Inferring Visual Task from Eye Movements

Amin Haji-Abolhassani

Doctor of Philosophy

Department of Electrical & Computer Engineering

McGill University

Montreal, Quebec

2014-1-20

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

DEDICATION

This thesis is dedicated to my family for their love, endless support and encoragement.

To the memory of my grandmother, Ana.

ACKNOWLEDGEMENTS

Though only my name appears on the cover of this dissertation, many people have contributed to its production. I owe my gratitude to all those people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

I owe my deepest gratitude to my advisor, Professor Clark. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own and at the same time the guidance to recover when my steps faltered. He taught me how to question thoughts and express ideas. His valuable scientific advice, knowledge, support, and mentorship have highly helped me in completing my research project. I am grateful to his insightful discussions and suggestions regarding this project. Without his knowledge and assistance this study would not have been successful. Professor Clark is and will remain my best scientific role model, and I will follow his guidance throughout my life.

I am very grateful to the members of my committee, Professor Ferrie, Professor Precup and Professor Pineau for their time, encouragement, and expertise throughout my research.

I am indebted to my many colleagues and who supported me during my quest at the Center of Intelligent Machines. I thank Prasun Lala, John Harisson, Cynthia Davidson, Marlene Gray, Jan Binder, Patrick McLean, Nick Wilson and the members of Visual Motor Research Lab, who have all provided much needed advice, support and friendship during the course of my graduate studies.

I also thank Leon and Suzanne Fattal, *Fonds Qubcois 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.

Last but not least, I would like to express my love and gratitude to my beloved family for their understanding, love and support, through the duration of my studies.

ABSTRACT

From the whole amount of visual information impinging on the eye, only a fraction ascends to the higher levels of visual awareness and consciousness in the brain. *Attention* is the process of selecting a subset of the available sensory information for further processing in short-term memory, and has equipped the primates with a remarkable ability to interpret complex scenes in real-time, despite the limited computational capacity. In other words, attention implements an information-processing bottleneck that instead of attempting to fully process the massive sensory input in parallel, realizes a serial strategy to achieve near real-time performance. This serial strategy builds up an internal representation of a scene by successively directing a spatially circumscribed region of the visual field corresponding to the highest resolution region of the retina, the so-called *fovea*, to conspicuous locations and creating eye trajectories by sequentially fixating on attention demanding targets in the scene.

Directing the fovea to visual targets in a scene is done through rapid eye movements called *saccades* that typically occur between two and five times per second. Pattern information is only acquired during periods of relative gaze stability in between the saccades, called *fixations*, owing to the brain's suppression of information during the saccades. Gaze planning, thus, is the process of directing the fovea through a scene in real-time in the service of ongoing perceptual, cognitive and behavioral activity. The question of exactly what is happening during fixations is still something of a puzzle, but the effect of visual-task on the pattern and specifications of eye movements has been long studied in the literature.

Although the effect of visual-tasks on eye movement pattern has been investigated for various tasks, there is not much done in the area of visual-task inference from the eye movements. In this work, we develop a probabilistic method to infer the visual-task of a viewer by analyzing the eye movements. To do so, a task-dependent attention model is developed to infer the attention location from the gaze position. Given the attentional spot and the stimuli we can locate the targets visited during an eye trajectory and infer the ongoing task based on the attended targets.

Two different scenarios are studied in the thesis. First a method is developed to infer the tasks in synthetic stimuli, where the location of all objects in the image is given to the model. In the second group of models, the tasks are executed on a set of natural images, where no prior information about the location of targets is provided to the model.

A probabilistic approach for task inference is presented, that is based on the theory of Hidden Markov Models (HHM). A HMM is a statistical model based on Markov processes. The proposed model is used in the context of Bayesian learning to realize a fully statistical inference that can incorporate different sources of information about the ongoing task. An alternative approach for the Bayesian inference is also proposed that incorporates the a-priori sources of information about the tasks into the inference model and builds a full-scale visual-task recognition framework.

In order to evaluate the performance of the model, the results of task inference are presented in form of confusion matrices and are compared to the inference results of other models. The results support the idea of attention modeling using the HMMs and suggest a solid probabilistic framework for task inference using the HMMs.

ABRÉGÉ

Sur la quantité d'informations visuelles empiéter sur l'œil, seule une fraction monte à des niveaux supérieurs de la conscience visuelle et de la conscience dans le cerveau. L'*attention* est le processus de sélection d'un sous-ensemble de l'information sensorielle disponible pour un traitement ultérieur dans la mémoire à court terme, et a équipé les primates avec une remarquable capacité à interpréter des scènes complexes en temps réel, en dépit de la capacité de calcul limitée. En d'autres termes, à l'attention implémente un goulot d'étranglement traitement de l'information, au lieu de tenter de traiter la totalité des entrées sensorielles massif en parallèle, réalise une stratégie de série pour atteindre une performance quasi-temps réel. Cette stratégie de série construit une représentation interne d'une scène en dirigeant successivement une région spatialement circonscrits du champ visuel correspondant à la région de la plus haute résolution de la rétine, la soi-disant *fovéa*, à des endroits bien visibles et en créant des trajectoires oculaires par fixateur séquentiellement sur l'attention objectifs exigeants de la scène.

Dans les activités visuelles, les yeux font des mouvements rapides, appelées *saccades*, généralement entre deux et cinq fois par seconde, afin d'apporter de l'information environnementale dans la fovéa. Informations relatives aux signatures est acquise uniquement pendant les périodes de stabilité relative, du regard appelés *fixations*, en raison de la suppression par le cerveau de l'information pendant les saccades. Admirez la planification, donc, est le processus de réalisation de la fovéa à travers une scène en temps réel au service de l'activité perceptive, cognitive et comportementale

en cours. La question d'exactement ce qui se passe au cours de fixations est toujours quelque chose d'un puzzle, mais l'effet de la tâche visuelle sur le modèle et les spécifications des mouvements oculaires a été longuement étudié dans la littérature.

Bien que l'effet des tâches visuelles comme la lecture, le comptage et la recherche, sur le modèle des mouvements oculaires a été étudiée pour différentes tâches, il n'ya pas beaucoup fait dans le domaine de l'optique-tâche inférence à partir des mouvements oculaires. Dans ce travail, nous développons une méthode probabiliste de déduire le visual-tâche d'un spectateur par l'analyze des mouvements oculaires. Deux scénarios différents sont étudiés dans la thèse. D'abord une méthode est développée pour déduire les tâches de stimuli synthétiques, où l'emplacement de tous les objets dans l'image sont donnés au modèle. Dans le second groupe de modèles, les tâches sont exécutées sur un ensemble d'images naturelles, où aucune information préalable sur l'emplacement des cibles sont fournis pour le modèle.

Une approche probabiliste pour tâche inférence est présenté, qui est basé sur la théorie des modèles de Markov cachés (HHM). Un HMM est un modèle statistique basé sur les processus de Markov. Le modèle proposé est utilisé dans le contexte de l'apprentissage bayésien pour réaliser une inférence statistique pleinement pouvant intégrer différentes sources d'information sur la tâche en cours.

Afin d'évaluer la performance du modèle, les résultats de la tâche inférence sont présentés sous forme de matrices de confusion et sont comparés aux résultats de l'inférence des modèles d'attention classiques. Les résultats soutiennent l'idée de la modélisation de l'attention en utilisant les HMM et proposent un cadre probabiliste solide pour tâche inférence utilisant les HMM.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

CMPD	Carnegie Mellon Pronouncing Dictionary
COG	Center of Gaze
DSCHMM	Double-State Character HMM
DSWHMM	Double-State Word HMM
DTMC	Discrete-Time Markov Chain
EM	Expectation Maximization
FOA	Focus of Attention
FSM	Finite State Machine
GMM	Gaussian Mixture Model
HDP-HMM	Hierarchical Dirichlet Process HMM
HMM	Hidden Markov Models
IOR	Inhibition of Return
LM	Language Model
LOOCV	Leave-One-Out Cross-Validation
MAP	Maximum A-Posteriori
ML	Maximum Likelihood
SEM	Standard Error of the Mean
SSHMM	Single-State Hidden Markov Model
STD	Standard Deviation
TD	Top-Down
TSHMM	Tri-State Hidden Markov Model

CHAPTER 1
Introduction

1.1 Problem Statement

Human vision is an active, dynamic process in which the viewer seeks out specific visual inputs according to the ongoing cognitive and behavioral activity. A critical aspect of active vision is directing a spatially circumscribed region of the visual field (about 3^o) corresponding to the highest resolution region of the retina, the so-called fovea, to the task-relevant stimuli in the environment. In this way our brain gets a clear view of the conspicuous locations in an image and will be able to build up an internal, task-specific, representation of the scene.

While low-level visual features are shown to influence eye movements [35, 130], visual-task can also influence the pattern of eye movements. This effect was shown in the celebrated study of Yarbus [128] who recorded the eye movements of people while viewing a painting. As shown in figure 1.1, different trajectories emerged depending on the task that the viewers were given. By his experiment he showed that visual-task has a great influence on specific parameters of eye movement control and that eye fixations are not randomly distributed in a scene, but instead tend to cluster on some regions at the expense of others. In this figure we can see how visual-task modulates the conspicuity of different regions and as a result change the pattern of eye movements.

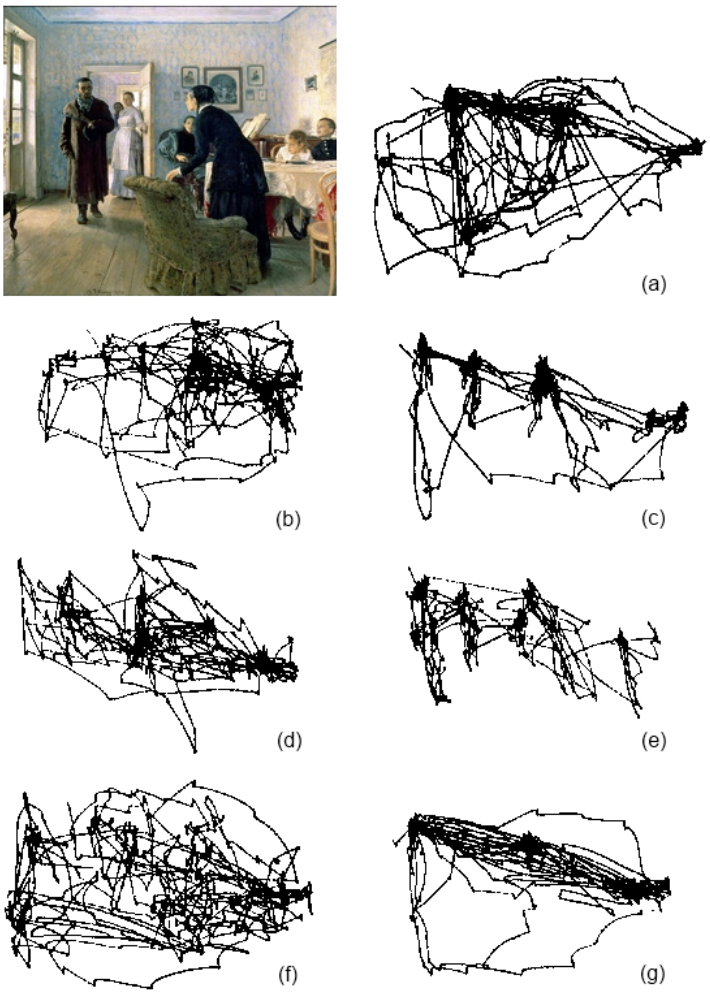


Figure 1–1: Eye trajectories of viewers carrying out different tasks measured by Yarbus. a) No specific task was given to the viewer and the subject carried out free viewing. b) "Estimate the wealth of the family". c) "Give the ages of the people in the painting". d) "Summarize what the family had been 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 [128] with permission from Springer Publishing Company. The photo of the painting is courtesy of www.ilyarepin.org.

Yarbus effect has been re-investigated in the literature related to the eye movement analysis, the common question of which being whether or not visual-task influences the pattern of eye movements. The studies of eye movements during natural behavior unanimously indicate a bond between the gaze location and informative locations to the immediate task goals [32, 52, 75, 74, 92, 94]. In the visual attention model of Schneider [106], the target selection is partially governed by action. This selection for action is particularly highlighted by the fact that the gaze targets are concentrated in the task-relevant areas in an image in the presence of a visual-task [52, 74], whereas before beginning the task, eye fixations are scattered over the image [52, 101].

To better demonstrate the gaze deployment under the influence of task, Rothkopf et al. [101] devised a series of experiments conducted in a purely virtual environment, where subjects executed two tasks of "approaching" and "avoiding" objects while navigating along a walkway. In the experiments they showed that the fixation distribution on an object changes according to the task and suggested that human gaze is directed toward the regions in a scene that are determined primarily by the task requirements. Several other studies have also reproduced the original finding of Yarbus by using new equipment and stimuli in their experiments. For instance, in [114] the results obtained by Yarbus is confirmed in an experiment that studied the effect of *instructions* in viewing a portrait of Yarbus. All of these experiments reproduce the results of Yarbus using new instruments and broader number of subjects (20 subjects in [114]) and emphasize the fact that visual task in fact does affect the pattern of eye movement.

Besides affecting the distribution of the gaze location, visual-task influences other metrics of eye movements as well. Tatler et al. [111] show that visual task affect the temporal statistics of eye movements in viewing natural images. Castelhano et al. [16] studied the effect of task on eye movement in tasks of *memorization* and *search* and showed that visual-task influence a number of eye movement measures, such as number of fixations and gaze duration on specific objects, while leaving other parameters, such as average saccade amplitude and individual fixation durations, unchanged. They also showed that viewing task biases the selection of scene regions and temporal measures on those regions. In [64] a temporal coupling between vision and action is demonstrated. In their experiment they measured the gaze onset towards the next target relative to the hand movements as the subject maneuvered an object past an obstacle. The recorded departure time of the eye was shown to be linked with the execution of the task as the gaze moved onto the next target as soon as the object cleared the obstacle. Similar temporal coupling between action and vision is also demonstrated for the tasks of driving [78, 76], tea making [52], sandwich making [74], music sight reading [39], walking [93] and reading aloud [11]. In [77] the eye movements of cricket players are studied and it is shown that different skill levels of the players in performing the task entail different latency in directing the gaze towards predicted locations of the incoming ball. This temporal coupling between action and vision highlights the importance of considering task influence on temporal characteristics of eye movement, beside its spatial characteristics, in task-dependent models of visual motor systems.

The effect of task on the pattern of eye movements or attentional deployment has been studied for different tasks, such as copying arrangements of blocks [2], making tea [74], making sandwich [52] and driving [76]. In [77] it is shown that while watching a cricket game the gaze is directed through the video according to the ongoing events in the game. In another experiment eye movements of subjects were recorded while watching a person who stacked a set of blocks. In this block-sorting task, gaze was shown to be anticipating the expected points of interaction and directed to it [37]. In another block-copying experiment conducted by Ballard et al. [2], 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 [17] studied the spatial characteristics of eye movements for the task of reading and showed that when reading a text, the *centre 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 [117] developed the *feature integration theory* that studies the parameters that influence the attentional deployment in the task of visual search. In [126] this model is improved by Wolfe et al., who developed a model called "the Guided Search" that also study how our brain directs attention through a scene during a search task. Hayhoe and Ballard [50] did a review on the goal-directed behavior of the visual-motor system, where a comprehensive set of references to the relevant studies in task influence on eye movements is presented.

Although the effect of visual-task on eye movement pattern has been investigated for various tasks, there has not been much done in the inverse process, that is to infer the visual-task from the eye movements. If we consider the classical Yarbus

process, which we refer to as *the forward Yarbus process*, as a function, the task is given as the input and task-dependent eye trajectories are the output of the process. In this thesis, on the other hand, the goal is to develop a method to realize an *inverse Yarbus process*, whereby the ongoing task can be inferred given the eye movements of a viewer. In other words, in an inverse Yarbus process an eye trajectory is given as input and the output is the visual-task that has led to such trajectory.

In a study by Greene et al. [46, 45] an unsuccessful attempt was made to solve the inverse Yarbus problem. They used supervised learning techniques to train three different classifiers based on linear discriminant analysis [84], correlational methods [49] and support vector machines [54]. The classifiers were trained on the so-called *summary statistics* of eye movements used in scanpath analysis [15, 85]. Seven features of eye trajectories including the number of fixations, mean fixation duration, mean saccade amplitude and percent of image covered by fixations were used as the observation vector of each trajectory. The results showed that the classifiers can only infer the task at the chance level and fail to consistently reveal the underlying task of test trajectories. Thus, based on the results they 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."*.

1.2 Approach

While the results of inferring the task from summary statistics is shown to be disappointing, concluding that the scanpaths cannot be used to identify the visual-task is overstated. In Greene's study only specific features that are based on aggregate characteristics of eye movements are used to train the classifiers and failing to infer

the visual-task does not mean that task inference in general is infeasible. In fact in another work, Castelhano et al. [16] studied the influence of task on a group of summary statistics (including the ones used in Greene's experiment) for 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 is obtained in [85], where they also showed that the effect of visual search task on the same features as in [46] is insignificant.

Predicting the states of observers from eye movements have been studied in several works. Bulling et al. [9, 10] 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. Detection of tiredness or distraction of the driver from eye movements is another example of using eye movements for predicting the mental state of the observer [22]. In another work by Di Stasi et al. [23], the maximum velocity of the eyes during saccadic movements was shown to have an inverse relationship with the *mental workload* of subjects in a simulated driving task. In a psychophysiological study by Benson et al. [5] eye movement analysis is used to detect schizophrenic patients. In [105] the blink duration, delay of lid reopening, blink interval and standardized lid closure speed were identied as indicators of mental fatigue.

In all of the above studies there exists a common conclusion, which indicates the possibility of predicting the observer's cognitive state by analyzing the eye movement behavior. Although in these works recognizing the human activity, mental workload,

mental health and fatigue were addressed as examples of the cognitive state, we believe the same rational could be applied to describe other mental states of the observer. Particularly, we consider visual-tasks as another indication of the cognitive state and devise a method to infer the task by analyzing the eye movements of observers.

Although certain statistical features of eye movements seem to remain unchanged across the tasks, the COG tends to fixate on targets that are relevant to the task at hand. This effect can be seen in the eye trajectories of Yarbus, in which the viewers seem to be fixating on the targets that are more informative according to the task. For instance, for the task of age estimation, faces are more likely to get fixated, while for the task of wealth estimation the objects become of more interest to the viewer. Thus, in this thesis the spatial information of the eye trajectories are taken as an indication of the visual-task and are used to infer the ongoing task.

The process that is responsible for directing the COG to different parts of a visual scene is called *the visual attention*. Thus, a task-specific attention model is proposed to be used for revealing the attentional spots given the eye trajectories. Then the *foci of attention* (FOA) are used to extract information about the task-relevant objects in a scene, which in turn is used to recognize the executed visual-task.

The proposed attention model is based on the generative model of *Hidden Markov Models* (HMM). For each task, we train a task-specific HMM to model the cognitive process in the human brain that generates eye movements given the task. The output of each HMM would be task-dependent eye trajectories along with their respective likelihoods. In order to infer the ongoing task, we use this likelihood

term in a Bayesian inference framework that incorporates the likelihood with a-priori knowledge about the task and gives the posterior probability of various tasks given the eye trajectory.

1.3 Overview of the Thesis and Its Contributions

In this thesis, the overarching goal is to explore the notion of visual-task inference from the eye movement trajectories. To this end, an effort is made to realize a model for visual attention to be used in a Bayesian inference framework that gives rise to the task that best matches the attended objects in a scene. Different models are proposed for synthetic and natural images, followed by a chapter on how to fuse different sources of information in the inference.

In **Chapter 2** an overview of the classical models of visual attention is given. In the theories of visual attention there are two major viewpoints that either emphasize *bottom-up*, image-based, and task-independent effects of the visual stimuli on the attention allocation, or *top-down*, volition-controlled, and task-dependent modulation of attention [19]. Both of these classical attention models are studied along with their theoretical details and will be used in the evaluation of the proposed attention model in later chapters.

At the end of Chapter 2 a review is given on the theory of the HMMs as the foundation of the proposed attention model that is elaborated in the following chapters. In particular, some light is shed on three fundamental problems in the model description of the HMMs that concern the likelihood evaluation, decoding and training of the HMMs given a set of observation sequence. The solution to these problems

will come in handy in training and using the proposed HMM-based attention models in evaluating the likelihood term of the Bayesian inference.

Although image salience models have been extensively researched and are quite well-developed, empirical evaluation of such models shows that they are disappointingly poor at accounting for actual attention allocations. In **Chapter 3** we highlight the shortage of the classical models in modeling the phenomenon called *the covert attention*. In this chapter it is shown that one of the main deficits of the classical models is that they assume that tracking the FOA is equivalent to tracking the COG. However, the COG does not necessarily follow the FOA and in fact they can be well away from each other. The attentional spot that is diverted away from the COG is called the covert attention and requires a different technique, rather than simply tracking the COG, to be located.

In the chapter it is explained how the "hiddenness" of the HMMs and the "covertness" of the attentional spot can be mapped to each other, on which the theoretical specification of an HMM-based attention model is established. In the proposed model, covert attention is represented by the hidden states of a task-dependent HMM. Fixation locations, thus, will correspond to the observations of an HMM and can be used in training task-dependent models. By this interpretation of variables we can use a sequence of eye positions to represent the hidden sequence of covert attention locations, which is useful in spotting the task-specific targets based on the eye movement trajectories.

In **Chapter 4** the proposed HMM-based attention model is used in inferring the visual-task. To this end, we show how the task-dependent HMMs can be used

to evaluate the likelihood term in the Bayesian inference formulation. Different attention models are proposed for task inference in synthetic and natural images. In synthetic images the location of targets are pre-defined in the image, while in the natural images we usually do not have the knowledge about the target. The HMMs are trained on a database of task-dependent eye movements created for synthetic and natural images. Having trained the task-dependent attention models for different tasks, they are used to evaluate the likelihood of a given test eye trajectory, which is in turn used in the Bayesian inference framework to recognize the visual-task.

In **Chapter 5** a-priori information is used in order to better refine the results of the posterior probability. It is shown that using Bayesian formulation allows us to better fuse different sources of information to improve the task recognition results. As an example, in this chapter an application is developed to recognize the word that a user is typing by directing the FOA to different sequence of characters shown on a soft keyboard. In this case a dictionary of lexicons is used as the a-priori source and the test eye movement trajectories are fed into the HMMs trained for the typing application to evaluate the likelihood of the trajectories given the parameters of the HMMs. The results are compared to those of other techniques that use the classical models of attention, which show that HMMs can be successfully applied to the problem of task inference and produce better results compared to the classical models.

Finally, **Chapter 6** concludes the thesis with a review of its main findings and introduces some remaining related questions to be explored.

CHAPTER 2
Background

2.1 Visual-Task Inference Using Bayesian Inference

Bayesian Inference is a class of parametric learning technique that classifies data in a probabilistic manner [81]. By applying a probabilistic inference, a mathematical framework can be developed for merging all sources of information and presenting the result in the form of a probability density function. In the case of developing an inverse projection from eye movement space to visual-task space, this structure will be useful in the sense that it can incorporate prior knowledge about the tasks. Moreover, the inference gives us a probability distribution over different possible tasks rather than providing us with a single task as the output. In this way a higher level process can be designed to make decisions about the task and provide us with the degree of confidence in the decision.

Suppose we have data of the form $< \mathbf{O}, \theta >$, where $\theta \in \Theta$ is the class label, which is selected from a set of all class labels Θ, and $\mathbf{O}$ is the observation vector containing the observations $\vec{O}_t$ at time $1 \leq t \leq T$ i.e.:

$$\mathbf{O} = < \vec{O}_1, \vec{O}_2, ..., \vec{O}_T > . \tag{2.1}$$

In general, Bayesian inference models two entities:

- $P(\theta)$: The *prior probability* of each class $\theta \in \Theta$.

- $P(\mathbf{O}|\theta)$: The class conditional distribution, which is also referred to as the *likelihood* function.

The probability of class θ given an observation sequence $\mathbf{O}$ can be written by applying the Bayes's rule:

$$P(\theta|\mathbf{O}) = \frac{P(\mathbf{O}|\theta)P(\theta)}{P(\mathbf{O})} = \frac{P(\mathbf{O}|\theta)P(\theta)}{\sum_{\theta' \in \Theta} P(\mathbf{O}|\theta')P(\theta')}. \tag{2.2}$$

Thus, in order to make an inference, the likelihood term should be modulated by the prior knowledge about the class.

In visual activities, every 100 to 800 $msec$, the eyes make rapid movements, called *saccades* in order to foveate different objects in a scene. Pattern information, though, is only acquired during periods of relative gaze stability, called *fixations*, owing to the brain's suppression of information during the saccades [83]. Therefore, in the context of visual-task inference from the eye trajectories, the problem can be defined in terms of the Bayesian inference formulation by defining the observation vector ($\mathbf{O}$) as the consecutive fixations the eyes make on different locations in an image in order to perform a task. Also the visual-task can be mapped to the class labels (θ) that are selected from a pool of class labels (Θ). Therefore, the posterior probability $P(\theta|\mathbf{O})$ is expressed as the probability of a visual-task, given the fixation locations in an image.

By this new interpretation of variables, each $\vec{O}_t$ in equation (2.1) is a vector containing the coordinates of the fixations that are sampled from a stochastic process $\{\vec{O}_t\}$ at discrete times $t = \{1, 2, ..., T\}$ over random image locations. Each $\vec{O}_t$, then, is a vector defined by (x_t, y_t), where x_t and y_t are the x and y coordinates of the

t^{th} fixation, respectively. $P(\theta)$ is the prior probability of each task $\theta \in \Theta$, which assigns a probability distribution to the tasks based on our prior knowledge about them. This is where we can apply other sources of information about the tasks and improve the inference.

$P(\mathbf{O}|\theta)$ is the task conditional distribution, which is also referred to as the likelihood function. The likelihood term defines the probability of observing the sequence $\mathbf{O}$ while executing the task θ and can be considered as an objective evaluation of the forward Yarbus process in the sense that it evaluates the probability of seeing an observation given a task. The likelihood term can be broken down to the conditional probabilities:

$$\begin{aligned} P(\mathbf{O}|\theta) =& P(\vec{O}_1, \vec{O}_2, \ldots, \vec{O}_T|\theta) \\ =& P(\vec{O}_1|\theta)P(\vec{O}_2|\vec{O}_1, \theta)\ldots P(\vec{O}_T|\vec{O}_1\ldots\vec{O}_{T-1}, \theta). \end{aligned} \tag{2.3}$$

The standard approach for quantifying the likelihood is to use a *saliency map* as an indication of how attractive a given part of the field-of-view is to the attention [62]. In classical attention models, the likelihood is quantified as proportional to the amplitude of the saliency map on different targets in an image. In the following section we will review the classical models of attention and explain how they are used to evaluate the likelihood of an eye trajectory.

2.2 Related works on Attention modeling

In the theories of visual attention, there are two major viewpoints that either emphasize bottom-up, image-based, and task-independent effects of the visual stimuli on the saliency map or top-down, volition-controlled, and task-dependent modulation

of such maps. In the following sections we show how these models posit different assumptions to evaluate the likelihood term.

2.2.1 Bottom-Up Saliency Maps

In the bottom-up 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 (i.e., $P(\mathbf{O}|\theta)$ is assumed to be equal to $P(\mathbf{O})$). Moreover, in this model, it's assumed that observations $\vec{O}_t$ are conditionally independent, which reduces the likelihood term of equation (2.3) to:

$$\begin{aligned} P(\mathbf{O}|\theta) &= P(\mathbf{O}) = P(\vec{O}_1, \vec{O}_2, \ldots, \vec{O}_T) \\ &= P(\vec{O}_1)P(\vec{O}_2)\ldots P(\vec{O}_T). \end{aligned} \tag{2.4}$$

This assumption is called the *naïve Bayes assumption* and only needs the probability of directing the FOA to the fixation locations appearing in each trajectory to obtain the likelihood term.

In this model, attention tracking is typically based on a model of image salience. One can take the location with the highest salience as the estimate of the current FOA. Current salience models are based on relatively low-level features, such as color, orientation and intensity contrasts.

One of the most advanced saliency models is the one proposed by Itti and Koch [60]. 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*, that are the outputs from different filters tuned to simple visual attributes, such as color, intensity and orientation (see figure 2–1).

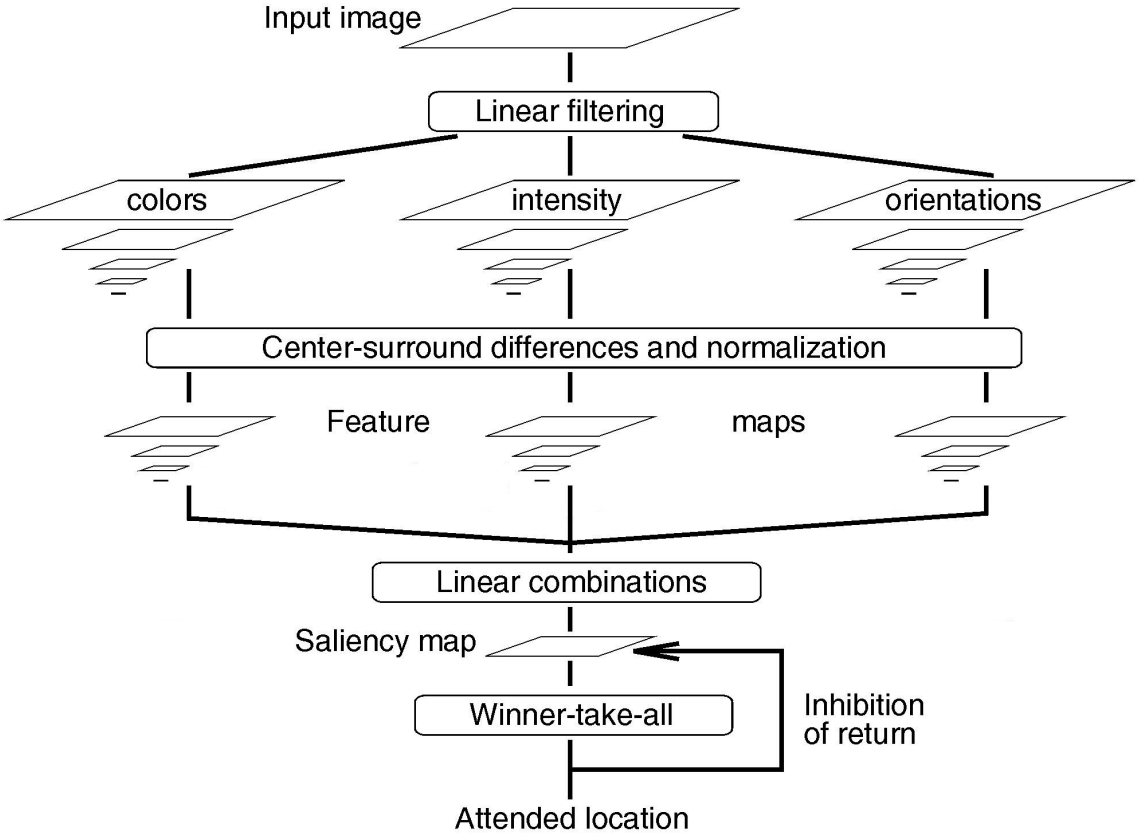
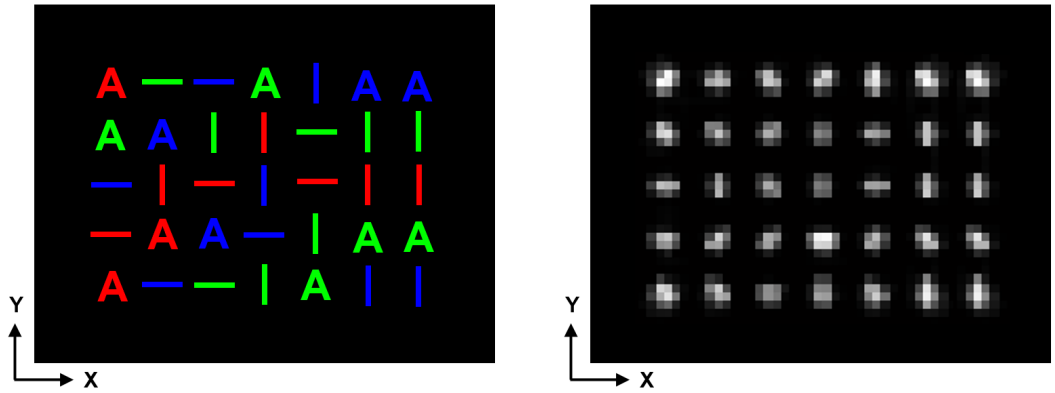


Figure 2–1: General architecture of the bottom-up attention model by Itti and Koch [60]. The saliency map is built by linearly combining the feature maps, that are the outputs from different filters tuned to simple visual attributes, such as color, intensity and orientation (the figure is a reproduction of figure 1 in [60]).

