Behavior Research Methods (2020) 52:1225–1243
https://doi.org/10.3758/s13428-019-01313-2

# A hidden Markov model for analyzing eye-tracking of moving objects

## Case study in a sustained attention paradigm

Jaeah Kim[1] · Shashank Singh[2] · Erik D. Thiessen[1] · Anna V. Fisher[1]

Published online: 2 January 2020


**Abstract**

Eye-tracking provides an opportunity to generate and analyze high-density data relevant to understanding cognition. However, while events in the real world are often dynamic, eye-tracking paradigms are typically limited to assessing gaze toward static objects. In this study, we propose a generative framework, based on a hidden Markov model (HMM), for using eye-tracking data to analyze behavior in the context of multiple moving objects of interest. We apply this framework to analyze data from a recent visual object tracking task paradigm, TrackIt, for studying selective sustained attention in children. Within this paradigm, we present two validation experiments to show that the HMM provides a viable approach to studying eye-tracking data with moving stimuli, and to illustrate the benefits of the HMM approach over some more naive possible approaches. The first experiment utilizes a novel 'supervised' variant of TrackIt, while the second compares directly with judgments made by human coders using data from the original TrackIt task. Our results suggest that the HMM-based method provides a robust analysis of eye-tracking data with moving stimuli, both for adults and for children as young as 3.5–6 years old.

**Keywords** Eye-tracking · Visual object tracking · Hidden Markov model · TrackIt · Selective sustained attention

## Introduction

Eye-tracking provides temporally rich behavioral data (gaze) that is closely linked to many cognitive functions. It has been widely used to study cognition, in diverse research areas including category learning (e.g., Rehder & Hoffman 2005), visual attention (e.g., Doran, Hoffman, & Scholl, 2009), sports expertise (e.g., Smuc, Mayr, & Windhager, 2010), visual perception (e.g., Gegenfurtner, Lehtinen, & Säljö, 2011), implicit bias and stereotype (e.g., Pyykkönen, Hyönä, & van Gompel, 2009), language processing (e.g., Barr 2008) and psychological disorders such as schizophrenia (e.g., Holzman et al. 1974). Beyond psychology, eye-tracking applications include safety evaluation in driving (e.g., Palinko, Kun, Shyrokov, & Heeman, 2010), usability studies in human–computer interaction (e.g., Jacob & Karn 2003), and diagnosis of Alzheimer's disease (e.g., Fernández, Castro, Schumacher, & Agamennoni, 2015).

Most of these applications rely on the extensive work that has been done assessing two important components of gaze: *fixation* (maintenance of gaze on a single location) and *saccade* (quick movement of gaze between two fixations) (Cassin, Solomon, & Rubin, 1984). There exist well-documented standards for identifying and analyzing fixations and saccades in eye-tracking data (Duchowski, 2017), and a meta-analysis study has shown that the most commonly used eye-tracking measures are number of fixations, mean fixation duration, and gaze duration (a function of multiple fixations) (Jacob & Karn, 2003). These have been incorporated into user-friendly analysis software built into commercial eye-trackers, and there

✉ Jaeah Kim
jaeahk@andrew.cmu.edu

[1] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[2] Machine Learning Department, Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

also exists open-source software for fixation-and-saccade-based analyses of generic eye-tracking data (e.g., Dink & Ferguson 2015). These analytical advances have facilitated adoption of fixation- and saccade-based eye-tracking methods as standard tools in investigating cognition and behavior.

While fixations and saccades describe most human eye movement in response to stationary or rapidly moving visual stimuli, tracking of smoothly moving stimuli obeys a different dynamic, namely *smooth pursuit*—slow eye movement that maintains the image of a moving object on the fovea (Cassin et al., 1984). Far less research using eye-tracking methods has studied smooth pursuit, in part due to a relative lack of analysis tools. A recent comprehensive review of eye-tracking methodology mentioned smooth pursuits only three times and noted that "*a robust and generic algorithm for their detection is currently an open research problem*" (Duchowski, 2017, p. 176). As a result, many eye-tracking studies rely on hand-coding by trained human coders (e.g., Franchak, Kretch, Soska, & Adolph, 2011; Bambach, Crandall, Smith, & Yu, 2018). In part because eye-tracking samples so densely over time, this can be costly in terms of time and effort (for example, the human coding in Experiment 2 of the present study took 45–50 human hours, involving over half a million human judgments), and can be subject to inconsistencies between coders.

It may not be immediately apparent why analysis of smooth pursuit eye movements can be so much more challenging than analysis of fixations and saccades. To better understand the challenges involved, consider an example scenario illustrated in Fig. 1. If objects are moving, they may overlap for brief periods of time (Fig. 1, panel 2). The eye-tracking information during this period of overlap can be insufficient to infer which object the participant is tracking, and this information must therefore be aggregated with information from before and/or after the period of overlap. Object overlap is a problem even when the eye-tracker perfectly captures the participant's gaze, but, especially in crowded visual environments, the complications for analysis are dramatically exacerbated by noise intrinsic to both eye-trackers and human behavior (including inaccurate eye-tracker calibration, oculomotor control, blinking/head movement, etc.), which effectively increase the overlap area between objects. This is especially true in children, since the human smooth pursuit system develops much more slowly than the saccadic system, reaching maturity only in adulthood (Ross, Radant, & Hommer, 1993; Katsanis, Iacono, & Harris, 1998; Luna, Velanova, & Geier, 2008). As a result, simple methods of identifying what object participants are tracking can be quite inaccurate; for example, as we will discuss later, our data from a visual object tracking task suggest that, as much as 1/3 of the time, the object closest to the participant's measured gaze is not the object they are tracking.

In this paper, we propose a novel hidden Markov model (HMM) approach to analyzing eye-tracking data in the context of multiple moving objects of interest. Given continuous gaze data collected from a participant tracking moving objects with known positions over time, our model can accurately determine the object being tracked at each time point. Because our method uses raw gaze data instead of pre-classified fixation/saccade data, our method works in contexts that include either or both of smooth-pursuit and fixation/saccade eye-movements, bypassing the difficult problem of identifying smooth pursuit movements. We anticipate that our model may be useful for researchers in cognitive science and related areas and have made an open-source Python implementation freely available.

## A hidden Markov model approach

Hidden Markov Models (HMMs) are a popular generative model for time series data, in which observed data are assumed to be drawn, at each time point, from a distribution depending on an unobserved *hidden state*. To make learning the model tractable, a "Markov" assumption is made; namely, the hidden state is assumed to depend only on

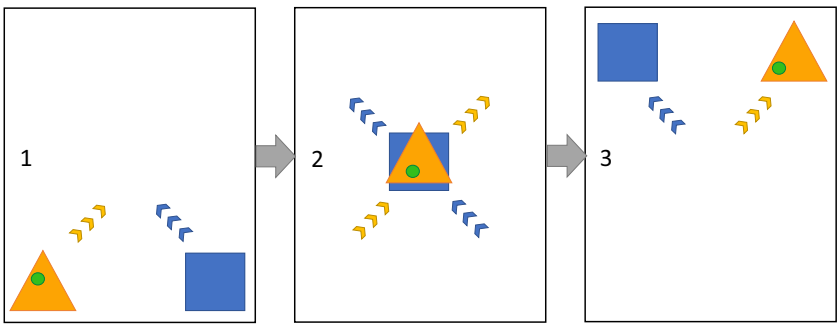


**Fig. 1** An example on an object collision (Panel 2), during which the object being tracked is ambiguous, without using information from the past (Panel 1) or future (Panel 3)

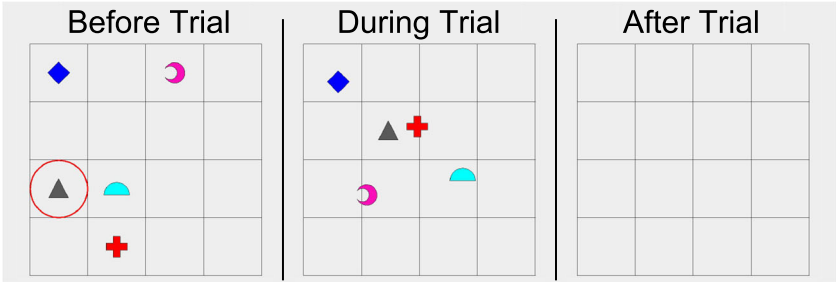


**Fig. 2** An example trial of the standard TrackIt task (endogenous condition), on a $4 \times 4$ grid with 4 distractor objects. The target object here is the grey triangle, as indicated before the trial. A video of an example TrackIt trial can be found at https://osf.io/utksa/

temporally proximal hidden states, and not on distant hidden states. An HMM is a natural choice for a simple model of human visual object tracking; at each time point $t$, the participant is looking at *something* $S(t)$ (the hidden state), and we observe eye-tracking data $E(t)$ that is primarily a function of $S(t)$ and random noise. Because humans tend to follow individual objects for at least short periods of time (rather than constantly switching between objects), at least for short timesteps, the state $S(t)$ is strongly related to the preceding and successive states ($S(t-1)$ and $S(t+1)$).[1] Unlike simpler models that consider data at each time point independently, the HMM uses this short-term dependence to mitigate noise and handle complex scenarios such as object collisions (when multiple objects briefly occupy the same space), without sacrificing the fine temporal resolution of eye-tracking data.

## Selective sustained attention and TrackIt

Selective sustained attention (SSA) is an important cognitive process that enables everyday functioning and task performance by allowing us to: (1) choose components of our environment to process at the exclusion of others and (2) maintain focus on those components over time. SSA relies on both endogenous factors (e.g., internal goals) and exogenous factors (e.g., stimulus salience), and studying how these factors develop and interact in guiding attention during childhood is of special interest for SSA development research (O'Connor, Manly, Robertson, Hevenor, & Levine, 2004).

Unfortunately, quantifying SSA in young children has proven challenging. Though studies have attempted to downward extend the Continuous Performance Task (CPT; the standard task for measuring sustained attention in adults) to make it easier and more engaging for children, 50% of children 4.5 or younger were still unable to complete the task and provide usable data (for review see Fisher & Kloos 2016).

TrackIt, introduced by Fisher, Thiessen, Godwin, Kloos, and Dickerson (2013), is a child-appropriate visual object-tracking task recently developed to measure SSA, that can capture differential contribution of exogenous and endogenous control of attention and allow flexible assessment over a range of developmental years (including pre-school years), with parameters for adjusting difficulty with age (Kim, Vande Velde, Thiessen, & Fisher, 2017). In the TrackIt task (illustrated in Fig. 2), participants visually track a single target object moving about on a grid, among other moving distractor objects. At the end of each such trial, all objects vanish from the grid, and participants are asked to identify the target's final grid cell location the target occupied before vanishing. Previous work has shown that children as young as 3 years old can consistently complete the TrackIt task and provide usable data (Fisher et al., 2013).

