Vision Research 103 (2014) 127–142

Contents lists available at ScienceDirect

Vision Research

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

An inverse Yarbus process: Predicting observers' task from eye movement patterns

Amin Haji-Abolhassani *, James J. Clark

Centre for Intelligent Machines, Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec H3A 0E9, Canada

ARTICLE INFO

Article history:
Received 25 August 2013
Received in revised form 30 July 2014
Available online 28 August 2014

Keywords:
Visual-task inference
Attention cognitive model
K-means clustering
Visual search
Eye movement
Hidden Markov model

ABSTRACT

In this paper we develop a probabilistic method to infer the visual-task of a viewer given measured eye movement trajectories. This method is based on the theory of hidden Markov models (HMM) that employs a first order Markov process to predict the coordinates of fixations given the task. The prediction confidence level of each task-dependent model is used in a Bayesian inference formulation, whereby the task with the maximum a posteriori (MAP) probability is selected. We applied this technique to a challenging dataset consisting of eye movement trajectories obtained from subjects viewing monochrome images of real scenes tasked with answering questions regarding the scenes. The results show that the HMM approach, combined with a clustering technique, can be a reliable way to infer visual-task from eye movements data.

1. Introduction

It is well known that low-level visual features, such as color and intensity contrasts, influence eye movements Findlay (1981), Zelinsky et al. (1997). However, it is also observed that the task being performed by the viewer can also influence the pattern of eye movements. For example, someone that is viewing a web page on a computer monitor could be engaged in, among others, the tasks of reading text, searching for a specific object, counting objects, or recognizing faces. Each of these tasks would produce a different pattern of eye movements.The influence of task on eye movements was vividly demonstrated in the celebrated study of Yarbus (1967) who recorded the eye movements of a subject while viewing a painting. The subject was asked different questions regarding the painting, such as to determine the wealth of the family depicted in the painting'. As shown in Fig. 1, different trajectories emerged depending on the specific question that the viewer was answering.

Several other studies have also reproduced the original finding of Yarbus using new equipment and stimuli, and with larger numbers of subjects. For instance, in Tatler et al. (2010) the results obtained by Yarbus were confirmed in an experiment that studied the effect of *instructions* in viewing a portrait of Yarbus. While the effect of visual-task on eye movement pattern has been thoroughly investigated, there has been little done for the inverse process – to infer the visual-task from the eye movements. Knowledge of the visual-task being carried out by a viewer has many potential uses. For example, one can envisage an 'intelligent display' which modifies what is being displayed in a way which facilitates the task. An intelligent web page could detect if a viewer is reading text and highlight or magnify the text, or if it detected the viewer was engaged in a counting or search behavior, it could highlight the target object. The goal of the work described in this paper is to develop such an *inverse Yarbus process*, whereby the visual-task is inferred given measurements of the eye movements of the viewer.

There is some doubt as to whether development of such an inverse Yarbus process is possible at all. In a study by Greene, Liu, and Wolfe (2012), Greene, Liu, and Wolfe (2011) two attempts were made to produce the inverse Yarbus problem. The first approach attempted to train humans to solve the inverse Yarbus problem, while the second tried to train a machine learning system to solve the problem. To obtain data for training and testing they recorded eye movements of several subjects, each performing a specific visual task on an image, and extracted a feature vector from the eye movement records. The feature vector used was a set of seven summary statistics of eye movements, which are often used in scanpath analysis (Castelhano & Henderson, 2008; Mika et al., 1999). This feature vector included, among others, the number of fixations, the mean fixation duration, the mean saccade amplitude and the portion of the image covered by fixations. The machine learning approaches used three different classifiers based on linear discriminant analysis (Mika et al., 1999), correlational

* Corresponding author.
E-mail addresses: amin@cim.mcgill.ca (A. Haji-Abolhassani), clark@cim.mcgill.ca (J.J. Clark).

http://dx.doi.org/10.1016/j.visres.2014.08.014
0042-6989/

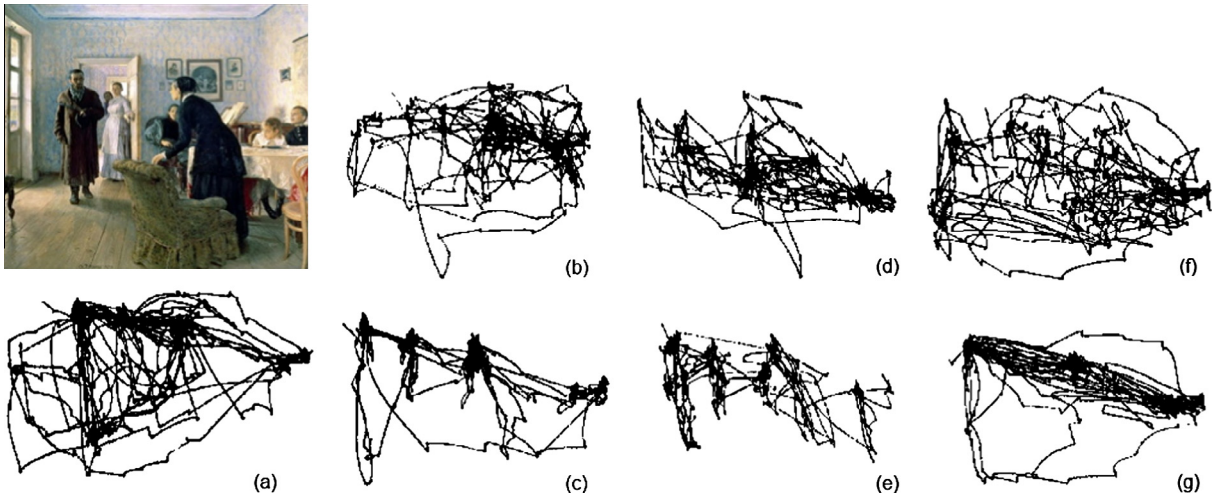


Fig. 1. Eye trajectories measured by Yarbus by viewers carrying out different tasks. (a) No specific task. (b) Estimate the wealth of the family. (c) Give the ages of the people in the painting. (d) Summarize what the family had been doing before the arrival of the "unexpected visitor". (e) Remember the clothes worn by the people. (f) Remember the position of the people and objects in the room. (g) Estimate how long the "unexpected visitor" had been away from the family. Image adapted from Yarbus (1967) with permission from Springer Publishing Company.

methods Haxby et al. (2001) and support vector machines (Hearst et al., 1998). The results showed that both humans and the machine classifiers can only infer the task at a chance level. Based on these results (Greene, Liu, and Wolfe (2011)) concluded that: *"The famous Yarbus figure may be compelling but, sadly, its message appears to be misleading. Neither humans nor machines can use scanpaths to identify the task of the viewer."*. A similar result was obtained in Kanan et al. (2014), where a radial-basis kernel function support vector machine (C-SVN) (Gunn, 1998) was used to classify the eye trajectories represented by their summary statistics. In their results (Kanan et al., 2014) could only achieve an accuracy of 26.3% (95% CI = 21.4–31.1%, $p = 0.61$) which is not significantly better than the chance level.

Summary statistics of eye movements are not sufficient to identify the visual task that was performed by the subject. Castelhano, Mack, and Henderson (2009) looked at the influence of task on a group of summary statistics (including the ones used in Greene's experiment) for the two tasks of memorization and visual search. After considering various features of eye trajectories, they came to the conclusion that the visual-task does not influence the features obtained from individual fixations. A similar result was obtained in Mika et al. (1999), where they also used the same features as in Greene, Liu, and Wolfe (2012). However, even though it is evident that summary statistics are not well suited for implementing an inverse Yarbus process, it may still be the case that other, more informative, features could do the job. For instance, it is shown in Borji and Itti (2014) that using the spatial information along with the summary statistics of the eye movements can marginally improve the results. In their experiment, Borji and Itti (2014) replicated Greene's experiment and showed that by adding the spatial information to the aggregate eye movement features a slightly, but significantly (34.12% correct versus 25% chance level; binomial test, $p = 1.07 \times 10^{-4}$), better accuracy can be obtained in decoding the observers' task.

To motivate our method for implementing the inverse Yarbus process, it is worthwhile to first examine the *forward Yarbus process*, in which the task is given as the input and the measured task-dependent eye trajectory is the output. The first question to ask regarding the forward Yarbus process is what, if anything, determines the gaze direction while viewing a scene. The fundamental premise in this regard is that gaze follows the allocation of *selective visual attention*. Then, the assumption is that viewer task modulates, in some fashion, the allocation of attention, which is then reflected in the overt gaze shifts. Let us first review the approaches that have been developed for modeling visual attention, and then consider how task modulates attention.

1.1. Attention modeling

In every second a vast quantity of visual information enters our eyes, only a fraction of which can be processed by the limited neuronal hardware available to our visual system. However, the human brain has the ability to process the visual information in real time thanks to the mechanisms of *visual attention*. Visual attention is the process that is responsible for selecting a subset of information to be processed in the higher levels of the visual system (Desimone & Duncan, 1995). This selection process can be interpreted as the directing of a *focus of attention* (FOA) to a circumscribed region in the visual field (Niebur & Koch, 1998, chap. 9).

An influential concept in attention modeling is that of *salience*, a term which can be loosely defined as the prominence or conspicuity of region or object in a scene. Salient regions are, in this view, *attractive* to attention, and attention will therefore be preferentially directed to these regions. Gaze shifts would then be expected to follow the attention shifts to these salient points. The extent to which a salience-based model of attention predicts the direction of gaze is often used as a measure of performance for that model.

The earliest saliency-based attention models were *bottom-up* models, which defined salience solely on features derived from the visual input. These models were typically task-independent. In the case of bottom-up attention models, the allocation of attention is based on the characteristics of the visual stimuli, and does not employ any top-down guidance or task information to shift attention. One of the most advanced saliency models is the one proposed by Itti and Koch (2001). In this model the FOA is guided by a map that conveys the saliency of each location in the field of view. The saliency map is built by linearly combining the *feature maps*, which are the outputs from different filters tuned to simple visual attributes, such as color, intensity and orientation (see Fig. 2a).

Although image salience models have been extensively researched and are quite well-developed, empirical evaluation of such models show that they are poor at accounting for actual

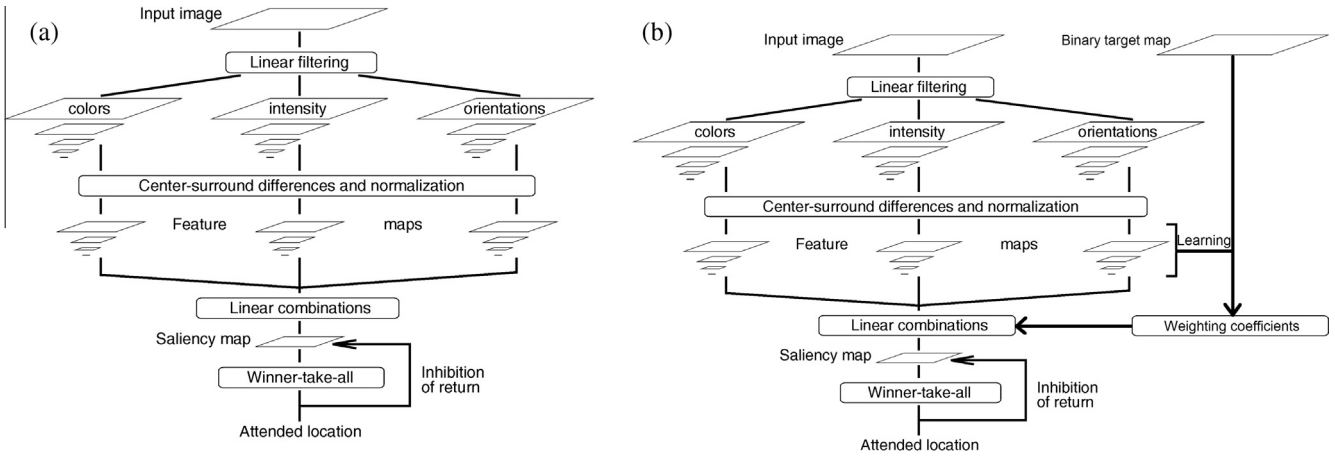


Fig. 2. (a) General architecture of bottom-up attention model by Itti and Koch (2001). The saliency map is built by linearly combining the feature maps which are the outputs from different filters tuned to simple visual attributes, such as color, intensity and orientation. (b) Influence of top-down, task-dependent priors on bottom-up attention models. The influence can be modeled as a weight vector modulating the linear combination of the feature maps (Itti & Koch, 2001).

attention allocations when a visual-task is involved (Einhäuser, Rutishauser, & Koch, 2008). In a study by Judd, Durand, and Torralba (2012) a database of 11,700 eye trajectories obtained from 39 subjects viewing a dataset of 300 natural images was created. The study compared the performance of 10 different salience-based models of visual attention in predicting the eye trajectories of the database. The results of this study indicate that early bottom-up attention models and their variations, such as the saliency model of Itti and Koch (2001), perform only slightly above the chance level when it comes to predicting fixations on natural images. The study also showed that although more recent models, such as the information-theoretic model of Bruce and Tsotsos (2009) and the context-based models of Torralba et al. (2006) and Goferman, Zelnik-Manor, and Tal (2012), perform better than the early models, their performance in predicting the location of fixations is still less than that obtained simply by using the fixations of a single human viewer. This conclusion was based on an evaluation of the accuracy of a fixation map from one human observer in predicting the fixation map of the other 38 observers. The accuracy was averaged over all images in the database and indicated how well the fixation map of a single human can predict an average fixation map of humans. This indicates the importance of incorporating actual eye movement patterns into attention models.