Figure 2–2 shows an example of a bottom-up saliency map obtained using the *Saliency Toolbox* [120]. Figure 2–2a shows a synthetic image that comprises a combination of character "A" and horizontal and vertical bars in three different colors. The objects are aligned on a 5×6 grid on a plain black background. Figure 2–2b shows the bottom-up saliency map according to the attention model of [60] shown in the block diagram of figure 2–1. The feature maps are obtained by color, intensity and orientation filters and combined into the saliency map by a linear combination.



(a) (b)

Figure 2–2: a) Original Image. b) Saliency map of the bottom-up attention model presented in [60].

2.2.2 Top-Down Saliency Maps

Although image-based salience models have been extensively researched and are quite well-developed, empirical evaluation of such models have shown that they are disappointingly poor at accounting for actual attention allocations when a visual-task is involved [27]. In our view the bulk of this shortfall is due 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 ongoing task. In modeling the task-dependant visual attention, two contrasting views exist that emphasize the visual salience or the cognitive relevance hypotheses. In the cognitive relevance models, an object-based representation of the scene is used to select the 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 [89]. In saliency-based models, attention maps are image-based and are derived from the spatial attributes of an image. In some hybrid models, these two models 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 behaviour of the visual system in the presence of visual task [122, 123].

Although visual salience and cognitive relevance models hypothesize different spatial representations for directing the attentional spot, empirical evidence supports their connection, due to a correlation between objects and salience [28, 56]. In other words, cognitive relevance models are based on object-based representations and objects generally differ from the scene background in their image properties, highlighting their locations in the saliency maps used in the saliency-based models.

The second major group of saliency-based, visual attention models is the top-down, task-dependent one that modulates the bottom-up saliency maps according to the viewer's visual-task. Figure 2–3 shows an illustration of the interaction between the top-down and the bottom-up models proposed in [61, 102]. In this model different tasks enforce different weight vectors ($\mathbf{W}$) in the linear combination phase.

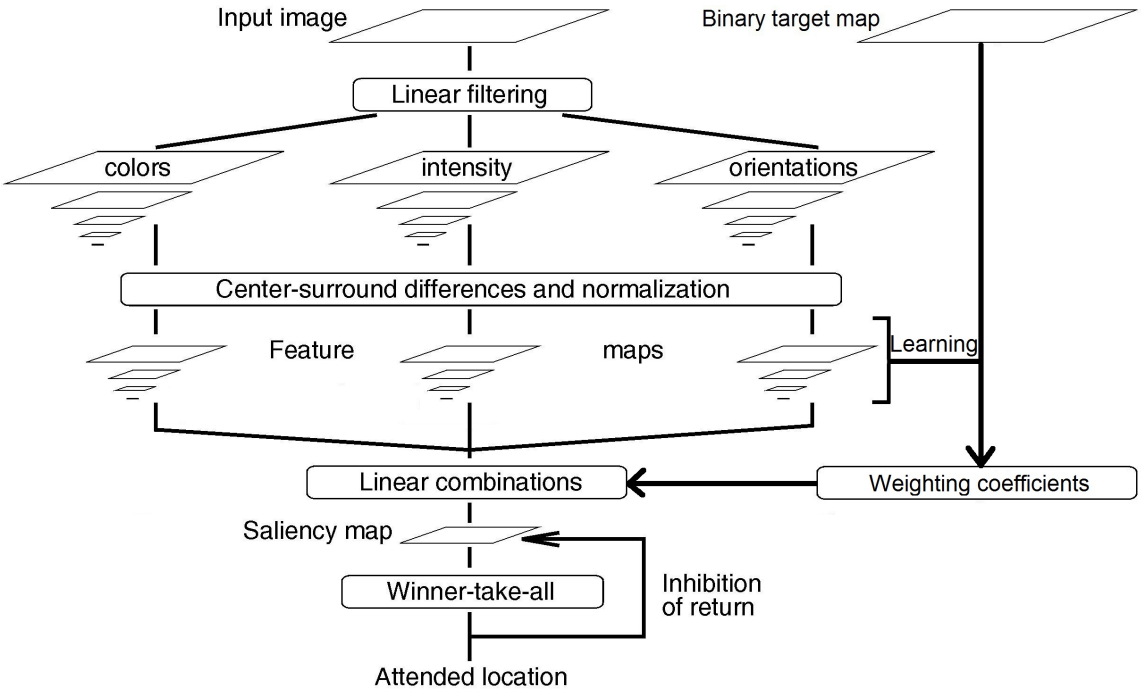


Figure 2–3: Influence of top-down, task-dependent priors on bottom-up attention models. The influence can be modeled as a weight vector modulating the linear combination of the feature maps (the figure is a reproduction of figure 1 in [61]).

Therefore, each task can be denoted by a specific weight vector that emphasizes a certain combination of feature maps in order to make the relevant targets in the saliency map become more conspicuous. For instance, in the task of searching for red objects, the weight vector would put more emphasis on the color feature maps rather than intensity or orientation maps.

In order to obtain the task-dependent weights, Itti and Koch [61] used a supervised learning scheme that trains on a manually labeled *binary target map* that indicates which locations are conspicuous according to the task. Thus, in the task of searching for red objects, the binary target map highlights the red areas (or objects)

to mark them as relevant locations to the task. To train a task-dependent model, the weights are iteratively updated to minimize an error function, E, according to the following optimization problem:

$$\mathbf{W} = \arg\min_{\mathbf{W}} \; E, \quad \text{where } E = \frac{1}{1 + \frac{\min(\mathbf{M}_{in}(\mathbf{W}))}{M_{out}(\mathbf{W})}}. \tag{2.5}$$

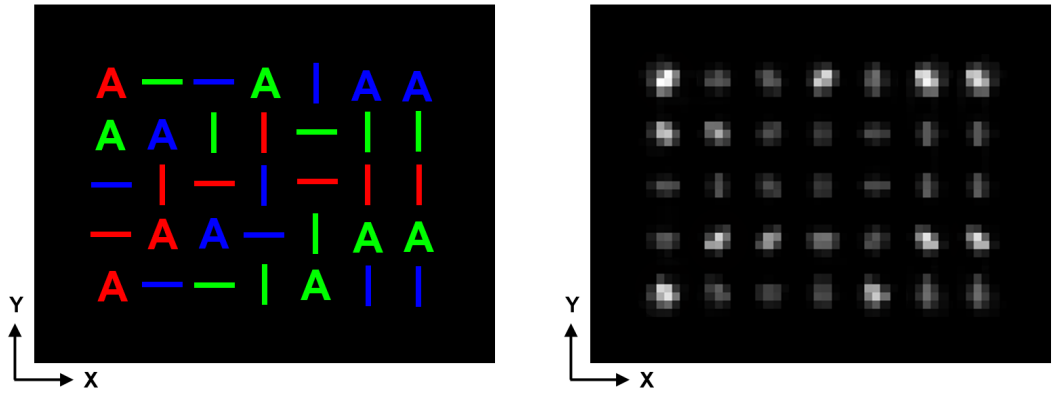
In this equation $\mathbf{M}_{in}$ is the saliency map's maxima inside the manually outlined target regions (one maximum per region) and M_{out} is the maximum saliency outside those regions.[1] In order to avoid divergence of the weights, a constraint is applied to the weights that sets their sum to a fixed number [61].

The optimization process of equation (2.5) will result in a $\mathbf{W}$ that increases the ratio of $\frac{\min(\mathbf{M}_{in}(\mathbf{W}))}{M_{out}(\mathbf{W})}$, which in turn gives more weight to the feature maps that highlight the target areas in the resulting saliency map. In other words, as soon as the error function, E, goes below %50, all targets become more conspicuous and the weight vectors can be used to generate the top-down, task-dependent saliency map.

Figure 2–4 shows an example of the top-down saliency map obtained using the *Saliency Toolbox* [120] for the task of "counting the number of characters". While the bottom-up models combine the maps with constant weighs, the top-down models (shown in the block diagram of figure 2–3) modulate the weights according to the task. Figure 2–4a shows the same synthetic image used in figure 2–2, and figure 2–4b

[1] Note that $\mathbf{M}_{in}(\mathbf{W})$ is a vector containing the saliencies of the maxima at all target regions, whereas $M_{out}(\mathbf{W})$ is the saliency at the non-target maximum. The values of both $\mathbf{M}_{in}$ and M_{out} are dependent on the weight vector $\mathbf{W}$, since the weight vector modulates the saliency of the objects in an image.

shows its saliency map tuned to the task of "counting the number of characters". As we can see, the locations of the characters are more conspicuous (lighter) in the top-down saliency map.



(a) (b)

Figure 2–4: a) Original Image. b) Saliency map of the same image using a top-down attention model [61].

Other combinations of the sources of guidance maps into a unified, task-dependent attentional maps are also studied in the literature. Ehinger et al. [26] achieved a 94% accordance with human eye movements in a visual search task by combining the saliency maps with the scene context and target features. Also Torralba et al. [116] use contextual information for facilitating object search in natural scenes. The *contextual guidance* model of attention use the bottom-up saliency map, scene context, and top-down mechanisms at an early stage of visual processing and combine them into a unified attention map. Kanan et al. [66] use the knowledge about how and where objects tend to appear in a scene in order to derive an appearance-based saliency model. In a recent study by Borji and Itti [7] 65 state-of-the-art salience models of attention are studied and categorized into either bottom-up or top-down

classes. In each category the models are qualitatively compared over 13 experimental criteria. One of these criteria, based on which the saliency models are evaluated, is the accountability for real-world eye movement datasets in terms of spatial correlation coefficient. In order to evaluate the statistical relationship of the saliency models with the eye movement datasets, the eye trajectories can be modeled into a so-called *ground-truth saliency map* by combining recorded eye fixations from all subjects into a map similar to that of the saliency-based attention models. This feature, along with other features that are studied in [7], are typical criteria that are used in the studies on attention modeling to objectively evaluate the models.

Overall, the task-dependent attention models improve the bottom-up models by incorporating the task-dependency and can be used to generate the likelihood term of equation (2.3) by the following equation:

$$\begin{aligned} P(\mathbf{O}|\theta) &= P(\vec{O}_1, \vec{O}_2, \ldots, \vec{O}_T|\theta) \\ &= P(\vec{O}_1|\theta)P(\vec{O}_2|\theta)\ldots P(\vec{O}_T|\theta), \end{aligned} \tag{2.6}$$

where $P(\vec{O}_t|\theta)$ is the normalized saliency of the location in the image visited at time t given a task θ. Similar to the bottom-up models, each $\vec{O}_t$, itself, is consisted of the x and y coordinates of the COG at time t.

As we can see, in top-down models it is still assumed that the naïve Bayes assumption holds and that the observations are conditionally independent. Thus, artificial eye trajectories can be created by successively selecting a target with a chance that is proportional to its conspicuity in the saliency map, regardless of its

proceeding fixation location. In the next section we will show that this assumption is in conflict with some of the known phenomena in the theories of eye movement.

2.3 Hidden Markov Models

Although salience-based top-down models have somewhat addressed the problem of task independence of the bottom-up models, they are based on some assumptions that degrade their performance. Tatler et al. [112] show that gaze allocation models that are based on the salience models are limited at accounting for many aspects of free viewing and can fail dramatically in the context of natural task performance. In their study they show that there are a set of reasons that limit the salience models, which suggest moving away from the picture-viewing paradigms to the models that can generalize to a broader range of experimental contexts. They argue that some of these limitations lie in the basic assumptions at the heart of such studies that are problematic if we wish to try to generalize these models to how gaze is allocated in natural behavior.

One of these arguments about the applicability of salience maps in modeling natural visual behavior is their intrinsic interpretation of cognitive relevance with spatial deviations of low-level features from local surround. While the contrast of low-level feature in fixated locations are shown to be statistically higher than control locations in an image, this correlation between the features and fixation is relatively weak [82, 91, 99]. This lack of explanatory power of image salience in the context of active tasks becomes particularly evident in the studies of the tasks of hitting a ball [1, 77], tea making [74] and sandwich making [52], where saccades are directed to the expected point of contact with no particular contrast in low-level visual features. Due

to this lack of explanatory power of image salience models, another group of models of task-dependant visual attention is emerging that emphasizes the cognitive relevance hypotheses in predicting fixation locations. In the cognitive relevance models, an object-based representation of the scene is used to select the 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 [89]. 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 [122, 123].

Another deficiency of salience models highlighted in [112] is that the decision about where to fixate is made by a winner-takes-all process that selects the most conspicuous location on a salience map. In this selection criterion, though, what has not been accounted for is the retinal position of image information, which leads to neglecting the fact that the retinal acuity decays in peripheral vision. 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 [69, 70, 71, 97], recent empirical evidence in viewing photographic images rejects such an effect in the eye trajectories [109, 113]. As another limitation of salience models, Tatler et al. [112] highlight the importance of temporal information about the eye movements besides their spatial characteristics, which is usually neglected in these models. In other word, the primary

goal of salience models is to spatially model the fixations and they usually disregard the temporal aspects of viewing behavior. This is while evidence from natural tasks emphasizes the need to consider fixation duration as well as fixation location in understanding the mechanism of the visual system [25, 51, 74].

One of the other limitations of salience models highlighted in [112] that is particularly of interest to this work is postulating that the saccades are precisely directed to the target locations for processing. While this seems to be a plausible assumption from the perspective of eye movement behavior in simple viewing tasks, in the context of natural tasks this assumption might not be true. For instance, Johansson et al. [64] 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 [64] corrective saccades of amplitude less that 2.5 degrees were infrequent suggesting that in natural behavior the fixations land only close to the attention demanding targets and they are not always precisely following the focus of attention.

In our proposed model we use HMMs to relax the inherent assumptions in the salience models and use real-world eye movements to train task-dependent models that can infer the visual-task on natural images. Particularly we will show how HMMs relax the assumption of precise target fixations that is assumed in salience models by modeling the fixations by a Gaussian distribution function that allows for fixations well away from the target. Moreover, by analyzing the eye trajectories as time-series we give temporal features of eye movements the same weight as their spatial features. The notion of cognitive relevance to the low-level features is also

loosened in the HMM models as the Gaussian distributions are allowed to move away from salient objects to more cognitive relevant targets in an image.

The theory of HMMs have been used in different fields, such as speech recognition [98], anomaly detection in video surveillance [86] and hand writing recognition [58]. HMMs have also been used in analysis of eye movements. In [104] HMMs are used to automatically label the recorded eye movements as fixations and saccades. In another study, Salvucci and Anderson [103] developed a HMM-based model for analysis of eye movements during the task of equation solving. Simola et al. [108] modeled three cognitive states of visual process during a reading task by the hidden states of HMMs. Van Der Lans et al. [118] 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.

The successful application of the HMMs in time series analysis, such as the speech signal, makes it a good candidate for our goal of analyzing the FOA sequences. In order to develop the HMM-based attention model an introduction is given here on the theory of the HMMs. The solutions for training the HMMs is used in the later chapters to train a task-dependent attention model, which is later used in a Bayesian inference to reveal the ongoing visual-task.

2.3.1 Model Definition

In the previous section we showed that classical models of attention are limited in terms of accounting for real-world eye movements of observers while viewing natural images. This could be concluded from the benchmark presented in [65] that compares 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 is 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, in this section we present a model that is based on real, 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. Therefore, one of the benefits of our HMM model is their 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 [98]. Similar to the first-order, finite-state, discrete-time Markov chain (DTMC), HMMs govern the transitions 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*, where

$$a_{ij} = P(q_{t+1} = s_j | q_t = s_i), \qquad 1 \leq i, j \leq N. \tag{2.7}$$

- $q_t \in S$ and $1 \leq t \leq T$ 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.
- $\Pi = \{\pi_i\}$ is *the initial state distribution*.

- π_i is the probability of starting a sequence at state i, i.e.,

$$\pi_i = P(q_1 = s_i), \qquad 1 \leq i \leq N. \tag{2.8}$$

In a more general view, both HMMs and DTMCs are classes of finite state machines (FSMs) [4] that at each time step generate 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 $\mathbf{O}$, where:

- $\mathbf{O}$ 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 the DTMCs, each observation vector $\vec{O}_t$ can only be generated by a specific state, meaning that there is no overlap between the observation vectors of different states. Figure 2–5a shows a DTMC with two states (i.e., $N = 2$). At the time step $t = 0$, the process starts with entering one of the states s_1 or s_2 with the probability of π_1 or π_2, respectively. In the following time steps, the process chooses the next state sequentially according to the transition probabilities a_{ij} and at each time step an observation is generated, which is specific to the state being visited (i.e., there is no overlap between the observation generating functions' outcomes).

Figure 2–5b shows a sample state sequence of the process, $\{q_t : 1 \leq t \leq 3\}$, where $q_t \in \{s_1, s_2\}$ is the state that the sequence is visiting at time t. The observation sequence is in the form $\{\vec{O}_t : 1 \leq t \leq 3\}$, which ,due to the non-overlapping

characteristic of the observation space between the states, can also be used to represent the unique state sequence.

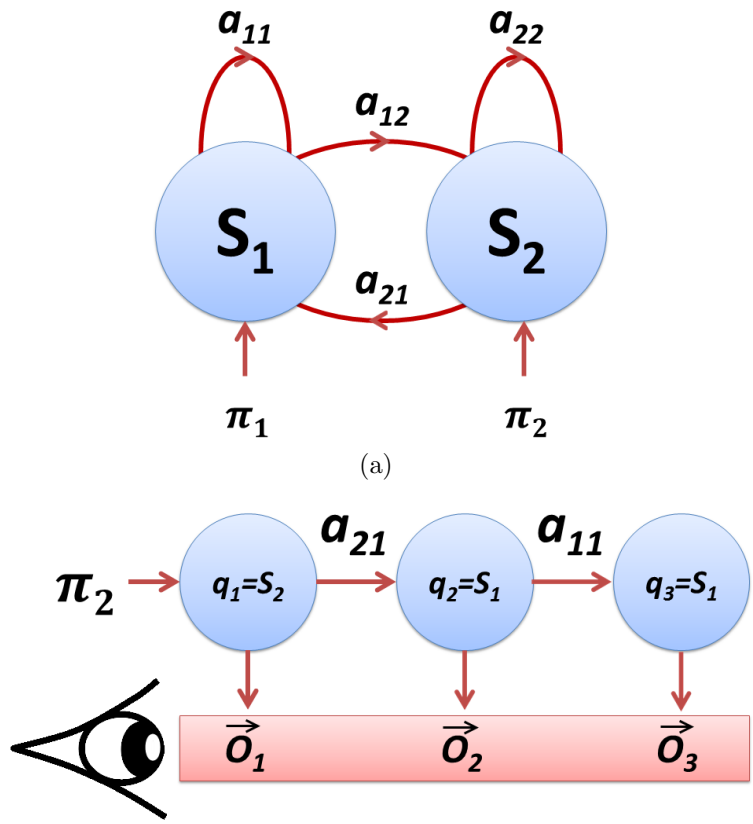


Figure 2–5: a) A first-order, finite-state, discrete-time Markov chain (DTMC) with two states (i.e., $N = 2$). The DTMC is defined by a *state space* $S = \{s_i : 1 \leq i \leq N\}$, a *state transition matrix* $A_{N \times N} = \{a_{ij} : 1 \leq i, j \leq N\}$ and a set of *initial state distribution* $\Pi = \{\pi_i : 1 \leq i \leq N\}$. 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.

The Markov process of a HMM is also defined by the parameters of the underlying DTMC. The only difference between the DTMCs and the HMMs is that in the HMMs the observations are generated according to a state-specific density function,

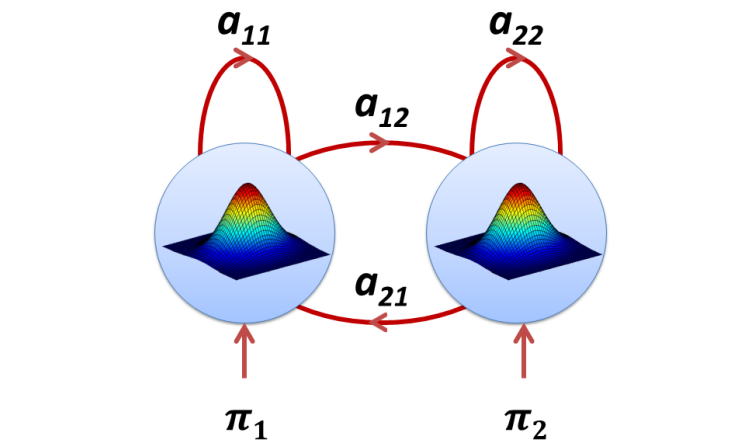
called *the observation pdf* (B). In contrast to the observations of a DTMC, in a HMM the observation pdf of different states can overlap and might generate the same observation as the output. Therefore, in HMMs we cannot directly map an observation to a unique state, which makes the states hidden to 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

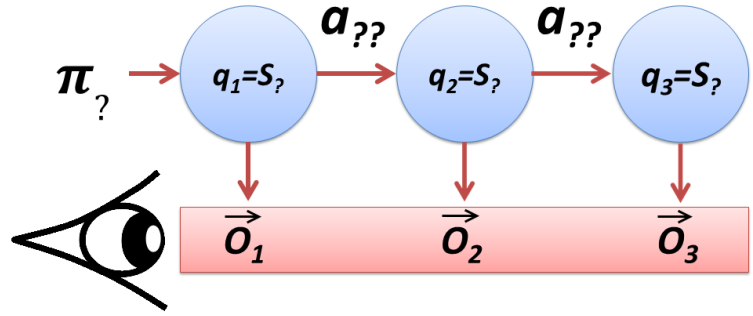
$$b_j(\vec{O}_t) = P(\vec{O}_t | q_t = s_j), \qquad 1 \leq j \leq N, 1 \leq t \leq T \tag{2.9}$$

Figure 2–6a shows a HMM with two states, similar to the DTMC shown in Figure 2–5a. In this example, similar to what we will see in the HMM-based attention model, each observation (i.e., $\vec{O}_t$) is a 2D vector generated according to the state-specific, 2D Gaussian distribution functions.

Figure 2–6b shows a sample outcome of the HMM of figure 2–6a. The outcome of the process is an observation sequence $\{\vec{O}_t : 1 \leq t \leq 3\}$, where $\vec{O}_t$ is the observation at time t. As it is shown in this figure, the states are hidden to the observers and only the observations are overt.



(a)



(b)

Figure 2–6: 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), a 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 the density function, B.

HMMs can be used as a generative model to reproduce sequences of observations that are consistent with the implicit Markov structure of their model. To generate a sample trajectory of length T, we have to choose an initial state according to the initial state distribution Π, choose an observation ($\vec{O}_t$) according to the observation parameters (B) of the current state, choose the next state according to the state transition probability (A) and repeat the process for T times (see figure 2–7).

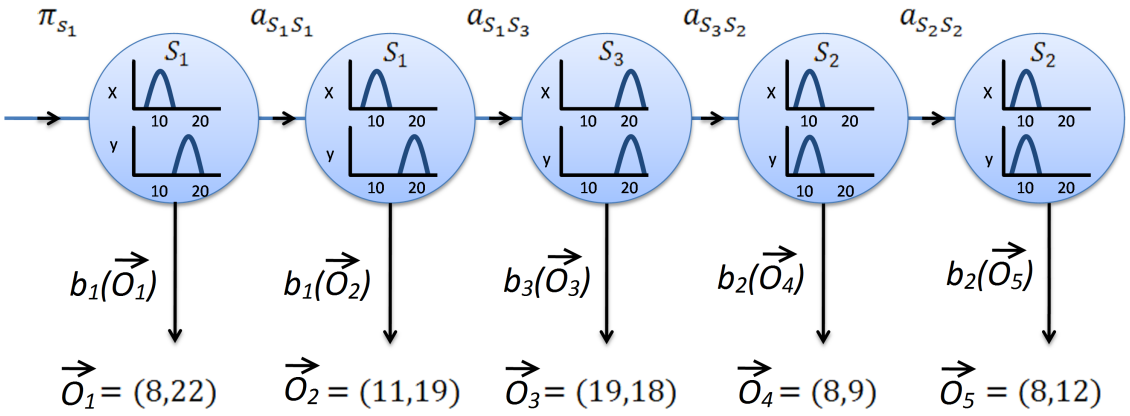


Figure 2–7: HMMs as generating models. We can generate a sample observation sequence $(\vec{O}_1, \vec{O}_2, \ldots, \vec{O}_T)$ by using the probabilistic parameters of HMMs $\{A, B, \Pi\}$. In this example we chose $M = 2$ and $\vec{O}_t = (o_{t,1}, o_{t,2})$.

In the literature related to the HMMs we can always find three fundamental problems that are of main interest: evaluation, decoding and training. Assume we are given an HMM, λ, and a sequence of observation, $\mathbf{O}$. Evaluation or scoring is the computation of the probability of the observation sequence given the HMM, i.e., $P(\mathbf{O}|\lambda)$. Decoding finds the best state sequence that maximizes the probability of the observation sequence given the model parameters. Finally, training adjusts

model parameters to maximize the probability of generating a given observation sequence (training data). The algorithms that cope with evaluation, decoding and training problems are called forward (or backward), Viterbi and Baum-Welch algorithm, respectively. In the rest of this chapter we review the methods for evaluation and training, as these two problems will be used in the proposed attention model introduced in the following chapters.

2.3.2 Evaluation

In the evaluation we want to calculate the probability of $P(\mathbf{O}|\lambda)$ for a given sequence of observation, $\mathbf{O}$. One method to do that would be to evaluate it exhaustively over all possible state sequences:

$$\begin{aligned} P(\mathbf{O}|\lambda) =& \Sigma_Q P(\mathbf{O}, Q|\lambda) \\ P(\mathbf{O}, Q|\lambda) =& P(\mathbf{O}|Q, \lambda) P(Q|\lambda) \end{aligned} \tag{2.10}$$

For a given state sequence $Q = q_1 q_2 ... q_T$ we would have:

$$\begin{aligned} P(\mathbf{O}|Q, \lambda) =& b_{q_1}(\vec{O}_1) b_{q_2}(\vec{O}_2) ... b_{q_T}(\vec{O}_T) \\ P(Q|\lambda) =& \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} ... a_{q_{T-1} q_T}. \end{aligned} \tag{2.11}$$

Therefore:

$$P(\mathbf{O}|\lambda) = \Sigma_{q_1 q_2 ... q_T} \pi_{q_1} a_{q_1 q_2} b_{q_1}(\vec{O}_1) a_{q_2 q_3} b_{q_2}(\vec{O}_2) a_{q_2 q_3} ... a_{q_{T-1} q_T} b_{q_T}(\vec{O}_T) \tag{2.12}$$

However with T observations and N states in the model we have N^T possible states and approximately $2TN^T$ operations. A more efficient way to calculate the term $P(\mathbf{O}|\lambda)$ is to use iterative algorithms of *forward algorithm* or *backward algorithm*.

Forward Algorithm

In this method we define $\alpha_t(i)$ as the probability of observing the first t observations ($\vec{O}_1$ to $\vec{O}_t$) with the state sequence terminating in state $q_t = s_i$, given the parameters of the HMM, λ, i.e.:

$$\alpha_t(i) = P(\vec{O}_1\vec{O}_2...\vec{O}_t, q_t = s_i|\lambda) \tag{2.13}$$

Given this definition we can calculate the probability of a given observation sequence (i.e., $P(\mathbf{O}|\lambda)$) by iteratively calculating the next value of $\alpha_t(i)$ given its $t-1$ previous values. Eventually we can terminate the process when reaching the end of the observation sequence, where we can calculate the observation probability by summing over different final states:

- *Initialization*: $\alpha_1(i) = \pi_i b_i(\vec{O}_1), 1 \leq i \leq N$
- *Induction*: $\alpha_{t+1}(i) = [\Sigma_{j=1}^{N}\alpha_t(j)a_{ji}]b_i(\vec{O}_{t+1}), 1 \leq t \leq T-1$ and $1 \leq i \leq N$
- *Termination*: $P(\mathbf{O}|\lambda) = \Sigma_{i=1}^{N}\alpha_T(i)$

This way with T observations and N states, we require approximately N^2T operations.