Prior studies using TrackIt have measured task performance mainly in terms of this final response—whether the final grid cell was correctly identified. However, this measure has several limitations. For example, Kim et al. (2017) suggested that many behavioral 'errors' may be attributable to participants' limited visual resolution when identifying the final grid cell location of the target (thereby clicking an adjacent cell). Also, this measurement is made *after* task and only indirectly tells us what participants do *during* task.

To address these limitations of data currently available directly from TrackIt, we began collecting eye-tracking data. Analyzing these rich data, however, involved addressing the non-trivial technical challenge described above, namely that of robustly identifying the object a participant is tracking from noisy eye-tracking data, even when objects are moving, crowded, and potentially overlapping. This problem motivated the development of the new method we propose in this paper; this new method can facilitate analyses of smooth pursuit eye movements in the context of the TrackIt task and can be useful for analyzing eye-tracking data in more general experimental contexts.

[1] While human attention is likely not really Markovian (i.e., the attentive state at a time $t$ may depend directly on attentive states at very distant timepoints), these dependencies vary widely with context (e.g., the types of objects and the task at hand), and modeling long-term dependencies is beyond the scope of this work.

## Related work

Several prior studies have reported using HMMs to analyze eye-tracking data. Kärrsgård and Lindholm (2003) used HMMs for an eye-typing application (in which users form words by fixating on characters on a display). More recently, Haji-Abolhassani & Clark (2013, 2014) used HMMs to predict the visual tasks being performed by participants viewing a painting. Although not using eye-tracking data, Kumar, Harding, and Shiffrin (2018) used a similar algorithm to analyze computer mouse movements in adults, in another task recently proposed to measure SSA. Finally, a substantial line of work has used HMMs to study eye movement patterns involved in face recognition (Chuk, Chan, & Hsiao, 2014, 2015; Chuk, Chan, Shimojo, & Hsiao, 2016; Chuk, Crookes, Hayward, Chan, & Hsiao, 2017b; Chuk, Chan, & Hsiao, 2017; Brueggemann, Chan, & Hsiao, 2016). A MATLAB toolbox has also been published implementing these analyses (Coutrot, Hsiao, & Chan, 2018).

All of the above studies share several features that contrast them from the current study. First, the stimuli presented were static images. While Coutrot et al. (2018) used conversational video stimuli, the regions of interest, which were the faces of speakers, were essentially stationary relative to the display. In contrast, our stimuli are videos of moving objects, and so the parameters of our HMMs evolve over time as objects move. Second, all prior analyses were based on first identifying fixations and then modeling these fixations using HMMs, whereas the HMM in the current study directly models continuous eye-tracking data. Thus, the approach that we present below is more appropriate for measuring smooth pursuit, which is not composed of fixations. Finally, the prior studies used repetitive tasks (e.g., face recognition with aligned face stimuli) or identical tasks performed by different participants, so that many identically distributed samples can be combined (across stimuli or across participants) to learn a single HMM. This was a good fit for the studies that investigated *where* most humans gaze when presented with certain kinds of stimuli; however, this approach is not a good fit for the current study or other studies that involve smooth pursuit of objects moving in a non-predetermined fashion. In the current study, object trajectories are randomly generated before each trial, and we are interested in studying broad patterns of behavior, independent of specific stimuli and locations presented. As a result, each trial is distinct, and an HMM must be fit for each trial *using data from only that trial*. The approach we describe below makes this possible because positions of objects of interest over time are known.

To the best of our knowledge, HMMs have been used only a few times in the context of tracking *moving* objects. Citorík (2016) used a rather different HMM-based approach for analyzing eye-tracking data in the classic multiple object tracking paradigm of Pylyshyn and Storm (1988). Their approach utilized a separate HMM for each stimulus object, with two states indicating whether or not that object is being tracked. Beyond behavioral studies, Mantiuk, Bazyluk, and Mantiuk (2013) described the use of an HMM algorithm similar to the one described in this paper in a real-time 3D scene rendering system.

Finally, a few other approaches have been considered for automated analysis of eye-tracking data in dynamic contexts. Most relevantly, Zelinsky and Neider (2008) proposed a shortest-distance model (SDM), which assumes, at each time point, that participants are tracking the object closest to their gaze. This model, which we use as a baseline for comparison in our experiments, does not leverage temporal information, and our experiments consistently show that our proposed HMM method outperforms the SDM in terms of correctly identifying the tracked object and detecting switches between objects.

Other methods have been proposed based on determining dynamic areas of interest (Papenmeier & Huff 2010; Friedrich, Rußwinkel, & Möhlenbrink, 2017). These papers focus on precisely specifying the *spatial* regions in which gaze corresponds to tracking a particular object. Our HMM method, on the other hand, focuses on using *temporal* structure to improve tracking classification. Thus, these methods are complementary; for example, our HMM can be used with more complex emission distributions based on AOIs computed using the methods in these papers (from the geometry of displayed shapes), instead of the spherical Gaussian distributions we describe in "Hidden Markov model". As an example application, in the three-dimensional setting studied by Papenmeier and Huff (2010), the HMM could serve to distinguish between tracking two objects when they overlap (due to the two-dimensional viewing projection). A more subtle but important distinction is that these papers use a "strict" spatial criterion (gaze point inside the AOI) for matching, whereas we use a "soft" spatial criterion (quantified by the likelihood of a gaze point under a Gaussian distribution around the object). The "strict" criterion may be appropriate, for example, when studying constraints of the visual system during object tracking. On the other hand, when studying *attention*, for which gaze is a less precise proxy, the more lenient "soft" criterion may be more appropriate (especially for child participants, due to noisier oculomotor control Ross et al. 1993; Katsanis et al. 1998; Luna et al. 2008). The HMM can also be made to enforce the "strict" criterion by using a uniform emission distribution over the object (0 probability outside the object).

## Contributions and organization of this paper

The main contribution of this paper is to propose and evaluate a method for using continuously sampled gaze

data to identify a sequence of objects that a participant tracks during an experiment, given the positions of possible objects of interest over time. In particular, we present an HMM-based method that can handle smoothly moving, crowded, and potentially overlapping target objects, and can identify tracked objects densely over time with high temporal precision. As discussed above, we do not know of other previously proposed methods that can handle this kind of data; therefore, we compare the proposed new approach to both a simpler baseline model and human coding of smooth pursuit eye-tracking data.

In "Hidden Markov model" we formally present our proposed HMM approach. Section "The TrackIt task" describes the TrackIt task, a task paradigm recently used for studying sustained attention development in young children, which we use as a setting for validating the proposed HMM model. In "Experiment 1: supervised TrackIt" and "Experiment 2: comparison with human coding", we present the results of two validation experiments designed to evaluate the proposed method. Section "Measuring HMM model fit" briefly discusses diagnostic methods for evaluating some of the assumptions underlying our proposed HMM analysis. Finally, Section "Conclusions and future directions" discusses some implications of our results, as well as possible extensions of the proposed model.

### Source code and reproducibility

Supporting materials for both experiments reported in this paper are freely available via the Open Science Framework at https://osf.io/u8jbs/. Specifically available are:

1. Python scripts for reproducing all our analyses, results, and figures.
2. All eye-tracking and TrackIt data used.
3. Videos of an example Supervised TrackIt trial (used in Experiment 1) and an example standard TrackIt trial (used in Experiment 2).
4. All human-coded data and human coder materials (including coding protocol script, Solomon Coder configuration file, and an example trial video reconstruction used by the coders) for Experiment 2.
5. The Python executable used to collect all eye-tracking data with the SMI RED-250 mobile eye tracker (SMI, 2009).

Finally, a TrackIt executable and its source code are freely available at http://www.psy.cmu.edu/~trackit/.

## Hidden Markov model

**Overview of hidden Markov model** We model the participant as being, at each time point $t$, in a state $S(t) \in \mathcal{S}$. When in the state $S(t)$, we model the participant's eye-tracking data with a Gaussian emission distribution centered at the center $X_S(t)$ of the object $S(t)$. In the case of TrackIt, if $N_D$ denotes the number of distractors (e.g., in Experiment 1, $N_D = 4$), $N = N_D + 1$ (1 target, $N_D$ distractors). Figure 3 illustrates the main components of our model in this context. Note, however, that the model is quite general. For example, the Gaussian emission distribution can be easily generalized for non-elliptical objects. The model might even be adaptable to a multiple object tracking setting by using centroids of sets of objects rather than the objects themselves (Fehd & Seiffert 2008, 2010; Hyönä, Li, & Oksama, 2019).

**Notation** Spatial coordinates are measured in pixels ($\approx 0.02°$ of visual field) with $(0, 0)$ denoting the bottom left corner of the display. $x_{min}, x_{max}, y_{min}$, and $y_{max}$ respectively denote the minimum and maximum horizontal

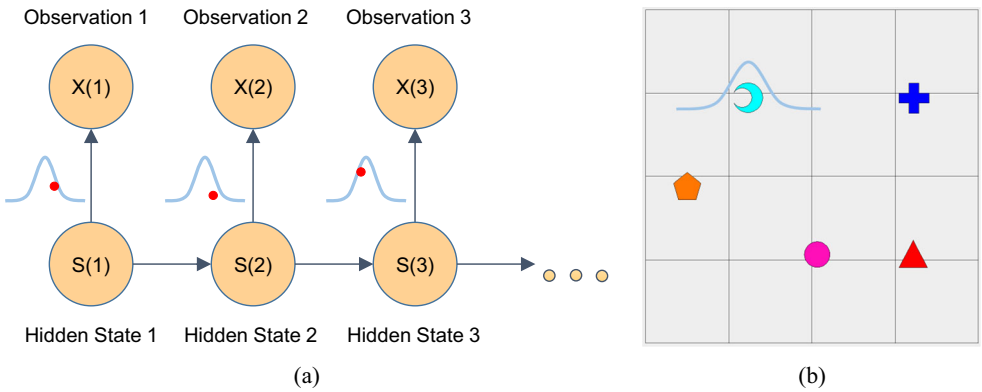


**Fig. 3** **a** Graphical model schematic of HMM. The initial state (object) $S(1)$ is sampled uniformly at random. At each time point $t$, we observe a gaze data point $X(t)$, distributed according to a Gaussian centered around the state $S(t)$. At the next time point $t + 1$, a new state $S(t + 1)$ is sampled according to a distribution depending on $S(t)$, and the process repeats. **b** Example conditional distribution of $E(t)$ given $S(t) =$ "Blue Moon"

and vertical coordinates observable by the eye-tracker. The observable region $R := [x_{min}, x_{max}] \times [y_{min}, y_{max}]$ is a rectangle including the entire grid traversable by TrackIt objects. Within the context of any particular trial, $T$ denotes the trial length (in 60 Hz frames), and $t \in [T] := \{1, 2, ..., T\}$ indexes individual frames.