Much of the shortfall in performance of the approaches considered in the Judd, Durand, and Torralba (2012) study can be ascribed to the lack of task-dependence in the models. Attention is not just a passive enhancement of the visual stimuli, rather, it actively selects certain parts of a scene based on the needs of the ongoing visual task. This has led to the development of *top-down* models, which modulate the bottom-up features based on high level reasoning, volition and viewer task (Connor, Egeth, & Yantis, 2004). In a recent study by Borji and Itti (2013) 65 state-of-the-art models of attention were studied and categorized as either bottom-up or top-down. In each category the models were qualitatively compared over 13 experimental criteria. One of these criteria was the accuracy with which a model predicts real-world eye movement patterns as quantified by a spatial correlation coefficient. In order to evaluate the statistical relationship of the saliency models with the eye movement datasets, the recorded eye trajectories can be combined to form a *ground-truth saliency map* that takes on a form similar to that of the saliency map produced by saliency-based attention models. This map, along with other features that are studied in Borji and Itti (2013), are often used in studies of attention to objectively evaluate the models.

Fig. 2b shows an illustration of the interaction between top-down and bottom-up models as proposed by Itti and Koch (2001) and Rutishauser and Koch (2007). In this model different tasks enforce different weight vectors in the linear combination phase. Ehinger et al. (2009) achieved a 94% agreement with human eye movements in a visual search task by combining saliency maps with scene context and target features. Torralba et al. (2006) also used contextual information for facilitating object search in natural scenes. The contextual guidance model of attention uses the bottom-up saliency map, scene context, and top-down mechanisms at an early stage of visual processing and combines them into a unified attention map. Kanan et al. (2009) used the knowledge about how and where objects tend to appear in a scene in order to derive an appearance-based saliency model.

Although salience-based top-down models address the problem of task independence of the bottom-up models, they are based on assumptions that can degrade their performance. The development of salience-based attention models generally proceeded under a *picture-viewing* paradigm, wherein a static 2-dimensional image or photograph was viewed. In addition, such models were typically created to handle simple situations where viewers performed search or detection tasks where the targets can be defined by simple conjunctions of high contrast visual features. Tatler et al. (2011) showed that gaze allocation models that are based on salience are limited in accounting for many aspects of free viewing of complex scenes and often fail when applied in the context of natural (as opposed to artificially constrained search) task performance. They argued for moving away from models based on the picture-viewing paradigm and focusing on the principles governing gaze allocation in a broader range of experimental contexts.

1.2. Linking attention and gaze direction

An important aspect of the (forward) Yarbus process is that attention allocation (suitably modulated by task) determines the direction of gaze. The most straightforward implementation of this is to direct the gaze to the most salient scene point. There is compelling evidence that the mammalian visual-motor system employs such a targeting scheme (Henderson, 1992; Clark (1999)) at least in simple constrained situations. One of the arguments for the use of salience maps in modeling natural visual behavior is that spatial deviations of low-level features from the local surround are cognitively relevant. However, while the contrast of low-level features in fixated locations are shown to be statistically higher than control locations in an image, this correlation

is relatively weak in more complex situations (Mannan, Ruddock, & Wooding, 1997; Parkhurst, Law, & Niebur, 2002; Reinagel & Zador, 1999). This lack of explanatory power for image salience in the context of active tasks is evident in studies of natural tasks such as hitting a ball (Ballard & Hayhoe, 2009; Land & McLeod, 2000), tea making (Land, Mennie, & Rusted, 1999) and sandwich making (Hayhoe et al., 2003). In these tasks saccades are often directed to the expected points of contact, which can exhibit low salience. Due to this lack of explanatory power of image salience models, another class of task-dependent visual attention models is emerging which emphasizes cognitive relevance hypotheses in predicting fixation locations. In cognitive relevance models an object-based representation of the scene is used to select fixation locations based on the needs of the cognitive system in relation to the current task, and saccade targets are ranked based on the cognitive relevance of the objects to the task (Nuthmann & Henderson, 2010). In some hybrid models, the cognitive relevance and image salience are combined to include both low-level, image-based and medium-level, proto-object-based representations of the attentional map into a coherent architecture based on real cognitive behavior of the visual system in the presence of visual task (Wischnewski et al., 2010; Wischnewski et al. (2009)).

Tatler et al. (2011) highlighted another deficiency of simplistic salience models, which is that the decision about where to fixate in these approaches is commonly made by a winner-takes-all process that selects the most conspicuous location on a salience map. This selection criterion, however, fails to account for the decrease in acuity with eccentricity. Moreover, in order to allow attention to move on from the most salient location in the map, these models assume a process known as *inhibition of return (IOR)* to inhibit the focus of attention from returning to the recently attended locations. Although IOR is supported by many classical psychophysical studies (Klein, 1980, 2000; Klein & MacInnes, 1999; Posner & Cohen, 1984), recent empirical evidence in viewing photographic images argues against such an effect (Smith & Henderson, 2009; Tatler & Vincent, 2008). Tatler et al. (2011) wrote of the importance of temporal information about the eye movements, which is usually neglected in the simple salience-based models. The primary goal of salience models is to spatially model fixations, and the temporal aspects of viewing behavior is usually ignored. Evidence from studies of gaze during the performance of natural tasks emphasizes the need to consider fixation duration as well as fixation location (Droll et al., 2005; Hayhoe, Bensinger, & Ballard, 1998; Land, Mennie, & Rusted, 1999).

Another limitation of current salience models lies in their postulating that saccades are precisely directed to the target locations for processing (Tatler et al., 2011). While this appears to be a plausible assumption in simple viewing tasks, in the context of natural tasks this assumption is generally invalid. For instance, Johansson et al. (2001) showed that, for a task of moving an object past an obstacle, foveating the target within 3 degrees of visual angle was sufficient. Similarly, in a tea making task (Johansson et al., 2001) corrective saccades of amplitude less than 2.5 degrees were infrequent, suggesting that, in natural behavior, fixations land close to the targets only in the case of attention demanding targets but typically do not precisely follow the focus of attention. It has long been known that short latency saccades, in which target-directed eye movements are made quickly in response to the onset of a target, frequently miss the target, instead being directed to the *center-of-mass* of the visual grouping of the target object and its surround (Coëffé & O'regan, 1987).

The final aspect of the (forward) Yarbus process to be considered is the link between the gaze direction and the visual-task. While certain statistical features of eye movements remain unchanged across different tasks, the COG tends to be directed to targets that are relevant to the task at hand. This effect can be seen in the eye trajectories of Yarbus, in which the viewers fixated on the targets that were informative for the task. For instance, in the task of age estimation, faces were more likely to get fixated, while for the task of wealth estimation inanimate objects in the room became of more interest to the viewer. Many other studies of eye movements during natural behaviors have likewise indicated that there is a link between the gaze location and informative locations and the immediate task goals (Epelboim et al., 1995; Hayhoe et al., 2003; Land & Furneaux, 1997; Land, Mennie, & Rusted, 1999; Patla & Vickers, 1997; Pelz & Canosa, 2001). In the visual attention model of Schneider (1995), target selection was partially governed by the action being performed. This selection for action was highlighted by the fact that the gaze targets were concentrated in the task-relevant areas in an image while a visual-task was being performed (Hayhoe et al., 2003; Land, Mennie, & Rusted, 1999), whereas before beginning the task, eye fixations were scattered over the image (Hayhoe et al., 2003; Rothkopf, Ballard, & Hayhoe, 2007).

To better demonstrate the gaze deployment under the influence of task, Rothkopf, Ballard, and Hayhoe (2007) carried out a series of experiments conducted in a virtual environment, where subjects executed the two tasks of "approaching" and "avoiding" objects while navigating along a walkway. In these experiments they showed that the distribution of fixations on an object changes according to the task and suggested that human gaze is directed toward regions in a scene determined primarily by the task requirements.

Besides the distribution of gaze locations, visual-task influences other metrics of eye movements. Tatler, Baddeley, and Vincent (2006) showed that visual task also affects the temporal statistics of eye movements in viewing natural images. Castelhano, Mack, and Henderson (2009) looked at eye movements during *memorization* and *search* tasks and showed that the task influences a number of eye movement measures, including the number of fixations and gaze duration on specific objects, while leaving unchanged other parameters, such as the average saccade amplitude and individual fixation durations. They also showed that the task biases the selection of scene regions and temporal measures on those regions. In Johansson et al. (2001) a temporal coupling between vision and action was demonstrated. In their experiment they detected the onset of gaze shifts towards the next target relative to the hand movements as the subject maneuvered an object past an obstacle. The gaze shift was shown to be linked with the execution of the task, as the gaze moved to the next target as soon as the object cleared the obstacle. Temporal coupling between action and vision was also demonstrated for the tasks of driving (Land & Lee, 1994; Land & Tatler, 2001), tea making (Hayhoe et al., 2003), sandwich making (Land, Mennie, & Rusted, 1999), music sight reading (Furneaux & Land, 1999), walking (Patla & Vickers, 2003) and reading aloud (Buswell, 1920). In Land and McLeod (2000) the eye movements of cricket players were studied and it was shown that different skill levels of the players in performing the task generally result in different latencies in directing the gaze towards predicted locations of the incoming ball. This temporal coupling between action and vision shows that models of visual-motor system function must consider task influence on the temporal characteristics of eye movement as well as on the spatial characteristics.

The task also affects the pattern, or sequencing, of eye movements. In the aforementioned study of Land and McLeod (2000) it was shown that while watching a cricket game the gaze is directed according to the ongoing events in the game. In another experiment, the eye movements of subjects were recorded while watching a person stack a set of blocks Flanagan and Johansson (2003). In this block-sorting task, the viewers' gaze was shown to be anticipating the expected points of interaction. In another block-copying experiment (Ballard, Hayhoe, & Pelz, 1995) the eye

movements showed similar patterns through the progression of the task that could be interpreted in terms of momentary information processing needs. Clark and O'Regan (1998) studied the spatial characteristics of eye movements for the task of reading and showed that when reading a text, the *center of gaze* (COG) lands on the locations that minimize the ambiguity of the word arising from the incomplete recognition of the letters. In a seminal study, Treisman and Gelade (1980) developed the *feature integration theory* that modeled the attentional deployment in the task of visual search. In Wolfe, Cave, and Franzel (1989) an improved model called *Guided Search* was suggested that studied how our brain directs attention through a scene during a search task. Hayhoe and Ballard (2005) reviewed the goal-directed behavior of the visual-motor system, and provided a comprehensive set of references to studies of task influence on eye movements.