Backward Algorithm

Another way to calculate the term $P(\mathbf{O}|\lambda)$ efficiently is to do it in the reverse direction of the forward algorithm. In this method we define:

$$\beta_t(i) = P(\vec{O}_{t+1}\vec{O}_{t+2}...\vec{O}_T, q_t = s_i|\lambda) \tag{2.14}$$

as the probability of observing the last $T-t$ observations ($\vec{O}_{t+1}$ to $\vec{O}_T$) with the state sequence being in the i^{th} state at the time step t (i.e., $q_t = s_i$), given the parameters

of the HMM, λ. Given this definition we can calculate the observation probability of a given observation sequence (i.e., $P(\mathbf{O}|\lambda)$) following the same procedure as in the forward method, but in the reverse direction:

- *Initialization*: $\beta_T(i) = 1, 1 \leq i \leq N$
- *Induction*: $\beta_t(i) = \Sigma_{j=1}^{N} a_{ij} b_j(\vec{O}_{t+1}) \beta_{t+1}(j), t = T-1, T-2, ..., 2$ and $1 \leq i \leq N$
- *Termination*: $P(\mathbf{O}|\lambda) = \Sigma_{i=1}^{N} \pi_i b_i(\vec{O}_1) \beta_1(i)$

Similar to the forward method the complexity of this method for T observations and N states is approximately N^2T.

2.3.3 Training

In order to train the parameters of an HMM we need to have sequences of training observations $\mathbf{O}$. The goal is to train the model parameters of an HMM, $\lambda = \{A, B, \Pi\}$, so that the observation probability $P(\mathbf{O}|\lambda)$ becomes maximum for the training database. In the theory of HMMs, a method based on dynamic programming, called Baum-Welch (also known as the Forward-Backward algorithm), is suggested to iteratively find a solution to this problem.

Baum-Welch Algorithm

In the Baum-Welch algorithm, given an initial HMM model, λ, we estimate a new set of model parameters, $\widehat{\lambda}$, so that $P(\mathbf{O}|\widehat{\lambda}) \geq P(\mathbf{O}|\lambda)$. In order to do this, we need to extensively use forward and backward methods and based on their evaluation iteratively improve the parameters.

We define $\xi_t(i, j)$ as the probability of being in state s_i at time t and state s_j at time $t + 1$, given the model λ and the observation sequence $\mathbf{O}$, i.e.,

$$\xi_t(i, j) = P(q_t = s_i, q_{t+1} = s_j|\mathbf{O}, \lambda) \tag{2.15}$$

Given this definition we will have:

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(\vec{O}_{t+1})\beta_{t+1}(j)}{P(\mathbf{O}|\lambda)} = \frac{\alpha_t(i)a_{ij}b_j(\vec{O}_{t+1})\beta_{t+1}(j)}{\Sigma_{i=1}^{N}\Sigma_{j=1}^{N}\alpha_t(i)a_{ij}b_j(\vec{O}_{t+1})\beta_{t+1}(j)} \tag{2.16}$$

Based on the above definition, we define two more posterior probabilities that will be used in the iterative method for finding the model parameters:

- $\Sigma_{t=1}^{T-1}\Sigma_{j=1}^{N}\xi_t(i,j)$: Expected number of transitions from s_i
- $\Sigma_{t=1}^{T-1}\xi_t(i,j)$: Expected number of transitions from s_i to s_j

Given these parameters we can re-estimate the model parameters as follows:

- *Initial State Probabilities:*

  Expected number of time instances in state s_i at time $t = 1$

  $$\widehat{\pi} = \Sigma_{j=1}^{N}\xi_1(i,j)$$

- *Transition Probabilities:*

  Expected number of transitions from s_i to s_j over expected number of transitions from s_i

  $$\widehat{a}_{ij} = [\Sigma_{t=1}^{T-1}\xi_t(i,j)]/[\Sigma_{t=1}^{T-1}\Sigma_{j=1}^{N}\xi_t(i,j)]$$

- *Observation Probabilities:*

  Expected number of times in state s_j and observing $\vec{O}_t$ over expected number of times in state s_j

  $$\widehat{b}_j(\vec{O}_t) = [\Sigma_{t'=t}^{T-1}\Sigma_{i=1}^{N}\xi_{t'}(j,i)]/[\Sigma_{t=1}^{T-1}\Sigma_{j=1}^{N}\xi_t(j,i)]$$

Therefore the Baum-Welch algorithm tunes the parameters to the training set using the following iterations:

- *Initialization*:

  Obtain initial estimation for $\lambda = \{A, B, \Pi\}$

- *Likelihood Computation*:

  Compute likelihoods $\alpha_t(i)$ and $\beta_t(i)$ and posterior probabilities $\xi_t(i,j)$ for $i,j = 1,...,N$ and $t = 1,...,T$

- *Parameter Update*:

  Given the likelihoods and the posterior probabilities, compute $\widehat{\pi}$, $\widehat{a}_{ij}$ and $\widehat{b}_j(\vec{O}_t)$

- *Termination*:

  Repeat Steps *Likelihood Computation* and *Parameter Update* steps until convergence, i.e.,

$$(\log P(\mathbf{O}|\widehat{\lambda}_{i+1}) - \log P(\mathbf{O}|\widehat{\lambda}_i)) < \epsilon$$

The algorithms for training the HMMs and using them to evaluate the probability of an observation sequence give us the necessary tools for developping our HMM-based attention models that will be introduced in the next chapter.

CHAPTER 3
Using HMMs as Attention Models

As shown in chapter 2, classical attention models are based on a spatial map that defines the conspicuous locations that are potential targets of the fixations. However, there are several factors that make the saliency-based models inaccurate when it comes to modeling the real eye movements while viewing an image.

Higher-order processes are shown to influence the selection of targets in an eye trajectory. These processes affect the selection of the next target based on the recently fixated ones, which is in conflict with the independence assumption in the likelihood estimation of the saliency-based models used to derive equations (2.4) and (2.6). For instance, *Proximity preference* is a cognitive process that facilitates fixations near the currently fixated target and *similarity preference* is a cognitive process that favors the similar objects to the one that is currently being fixated [72]. *Inhibition of return* (IOR) [70] is another high-level process that discourages fixation on the target that has just been visited.

Another issue with the classical attention models is the implicit assumption of overtness of visual attention. While most of the times the FOA follows the COG, this is not always the case. When a visual-task is given to an observer, although correctly executing the task needs directing the FOA to certain targets in an image,

the observed COG trajectory can vary from subject to subject.[1] In other words, eye position does not tell the whole story when it comes to tracking attention [14].

In this chapter a new attention model based on the theory of HMMs is proposed that addresses these issues by relaxing the independence and overtness assumptions that are implicit in the classical attention models. Before describing the proposed model, we elaborate on the dificiencies of the classical models and show sample eye trajectories that do not fully comply with these assumptions.

3.1 Eye Movements are Sequential

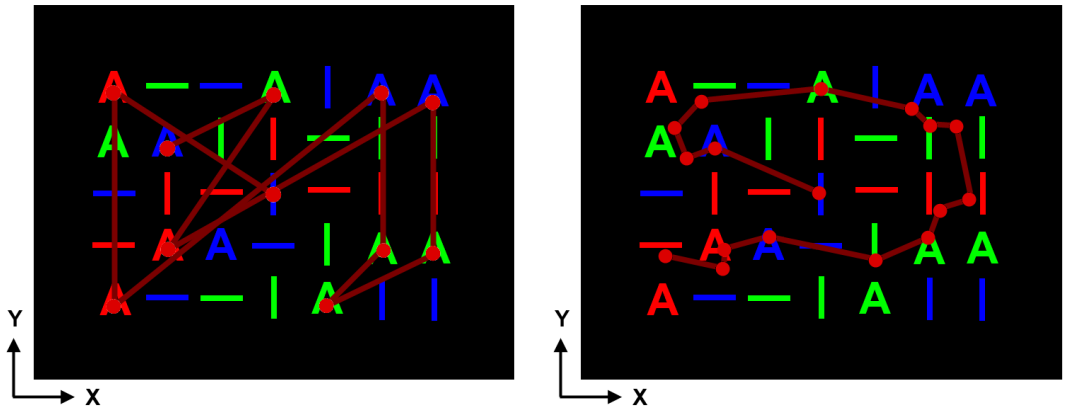
In both bottom-up and top-down viewpoints in modeling the visual attention, the probability of successive observations are assumed to be mutually independent and an application of naïve Bayes assumption is implicated in equations (2.4) and (2.6). In other words, in classical models the probability of fixating on a certain location in the image only depends on the saliency of that point in the saliency map and it is assumed to be independent of the previously fixated targets in the scene. However, this assumption is inconsistent with what has been demonstrated in the human visual psychophysical experiments [97]. For instance, early studies by Engel [30, 31] indicated that in a visual search task, the probability of detecting a target depends on its proximity to the location currently being fixated (*proximity*

[1] So far we used COG and FOA interchangeably, but from now on, after explaining the difference between these two phenomena, we will distinguish between these two terms.

preference). Although Dorr et al. [24] suggest that in a free viewing of a scene, low-level features at fixation contribute little to the choice of the next saccade, Koch et al. [72] and Geiger et al. [40] suggest that in a task-involved viewing, the processing focus will preferentially shift to a location with the same or similar low-level features as the presently selected location (*similarity preference*).

Perhaps the discrepancy between artificial and natural eye trajectories can best be demonstrated by comparing a trajectory produced by a saliency-based model against a recording of the eye movements. Figure 3–1a shows an artificial eye trajectory produced by a top-down saliency map and figure 3–1b shows a recording of the eye movements of a subject. In both cases the task was to count the number of characters and an identical stimulus was used for both scenarios. For the top-down model the saliency map of Itti and Koch [61] is used (for the diagram see figure 2–3 and for the saliency map based on which the eye trajectory is generated see figure 2–4b). As we can see the COGs in figure 3–1a (artificial) are sparse, while those of figure 3–1b (natural) are more correlated to their predecessors'.

The successive COGs of the real eye trajectory show dependence to each other in the sense that the local transitions between them are encouraged while the remote transitions to a distant location in the image are discouraged. In the eye trajectory generated by the saliency model, however, the COGs are mutually independent, resulting in a random exploration of the "A"s in the trajectory. This effect is the direct result of the naïve Bayes assumption that is embedded in the likelihood terms of the classical models of attention.



(a) (b)

Figure 3–1: Comparison of an eye trajectory produced by the classical models of attention vs. natural task-dependent eye movements. The task is to count the number of characters in the image. a) The trajectory produced by the top-down model. b) The trajectory obtained by recording a subject's eye position while performing the task.

An alternative approach to obtain the density function of the likelihood term is to allow for dependence between the attributes $\vec{O}_1 \ldots \vec{O}_T$. *Bayesian belief networks* are means to take into account the dependence between attributes in a graphical way. Since the observations $(\vec{O}_i : i \in [1, T])$ are sequentially sampled in time, we propose to use the *dynamic Bayesian networks* to represent sequences of observations [41].

The simplest form of dynamic Bayesian networks is the *Markov process*. In a study Hacisalihzade et al. [47] used Markov processes to model the visual fixations of observers during the task of recognizing an object. They showed that the eyes visit the features of an object cyclically, following somewhat regular scanpaths rather than

crisscrossing it at random.[2] In another study, Elhelw et al. [29] also successfully used a first-order, discrete-time, discrete-state-space Markov chain to model eye movement dynamics. Stark et al. [110] also came up with a Markov process as a general model of fixation placement during the task of reading. Pieters et al. [95] also observed a similar pattern in the scanpaths of the observers while looking at printed advertisements.

Based on the successful application of Markov process to model the eye-position trajectories, which is correlated to the attention location trajectories [100], we propose a first-order, discrete-time, discrete-state-space Markov chain to model the attention cognitive process of the human brain. Such a process is called a first order Markov chain, if for any $t > 0$:

$$P(q_{t+1}|q_t, q_{t-1}, \ldots, q_1) = P(q_{t+1}|q_t). \tag{3.1}$$

In this equation each eye fixation is assigned to one of the pre-defined states in an image and the Markov process defines the probability of transitions from a state to another in an eye trajectory. This interpretation forms a finite-state, *discrete-time Markov chain* (DTMC) that gives us the likelihood of an eye trajectory based on the loci of fixations (details in section 3.3). Moreover, if we posit a first-order Markov process as the underlying process that governs the transitions between the states (which is shown to be a valid assumption for eye movements [47]), we can train a

[2] Repetitive and idiosyncratic eye trajectories during a recognition task is called scanpath [88].

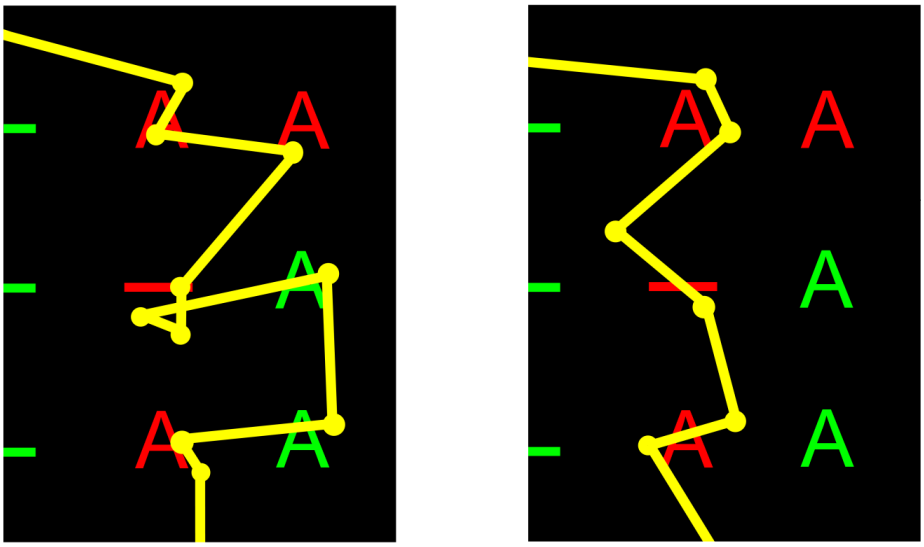
first-order DTMC for each task. This model is used by Elhelw et al. [29], where they successfully used a first-order DTMC to model eye movement dynamics.

3.2 Overt vs. Covert Visual Attention

While it is well known that there is a strong link between eye movements and attention, the attentional focus is nevertheless frequently well away from the current eye position [36]. The classical attention models that are based on eye tracking may be appropriate when the subject is carrying out a task that requires foveation. However, these methods are of little use (and even counter-productive) when the subject is engaged in tasks requiring peripheral vigilance. Moreover, due to the noisy nature of the eye-tracking equipment, the actual eye position itself is usually different from what the eye-tracker shows, which will bring in systematic error to the estimations.

Figure 3–2 shows two different eye trajectories recorded while viewers were counting the number of characters in a synthetic image. As can be seen, these two images illustrate different levels of linkage between the COG and FOA. In figure 3–2a, fixation points mainly land on the targets of interest (*overt attention*), whereas in the other instance (figure 3–2b), 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 first scientist to provide an experimental demonstration of covert attention is known to be Hermann Von Helmholtz [55]. In his experiment, Helmholtz briefly illuminated inside a box by lighting a spark and looked at it through two pinholes. Before the flash he attended to a particular region of his visual field without moving



(a) (b)

Figure 3–2: Eye trajectories recorded while executing a task on the same stimulus. In the trajectories straight lines depict saccades between two consecutive fixations (shown by dots). In this figure two snapshots of the eye movements during the task of "counting the number of characters" is shown. The results from counting the characters were correct for both cases. Thus, the target that seems to be skipped over (the middle right "A" in b) has been attended at some point.

his eyes in that direction. He showed that only the objects in the attended area could be recognized implying attention can be away from the eye movements.

In real life, human attention often deviates from the locus of fixation to give us knowledge about the parafoveal and peripheral environment. This knowledge can help the brain decide about the location of the next fixation that is most informative for building the internal representation of the scene. This discrepancy between the FOA and the COG helps us efficiently investigate a scene, and at the same time makes the FOA covert and consequently hard to track.

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, equipment bias, undershooting or overshooting of the target [3], or the phenomenon of *center-of-gravity fixations* [130, 53, 87] are some of the most common sources of recurrent divergence between the COG and the FOA, which also cause the occurrence of dissimilar eye trajectories for a given task.

3.3 Fundamentals of a HMM-Based Attention Model

One solution to the problem of detecting the covert attention is to force the binding of attention to eye movement by decreasing the signal-to-noise of the image, i.e., lowering the ratio between signal strength (target salience) and noise strength (distractor salience). In this way, the targets become more resolution demanding and will entail foveation in order to be distinguished from their surrounding distractors. The resulting COG, then, will be the same as the FOA trajectory.

The manipulation of SNR has been studied before by Koch et al. [72] as a way to attract attention to a certain target in an image. In another study, Wolfe et al. [125] proposed maximizing the SNR to decrease the search time. In their experiments, they noticed an increase in the search time by lowering the SNR in synthetic stimuli.

Although by decreasing the SNR we could obtain attention trajectories rather than eye trajectories, manipulation of the image SNR is not always feasible. For instance, in natural images, or more generally in non-synthetic stimuli, we have limited control over the image to adjust the saliency of targets. Figure 3–3 shows an eye trajectory recorded while a viewer was performing the task of "counting the number of people" in a natural image, which shows that the covert attention could

happen in natural images as well, where the application of the SNR manipulation technique to bind the COG and the FOA is not straightforward.

(a) (b)

Figure 3–3: a) The original image, on which the task is performed. b) An eye trajectory recorded while executing the task of "counting the number of people in the image". In the trajectory the 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) seems to be overlooked. The target that did not get any fixations is either attended covertly or has been fixated by the parafoveal vision.

In order to relax the overtness and independence assumptions in the classical models of attention, in this thesis we propose to use the HMMs as a better alternative to the classical models of attention in tracking the overt and covert shifts of attention. As opposed to the classical methods of attention tracking, the proposed model relaxes the overtness constraint postulated in both bottom-up and top-down methods by using the state-specific Gaussian observation probability density functions to model the task-dependent attention process. Moreover, the underlying Markov process of

the HMMs implicates the sequentiality of the attention and allows for dependence between consequetive fixations.

HMMs allow for covert attention by postulating the fixations to be the outcome of the observation distributions, which can be a point away from the FOA, whereas in the top-down and bottom-up attention models the fixation location is assumed to be on attention demanding spots in an image. When entering a state of a HMM, a Gaussian distribution function generates a fixation as an observation that is overt to the viewer (recall figure 2–6b) while the attentional state is covert to the viewer. Thus, in our proposed model the states represent the FOAs, and the COGs form the observation sequences.

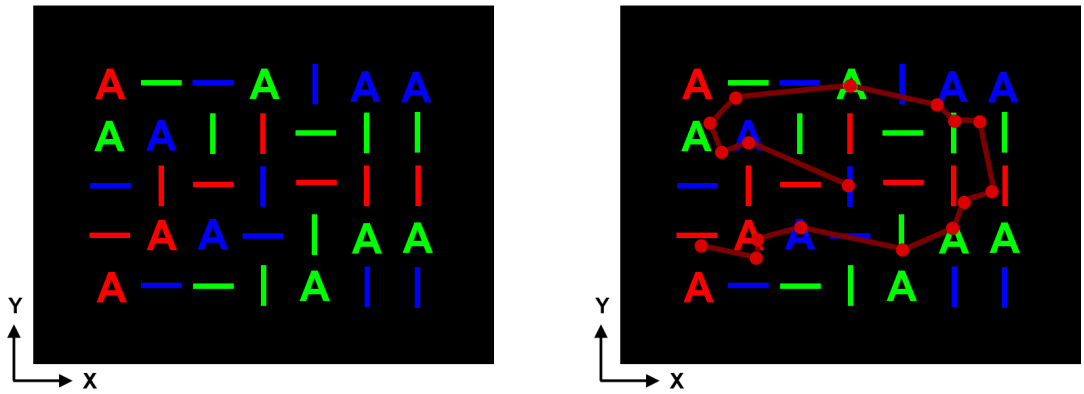
In the model each state is designated to one or several potential attentional targets and each observation density function is defined by a 2D Gaussian function centered on each target (for now we assume all of the objects in an image are potential attentional targets for a given task). In other words directing attention (covert or overt) to a target is equivalent to going to the state designated to that target and recording an eye position is equivalent to generating a random outcome from the 2D Gaussian observation pdf of that state that is centered on the target. The location of the COGs, thus, can be away from the target that is being attended.

To train the HMMs, the task-dependent eye movement trajectories are used. In the training phase the transition probabilities (A), initial state distributions (Π) and the observation pdfs (B) are trained to form task-dependent models λ_θ, which in turn can be used to evaluate the probability $P(\mathbf{O}|\lambda_\theta)$ for the test trajectories.

In order to better understand how the concepts of hidden state and observation in a HMM relate to the covert FOA and overt COG, respectively, here we sketch a prototype that employs HMMs as a cognitive process model of attention for both a synthetic and a natural image.

Figures 3–4a shows a sample synthetic stimulus comprising characters among horizontal and vertical bars. All items are in blue, red or green colors and they are placed on a 5×7 grid with a plain black background. Figure 3–4b shows eye fixation locations of a subject while performing a task (counting the number of characters) superimposed on the original stimulus. The fixation locations sometimes undershoot or overshoot the targets due to the oculomotor properties of the human eyes or noisiness of the eye tracker. In our model we posit that these fixation locations constitute the observations in the HMM context. We postulate that these observations are random outcomes of a 2D Gaussian probability density function (with features x and y in Cartesian coordinates), which is maximum on the target location and fades off as we become more distant (Euclidean) from the targets.

To build a database of task-dependent eye trajectories, we ran 1080 trials and recorded the eye movements of six subjects while performing a set of pre-defined simple visual-tasks. Five different visual stimuli were generated by a computer, each of which containing 30 objects randomly selected from a set of nine objects (horizontal bar, vertical bar and character "A" in red, green and blue colors) that were placed at the nodes of an imaginary 5×6 grid ($6.75^o \times 8.1^o$) superimposed on a black background. The visual-tasks were counting red bars, green bars, blue bars, horizontal bars, vertical bars or characters (six tasks in total). Each of the tasks



(a) (b)

Figure 3–4: a) The original synthetic image, on which the task is performed. b) Eye trajectories recorded while executing the task of "counting the number of characters", superimposed on the image.

was defined so that their corresponding targets can be distinguished from distractors by a single feature. We used an eye tracker (ISCAN RK-726PCI) to record the participant's left eye positions and classified the eye movement data into saccades and fixations using the velocity-threshold identification (I-VT) method [34] with a 50 deg/sec threshold.

For training the task-dependent HMM for the task of "counting characters" the Baum-Welch algorithm is applied on a training set selected from the database of task-dependent eye trajectories. For the initial values used in the Baum-Welch training, we used an ergodic HMM comprising N states, each of which is dedicated to one of the targets in the image (in this example characters, thus $N = 12$). Random values (with a normal distribution) were used as the initial values of the transition matrix (A) and the initial state probabilities (Π). As for the observation pdfs (B), we used 2D Gaussian distributions and centered them on each of the targets in the image.

During the training we optimized the covariance and mean of the observation pdfs, as well as the transition and initial state distributions to maximize the likelihood of the training set.

This layout is used as a simple structure for a task-dependent HMM to demonstrate the idea of training HMMs as attention models and use them to account for the off-target fixations in a natural eye trajectory. That said, the model selection and initialization of the parameters have a significant effect on the resulting attention models and their accuracy to model the attentional deployment on an image and need to be more specific to the characteristics of the tasks and the stimuli. In later chapters, we will see that defining the task-specific targets in advance and using that information to train the HMMs is not always feasible. Thus, in chapters 4 we will elaborate on different structures of the HMMs given different types of tasks and stimuli.

Figure 3–5 shows a depiction of the observation density associated with the HMM that is trained on the eye-trajectories recorded while executing the task of counting the number of characters in the synthetic image using the above training scheme. Each 2D Gaussian pdf is demonstrated by a top-view heat map, where the heat shows higher probabilities. Each Gaussian pdf represents an attentional state and at each time step a COG coordinate pair is generated by drawing a random outcome from a pdf that is randomly selected according to the transition probabilities. For instance, directing the FOA to the bottom left character can result in a fixation that is further away from the physical boundaries of the character. The capability of Gaussian HMMs in representing off-target fixations is illustrated in this image by

overlaying the trajectory of figure 3–4b on the image. While the classical models seemed to fail to account for off-target fixations, here we show that the Gaussian observation function can properly justify them.

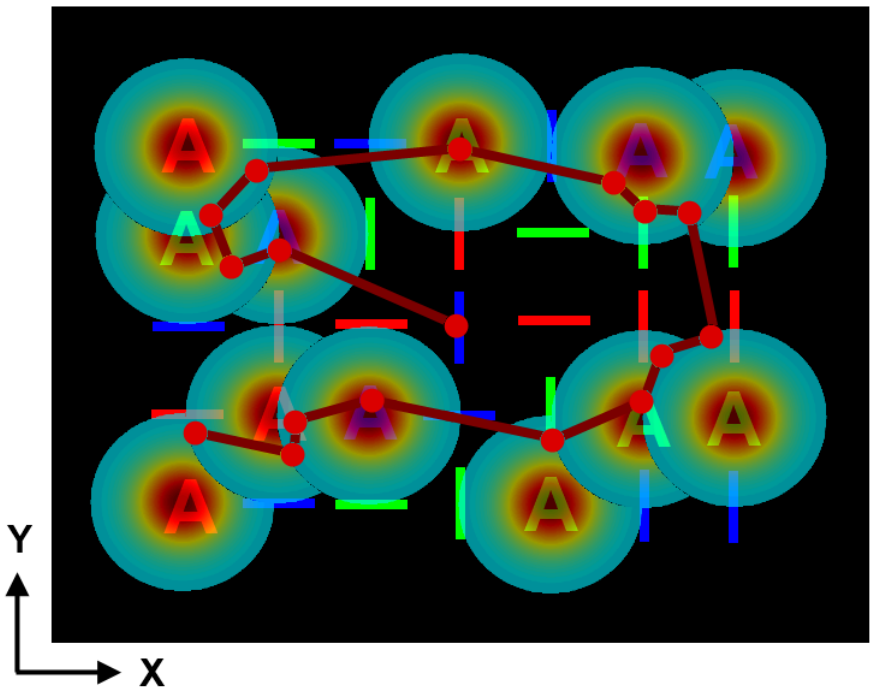


Figure 3–5: Overlaying the 2D observation Gaussian distributions on a synthetic image. The combination of the Gaussian pdfs form the HMM that is trained for the task of counting the number of characters in the image. The overall model can generate synthetic eye-trajectories based on the parameters of 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.

Figure 3–6 shows a HMM that is trained on the eye-trajectories recorded while executing the task of counting the number of people in a natural image. Similar

to the HMM of figure 3–5, the task-specific attention model allows for off-target fixations by using the Gaussian observation pdfs as the generative models of the fixation locations. In this attention model, a similar training scheme was used and each Gaussian represents an attentional state. The resulting HMM is also a fully ergodic HMM that allows for transitions from each state to one another.

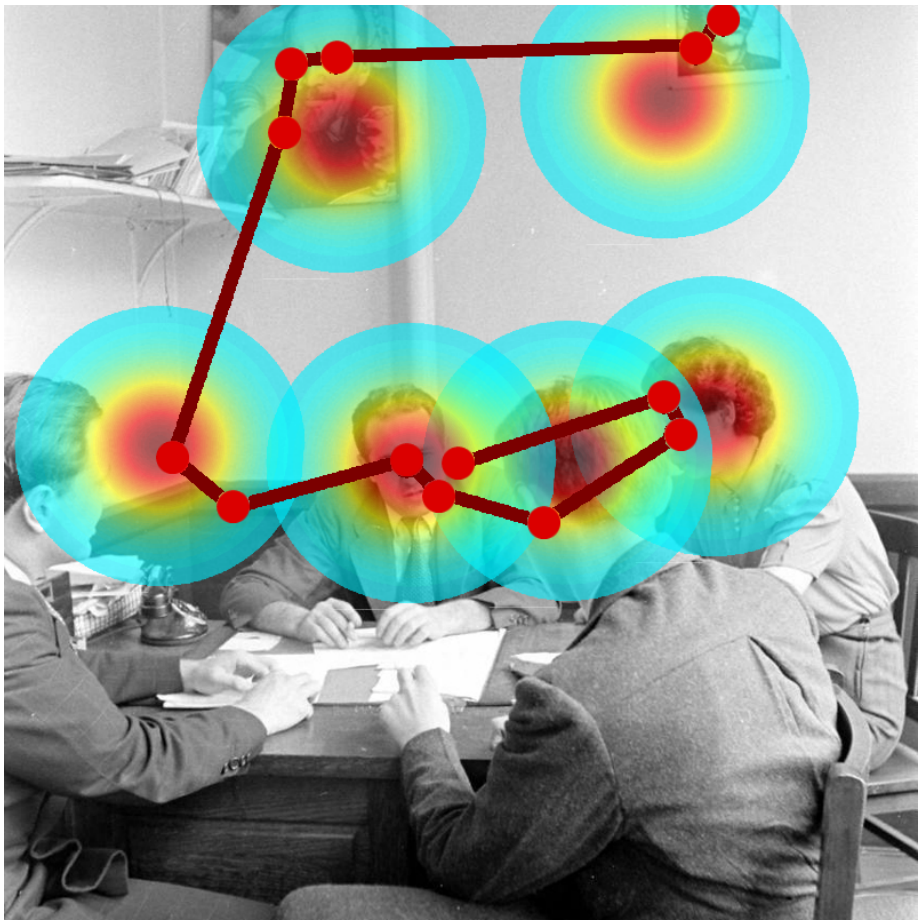


Figure 3–6: Overlaying the 2D observation Gaussian distributions on an image. The combination of the Gaussian pdfs form the HMM that is trained for the task of counting the number of people in a natural image. The overall model can generate synthetic eye trajectories based on the parameters of the HMM. 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.

3.4 Discussion

The classical models of eye movement analysis are limited in terms of accounting for real-world eye movements of natural vision due to their pure dependence on low-level image features. Low-level features, however, are not always conspicuous in the fixation locations of a task-dependant eye movement as shown in the eye recordings of observers in natural task execution [1, 77, 74, 52]. As an alternative to the classical, salience-based, attention models, it is shown in the eye movement studies that the fixation map of human observers outperforms even the best salience-based models in terms of prediction rate of the eye trajectories [65]. Thus, in this chapter we studied HMMs as a model that is based on real, task-dependent eye trajectories recorded while viewing natural images. To do so, we used 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.

Using the HMM-based method not only allows us to track overt foci of attention, but also allows for covert attention and other sources of discrepancy between the center of gaze (COG) and the focus of attention (FOA). The HMMs allow for decoupling the COG and the FOA by means of the state-specific Gaussian distribution functions. The Gaussian distribution functions used in model definition of the HMMs span an area around the attentional spots, the random outcome of which can be well away from its center. By this interpretation of variables the HMMs allow for decoupling between the COG and the FOA.

A benefit of using time-series analysis to model eye movements is incorporating the temporal information as well as the spatial features of fixations into the model.

In analyzing the eye movement behavior, spatial information is usually taken into account and temporal information of fixations is simply omitted from many models of eye movement analysis. However, it is becoming increasingly more clear that temporal analysis of eye movement is as important as its spatial aspect in describing the underlying mechanism of the visual behavior. This is while evidence from natural tasks emphasizes the need to consider fixation duration as well as fixation location in understanding the mechanism of the visual system [25, 51, 74].