**Hidden state model** The sequence of underlying hidden states is modeled as a Markov chain with a fixed initial distribution $\pi \in [0, 1]^{\mathcal{S}}$ (such that $\sum_{S \in \mathcal{S}} \pi_S = 1$) and transition matrix $\Pi \in [0, 1]^{\mathcal{S} \times \mathcal{S}}$ (such that, for each $S \in \mathcal{S}$, $\sum_{S' \in \mathcal{S}} \pi_{S,S'} = 1$). Since, in this study, we are interested in using our model to classify participants' behavioral states over time, to avoid biasing the model, $\pi$ is constrained to be uniform (i.e., $\pi_{s_1} = \cdots = \pi_{s_N}$), and $\Pi$ is constrained to have identical diagonal values $c_1$ and identical off-diagonal values $c_2$; i.e.,

$$\Pi = \begin{bmatrix} c_1 & c_2 & \cdots & c_2 \\ c_2 & c_1 & \cdots & c_2 \\ \vdots & \vdots & \ddots & \vdots \\ c_2 & c_2 & \cdots & c_1 \end{bmatrix}.$$

We set $c_1 = \frac{599}{600}$ and $c_2 = (1 - c_1)/N$, corresponding to an average of 1 uniformly random transition per 600 frames ($\approx$ 10 s); this choice is due to the tuning procedure used to learn the model hyperparameters (see "Supervised TrackIt").

**Emission distributions** Let $S : [T] \rightarrow \mathcal{S}$ denote the sequence of states assumed by the participant. At each time point, if the participant is in the state corresponding to tracking the object $s$, the model assumes the eye-tracking data of the participant is distributed according to an isotropic Gaussian centered at the center of $S$; that is, for each $t \in [T]$ and $s \in \mathcal{S}$,

$$E(t)|S(t) = s \sim \mathcal{N}\left(X_s(t), \sigma^2 I_2\right),$$

where $E : [T] \rightarrow R$ denotes the eye-tracker trajectory, and, for each $S \in \mathcal{S}$, $X_S : [T] \rightarrow R$ denotes the trajectory of the object corresponding to state $S$. The spherical standard deviation $\sigma$, which we model as common across objects, is an important hyperparameter whose selection is discussed below.

**Model fitting** Because, when analyzing eye-tracking data from TrackIt, we have no *a priori* knowledge of the true state sequence $S$, the model is trained in an unsupervised manner, using a maximum likelihood estimate (MLE); that is, the estimated sequence of states is that which maximizes the likelihood of the observed eye-tracking data. Our implementation uses the Viterbi algorithm (Forney, 1973), a standard dynamic programming algorithm for efficiently computing the MLE of an HMM.

**Parameter selection** The main free parameters in the model are the transition probability $c_2$ and the spherical standard deviation $\sigma$ of the Gaussian emission distributions. The optimal values for these parameters depend on context-specific factors such as the size and density of objects, as well as properties of the participant (e.g., have less precise smooth pursuit eye-movements than adults (Ross et al., 1993; Katsanis et al., 1998; Luna et al., 2008), corresponding to larger $\sigma$). Since, when determining whether to make a transition at a particular frame, the model essentially trades-off between the cost of transitioning and the cost of selecting an object far from the gaze point, both of these parameters essentially modulate how often the model transitions between objects. Hence, to keep analysis simple in this paper, we fix $c_2$ to a sensible value (corresponding to the 0.1-Hz mean transition rate we enforce in Experiment 1), and report results across a large range of $\sigma$ parameters, highlighting results at some values of $\sigma$ that we found to provide the best results. We also varied experiment parameters between Experiments 1 and 2, giving some indication of how different $\sigma$ should be used in different settings. At present, we do not have an automatic method for calibrating $\sigma$, and we suggest that users either consider results over a range of $\sigma$ values or calibrate $\sigma$ by having human coders manually code a small subset of data.

## The TrackIt task

TrackIt is a recently developed task paradigm for measuring SSA in children (Fisher et al., 2013). TrackIt has been shown to have good psychometric properties for measuring SSA, and research in several labs has linked performance on TrackIt to classroom learning, numeracy skills, prospective memory, and proactive control (Fisher et al. 2013; Erickson, Thiessen, Godwin, Dickerson, & Fisher, 2015; Doebel et al. 2017; Doebel, Dickerson, Hoover, & Munakata, 2018; Brueggemann & Gable 2018; Mahy, Mazachowsky, & Pagobo, 2018).

In the standard TrackIt task, participants visually track a single target object as it moves on a grid, among moving distractor objects. For each trial, the target and distractor objects are constructed with random colors (selected without replacement from a set of nine distinct colors) and shapes (selected without replacement from a set of nine distinct shapes); that is, out of 81 possible objects, target and distractor objects are selected randomly under the constraint that no color or shape is repeated within a trial. See Fig. 2 for an example. At the beginning of each trial, objects appear on a grid, centered in random, distinct grid cells, and the target object is indicated by a red circle around it.

Upon starting the trial (by button press), the red circle disappears, and the objects begin to move in piecewise-linear

trajectories from grid cell to grid cell at a constant speed (500 pixels, or 10°, per second). At the end of each trial, all objects vanish, and the participant is asked to indicate with their finger the grid cell the target object last occupied before disappearing.

The path of each object is randomized, with the constraint that the target has to be in the center of a grid cell at the end of the trial, to reduce ambiguity for the participant in determining its final location. Due to this constraint, trial length is not fixed, but varies slightly between trials (to allow the target to reach the center of a grid cell), with a minimum of 10 s.

The grid size, object speed, number of distractors, and minimum trial length, are experimenter-selected TrackIt parameters; the above values were suggested by prior work as appropriate for young children. In Experiment 1, we used an "easy" 4 × 4 grid size with four distractor objects, while in Experiment 2, we used a "hard" 6 × 6 grid with six distractors. These settings span the range of parameters recommended for use with young children by previous work (Kim et al., 2017); all other parameters were set to the default values in TrackIt.

## Experiment 1: supervised TrackIt

Evaluating the performance of the HMM model requires comparing its predictions to a "ground truth" estimate of the object the participant is tracking. In this section, we report results from one approach to obtaining such ground truth. Specifically, we conducted TrackIt experiments in which we used several features to amplify the salience of the target object relative to distractors (see details below). The core assumption inherent to this approach is that by making the target object highly salient, we make the task relatively easy such that the participants are able to successfully track the prescribed target at all time points; thus, we use the target object itself as an estimate of ground truth. Additionally, since we are interested in the HMM's performance in the context of possible attention switches among different objects, rather than using a single target for the entire trial, we changed the target periodically throughout the trial. We also lengthened trials to ensure several object transitions would take place.

### Supervised TrackIt

To tune the parameter $\sigma$ and evaluate model performance, we designed a 'supervised' variant of TrackIt, in which we know, with relatively high confidence, what object the participant is looking at (i.e., the 'true state') at most time points. To do this, we made the target flash white repeatedly (for 100 ms, separated by 200 ms) during the entire trial, making it salient and easy to track. Participants were instructed to follow the flashing object with their eyes. Rather than using a single target for the entire trial, the flashing target changed at random intervals (uniformly between 5 s and 15 s). To allow multiple target changes, trials were lengthened to a minimum of 30 s (from 10 s in Unsupervised TrackIt). Changing the target within trials was essential to ensure the fitted model could accurately detect transitions between objects; without this, the model would learn to always estimate a single most likely target during each trial (i.e., the selected $\sigma$ would be too large). As in Unsupervised TrackIt, the target was circled in red and flashed before trial start, so participants could begin the trial tracking the correct object. Other parameters and preprocessing steps were identical to the Unsupervised TrackIt setup. TrackIt recorded the flashing target's identity in each frame, allowing us to compare model predictions to this 'ground truth'. Some error is introduced by the delay with which participants transition after the blinking object changes. Better results might be obtained by ignoring a few frames after each change when measuring error, but our results are robust without doing this.

### Experimental procedure

**Participants** Fifteen healthy adult volunteers aged 18 to 31 ($M = 22.5$; $SD = 3.4$; 13 female, two male) and 15 typically developing 5-year-old children aged 5.1 to 5.9 ($M = 5.3$; $SD = 0.23$; seven female, eight male) each performed 12 trials of Supervised TrackIt, including two initial practice trials during which the experimenter explained the task. Practice trials were not analyzed, giving ten usable trials/participant.

**Materials and apparatus** Stimuli were presented on a Lenovo laptop screen with physical dimensions 19.1 cm × 34.2 cm and pixel dimensions 1080 × 1920 pixels (approximately 22° × 40° of visual field). Participants were seated at a desk facing the screen with their heads about 0.5 m away from the screen. The SMI RED-250 mobile eye-tracker (SMI, 2009) was used to record continuous gaze positions at 60 Hz during TrackIt trials. After using SMI's iView X software to calibrate the eye-tracker, we used a custom Python script (available in the supplementary material https://osf.io/vqjgs/) to collect eye-tracking data synchronized with TrackIt.

**TrackIt parameters** We used a 4 × 4 grid size, object speed of 500 pixels/second, four distractors, and minimum trial length of 10 s (as suggested by prior work as appropriate for young children; (Kim et al., 2017)). These parameters are recommended for use with children younger than 5 by previous work (Kim et al., 2017), but we used these settings

**Table 1** Missing gaze data before and after preprocessing

| Population | Proportion of Frames | | |
|---|---|---|---|
| | Raw Data | After Interpolation | After Interpolation & Filtering |
| Adults | 7.9% | 4.1% | 2.4% |
| 5 year olds | 41.1% | 30.3% | 15.0% |

with 5-year-old participants since this validation experiment necessitated high participant performance to simulate an accurate ground truth. All other parameters were set to the default values in TrackIt.

**Data preprocessing** Child eye-tracking data contains a large proportion of missing values (due to children looking away from task or moving excessively), and so we preprocessed data to mitigate this. Whenever a short interval of at most $\leq$ 10 consecutive frames ($\approx$ 167 ms) of eye-tracking data was missing, we linearly interpolated gaze during those frames from non-missing data immediately before and after that interval. Then, we discarded all data from participants for whom more than 50% ($>$ 5 trials) were missing more than more than 50% of frames (3 children); our reported results are on data from the remaining 12 children. Even after these steps, intervals of ($>$ 10 frames of) eye-tracking data may still be missing. For these frames, the HMM automatically assigns a 'null' state, and the frames before and after each such interval are fit independently by the Viterbi algorithm. When evaluating model performance, we report results both treating these frames as incorrect classifications (giving a conservative 'worst-case' lower bound on performance) and ignoring these frames (giving a less conservative 'average-case' performance estimate).

Table 1 shows, for each population and condition, the proportion of frames missing eye-tracking data, in the raw data, after interpolating short intervals of missing data, and after filtering participants with excessive missing data. As expected, the proportion of missing data was far larger for children than for adults. Both preprocessing steps significantly improved data quality, especially with data from child participants.