It can be seen from the material presented in this section that visual task does influence the spatial and temporal patterns of eye movements. It is therefore conceivable that it should be possible to invert this process. At a more general level, eye movement patterns can serve as a window into the brain, and be used to infer the mental states of observers. This has been studied by many researchers. Bulling et al. (2009), Bulling et al. (2011) successfully used eye movement analysis for recognizing the physical activity of subjects while copying a text, reading a printed paper, taking hand-written notes, watching a video, browsing the web or being idle. It would be of obvious utility to know the mental state of people engaged in safety–critical attention-demanding activities such as driving a car or flying a plane. Detection of tiredness or distraction of the operator could be used to trigger alarms or machine backup systems (Di Stasi et al., 2012). As an example of how this could be done, in Di Stasi et al. (2010) the maximum eye velocity during saccadic movements was shown to be inversely proportional to the *mental workload* of subjects in a simulated driving task. In a study by Benson et al. (2012) eye movement analysis was used to detect schizophrenia. In Schleicher et al. (2008) blink duration, delay of lid reopening, blink interval, and standardized lid closure speed were identified as indicators of mental fatigue. These studies share a common conclusion, which is that it is possible to predict an observer's cognitive state by analyzing his eye movement behavior. Continuing along this line of thinking, we consider the visual-task being carried out as an aspect of the cognitive state and therefore aim to predict or infer the task by analyzing the observer's eye movement behavior.

The inversion process should use features of the eye movement trajectories that are more informative than summary statistics, and should be able to model *covert* attention allocation rather than just the position of the eyes (or *overt* attention). The inversion technique should be applicable to complex natural scenes and abstract tasks such as those in the original Yarbus experiment. To this end, we propose to use Hidden-Markov-Models (HMMs) to relax the inherent assumptions in the simplistic salience models and use real-world eye movements to train task-dependent models that can infer the visual-task on natural images. In the following sections of the paper we will show how HMMs accomplish this by modeling the fixation distributions with a Gaussian distribution function that allows for fixations well away from the target (assumed to be associated with the covert attentional locus). Moreover, by analyzing the eye trajectories as time-series we give the temporal features of eye movements the same importance as the spatial features. The modeling of the cognitive relevance of the low-level features is facilitated by the HMM approach, as the Gaussian distributions are allowed to move away from salient objects to more cognitively relevant targets in an image. The Gaussian distributions also account for overshooting and undershooting of targets when directing the gaze. Consequently, the assumption of precise targeting inherent in the salience models is relaxed in the HMMs by using the observation distributions over the targets. Moreover, using HMMs to model the transition of the attentional focus from one location to another overcomes the inherent shortcomings of the target selection processes used in the salience models. In an HMM-based model the target selection is governed by a statistical process that is trained on natural eye trajectories measured during task execution, which replaces processes such as winner-takes-all and inhibition of return that are associated with target selection in salience-based models.

1.3. Attention tracking using Hidden Markov models

In the previous section we observed that classical models of attention are limited in terms of accounting for real-world eye movements of observers while viewing natural images. This can be seen in the benchmark presented in Judd, Durand, and Torralba (2012), which compared the performances of salience models in predicting eye fixations made on natural images. One of the most striking experiments done in this study was to compare the performance of the best salience model and a model based on real eye trajectories. It was shown that even the best model performs worse than the fixation map of just one human observer in terms of prediction rate of the eye trajectories. Thus, we base the development of our attention model on actual task-dependent eye trajectories recorded while viewing natural images. To do so, we use Hidden Markov models (HMMs) as a tool for time-series analysis of the eye trajectories to encode the dynamics of natural eye movements into task-dependent models. One of the benefits of our HMM model is its trainability on natural eye movements to capture their spatial and temporal patterns rather than purely depending on analyzing the patterns of image features in fixated regions, as done in the salience models.

Hidden Markov models (HMMs) are a group of generative models that are used in supervised and semi-supervised learning (Rabiner, 1990). Similar to the first-order, finite-state, discrete-time Markov chain (DTMC), HMMs govern the transition between the states by a first-order Markov process.

A typical DTMC can be defined by a set of parameters, $\gamma = \{A, \Pi\}$, where:

- $A = \{a_{ij}\}$ is *the state transition probability distribution*

$$a_{ij} = P(q_{t+1} = s_j | q_t = s_i), \qquad 1 \leqslant i, j \leqslant N \tag{1}$$

- $\Pi = \{\pi_i\}$ is *the initial state distribution*
- π_i is the probability of starting a sequence at state i

$$\pi_i = P(q_1 = s_i), \qquad 1 \leqslant i \leqslant N \tag{2}$$

- $q_t \in S$ and $1 \leqslant t \leqslant T$ is the state at time t
- $S = \{s_1, s_2, \ldots, s_N\}$ is the *state space*
- N is the number of states in the model

In a more general view, both HMMs and DTMCs are classes of finite state machines (FSMs) (Bengio & Frasconi, 1995) that at each time step generates an observation sample vector $\vec{O}_T$ $(t \in [1, T])$ according to the state currently being visited. Therefore, in each traverse of these FSMs we will obtain an observation sequence **O**, where:

- **O** is a sequence of T observations $(\vec{O}_1, \vec{O}_2, \ldots, \vec{O}_T)$
- $\vec{O}_t$ $(t \in [1, T])$ is an observation sample vector consisted of M feature values $(o_{t,1}, o_{t,2}, \ldots, o_{t,M})$
- M is the number of feature values in each observation.

In a DTMC, each state can only generate a specific set of observation vectors, meaning that there is no overlap between the

observation vectors of different states. Fig. 3a shows a DTMC with two states (i.e., $N = 2$). At time step $t = 0$, the process starts by entering one of the states s_1 or s_2 with the probability of π_1 and π_2, respectively. In the following time steps, the process chooses the next state according to the transition probabilities a_{ij}. At each time step an observation is generated according to the current state, which reveals the current state of the DTMC.

Fig. 3b shows a sample state sequence of the process, $\{q_t : 1 \leqslant t \leqslant 3\}$, where $q_t \in \{s_1, s_2\}$ is the state that the sequence is visiting at time t. The overall observation sequence is in the form $\{\vec{O}_t : 1 \leqslant t \leqslant 3\}$, which is equivalent to a unique state sequence due to the non-overlapping characteristic of the observation space between the states.

The Markov process of an HMM is also defined by the parameters of the underlying DTMC. The only difference between the DTMC and the HMM is that in HMMs the observations are generated according to a state-specific density function, B, called *the observation pdf*. In contrast to the observations of a DTMC, in an HMM the observation pdf of different states can overlap and might generate the same observation as the output. Therefore, in HMMs we cannot directly map an observation to a unique state, which makes the states hidden to the observer.

A typical discrete-time, continuous HMM, λ, can be defined by a set of parameters, $\lambda = \{A, B, \Pi\}$, where $B = \{b_j(\vec{O}_t)\}$ is the *observation probability density function* in the state j and

$$b_j(\vec{O}_t) = P(\vec{O}_t | q_t = s_j), \qquad 1 \leqslant j \leqslant N, 1 \leqslant t \leqslant T \tag{3}$$

Fig. 4a shows an HMM with two states, similar to the DTMC shown in Fig. 3a. In this example, each observation (i.e., $\vec{O}_t$) is a 2D vector generated according to the state-specific, 2D Gaussian distribution functions.

Fig. 4b shows a sample outcome of the HMM of Fig. 4a. The outcome of the process is an observation sequence $\{\vec{O}_t : 1 \leqslant t \leqslant 3\}$, where $\vec{O}_t$ is the observation at time t.

As mentioned in Section 1.1, classical attention models are based on a spatial saliency map that defines the conspicuous locations, which are potential targets of the fixations. In addition to the saliency maps, high-order processes are also observed to influence the selection of targets in an eye movement trajectory and are used as a source of information for attention allocation. *Proximity preference* is a cognitive process that facilitates fixations near the currently fixated target and *similarity preference* is a cognitive process that favors objects similar in appearance to the one that is currently being fixated (Koch & Ullman, 1985). *Inhibition of return* (IOR) (Klein, 2000) is a high-level process that discourages fixation on the target that have been visited in the preceding period of time.

These higher-level processes affect the selection of the next target based on the recently fixated ones, which suggests a Markov cognitive process as the target selection model of the visual motor mechanism. The HMMs use Markov processes as their underlying

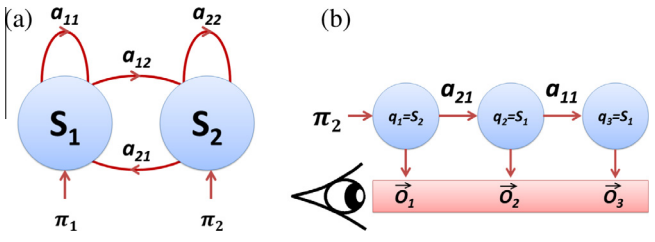


Fig. 3. (a) A first-order, finite-state, discrete-time Markov chain (DTMC) with two states (i.e., $N = 2$). The DTMC is defined by a *state space* $S = \{s_i : 1 \leqslant i \leqslant N\}$, a *state transition matrix* $A_{N\times N} = \{a_{ij} : 1 \leqslant i,j \leqslant N\}$ and a set of *initial state distribution* $\Pi = \{\pi_i : 1 \leqslant i \leqslant N\}$. (b) A sample trajectory that is generated by the DTMC. In the trajectory the states are overt and the observer can see which state is visited at each time step.

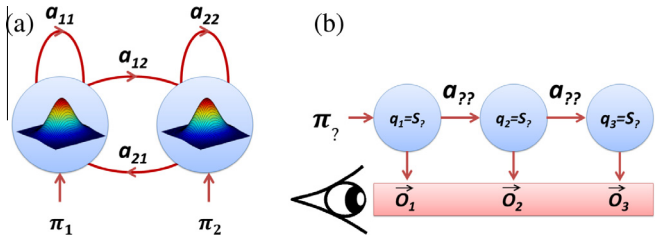


Fig. 4. (a) A HMM with two states (i.e., $N = 2$) is shown in this figure. In addition to the parameters of the underlying DTMC (i.e., A and Π), an HMM has an extra parameter called *the observation pdf*, B, which gives the probability distribution over different observations in each state. (b) In the trajectories generated by HMMs, the state sequence is hidden to the observer and at each time step, an observation is generated according to a density function, B.

models for generating time-series observations similar to the eye trajectories. This feature of HMMs allows us to incorporate these higher-level processes into a coherent model for eye movement analysis.

Markov processes have been considered in prior studies on eye movement generation, and have been shown to generate similar patterns to those produced by the mammalian visual-motor system. Hacisalihzade, Stark, and Allen (1992) used Markov processes to model visual fixations of observers. They showed that the eyes visit the features of an object cyclically, following somewhat regular scanpaths[1] rather than crisscrossing it at random. Stark and Ellis (1981) also proposed using Markov processes as a general model of fixation placement during the task of reading. Pieters, Rosbergen, and Wedel (1999) observed a similar pattern in the scanpaths of the observers while looking at printed advertisements.

If we consider each target in an image as a state, the saliency map and the Markov process define the probability of transitions from one state to another in an eye trajectory. This interpretation forms a finite-state, discrete-time Markov chain that gives us the likelihood of an eye trajectory based on the loci of fixations. Moreover, if we posit a first-order Markov process as the underlying process that governs the transitions between the targets (which was shown to be a valid assumption for eye movements (Hacisalihzade, Stark, & Allen, 1992)), we can train a first-order DTMC for each task. This model was used by Elhelw et al. (2008), where they successfully used a first-order DTMC to model eye movement dynamics.

One of the main deficits of classical models is that they assume that tracking the FOA is equivalent to tracking the COG. However, as noted in the introduction, the COG does not necessarily follow the FOA and in fact they can be quite some distance from each other. Fig. 5a shows an eye trajectory recorded when a viewer was asked to count the number of people in the image. While fixations mainly land on the targets of interest (*overt attention*), the person on the left does not get any fixation. The fact that the answer given by the viewer to the question was correct suggests that the COG does not necessarily follow the FOA and sometimes our *awareness* of a target does not imply foveation on that target (*covert attention*).