In both natural and synthetic images we showed that contrary to the classical methods of attention tracking, the proposed model relaxes the overtness constraint postulated in both bottom-up and top-down methods by using HMMs to model the task-dependent attention process. HMMs allow for covert attention by postulating the fixations to be the outcome of observation distributions, which can be a point away from the FOA, whereas in the top-down and bottom-up attention models the fixation location is assumed to be on attention demanding spots in an image. Therefore, by using the parameters of the task-specific HMMs (λ_θ) in the forward algorithm we could calculate the likelihood term of $P(\mathbf{O}|\lambda_\theta)$.

Although in both examples presented for natural and synthetic images the task-specific objects were easy to spot (characters in figure 3–5 and faces in figure 3–6), this is not always the case. For abstract tasks, such as the ones used in the original experiment by Yarbus [128], defining the targets is not straightforward. In synthetic images, we probably know the number of objects in the image when generating the image, which leaves us the option to assign a state to each of them for training the HMM. However, in case of natural images we usually do not have a clear notion of

the objects and targets are not separated in the space, which complicates the design and initialization of the HMM (if not making it impossible).

In the next chapter we study how we can apply the idea of the HMM-based attention models to infer the task in synthetic and natural images. For the natural images a solution is proposed to design and initiate the HMMs for abstract tasks, where defining task-specific targets is not straightforward.

One important point to be noted is that an "off-target" fixation does not necessarily mean that the FOA is away from the fixation (i.e. covert). For instance, in a phenomenon known as *the center-of-gravity* (also known as the *global effect*), the stimulus is actually the collection of features, and the location of the stimulus is the center-of-mass of this collection [130, 53, 87]. Thus, the FOA is in this case overt, although the COG lands somewhere in between several objects rather than fixating on an object. However, in the proposed model we do not distinguish between the covert and overt attention and the goal is to track the attention in its general notion. That said, in the next chapter we will see that the hidden states of the HMMs are correlated to the attention demanding targets in executing a task and the states can be postulated as a good estimation of the attention location, even if the fixations are away from the state centers.

CHAPTER 4
Visual-Task Inference Using the HMM Attention Models

Inferring the visual-task from fixation positions is a hard problem in the sense that executing a task twice does not necessarily result in two similar trajectories and in fact the outcomes are usually quite different from each other. Also two similar trajectories may have been generated by two different tasks, which means that the mapping from the feature space to the task space is not a one-to-one mapping.

Several efforts have been made to infer the task by classifying the patterns of the task-dependent eye trajectories and recognizing it in a test eye trajectory. However, in most of the cases no significantly different pattern could be found in the features of the eye trajectories of a specific task, which makes a direct mapping from the feature space into the task space an ill-posed problem.

One of the most recent failed attempts to implement a direct mapping from the feature space into the task space is the study by Greene et al. in [45] and [46]. In these works, Greene et al. attempted to find a discriminative function to map eye trajectories to their visual-tasks. They used supervised learning techniques to train three different classifiers based on linear discriminant analysis [84], correlational methods [49] and support vector machines [54]. The classifiers were trained on the so-called *summary statistics* of eye movements used in scanpath analysis [15, 85]. Seven features of eye trajectories including the number of fixations, mean fixation duration, mean saccade amplitude and percent of image covered by fixations were used as the

observation vector of each trajectory. The results showed that the classifiers can only infer the task at the chance level and fail to consistently reveal the underlying task of test trajectories. Thus, based on the results they 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."*.

While the results of inferring the task from summary statistics is shown to be disappointing, in our view concluding that the scanpaths cannot be used to identify the visual-task is overstated. In Greene's study only specific features derived from the aggregate characteristics of eye movements are used to train the classifiers and failing to infer the visual-task does not mean that task inference in general is infeasible.

In fact, in another study Castelhano et al. [16] studied the influence of task on a group of summary statistics (including the ones used in Greene's experiment) for 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, such as the summary statistics features. A similar result is obtained in [85], where they also used the same features as in Greene's study. Thus, almost all the studies dedicated to the realization of an inverse Yarbus process unanimously agree on the infeasibility of an inverse Yarbus mapping from the feature space to the task space by using the summary statistics of the eye movements.

Although executing a task may generate dissimilar eye trajectories and eye movement features, it usually requires directing attention to specific targets in an image

that are relevant to the task. By directing the focus of attention (FOA) to the task-relevant targets in an image, our eyes direct the processing power of the brain to the informative areas in the scene. Therefore, in our model we propose to first infer the attentional spots from an eye trajectory and then infer the task given the attended target. For instance, if an attention model shows that red objects in an image are attended, "Counting the red objects" can be a possible inference about the task.

As explained in the previous chapter, the task-dependent attention models give us possible attentional spots in a stochastic manner. Thus, in our proposed model we suggest a Bayesian inference that uses attention models to evaluate the likelihood term and as a result give the posterior probability of different tasks given an eye trajectory.

In section 2.2.1 and 2.2.2 we showed how attention models can serve to evaluate the likelihood term of equation (2.2). However, classical attention models have several insufficiencies that fail them to accurately predict the attentional spot and evaluate the likelihood term.

In the previous chapter we showed how HMMs can be used as a model for the visual attention. The application of HMMs in attention tracking addressed many deficiencies of the classical models. HMMs do not require the FOAs to be overt and allow for covert shifts of attention. Moreover, HMMs impose higher level processes to the transitions from a target to another, which is modeled by the underlying Markov process of the HMMs. In this chapter we will use the HMM-based attention model to evaluate the likelihood of a given eye trajectory and use it in the Bayesian inference framework to infer the visual-task.

In section 2.3 we showed how we can train HMMs by using the Baum-Welch algorithm. Thus, if we train the HMMs with task-dependent eye trajectories we end up with task-dependent HMMs, which can be denoted by λ_θ. In this notion, θ indicates the task associated with the trajectories of the training data. Given the task-dependent HMMs, the forward algorithm can evaluate the likelihood of an observation sequence, which can be shown by the term $P(\mathbf{O}|\lambda_\theta)$. By substituting the likelihood term into the Bayesian equation, equation (2.2) becomes:

$$P(\lambda_\theta|\mathbf{O}) = \frac{P(\mathbf{O}|\lambda_\theta)P(\lambda_\theta)}{P(\mathbf{O})}. \tag{4.1}$$

The training of the model parameters includes finding the parameters of a HMM (i.e., A, B and Π) that maximize the likelihood of the observation sequences of the eye trajectories (i.e., $P(\mathbf{O}|\lambda_\theta)$). In order to iteratively improve the observation likelihood (using the Baum-Welch algorithm), we need to start off from a generic structure for the HMM. This generic structure has to define the number of states, number of Gaussian observation pdfs in each state, transition pattern and initial values for the parameters of the HMMs. However, based on the task set and the characteristics of the stimuli we could require different structures in the generic HMM.

One of the key factors in model definition of the generic HMMs is the definition of the observation pdfs of the states. As explained in section 2.3.1, the observation pdfs (B), are composed of 2D Gaussian distribution functions that are located around the attentional targets in an image.

One of the main characteristics of synthetic images is that the number, location and specifications of the targets are usually known in advance or can be extracted

from the stimuli. Thus, in synthetic images we are able to spot the targets and initially deploy the Gaussians on them. In natural images, however, we do not usually know the location or number of the targets in an image. Therefore, before training the task-dependent attention models for natural images we need an approach to define the initial, generic HMM to be trained by the training database.

Despite the type of the stimulus, easy or complex visual-tasks also require different generic HMM structures. In easy tasks, attention is mostly directed to the task-relevant objects in the scene, whereas in the more complex tasks the difficulty of the tasks increases the response time and causes several attentional deployments on non-target objects (*off-target FOAs*) in order to examine their task-relevant features and dismiss them from potential target locations. These off-target FOAs on objects that are not directly relevant to the ongoing task will require a more complex HMM structure to be able to distinguish them from the FOAs made on targets and make the inference based on the *on-target FOAs*.

In the following sections we will progressively improve the result of task inference by amending the generic structure of the HMMs based on the level of complexity of the tasks and the type of stimuli. To this end, we study the natural and synthetic images and address the challenges related to each type of image, separately. First we start with implementing a model that infers the task among a group of simple tasks. Then we improve the model to deal with complex situations, as well. In the end we apply the model to the natural images and evaluate the recognition rate in images similar to those of Yarbus'.

4.1 Inferring Simple Tasks in Synthetic Images

In this section we use the proposed HMM-based model to infer the ongoing task in a simple visual search. The inference is made by applying the Bayes rule (equation (4.1)) to the observation sequences of a database of task-dependent eye movements. In order to obtain the HMMs for each visual-task, we need to train the parameters by using the eye movement database of the corresponding task. To do so, first we need to define the structure of our desired HMMs by creating the generic HMM.

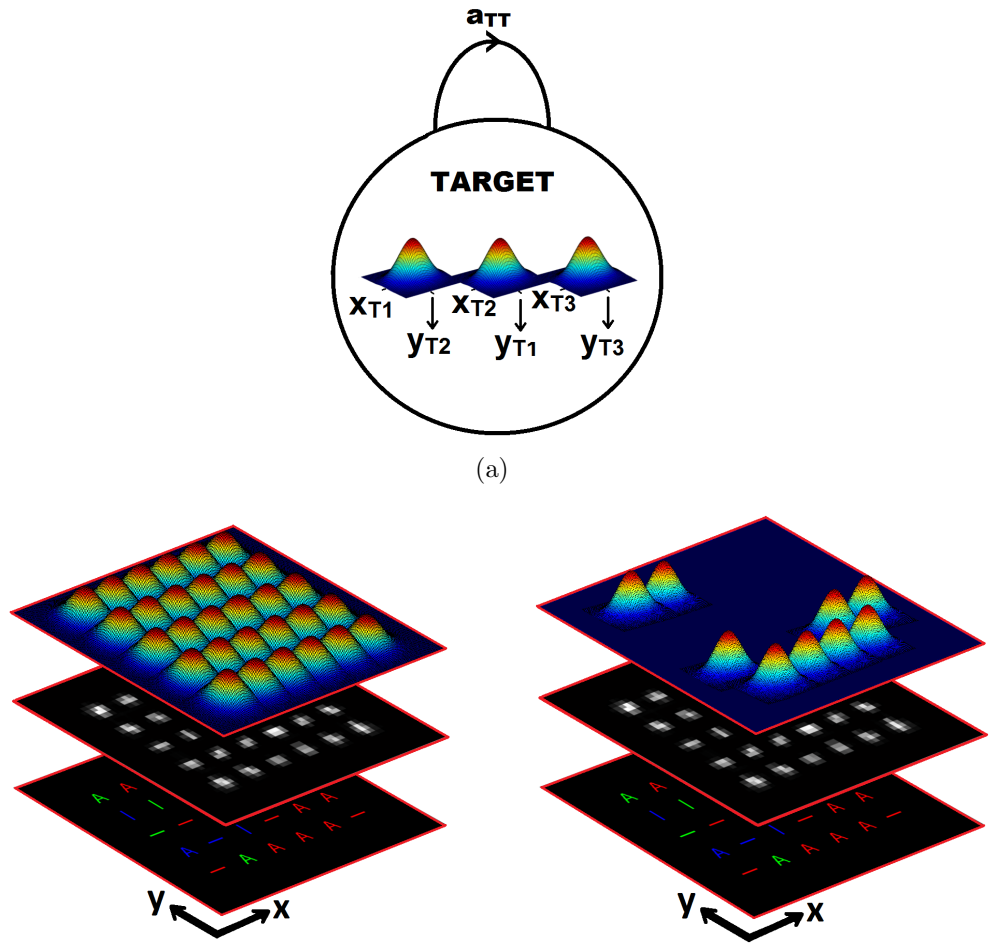
Since in a simple search task the targets are usually distinguished from the distractors by a single feature, attention is mostly directed to the task-relevant objects in the scene. This gives us a good basis to compare the results of the HMMs and the TD models, because TD models are ineffective in dealing with off-target attention deployment and assume the FOA to be mostly on targets. To be able to compare the models, in our HMM-based model for the easy search we also assume the FOAs to be mostly on targets and degenerate the conventional structure of the HMMs to a single-state, self-returning one, which results in a model that allows for covert attention and is similar to TD models, otherwise.

In the generic structure of the proposed *single-state HMM* (SSHMM) the state represents the target locations for different tasks. For the target locations we postulate that the observations are random outcomes of a mixture of 2-D Gaussians with features x and y in Cartesian coordinates that are maximum on the centroids of the targets and fade away as we become more distant, as measured by a Euclidean metric, from them (see figure 4–1a). In figure 4–1b we put the observation pdfs of all the

objects together and superimposed them on the original image and its corresponding bottom-up saliency map. It is from this grid of Gaussians that we select the ones related to the task and combine them into a Gaussian mixture model (GMM) with equal weights to represent the single state's observation pdf. Figure 4–1c shows an example of the GMM distribution for the task of counting the characters. To build the HMM, from the pool of the pdfs in figure 4–1b we only select the ones centered on the characters and remove the others from the generic model.

Although here we assume that the targets can be selected according to their task relevance, this assumption does not always hold. For some abstract tasks, no specific target might be directly relevant to the task. As mentioned in section 3.3, if for a given task it cannot be decided whether a target is directly relevant to a task or not, we should select all the targets to train the HMMs. In the training process less weight will be given to the transitions to the Gaussian pdfs around the targets that are less relevant to the task in exchange for a greater favor to the transitions to the Gaussians on the targets that are more relevant to the task.

However, postponing the target selection to the training phase causes a longer training time and less accurate results for a small training database. Moreover, this option is only valid for the stimuli with discrete objects. In section 4.3 we will propose a method to address this issue in natural images, where targets are spread around in an image and overlap with each other.



(a)

(b)

(c)

Figure 4–1: The structure of the single-state HMM (SSHMM). a) The generic SSHMM for simple task inference. The transition matrix (A) is composed of a deterministic loop from the state to itself ($a_{TT} = 1$) and the observation pdf comprises mixture of Gaussians centered on target-relevant objects in the image. b) Observation pdfs give us the probability of seeing an observation given a hidden state. In this figure we put the fixation location pdfs of all the targets together and superimposed them on the original image and its corresponding bottom-up saliency map. c) In this figure we show the GMM that constitutes the SSHMM of the task of character counting. From the pool of pdfs in the middle figure, only the Gaussians centered on the characters are selected and the rest are removed from the final model.

As we can see, in this model only the *hidden-ness* of the HMMs is emphasized and the *Markov-ness* of the sequences is marginalized to make a maximum likelihood estimator with a mixture distribution.

In other words, the main difference between the classical TD model and the proposed SSHMM is that in the TD model we associate each fixation to the nearest neighbor target, whereas in the SSHMMs a fixation on an object might be a noisy observation of an attentional focus on an adjacent target. In this way, by comparing the results of the TD model and the SSHMM we can examine the importance of using the observation distribution in the HMMs and highlight the significance of considering the covert attention in task inference.

Having defined the generic structure of the SSHMM, we can obtain task-dependent HMMs by training the generic HMM with task-specific eye trajectories by using the expectation maximization-based (EM-based) algorithm of Baum-Welch. In the training, we fix the means of the Gaussians to align with the center of the task-relevant objects and use a uniform distribution for the mixture of Gaussians' prior class probabilities to remove any spatial bias towards any target in stimuli. Moreover, since we have a deterministic state transition (A) and initial state distributions (Π), the only parameters to be trained are the covariances of the observation pdfs (C).

After training the task-dependent SSHMM for each task (λ_θ), we can calculate the likelihood term ($P(\mathbf{O}|\lambda_\theta)$) by applying the parameters of λ_θ to the forward algorithm. In this way, we will be able to make inferences about the tasks given an eye trajectory by substituting the likelihood term into equation (4.1).

4.1.1 Evaluation

In order to perform the evaluation, we compare the results of our model with those of the top-down (TD) model. To build a database of task-dependent eye trajectories, we ran 1080 trials and recorded the eye movements of six subjects while performing a set of pre-defined simple visual-tasks. Six McGill graduate students (three females and three males), aged between 18 and 30, with normal or corrected-to-normal vision volunteered to participate in this experiment and all were naive about the purpose of the experiment.

Five different visual stimuli were generated by a computer and displayed on a 1280×800 pixel screen at a viewing distance of 45 centimetres (1^o of visual angle corresponds to 30 pixels, approximately). Each stimulus was composed of 30 objects each randomly selected from a set of nine objects (horizontal bar, vertical bar and character "A" in red, green and blue colors) that were placed at the nodes of an imaginary 5×6 grid ($6.75^o \times 8.1^o$) superimposed on a black background (see the lower layer of figure 4–1b).

The visual-tasks were counting red bars, green bars, blue bars, horizontal bars, vertical bars or characters (six tasks in total). Each of the tasks was defined so that their corresponding targets can be distinguished from distractors by a single feature. For instance, characters are the only objects responding to slanted orientation filters (see figure2–3) and red objects can be detected by red feature maps alone [60].

At the beginning of each trial, a textual message defining the targets to be sought (e.g., red, green or characters) followed by a fixation mark of size $0.26^o \times 0.26^o$ appeared at the center of the screen. After foveating the fixation mark, the

participant initiated the trial with a key-press. Once the trials were triggered, one of the five stimuli was shown on the display and the eye movements of the subject were recorded while performing the specified visual-task.

An eye tracker (ISCAN RK-726PCI) was used to record the participant's left eye positions at 60 Hz and a chin rest was used to minimize head movements. According to the manufacturer's device description [59], the eye tracker's resolution is approximately 0.3^o over ± 20 degree horizontal and vertical range using the pupil/corneal reflection difference (the actual accuracy is likely to be poorer). An LCD monitor was used for displaying the images and the subjects used both eyes to conduct the experiments.

Each subject did six segments of experiments, each of which consisted of performing the six tasks on five stimuli resulting in 180 trials for each subject (1080 trials in total).[1]

At the beginning of each session, we calibrated the eye tracker by having the participant look at a 20-point calibration grid (4×5) that extended to $8^o \times 10^o$ of the visual angle. The area covered by the calibration grid is stretched beyond the stimuli, which spans $6.75^o \times 8.1^o$ of the visual angle.

[1] In order to reduce the memory effect, we set up the experiments so that each stimulus is displayed once in every 5 trials. Moreover, in each segment, each combination of stimulus-task is executed only once, which results in a repetition in executing a task on an image only in every other 30 trials. Given the nature of the synthetic stimuli that comprise similar objects at random locations, we believe the effect of memory is not significant.

After recording the eye movements, data analysis was carried out on each trial, wherein we removed the blinks, outliers and trials with wrong answers in the verification phase from the data and classified the eye movement data into saccades and fixations using the velocity-threshold identification (I-VT) method [34] with a 50 deg/sec threshold. It is generally agreed upon that visual and cognitive processing occur during fixations and little or no visual processing can be achieved during a saccade [38], therefore, in our analysis we only considered the fixation points.

To define the covariance of the Gaussian distributions, we use a technique called *parameter tieing* [98] 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:

$$COV(B) = \sigma^2 I(N), \tag{4.2}$$

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 reduces from $3K$ to 1. Moreover, a fully diagonal covariance matrix results in a round Gaussian distribution, which is similar to the quasi-circular COG of visual system [33].

For the training phase we use nearest neighbor to find the closest state to each fixation point in the training database and use the sample covariance as the initial estimate of the covariance matrix in the generic HMM. During the training, we use this generic structure to fine tune the covariance value so that the final model results

in a higher observation likelihood. Eventually, it is the outcome of the training phase that is used as the task-specific model of a given task and that is used in the test phase to infer the task of a test trajectory.

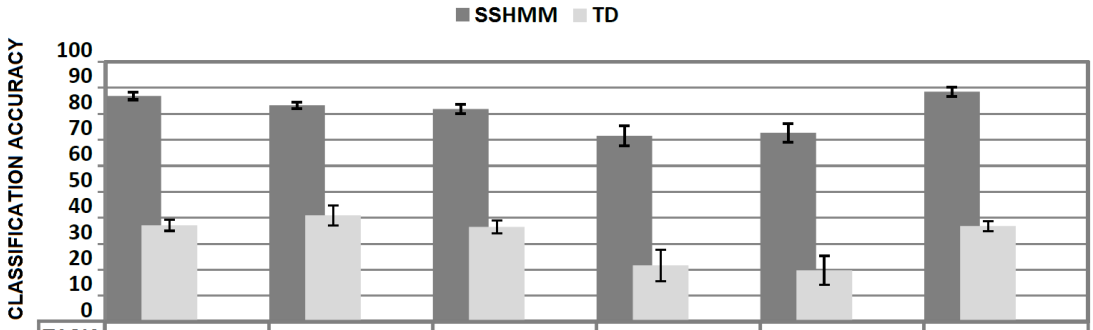
In the TD attention model, the viewer's task emphasizes the conspicuity of the relevant targets and provides us with a task-modulated saliency map. Since top-down methods only model overt attention, we used a nearest neighbor approach to find the closest target to each fixation location to represent its attentional allocation. To train the TD model, we set the target maps by manually selecting the to-be-counted objects in each stimulus (e.g., red objects in the task of counting the number of red objects).

For the optimization of equation (2.5) we used MATLAB optimization toolbox function FMINSEARCH. The training was done in batch mode and a fixed sum equal to one was used for the weight vector (see figure 2–3) to avoid divergence. The resulting weight vectors were used to acquire the task-dependent saliency maps using a saliency toolbox [120] based on Itti and Koch's model [60]. By normalizing the resulting saliencies to one, we could use them as probabilities ($P(\vec{O}_i|\theta)$) and calculate the likelihood term of equation (2.6). The viewer's task, then, was obtained by finding the task which maximizes the posterior probability of equations (2.2).

Since we used equal a-priori probabilities for all tasks, the inference was reduced to a maximum likelihood (ML) estimator. In chapter 5 we will show that with unequal prior probabilities we could use our prior knowledge about the tasks and turn the inference to a maximum a-posteriori (MAP) estimator to obtain better accuracy in the inferences.

4.1.2 Results

Figure 4–2 shows the accuracy of the SSHMM and the TD model in inferring the viewer's task in terms of the number of correctly classified instances of a task. Each bar summarizes the accuracy of its corresponding model by representing the mean (%) along with its standard error of the mean (SEM) in correctly inferring the visual-task. For each bar we ran a 10-fold cross-validation [6] on a dataset of 1080 task specific eye trajectories to train/test the model and compared the performance of the models by drawing their corresponding bars for each visual-task. As can be seen, the SSHMM significantly outperforms the TD method in all six cases, and that is mostly due to relaxing the overtness of attention constraint imposed in the TD models.



	TASK	R	G	B	H	V	C
MEAN	HMM	86.8	83.3	81.8	71.5	72.7	88.5
	TD	37.1	40.9	36.5	21.7	19.8	36.8
SEM	HMM	1.4	1.3	1.8	3.9	3.5	1.7
	TD	2.1	3.8	2.5	6.1	5.6	1.9

Figure 4–2: Comparison of the accuracy of visual-task inference using the single-state HMM (SSHMM) and the Top-Down (TD) models. Each bar demonstrates the recognition rate (%) of inferring simple visual-tasks of counting red (R), green (G), blue (B), horizontal (H) and vertical (V) bars as well as counting the characters (C). The mean value and the standard error of the mean (SEM) are represented by bars and the numerical values are given in the lower table.

4.2 Inferring Complex Tasks in Synthetic Images

In the previous section we successfully applied the idea of using HMMs to infer the ongoing task in simple visual-tasks, where the targets differed from the surrounding distractors by a unique visual feature, such as color, orientation, size or shape, and could be located in a stimulus within a short period of time. Nevertheless, in real-life we usually encounter situations where the target is surrounded by distractors with similar features and can be distinguished from them only by comparing a combination of visual features. In this section we extend our method to a more complicated group of tasks and investigate the applicability of our HMM-based method in task inference in a complex visual-task.

One way of making the tasks more difficult is to make the targets distinguishable by a combination of features rather than a single feature. The difficulty of the task increases the response time and causes several attentional deployments on non-target objects (*off-target FOAs*) in order to examine their task-relevant features and dismiss them as potential target locations. These off-target FOAs on objects are not fully in accordance with the structure of the SSHMM model and will presumably cause attenuation in the accuracy of task inference. In this section we tailor the generic structure of the SSHMM to allow for off-target FOAs in a difficult search paradigm.

To investigate the task inference in a complex set of tasks, we develop an *eye-typing* application, where users can type a character string by directing their gaze to an on-screen keyboard. In this scenario inferring the task is equivalent to determining what word has been eye-typed by observing the eye movements of the subject while performing the task, hence there are a wide range of potential tasks. In order to force visual search, we randomized the location of characters in the keyboard layout in each trial. After each trial we also ran a verification phase where a question is asked about the location of one of the characters in the word the subject has already typed to monitor the correctness of the process (see figure 4–3).

In the experimental results in section 4.2.4 we will show that the SSHMMs are not as effective for inferring task in a complex search task. The bulk of this shortfall is due to the off-target FOAs that take place as a result of the increase in task difficulty. In the new models proposed here we add an extra state to the model that represents the off-target FOAs in the trajectories. This adds to the training burden due to the introduction of extra parameters to the model for the new states' observation

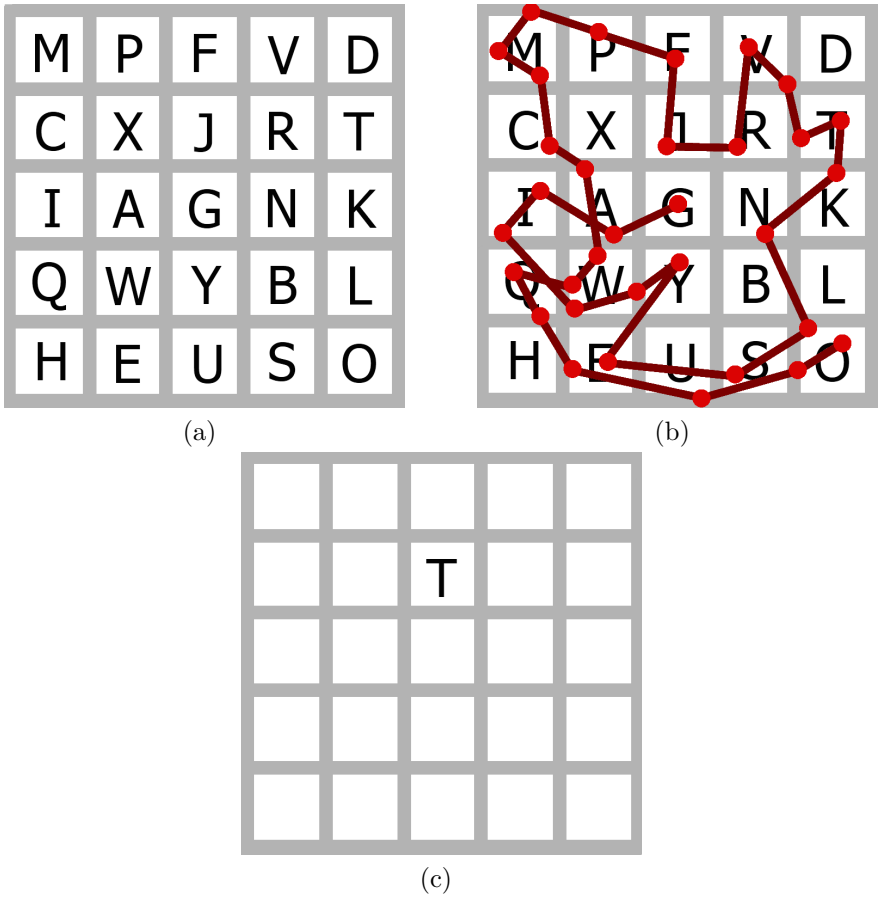


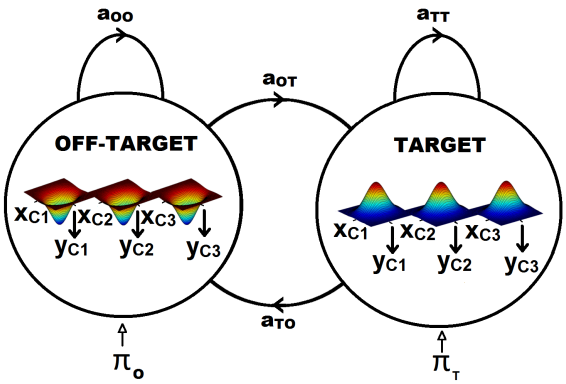
Figure 4–3: Eye-typing application setup. a) The schematic of the on-screen keyboard used to enforce complex visual search tasks. We removed a letter ("Z") in order to have a square layout to reduce directional bias. Also the location of each character is randomized in each trial so that the user has to search for the characters. b) Eye movements of a subject are overlaid on the keyboard layout, on which the trial was executed. The subject searches among the characters to eye-type the word "TWO". The dots indicate the fixations and the connecting lines between the dots show the saccades that brings the new fixation location into the COG. c) After each trial a question is asked of the user about the location of a character that appeared in the word to validate the result.

matrices and state transition probabilities, but the advantage is two-fold: first, we allow for the off-target FOAs both in training and testing, second, the transition matrix becomes stochastic as opposed to the deterministic, self-returning transitions we had in the SSHMM.

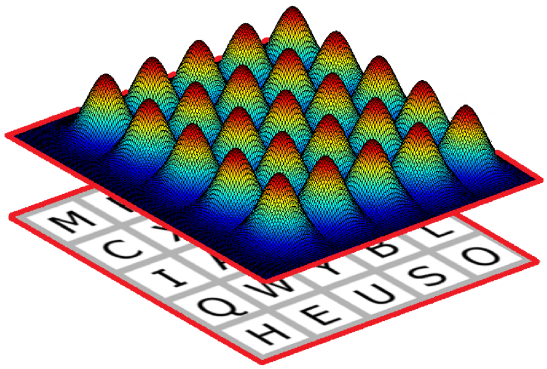
A stochastic transition matrix introduces another source of information to the model by capturing the pattern of transitions between the states. This information can be matched against the test data to see how well the model accords with it. In this way we can consider the dynamics of attention and use pattern matching to locate the targets and predict the ongoing task.

4.2.1 Double-state word HMM (DSWHMM)

In the first attempt to develop a generative model of eye movement trajectories in complex tasks, we add another state to the structure of the SSHMM. Figure 4–4a shows the amendment we made to the model in order to make it capable of dealing with the off-target FOAs. Here in each state we suggest that a mixture of Gaussian distributions demonstrates the possibility of observing different fixation locations. For the *on-target* state, we postulate that fixation locations are random outcomes of a 2D GMM with equal weights that peaks on target character locations.