**Evaluating model performance** We compared our HMM's performance to that of a 'Shortest Distance Model' (SDM; Zelinsky & Neider 2008) that assumed that, at each time point, the participant was looking at the object closest to their gaze. This model is equivalent to a variant of our HMM with uniform transition matrix $\Pi$, thus ignoring the underlying Markov model and using only emission probabilities.

Our main measure of model (HMM or SDM) performance is *decoding accuracy*, the proportion of frames (across all participants and trials) on which the model agrees with the "ground truth" (location of the target in the supervised version of the task).

Recall that the HMM has a free parameter $\sigma$ that must be selected by the user. In this experiment, we report results for 50 logarithmically spaced values of $\sigma$ between 10 and $10^4$ pixels ($\approx$ 0.2°-24° of visual field).

## Results

Figure 4 shows the HMM's accuracy, as a function of $\sigma$, along with that of the SDM and 'chance' of 20% (one out of five total objects), for adult and child participants, respectively.

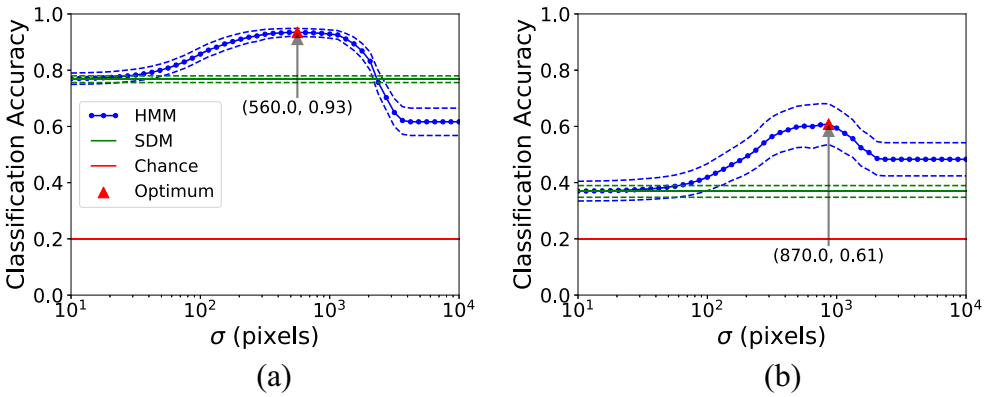


**Fig. 4** **a** Semi-log plot of HMM, SDM, and chance accuracies for adult data, as functions of HMM parameter $\sigma$. Dashed lines indicate bootstrapped 95% confidence bands. The point of optimal HMM performance (our suggested value of $\sigma$) is indicated by a triangle. Only accuracies on non-missing frames are shown, but curves computed using all frames were qualitatively similar. **b** Same plot for child data

While both the HMM and the SDM perform much better on adult data than on child data, curves are qualitatively similar for both populations. For very small $\sigma$ (e.g., $< 100$ ($\approx 2°$)), the cost of selecting an object even slightly further than the closest object outweighs the cost of transitioning states, and so the HMM behaves essentially like the SDM. For very large $\sigma$ (e.g., $> 2000$ ($\approx 49°$)), the emission distributions of all objects become similar, and the HMM may fail to ever transition, performing worse than the SDM. As we expected, the optimal $\sigma$ for children was much larger than that for adults (870 pixels ($\approx 18°$) versus 490 pixels ($\approx 10°$)), reflecting less precise visual tracking of the target object. However, for both adults and children, in a large range of approximately $\sigma \in [10^2, 10^3]$ ($\approx 2° - 21°$), the accuracy of the HMM is significantly higher than that of the SDM, with a mean performance difference larger than the statistical uncertainty (in terms of the radius of 95% confidence intervals around the mean accuracy; see Fig. 4).

This analysis suggests that superiority of the HMM decoder depends on the value of $\sigma$, albeit quite robustly. Hence, to objectively evaluate decoder performance independently of tuning, we next used leave-one-out cross-validation: For each of the 15 participants, we measured the accuracy of the HMM on this "held-out" participant when using the $\sigma$ value that maximized the mean accuracy over the other 14 participants. Table 2, which reports the average of this "held-out" accuracy over participants, indicates that the HMM provides a large mean improvement ($\geq 16.1\%$ in adults, $\geq 20.9\%$ in children) in accuracy over the SDM.

## Experiment 2: comparison with human coding

Our results with Supervised TrackIt in Experiment 1 suggested that the HMM provides a significant improvement over the accuracy of a SDM baseline model which simply selects the closest object to the eye gaze at each time point. While this is a promising first result in favor of the HMM model, the results of Experiment 1 are insufficient to fully provide a confident assessment of the HMM's performance, for a number of reasons.

First, the sample of 15 children and 15 adults is fairly small. Second, the results from Experiment 1 (Table 2) allow for the possibility that the HMM's decoding accuracy for child data might be as low as 50-60%. Given the much higher accuracy measured in adults, it is possible that a significant proportion of the measured "error" of the HMM model stemmed not from true model errors, but rather from an inaccurate assumption about ground truth, because children may have struggled to continuously follow the target object even under the condition of high salience.

Third, Experiment 1 involved only data from 5-year-old children, whereas TrackIt is intended for use with children as young as 3 years old (Kim et al., 2017) (for whom decoding accuracy might be even lower). Finally, the Supervised TrackIt might differ from the Unsupervised TrackIt task in ways that affect the performance of the HMM. For example, the transition probabilities of the HMM were calibrated to match the mean transition frequency of the Supervised TrackIt task, whereas the transition frequencies of children in Unsupervised TrackIt are unknown.

Experiment 2 was designed to address these limitations and provide a more direct assessment of the HMM as a tool for decoding the object of attention in the standard TrackIt task, and to do so over a larger sample of child participants, with a larger range of ages. To accomplish this, we used the output of TrackIt and the eye-tracker to construct video recordings of eye-gaze data overlaid on the original TrackIt task, and used human coders to estimate the object of attention from these videos. We then compared the output of the HMM and SDM algorithms to these human judgments.

### Methods

**TrackIt settings** In the previous experiment, it was necessary for the participant to perform the task successfully, and so we used TrackIt settings that are known to yield

**Table 2** Mean (across participants) proportion of supervised frames correctly classified, based on using the other 14 participants to select the optimal $\sigma$

| Population | HMM | SDM |
|---|---|---|
| All frames | | |
| Adult | 91.4%(2.7%) | 75.3%(2.5%) |
| Child | 52.7%(3.9%) | 31.8%(2.3%) |
| Non-missing or interpolated frames only | | |
| Adult | 93.5%(1.3%) | 76.8%(1.5%) |
| Child | 60.7%(2.2%) | 36.8%(2.1%) |

Numbers in parentheses are radii of 95% normal confidence intervals, based on standard errors across participants

high performance in 5-year-old participants (Kim et al., 2017). In Experiment 2, we used parameter settings that are age-appropriate for 4 to 6-year-old participants based on prior research (Kim et al., 2017). Specifically, the following TrackIt settings were used: object speed was 500 pixels per second, grid size was $6 \times 6$, number of distractors was six. Thus, we expected the task to be challenging for the 3-year-old children, but wanted to test if, with the help of eye-tracking measurement, informative features of their attention may be still be retrieved. Since we did not need to accommodate multiple object transitions within each trial, we also used more a conventional minimum trial length of 10 s.

Each participant performed two conditions of the TrackIt task – the "Exogenous" condition and the "Endogenous" condition – designed to differentially measure top-down vs. bottom-up attentional components which have been distinguished in the attention development literature (Oakes, Kannass, & Shaddy, 2002; Colombo & Cheatham 2006). This condition difference was not relevant for the present paper (these data were collected as part of a larger project and the effects of condition on attention will be reported separately elsewhere). To ensure that condition did not affect any of the conclusions of this experiment, we ran all analyses on the data from each condition separately; all results were qualitatively identical, with minor quantitative differences. To simplify presentation of results, in this paper we present average results over the two conditions.

For completeness, we briefly describe the two conditions here. In the endogenous condition, the target is differentiated from the distractors only by being circled *before the start of the trial*, as described in "The TrackIt task". Hence, during the trial, participants must rely primarily on their internal goal representation to support their SSA in a top-down fashion, in the absence of external support. In the Exogenous condition, the target object is differentiated from the distractor objects *throughout the trial* in two ways. First, the distractor objects are constrained to all be identical (i.e., the same shape and color), and distinct from the target. Second, the target rhythmically "shrinks" and "unshrinks" (specifically, it alternates between its default size and a 50% reduced size, at 3 Hz) throughout the trial. These features increase the salience of the target relative to the distractors, thereby exogenously supporting maintenance of attention on the target. Participants performed the two conditions on two separate days (approximately 1 week apart), with order counter-balanced.

**Participants** Fifty typically developing children, aged 3.5 to 6 years ($M = 4.60$, $SD = 0.67$), each performed 11 TrackIt trials, including one initial practice trial during which the experimenter explained the task. Practice trials were not analyzed, giving ten usable trials per participant per condition. After removing eight participants' data due to eye-tracking data quality issues (as described in Experiment 1 under "Data preprocessing"), 42 children, ages 3.5 to 6 years ($M = 4.65$, $SD = 0.71$) contributed data to the analysis.

**Materials and apparatus** Stimulus display and eye-tracking setup were identical to those in Experiment 1.

**Data preprocessing** Data preprocessing (to reduce missing eye-tracking data) was identical to that in Experiment 1. Frames that were missing even after preprocessing were excluded from the evaluation of model performance. After preprocessing, data from 42 children (840 unique trials) remained. As noted below, unlike in Experiment 1, when evaluating model performance, we ignored missing frames.

**Video coding procedure** Here, we describe our procedure for coding videos of participants' eye-tracking data. A detailed protocol can be found at https://osf.io/54kyd/. An example of a trial video reconstruction used for video coding can be viewed at https://osf.io/m6kru/. After preprocessing, 84 sessions (one session per participant in each condition) of ten trials each were analyzed. For each trial, using the outputs of TrackIt and the eye-tracker, we generated a video reconstruction (at 1/10 the original speed) consisting of the participants' gaze overlaid on a video of the original TrackIt object trajectories. Two human coders then used the videos to identify which object, if any, the participant was tracking at each time point.

Each of the 84 sessions was randomly assigned to either Coder 1 or Coder 2. Additionally, to assess inter-coder reliability, a randomly selected 20% of sessions were coded by both coders. Ultimately, 45 sessions were coded only by Coder 1, 21 sessions were coded only by Coder 2, and 18 sessions (the "overlap dataset") were coded by both coders.