The disparity between the FOA and the COG can be attributed to several other factors other than covert attention. Accidental attention-independent movement of eye, eye-tracking equipment bias, undershooting or overshooting of the target (Becker, 1972), or the phenomenon of *center-of-gravity fixations* (Zelinsky et al., 1997; He & Kowler, 1989; Najemnik & Geisler, 2005) are some of the most common sources of recurrent divergence between the COG and the FOA.

[1] Repetitive and idiosyncratic eye trajectories during a recognition task is called scanpath (Noton & Stark, 1971).

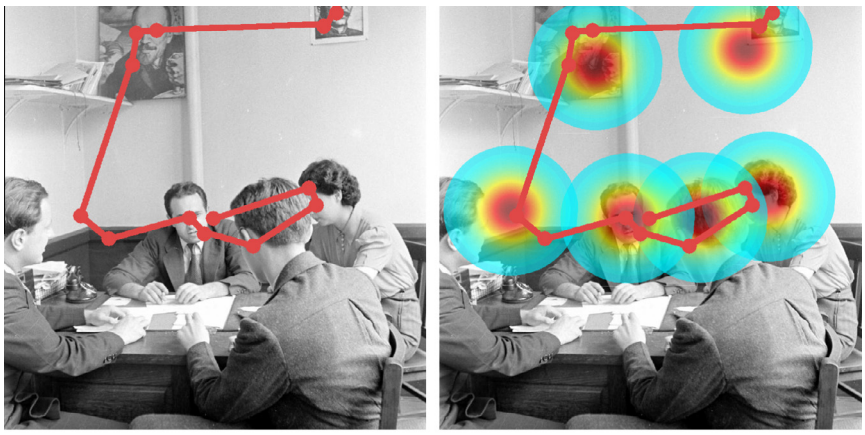


Fig. 5. (a) Eye trajectories recorded while executing the task of "counting the number of people in the image". In the trajectories straight lines depict saccades between two consecutive fixations (shown by dots). While the viewer gave the correct answer in the trial, one of the targets (the leftmost person in the image) is not fixated. The target that did not get any fixations is assumed to have been attended covertly. (b) Overlay of the 2D observation Gaussian distributions on top of an image. The combination of the Gaussian pdfs form the HMM that is trained for the task of counting faces on the image. The overall model can generate synthetic eye-trajectories based on the parameters of the HMM. The transitions between the states are governed by the transition probabilities, and at each time step, the state's observation pdf generates the 2D coordinates of the next fixation. The trajectory shown in the image is the real eye movements of a viewer while performing the task. As we can see, all fixations are covered by the observation pdfs, which makes the whole trajectory a plausible outcome of the HMM.

In terms of a DTMC-based model of attention, the coordinates of the COGs recorded by the eye tracker comprise the observation sequence, which is taken as equivalent to the attentional states. Thus, in the DTMC approach covert attention, where the attentional state is different than the gaze coordinates, cannot be modeled by classical models of attention. However, in HMMs the states are hidden and can be different than the overt observations. Therefore, in our view, the HMMs can serve as a better alternative to the DTMCs in modeling the overt and covert shifts of attention. When entering a state of a HMM, a Gaussian distribution function generates an observation that is overt to the viewer (Rabiner, 1990). Thus, in our proposed model the states represent the FOAs, and the COGs form the observation sequences. In terms of the problem at hand, the hidden states of the HMM correspond to the covert attention loci and the observations of the HMM correspond to the eye positions or overt attention loci.

Fig. 5b shows an HMM that is trained on the eye trajectories recorded while executing the task of counting people in the image. Each 2D Gaussian probability density function (pdf) is depicted by a *heat map*, where the heat represents probability values. Each Gaussian pdf represents the distribution or probability of an attentional state, and at each time step a COG coordinate pair is generated by drawing a random outcome from these pdfs. For instance, directing the FOA (covert attention) to the face of the person on the left can result in a fixation (overt attention) that is further away from the physical boundaries of the face. The capability of Gaussian HMMs in representing off-target fixations is illustrated in this image by overlaying the trajectory of Fig. 5a on the image. While the classical salience-base attention models fail to account for off-target fixations, here we show that the Gaussian observation function can properly model them.

The theory of HMMs has been used in different fields, such as speech recognition (Rabiner, 1990), anomaly detection in video surveillance (Nair et al., 2002) and hand writing recognition (Hu, Brown, & Turin, 1996). HMMs have also been used in analysis of eye movements. In Salvucci and Goldberg (2000) HMMs were used to automatically label the recorded eye movements as either fixations or saccades. In another study (Salvucci & Anderson, 2001) developed an HMM-based model for analysis of eye movements during the task of equation solving. Simola, Salojärvi, and Kojo (2008) modeled three cognitive states of visual process during a reading task by the hidden states of HMMs. Van Der Lans et al. (2008) split a visual search task into two stages of *localization* and *identification* and mapped each of these cognitive states into one of the states of a two-state HMM.

Recently Haji-Abolhassani and Clark (2013) showed that HMMs can also serve as a good model for the visual attention process. They proposed an attention model that allowed for covert shifts of attention as well as overt ones. They used their model in tracking attention during visual search tasks that were conducted on synthetic images. However, in their model they assumed that the targets can be defined in advance and built their model based on the known location of targets. For instance, in the task of counting faces in Fig. 5b, they assumed that the foci of attention will be on the faces. This assumption is valid for simple tasks with objective results (such as number of red objects and number of horizontal bars), but in more abstract tasks, such as the one used in the Yarbus (1967) and the Greene, Liu, and Wolfe (2012) experiments, defining the potential targets of attention is not straightforward. Another problematic aspect of the Haji-Abolhassani and Clark model is that the number of states has to be defined before training. This is only possible in images with a predefined number of targets (as in the synthetic images used in their experiments). However, in natural scenes the targets can appear anywhere in the image and typically no prior information about the location of the targets is available to the model.

In this paper we present an HMM-based attention model that can be applied on natural images. The approach begins by first using the *K*-means clustering technique (Kaufman & Rousseeuw, 2009) to locate potential targets in an image and then using the HMM-based method to decode the eye trajectories. The overall method is then used to infer the visual-task in the same dataset that was used in Greene, Liu, and Wolfe (2012).

2. An Inverse Yarbus process via Bayesian inference

HMMs are a class of semi-supervised learning methods. Being generative models, they classify the test data in a probabilistic manner that can be readily applied to the Bayesian inference framework. One of the many advantages of Bayesian inference is the ability to merge other sources of information in the a priori

term and give an a posteriori probability distribution function over the possible outcomes, whereby a higher level process can make an inference and select a task as the result.

Suppose observations obtained from the eye tracker are in the form of $\langle \mathbf{O}, \theta \rangle$, where $\theta \in \Theta$ is the task label, selected from the set of all task labels Θ, and $\mathbf{O}$ is the observation sequence of fixation locations $(\vec{O}_1, \vec{O}_2, \ldots, \vec{O}_T)$. Each $\vec{O}_i$, itself, is a vector (x_i, y_i) containing the coordinates of a fixation at time i.

In order to infer the visual-task from an eye trajectory we need to evaluate the posterior probability of different tasks, $\theta \in \Theta$, given an observation sequence, $\vec{O}$. According to Bayes rule, this can be calculated as:

$$P(\theta|\mathbf{O}) = \frac{P(\mathbf{O}|\theta)P(\theta)}{P(\mathbf{O})} = \frac{P(\mathbf{O}|\theta)P(\theta)}{\sum_{\theta' \in \Theta} P(\mathbf{O}|\theta')P(\theta')}. \quad (4)$$

In this equation $P(\theta)$ is the *prior probability* of each task $\theta \in \Theta$, and $P(\mathbf{O}|\theta)$ is the task conditional distribution, which is also referred to as the *likelihood* function. The prior distribution assigns a probability distribution to the tasks based on our prior knowledge. This is where we can apply other sources of information about the tasks and improve the inference. The likelihood term of the equation gives the probability of observing the sequence $\mathbf{O}$ while executing the task θ. The likelihood term can be broken down to the conditional probabilities:

$$\begin{aligned} P(\mathbf{O}|\theta) &= P(\vec{O}_1, \vec{O}_2, \ldots, \vec{O}_T|\theta) \\ &= P(\vec{O}_1|\theta)P(\vec{O}_2|\vec{O}_1, \theta), \ldots, P(\vec{O}_T|\vec{O}_1, \ldots, \vec{O}_{T-1}, \theta). \end{aligned} \quad (5)$$

In classical saliency-based attention models, the likelihood can be quantified as proportional to the amplitude of the saliency map on different targets in the image. Fig. 6 shows an example of a saliency map obtained using the *Saliency Toolbox* Walther and Koch (2006). Fig. 6a shows a synthetic image that comprises a combination of "A" symbols and horizontal and vertical bars in three different colors. The objects are placed at the vertices of a 5×6 grid, on a featureless black background. Fig. 6b shows the bottom-up saliency map according to the attention model of Itti and Koch (2001) shown in Fig. 2a. The feature maps are obtained by applying color, intensity and orientation filters to the input image and integrated into the saliency map by a linear combination.

While bottom-up models combine the maps with constant weights, top-down models (shown in the block diagram of Fig. 2b) modulate the weights according to the task Itti and Koch (2001). Fig. 6c shows the saliency map of the same image tuned to the task of "searching for the characters". As we can see, the locations of the characters are more conspicuous (lighter) in the top-down saliency map.

Since in the bottom-up models the allocation of attention is merely based on the characteristics of the visual stimuli, the fixation locations are independent of the ongoing task (i.e., $P(\mathbf{O}|\theta)$ is assumed to be equal to $P(\mathbf{O})$) and the likelihood term becomes:

$$P(\mathbf{O}|\theta) = P(\mathbf{O}) = P(\vec{O}_1, \vec{O}_2, \ldots, \vec{O}_T). \quad (6)$$

In top-down models, saliency of the targets is modulated by the task, which makes the likelihood term task-dependent. Moreover, if we use a discrete-time Markov chain (DTMC) to model high-level processes (Hacisalihzade, Stark, & Allen, 1992) such as inhibition of return (Klein, 2000), proximity and similarity preference (Koch & Ullman, 1985), the likelihood term of Eq. (5) reduces to:

$$\begin{aligned} P(\mathbf{O}|\theta) &= P(\vec{O}_1, \vec{O}_2, \ldots, \vec{O}_T|\theta) \\ &= P(\vec{O}_1|\theta)P(\vec{O}_2|\vec{O}_1, \theta), \ldots, P(\vec{O}_T|\vec{O}_{T-1}, \theta). \end{aligned} \quad (7)$$

In Eq. (7) each task, θ, is represented by the corresponding DTMC that is trained on the eye movements of the viewers who performed that task. The task-dependent DTMCs are represented by $\gamma = \{A, \Pi\}$, a 2-state example of which is shown in Fig. 3a. Since in DTMCs each sequence of fixations corresponds to a unique sequence of states having the parameters of the DCMCs, calculating the likelihood term is a matter of multiplying the state transitions that emerge in the trajectory.

$$\begin{aligned} P(\mathbf{O}|\theta) &= P(\vec{O}_1|\theta)P(\vec{O}_2|\vec{O}_1, \theta), \ldots, P(\vec{O}_T|\vec{O}_{T-1}, \theta) \\ &= P(q_1|\theta)P(q_2|q_1, \theta), \ldots, P(q_T|q_{T-1}, \theta) \\ &= P(q_1|\theta)P(a_{q_2 q_1}|\theta), \ldots, P(a_{q_T q_{T-1}}|\theta). \end{aligned} \quad (8)$$

In the HMMs, however, the states are hidden and the likelihood term cannot be evaluated directly. In the theory of HMMs there are three fundamental problems: *evaluation*, *decoding* and *training*. Assume we have an HMM λ and a sequence of observation $\mathbf{O}$. Evaluation or scoring is the calculation of the probability of the observation sequence given the HMM, i.e., $P(\mathbf{O}|\lambda)$. Decoding is the process of finding the best state sequence that can give rise to the observation sequence. Finally, training is the adjusting of model parameters to maximize the probability of generating a given training observation sequence. The algorithms that cope with evaluation, decoding and training problems are called the forward, Viterbi and Baum–Welch algorithms, respectively (see Rabiner (1990) for details).