(a)



(b)

Figure 4–4: Structure of the double-state word HMM (DSWHMM). a) For each state (on-target and off-target) a GMM with equal weights defines the pdf of fixation locations. The transition probabilities $(a_{ij}|i,j \in \{O,T\})$ give us the probability of directing the attention from a target/non-target object to a target/non-target object in the image. For instance, if $a_{TT} >> a_{TO}$ it means that the targets are easy to spot and chances of off-target fixations are low and if $a_{TT} << a_{TO}$, it means targets are hard to spot and finding them involves fixations on distractors. The initial state probabilities $(\Pi_i|i \in \{O,T\})$ give us the probability of starting a search from each of the states. For instance, in the case that targets are hard to spot, the probability of starting the quest from the O state (off-target) is higher. b) In this figure we put the Gaussian pdfs of all the characters together and superimposed them on the original keyboard. It is from this pool of pdfs that we select the task-relevant ones for the target state to define its GMM observation distribution.

Having defined the 2D GMM of the on-target state, we can simply define the pdf of the off-target state by complementing the distribution function of the target state and normalizing it. The rationale behind this is that the probability of fixating a point in the close vicinity of a character is usually higher than other locations when we find it as the target. On the other hand, the farther from the character we fixate, the more likely it is that we are attending a non-target location. In figure 4–4b we superimpose the observation pdfs on the keyboard layout.

Beyond the spatial provisions for the off-target FOAs in the design of the DSWHMM, the new design encapsulates another aspect of the visual attention that scrutinizes attention dynamics from a temporal point of view. The relation between working memory and visual attention is established in neurophysiological and psychophysiological studies [20, 79]. It has been known that attention interacts with working memory and the intensity of this interaction depends on the memory requirement of the ongoing task. In the eye-typing application, after finding each character, the cognitive process that is responsible for driving the visual attention retrieves the next character in the word string from short-term memory to direct attention towards relevant features in the image.

Different variations of this interaction between the attentional system and the short-term memory can be seen in most of the real-world visual search tasks as well. In another example, when a viewer counts the number of objects in a scene, the interaction is invoked after finding each target to increase the count by one in the working memory. In the eye-typing application, as shown in figure 4–3, after each trial a question is asked of the user to verify if the characters are attended

correctly or not. Therefore, when the target character is found, the short-term memory keeps track of the coordinates of the characters in order to correctly respond to the verification question.

Due to different neural circuitry of visual attention and working memory in the brain [43], more interaction between these two functionalities will cause a longer stay on memory demanding targets. A longer stay on the targets, however, does not necessarily mean increasing the time of fixations. In a study Mills et al. [85] investigated the influence of task on temporal characteristics of eye movements during scene perception. They showed that in a visual search, the task affects the spatial parameters of eye movement (e.g., saccade amplitude), rather than the temporal parameters (e.g., latency). This effect is also demonstrated in a study by Castelhano et al. [16], where they examined the influence of task on fixation duration and did not find a significant difference in average fixation duration when memory is involved more heavily in performing the task. However, they showed that for objects that require more thorough processing to be encoded into memory, a strategy is adopted to increase the number, rather than duration, of fixations on them. This conclusion is also supported by an earlier study by Loftus [80] who showed that an increase in the interaction between attention and memory does not affect the duration of fixation, but rather increases the number of fixations made in the region.

The design of the double-state word HMM (DSWHMM) allows us to exploit this phenomenon to spot a pattern in memory interaction and locate the targets that have potentially invoked a more intense interaction between the attention and the working memory. In a difficult search task, where the level of memory involvement changes

significantly during the search, once a target is found, the level of interaction with the memory rises, which in turn causes multiple fixations on the same object. While training a task-dependent model, this pattern is reflected in the transition matrix, which will show a bias in the transitions from the on-target state to itself. This information, along with spatial information embedded in the observation pdf, can be used for locating the task-relevant objects in a scene and inferring the task based on them.

4.2.2 Double-state character HMM (DSCHMM)

In the previous section we showed how introducing a second state to the SSHMM can capture the temporal dynamics of the visual attention and at the same time allow for off-target FOAs. Although in section 4.2.5 we show that DSWHMM is a practical method for task inference in a small-size dictionary, it is not clear what portion of the obtained accuracy is due to the a-priori constraints set by the dictionary and what portion is due to the likelihood provided by the HMM structure. The dictionary of the possible words in the eye typing application, forces a-priori constraints on the tasks by giving zero probabilities for non-existing words and equal probabilities for the existing ones as the a-priori term in the Bayesian inference of equation (4.1).

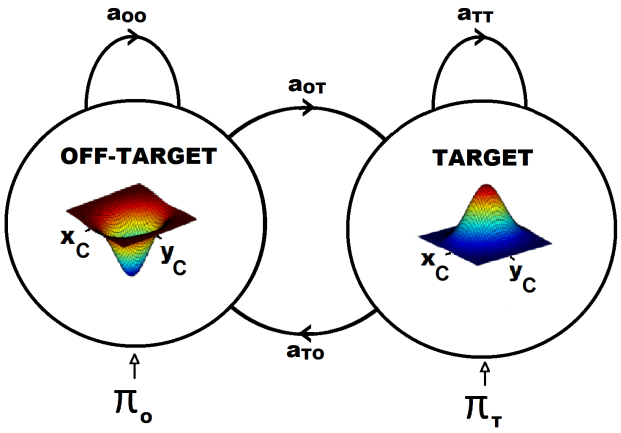
The dictionary size plays an important role in the classification accuracy of the DSWHMMs. On one hand, the word model proposed in the DSWHMM is insensitive to the order of characters and cannot tell anagrams apart[2] . On the other hand, the model performs best when the dictionary contains short words. For longer words,

[2] anagrams are words with the same characters but in different order

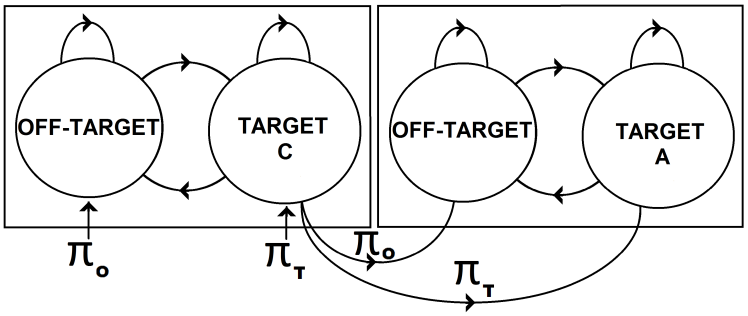
since the target GMMs include more Gaussians, the area covered by the GMMs expands to a larger area on the keyboard, which leads to classifying most of the fixations as on-targets. Thus, by increasing the size of the dictionary, the chances of having long words or anagrams increase, which may cause attenuation in the classification accuracy.

The main reason behind this dependency on the dictionary size is the fact that in the DSWHMMs all of the characters are treated the same and all of them are included in the model by a unique GMM. This fact makes the model insensitive to the order of comprising characters. Moreover, in long words, since the GMM of the on-target state spans a larger area of the keyboard surface, off-target fixations become harder to spot.

The solution we suggest is to assign the double-state HMMs to the characters rather than the whole word. Namely, we suggest to train 25 character models (the letter "Z" is omitted from the keyboard layout) and concatenate the HMMs of the characters that constitute a word to build the word's model. In this way we not only make the model robust to the length of the words by treating each character independent of the proceeding or the following characters, but also we respect the order of the comprising characters while building up the word models. Thus, we expect the model to be more robust to the dictionary size by modeling the sub-word units of characters rather than the whole word itself.



(a)



(b)

Figure 4–5: Structure of the double-state character HMM (DSCHMM). a) shows the general structure of the DSCHMM for character "C". The parameters a_{ij} and Π_i are defined exactly the same as in the DSWHMM the difference being that here we train a model for searching each character, separately. Thus, in the target state we will only have one Gaussian observation pdf around the location of the character in the image. Therefore, in order to build a word model we have to concatenate the constituting characters of the word. b) shows how to concatenate the character models to build up a word model (here for a hypothetical word "CA"). In this model (x_C, y_C) and (x_A, y_A) are the coordinates of characters "C" and "A", respectively. For the transitions between the sub HMMs (for each character), we use the initial state probabilities Π_i, since looking for the next character after finding the proceeding one can be postulated to be roughly similar to start a new search for the new character.

In figure 4–5a we show the general structure of a double-state character HMM (DSCHMM) for the character "C". While the structure is the same as the DSWHMMs, the observation distributions are centered on the characters rather than the whole word. Similar to the DSWHMM, the transition matrix governs the transitions between the off-target and the on-target states and the initial state distribution defines the odds of starting off from each state.

Figure 4–5b shows how to concatenate the character models to build up a word model. If we posit that the character models are independent of their proceeding characters, we can set the transition probability between two arbitrary characters to the corresponding initial state probabilities of the target character. By this assumption, a character's model becomes independent of its place in a word, its proceeding characters and its following characters. In this way all of the words in the training database that include that character can contribute to training the model for that character, which can significantly reduce the minimum size of the training data for obtaining convergence in the training of the model parameters.

Although here we assume that the character models are independent of the following and the proceeding characters in a word, in chapter 5 we will show how we can impose higher order constraints on the models during concatenation of the sub-word models and improve the inference accuracy by relaxing this assumption.

In section 4.2.5 we will show that the proposed character model improves the results of the word model HMMs even with a few number of training trajectories.

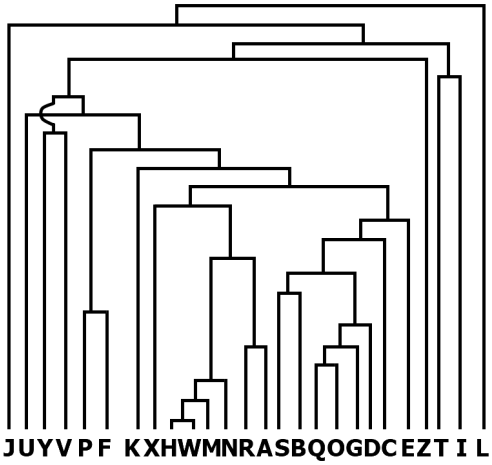
4.2.3 Tri-state HMM (TSHMM)

Modeling sub-word units (i.e., characters), rather than words, allows us to investigate the fixations in more detail. In the DSCHMM structure we classified the attention deployment into the ones on target or the ones on non-target characters. However, we believe that even attended non-target characters carry information about the sought character. Figure 4–6b shows the top nine bins of the histogram of fixations on characters (fixation distribution histogram) when looking for character "W". It can be seen that even off-target fixations show a pattern in the sense that seemingly similar characters tend to draw attention towards themselves more often than the dissimilar ones.

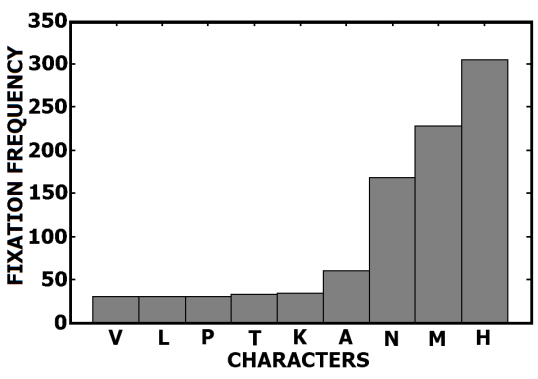
This phenomenon is studied in the psychological literature related to perceptual measurement of image similarity [68]. Particularly, our finding is in accordance with a psychophysical experiment in [42, Figure 1] that classifies uppercase English letters according to their similarity in appearance. In figure 4–6a the result of this classification is shown in form of a hierarchical cluster that classifies the characters into clusters. The lower the connecting line between the clusters, the higher the similarity between them.

In our database of eye movements we saw a similar pattern when analyzing other characters as well. Figure 4–6c shows the average of top nine fixation location histogram bins when looking for different characters. This trend suggests that off-target fixations can also be used as another source of information. Namely, when

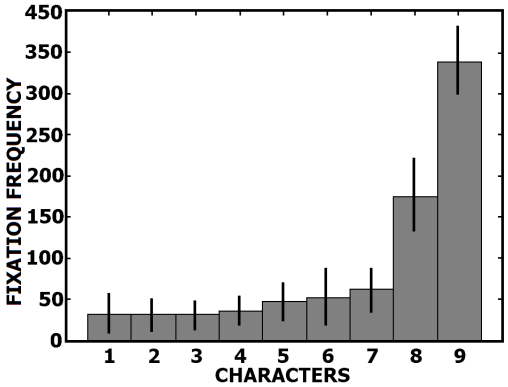
looking for a target, similar characters are more likely to be found among the off-target fixations, which can help us narrow down our choices in the inference process.



(a)



(b)



(c)

Figure 4–6: Spatial distribution of fixations (fixation distribution histogram) while searching for a character. a) Shows the result of the experiment in perceptual measurement of image similarity (based on (Gilmore, Hersh, Caramazza, & Griffin, 1979)). The figure is reproduced with permission from Springer Publishing Company. b) shows the top nine bins of the fixation distribution histogram when looking for character "W". Similar characters tend to draw attention towards themselves, which is in accordance with the psychological experiments. c) shows the average of top 9 fixation location histogram bins, along with their respective standard error of the mean (SEM), when looking for different characters in the keyboard.

Figure 4–7a shows the new structure we propose to be used for task inference in the character recognition application. In this new setup, we split the off-target state to *dissimilar state* (D-state) and *similar state* (S-state) according to the similarity of the attended object to the target character. The tri-state HMM (TSHMM) for character recognition allows us to investigate the fixations in more detail and gives us more information about the sought character. In this model not only the dynamics of on-target FOAs are taken into account, but also the off-target FOAs play an important role in revealing the target character.

Figure 4–7b shows how we build a word model by concatenating the HMMs of the comprising characters. As can be seen, the transitions between the characters are made from the target state and the structure is similar to that of the DSCHMM, otherwise. We heuristically select the top two characters of the fixation distribution histogram of each target to model the GMM of its S-state. The distribution function of the D-state is obtained by complementing the mixture of target Gaussian pdf and GMM of the S-state (all with the same weight). For instance for modeling the character "W", we put the top two similar letters (i.e., "H" and "M") into the S-state and the rest of the letters (i.e., all the letters excluding "W", "H", "M" and "Z") constitute the D-state.

Similar to the DSCHMM, we posit that the character models are independent of their proceeding characters and set the transition probability between two arbitrary characters to the initial state probabilities of the target character. By this assumption, a character's model becomes independent of its place in a word, its proceeding characters and its following characters. In this way all of the words in the training

database that include that character can contribute to training the model for that character, which can significantly reduce the minimum size of the training data for obtaining convergence in the training of the model parameters.

Although here we assume that the character models are independent of the following and the proceeding characters in a word, in chapter 5 we will show how we can impose higher order constraints on the models during concatenation of the sub-word models and improve the inference accuracy by relaxing this assumption.

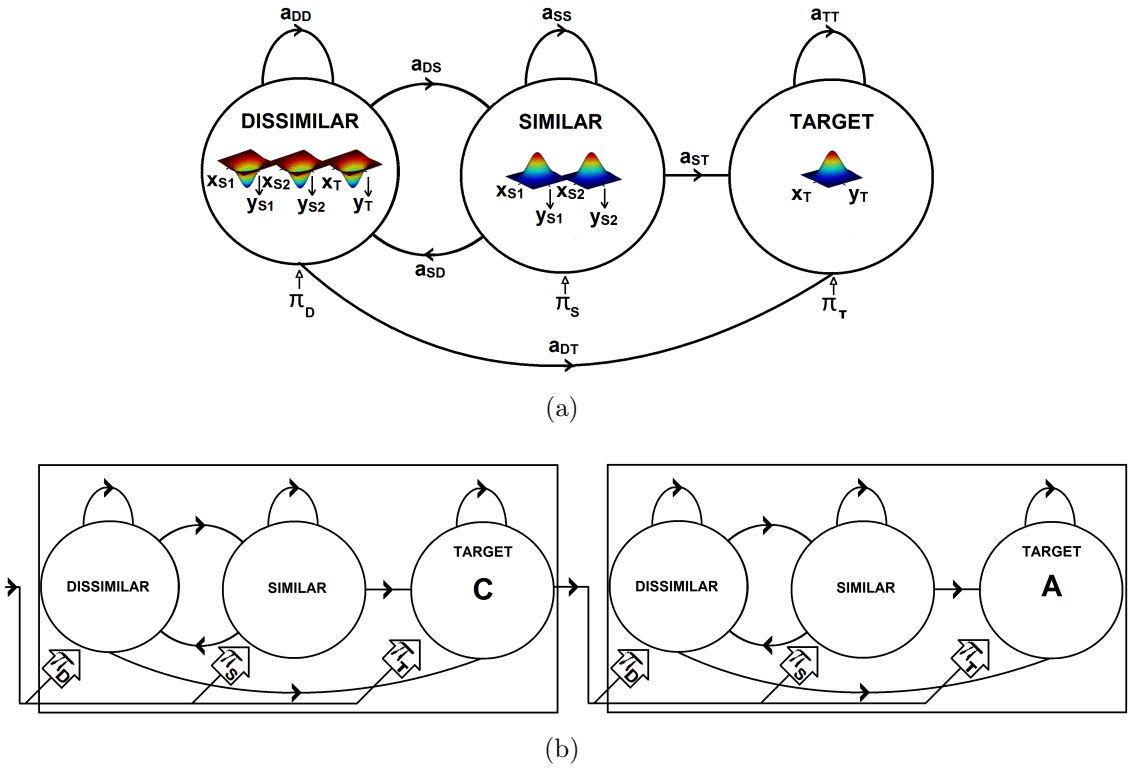


Figure 4–7: The structure of the tri-state HMM (TSHMM) for character recognition. a) The TSHMM for a single character. The mean vector of the S-state's GMM is centered on the top two characters in the fixation distribution histogram of the fixations made during the search for the target character in the training data. Similar to the other models, the transition probabilities a_{ij} governs the transitions between the states and the initial state distributions Πi gives us the odds of starting a search from each state. Also similar to the DSCHMM, the HMMs are trained for each character separately and concatenated in order to make a word model. b) Concatenating the character models to build up the word model. Similar to the DSCHMM, the transitions between the states are governed by the initial state probabilities.

4.2.4 Evaluation

In this section we will evaluate the task inference made by the suggested attention models as well as the classical, top-down attention model. For the sake of comparison, we first build a database of task-dependent eye trajectories and then apply each of the techniques and compare their accuracies in inferring the visual-task.

To build a database of task-dependent eye trajectories, we ran a set of trials and recorded the eye movements of six McGill graduate students (three females and three males), aged between 18 and 30, while eye-typing 26 different 3-character words. The subjects had normal or corrected-to-normal vision and all were naive about the purpose of the experiment. The trials started with a fixation mark of size $0.26^o \times 0.26^o$ appearing at the center of the screen. After foveating the fixation mark, the participant initiated the trial with a key-press. Once a trial was triggered, a textual message at the center of the screen showed the word to be eye-typed. Once the subject indicated his readiness by pressing a key, another fixation mark appeared at the center of the screen followed by an on-screen keyboard similar to the one shown in figure 4–3a. At this phase, subjects *eye-typed* the word by searching for the characters appearing in it as quickly as possible and signaled when they were done by pressing a key (subjects were only told to eye-type the words as quickly as possible and press a key when done).

Each trial was followed by a verification, wherein a question about the location of a randomly selected character in the word, in form of a forced choice paradigm was asked. The selected character appeared as the label of two keys in the keyboard, of which only one corresponded to the original location of the character in the keyboard

layout (see figure 4–3c). The viewer selected one of the keys as the correct location of the character by fixating it and pressing a button. In the data processing phase, we took the result of the question as an indication of whether the subjects had performed the task attentively or not. Once the question was answered, the next word was shown and the trial carried on.

Each keyboard was composed of 25 uppercase English characters randomly located on a 5×5 grid superimposed on a gray background (we removed the letter "Z" in order to have a square layout to reduce directional bias). For the experiment, 26 three-letter words that did not have any repeated characters in them were selected. At the beginning of every experimental session, we calibrated the eye tracker by having the participant look at a 16-point calibration display (4×4) that extended to $10^o \times 10^o$ of visual angle. The area covered by the calibration grid was stretched beyond the stimuli, which spans $6.7^o \times 6.7^o$ of visual angle.

An eye tracker (ISCAN RK-726PCI) was used to record the participant's left eye positions at 60 Hz and a chin rest was used to minimize head movements. According to the manufacturer's device description [59], the eye tracker's resolution is approximately 0.3^o over ± 20 degree horizontal and vertical range using the pupil/corneal reflection difference (the actual accuracy is likely to be poorer). An LCD monitor was used for displaying the images and the subjects used both eyes to conduct the experiments.

After recording the eye movements, data analysis was carried out on each trial, wherein the blinks, outliers, fixations made after finishing the task and trials with wrong answers in the verification phase were removed from the data and the eye

movement data was classified into saccades and fixations using the velocity-threshold identification (I-VT) method [34] with a 50 deg/sec threshold. For the same reason as in the previous experiment, in our analysis we only considered the fixation points and removed the eye positions in between the fixations.

After the preprocessing, we obtained a database of 145 trajectories $\{\mathbf{O}_1, \mathbf{O}_2, ..., \mathbf{O}_{145}\}$, each of them of the form $(\vec{O}_1, \vec{O}_2, ..., \vec{O}_T)$, where $\vec{O}_i$ contains coordinates of the fixation at time step i, in the form of a tuple (x_i, y_i). Each tuple gives us the information regarding the x-coordinate and y-coordinate of the i^{th} fixation, respectively.

In order to train the models in the DSWHMM, DSCHMM and TSHMM, we have to adjust the mean vector of the 2D Gaussians according to the training word so that they align with the center of the respective target character locations. According to [98] a uniform (or random) initial estimation of Π and A is adequate for giving useful re-estimation of these parameters (subject to the stochastic and non-zero value constraints). Thus, in the generic HMM, we set random initial values for the transition and the initial state probabilities and run the Baum-Welch algorithm on the training set to obtain the final model. As in the HMM-based model for simple task inference in section 4.1.2, we use the parameter tieing technique [98] to force a unique, covariance matrix across all the Gaussian distributions in the mixtures in order to reduce the training burden and obtain a uniqie, circular pdf around the targets (similar to the foveal area of the visual system).

The resulting HMM is used to evaluate the likelihood term of equation (2.3) for all the words in the dictionary by using the forward method on the trained model.

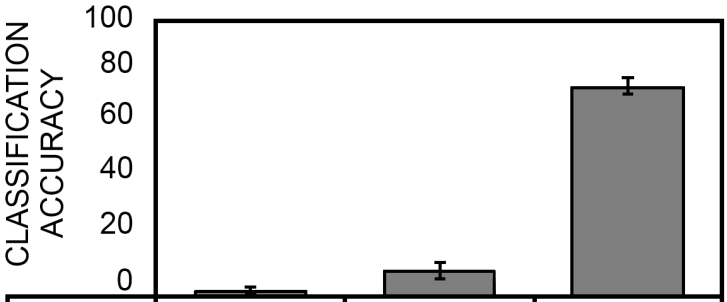
By estimating the likelihood term we can calculate the posterior probability of each word by substituting the likelihood term into equation (4.1). For now we assume a uniform distribution as the a-priori term and postpone using prior knowledge about the tasks to chapter 5.

4.2.5 Results

In the first attempt to infer the task in a difficult search, we compare the results of using the TD, SSHMM and DSWHMM models for task inference and show the results together in figure 4–8. We can see that although the SSHMM performs slightly better than the TD technique, the accuracy decreases significantly compared to the results of task inference in the easy search. The rightmost bar shows the result of task inference when using the DSWHMM. Each bar summarizes the accuracy of its corresponding model by representing the mean (%) along with its standard error of the mean (SEM) in correctly inferring the visual-task. For each bar we ran a 10-fold cross validation on our database of 145 trajectories in order to define the training and test sets and used the same epochs across all the methods. We also used equal probabilities as the word priors, which converts equation (4.1) to a ML estimator.

As can be seen, the DSWHMM significantly outperforms the SSHMM and the TD methods. The parameter estimates after training are shown in Table 4–1.

A very interesting phenomenon seen in the training results is the standard deviation of the Gaussian distributions around the characters (σ_{2D}), which expands to an area of about 3.6^o of the visual angle. This angle appears consistent with previous estimates of the size of the *operational fovea* as the central 3^o of vision [64]. In [13] it is also shown that targets within 4^o of central vision are still perceived at 50%



METHOD	TD	SSHMM	DSWHMM
MEAN	2.2	10.1	76.4
SEM	1.3	2.7	2.8

Figure 4–8: Comparison of task classification accuracy using different models in a difficult visual search. Each bar demonstrates the mean classification rate (%) of correctly recognizing the intended word in the eye-typing application. The mean value and the standard error of the mean (SEM) are represented by bars and the numerical values are given in the table.

of maximal acuity. Although, based on the current evidence we cannot tell whether this finding is a real effect or merely a coincidence, another experiment where the distance between the observer and the screen is altered can help us determine that.

In this experiment we also study the effect of dictionary size on the accuracy of task inference in the difficult task. We test the accuracy of task inference using the DSWHMM, DSCHMM and TSHMM for different dictionary sizes. We created four sets of dictionaries of 26, 52, 104 and 312 English words using the Carnegie Mellon pronouncing dictionary (CMPD) [121]. All dictionaries were built so that they all have the 26 original, three-letter words that were used in the previous experiment. Moreover, all dictionaries included all the words of the smaller ones and the new words were selected randomly from the CMPD with the length varying between

Table 4–1: Parameters of the DSWHMMs after training.

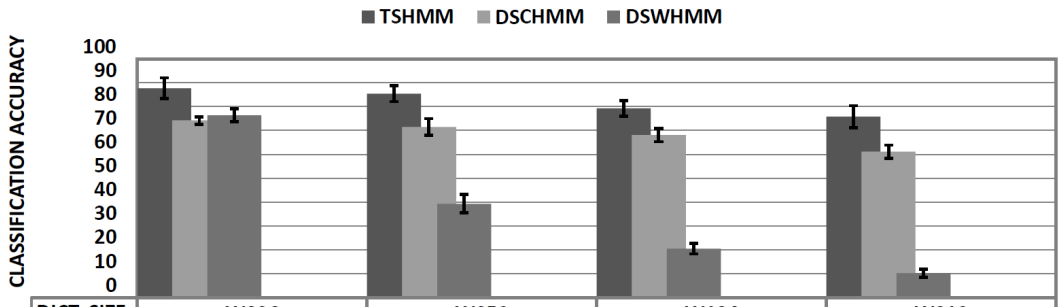
Parameter	Value
a_{OO}	71%
a_{OT}	29%
a_{TT}	67%
a_{TO}	33%
π_O	95%
π_T	5%
σ_{2D}	3.6^o

three to five characters. In all of the HMMs, we bound the covariances to a unique diagonal covariance matrix and obtained the variance, transition matrix, initial state distribution and the mean vectors from the training set. We used 10-fold cross validation to set the training set and the test set on the database of eye movements. For the DSCHMM and TSHMM we built the word models according to the templates shown in figures 4–5b and 4–7a, respectively.

In the TSHMM the characters selected in the S-state represented the top two bins in the fixation distribution histogram of the target character obtained in the training phase. Similar to figure 4–6, the fixation distribution histogram was created by counting the number of fixations on each character (using the nearest neighbor clustering method) when seeking a target. To do so, we manually labeled all 145 eye trajectories and split them into three parts, each of which representing the eye movements of the subject while looking for a character. The distribution function of the D-state is obtained by complementing the mixture of the target Gaussian pdf and the GMM of the S-state (all with the same weight).

Figure 4–9 shows the accuracy of word inference using the DSWHMM, DSCHMM and TSHMM methods ranging over four dictionary sizes. Although the accuracy of

the DSCHMM is slightly less than that of the DSWHMM (74.2% vs. 76.4%), for the 26-word case, it shows less decline as the dictionary size increases. As expected, the TSHMM starts off even better than DSWHMM and stays in the same range over different dictionary sizes. The table below the figure shows the accuracy and the standard error of the mean (SEM) of the corresponding bar.



	DICT. SIZE	W026	W052	W104	W312
MEAN	TSHMM	87.7	85.4	79.3	75.8
	DSCHMM	74.2	71.5	68.1	61.1
	DSWHMM	76.4	39.3	20.5	10.2
SEM	TSHMM	4.4	3.3	3.2	4.6
	DSCHMM	1.6	3.5	2.8	2.8
	DSWHMM	2.8	3.9	2.2	1.6

Figure 4–9: Comparison of task classification accuracy using the TSHMM, DSCHMM and DSWHMM methods in a difficult visual search task. Each bar demonstrates the mean classification rate (%) of correctly recognizing the intended word in the eye-typing application. The mean value and the standard error of the mean (SEM) are represented by bars and their numerical values are given in the table.

4.3 Task Inference in Natural Images

So far we showed that HMMs can serve as a good model for the visual attention process in simple and complex tasks executed on synthetic images. We progressively implemented several HMM models for inferring the task during a visual search that was conducted on synthetic images. However, in all of the models presented so far we

assumed that the targets can be defined in advance and built the models based on the locations of the targets. While this assumption holds for tasks with objective results (such as the number of red objects, horizontal bars, etc.) executed on synthetic images, for more abstract tasks executed on natural images, such as the ones used in the Yarbus [128] and Greene's experiment [46], defining the potential targets of attention is not always straightforward.

Another specification of the previous models that stops us from using them on natural images is that the number of states has to be defined before training. This is only possible in images with a predefined number of targets (like synthetic images used in the previous experiments). However, in natural scenes the targets can appear anywhere in the image and usually no prior information about the location of the targets is available to the model in advance.

Here we will develop a HMM-based attention model that can be applied on natural images. To do so, we first use the K-means clustering technique [67] to locate the potential targets in an image and then use 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 [46].

4.3.1 State positioning of HMMs Using the K-means Clustering

So far we explained how we can use the parameters of a task-dependent HMM (i.e., λ_θ) to infer the underlying task of an eye trajectory (i.e., $\mathbf{O}$), executed on a synthetic image. We used the states to represent the FOA and used the coordinates of the COG as the observations. In the model, each state was composed of a mixture of 2D Gaussian observation pdfs that were centered on state-dependent targets. In

synthetic images, such as the one shown in figure 4–3a, the targets were defined in advance and their coordinates were given to the model as prior knowledge. When executing a task with objective targets on a natural image, we might also define the position of the targets in advance and use that information in training the HMMs. For instance, in figure 3–6 we showed an example of a HMM trained for the task of "counting the number of people in the image". As we saw, the targets learned for this task roughly corresponded to the faces in the image.

However, positioning of the states is not always trivial. When executing tasks, such as the ones used in [46] (e.g., "Memorizing the picture" or "determining the wealth of the people in the picture"), on natural images, pre-defining 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 propose to 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 [67] 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 spots in the image for that task.