To make the task manageable for human coders, sessions were coded every six frames (yielding ten judgments/second), and, accordingly, the HMM and SDM classifications were down-sampled by a factor of 6. Coding was performed using Solomon Coder (Péter, 2017). Each 10-Hz timepoint was coded as one of

{"Object 0", "Object 1", ..., "Object 6", "Off Screen", "Off Task"},

with each "Object" code corresponding to one of the seven displayed objects, "Off Screen" corresponding to missing eye-tracking data, and "Off Task" corresponding to the coder being unable to identify the object being tracked. Altogether, 501,147 total judgments were made. Inter-coder reliability in terms of joint proportion of agreement was 84.5% (95% confidence interval (80.6%, 88.5%)), out of 25075 total judgments per coder on the overlap dataset. When we excluded "Off Task" frames (as we

do when comparing HMM and SDM to human coding) agreement increased to 95.3% (95% confidence interval (93.0%, 97.6%)).

**Evaluating model performance** As in Experiment 1, we compared our HMM's performance to that of a 'shortest distance model' (SDM) that assumed that, at each time point, the participant was looking at the object closest to their gaze.

Also as in Experiment 1, our first measure of model performance was the proportion of frames agreeing with ground truth (where "ground truth" is now human coding instead of Supervised TrackIt object locations). However, this measure does not capture more specific attention dynamics that may play out over a finer temporal scale, such as, attentional switches between objects.

Thus, for Experiment 2, we additionally evaluated how well the models can identify attentional switches. That is, for each pair $(t, t+1)$ of consecutive timepoints (with non-missing eye-tracking data), we identified whether the model (HMM, SDM, or human coding) identified an attentional switch (i.e., whether $\hat{S}(t) = \hat{S}(t+1)$). This resulted in a binary sequence of "switch predictions" (i.e., "Switch" or "No Switch") for all non-missing timepoints. We then compared the HMM and SDM switch predictions with the human-coded switch predictions using a variety of common binary classification performance measures. Note that, because the classification problem is strongly imbalanced (i.e., 96.9% of frames were classified as "No Switch" by human coders), accuracy (i.e., the proportion of frames agreeing with human coders) is a poor measure of switch detection performance—for example, a trivial model that always predicts "No Switch" achieves an accuracy of 96.9%. Instead, we measured:

1. Precision: proportion of detected switches that are true
2. Recall: proportion of true switches that are detected
3. Matthews' correlation coefficient (MCC): Pearson correlation between predicted switches and true switches
4. F1 score (a.k.a., Dice coefficient): harmonic mean of Precision and Recall

Of these, Precision and Recall are one-sided performance measures, in that a model that predicts only "Switch" would have perfect Recall and very low Precision, and a model that predicts only "No Switch" would have perfect Precision and very low Recall. MCC and the F1 score are balanced, in that they yield a score of 1 only if the predicted switch sequence is exactly identical to the true switch sequence. In this sense, MCC and F1 are better measures of performance in practical settings, and we chose to present Precision and Recall results because they illustrate how performance depends on the parameter $\sigma$, and why the HMM outperforms the SDM in practice.

Each statistic above was calculated separately for each of the 84 sessions; below, we report means and normal confidence intervals over the 84 sessions. Note that, unlike in Experiment 1, where Supervised TrackIt provided a "ground truth" value for every frame, in Experiment 2, no ground truth is available for frames with missing eye-tracking data. For this reason, we only report numbers with missing data frames removed (as opposed to treating them as "incorrect" predictions).

## Results

**Proportion of frames agreeing with human coding** Figure 5 shows the HMM accuracy, as a function of $\sigma$, and that of the SDM, as well as the joint proportion of agreement for human coders, when omitting frames classified as "Off Task" by either coder. Both models performed far above 'chance' accuracy of $\approx$ 14.3% (1/7 total objects). For very small $\sigma$ (e.g., $<$ 50 ($\approx$ 1°)), the HMM behaves essentially like the SDM. For very large $\sigma$, the HMM has trouble detecting attention switches, and so performance decays. However, for nearly all $\sigma$ considered, the HMM significantly outperforms the SDM (65.7% $\pm$1.2% accuracy), reaching peak accuracy (85.4% $\pm$ 1.8%) at $\sigma = 300$ pixels ($\approx 6°$). Figure 5 also shows two estimates of agreement between human coders, to which performance

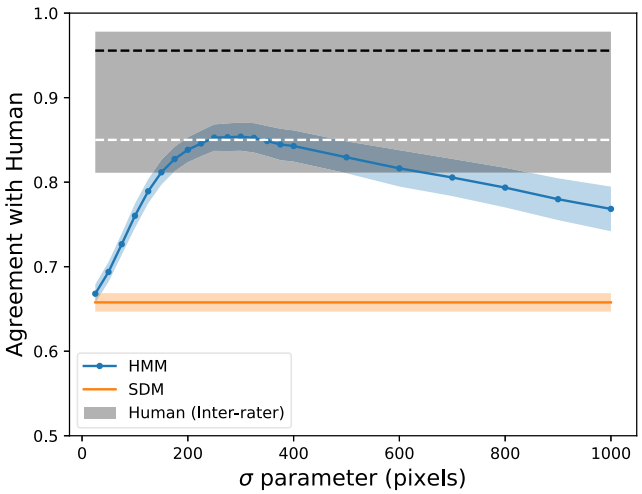


**Fig. 5** Accuracy (in terms of proportion of frames agreeing with hand-coding) for HMM (as a function of $\sigma$ parameter) and SDM, as well as joint proportion of agreement between human coders. Markers indicate $\sigma$-values for which the HMM was actually computed; other values are linearly interpolated. Shaded regions indicate 95% normal confidence intervals, also linearly interpolated between $\sigma$-values for which the HMM was actually computed. As motivated in "Discussion", two versions of joint proportion of agreement between human coders are plotted: for the dashed black line, frames in which either coder gave an "Off Task" coding were omitted, while, for the dashed white line, these frames were included

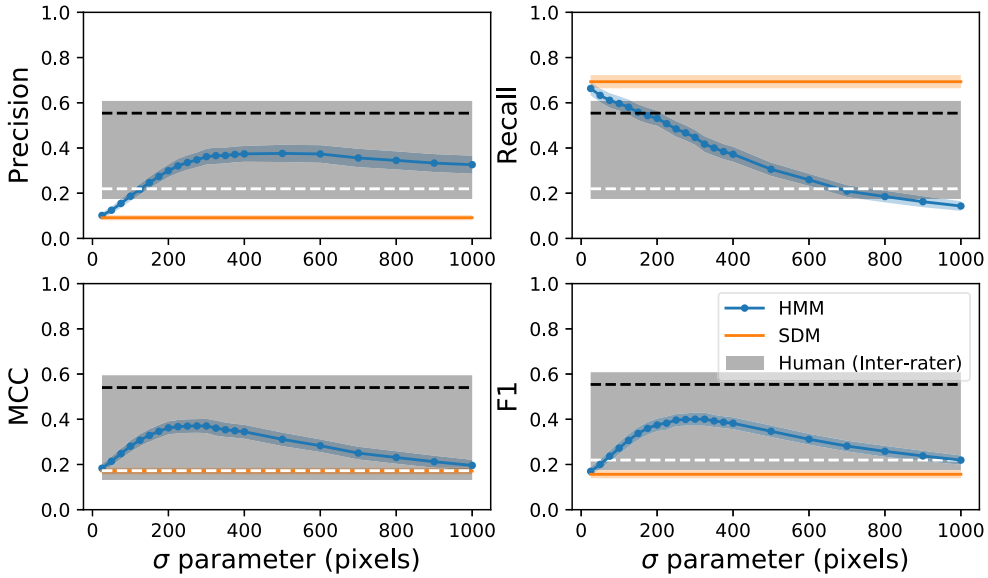


**Fig. 6** Precision, Recall, Matthews' correlation coefficient (MCC), and F1 score for predicting attentional switches using the HMM and SDM, as well as for human coders (using each coder as a ground truth for the other, and then averaging over coders). Blue markers indicate $\sigma$-values for which the HMM was actually computed; other values are linearly interpolated. One may note that, in terms of Recall, the SDM exhibits higher performance than the HMM—this makes sense given that the SDM labels frames as switches much more liberally in general. Hence, correspondingly, the SDM performs poorly in terms of Precision, which penalizes incorrect "switch" predictions. In terms of MCC and F1 score, which incorporate precision and recall into more balanced measures of accuracy, the HMM outperforms the SDM for $\sigma$ values considered. Blue and orange shading indicate 95% normal confidence intervals, also linearly interpolated between $\sigma$-values for which the HMM was actually computed. As motivated in "Discussion", two versions of human performance (inter-rater agreement) are plotted: for the dashed black line, frames in which either coder gave an "Off Task" coding were omitted, while, for the dashed white line, these frames were included. Grey shading indicates the region between the lower 95% confidence bound of the white line and the upper 95% confidence bound of the black line

of the HMM may be compared with some care, as discussed in "Discussion".

We note that this accuracy for the HMM is much higher than the $\approx$ 65% lower bound estimated in Experiment 1. This is despite the fact that we use more challenging parameters in the TrackIt task in Experiment 2 (6 distractors in Experiment 2 vs. 4 distractors in Experiment 1). This finding supports the possibility that, in Experiment 1, we were able to reasonably approximate the "ground truth" for adults but not for children.

**Detection of attentional switches** Figure 6 displays several performance measures of the HMM's and SDM's abilities to detect attentional switches. As we expected, the SDM has a reasonable Recall of 0.68 (i.e., it detects 68% of true attentional switches). However, it has a Precision of only 0.07 (i.e., 93% of the switches it predicts are spurious). This is only slightly higher than the "chance" Precision (0.03) of a model that predicts "Switch" in every time frame, making it hardly usable for researching attentional switches. In the HMM, as $\sigma$ increases from 0, Recall decays gradually and Precision increases significantly, at least up to $\sigma = 300$, at which point Precision plateaus. For this value of $\sigma = 300$, the HMM offers much more balanced Precision of 0.36 and Recall of 0.45. According to the more balanced performance metrics, the HMM is far more informative of attentional switches than the SDM for essentially all values of $\sigma$ considered, with an MCC of 0.37 (compared to 0.17) and an F1 score of 0.40 (compared to 0.16), for $\sigma = 300$. Reassuringly, the optimal value of $\sigma = 300$ under both of these measures is the same as the value optimizing the proportion of frames agreeing with human coding, as described above, suggesting that the same decoding model reliably approximates human coding under both these measures. Table 3 gives precise numerical confusion matrices for the HMM (with $\sigma = 300$), and for the SDM, over all 501147 judgments made in this experiment.