In order to find the likelihood term of Eq. (7) we need to solve the evaluation problem for λ_θ, which is the HMM trained to the task θ using the Baum–Welch algorithm on the training database of task-dependent eye trajectories. The method used in Rabiner (1990) to calculate the term $P(\mathbf{O}|\lambda_\theta)$ is an iterative method based on dynamic programming called *forward algorithm*. In this method we define $\alpha_t(i) = P(\vec{O}_1 \vec{O}_2, \ldots, \vec{O}_t, q_t = i|\lambda_\theta)$ as the probability of observations $\vec{O}_1$ to $\vec{O}_t$ with state sequence terminating in state $q_t = s_i$, given HMM λ_θ. We can, then, estimate the probability $P(\mathbf{O}|\lambda_\theta)$ by iterating over the following steps until the termination criterion is met:

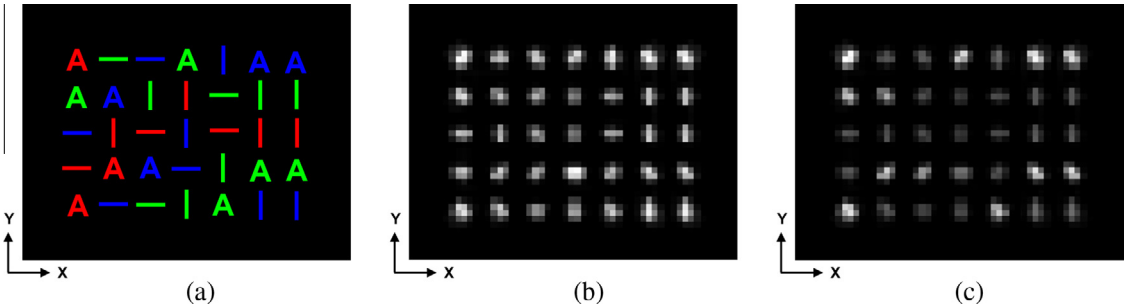


Fig. 6. (a) Original Image. (b) Saliency map of the bottom-up attention model presented in Itti and Koch (2001). (c) Saliency map of the same image using a top-down attention model (Itti & Koch, 2001).

- *Initialization*: $\alpha_t(i) = \pi_i b_i(\vec{O}_1), 1 \leqslant i \leqslant N$,
- *Induction*:
  $\alpha_{t+1}(j) = [\Sigma_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(\vec{O}_{t+1}), 1 \leqslant t \leqslant T-1 \quad and \quad 1 \leqslant j \leqslant N$,
- *Termination*: $P(\mathbf{O}|\lambda_\theta) = \Sigma_{i=1}^{N} \alpha_T(i)$.

With T observations and N states, we require approximately N^2T operations.

3. State positioning of HMMs using *K*-means clustering

So far we have explained how we can use the parameters of a task-dependent HMM λ_θ to infer the underlying task of an eye trajectory $\mathbf{O}$. However, the training to obtain the parameters of λ_θ still remains to be explained. In order to train task-dependent HMMs, we first need to come up with a design for a *generic HMM*. We can then use task-dependent eye trajectories to train the generic HMM by using the Baum–Welch method to make *task-specific* HMMs.

As we explained before, in HMMs the states are hidden and only the observations are overt to the viewer. Therefore, in application to the problem of attention tracking, we used the states to represent the FOA and used the coordinates of the COG as the observations. In our model, each state is composed of a 2D Gaussian observation pdf that is centered on a target. Fig. 5b shows an example of an HMM trained for the task of "counting the number of faces in the image". As we can see, targets learned for this task roughly correspond to the faces in the image. The positioning of the Gaussian pdfs in a synthetic image with discrete targets, such as the one shown in Fig. 6a, is also straightforward as we can assign a state to each of the targets in the image and remove the task-irrelevant ones during the training.

However, positioning of the states' observation pdfs is not always trivial. When executing tasks, such as the ones used in Greene, Liu, and Wolfe (2012) (e.g., "Memorizing the picture" or "determining the wealth of the people in the picture"), on natural images, predefining the attentional targets in the generic HMM needs to be done manually and requires knowledge about the relevance of the objects in the image to the task.

In order to automatically position the observation pdfs of the generic HMM on task-relevant objects, we use a clustering technique to locate the "hot spots" that are informative for execution of the task. To do so, we propose to use *K*-means clustering (Kaufman & Rousseeuw, 2009) on the ensemble of the fixations of the training set. Since the training set comprises all the fixations of the subjects performing a specific task, the ensemble reveals the potential attention demanding targets in the image for that task.

Fig. 7b and c shows the gaze opacity maps of a training set of eye movements recorded while performing the task of "determining how well the people in the picture know each other (people)" and "determining the wealth of the people in the picture (wealth)" on the image of Fig. 7a. The gaze opacity map is obtained by applying a mask overlaying the image. The opacity of this mask at a given point in the image is inversely proportional to the number of fixations in a region about this point. In these maps the areas with a large number of fixations are shown clearly, whereas the areas with no, or few, fixations are masked. As can be seen, the areas near the faces get more fixations in the *people* task and the areas around objects such as the telephone, tie, pipe and the objects on the desk, are more likely to get fixated in the *wealth* task.

By using this simple technique we can get a sense of the conspicuous locations for different tasks with a computational complexity of $O(n)$ Xu and Wunsch (2005). The *K*-means clustering will provide us with K points that indicate the centroids of the top K fixated areas in the training set. In the generic HMM, we will use these centroids as the initial means of the observation pdfs of K states. This initial placement of the 2D Gaussians of the generic HMM on the image, however, is only an estimate of their eventual positions, which may change during the Baum–Welch training.

4. Experiment

To validate our HMM-based approach, we carried out an experiment in which human observers carried out abstract tasks while viewing photographs of complex natural scenes. In order to benchmark our results against those of Greene, Liu, and Wolfe (2012), we used the same database of natural images as they used in their experiment. The image set comprises 64 gray-scale photographs taken from the LIFE magazine photo archive hosted by Google and photo archive hosted by Google (2013), an example of which is shown in Fig. 7a. The date of the images span the years between 1930 and 1979. In each image there are at least two people, and the images do not display faces or locations that were familiar to our test subjects.

For the sake of comparison of the results, in building a database of task-dependent eye trajectories, we followed the same procedure as in Greene, Liu, and Wolfe (2012). In total, we ran 1280 trials and recorded the eye movements of five subjects while performing a set of pre-defined visual-tasks. Five graduate students (one female and four males), aged between 18 and 30, with normal or corrected-to-normal vision volunteered to participate in this experiment. We used the same four tasks as in the Greene et al. experiment:

- Memorize the picture (*memory*).
- Determine the decade in which the picture was taken (*decade*).
- Determine how well the people in the picture know each other (*people*).
- Determine the wealth of the people in the picture (*wealth*).

The images were displayed on a 1920×1080 pixel LCD monitor with a screen size of 53.3×30 cm. The viewing distance was 45 cm. Each image had a resolution of 800×800 pixels, which subtended 28 degrees of visual angle. The background pixels were all set to black.

Each subject did four segments of trials during his/her experiment. Each segment consisted of four blocks of 16 images. The subject was informed of the task by an instruction image at the beginning of each block. During each segment, each of the 64 images were displayed once and subjects had 10 s to view each image. In order to better engage the subjects in the tasks, after each image in the "decade", "people" and "wealth" blocks, a question in form of a five-alternative-forced-choice was presented to the subject. The subjects were asked to select the best answer by clicking on one of the five choices. (We used the same routine and questions as in the Greene et al. experiment.)

After each segment, a mandatory rest period was assigned to the subject, followed by the next segment of 64 images. In each segment we rotated the task order so that each subject performs all the tasks on all the images. In the end, we obtained five trajectories per task, per image, from which we selected the test and training set using leave-one-out (LOO) cross-validation.

A Tobii X120 eye tracker was used to record the participants' eye positions, running at an acquisition rate of 120 Hz. The eye tracker's spatial resolution is approximately 0.2° and its accuracy following calibration is about 0.5^o. The subjects used both eyes when conducting the experiments.

At the beginning of each segment, we calibrated the eye tracker using the built-in, five-point, changing diameter, moving dot calibration routine in Tobii Studio software (ver. 3.2.0, Tobii

Fig. 7. Compilation of the fixation spots during two visual-tasks in the form of opacity maps. (a) The original image on which the tasks were executed. (b) The gaze opacity for the task of "determining how well the people in the picture know each other (people)". (c) The opacity map for the task of "determining the wealth of the people in the picture (wealth)".

Technology, Stockholm, Sweden). The calibration grid spanned the entire display.

After the recording of the eye movements, data analysis was carried out on each trial, wherein we removed the blinks and outliers from the data and classified the eye movement data as either saccades or fixations using the velocity-threshold identification (I-VT) method provided in the Tobii Studio software. The outliers were all fixations that appeared to be out of the screen area, which might have been caused by errors in the interpolation or real fixations at points outside of the screen boundary. It is generally agreed that visual and cognitive processing primarily occurs during fixations and little or no visual processing can be achieved during a saccade (Fuchs, 1971). Therefore, in our analysis we only considered the fixation points.

4.1. Methods

In this experiment we use the proposed HMM-based model to infer the visual-task. The inference is made by applying Bayes rule (Eq. (4)) to the likelihood term calculated by the forward algorithm. A uniform distribution is used for the a priori task probabilities, which makes the inference a maximum likelihood estimation of the task. However, in practical applications we typically would have some prior information about the tasks, which can be applied to the a priori term and increase the accuracy of the inference.

In order to obtain the likelihood term ($P(\mathbf{O}|\theta)$), we need to train the parameters of an HMM for each task (θ) by using the training eye movements of the corresponding task. To do so, first we need to define the structure of the generic HMM and then customize it by training it with eye movements of that task. For the generic HMM we assign an ergodic, or fully connected, structure wherein we can go to any state of the model in a single step no matter what the current state of the model.[2] This is consistent with the characteristics of eye movement, where we can also move our COG to any target in a given stimulus.

As explained in Section 3, we use K-means clustering to define the initial locations (means) of the observation pdfs in the generic HMM. For each task-image pair, we examine different values for the number of clusters ranging from $K = 2$ to 10 and use the value that gives the maximum a-posterior probability (MAP) to the training data, i.e.:

$$K = \arg\max_{K=2:10} P(\langle \mathbf{O}, \theta \rangle_{training} | N = K), \tag{9}$$

where N is the number of states. If we use a very small number of clusters, the HMM will not be able to capture the transition patterns between the objects and will be less task-dependent. On the other hand if we assign a large number to K, the training algorithm will diverge and will not find a feature set that maximizes the likelihood of the training set. We expect that the value of K will be highly dependent on the number of task relevant targets in an image. For instance, for the people task model ($\theta = people$) of the image in Fig. 8, where we have six faces, $K = 6$ gives us the best result, suggesting that a 6-state HMM would be the best choice for λ_{people} of the image.

To define the covariance of the Gaussian distributions, we use a technique called *parameter tieing* (Rabiner, 1990) to force a unique covariance matrix across all the Gaussian distributions. We also fix the off-diagonal elements of the covariance matrix to zero, which leads to fully circular Gaussian observation distributions:

$$COV(B) = \sigma^2 I(N), \tag{10}$$

where $I(N)$ is the identity matrix of size $N \times N$. These two provisions allow us to obtain convergence in training the HMMs with the very limited number of observations in the training database, since the number of parameters to train the covariance matrices decreases from $3K$ to 1. Moreover, a fully diagonal covariance matrix results in a circularly-symmetric Gaussian distribution, which is similar to the quasi-circular FOA of the human visual system (Eriksen & James (1986)).