Figure 4–10b and 4–10c show the gaze opacity maps of two training sets 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)", respectively. The tasks were executed on the image shown

in figure 4–10a. In these two maps, the areas with higher number of fixations are shown clearly, whereas the areas with no or small number of fixations are masked out by a dark filter and the level of darkness in each segment is inversely proportional to the number of fixations on that segment. As we can see, the areas near the faces get more fixations in the *people* task and the areas around the objects, such as the telephone, tie, pipe and the objects on the desk, are more likely to get fixated in the *wealth* task.

(a) (b) (c)

Figure 4–10: 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)".

By using this simple technique we can get a sense of "conspicuous" locations for different tasks with a linear complexity [127]. The K-means clustering will provide us with K points that indicate the centroid coordinates 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 the K states, each consisting of one Gaussian observation pdf. The initial placement of the 2D Gaussians of the generic HMM on the image,

however, is only an estimate of the eventual positions and might alter during the Baum-Welch training.

4.3.2 Ergodic HMM

Similar to what we did in the synthetic images, we use the proposed HMM-based model to infer the visual-task. The inference is made by calculating the likelihood of an observation, calculated by the forward algorithm, and substituting it into the Bayes rule (equation (4.1)).

In order to obtain the likelihood term ($P(\mathbf{O}|\lambda_\theta)$), we train the parameters of a HMM for each task (λ_θ) by using the training eye movements of the corresponding task. To do so, first we define the structure of the generic HMM and then customize it by training it with the eye movements of the task.

In the generic model definition for the synthetic images, we always had the position information of the targets related to the given tasks. Therefore, in the models proposed in sections 4.2.2 and 4.2.3, we were able to separate the target from non-target objects and assign different states to them. However, in the model definition of the abstract tasks, such as the one used here, we cannot tell in advance whether a fixated object was relevant to the task or it was one of the informative, but non-target objects (such as the ones defined in the *Dissimilar* state in figure 4–7a). Thus, 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, regardless of the current

state of the model [3] . The same solution can be applied to the natural images, where the task does not necessarily specify certain objects in a scene as the targets. For instance, for more abstract tasks, such as the one used in the Yarbus experiment, it is not always straightforward to define some of the objects as the task-relevant ones and put the others in a non-target state.

By defining an ergodic HMM, what we do is to put off the target allocation to the training phase. In the training, the transition matrix will reward transitions to the informative areas and will give lower weight to transitions to non-informative objects in an image. The drawback of defining an ergodic HMM as the generic model, though, is adding to the training burden and increasing the chances of divergence for a small-size training set.

As explained in section 4.3.1, we use K-means clustering to define the initial locations (means) of the observation pdfs and in the generic HMM we assign a state to each of the Gaussian observation pdfs. For each task-image pair, we examine different values for the number of clusters ranging from $K = 2$ to 10 and use this value as the number of clusters in the generic HMM. Then after the training the model that gives the maximum observation likelihood to the training data is selected as the task-dependent model, i.e.:

$$K = \arg \max_{K=2:10} P(\mathbf{O}_{training}|\lambda_{\theta,K}), \tag{4.3}$$

[3] strictly speaking, an ergodic model has the property that every state can be reached from every other state in a finite number of steps. [98]

where $\lambda_{\theta,K}$ is the model that is trained based on a K-state generic HMM. If we use a very small number of clusters, the HMM will not be able to spot all of the task-relevant targets in an image and the resulting HMM will be less task-dependent. On the other hand, if we use a large number of clusters, the training algorithm will diverge due to the increase in the parameter set in the training. Moreover, a large number of states will cause the states to overlap with each other, which voids the effect of having more states to cover more targets in the HMM. We expect that the value of K is highly dependent on the number of task-relevant targets in an image. For instance, for the people task ($\theta = people$) on the image in figure 4–11, 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} on the image.

To define the covariance of the Gaussian distributions, we use the *parameter tieing* technique [98] 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:

$$COV(B) = \sigma^2 I(N), \tag{4.4}$$

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 reduces from $3K$ to 1. Moreover, a fully diagonal covariance matrix results in a round Gaussian distribution, which is similar to the quasi-circular COG of visual system [33].

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

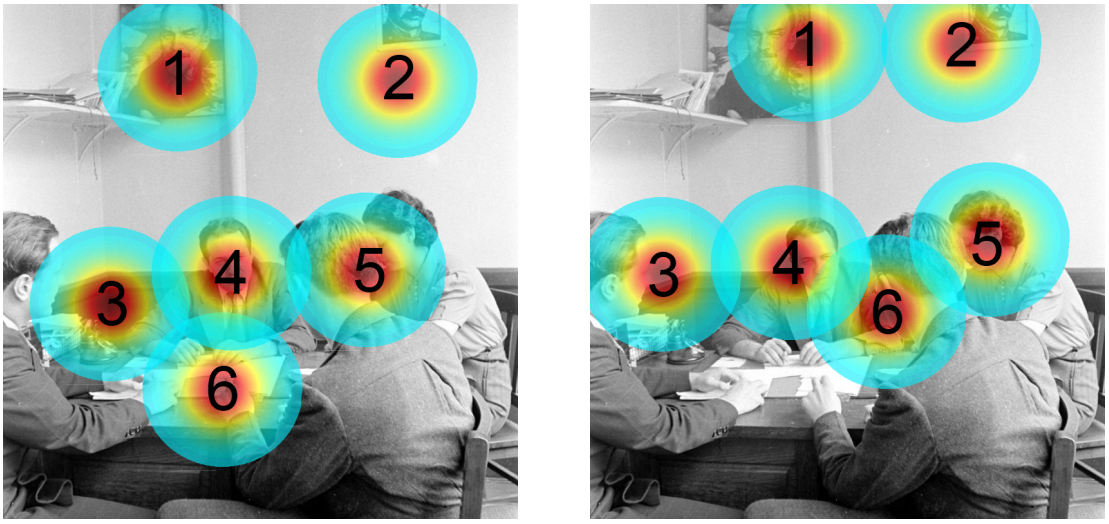
As stated in [98], a uniform distribution assumption suffices as the initial pdf of the *initial state distribution* (Π) and *the state transition probability distribution* (A).

Having defined the structure of the generic HMM, we can obtain task-dependent HMMs by training it with task-specific eye trajectories by using the expectation maximization-based (EM-based) algorithm of Baum-Welch (see section 2.3 for details).

Figure 4–11a shows the generic HMM for the task of *people*, superimposed on the original image. The STD 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 is shown in figure 4–11b. As we can see, the states (pdf means) move around to give rise to the observations in the training set.

4.3.3 Evaluation

In this experiment we concern ourselves with the implementation of a HMM-based inverse Yarbus process, whereby we can infer the visual-task given an eye trajectory of a viewer performing a task on natural images. In order to be able to benchmark our results against those of Greene et al. [46], we used the same database of natural images they used in their experiment. The image set comprises 64 gray-scale photographs taken from the Life photo archive hosted by Google [44], an example of which is shown in figure 4–10a. The date of the images span the years



(a) (b)

Figure 4–11: a) The generic HMM that is used as the generic model for the task of "determining how well the people in the picture know each other (people)". b) The task-dependent HMM after training the generic HMM on the training data.

between 1930 and 1979. In each image there are at least two people, and images do not include photographs of familiar faces or locations.

To build a database of task-dependent eye trajectories, we followed the same procedure as in [46] for the sake of comparison of the results. Overall we ran 1280 trials and recorded the eye movements of five subjects while performing a set of predefined visual-tasks. Five McGill 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 tasks as in Greene's experiment as follows:

- 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 screen of size $53.3 \times 30cm$ at a viewing distance of 45 centimetres (1^o of visual angle corresponds to 28 pixels, approximately). Each image had a resolution of 800×800 and was shown in its original size, which extended to 28×28 degrees of visual angle on a plain black background.

Each subject did four segments of trials during his/her experiment. Each segment consisted of four blocks of 16 images, during which the subject was asked to perform the task stated in an instruction image at the beginning of each block. During each segment, all 64 images were displayed once and subjects had 10 seconds to look at each image. In order to better engage the subjects to the tasks, after each image in the "decade", "people" and "wealth" blocks, a question in form of a forced choice paradigm was presented to the subject. The subjects were asked to select the best answer by clicking one of the five choices. (We used the same routine and questions as in Greene's 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 in all the images. In the end, we obtained five trajectories per task, per image, from which we selected the test and training set using the leave-one-out cross-validation (CV).

An eye tracker (Tobii X120) was used to record the participant's eye positions at 120 Hz. The eye tracker's spatial resolution is approximately 0.2^o and its accuracy is about 0.5^o. The subjects used both eyes to conduct the experiments. The experimental setup is shown in figure 4–12.

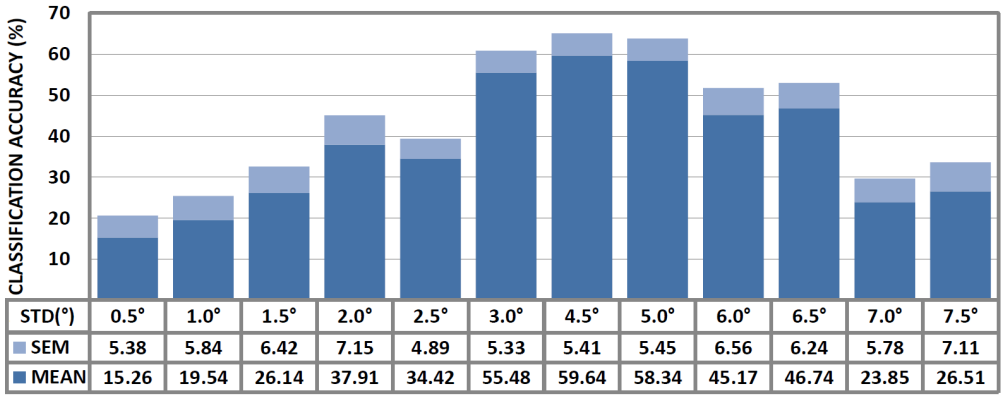
At the beginning of each segment, we calibrated the eye tracker using the built-in, five point, changing diameter, moving dot calibration routine in Tobii Studio software (version 3.2.0, Tobii Technology, Stockholm, Sweden). The calibration grid spanned the whole display.

After recording 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 into saccades and fixations using the built-in velocity-threshold identification (I-VT) method in the Tobii studio software. It is generally agreed upon that visual and cognitive processing occur during fixations and little or no visual processing can be achieved during a saccade [38], therefore, similar to all previous experiments, we only considered the fixation points and removed the eye positions in between the fixations.

Figure 4–12: Experimental setup using the Tobii X120 eye tracker and the LCD display.

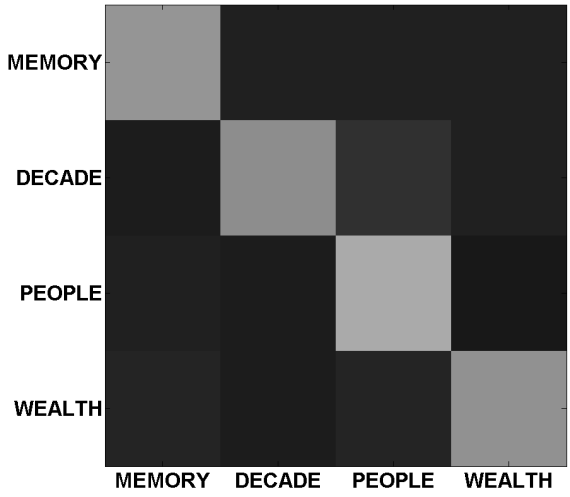
4.3.4 Results

In section 4.3.2 we remarked that the best value for the standard deviation is $\sigma = 4.5^o$. Figure 4–13a shows the accuracy of task classification versus different values of the standard deviation (STD) of the Gaussian observations. As we can see, the maximum accuracy occures when $\sigma = 4.5^o$. 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 (CV) 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.



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)

Figure 4–13: 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 the SEMs. b) Confusion matrix of task inference using the HMM-based model.

As mentioned in the introduction, in the study by Greene et al. [46], their confusion matrix for task inference was at the chance level. Figure 4–13b shows the confusion matrix we obtained using our HMM model. Also the numerical values of the confusion matrix is shown in table 4–2. As we can see, the values are well above the chance level (25 %) (that is obtained in [46]) and the model can infer the visual-task with average accuracy of 59.64 %, as given by averaging the diagonal elements of the confusion matrix.

In section 2.3 we showed that the only difference between the HMMs and the discrete-time Markov chains (DTMC) is that the HMMs allow for covert shift of attention by postulating the COGs to be the outcome of an observation pdf that covers the area around the attentional spots. In order to show the advantage of allowing for covert shift of attention, we use the same database and do the task inference using the DTMCs. To do so, we use the exact same set up as in the HMMs (using K-Means clustering), but rather than setting an observation pdf to each state, we use Euclidean nearest neighbor to select the current state of a fixation. The confusion matrix obtained by using DTMC has an average accuracy of 31.54 % and is shown in table 4–3. 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 4–2: 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 leave-one-out 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**

Table 4–3: 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 leave-one-out 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**

4.4 Discussion

In the beginning of this chapter we applied the proposed attention model based on the theory of Hidden Markov Models (HMMs) to two tasks of simple and complex search. For the simple search task, the application of a single-state HMM (SSHMM) with mixture of Gaussian observation distribution functions gave good results in the task inference. However, the experiment showed that the SSHMMs are not as effective for task inference in more complex tasks, due to the frequent off-target deployment of attention.

Based on the literature related to the effect of task on eye movements, it is known that complex tasks impose patterns on transitions rather than changing the aggregate measures of the eye movements. In the new model, along with the spatial information of 2D Gaussian mixture models (GMMs), the transition pattern information is used to elicit information about the task. By introducing a second state, the double-state word HMMs (DSWHMMs) were able to capture the transition dynamics of eye movement data and use self-returning transitions as a sign of interaction intensity with short-term memory, which in turn were used as an indication of whether the FOA is on target or non-target object.

In another attempt an effort was made to reduce the a-priori constraint set by using a small-size dictionary in the DSWHMM. To do so, we proposed to model the attention cognitive process that drives eye movements for seeking characters as the targets rather than the word in whole. In this way not only we were able to respect the order of characters in a word, but also allowed for longer words in the dictionary. The results showed that modeling character using a double-state character HMM (DSCHMM) increases the consistency of word inference across different dictionary sizes, whereas DSWHMM showed to be sensitive to the dictionary size.

In another model we proposed that even off-target FOAs show a specific pattern given the target. Namely, we found out that in a visual search for a character, attention tends to land on characters similar to the target to narrow down the potential locations of the target. Thus, we proposed to split the off-target state to two separate states representing the FOAs on similar and dissimilar characters to the target. The results showed not only that the tri-state HMM (TSHMM) is robust to the size of

the dictionary, but also that the additional information elicited from the off-target fixations helps us better infer the task.

In general, we used the SSHMMs to model the covert attention in the easy visual search, where the fixations are mainly on the targets. In the DSHMMs we showed how we can capture the temporal dynamics of human attention and at the same time allow for off-target deployment of attention, while maintaining the support for the covert attention, inherited from the SSHMMs, by introducing a second state to the HMM structure. Two variations of the DSHMM were used to infer the task in whole (DSWHMM) or in part (DSCHMM). In the TSHMMs we took advantage of the information hidden in the off-target FOAs by introducing a third state to the model. Overall, the results supported the idea of attention modeling using the HMMs and suggested a solid probabilistic framework for task inference in synthetic images.

While the results presented in section 4.1 were very promising, further investigations was necessary to extend the idea to natural scenes and more realistic situations like those of Yarbus. In the rest of the chapter we concerned ourselves with inferring the tasks with no objective (or not straightforward) targets executed in natural images. The main challenge in abstract task inference in natural images is that the targets are not known in advance.

In section 4.3 we presented a probabilistic framework for task inference in natural images. This work was particularly motivated after previously encountered difficulties in developing a reliable model for the inverse Yarbus process. As a reference we based our experiment on a setup used in a recent study by Greene et al.

[46, 45], who concluded that the Yarbus finding is evocative and visual-task cannot be inferred using the eye movements.

In order to benchmark our results against theirs, we used the same database of natural images they used in their experiment. We also used the same experimental protocol as in their experiment to limit the effect of other factors on the results.

The method we proposed was to first estimate the attention-demanding spots in an image according to the task and then use the HMMs to model the task-dependent eye movements for the given task and image. To find the attention demanding locations in an image given a task, we used K-means clustering technique on the ensemble of the training set and used the centroids as the potential targets. Due to the lack of knowledge about task relevance of these potential targets, though, we cannot split the targets in our models as was done in the other models. Thus, we used an ergodic structure for the 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 equation (4.1) and make inference about the task.

The results show a significant improvement over the results obtained by Greene et al. [46].

In our view there are several reasons behind our improved results as compared to those in [45]. In Green's experiment only aggregate features of eye movements, such as the number of fixations, duration of fixations, etc., were used to classify the trajectories. These features, however, have been shown previously (e.g., [15])

to be unreliable in task inference and therefore cannot be used to regularize the ill-posedness of the problem.

Another reason behind the failure of the aggregate-based method in inferring the task is that no information about the image is used in classifications. This is while there is a well-studied relation between these two and the image context is proven to have a major effect on eye movements [116].

In analyzing the eye movement behavior, the temporal order of fixations is usually omitted from many models of eye movement analysis including the one by Greene et al. [46], where the temporal order of fixations is not used in the summary statistics of the eye movements. However, it is becoming increasingly more obvious that 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 raised since the early studies by Yarbus [128] and Buswell [12]. In the salience-based models of attention the temporal order of fixations is not usually considered in training the models [7, Figure 7]. From the statistical point of view, these models postulate a *naïve Bayes assumption* in evaluating the likelihood probability of equation (4.1), which assumes independence between consecutive fixation locations. In contrast to these models, consecutive fixations have shown to be highly dependent on each other. In a study by Hacisalihzade et al. [47] they recorded the eye movements of observers during the task of recognizing an object and showed that the fixations follow a somewhat Markov process. They showed that the eyes visit the features of an object cyclically, following a regular scanpaths rather than crisscrossing it at random. Elhelw et al.

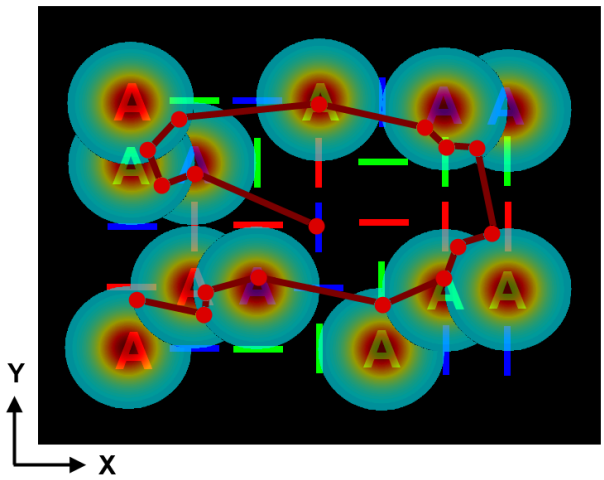
[29] used a first-order, discrete-time, discrete-state-space Markov chain to model eye movement dynamics. Stark and Ellis [110] also came up with a Markov process as a general model of fixation placement during the task of reading. Pieters et al. [95] also observed a similar pattern in the scanpaths of the observers while looking at printed advertisements.

The temporal order of the fixations plays an important role in decoding the pattern of eye movements in the HMMs. In fact, the transition matrix of the HMMs (A) adjusts its elements according to the order of fixations the subjects 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 an observation of that task. In [46], however, the temporal order of fixations is not used in the summary statistics of the eye movements. Eliminating the temporal order information from the feature set prevents the classifiers from improving the results based on that information and probably is one of the reasons the classification based on the summary statistics leads to a poor result.

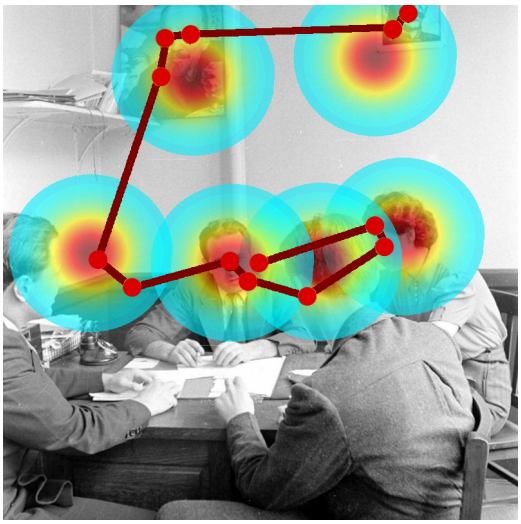
Both temporal and spatial information could be extracted in the original Yarbus experiment [128, figures 107 to 124], as well as in the experiment by Greene et al. [46, figure 3]. However, in order to replicate the Greene et al. experiment using their feature set, we removed the temporal information 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. The performance was still above chance, however. Moreover, the HMMs completely fails in inferring the task when spatial information from the eye trajectories is removed. Thus, we can hypothesize that spatial and temporal information is crucial to solving the inverse Yarbus problem, and the lack of such information may be the reason that the Greene et al. approach did not work.

An important point to consider is that the purpose of this chapter was to infer the visual-task based on the recordings of the fixations made on an image and not to infer the FOA. This can specifically be noticed in the first experiment, where all the fixations are assumed to be on target, while in a real word situation, off-target fixations, even for simple tasks, are inevitable. However, in the results of this experiment, the hidden states of the model intuitively become closer to our expectation of the FOA. For instance, in the second experiment the states are dedicated to the on-target and off-target fixations, which can be safely assumed to happen in the FOA trajectories during a task execution. In the third experiment we further break down the off-target fixations to the ones on similar and dissimilar targets, which is shown to be in line with the experimental results of [42] that is shown in figure 4–6. Thus, the MAP estimation of the hidden state sequence given an eye trajectory (that can be obtained using the Viterbi method) can be used as an estimate of the possible targets of the visual attention for a given task (figure 4–14).



(a)



(b)

Figure 4–14: a) The hidden states of a task-dependent HMM visited during an eye trajectory of a subject executing the task of "counting the number or characters" on a synthetic image. b) The hidden states of a task-dependent HMM visited during an eye trajectory of a subject executing the task of "counting the number or people" on a natural image.

The increase in the accuracy of the HMMs compared to the DTMCs not only shows the advantage of allowing for the decoupling in the model, but also implies instances of covert shifts of attention in the eye movement data. The indication of the COG deviation from the locus of attention is an important by-product of the model, since it is not easy to demonstrate 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. [90] this COG-FOA decoupling is implied from observers' unawareness of changes in an image in 40% of the time, even though they were directly fixating the change location. Measuring the reaction time (RT) is another indirect implication of covert attention that is used in psychophysical studies [57, 73, 21, 107], where a decline in the reaction time to a non-fixated stimulus is associated to the covert attention.

Although the comparison between the results of the HMMs and the DTMCs implies the existence of off-stimulus attention, we have to note that off-stimulus fixations do not necessarily mean that the FOA is away from the fixation (i.e. covert). For instance, in a phenomenon known as *the center-of-gravity* (also known as the *global effect*) [130, 53, 87], the stimulus is actually the collection of features, and the location of the stimulus is the centre-of-mass of this collection. Hence the FOA is in this case overt. If we were to accurately compare the hidden states of our models with the focus of attention, we would have to use methods such as *attentional-probing*, detecting microsaccades [48] or fMRI recording [124], in order to locate the FOA

and measure the correlation between their estimation and the centroids of the HMM states.

CHAPTER 5
Information Fusion in Visual-Task Inference

In the previous chapters we developed a model based on the theory of the HMMs to infer the visual-task for simple and complex tasks, executed on synthetic or natural stimuli. In the task inference, we used the HMMs to calculate the likelihood term of $P(\mathbf{O}|\lambda_\theta)$ and substituted it into equation (4.1) compute the posterior. We then determined the task that maximized the posterior probability. In the Bayesian inference, however, we used equal probabilities for all tasks, which reduced the *maximum a-posteriori* (MAP) inference to a *maximum likelihood* (ML) inference, i.e.:

$$\arg\max_{\theta\in\Theta} P(\lambda_\theta|\mathbf{O}) = \arg\max_{\theta\in\Theta} P(\mathbf{O}|\lambda_\theta). \tag{5.1}$$

In real life, tasks do not happen according to a uniform distribution (as assumed in the ML estimator) and have different a-priori probabilities (as assumed in the MAP estimator). In the first section of this chapter we show how to fuse different sources of information and improve the results by applying the higher order constraints to the Bayesian inference in terms of the prior probability, $P(\lambda_\theta)$.

In this section we use the tri-state HMM (TSHMM) proposed in section 4.2.3 as the baseline and see the effect of prior knowledge about the tasks on its accuracy. The same technique can be used in task inference in other HMM-based inference models, since in the Bayesian inference the a-priori knowledge is marginalized from the likelihood term and can use any model to evaluate the likelihood. Moreover, in

this section we use the eye-typing application for the task inference. In the eye-typing application the tasks correspond to the dictionary tokens (words or characters) and inferring the tasks corresponds to classifying or recognizing the eye-typed token.

Although Bayesian inference provides a nice framework for fusing the prior knowledge about the task and inferring the visual-task from the eye movements, there is a limitation that stops us from using the classical Bayesian inference for larger number of tasks in the task space. So far the inferences we made were in fact a classification, where we selected the task from a pool of tasks that was most consistent with the observation vector. For instance, in the eye-typing application, we had a dictionary of possible words, from which we selected the best word (task) that gave rise to the observation vector of the fixations. To this end, we had to build the models for each of the words in the dictionary in advance and evaluate the likelihood of a given observation for each of the models. Then by using the Bayesian inference we picked the task that gave the maximum a-posterior probability as the eye-typed word.

This method proved to be able to infer the task for a small-size dictionary of possible words. However, if the dictionary of the possible words is very large, building the model for each of the character combination takes a lot of processing and memory. Moreover, for a free vocabulary dictionary that has unlimited number of possible words, building the models in advance is not an option and we have to recognize the words rather than classifying them.

In order to address this issue and *recognizing* the task rather than *classifying* it, in the second section of the chapter we propose a variation of the classical Bayesian

inference that uses the HMM models within a simple conceptual model of recognition called *the word lexicon* that incorporates the sub-task HMMs in a transition network structure. We will use this model to incorporate the prior information about the tasks into the recognition and build a complete model of recognizing the ongoing task for the eye-typing application. In the word lexicon framework, rather than building the models in advance, we can incorporate the models on the fly. Besides dynamically building the models, using the lexicon allows us to naturally incorporate the a-prior information about the tasks into the recognition in terms of rewards and penalties imposed on the transitions.

So far we used the forward algorithm to evaluate the likelihood of an observation given different models. In the Bayesian inference we used the result of the forward algorithm as the likelihood term and evaluated the a-posterior probability of different tasks given the observation. A variation of the forward algorithm called *token passing* will be introduced in section 5.2.1 to make inferences in the lexicon model. Using the token passing technique we can infer the task by recognizing the best path throughout the lexicon that better matches the observation.

5.1 Information Fusion in The Bayesian Inference

Although in section 4.2.3 we showed that the structure of the TSHMMs is more compatible with the nature of the complex task of eye-typing on a soft keyboard resulting in a better classification accuracy in the task inference, there are other sources of information that could be applied to the inference to improve the performance of the model. The prior probability distribution of tasks is a source of information that we use on a daily basis to make inferences about our observations. For instance,

in a reading task when we encounter the character "Q", our brain gives rise to the character "U" as the following character, since in common English words, most of the time "Q" is followed by "U". Therefore, in our eye-typing application, recognizing "Q" as the current character could a-priori increase the chances of recognizing "U" as the next eye-typed character.

A similar technique is used in the speech processing literature to improve the result of a recognizer by applying high-level constraints to the sequences of speech unit (i.e., words, characters, phoneme, etc.) [98]. The constraint is imposed to the decision making engine in the form of a lexicon dictionary, called the *language model* (LM), that provides us with prior probabilities of seeing different speech units given the current one. For instance, if the recognition units are the words, the LM governs the transitions between the words according to the sentences that exist in the dictionary and gives different weights to the transitions depending on the combination of the words. In short, the units of speech themselves are modeled by the HMMs and then concatenated according to the LM to build up the general model that is used for recognition of the speech units.

In this section we apply the same technique to our problem of task inference and fuse the high-level constraints to the Bayesian inference framework that we used for the task inference. To do so, we define the task as typing a single character and define the inference as finding the character that is eye-typed given the observation vector of the eye fixations made during the execution of the task. Thus, a word can be modeled by a string of characters, each of which is represented by the HMM-based model (e.g., TSHMM) that was generated for each character, and can be represented

as follows:

$$\Lambda_\theta = \{\lambda_{\theta_1}, \lambda_{\theta_1}, \ldots, \lambda_{\theta_n}\}, \tag{5.2}$$

In this representation n is the number of characters in the word and λ_{θ_i} is the character that appears in the i^{th} place of the word. Each character (θ_i) is denoted by its HMM model, λ_{θ_i}, to indicate the parameters of the attention model that are trained for that character (training the parameters of λ_{θ_i} was explained in chapter 4).

In the Bayesian inference we defined the recognition of the i^{th} character (λ_{θ_i}) given the observation vector of the fixations ($\mathbf{O}$) as the solution to the following optimization problem:

$$\hat{\lambda}_{\theta_i} = \arg\max_{\lambda_{\theta_i}} \; P(\lambda_{\theta_i}|\mathbf{O}, \Lambda_\theta) = \arg\max_{\lambda_{\theta_i}} P(\mathbf{O}|\lambda_{\theta_i}) P(\lambda_{\theta_i}|\Lambda_\theta) \tag{5.3}$$

So far we used a uniform distribution for the a-priori term of $P(\lambda_{\theta_i}|\Lambda_\theta)$, which reduced the inference to a maximum likelihood estimation of:

$$\hat{\lambda}_{\theta_i} = \arg\max_{\lambda_{\theta_i}} P(\mathbf{O}|\lambda_{\theta_i}). \tag{5.4}$$

However, prior knowledge about the task, such as the one imposed by the language models, can improve the results by giving rise to the tasks with maximum a-posterior probability, making the inference a maximum a-posteriori (MAP) estimation of the task.

Since in our application we are dealing with common English words, we use a similar technique used in speech recognition [98] to apply higher order constraints on the recognizer. In our application, the recognition units are the TSHMMs that

are trained for each characters and the higher level process that concatenates them to build up each word in the dictionary is modeled by the LM. The TSHMMs and the LM are denoted in equation (5.3) by the terms λ_{θ_i} and $P(\lambda_{\theta_i}|\Lambda_\theta)$, respectively.

Several models exist for defining the LM, the main feature of them being the order of dependency between the models of task units (characters in our case). As a common model, in the LMs we assume a Markov chain that governs the transitions between the units of recognition and depending on the order of the Markov chain used in the LM, the model is called an n-gram LM (n defines the order of the Markov chain). Choosing Markov chains as the underlying model of the LM is particularly in full accordance with the structure of the character units, which is based on the HMMs.