**Table 3** Confusion matrices for attentional switch detection for HMM (with $\sigma = 300$) and SDM, using human-coding as ground truth

| | HMM | | SDM | |
|---|---|---|---|---|
| | "Switch" | "No Switch" | "Switch" | "No Switch" |
| Human "Switch" | 7159 | 13019 | 11173 | 5012 |
| Human "No Switch" | 9026 | 471943 | 126077 | 358885 |

Green and red cells indicate correct and incorrect predictions, respectively

**Table 4** Confusion matrices for attentional switch detection for HMM (with $\sigma = 300$) and SDM, allowing for a slack of 200 ms between detected and true switches

| | HMM | | SDM | |
|---|---|---|---|---|
| | "Switch" | "No Switch" | "Switch" | "No Switch" |
| Human "Switch" | 11827 | 8351 | 30118 | 55 |
| Human "No Switch" | 4837 | 476132 | 107132 | 363842 |

Since the observed precision and recall numbers were not extremely high (e.g., Precision and Recall $< 0.5$ for the best HMM model, and $< 0.6$ for human-to-human reliability), we considered the possibility that our measure of prediction correctness (i.e., an exact match between the model and human predictions on each frame) was too stringent. This consideration was motivated by the anecdotal observation that even human coders sometimes disagreed on the exact frame at which an attentional switch occurred, but more often agreed on whether or not a switch occurred within a few frames. Furthermore, Solomon Coder appears to have a limited temporal precision, of 3–10 eye-tracking frames ($\approx$ 50–170 ms seconds), when sub-sampling the video for coding, potentially causing temporal ambiguity when lining up the human labels with the model predictions. For these reasons, we also considered a more lenient measure of correctness, identical to the first, except that model "Switch" predictions were considered to agree with human "Switch" predictions if they were within two video-coding frames (200 ms).

By all metrics, performance of both HMM and SDM, as well as inter-coder agreement, improved under this more lenient measure; detailed results are given in Table 4 and Fig. 7. While the precision and recall of the SDM both improved (precision from 0.07 to 0.24 and recall from 0.68 to 0.99), the precision of the SDM was still very low. The precision and recall of the HMM (with $\sigma = 300$) also both improved (precision from 0.36 to 0.59 and recall from 0.45 to 0.71). Again, according to the more balanced performance metrics, the HMM significantly outperforms the SDM for essentially all values of $\sigma$ considered, with an MCC of 0.62 (compared to 0.41) and an F1 score of 0.63 (compared to 0.37), for $\sigma = 300$. These results can be interpreted as a trade-off between the *temporal precision* and the detection performance of the model—the model is more reliably able to detect switches to within 250 ms than to within 50 ms.

## Measuring HMM model fit

The HMM proposed in this paper relies on a number of assumptions about participant behavior; for example, it assumes that, on any frame, the participant is tracking exactly one of the displayed objects. In reality, participants may behave in many other ways. For example, they may simultaneously track multiple objects (Pylyshyn & Storm 1988; Meyerhoff, Papenmeier, & Huff, 2017), or they may gaze towards empty portions of the display (e.g., former

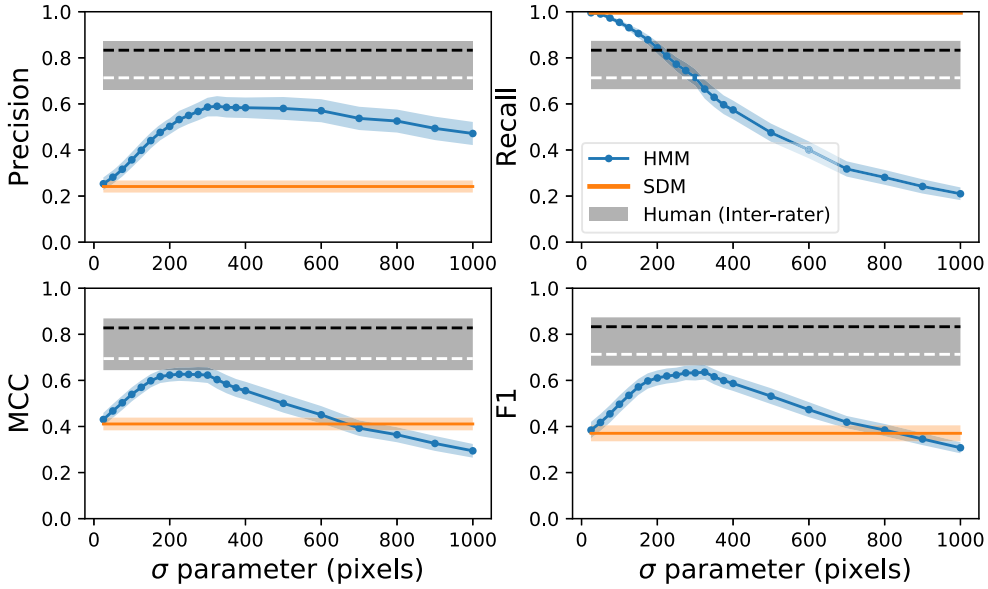


**Fig. 7** Precision, Recall, Matthews' correlation coefficient (MCC), and F1 score for predicting attentional switches using the HMM and SDM, *allowing for a slack of 200 ms between detected and true switches*. Allowing for this slack improves performance of all models according to all measures (compare Fig. 6). As in Fig. 6, the SDM exhibits good (almost perfect) Recall, but at the cost of very low Precision, while, for most $\sigma$ values considered, the HMM performs better according to Precision, MCC, and F1 score

positions of objects (Ferreira, Apel, & Henderson, 2008; De Groot, Huettig, & Olivers, 2016). An important feature of generative models such as the HMM is the ability to explicitly compute the likelihood of observed data under modeling assumptions, and thereby to detect violation of those modeling assumptions. In this section, we discuss a likelihood-based method for detecting trials in which behavior deviates from the single-object-tracking behavior assumed by the HMM.

To motivate our approach, recall the null-hypothesis testing framework used in statistics, in which one justifies conclusions drawn from the data by assuming a simple "null" model and then comparing statistics of the observed data to those predicted (via calculation or simulation) under the null model. This allows standardized quantification (e.g., a $p$ value) of how unusual the (statistics of) the data are under the null model, which is then considered indicative of how well the null model fits the data.

**A trial log-likelihood statistic** As a simple example, which we investigate here, one can ask how well a particular state sequence (such as the maximum likelihood sequence our HMM outputs) explains the gaze data in a trial. To answer this, we propose a "trial log-likelihood" (TLL) statistic of a state sequence $S$, defined as the log of the likelihood of the observed gaze data in a trial given the state sequence $S$. Due the Markov assumption and the Gaussian emission distributions, the log-likelihood of a trial is simply proportional to $-\frac{1}{T}\sum_{t=1}^{T}\frac{\|E(t)-X_{\hat{S}(t)}(t)\|^2}{\sigma^2}$ (i.e., the negative mean of the squared distances between the gaze points and the corresponding object centers, normalized by the squared $\sigma$ parameter). This simple form lends a clear intuition for the TLL statistic: TLL tends to be low when the participants gaze tends to be far from the most likely object.

We implemented the TLL statistic (included in the supporting materials at https://osf.io/ysgcz/), and we performed a simple validation of the TLL statistic as follows. Since coders used the "Off Task" classification to code frames on which the participant did not appear to be tracking any single object, we hypothesized that TLL of the MLE state sequence should correlate negatively with the proportion of frames in a trial that were classified as "Off Task; i.e., trials with more frames classified as "Off Task" should be unlikely under the HMM, which only models single-object tracking. Indeed, the Pearson correlation (across trials) between TLL and the proportion of "Off Task" frames was $-0.36$ (with 95% confidence interval $(-0.42, -0.30)$ according to Fisher $Z$-transformation and $(-0.44, -0.30)$ according to bootstrapping with $10^4$ repetitions).

**Addressing poor model fit** Having identified trials with low goodness-of-fit, a researcher can handle these trials in one of several ways, including (a) omitting these trials from downstream analysis, (b) manually coding these trials, or (c) adding a new state to the HMM to account for behavior during these trial when a clear behavioral pattern (e.g., following the centroid of some objects, corresponding to a Gaussian distribution around that centroid, or "looking at nothing" (Ferreira et al., 2008; De Groot et al., 2016), corresponding perhaps to a Gaussian distribution around the former position of an object that has moved or disappeared from the display) can be identified.

To pursue option (c) in a principled manner, given an HMM with a particular set of states and transition matrix, one can leverage the fact that the HMM is a fully-specified generative model to test the null hypothesis that the data were generated by *that* HMM under *any* possible state sequence. Specifically, one could compare the maximum likelihood (over state sequences) of the data to the maximum likelihood of simulated data from the HMM. This would allow one to objectively compare the fit of HMMs with different sets of states. Since, in the current paper, we do not study modifications of the HMM states, we leave investigation of this idea for future work.

## Discussion

While it is encouraging to see that the HMM quite reliably outperforms the SDM, ultimately, the purpose of this evaluation is to try to understand whether the performance of the HMM (in terms of frame classification accuracy, switch detection, or some other metric) is "good enough" for it to be used in place of human coders in real experimental settings. In addition to noting that this will depend on the particulars of the experiment, it is important to note some limitations of the performance measures provided in our experiments. As illustrated by imperfect inter-coder agreement, the classifications provided by human coders are a "noisy" ground truth. For this reason, rather than "perfect performance" (e.g., 100% accuracy or $F_1$ score of 1), we cautiously suggest comparing the HMM's performance under each measure to corresponding measures of inter-coder reliability, which suggest how well the HMM could possibly perform in our evaluation. For this reason, Figs. 5, 6, and 7 include plots of inter-coder reliability under each measure (computed by using each coder as a "ground truth" for the predictions of the other (on the overlap dataset) and then averaging over coders).

Even this comparison must be interpreted with some care. In our evaluation, we required the HMM and SDM to provide an object classification for every frame that the human "ground truth" did not identify as Off Task or Off Screen. However, evaluation of inter-coder agreement suggested that the majority of frames on which coders disagree are those for which one, *but not both*, coders

classified the frame as "Off Task"; when *omitting* these frames, human performance (plotted in dashed black) is significantly above that of the HMM and SDM, whereas, when *counting these frames as incorrect*, human performance (plotted in dashed white) was comparable to that of the HMM. While omitting these frames gives humans a much easier task than the model HMM (thus potentially overestimating agreement), since these are typically the most difficult (ambiguous) frames to classify, the alternative of counting these frames as incorrect may conversely underestimate human agreement, since humans operated under the assumption that "Off Task" was a valid judgment. Thus, a fair measure of human performance to which to compare model performance likely lies somewhere in between these two lines.

These results suggest that it may be desirable to allow the HMM an equivalent of the "Off Task" classification, or, more specifically, to allow it to explicitly abstain from classifying some difficult frames, which are ambiguous even for human coders. While we have some initial thoughts on how this might be achieved (e.g., adding an explicit "Ambiguous" or "Off Task" state with emission distribution uniform over the display), this would require, at the very least, tuning a new hyperparameter and determining how to evaluate classifications in this state; hence, we leave this for future work. For the time being, we have proposed a trial log-likelihood (TLL) statistic, which can be used as a indicator for the quality of the fit of the HMM to the data, and we have shown that the TLL statistic correlates with human coders' "Off Task" classifications.