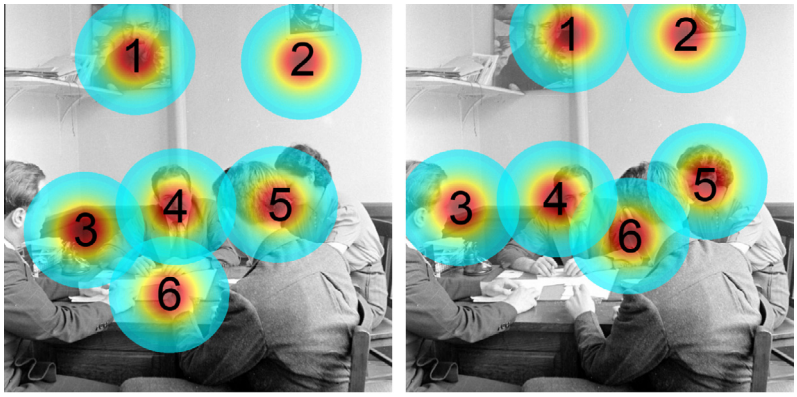
For defining the standard deviation (σ) used in the covariance matrix we tested several values ranging from 14 pixels (0.5°) to 210 pixels (15°) in 14 pixels steps (0.5°) and obtained the best result for 126 pixels (4.5°).

As stated in Rabiner (1990), a uniform distribution assumption suffices as the initial pdf of the *initial state distribution* (Π) and *the state transition probability distribution* (A).

Having defined the structure of the generic HMM, we can obtain a task-dependent HMM by training it with task-specific eye trajectories by using the expectation maximization-based (EM-based) algorithm of Baum–Welch Rabiner (1990).

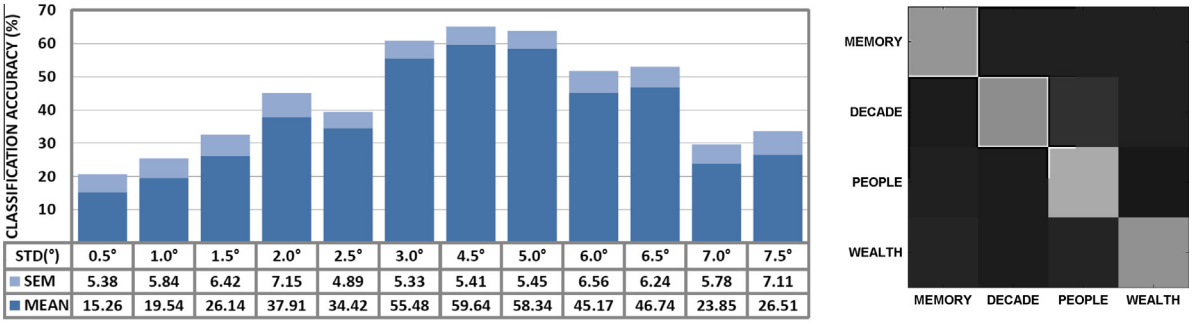
Fig. 8a shows the generic HMM for the task of *people*, superimposed on the original image. The standard deviation of the observation distribution is set to 126 pixels and $K = 6$ centroids are used for clustering. The result of training the generic HMM to the task-specific trajectories of the *people* task is shown in Fig. 8b. As we can see, the states (pdf means) move from their positions in the generic HMM to be compatible with the observations in the training set.

[2] "Strictly speaking, an ergodic model has the property that every state can be reached from every other state in a finite number of steps." Rabiner (1990).



(a) (b)

Fig. 8. In this figure it is shown how an HMM is trained for the task of "determining how well the people in the picture know each other (people)" on a given picture. To begin the training we first need to define the main structure of the HMM in the form of a generic HMM. The generic HMM is composed of six 2D Gaussian pdfs centered on the centroids of the *K*-means clustering conducted on the training set. The standard deviation used in the covariance matrix of each Gaussian is set to 126 pixels and a uniform transition matrix is used for governing the transitions between the states (each Gaussian represents a state). Part (a) shows the model that is used as the generic HMM for training the task-dependent HMM. Each Gaussian observation pdf is shown by a heat-map, centered on the centroids of its corresponding cluster. The generic HMM is used in the Baum–Welch algorithm to train the task-specific HMM. Part (b) shows the final, task-specific, model after training the generic HMM by the eye trajectories of the training set.



STD(°)	0.5°	1.0°	1.5°	2.0°	2.5°	3.0°	4.5°	5.0°	6.0°	6.5°	7.0°	7.5°
SEM	5.38	5.84	6.42	7.15	4.89	5.33	5.41	5.45	6.56	6.24	5.78	7.11
MEAN	15.26	19.54	26.14	37.91	34.42	55.48	59.64	58.34	45.17	46.74	23.85	26.51

(a) (b)

Fig. 9. (a) Accuracy of task classification versus standard deviation (STD) of the Gaussian observations. The accuracy is obtained by averaging the diagonal elements of the confusion matrix of all 64 images and the error bars show the standard error of the mean (SEM). The table at the bottom of the figure shows the values of the means and SEMs. (b) Confusion matrix of task inference using the HMM-based model.

4.2. Results

In Section 4.1 we remarked that the best value for the standard deviation is $\sigma = 4.5^{\circ}$. This value is the standard deviation used in the covariance matrix which defines the area covered by each of the 2D Gaussian observation pdfs. In other words, this value shows the best scale for the diameters of the Gaussians shown in Fig. 8. Increasing this value will expand the overlapped area between different observation pdfs, which in turn relaxes the overtness constraint of the attentional spot. However, too large values of the standard deviation causes too much overlap in the observation pdfs, which flattens the likelihood distribution function of Eq. (5).

As mentioned in the introduction, the confusion matrix given in the study by Greene, Liu, and Wolfe (2012) indicated that task inference was at the chance level (25%). Fig. 9b shows the confusion matrix obtained using our HMM model. The numerical values of the confusion matrix are shown in Table 1. The diagonal elements of the confusion matrix show the percentage of trajectories whose task labels were correctly classified (*hits*) and the off-diagonal elements comprise the *misses* in the classification.

Fig. 9a shows the accuracy of task classification versus standard deviation (STD) of the Gaussian observations. The accuracy is obtained by averaging the diagonal elements of the confusion matrix and the error bars show the standard error of the mean (SEM). The table at the bottom of the figure shows the values of the means and the SEMs. In the experiment we use a leave-one-out cross-validation to define the training set and use the average accuracy across all images to represent the overall accuracy. The SEMs are the sample estimate of the population standard deviation of the accuracies across all images divided by the square root of the number of images.

The diagonal values of the confusion matrix in Fig. 9b are well above the chance level. The model is able to infer the visual-task with average accuracy of 59.64%, as given by averaging the diagonal elements of the confusion matrix.

In order to show the advantage of HMMs over DTMCs, we used the same database and did the task inference using DTMCs. To do so, we used the same set up as in the HMMs (using *K*-Means clustering), but rather than setting an observation pdf to each state, we used Euclidean nearest-neighbors to select the current state of a fixation. This is equivalent to assuming that covert attention is the same as the overt attention (i.e. that attention is allocated to the same location as the eye fixation). The confusion matrix obtained when using DTMC has an average accuracy of 31.54% and is shown in Table 2. Comparing the results of HMM and DTMC highlights the importance of allowing for off-target fixations in our model for inferring the task in real images.

Table 1
Numerical values of the confusion matrix for task classification using the HMM-based model. To obtain the results we set $\sigma = 4.5^o$ and did LOO cross validation over all task dependent eye trajectories.

	MEMORY	DECADE	PEOPLE	WEALTH
MEMORY	**59.35**	13.76	12.98	13.91
DECADE	11.86	**55.91**	18.84	13.39
PEOPLE	12.56	11.57	**65.84**	10.03
WEALTH	15.44	11.64	15.46	**57.46**

Bold values indicate maximum number in each row.

Table 2
Numerical values of the confusion matrix for task classification using the DTMC-based model. To obtain the results we used the same setup (number of clusters) as in the HMMs and did LOO cross validation over all task dependent eye trajectories. In order to define the presumably overt state, we set the state to the closest state using the Euclidean nearest neighbor.

	MEMORY	DECADE	PEOPLE	WEALTH
MEMORY	**23.54**	28.64	32.57	15.25
DECADE	13.43	**27.68**	21.64	37.25
PEOPLE	10.63	28.47	**45.29**	15.61
WEALTH	24.74	16.47	29.14	**29.65**

Bold values indicate maximum number in each row.

5. Conclusion

In this article we presented a probabilistic framework for task inference in natural images. This work was motivated in part by previously reported difficulties in developing a reliable approach for implementing an inverse Yarbus process. In particular, we examined the study of Greene, Liu, and Wolfe (2011, 2012), who concluded that visual-task cannot be inferred using eye movements, and tried to understand why their approach was not successful. We hypothesized that the difficulty lay in the lack of explanatory power of the summary statistics that were used, such as the number of fixations, and duration of fixations, that were used to classify the trajectories. These features, however, have been shown (e.g., Castelhano & Henderson (2008)) to be unreliable in task inference.

Another reason for the failure of the aggregate-based method in inferring the task is that no contextual information about the image is used in classification. This is in spite of the fact that image context has been shown to have a major effect on eye movement behavior (Torralba et al., 2006; Goferman, Zelnik-Manor, and Tal, 2012).

To handle these problems we used features that are more informative than summary statistics, and provided a way to incorporate local (contextual) information. To validate our approach in relation to the results of the Greene, Liu, and Wolfe (2011, 2012), experiments, we used the same database of natural images and the same experimental protocol.

One could argue that the negative results in the Greene, Liu, and Wolfe (2011, 2012), experiments imply that it is not possible to perform the inverse Yarbus process. However, there is evidence that such a process is in fact possible, provided by work on predicting the cognitive state of an observer from eye movements. Eye movement measurements have been used in the recognition of physical activity Bulling et al. (2009), Bulling et al. (2011), detection of tiredness or distraction (Di Stasi et al., 2012), estimating mental workload levels (Di Stasi et al., 2010), diagnosis of schizophrenia (Benson et al., 2012) and detection of mental fatigue (Schleicher et al., 2008). As the cognitive state of a viewer carrying out a visual task is presumably affected by the nature of the task it is reasonable to expect that viewer task can likewise be detected from eye movement measurements. The results of applying our technique support this conclusion.

Our approach is based on the idea that visual task is revealed by the spatio-temporal patterns of the allocation of visual attention. In practice, attention has most often been tracked using eye movements, and models of attention are frequently evaluated based on how well they can predict eye trajectories. However, classical salience-based models of eye movement generation exhibit limited performance in accounting for eye movements in real-world situations of viewing complex natural scenes. This is due, in part, to their pure bottom-up dependence on low-level image features. In such situations single human observers outperform even the best salience-based models in predicting eye trajectories (Judd, Durand, and Torralba (2012)). Low-level features often have low salience in areas near fixations (Ballard & Hayhoe, 2009; Hayhoe et al., 2003; Land & McLeod, 2000; Land, Mennie, & Rusted, 1999). Therefore we developed our model based on real, task-dependent eye trajectories recorded while viewing natural images. To go beyond the simple salience-based approaches we used Hidden Markov models (HMMs) as a tool for time-series analysis of the eye trajectories. This allows us to encode the dynamics of natural eye movements into task-dependent models.

The HMM-based method not only allows us to track overt foci of attention (i.e. fixation locations), but also allows for the tracking of covert attention and other sources of discrepancy between the center of gaze (COG) and the focus of attention (FOA). A deviation between the COG and FOA can arise by a variety of mechanisms. For example, in the phenomenon known as *the center-of-gravity* (also known as the *global effect*) (Zelinsky et al., 1997; He & Kowler, 1989; Najemnik & Geisler, 2005), the target of the eye movement is actually the center-of-mass of a set of visually-salient objects, one of which would correspond to the FOA. The resulting COG at the center-of-mass location would generally not correspond to a location of high salience. As Coëffé and O'regan (1987) point out, the global effect is less pronounced when saccadic latencies are long, as is typically the case when visual search is being carried out in a slow, deliberate manner. But when a task is being done quickly, then significant deviations between the COG and FOA can rise. The advantage of decoupling the COG and the FOA becomes clear by comparing the results of our Discrete-Time-Markov-Chain (DTMC) and HMM models, since the only difference between these two models is the linkage between the FOA and the COG. The HMMs allow for decoupling the COG and the FOA by means of the state-specific Gaussian distribution functions, whereas in the DTMC they are assumed to be the same. The Gaussian distribution functions used in the model definition of the HMMs span an area around the covert attentional loci. The actual eye fixation points are then considered as a random outcome of the Gaussian process, which can be result in locations well away from the covert attention locus. The experimental results show that by separating the COG and FOA leads to better performance in inferring viewer task.