Here we use a unigram LM to model the transitions between the characters ($n = 1$), but the same principal applies for higher order LMs, as well. In a unigram LM, each character only depends on the previous character, i.e.,:

$$P(\lambda_{\theta_i}|\Lambda_\theta) = P(\lambda_{\theta_i}|\lambda_{\theta_{i-1}}). \tag{5.5}$$

Thus, By using the unigram LM as the prior task probability, equation (5.3) becomes:

$$\hat{\lambda}_{\theta_i} = \arg\max_{\lambda_{\theta_i}} P(\mathbf{O}|\lambda_{\theta_i}) P(\lambda_{\theta_i}|\lambda_{\theta_{i-1}}) \tag{5.6}$$

In order to build the language model (LM), we need to train the LM by assuming the first order Markov chain as the underlying process of the character sequences that appear in the dictionary words. The training is done by counting the number of each pair of transitions in the dictionary and therefore the probabilities are based on

frequencies and counts of each pair, i.e.,:

$$P(\lambda_{\theta_i}|\lambda_{\theta_{i-1}}) = \frac{c(\lambda_{\theta_{i-1}}, \lambda_{\theta_i})}{c(\lambda_{\theta_{i-1}})}, \tag{5.7}$$

where $c()$ denotes the count function. Eventually the language model will give us the probability of $P(\lambda_{\theta_i}|\lambda_{\theta_j})$ for each pair of characters $(\lambda_{\theta_i}, \lambda_{\theta_j})$.

In section 4.2.3 we showed how we can train the TSHMMs for each character. Therefore, by training the LM we have a complete set of probability distribution functions to evaluate the a-posterior probability of equation (5.6).

In the experiments we will show how prior task probabilities can improve the result of task inference in character recognition. The significant improvement we get by applying the LM shows the importance of the task priors and their effect in disambiguating the likelihood of an observation given different models. In this context, the LM works as a source of *experience*, whereby we can apply our knowledge acquired through history into the inferences we make. This effect is comparable to how our brain improves its inferences by applying the experience it gains through previous practices.

5.1.1 Evaluation

In this experiment we show the effect of using a-priori information on the Bayesian inference. The goal is to infer what character was being eye-typed given an eye trajectory. In other words, the tasks are composed of the characters and the observations are the eye trajectories of the subjects while performing the tasks. To do so, we use the same database of eye movements that was used in chapter 4. The

database of task-dependent eye trajectories is composed of the recordings of the eye movements of six subjects while eye-typing 26 different 3-character words.

In order to extract the character trajectories, we manually split each trajectory into three sub-trajectories, each dedicated to find one of the three comprising characters of a word. In chapter 4 we showed that in the TSHMM models the fixations denoting the final fixations of a character quest are usually located around the target character and can be modeled by a Gaussian observation pdf (see figure 4–7). Therefore, in order to split the word trajectory, we spotted consecutive fixations around each of the three characters (target state) and manually split the trajectory at those fixations. Figure 5–1 shows a sample eye trajectory of a subject while typing the word "TWO". After spotting the consecutive fixations on each of the characters "T", "W" and "O", we split the word trajectory into three character trajectories. In this figure, the fixations related to characters "T", "W" and "O" are colored in red, green and blue, respectively.

After the preprocessing, we obtained a database of 435 character trajectories each of the form $(O_1, \ldots, O_m)$, containing the observation sequences of coordinates of fixations while performing the eye-typing, where $O_i = (x_i, y_i)$ represents x-coordinate and y-coordinate of the i^{th} fixation, respectively. For the character HMMs, we used the TSHMM models introduced in section 4.2.3 on the training data to build a character model for each of the characters on the keyboard. To separate between the training and testing data, we used a 10-fold cross validation on the whole database of eye trajectories.

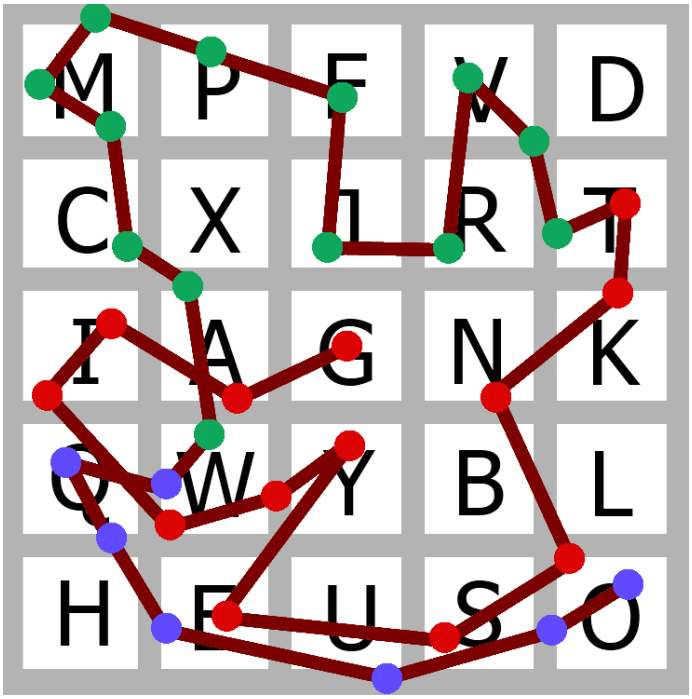


Figure 5–1: A Sample eye trajectory of a subject while typing the word "TWO". After spotting the target states for each of the comprising characters, we split the trajectory into three sub-trajectories, each of which denoting the eye movements while looking for the respective character. In this figure we colored the fixations dedicated to the characters "T", "W" and "O" in red, green and blue, respectively.

In order to train the language model, we created four sets of dictionaries of 26, 52, 104 and 312 English words using the Carnegie Mellon pronouncing dictionary (CMPD) [121]. All dictionaries were built so that they all included all the words of the smaller dictionaries (the same dictionaries used in chapter 4 were used here). The words were selected randomly from the CMPD and the words length varied between three to five characters. The language model was also created using equation (5.3) on the words in each dictionary. To that end, we used the CMU-Cambridge toolkit

[18] that uses the same technique as in equation (5.3) to train four LMs given each of the dictionaries.

5.1.2 Results

Figure 5–2 shows the accuracy of the character inference using the Bayesian inference with a-priori knowledge (+LM) and without a-priori knowledge (-LM). The a-priori knowledge is applied to the Bayesian inference in form of a LM trained over four dictionaries of 26, 52, 104 and 312 English words. In both +LM and -LM models we used the same TSHMM models trained on the same training folds of the eye movement trajectories to calculate the likelihood term of the Bayesian formulation.

As expected, the +LM performs better than -LM due to the fusion of information provided by the +LM. Particularly, for smaller size dictionaries, where the LM can better predict the characters, the improvement is much higher than the -LM baseline. As the size of the LM grows, the accuracy of the +LM drops, which is due to the ambiguity brought to the LM by including out-of-dictionary words in the LM training. That said, since all of the dictionaries include the core 26 words that are used in the eye-typing experiment, the +LM consistently performs better than the -LM in all dictionary sizes. For the small-size dictionaries, the entropy of the LM decreases, which results in higher probabilities for the character combinations in the dictionary and lower probabilities for the out-of-dictionary character combinations. The lower entropy of the LM, highlights the a-priori effect in the Bayesian inference with respect to the likelihood term. As a result we can see higher accuracies in

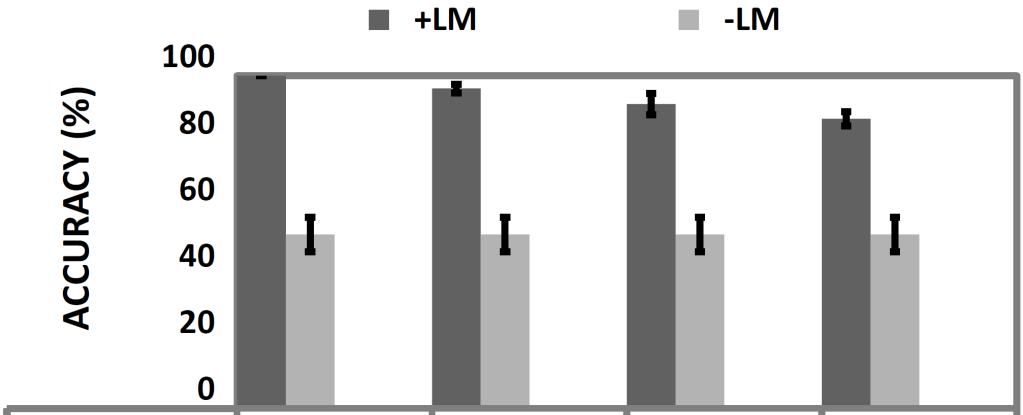
smaller dictionaries, so much so that the 26-word dictionary correctly classifies the characters in all cases.

Figure 5–2 shows the results of information fusion in the Bayesian inference for the character classification task. Each column is dedicated to each of the four dictionaries on which the corresponding LM is trained. The table below the figure shows the average accuracy and the standard error of the mean (SEM) of the task inference by each of the +LM and -LM methods. For each bar we ran a 10-fold cross validation on our database of 435 trajectories in order to define the training and test sets and used the same epochs across both methods. As we can see, the +LM outperforms the -LM in all four dictionary sizes by bringing new source of information into the inference.

5.2 Information Fusion Using the Lexicon

In the previous section we showed how the prior knowledge about the task can be fused into the task inference in the Bayesian inference framework. We represented the prior probabilities about the character being recognized in the form of a LM, which is a transition network that gives different weights to the characters, given the previous one. In the classical Bayesian inference, though, we needed to build the task models for each of the tasks in the dictionary in advance and select the one that gives the best a-posterior probability for a given observation. While this method works fine for a small-sized dictionary of possible tasks, we need an alternative approach for inferring the task in a large-sized or unlimited dictionary of tasks.

In this section we will see how the LM can be merged with the HMM units of character and build a full-scale unit of recognition that incorporates the likelihood



DICT. SIZE	W026	W052	W104	W312
+LM (MEAN)	100.0	96.2	91.5	87.1
-LM (MEAN)	52.3	52.3	52.3	52.3
+LM (SEM)	0.0	1.3	3.2	2.1
-LM (SEM)	5.2	5.2	5.2	5.2

Figure 5–2: Results of information fusion in the Bayesian inference for the character classification task. Each column shows the results of character classification with an LM trained over a specific number of words in the LM training dictionary. The dictionary size varies from 26 to 312 words and the average classification accuracy of character classification of the Bayesian inference with a-priori knowledge (+LM) and without a-priori knowledge (-LM) are shown next to their standard error of the mean (SEM).

and prior terms into one unique model of attention that can be used for making inferences about the visual-task. To do so, we merge the TSHMM character models with the LM to build up a unique attention model for all of the tasks in the dictionary, called *the lexicon* [63]. In the Bayesian approach we created a model for each of the words in the dictionary and picked the one that better matched the pattern of the observation vector. In the word lexicon, though, we create a unique model for all of the words in the dictionary (and even the ones that are not in the dictionary),

where the transitions between the characters are governed by the LM. In order to infer the task, a new method called *the token passing* is introduced in this section that can traverse the word lexicon and give us the task that best matches the observation vector. The new model has many advantages, one of the most important of which being the elimination of pre-building the models for each of the words in the dictionary.

The word lexicon is composed of the character TSHMMs put together in a parallel state machine. The connections between the characters are governed through a state specified to the LM. The complete structure of the word lexicon is shown in figure 5–3. In order to generate a sample observation vector, first we select a character according to the LM and start from a state according to the initial state probabilities of that character's HMM. By following the transition probabilities we can choose the next states at each time step and generate observations according to the observation probabilities. When getting to the final state of the character (target state), it is the language model that suggests which character, by what probability, can follow the current one. The LM weights to the combination of characters are applied via a hypothetical state (that neither generates an observation nor represents a time-step) that applies different probabilities for the next character according to the current one (unigram LM).

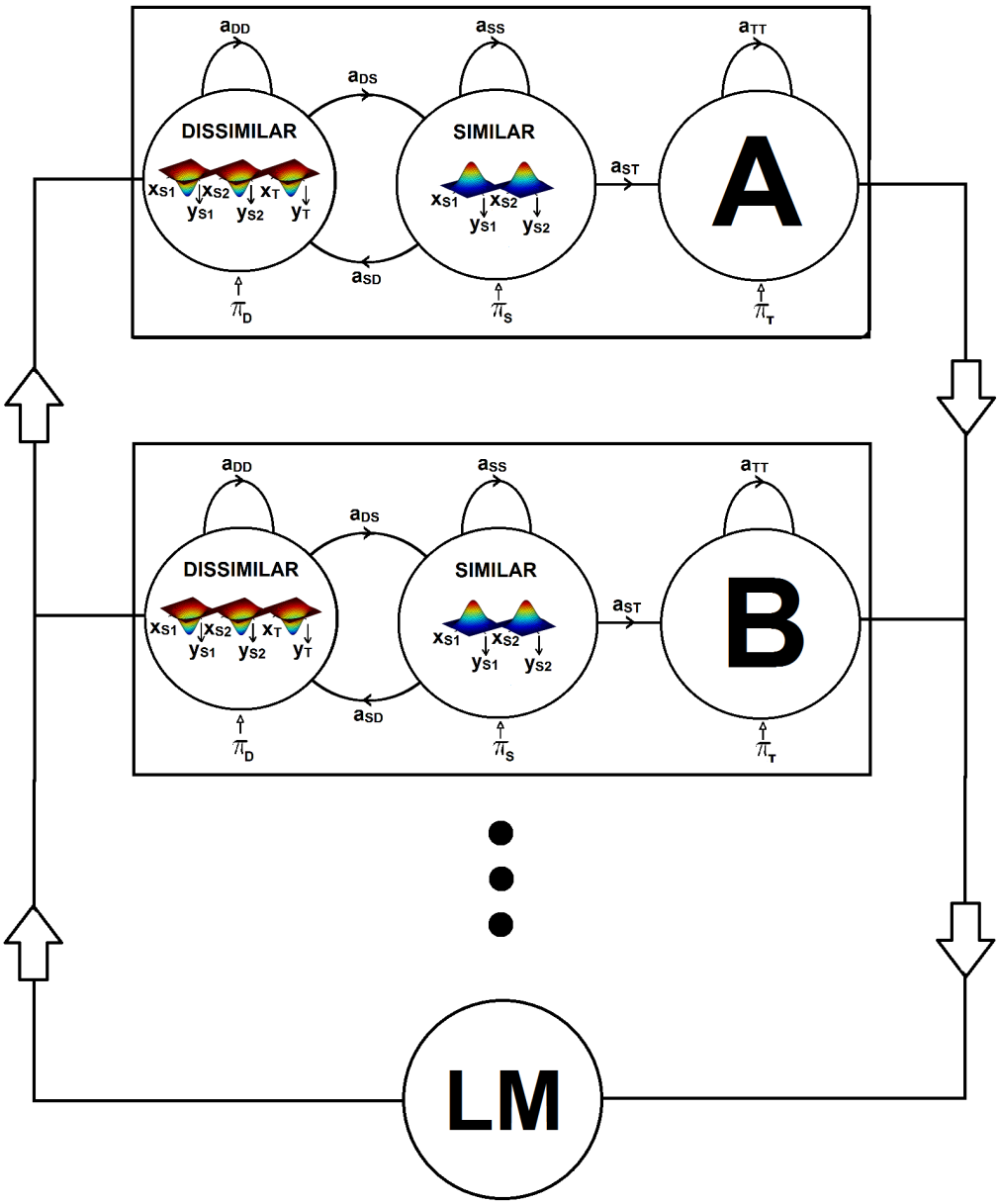


Figure 5–3: The word lexicon. The state at the bottom is where language model parameters are applied to the transitions.

Given the structure of the word lexicon, the model can generate words with arbitrary number of characters. Also in the new model the number of words in the dictionary will not impose any computational overhead in the inference. This feature is due to the unified representation of the word models in the word lexicon as opposed to the multiple model creation for the words in the dictionary in the classical Bayesian approach. Therefore, in the word lexicon model we can optionally have a large or unlimited number of words in the dictionary resulting in recognizing the task rather than classifying it.

Although the complexity of the evaluation phase is independent of the number of words in the dictionary and the complexity of the LM training is linear to it, the larger dictionary will result in more uncertainty in the LM when transitioning from a character to another, which is the direct effect of the larger number of character pairs that appear in the LM training set. In other words, training the LM on a large dictionary will broaden the choices of transitions from a character to another and will increase the number of pairs that are counted in the training word set leading to a larger "$c()$" in equation (5.7).

5.2.1 The Token Passing Technique

Training the word lexicon is similar to what we had in the character TSHMMs. To build up a word lexicon, we need to train the character TSHMMs as was done in section 4.2.3 and concatenate them according to figure 5–3 using the LM.

For training the LM, we also use the same method as used in section 5.1 and find the probabilities by counting the frequencies of seeing different character pairs in the training set. That said, the unified structure of the lexicon for all of the words

in the dictionary necessitates a different technique than the forward algorithm that was used for evaluating the likelihood term in the Bayesian inference (section 2.3.2).

In this section we introduce a variation of the forward algorithm, called *token passing*, whereby we directly infer the tasks in the lexicon models. We borrow this technique from the literature related to speech processing [129], where they use token passing to find the best sequence of recognition units (phonemes, characters, words, etc.) that better matches the feature vector elicited from the observation audio signal.

In the token passing technique, finding the best sequence of states that matches the observation sequence is based on minimizing a cost function C (or maximizing a reward function) that incorporates the transition and observation probabilities. The tokens can be thought of as objects that can move in the lexicon network from state to state and keep track of the traversed state sequence and the cost function inside themselves. The cost is incurred by either transitions from a state to another (*transition cost* C_t) or by generating an observation vector at each time step (*local cost* C_l).

The transition cost is denoted by $C_t(i, j)$, which can be calculated by the following equation:

$$C_t(i, j) = -\log P_{ij} \tag{5.8}$$

Here P_{ij} is the probability of the transition from state i to state j. As we can see, lower transition probabilities will result in higher costs and vice versa. When entering an HMM, P is equal to the initial state probability (Π) as denoted in equation (2.8). If the transition is within a HMM, P is equal to the transition probability of the

HMM (a_{ij}) as denoted in equation (2.7). For transitions between the HMMs, in addition to the initial state probabilities of the target state, the cost includes that of imposed by the LM, i.e.:

$$P_{ij} = P(\lambda_{\theta_j}|\lambda_{\theta_i}) * \pi_j. \tag{5.9}$$

In this equation $P(\lambda_{\theta_j}|\lambda_{\theta_i})$ is the weight imposed by the LM as denoted in equation (5.7) and π_j is the initial state probability. Thus, the transition cost function in transitions between HMMs becomes:

$$C_t(i,j) = -\log P(\lambda_{\theta_j}|\lambda_{\theta_i}) - \log \pi_j \tag{5.10}$$

The second type of cost is the local cost (C_l), which is defined by the level of similarity between the observation vector at time t and the observation pdf of the current state. In order to define the local cost, we quantify it in terms of the observation probability density function (b_j) as follows:

$$C_l(j,t) = -\log b_j(\vec{O}_t) \tag{5.11}$$

where $b_j(\vec{O}_t)$ is the likelihood of generating the observation vector $\vec{O}_t$ at state j, whose value can be calculated according to equation (2.9). Therefore, the overall cost for a transition from state i to state j at time t is calculated as:

$$C(i,j,t) = C_t(i,j) + C_l(j,t). \tag{5.12}$$

Algorithm 1 shows how we can use this cost function in order to decode a sequence of eye movements by finding the path with the lowest cost in the lexicon of figure 5–3. We start the algorithm by assigning a token to all of the states in the

lexicon. At the time step $t = 0$ the cost value of each of the tokens is defined by the initial state probability of their respective states. After initialization, we copy each token to all connecting states and increment its cost by $C(i, j, t)$ according to equation (5.12). Once all tokens are copied to all connecting states, for each state we discard the previous token and from the imported tokens keep the one with the minimum cost function. This process is repeated until we reach the end of the observation sequence. At that time we compare the cost of the tokens in the final states of all of the HMMs (the target state) and pick the one with the minimum cost value. The sequence of states which has yielded that token, then, is selected as the loci of attention during the visual-task execution. Therefore, the state history of the final token reveals the characters that were most likely visited during the task execution and can be used as the inference about the visual-task. In the experiments we compare the accuracy of inferences made by the Bayesian formulation and the token passing technique and show that with a slight sacrifice in the accuracy we can use the unified model of the lexicon to recognize the task being executed.

5.2.2 Evaluation

In order to compare the results of the task inference using the lexicon and the Bayesian frameworks, we used the same database of eye movements that was used in chapter 4. The database of task-dependent eye trajectories is composed of the recordings of the eye movements of six subjects while eye-typing 26 different 3-character words. The words were selected so that there was no repetition of characters in them.

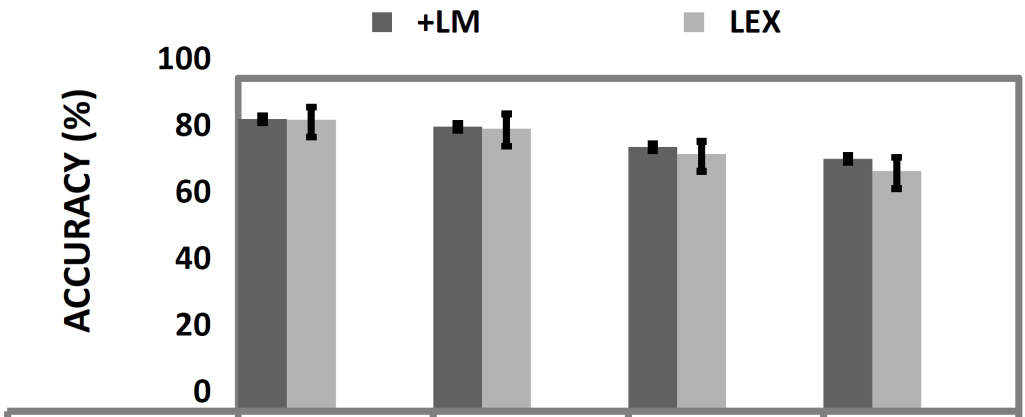
After the preprocessing, we obtained a database of 145 trajectories each of the form $(O_1, \ldots, O_m)$, containing observation sequences of coordinates of fixations

Algorithm 1 Token Passing

Initialize:
Assign a zero valued token to the initial states of the word models.
Assign an infinity valued token to all other states.
Algorithm:
for t:=1 to T **do**
  for each state i **do**
    Copy the token in each state i to the connecting state j and increment its value by $C(i,j,t) = C_t(i,j) + C_l(j,t)$ (see equation (5.12))
  end for
  Discard the original tokens.
  for each state i **do**
    Keep the token with the minimum value and discard the rest.
  end for
end for
Termination:
In the final states, the token with the smallest value corresponds to the best match.

while performing the eye-typing, where $O_i = (x_i, y_i)$ represents x-coordinate and y-coordinate of the i^{th} fixation, respectively.

As for the a-priori information, we used the same LMs as in the previous experiment (section 5.1.1). The LMs were trained on four sets of dictionaries of 26, 52, 104 and 312 English words all including the core 26 words of the smallest dictionary (the same dictionaries used in chapter 4 were used here). The words were selected randomly from the CMPD ([121]) and the words length varied between three to five characters. Also in the FSM structure of the lexicon, for each character, we used the TSHMM of the character, which corresponds to the model used in the previous experiment.



DICT. SIZE	W026	W052	W104	W312
+LM (MEAN)	87.7	85.4	79.3	75.8
LEX (MEAN)	87.5	84.8	77.2	72.1
+LM (SEM)	4.4	3.3	3.2	4.6
LEX (SEM)	3.8	4.5	3.9	4.1

Figure 5–4: Comparison of task classification accuracy using the Bayesian classifier (+LM) and the word lexicon (LEX) in a word classification task. The size of the dictionary on which the LM is trained is written at the first row of the table and their respective accuracies obtained using each of the models follow them. Each bar shows the mean classification rate (%) of correctly recognizing the intended word in the eye-typing application. The mean value and the standard error of the mean (SEM) are represented by bars and the numerical values are given in the following table.

5.2.3 Results

Figure 5–4 shows the accuracy of the word inference using the Bayesian inference and the lexicon models with the LMs trained over dictionaries ranging over four sizes. The result of the Bayesian inference is the same as the results shown in the TSHMMS (see figure 4–9). As we can see with a little sacrifice in accuracy, we can obtain the same results using the word lexicon model and token passing technique. The small

decrease in the accuracy is due to the fact that in the token passing technique, at each time step we only keep the best token of each state and eliminate the rest of them. This reduces the accuracy of the recognition model for the trajectories with noisy observations. Nevertheless, the lexicon obtains similar accuracy in word recognition as compared to the Bayesian inference while eliminating the requirement for pre-building the word models in advance.

One important point to be noted is that although in the Bayesian inference we do not explicitly use the LM as the task priors, the a-priori information is implicitly applied in the inference by limiting the classification results to one of the words in the dictionary. Thus, as opposed to the previous experiment for character classification where the -LM model was a Bayesian classifier with a flat task priors, here the Bayesian classifier is implicitly using the LM and falls under the +LM models of the previous section (hence using +LM to denote the Bayesian classifier).

The table in figure 5–4 shows the accuracy and the standard error of the mean (SEM) of the corresponding bars. For each bar we ran a 10-fold cross validation on our database of 145 trajectories in order to define the training and test sets and used the same epochs across both methods.

5.3 Discussion

In this chapter we showed how different sources of information can be merged into the Bayesian inference. We showed that the prior information about the tasks can improve the accuracy of the inference by turning the maximum likelihood (ML) inference into a maximum a-posteriori (MAP) inference. The a-priori information can originate from different sources. For instance, in English words the current

character can give prior information about the next character. In section 5.1 we used the frequency of character pairs to capture the pattern of transitions from a character to another. The information was encapsulated in a character network, called the language model (LM), which gives us chances of seeing different character given the current one.

The LM is a n^{th}-order Markov chain that assigns probabilities to transitions from one character to another based on the previous history of them. Depending on the order of the Markov chain an n-gram LM is generated. The order of the LM varies the accuracy of the LM in predicting the next character at the expense of training complexity. If we increase the order of the LM, longer patterns can be captured in the LM, which can result in a better prediction of the upcoming characters. However, a higher order LM necessitates a larger training database or otherwise there is a risk of divergence in the training phase. The number of character combinations grows exponentially with the order of the LM, which also entails an exponential growth in the training set in order to have all of the character sequence combinations in the training database.

In section 5.1.1 we showed that incorporating the knowledge from the LM into the inference leads to a MAP classifier, which significantly mitigates the accuracy of the Bayesian inference with a flat a-priori pdf (ML classifier). In order to benchmark the MAP classifier against the ML classifier, we compared the results of both classifiers in task inference on a unique database. The database was extracted from the eye-typing trajectories of the virtual keyboard experiment of chapter 4. The goal in our experiment was to infer the character that was being eye-typed given

the eye movement trajectories. In order to extract the character trajectories from the word trajectories, we approximated them by manually splitting the fixations into three sections, each dedicated to one of the characters in the word. The splitting was inspired after the TSHMM model of chapter 4, where we modeled the terminal fixations while searching a character by a Gaussian probability density function around the target character. Thus, in order to split a trajectory, we looked for fixations around the target characters and split the trajectories at those fixations.

Although in extracting the character database we implicitly postulate that searching for the characters of a word are independent of each other, this assumption has a similar effect on both the MAP and the ML classifiers. Therefore, even though this assumption might occasionally be violated by the memory effect, we can safely compare the results of the classifiers with each other and the improvement achieved by the MAP classifier is independent of any potential effect of the assumption on the results of the ML classifier. In other words, since in the Bayesian inference we marginalize the a-priori term from the likelihood term, we can examine the effect of the a-priori term regardless of the accuracy of the likelihood term.

One of the shortcomings of the classical Bayesian inference method is that for each task in the task set, we need to build a separate HMM and find out which model better gives rise to the observation vector to select it as the inferred task. While building a model for a small-size dictionary of tasks is not complicated, the complexity of the inference grows with the number of tasks in the dictionary. Another shortcoming of the classical Bayesian approaches is that they can only be used for classification. In classification we have a limited number of tasks in the dictionary

and we need to find out which task was being executed during the recording of the observation vector. However, in recognition, the number of possible tasks is not limited to a specific number and the dictionary can be infinitely large.

In section 5.2 we introduced a variation of Bayesian inference that uses a unified model for all of the tasks in the dictionary. By using the lexicons, not only we were able to eliminate the need for building a new models for the new tasks in the dictionary, but also we could generalize the inference to recognition of the tasks. To do so, we used a technique called *token passing* to decode the underlying states of an observation vector in a lexicon.

The lexicon and the token passing technique introduce a straightforward way of combining the HMMs with the LM into a unified attention model that can recognize the task with a linear complexity with respect to the number of words in the dictionary. One of the benefits of using the explicit LM in the lexicon is that when new vocabularies are added to the dictionary we only need to amend the LM. Moreover, we can manually change the task priors in the unified lexicon model by changing the transition probabilities of the LM. This is especially helpful when recognizing the task rather than classifying it, where we should manually change the LM in order to allow for out-of-dictionary tasks.

In the token passing method we moved hypothetical tokens throughout the lexicon that updated an internal cost function assigned to each of them. The cost function reflects the likelihood of the token being at the current state at the current time given the current observation vector. In the end, the token with the lowest cost

shows the most potential sequence of states that could have resulted in the given observation trajectory.

Although the token passing algorithm is an efficient way of traversing the lexicon with a linear complexity and linear memory space, it is suboptimal compared to the Bayesian inference for a classification task. At each time step, the token passing only keeps the token with the lowest cost and discards the rest of the tokens in each state. Therefore, a noisy observation can degrade the performance of the algorithm. That said, even though a noisy observation can cause the best token to be eliminated, similar tokens with low costs still have a chance to survive and replace the eliminated, optimal token and make a correct inference about the task.[1]

The experimental results compare the classification results of the Bayesian and the lexicon models and show only a small degradation in the results of the lexicon compared to the Bayesian.

[1] other variations of the token passing algorithm exist that keep more than one token at a given time step at the expense of exponentially increasing the computational complexity and the memory usage of the algorithm [129].

CHAPTER 6
Conclusion

Predicting the cognitive state of an observer from eye movements has been studied before in several works. Recognition of the physical activity, detection of tiredness or distraction, estimating the mental workload, detecting schizophrenic patients and indicating the mental fatigue are some of the inferences about the cognitive state of an observer that have been successfully accomplished in the oculomotor studies of human. As another feature of the cognitive state of an observer, in this thesis we studied visual-task detection by observing the pattern of eye trajectories. Our main goal was to reveal the task by detecting the goal-driven behavior of the oculomotor mechanism of an observer that directs the eyes in a scene to gather task-relevant information. We showed that eye movement can be used to reveal the attention demanding targets in a scene, which in turn can be used to infer the task. This work was inspired after the experiment by Yarbus that showed the visual-task affects the pattern of eye movement.