## Conclusions and future directions

This paper proposed a novel algorithm, based on a hidden Markov model, to predict the object a participant is tracking in a dynamic visual scene, given their gaze position and the positions of possible objects of interest over time. The HMM converts noisy spatiotemporal eye-tracking data into a sequence of a small number of states, simultaneously denoising the data and making it more behaviorally interpretable. The model is flexible in that input data can be from any visual stimulus with known moving objects or areas of interest, and many analyses can be performed on its output. A Python implementation of the HMM is freely available online, and we invite other researchers to use it in their own studies.

We evaluated this model in the context of a child object tracking task, TrackIt, using both a supervised variant of TrackIt and judgments of human coders to provide ground truth labels. The main evaluation setting was a rather challenging setting, with noisy eye-tracking data provided by young (3–6 years old) children and a dense scene of seven fast-moving objects.

The findings of this validation study are as follows. First, compared to a shortest distance model (SDM) baseline that assumes the participant is attending to the object closest to their gaze, the HMM can consistently improve prediction accuracy on an average frame by at least 15–20%. Second, while the HMM requires the user to specify an additional hyperparameter $\sigma$, it outperforms the SDM baseline for a large range of values of $\sigma$. Third, for appropriate $\sigma$, the accuracy of the HMM on child data, in a fairly dense TrackIt environment with seven moving objects, is approximately 85%. Fourth, the HMM is able to detect attentional switches with far more precision than the SDM baseline, allowing for a slight loss in recall. Finally, by several measures, the agreement between the HMM and human coders is comparable to the agreement between two human coders, suggesting that the HMM method may be sufficiently accurate for use in behavioral experiments.

We reiterate that at present, we do not have a general, automatic method for calibrating the tuning parameter $\sigma$ in the HMM. $\sigma$ depends on both the physical properties (e.g., display size and resolution, viewing distance, object speed) of the experimental setup and characteristics of the participant (e.g., age). Practical solutions include considering results over a range of $\sigma$ values or calibrating $\sigma$, either by having human coders manually code a small subset of data from the task being studied or by directly estimating the variance of the participant's gaze data when tracking an object (e.g., using a calibration experiment consisting of TrackIt with no distractor objects). Statistical approaches, such maximum likelihood, may also be applicable. When in doubt, both intuition and our empirical results suggest that erring on the side of using a smaller $\sigma$ value will minimize potential bias introduced by the HMM model, while still outperforming the SDM.

### Applications to attention research

The temporal dynamics of attention span several timescales, and eye-tracking is among the few behavioral tools that allow researchers to probe the fastest of these timescales. Attention has been studied at sub-second timescales as well as on the scale of minutes or hours (Van Dongen & Dinges 2005; Aue, Arruda, Kass, & Stanny, 2009; Smith, Valentino, & Arruda, 2003; Arruda, Zhang, Amoss, Coburn, & Aue, 2009; Fiebelkorn, Pinsk, & Kastner, 2018). For example, recent work, based on both high-frequency (ECoG) neural data and behavioral data, has advanced an account of attention as a system that oscillates rapidly (at 4–8 Hz) between perceptual sampling and attentional switching/exploratory (motor) states modulated by intrinsic

neural oscillatory rhythms (Helfrich et al., 2018; VanRullen, 2018; Fiebelkorn & Kastner, 2018). A temporally precise behavioral measure of attentional *switches*, such as the TrackIt-eye-tracking combination studied here, may be especially useful for finely investigating the behavioral side of this high-frequency system.

In the context of SSA development research in children, the rich data and potentially greater sensitivity of the combined TrackIt and eye-tracking set-up may further address the measurement gap for SSA in young children. It may be possible, for example, to perform within trial time-course analyses or individual difference analyses that were previously infeasible due to limited density and quality of data provided by each participant. We believe this work could be useful towards building a normative account of sustained attention development, especially in young children, with potential implications for early detection of atypicalities in attention development.

## Extension to natural scenes with automatic object detection

The most limiting constraint of the proposed method is that it requires knowing the positions of all objects of interest. While readily available for artificially-generated stimuli, this information may be difficult to obtain in studies that use videos of natural scenes or are not computer-based. An especially interesting context is that of head-mounted video and gaze-tracking, which are becoming popular tools for studying behavior in natural environments (Smith, Yu, Yoshida, & Fausey, 2015). Many studies utilizing these technologies rely on human coding to identify what objects participants are viewing at each timepoint (Franchak et al., 2011; Bambach et al., 2018). Besides being slow, expensive, and difficult to replicate, this is infeasible in real-time feedback settings (discussed below).

To bypass this limitation, a promising approach, which we are currently pursuing, is to combine our HMM approach with algorithms for automated object detection in video, which have become quite fast and robust in recent years (Redmon, Divvala, Girshick, & Farhadi, 2016; Ren, He, Girshick, & Sun, 2015; Wang, Shen, & Shao, 2018). While further work will be needed to evaluate the effectiveness of the HMM method in natural scenes, this technology could accelerate behavioral research in natural environments by quickly identifying objects with which participants interact visually. Given the diversity possible in natural scenes, several additional challenges will likely be needed to make this technology robust, however. For example, rather than following a single object, viewing natural scenes often requires tracking multiple objects simultaneously (Meyerhoff et al., 2017). As noted previously, this may require introducing additional states in the HMM corresponding to subsets of visible objects. Extensive research in the multiple object and multiple identity tracking paradigms suggests that gaze may be concentrated around the centroid of the tracked objects, with occasional looks to the individual tracked objects (Hyönä et al., 2019). This could be incorporated into the HMM by adding a state whose emission distribution is Gaussian around the centroid, or a mixture of a Gaussian around the centroid and Gaussians around the individual objects.

## Online extensions for eye-tracking-based feedback

The HMM approach described in this paper maximizes the joint likelihood over the entire sequence of object-tracking predictions, based on a forwards-backwards algorithm that traverses the entire gaze data sequence twice. The method thus requires that the entire experimental data have already been collected. A number of innovative recent papers have utilized eye-tracking to provide real-time feedback to humans as they perform certain tasks, in contexts such as visual search (Drew & Williams, 2017), manual assembly (Renner & Pfeiffer, 2017), and medical (Ashraf et al., 2018) or programming (Sun & Hsu, 2019) education. Eye-tracking-based feedback has potential to be faster, cheaper, and more widely usable than similar feedback based on neural data collected as participants perform tasks in an fMRI (Awh & Vogel 2015; Faller, Cummings, Saproo, & Sajda, 2019), given the huge expense and practical constraints associated with fMRI, as well as the relatively slow timecourse of the BOLD signal. For these and other applications, it may be desirable to adapt our proposed model to the online setting, in which object-tracking predictions must be made rapidly and using only data from previous (as opposed to future) timepoints to inform the current prediction. This could likely be achieved by replacing the HMM in our proposed method with one of several variants of HMMs and the forwards-backwards algorithm that have been previously proposed for online settings (Stiller & Radons 1999; Liu, Jaeger, & Nakagawa, 2004; Mongillo & Deneve 2008). Nevertheless, further work may be necessary to ensure that predictions are sufficiently fast and accurate to provide helpful feedback.

## Towards a cognitive model of object tracking

Our decoder is based on a generative model of eye-tracking data. This model may be a suggestive first step towards linking eye-tracking data to the cognitive process of visual object tracking, and, perhaps, to the higher-level construct of visual SSA. Currently, the model plausibly encodes how eye-tracking data is generated *when following a particular object* $X$, but the model of how the object $X$ itself is selected is overly simplistic (fixed transition

probabilities, independent of object properties and other experimental parameters). Using such a model to study participant performance during task (as in this study) requires fixing the HMM with uniform initial and transition probabilities, so that the model does not intrinsically prefer some states over others (e.g., in the case of TrackIt, the model should treat the target identically to the other objects). Conversely, a realistic cognitive model should have non-uniform probabilities (e.g., preferring to follow the target over distractors, by virtue of SSA). Hence, a major step in developing such a cognitive model would be fitting its parameters to behavioral data. For Gaussian HMMs, this can be done using expectation maximization, specifically the Baum-Welch algorithm Bilmes & et al. 1998), which we suggest as a fruitful direction for future work.

**Acknowledgements** We thank Anna Vande Velde, Emily Keebler, Melissa Pocsai, and Oceann Stanley for their help collecting data. We thank Priscilla Medor and Kristen Boyle for help coding eye-tracking videos. We thank the children, parents, and teachers of the CMU Children's School, Amazing Scholars Academy Preschool, Beth Shalom Early Learning Center, and Glenn Avenue Preschool for making this work possible. We thank Dr. Frank Papenmeier and anonymous reviewers for several helpful suggestions that significantly improved the manuscript. This work was supported by the National Science Foundation (grant BCS-1451706 to AVF and EDT and graduate research fellowship DGE-1252522 to SS).

**Open Practices Statements** The data and materials for both Experiments 1 and 2 are available on the Open Science Framework (OSF) at https://osf.io/u8jbs/. None of the experiments reported here was preregistered.

## References

Arruda, J. E., Zhang, H., Amoss, R. T., Coburn, K. L., & Aue, W. R. (2009). Rhythmic oscillations in quantitative EEG measured during a continuous performance task. *Applied Psychophysiology and Biofeedback*, *34*(1), 7.

Ashraf, H., Sodergren, M. H., Merali, N., Mylonas, G., Singh, H., & Darzi, A. (2018). Eye-tracking technology in medical education: A systematic review. *Medical Teacher*, *40*(1), 62–69.

Aue, W. R., Arruda, J. E., Kass, S. J., & Stanny, C. J. (2009). Cyclic variations in sustained human performance. *Brain and Cognition*, *71*(3), 336–344.

Awh, E., & Vogel, E. K. (2015). Attention: Feedback focuses a wandering mind. *Nature Neuroscience*, *18*(3), 327.

Bambach, S., Crandall, D., Smith, L., & Yu, C. (2018). Toddler-inspired visual object learning. In *Advances in neural information processing systems*, (pp. 1209–1218).

Barr, D. J. (2008). Analyzing visual world eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474.

Bilmes, J. A., et al. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, *4*(510), 126.

Brueggemann, A., & Gable, S. (2018). Preschoolers' selective sustained attention and numeracy skills and knowledge. *Journal of Experimental Child Psychology*, *171*, 138–147.

Brueggemann, S., Chan, A. B., & Hsiao, J. (2016). Hidden Markov modeling of eye movements with image information leads to better discovery of regions of interest. In *Proceedings of the 38th annual conference of the Cognitive Science Society, Cognitive Science Society*.

Cassin, B., Solomon, S., & Rubin, M. L. (1984). *Dictionary of eye terminology*. Gainesville: Triad Publishing Company.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. *Journal of Vision*, *14*(11), 8–8.