The improvement in performance in inferring visual task with the HMM approach as compared to the DTMC approach also provides indirect experimental evidence for separation between the COG and the FOA in real-world picture viewing tasks. This indication of the separation of covert and overt attention is an important by-product of our approach, since it is not easy to demonstrate such separations in scene viewing eye movement recordings. The possibility of this dissociation between the COG and FOA has been raised before in oculomotor studies by indirectly tracing attentional spot on non-fixated targets. In a study by O'Regan et al. (2000) COG-FOA decoupling is implied from observers' lack of awareness of changes in an image 40% of the time, even though they were directly fixating the change location. Declines in reaction times to attentional probes away from fixation (Hoffman & Subramaniam, 1995; Kowler et al., 1995; Deubel & Schneider, 1996; Schneider & Deubel, 2002), have also been used to indicate

allocation of covert attention away from fixation. To avoid the problems that arise from using measurements of the COG to track the FOA, other more direct techniques for tracking covert attention could be incorporated in our model. These include techniques such as the dot-probe task (MacLeod, Mathews, & Tata, 1986), detecting microsaccades (Hafed & Clark, 2002), or fMRI recording (Wojciulik, Kanwisher, & Driver, 1998). However, these methods can interfere with the on-going task and provide limited spatial and temporal resolution. Thus they are not appropriate for practical applications of the inverse Yarbus process.

Time-series analysis incorporates temporal information about fixations as well as spatial into an attention model. In previous approaches to analyzing eye movement behavior, usually only spatial information is considered and temporal information of fixations is simply omitted. For example, the Greene, Liu, and Wolfe (2011, 2012), studies ignored the temporal aspect of the viewer eye movement trajectories. However, it is becoming increasingly clear that temporal analysis of eye movement is as important as its spatial counterpart in describing the underlying mechanisms. The temporal order of fixations is an important feature in describing the underlying mechanism of the visual behavior. The question of whether and how the temporal order of fixations matters in modeling eye movements has been considered often, beginning with the pioneering studies of Buswell (1935) and Yarbus (1967). In salience-based models of attention the temporal order of fixations is not usually considered in training the models (Borji & Itti, 2013, Figure 7). From a statistical point of view, these models postulate a *naïve Bayes assumption* in evaluating the likelihood probability of Eq. (5), which assumes independence between consecutive fixation locations. In contrast to these models, consecutive fixations have been shown to be highly dependent on each other. Hacisalihzade, Stark, and Allen (1992) recorded the eye movements of observers during the task of recognizing an object and showed that the fixations loosely follow a Markov process. They showed that the eyes visit the features of an object cyclically, following regular scanpaths rather than moving randomly. Elhelw et al. (2008) used a first-order, discrete-time, discrete-state-space Markov chain to model eye movement dynamics. Stark and Ellis (1981) also came up with a Markov process as a general model of fixation placement during the task of reading. Pieters, Rosbergen, and Wedel (1999) observed a similar pattern in the scanpaths of the observers while looking at printed advertisements. There is more information in the time-series of eye positions than just the ordering of fixation locations. Evidence from studies done with viewers carrying out natural real-world tasks emphasizes the need to consider fixation duration as well as fixation location in understanding the mechanism of the visual system (Droll et al., 2005; Hayhoe, Bensinger, & Ballard, 1998; Land, Mennie, & Rusted, 1999).

The HMM approach that we propose in this paper conveniently incorporates the temporal aspects of attention through its Markov modeling. The temporal order of the fixations plays an important role in decoding the pattern of eye movements in the HMMs. The transition matrix of the HMMs (A) adjusts its elements according to the order of the fixations viewers make on targets during the training. This information is later used by the HMM to match the pattern of state transitions against that of a test trajectory. The better the transition pattern of the test trajectory accords with that of a task-dependent HMM, the more likely the trajectory is to be an observation of that task.

It is possible to extract at least temporal order information from the eye trajectories in the original Yarbus experiment (Yarbus, 1967, Figs. 107–124), as well as in the experiment by Greene, Liu, and Wolfe (2012), figure 3. So, the machine learning method employed in the Greene et al. study could have used temporal features as well as the summary spatial statistics. It is possible that the human classifiers did (unconsciously) assume some sort of temporal order by tracing along the lines of the displayed eye tracks. The Greene et al. study therefore leaves open the question of whether temporal information can improve the task inference. To judge the influence of the temporal information on the ability to infer task, we created a constrained HMM method which lacked any temporal information. We removed the temporal information of the fixations from the trained HMMs by setting the transition matrix to equal values. In this way no knowledge of the temporal order of fixations that may be in the training set is incorporated into the HMM. Throwing away the temporal information in this manner resulted in a 15.51% average degradation on the diagonal elements of the confusion matrix. This is a significant drop in performance but it should be noted that the performance is still above chance, showing that the decoupling of the FOA and COG results in some improvement over the summary statistics. The HMMs fail completely in inferring the task when spatial information from the eye trajectories is also removed. Thus, we conclude that both spatial and temporal information is crucial in solving the inverse Yarbus problem, and the lack of such information may be one reason that the Greene et al. approach did not work.

The HMM task inference method we proposed requires that the location of attention targets be known beforehand. In the past, salience was used to define targets for attention shifts, but in real-world viewing of complex scenes, with abstract tasks, defining salience is difficult. The specification of task-dependent salience measures is an open research problem, and our paper only scratches the surface of what is necessary. Well-performing task-dependent salient point localization schemes will involve high-level symbolic reasoning about the scene directed by task knowledge. Our approach is very simple, but was sufficient for the restricted problem posed by the Greene, Liu, and Wolfe (2011, 2012) work – that of training a viewer-task classifier from a set of images and eye movement trajectories recorded while viewers examine these images under various task instructions. These training examples can be used to find a statistical model for the salient locations. We used the K-means clustering technique on the training set and used the centroids of these clusters as the salient points or potential targets. Due to the lack of knowledge about the task relevance of these potential targets, we cannot reject the possibility of next attending a given target given the currently fixated one. Thus, to model the temporal aspect of the eye movements we used an ergodic structure for a generic HMM that allows transitions from a state to any other one. The generic HMM undergoes a training phase to build attention models for each task-image combination, whereby we can calculate the likelihood term of Eq. (5) and make an inference about the task. While this approach to predefining attention targets worked well in this specific application, a more unconstrained and unsupervised problem would require a much more sophisticated approach to learning what the targets are. For example, if we applied our trained HMMs to inferring visual task for viewers looking at images that were not trained on, the method would fail miserably, since the targets will be in locations different than those in the training set. Some method for generalizing the location of the targets to different images would be needed.

An interesting phenomenon seen in the training results is the variation of performance with different standard deviations of the observation pdf (Fig. 9a). This figure shows a falloff in the task classification accuracy as the standard deviation moves away from a value of roughly 4° of visual angle. The optimal value is consistent with previous estimates of the size of the *operational fovea* as the central 3° of vision Johansson et al. (2001). Carpenter (1991) shows that targets within 4° of central vision are still perceived at 50% of maximal acuity. Based on the current evidence we cannot tell whether this merely a coincidence, but further experimentation could investigate this more deeply. Certainly,

the degree of spatial decoupling of the FOA and COG is worth quantifying, whether this information is used to tune a statistical attention model such as ours, or to gauge the level of acuity needed for carrying out specific visual tasks.

Task inference has many applications. Knowing what the user is seeking on a web page combined with a dynamic interactive design can lead to a smart web page that highlights the relevant information in a page according to the ongoing visual-task. The same idea applies to an intelligent signage that changes its contents to show relevant advertisements according to the task inferred from each viewer's eye movements. We believe that in each of these applications an HMM-based model can be used as a reliable model to infer the visual-task. Indeed, by increasing the amount of training data and using prior task knowledge in the Bayesian formulation we can improve the accuracy of the results. Other examples of interesting applications can be found in the literature. Vidal et al. (2012) implemented a pervasive healthcare system by using eye movements to infer the mental status of patients. Bulling, Roggen, and Troster (2011) used eye movements to obtain information about a person's context, and suggested a context-aware pervasive computing system based on the eye movements. As mentioned earlier, a by-product of our HMM model is that it can locate the focus of attention, whether it is overt or covert. This feature allows us to track the more informative attentional spot, rather than the simple motion of the gaze. Thus, in applications based on eye movements, performance gains might be obtained by using the attentional locus, which is task-oriented and robust, rather than the gaze information provided by standard eye trackers.

6. Ethical considerations

All experiments conducted in this research were approved by the McGill Ethics Review Board. The Research Ethics Boards of McGill University adhere to and are required to follow the Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada- Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans, December 2010. This policy espouses the core principles of Respect for Persons, Concern for Welfare and Justice, in keeping with leading international ethical norms, such as the Declaration of Helsinki.

6.1. Role of the funding source

The funding agencies had no involvement in the study design, collection, analysis and interpretation of data, writing of the report, nor in the decision to submit the article for publication.

Acknowledgment

The authors would like to thank Jeremy Wolfe and Ben Tatler, the reviewers of the manuscript, for their valuable comments, which helped us significantly improve the paper. We would also like to thank *Fonds Québecois de la Recherche sur la Nature et les Technologies* (FQRNT) and *Natural Sciences and Engineering Research Council of Canada* (NSERC) for their support of this work.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.visres.2014.08.014.

References

Ballard, D., & Hayhoe, M. (2009). Modelling the role of task in the control of gaze. *Visual Cognition, 17*(6-7), 1185–1204.
Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience, 7*(1), 66–80.
Becker, W. (1972). The control of eye movements in the saccadic system. *Cerebral Control Of Eye Movements And Motion Perception, 82*, 233–243.
Bengio, Y., & Frasconi, P. (1995). An input output HMM architecture. In *Advances in neural information processing systems* (pp. 427–434). Morgan Kaufmann Publisher.
Benson, P., Beedie, S., Shephard, E., Giegling, I., Rujescu, D., & St Clair, D. (2012). Simple viewing tests can detect eye movement abnormalities that distinguish schizophrenia cases from controls with exceptional accuracy. *Biological Psychiatry, 72*(9), 716–724.
Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207.
Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision, 14*(3), 29.
Bruce, N., & Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*(3), 5.
Bulling, A., Roggen, D., & Troster, G. (2011). What's in the eyes for context-awareness? *IEEE Pervasive Computing, 10*(2), 48–57.
Bulling, A., Ward, J., Gellersen, H., & Tröster, G. (2009). Eye movement analysis for activity recognition. In *Proceedings of the 11th international conference on Ubiquitous computing* (pp. 41–50). ACM.
Bulling, A., Ward, J., Gellersen, H., & Tröster, G. (2011). Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(4), 741–753.
Buswell, G. (1920). In *An experimental study of the eye-voice span in reading* (Vol. 17). University of Chicago.
Buswell, G. (1935). *How people look at pictures: A study of the psychology of perception in art*. Chicago: University of Chicago Press.
Carpenter, R. (1991). The visual origins of ocular motility. *Vision And Visual Function, 8*, 1–10.
Castelhano, M., & Henderson, J. (2008). Stable individual differences across images in human saccadic eye movements. *Canadian Journal of Experimental Psychology/ Revue canadienne de psychologie expérimentale, 62*(1), 1–14.
Castelhano, M., Mack, M., & Henderson, J. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision, 9*(3), 6.
Clark, J. (1999). Spatial attention and latencies of saccadic eye movements. *Vision Research, 39*(3), 585–602.
Clark, J., & O'Regan, J. (1998). Word ambiguity and the optimal viewing position in reading. *Vision Research, 39*(4), 843–857.
Coëffé, C., & O'regan, J. (1987). Reducing the influence of non-target stimuli on saccade accuracy: Predictability and latency effects. *Vision Research, 27*(2), 227–240.
Connor, C., Egeth, H., & Yantis, S. (2004). Visual attention: Bottom-up versus top-down. *Current Biology, 14*(19), R850–R852.
Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review Of Neuroscience, 18*(1), 193–222.
Deubel, H., & Schneider, W. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*(12), 1827–1837.
Di Stasi, L., Renner, R., Catena, A., Cañas, J., Velichkovsky, B., & Pannasch, S. (2012). Towards a driver fatigue test based on the saccadic main sequence: A partial validation by subjective report data. *Transportation Research Part C: Emerging Technologies, 21*(1), 122–133.
Di Stasi, L., Renner, R., Staehr, P., Helmert, J., Velichkovsky, B., Canas, J., et al. (2010). Saccadic peak velocity sensitivity to variations in mental workload. *Aviation, Space, and Environmental Medicine, 81*(4), 413–417.
Droll, J., Hayhoe, M., Triesch, J., & Sullivan, B. (2005). Task demands control acquisition and storage of visual information. *Journal of Experimental Psychology: Human Perception and Performance, 31*(6), 1416.
Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition, 17*(6-7), 945–978.
Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision, 8*(2), 2.
Elhelw, M., Nicolaou, M., Chung, A., Yang, G., & Atkins, M. (2008). A gaze-based study for investigating the perception of visual realism in simulated scenes. *ACM Transactions on Applied Perception (TAP), 5*(1), 3.
Epelboim, J., Steinman, R., Kowler, E., Edwards, M., Pizlo, Z., Erkelens, C., et al. (1995). The function of visual search and memory in sequential looking tasks. *Vision Research, 35*(23), 3401–3422.
Eriksen, C., & James, J. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics, 40*(4), 225–240.
Findlay, J. (1981). Local and global influences on saccadic eye movements. *Eye Movements: Cognition And Visual Perception*, 171–179.
Flanagan, J., & Johansson, R. (2003). Action plans used in action observation. *Nature, 424*(6950), 769–771.
Fuchs, A. (1971). The saccadic system. In C. Collins & J. Hyde (Eds.), *The control of eye movements* (pp. 343–362). New York: Academic Press.

