real-time simulation of arbitrary visual fields. In B. E. Rogowitz, & T. N. Pappas (Eds.). *Human vision and electronic imaging VII* (pp. 294–305). Bellingham, WA: Society of Photo-Optical Instrumentation Engineers (SPIE) ((reprinted) (2002)).

Peterson, M. S., Kramer, A. F., & Irwin, D. E. (2004). Covert shifts of attention precede involuntary eye movements. *Perception & Psychophysics, 66*(3), 398–405.

Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition, 32*(1), 65–97.

Pylyshyn, Z. W. (2004). Some puzzling findings in multiple object tracking: I. Tracking without keeping track of object identities. *Visual Cognition, 11*(7), 801–822.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3*(3), 179–197.

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology, 7*(1), 65–81. https://doi.org/10.1016/0010-0285(75)90005-5.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition, 7*(1–3), 17–42.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition, 76*(3), 269–295.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science, 5*(4), 195–200.

Shao, N., Li, J., Shui, R. D., Zheng, X. J., Lu, J. G., ... Shen, M. W. (2010). Saccades elicit obligatory allocation of visual working memory. *Memory & Cognition, 38*(5), 629–640.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology. Human Learning and Memory, 6*(2), 174–215.

Srivastava, N., & Vul, E. (2016). Attention modulates spatial precision in multiple-object tracking. *Topics in Cognitive Science, 8*(1), 335–348. https://doi.org/10.1111/tops.12189.

Theeuwes, J., Belopolsky, A., & Olivers, C. (2009). Interactions between working memory, attention and eye movements. *Acta Psychologica, 132*(2), 106–114.

Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology, 6*(2), 171–178.

Tripathy, S. P., & Barrett, B. T. (2004). Severe loss of positional information when detecting deviations in multiple trajectories. *Journal of Vision, 4*(12), 1020–1043. https://doi.org/10.1167/4.12.4.

Ungerleider, L. G., & Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Current Opinion in Neurobiology, 4*(2), 157–165.

Van der Stigchel, S., & Hollingworth, A. (2018). Visuospatial working memory as a fundamental component of the eye movement system. *Current Directions in Psychological Science, 27*(2), 136–143. https://doi.org/10.1177/0963721417741710.

Van Ettinger-Veenstra, H. M., Huijbers, W., Gutteling, T. P., Vink, M., Kenemans, J. L., ... Neggers, S. F. W. (2009). FMRI-Guided TMS on cortical eye fields: The frontal but not intraparietal eye fields regulate the coupling between visuospatial attention and eye movements. *Journal of Neurophysiology, 102*(6), 3469–3480. https://doi.org/10.1152/jn.00350.2009.

Vul, E., Frank, M. C., Tenenbaum, J. B., & Alvarez, G. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Advances in Neural Information Processing Systems, 22*, 1–9.

Williams, D. E., Reingold, E. M., Moscovitch, M., & Behrmann, M. (1997). Patterns of eye movements during parallel and serial visual search tasks. *Canadian Journal of Experimental Psychology, 51*(2), 151–164.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review, 1*(2), 202–238.

Wolfe, J. M., & Bennett, S. C. (1997). Preattentive object files: Shapeless bundles of basic features. *Vision Research, 37*(1), 25–43.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour, 1*(3), 58. https://doi.org/10.1038/s41562-017-0058.

Xu, Y., & Chun, M. M. (2009). Selecting and perceiving multiple visual objects. *Trends in Cognitive Sciences, 13*(4), 167–174.

Yeshurun, Y., & Carrasco, M. (1999). Spatial attention improves performance in spatial resolution tasks. *Vision Research, 39*(2), 293–306.

Young, A. H., & Hulleman, J. (2013). Eye movements reveal how task difficulty moulds visual search. *Journal of Experimental Psychology: Human Perception and Performance, 39*(1), 168–190. https://doi.org/10.1037/a0028679.