Chuk, T., Chan, A., & Hsiao, J. (2015). Hidden Markov model analysis reveals better eye movement strategies in face recognition. In *Proceedings of the Cognitive Science Society, Cognitive Science Society*.

Chuk, T., Chan, A. B., Shimojo, S., & Hsiao, J. (2016). Mind reading: Discovering individual preferences from eye movements using switching hidden Markov models. In *Proceedings of the Cognitive Science Society, Cognitive Science Society*.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2017). Is having similar eye movement patterns during face learning and recognition beneficial for recognition performance? Evidence from hidden Markov modeling. *Vision research*, *141*, 204–216.

Chuk, T., Crookes, K., Hayward, W. G., Chan, A. B., & Hsiao, J. H. (2017b). Hidden Markov model analysis reveals the advantage of analytic eye movement patterns in face recognition across cultures. *Cognition*, *169*, 102–117.

Citorík, J. (2016). Predicting targets in multiple object tracking task. Master's thesis, Univerzita Karlova, Matematicko-fyzikální fakulta.

Colombo, J., & Cheatham, C. L. (2006). The emergence and basis of endogenous attention in infancy and early childhood. In *Advances in child development and behavior*, (Vol. 34, pp. 283–322): Elsevier.

Coutrot, A., Hsiao, J. H., & Chan, A. B. (2018). Scanpath modeling and classification with hidden Markov models. *Behavior research methods*, *50*(1), 362–379.

De Groot, F., Huettig, F., & Olivers, C. N. (2016). Revisiting the looking at nothing phenomenon: Visual and semantic biases in memory search. *Visual Cognition*, *24*(3), 226–245.

Dink, J., & Ferguson, B. (2015). eyetrackingR: An R library for eye-tracking data analysis.

Doebel, S., Barker, J. E., Chevalier, N., Michaelson, L. E., Fisher, A. V., & Munakata, Y. (2017). Getting ready to use control: Advances in the measurement of young children's use of proactive control. *PloS one*, *12*(4), e0175072.

Doebel, S., Dickerson, J. P., Hoover, J. D., & Munakata, Y. (2018). Using language to get ready: Familiar labels help children engage proactive control. *Journal of Experimental Child Psychology*, *166*, 147–159.

Doran, M., Hoffman, J., & Scholl, B. (2009). The role of eye fixations in concentration and amplification effects during multiple object tracking. *Visual Cognition*, *17*(4), 574.

Drew, T., & Williams, L. H. (2017). Simple eye-movement feedback during visual search is not helpful. *Cognitive Research: Principles and Implications*, *2*(1), 44.

Duchowski, A. T. (2017). *Eye tracking methodology: Theory and practice*. Berlin: Springer.

Erickson, L. C., Thiessen, E. D., Godwin, K. E., Dickerson, J. P., & Fisher, A. V. (2015). Endogenously and exogenously driven selective sustained attention: Contributions to learning in kindergarten children. *Journal of Experimental Child Psychology*, *138*, 126–134.

Faller, J., Cummings, J., Saproo, S., & Sajda, P. (2019). Regulation of arousal via online neurofeedback improves human performance in a demanding sensory-motor task. Proceedings of the National Academy of Sciences, pp. 201817207.

Fehd, H. M., & Seiffert, A. E. (2008). Eye movements during multiple object tracking: Where do participants look? *Cognition*, *108*(1), 201–209.

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

Fernández, G., Castro, L. R., Schumacher, M., & Agamennoni, O. E. (2015). Diagnosis of mild Alzheimer disease through the analysis of eye movements during reading. *Journal of Integrative Neuroscience*, *14*(01), 121–133.

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

Fiebelkorn, I. C., & Kastner, S. (2018). A rhythmic theory of attention. Trends in cognitive sciences.

Fiebelkorn, I. C., Pinsk, M. A., & Kastner, S. (2018). A dynamic interplay within the frontoparietal network underlies rhythmic spatial attention. *Neuron*, *99*(4), 842–853.

Fisher, A. V., & Kloos, H. (2016). Development of selective sustained attention: The role of executive functions. In J. A. Griffin, P. McCardle, & L. S. Freund (Eds.) *Executive function in preschool-age children: integrating measurement, neurodevelopment, and translational research, American Psychological Association, Washington, DC, US*, (pp. 215-237).

Fisher, A. V., Thiessen, E., Godwin, K., Kloos, H., & Dickerson, J. (2013). Assessing selective sustained attention in 3-to 5-year-old children: Evidence from a new paradigm. *J of Experimental Child Psychology*, *114*(2), 275–294.

Forney, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, *61*(3), 268–278.

Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye tracking: A new method to describe infant looking. *Child Development*, *82*(6), 1738–1750.

Friedrich, M., Rußwinkel, N., & Möhlenbrink, C. (2017). A guideline for integrating dynamic areas of interests in existing set-up for capturing eye movement: Looking at moving aircraft. *Behavior research methods*, *49*(3), 822–834.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, *23*(4), 523–552.

Haji-Abolhassani, A., & Clark, J. J. (2013). A computational model for task inference in visual search. *Journal of Vision*, *13*(3), 29–29.

Haji-Abolhassani, A., & Clark, J. J. (2014). An inverse Yarbus process: Predicting observers' task from eye movement patterns. *Vision Research*, *103*, 127–142.

Helfrich, R. F., Fiebelkorn, I. C., Szczepanski, S. M., Lin, J. J., Parvizi, J., Knight, R. T., & Kastner, S. (2018). Neural mechanisms of sustained attention are rhythmic. *Neuron*, *99*(4), 854–865.

Holzman, P. S., Proctor, L. R., Levy, D. L., Yasillo, N. J., Meltzer, H. Y., & Hurt, S. W. (1974). Eye-tracking dysfunctions in schizophrenic patients and their relatives. *Archives of General Psychiatry*, *31*(2), 143–151.

Hyönä, J., Li, J., & Oksama, L. (2019). Eye behavior during multiple object tracking and multiple identity tracking. *Vision*, *3*(3), 37.

Jacob, R., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind*, *2*(3), 4.

Kärrsgård, I., & Lindholm, A. (2003). Eye movement tracking using hidden Markov models. Chalmers tek. högsk.

Katsanis, J., Iacono, W. G., & Harris, M. (1998). Development of oculomotor functioning in preadolescence, adolescence, and adulthood. *Psychophysiology*, *35*(1), 64–72.

Kim, J., Vande Velde, A., Thiessen, E. D., & Fisher, A. V. (2017). Variables involved in selective sustained attention development: Advances in measurement. In *Proceedings of the 39th annual conference of the Cognitive Science Society, Cognitive Science Society*.

Kumar, K., Harding, S., & Shiffrin, R. (2018). Inferring attention through cursor trajectories. In C.K.J.Z.M. Rau, & T. Rogers (Eds.) *CogSci*.

Liu, C. L., Jaeger, S., & Nakagawa, M. (2004). Online recognition of Chinese characters: The state-of-the-art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(2), 198–213.

Luna, B., Velanova, K., & Geier, C. F. (2008). Development of eye-movement control. *Brain and Cognition*, *68*(3), 293–308.

Mahy, C. E., Mazachowsky, T. R., & Pagobo, J. R. (2018). Do verbal reminders improve preschoolers' prospective memory performance? it depends on age and individual differences. *Cognitive Development*, *47*, 158–167.

Mantiuk, R., Bazyluk, B., & Mantiuk, R. K. (2013). Gaze-driven object tracking for real time rendering. In *Computer graphics forum, Wiley online library*, (Vol. 32, pp. 163-173).

Meyerhoff, H. S., Papenmeier, F., & Huff, M. (2017). Studying visual attention using the multiple object tracking paradigm: A tutorial review. *Attention, Perception, & Psychophysics*, *79*(5), 1255–1274.

Mongillo, G., & Deneve, S. (2008). Online learning with hidden Markov models. *Neural Computation*, *20*(7), 1706–1716.

Oakes, L. M., Kannass, K. N., & Shaddy, D. J. (2002). Developmental changes in endogenous control of attention: The role of target familiarity on infants' distraction latency. *Child Development*, *73*(6), 1644–1655.

O'Connor, C., Manly, T., Robertson, I., Hevenor, S., & Levine, B. (2004). An fMRI study of sustained attention with endogenous and exogenous engagement. *Brain and Cognition*, *54*(2), 113–135.

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 symposium on eye-tracking research & applications*, (pp. 141–144): ACM.

Papenmeier, F., & Huff, M. (2010). Dynaoi: a tool for matching eye-movement data with dynamic areas of interest in animations and movies. *Behavior Research Methods*, *42*(1), 179–187.

Péter, A. (2017). Solomon coder. https://solomoncoder.com/, beta version 17.03.22.

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

Pyykkönen, P., Hyönä, J., & van Gompel, R. P. (2009). Activating gender stereotypes during online spoken language processing. *Experimental Psychology*, *57*(2), 126–133.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 779–788).

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive psychology*, *51*(1), 1–41.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, (pp. 91–99).

Renner, P., & Pfeiffer, T. (2017). Attention guiding techniques using peripheral vision and eye tracking for feedback in augmented-reality-based assistance systems. In *2017 IEEE symposium on 3D user interfaces (3DUI)*, (pp. 186–194): IEEE.

Ross, R. G., Radant, A. D., & Hommer, D. W. (1993). A developmental study of smooth pursuit eye movements in normal children from 7 to 15 years of age. *Journal of the American Academy of Child & Adolescent Psychiatry*, *32*(4), 783–791.

SMI (2009). SMI: RED250 technical specification. SensoMotoric Instruments.

Smith, K. J., Valentino, D. A., & Arruda, J. E. (2003). Rhythmic oscillations in the performance of a sustained attention task.

*Journal of Clinical and Experimental Neuropsychology*, *25*(4), 561–570.

Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, *16*(3), 407–419.

Smuc, M., Mayr, E., & Windhager, F. (2010). The game lies in the eye of the beholder: The influence of expertise on watching soccer. In *Proceedings of the Cognitive Science Society, 32*.

Stiller, J., & Radons, G. (1999). Online estimation of hidden Markov models. *IEEE Signal Processing Letters*, *6*(8), 213–215.

Sun, J. C. Y., & Hsu, K. Y. C. (2019). A smart eye-tracking feedback scaffolding approach to improving students' learning self-efficacy and performance in a C programming course. *Computers in Human Behavior*, *95*, 66–72.

Van Dongen, H. P., & Dinges, D. F. (2005). Sleep, circadian rhythms, and psychomotor vigilance. *Clinics in Sports Medicine*, *24*(2), 237–249.

VanRullen, R. (2018). Attention cycles. *Neuron*, *99*(4), 632–634.

Wang, W., Shen, J., & Shao, L. (2018). Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, *27*(1), 38–49.

Zelinsky, G. J., & Neider, M. B. (2008). An eye movement analysis of multiple object tracking in a realistic environment. *Visual Cognition*, *16*(5), 553–566.