Furneaux, S., & Land, M. (1999). The effects of skill on the eye–hand span during musical sight-reading. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 266*(1436), 2435–2440.
Goferman, S., Zelnik-Manor, L., & Tal, A. (2012). Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(10), 1915–1926.
Google, (2013). LIFE photo archive hosted by Google. <http://images.google.com/hosted/life>.
Greene, M., Liu, T., & Wolfe, J. (2011). Reconsidering Yarbus: Pattern classification cannot predict observer's task from scan paths. *Journal of Vision, 11*(11), 498.
Greene, M., Liu, T., & Wolfe, J. (2012). Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research, 62*, 1–8.
Gunn, S. R. (1998). Support vector machines for classification and regression. ISIS Technical Report 14.
Hacisalihzade, S., Stark, L., & Allen, J. (1992). Visual perception and sequences of eye movement fixations: A stochastic modeling approach. *IEEE Transactions on Systems Man and Cybernetics, 2*(3), 474–481.
Hafed, Z., & Clark, J. (2002). Microsaccades as an overt measure of covert attention shifts. *Vision Research, 42*(22), 2533–2545.
Haji-Abolhassani, A., & Clark, J. (2013). A computational model for task inference in visual search. *Journal of Vision, 13*(3), 29.
Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425–2430.
Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188–194.
Hayhoe, M., Bensinger, D., & Ballard, D. (1998). Task constraints in visual working memory. *Vision Research, 38*(1), 125–137.
Hayhoe, M., Shrivastava, A., Mruczek, R., & Pelz, J. (2003). Visual memory and motor planning in a natural task. *Journal of Vision, 3*(1), 6.
Hearst, M., Dumais, S., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications, 13*(4), 18–28.
He, P., & Kowler, E. (1989). The role of location probability in the programming of saccades: Implications for center-of-gravity tendencies. *Vision Research, 29*(9), 1165–1181.
Henderson, J. (1992). Visual attention and eye movement control during reading and picture viewing. In *Eye movements and visual cognition* (pp. 260–283). Springer.
Hoffman, J., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics, 57*(6), 787–795.
Hu, J., Brown, M., & Turin, W. (1996). HMM based on-line handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(10), 1039–1045.
Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194–204.
Itti, L., & Koch, C. (2001). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging, 10*, 161–169.
Johansson, R., Westling, G., Bäckström, A., & Flanagan, J. (2001). Eye–hand coordination in object manipulation. *Journal of Neuroscience, 21*(17), 6917–6932.
Judd, T., Durand, F., Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. Mit Tech Report URL <http://hdl.handle.net/1721.1/68590>.
Kanan, C., Ray, N. A., Bseiso, D. N., Hsiao, J. H., & Cottrell, G. W. (2014). Predicting an observer's task using multi-fixation pattern analysis. In *Proceedings of the symposium on eye tracking research and applications* (pp. 287–290). Springer.
Kanan, C., Tong, M., Zhang, L., & Cottrell, G. (2009). SUN: Top-down saliency using natural statistics. *Visual Cognition, 17*(6-7), 979–1003.
Kaufman, L., & Rousseeuw, P. (2009). *Finding groups in data: An introduction to cluster analysis* (Vol. 344). Wiley Interscience.
Klein, R. (1980). Does oculomotor readiness mediate cognitive control of visual attention. *Attention and Performance VIII, 8*, 259–276.
Klein, R. (2000). Inhibition of return. *Trends in Cognitive Sciences, 4*(4), 138–147.
Klein, R., & MacInnes, W. (1999). Inhibition of return is a foraging facilitator in visual search. *Psychological Science, 10*(4), 346–352.
Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology, 4*(4), 219–227.
Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research, 35*(13), 1897–1916.
Land, M., & Furneaux, S. (1997). The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 352*(1358), 1231–1239.
Land, M., & Lee, D. (1994). Where do we look when we steer. *Nature, 369*, 742–744.
Land, M., & McLeod, P. (2000). From eye movements to actions: How batsmen hit the ball. *Nature Neuroscience, 3*(12), 1340–1345.
Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *PERCEPTION, 28*(11), 1311–1328.
Land, M., & Tatler, B. (2001). Steering with the head: The visual strategy of a racing driver. *Current Biology, 11*(15), 1215–1220.
MacLeod, C., Mathews, A., & Tata, P. (1986). Attentional bias in emotional disorders. *Journal of Abnormal Psychology, 95*(1), 15.
Mannan, S., Ruddock, K., & Wooding, D. (1997). Fixation sequences made during visual examination of briefly presented 2D images. *Spatial Vision, 11*(2), 157–178.
Mika, S., Ratsch, G., Weston, J., Scholkopf, B., & Mullers, K. (1999). Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop* (pp. 41–48). IEEE.
Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision, 11*(8) (article 17).
Nair, V., & Clark, J. (2002). Automated visual surveillance using hidden markov models. In *International conference on vision interface* (Vol. 93, pp. 88–93).
Najemnik, J., & Geisler, W. (2005). Optimal eye movement strategies in visual search. *Nature, 434*(7031), 387–391.
Niebur, E., & Koch, C. (1998). Computational architectures for attention. In R. Parasuraman (Ed.), *The attentive brain* (pp. 163–186). Cambridge, MA: MIT Press.
Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science, 171*(968), 308–311.
Nuthmann, A., & Henderson, J. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*(8).
O'Regan, J., Deubel, H., Clark, J., & Rensink, R. (2000). Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition, 7*(1-3), 191–211.
Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*(1), 107–123.
Patla, A., & Vickers, J. (1997). Where and when do we look as we approach and step over an obstacle in the travel path? *Neuroreport, 8*(17), 3661–3665.
Patla, A., & Vickers, J. (2003). How far ahead do we look when required to step on specific locations in the travel path during locomotion? *Experimental Brain Research, 148*(1), 133–138.
Pelz, J., & Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research, 41*(25), 3587–3596.
Pieters, R., Rosbergen, E., & Wedel, M. (1999). Visual attention to repeated print advertising: A test of scanpath theory. *Journal of Marketing Research, 36*(4), 424–438.
Posner, M., & Cohen, Y. (1984). Components of visual orienting. *Attention and Performance X: Control of Language Processes, 32*, 531–556.
Rabiner, L. (1990). A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in Speech Recognition, 53*(3), 267–296.
Reinagel, P., & Zador, A. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems, 10*(4), 341–350.
Rothkopf, C., Ballard, D., & Hayhoe, M. (2007). Task and context determine where you look. *Journal of Vision, 7*(14), 16.
Rutishauser, U., & Koch, C. (2007). Probabilistic modeling of eye movement data during conjunction search via feature-based attention. *Journal of Vision, 7*(6), 5.
Salvucci, D., & Anderson, J. (2001). Automated eye-movement protocol analysis. *Human-Computer Interaction, 16*(1), 39–86.
Salvucci, D., & Goldberg, J. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications* (pp. 71–78). ACM.
Schleicher, R., Galley, N., Briest, S., & Galley, L. (2008). Blinks and saccades as indicators of fatigue in sleepiness warnings: Looking tired? *Ergonomics, 51*(7), 982–1010.
Schneider, W. (1995). VAM: A neuro-cognitive model for visual attention control of segmentation, object recognition, and space-based motor action. *Visual Cognition, 2*(2-3), 331–376.
Schneider, W., & Deubel, H. (2002). Selection-for-perception and selection-for-spatial-motor-action are coupled by visual attention: A review of recent findings and new evidence from stimulus-driven saccade control. *Attention and Performance Xix: Common Mechanisms in Perception and Action* (19), 609–627.
Simola, J., Salojärvi, J., & Kojo, I. (2008). Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research, 9*(4), 237–251.
Smith, T., & Henderson, J. (2009). Facilitation of return during scene viewing. *Visual Cognition, 17*(6-7), 1083–1108.
Stark, L. W., & Ellis, S. (1981). Scanpaths revisited: Cognitive models direct active looking. In D. Fisher, R. Monty, & J. Senders (Eds.), *Eye movements and psychological processes* (pp. 192–226). Hillsdale, NJ: Lawrence Erlbaum Associates.
Tatler, B., Baddeley, R., & Vincent, B. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research, 46*(12), 1857–1862.
Tatler, B., Hayhoe, M., Land, M., & Ballard, D. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11*(5).
Tatler, B., & Vincent, B. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research, 2*(2), 1–18.
Tatler, B., Wade, N., Kwan, H., Findlay, J., Velichkovsky, B., et al. (2010). Yarbus, eye movements, and vision. *I-Perception, 1*(1), 7.
Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*(4), 766–786.
Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136. ISSN 0010-028.
Van Der Lans, R., Pieters, R., & Wedel, M. (2008). Eye-movement analysis of search effectiveness. *Journal of the American Statistical Association, 103*(482), 452–461.
Vidal, M., Turner, J., Bulling, A., & Gellersen, H. (2012). Wearable eye tracking for mental health monitoring. *Computer Communications, 35*(11), 1306–1311.
Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*(9), 1395–1407.

Wischnewski, M., Belardinelli, A., Schneider, W., & Steil, J. (2010). Where to look next? Combining static and dynamic proto-objects in a TVA-based model of visual attention. *Cognitive Computation, 2*(4), 326–343.

Wischnewski, M., Steil, J., Kehrer, L., & Schneider, W. (2009). Integrating inhomogeneous processing and proto-object formation in a computational model of visual attention. In H. Ritter (Ed.). *Human centered robot systems* (Vol. 6, pp. 93–102). Springer.

Wojciulik, E., Kanwisher, N., & Driver, J. (1998). Covert visual attention modulates face-specific activity in the human fusiform gyrus: fMRI study. *Journal of Neurophysiology, 79*(3), 1574–1578.

Wolfe, J., Cave, K., & Franzel, S. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance, 15*(3), 419–433.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645–678.

Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum Press (Translated from the Russian edition by Haigh, B).

Zelinsky, G., Rao, R., Hayhoe, M., & Ballard, D. (1997). Eye movements reveal the spatiotemporal dynamics of visual search. *Psychological Science, 8*(6), 448–453.