In a recent attempt Greene et al. [46] tried to realize an inverse Yarbus process, whereby we could detect the visual-task from the patterns of the eye movement. However, they could only achieve accuracies that were low or at the chance level. In this work we proposed that this failure originates from the feature set used to represent the eye movements. Particularly we showed the aggregate measures of eye

movements, such as the so-called *summary statistics*, do not retain either the temporal order, or the spatial information of the fixations, which leads to a insufficient representation of the eye trajectories for the task inference purpose.

In our proposed model we keep both spatial and temporal information of the fixations in a 3-tuple of the form (x, y, t), where x and y indicate the location of the fixation in a Cartesian space and t denotes the order of the fixation in the time space. Using this feature set, though, leads to a huge possible space of observations, which causes divergence in training of classifiers.

In chapter 2 we showed that although eye-trajectories can vary from subject to subject, they all serve to gather relevant information in service of the ongoing visual-task in the sense that the visual process directs the fixation to the informative locations in an image by means of a process known as *visual attention*.

In chapter 2 we showed that, even though the observation space spans a very large 3D space of possible feature sets (composed of the x, y and t information of the fixations), it can be mapped to a 2D space of attentional spots that indicates the areas attended during a task and their corresponding time stamps. The mapping from the eye movement space to the attentional space is done by models of visual attention. Therefore, as a secondary target of the thesis we investigated the visual attention models in order to track the foci of attention in an eye trajectory.

Several attention models were studied in the thesis, a common feature being the use of an integrated map of conspicuous locations in the image called the *saliency map*. Both *bottom-up* and *top-down* variations of the attention models were described in the background section and the disadvantages of each model were highlighted.

Based on the level of engagement between the fixation location and the focus of attention we have *covert* or *overt* types of visual attention. In an overt shift of attention, the center of gaze (COG) is directed to the focus of attention (FOA). In a covert shift of attention, however, the COG can be well away from the FOA, making the FOA harder to track.

Regardless of the overtness or covertness of attentional deployment, other factors can also contribute to the divergence between the measured COG and the FOA. These factors include noisy instruments, deliberate attentional focus in the parafoveal vision and visual phenomena such as the center of gravity.

One of the main assumptions in classical attention models is the overtness of attention. In other words, all of these models assume that the COG is the same as the FOA and use eye trajectories to represent the foci of attention. In chapter 3 we showed the difference between these two phenomena and showed how the COG can diverge from the FOA in many instances in a trajectory.

The disparity between the COG and the FOA can introduce noise to the mapping from the eye movement space to the attention space, which in turn degrades the accuracy of the task classifiers that are trained and tested on the noisy representation of the foci of attention. As a solution to this issue, we suggested a Gaussian probability distribution function (pdf) to relate the COG and the FOA in an eye trajectory. The Gaussian pdf represents the quasi-circular foveal region of the visual system that gives different probabilities to the fixations around an attentional target. This probability is higher at locations close to the target and diminishes as the COG moves away from the target. By this interpretation even fixations on the

neighboring targets could indicate a covert attention on the target, albeit with a lower probability.

In the literature related to eye movements, Markov chains are shown to be a good model for representing the dependence between consecutive fixations in a trajectory. Particularly, first order Markov chains have been successfully applied to model the pattern of eye movements in a variety of visual-tasks and are adopted by many studies as the visual cognitive model for directing the COGs. In this thesis we assumed that the model that governs the transitions between the FOAs is also a Markov chain and supported this assumption by reviewing some of the effects reflected in psychophysical experiments of visual attention, such as inhibition of return, similarity preference and proximity preference.

Assuming a first-order Markov chain as the generative model of the attentional spots and a Gaussian observation function as the generative model of fixations we can map the eye movements to the foci of attention by Hidden Markov models (HMMs). The HMMs are a class of supervised and semi-supervised learning models that generates different overt observations given the covert state. The observations are generated by a Gaussian pdf and the transitions between the states are governed by a first-order Markov chain. This description of the HMMs makes them a perfect fit for the purpose of modeling attention.

In chapter 3 we showed how we can use HMMs as the underlying model of visual attention. In the proposed models, the hidden states represent the FOA and the Gaussian observation pdfs model the COGs in an eye trajectory.

In chapter 4 we used the HMM attention models to infer the visual-task in synthetic and natural images. Several attention models were progressively developed to cope with the challenges in different conditions. The models differed from each other in the number of states, number of Gaussian pdfs in the Gaussian mixture models of each state and the definition of targets in the training phase. The evolution of the models started from a single-state HMM designed for inferring a set of simple tasks executed on synthetic images. Then we improved the model to the double-state and tri-state HMMs to infer more complex tasks on the images. Eventually we adopted a k-means clustering method in the training phase to come up with a model that can infer complex tasks executed on natural images, where we do not know the locations of the potential attentional targets in advance.

For the task inference we used the Bayesian inference formulation, whereby we could find the task-dependent attention model that resulted in the maximum a-posteriori probability. We used the HMMs trained for each task and used them in turn to evaluate the likelihood term for a given observation trajectory. Eventually, the HMM that better gave rise to the observation vector was selected as the ongoing task.

In chapter 5 we showed how prior knowledge about the tasks can refine our results by modulating the likelihood term of the Bayesian inference. To represent the a-priori knowledge about the task in the eye typing application, we introduced language models, whereby we trained a Markov chain on the pairs of characters appearing in the training set and used them to evaluate the a-priori chances of seeing a character given the current one. This probability was used in the Bayesian

inference as the a-priori term and modulated the likelihood term to refine the results of task inference.

Eventually we used an alternative variation of the Bayesian inference to eliminate the need for building the model for each of the tasks in the task set. In this unified model (called the *lexicon*) we combined the language model and the HMM attention models into a state machine that represented the task inference model for all of the tasks in the task set. We showed that the lexicon can reach accuracies close to those of the Bayesian inference by building the models on the fly and discarding the unlikely choices as the eye trajectory progresses.

6.1 Contributions

The contributions of the work described in this thesis are as follows:

1. Developing a Bayesian inference framework to infer the visual-task given the eye movements.
2. Developing a novel task-dependent attention model that represents the covert as well as the overt classes of attention shifts.
3. Incorporating the hidden Markov model attention models into the Bayesian inference for task inference.
    - Developing the single-state HMMs for simple task inference executed on synthetic images.
    - Developing the double-state and tri-state HMMs for complex task inference executed on synthetic images.
    - Incorporating a k-means clustering technique into an ergodic HMM for complex task inference executed on natural images.

- Building a database of task-dependent eye movements for simple tasks executed on synthetic images.
- Developing an eye typing application to build up a database of task-dependent eye movements for complex tasks executed on synthetic images.
- Building a database of task-dependent eye movements executed on natural images.

4. Developing a maximum a-posteriori Bayesian inference for visual-task inference.
    - Developing language models to evaluate the a-priori knowledge about the visual-task in the eye typing application.
    - Incorporating the language model into the Bayesian inference to obtain a maximum a-posteriori task inference.
5. Developing the lexicon model as a variation of the Bayesian inference to develop a unified task inference model.
    - Applying the token passing technique on the lexicon model to infer the task on the fly for a large-size task sets.

6.2 Future Directions

Due to the use of the observation pdfs, the HMMs allow for the discrepancy between the fixations and attention spots in a covert deployment of attention. Each observation pdf specifies an attentional spot, while a random outcome from the pdf specifies a fixation location while attending that spot. The observation pdf is a 2D Gaussian pdf that spans an area around the attentional spot, the random outcomes

of which can be well away from its center. By this interpretation of variables the HMMs allow for both covert and overt types of attention.

That said, the purpose of this work was to infer the visual-task based on the recordings of the fixations made on an image and not to infer the FOA. In order to accurately compare the hidden states of our models with the focus of attention, we should use methods such as *attentional-probing*, detecting microsaccades [48] or fMRI recording [124] in order to locate the FOA and measure the correlation between their estimation and the centroids of the HMM states.

An interesting phenomenon seen in the results is the accuracy of the task inference for different standard deviations of the observation pdf (figure 4–13a), which indicates a falloff in the task classification accuracy as the standard deviation diverges from a value of roughly 4^o of the visual angle. This effect is consistent with previous estimates of the size of the *operational fovea* as the central 3^o of vision [64]. In [13] it is shown that targets within 4^o of central vision are still perceived at 50% of maximal acuity. Although based on the current evidence we cannot tell whether this finding is a real effect or merely a coincidence, another experiment with a different distance between the observer and the screen can help us determine that.

Tracking covert attention can help us better predict the next target of eye fixation based on the currently attended target. This is based on the classical study of covert attention by Posner [96], which suggests that covert attention and eye movements are both drawn to exogenous (peripheral) stimuli, with covert attention moving more rapidly towards the stimulus. Thus, another interesting direction to be taken is to apply the proposed task inference model in real-world applications

to improve the user experience. For instance, knowing what the user is seeking on a web page combined with a dynamic 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 foci of attention inferred from each viewer's eye movements.

On a related topic, Vidal et al. [119] implemented a pervasive healthcare application by using the eye movements to infer the mental status of the patients. Bulling et al. [8] used eye movement to obtain information about a person's context suggesting a context-aware, pervasive computing system based on the eye movements. As mentioned earlier, a by-product of the HMM model is to locate the focus of attention, whether it is overt or covert. This feature allows us to track the informative attentional spot, rather than the noisy motion of the gaze. Thus, in applications that are based on eye movements, performance gains might be obtained by using the attentional locus, which is more task-oriented and robust, rather than the noise-prone gaze information.

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. Thus, one interesting direction would be to increase the accuracy of the results by using larger databases of task-dependent eye movements and more accurate models of task priors.

Other variations of HMMs could also be explored to see how they perform in the context of task inference. In particular a non-parametric variation of the HMMs, called the hierarchical Dirichlet process HMM (HDP-HMM) [115], has recently been

studied as a better replacement for the HMMs. The HDP-HMMs do not assume a pre-defined number of states in the HMMs and learn that information during the training process. This specification of the HDP-HMM is specifically helpful in task inference in video stimuli, where the number of states (targets) can change during a trial.

References

[1] D.H. Ballard and M.M. Hayhoe. Modelling the role of task in the control of gaze. *Visual Cognition*, 17(6-7):1185–1204, 2009.

[2] D.H. Ballard, M.M. Hayhoe, and J.B. Pelz. Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 7(1):66–80, 1995.

[3] W. Becker. The control of eye movements in the saccadic system. *Cerebral Control Of Eye Movements And Motion Perception*, 82:233–243, 1972.

[4] Y. Bengio and P. Frasconi. An input output HMM architecture. In *Advances in neural information processing systems*, pages 427–434. Morgan Kaufmann Publisher, 1995.

[5] P.J. Benson, S.A. Beedie, E. Shephard, I. Giegling, D. Rujescu, and D. St Clair. Simple viewing tests can detect eye movement abnormalities that distinguish schizophrenia cases from controls with exceptional accuracy. *Biological Psychiatry*, 72(9):716–724, 2012.

[6] C.M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer: New York, 2006.

[7] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):185–207, 2013.

[8] A. Bulling, D. Roggen, and G. Troster. What's in the eyes for context-awareness? *Pervasive Computing, IEEE*, 10(2):48–57, 2011.

[9] A. Bulling, J.A. Ward, H. Gellersen, and G. Tröster. Eye movement analysis for activity recognition. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 41–50. ACM, 2009.

[10] A. Bulling, J.A. Ward, H. Gellersen, and G. Troster. Eye movement analysis for activity recognition using electrooculography. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(4):741–753, 2011.

[11] G.T. Buswell. *An experimental study of the eye-voice span in reading.* Number 17. University of Chicago, 1920.

[12] G.T. Buswell. *How people look at pictures: A study of the psychology of perception in art.* Chicago: University of Chicago Press, 1935.

[13] R.H.S. Carpenter. The visual origins of ocular motility. *Vision And Visual Function*, 8:1–10, 1991.

[14] M. Carrasco. Visual attention: The past 25 years. *Vision Research*, 51(13):1484–1525, 2011.

[15] M.S. Castelhano and J.M. Henderson. Stable individual differences across images in human saccadic eye movements. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 62(1):1–14, 2008.

[16] M.S. Castelhano, M.L. Mack, and J.M. Henderson. Viewing task influences eye movement control during active scene perception. *Journal of Vision*, 9(3):6, 2009.

[17] J.J. Clark and J.K. O'Regan. Word ambiguity and the optimal viewing position in reading. *Vision Research*, 39(4):843–857, 1998.

[18] P. Clarkson and R. Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. In *Fifth European Conference on Speech Communication and Technology*, pages 2707–2710, 1997.

[19] C.E. Connor, H.E. Egeth, and S. Yantis. Visual attention: bottom-up versus top-down. *Current Biology*, 14(19):R850–R852, 2004.

[20] J.W. De Fockert, G. Rees, C.D. Frith, and N. Lavie. The role of working memory in visual selective attention. *Science*, 291(5509):1803, 2001.

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

[22] L.L. Di Stasi, R. Renner, A. Catena, J.J. Cañas, B.M. Velichkovsky, and S. Pannasch. 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, 2012.

[23] L.L. Di Stasi, R. Renner, P. Staehr, J.R. Helmert, B.M. Velichkovsky, J.J. Canas, A. Catena, and S. Pannasch. Saccadic peak velocity sensitivity to variations in mental workload. *Aviation, Space, And Environmental Medicine*, 81(4):413–417, 2010.

[24] M. Dorr, K.R. Gegenfurtner, and E. Barth. The contribution of low-level features at the centre of gaze to saccade target selection. *Vision Research*, 49(24):2918–2926, 2009.

[25] J.A. Droll, M.M. Hayhoe, J. Triesch, and B.T. Sullivan. Task demands control acquisition and storage of visual information. *Journal of Experimental Psychology: Human Perception and Performance*, 31(6):1416, 2005.

[26] K.A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, 17(6-7):945–978, 2009.

[27] W. Einhäuser, U. Rutishauser, and C. Koch. Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, 8(2):2, 2008.

[28] W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14), 2008.

[29] M. Elhelw, M. Nicolaou, A. Chung, G.Z. Yang, and M.S. Atkins. A gaze-based study for investigating the perception of visual realism in simulated scenes. *ACM Transactions on Applied Perception (TAP)*, 5(1):3, 2008.

[30] F.L. Engel. Visual conspicuity, directed attention and retinal locus (Visual conspicuity measurements, determining effects of directed attention and relation to visibility). *Vision Research*, 11:563–575, 1971.

[31] F.L. Engel. Visual conspicuity and selective background interference in eccentric vision. *Vision Research*, 14(7):459–471, 1974.

[32] J. Epelboim, R.M. Steinman, E. Kowler, M. Edwards, Z. Pizlo, C.J. Erkelens, and H. Collewijn. The function of visual search and memory in sequential looking tasks. *Vision Research*, 35(23):3401–3422, 1995.

[33] C.W. Eriksen and J.D. James. Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics*, 40(4):225–240, 1986.

[34] C.J. Erkelens and I. Vogels. The initial direction and landing position of saccades. *Eye Movement Research: Mechanisms, Processes, And Applications*, 6:133–144, 1995.

[35] J.M. Findlay. Local and global influences on saccadic eye movements. *Eye Movements: Cognition And Visual Perception*, pages 171–179, 1981.

[36] B. Fischer and H. Weber. Express saccades and visual attention. *Behavioral and Brain Sciences*, 16:553–553, 1993.

[37] J.R. Flanagan and R.S. Johansson. Action plans used in action observation. *Nature*, 424(6950):769–771, 2003.

[38] A.F. Fuchs. The saccadic system. In C.C. Collins and J.E. Hyde, editors, *The control of eye movements*, pages 343–362. New York: Academic Press, 1971.

[39] S. Furneaux and M.F. Land. 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, 1999.

[40] G. Geiger and J.Y. Lettvin. Enhancing the perception of form in peripheral vision. *Perception*, 15(2):119, 1986.

[41] Z. Ghahramani. Learning dynamic Bayesian networks. *Lecture Notes in Computer Science*, 1387:168–197, 1998.

[42] G.C. Gilmore, H. Hersh, A. Caramazza, and J. Griffin. Multidimensional letter similarity derived from recognition errors. *Attention, Perception, & Psychophysics*, 25(5):425–431, 1979.

[43] P.S. Goldman-Rakic. Cellular basis of working memory. *Neuron*, 14(3):477, 1995.

[44] Google. Life photo archive hosted by Google, June 2013.

[45] M.R. Greene, T. Liu, and J.M. Wolfe. Reconsidering Yarbus: Pattern classification cannot predict observer's task from scan paths. *Journal of Vision*, 11(11):498–498, 2011.

[46] M.R. Greene, T. Liu, and J.M. Wolfe. Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research*, 62:1–8, 2012.

[47] S.S. Hacisalihzade, L.W. Stark, and J.S. Allen. Visual perception and sequences of eye movement fixations: A stochastic modeling approach. *Systems Man and Cybernetics, IEEE Transactions on*, 22(3):474–481, 1992.

[48] Z. Hafed and J.J. Clark. Microsaccades as an overt measure of covert attention shifts. *Vision Research*, 42(22):2533–2545, 2002.

[49] J.V. Haxby, M.I. Gobbini, M.L. Furey, A. Ishai, J.L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

[50] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends In Cognitive Sciences*, 9(4):188–194, 2005.

[51] M.M. Hayhoe, D.G. Bensinger, and D.H. Ballard. Task constraints in visual working memory. *Vision Research*, 38(1):125–137, 1998.

[52] M.M. Hayhoe, A. Shrivastava, R. Mruczek, and J.B. Pelz. Visual memory and motor planning in a natural task. *Journal of Vision*, 3(1):6, 2003.

[53] P. He and E. Kowler. The role of location probability in the programming of saccades: Implications for center-of-gravity tendencies. *Vision Research*, 29(9):1165–1181, 1989.

[54] M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, and B. Scholkopf. Support vector machines. *Intelligent Systems and Their Applications, IEEE*, 13(4):18–28, 1998.

[55] H.V. Helmholtz. *Handbuch der Physiologischen Optik, Dritter Abschnitt.* Zweite Auflage ed., 1896.

[56] J.M. Henderson, J.R. Brockmole, M.S. Castelhano, M. Mack, M. Fischer, W. Murray, and R. Hill. Visual saliency does not account for eye movements during visual search in real-world scenes. In R.P.G. van Gompel, M.H. Fischer, W.S. Murray, and R.L. Hill, editors, *Eye movements: A window on mind and brain*, chapter 25, pages 537–562. Elsevier, 2007.

[57] J.E. Hoffman and B. Subramaniam. The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, 57(6):787–795, 1995.

[58] J. Hu, M.K. Brown, and W. Turin. HMM based on-line handwriting recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(10):1039–1045, 1996.

[59] Inc. ISCAN. Operation instruction: Rk-726pci pupil/corneal reflection tracking system, version 1.1.01.

[60] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–204, 2001.

[61] L. Itti and C. Koch. Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, 10:161–169, 2001.

[62] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998.

[63] F. Jelinek. *Statistical methods for speech recognition*. MIT press, 1997.

[64] R.S. Johansson, G. Westling, A. Bäckström, and J.R. Flanagan. Eye–hand coordination in object manipulation. *the Journal of Neuroscience*, 21(17):6917–6932, 2001.

[65] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. *Mit Tech Report*, 2012.

[66] C. Kanan, M.H. Tong, L. Zhang, and G.W. Cottrell. Sun: Top-down saliency using natural statistics. *Visual Cognition*, 17(6-7):979–1003, 2009.

[67] L. Kaufman and P.J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. Wiley-Interscience, 2009.

[68] G. Keren and S. Baggen. Recognition models of alphanumeric characters. *Attention, Perception, & Psychophysics*, 29(3):234–246, 1981.

[69] R. Klein. Does oculomotor readiness mediate cognitive control of visual attention. *Attention And Performance VIII*, 8:259–276, 1980.

[70] R.M. Klein. Inhibition of return. *Trends In Cognitive Sciences*, 4(4):138–147, 2000.

[71] R.M. Klein and W.J. MacInnes. Inhibition of return is a foraging facilitator in visual search. *Psychological Science*, 10(4):346–352, 1999.

[72] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227, 1985.

[73] E. Kowler, E. Anderson, B. Dosher, and E. Blaser. The role of attention in the programming of saccades. *Vision Research*, 35(13):1897–1916, 1995.

[74] M. Land, N. Mennie, and J. Rusted. The roles of vision and eye movements in the control of activities of daily living. *PERCEPTION*, 28(11):1311–1328, 1999.

[75] M.F. Land and S. Furneaux. The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1358):1231–1239, 1997.

[76] M.F. Land and D.N. Lee. Where do we look when we steer. *Nature*, 369:742–744, 1994.

[77] M.F. Land and P. McLeod. From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience*, 3(12):1340–1345, 2000.

[78] M.F. Land and B.W. Tatler. Steering with the head: The visual strategy of a racing driver. *Current Biology*, 11(15):1215–1220, 2001.

[79] N. Lavie. Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, 21(3):451, 1995.

[80] G.R. Loftus. Eye fixations and recognition memory for pictures. *Cognitive Psychology*, 3(4):525–551, 1972.

[81] D.J.C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.

[82] S.K. Mannan, K.H. Ruddock, and D.S. Wooding. Fixation sequences made during visual examination of briefly presented 2d images. *Spatial Vision*, 11(2):157–178, 1997.

[83] E. Matin. Saccadic suppression: A review and an analysis. *Psychological Bulletin*, 81(12):899–917, 1974.

[84] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.R. Mullers. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX,*

1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop, pages 41–48. IEEE, 1999.

[85] M. Mills, Hollingworth A., S. Van der Stigchel, L. Hoffman, and Dodd M.D. Examining the influence of task set on eye movements and fixations. *Journal of Vision*, 11(8), 2011. article 17.

[86] V. Nair and J.J. Clark. Automated visual surveillance using hidden markov models. In *International Conference on Vision Interface*, volume 93, pages 88–93, 2002.

[87] J. Najemnik and W.S. Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, 2005.

[88] D. Noton and L.W. Stark. Scanpaths in eye movements during pattern perception. *Science*, 171(968):308–311, 1971.

[89] A. Nuthmann and J.M. Henderson. Object-based attentional selection in scene viewing. *Journal of Vision*, 10(8), 2010.

[90] J.K. O'Regan, H. Deubel, J.J. Clark, and R.A. Rensink. Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition*, 7(1-3):191–211, 2000.

[91] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002.

[92] A.E. Patla and J.N. Vickers. Where and when do we look as we approach and step over an obstacle in the travel path? *Neuroreport*, 8(17):3661–3665, 1997.

[93] A.E. Patla and J.N. Vickers. 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, 2003.

[94] J.B. Pelz and R. Canosa. Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research*, 41(25):3587–3596, 2001.

[95] R. Pieters, E. Rosbergen, and M. Wedel. Visual attention to repeated print advertising: A test of scanpath theory. *Journal of Marketing Research*, 36(4):424–438, 1999.

[96] M.I. Posner. Orienting of attention. *Quarterly Journal Of Experimental Psychology*, 32(1):3–25, 1980.

[97] M.I. Posner and Y. Cohen. Components of visual orienting. *Attention And Performance X: Control Of Language Processes*, 32:531–556, 1984.

[98] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Readings In Speech Recognition*, 53(3):267–296, 1990.

[99] P. Reinagel and A.M. Zador. Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, 10(4):341–350, 1999.

[100] G. Rizzolatti, L. Riggio, and B.M. Sheliga. Space and selective attention. *Attention And Performance Xv: Conscious And Nonconscious Information Processing*, 15:231–265, 1994.

[101] C.A. Rothkopf, D.H. Ballard, and M.M. Hayhoe. Task and context determine where you look. *Journal of Vision*, 7(14):16, 2007.

[102] U. Rutishauser and C. Koch. Probabilistic modeling of eye movement data during conjunction search via feature-based attention. *Journal of Vision*, 7(6):5, 2007.

[103] D.D. Salvucci and J.R. Anderson. Automated eye-movement protocol analysis. *Human-Computer Interaction*, 16(1):39–86, 2001.

[104] D.D. Salvucci and J.H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78. ACM, 2000.

[105] R. Schleicher, N. Galley, S. Briest, and L. Galley. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, 51(7):982–1010, 2008.

[106] W.X. Schneider. 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, 1995.

[107] W.X. Schneider and H. Deubel. 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, 2002.

[108] J. Simola, J. Salojärvi, and I. Kojo. Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research*, 9(4):237–251, 2008.

[109] T.J. Smith and J.M. Henderson. Facilitation of return during scene viewing. *Visual Cognition*, 17(6-7):1083–1108, 2009.

[110] L. W Stark and S.R. Ellis. Scanpaths revisited: Cognitive models direct active looking. In D.F. Fisher, R.A. Monty, and J.W. Senders, editors, *Eye Movements and Psychological Processes*, pages 192–226. Lawrence Erlbaum Associates, Hillsdale, NJ, 1981.

[111] B.W. Tatler, R.J. Baddeley, and B.T. Vincent. The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, 46(12):1857–1862, 2006.

[112] B.W. Tatler, M.M. Hayhoe, M.F. Land, and D.H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11(5), 2011.

[113] B.W. Tatler and B.T. Vincent. Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, 2(2):1–18, 2008.

[114] B.W. Tatler, N.J. Wade, H. Kwan, J.M. Findlay, and B.M. Velichkovsky. Yarbus, eye movements, and vision. *I-Perception*, 1(1):7, 2010.

[115] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical dirichlet processes. *Journal Of The American Statistical Association*, 101(476), 2006.

[116] A. Torralba, A. Oliva, M.S. Castelhano, and J.M. Henderson. 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, 2006.

[117] A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[118] R. Van Der Lans, R. Pieters, and M. Wedel. Eye-movement analysis of search effectiveness. *Journal of the American Statistical Association*, 103(482):452–461, 2008.

[119] Mélodie Vidal, Jayson Turner, Andreas Bulling, and Hans Gellersen. Wearable eye tracking for mental health monitoring. *Computer Communications*, 35(11):1306 – 1311, 2012.

[120] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

[121] R. Weide. The carnegie mellon pronouncing dictionary [cmudict. 0.6], 2005.

[122] M. Wischnewski, A. Belardinelli, W.X. Schneider, and J.J. Steil. Where to look next? combining static and dynamic proto-objects in a tva-based model of visual attention. *Cognitive Computation*, 2(4):326–343, 2010.

[123] M. Wischnewski, J.J. Steil, L. Kehrer, and W.X. Schneider. Integrating inhomogeneous processing and proto-object formation in a computational model of visual attention. In H. Ritter, editor, *Human centered robot systems*, volume 6, pages 93–102. Springer, 2009.

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

[125] J.M. Wolfe, S.J. Butcher, C. Lee, and M. Hyle. Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons. *Journal of Experimental Psychology-Human Perception and Performance*, 29(2):483–501, 2003.

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

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

[128] A.L. Yarbus. *Eye movements and vision*. New York: Plenum Press, 1967. Translated from the Russian edition by Haigh, B.

[129] S.J. Young, N.H. Russell, and J.H.S. Thornton. Token passing: a simple conceptual model for connected speech recognition systems. *Cambridge University Engineering Department*, pages 1–23, 1989.

[130] G.J. Zelinsky, R.P.N. Rao, M.M. Hayhoe, and D.H. Ballard. Eye movements reveal the spatiotemporal dynamics of visual search. *Psychological Science*, 8(6):448–453, 1997.

Ethical considerations

All the 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.

Role of the student in publications

The publications on which this thesis is based on are the result of the student's research during his PhD program. The student conducted the experiments, obtained the results and put them into writing. The whole research and submissions were under supervision of professor James Clark, who is the second author of the aforementioned publications.

Role of the funding source

The funding agencies played 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.

Consent Forms

The following are the forms the participants were asked to sign before the experiments described in Chapter 4 and 5, respectively.

Consent form

I the undersigned, _____, have voluntarily agreed to participate as a subject in an experimental series undertaken by Amin Haji Abolhassani (Ph.D. Candidate, Dept. of Electrical and Computer Engineering, Centre for Intelligent Machines, McGill University, 3480 University Street, room 464, Montreal, Quebec, Canada H3A 2A7, Tel. 514-398-5856) and research supervisor J. Clark. The experiments are being conducted at the McGill Centre for Intelligent Machines, Department of Electrical and Computer Engineering, McGill University in the McConnell Engineering building.

The purpose of these experiments is to study how visual task affects the allocation of visual attention. In particular, this study examines the interactions amongst covert attention, eye and head movements, and visual task. Participants in the study will be asked to sit in a chair and their eye movements will be recorded while performing simple visual tasks. Eye and head movements will be measured using a light-weight, head mounted camera system while looking at a CRT monitor. No health hazard is associated with the study.

This study involves wearing a commercial device on the head that will track one's eye and head movements – this involves wearing a baseball cap to which a lightweight camera assembly is attached. Subjects should be aware that the eye tracker shines a low-power infrared (IR) beam on the left eye used for proper illumination of the eye for gaze tracking purposes. It is very safe. In order to avoid muscle and mental fatigue (boredom) from the procedure, subjects are strongly encouraged to rest as much as needed. The beginning of a trial is controlled by the subject. There will also be mandatory rest periods. If at any point the head-mounted tracking equipment is causing discomfort due to prolonged exposure – subjects may remove it during a rest period and are encouraged to do so.

Subject identity, age and gender will remain completely confidential in any report of the results of this study and the data will be stored on a password protected computer which is only accessible to the principal experimenter. We anticipate disseminating the information of this research through publication in scientific conference proceedings, journals, as well as a thesis. We may also have several talks in scientific seminars. Subjects will receive monetary compensation of $10/hour for participating within the experimental protocol.

Should I decide at anytime, prior or during experimentation, to withdraw from the study, I may do so without any prejudice to my present or future activities at McGill University.

I am not aware of any personal or hereditary family ailments, in particular epilepsy, and to the best of my knowledge have no deficits in my balance or orienting systems; I consider myself in good health at the present time. I am not at present under the influence of any therapeutic or other drugs.

Age: _____ Sex: M / F

Date: _____

Address: _____

Phone: _____ Signature: _____

Consent form

I the undersigned, _____, have voluntarily agreed to participate as a subject in an experimental series undertaken by Amin Haji-Abolhassani (Ph.D. Candidate, Dept. of Electrical and Computer Engineering, Centre for Intelligent Machines, McGill University, 3480 University Street, room 407, Montreal, Quebec, Canada H3A 2A7, Tel. 514-398-5856). The experiments are being conducted at McGill Centre for Intelligent Machines, Department of Electrical and Computer Engineering, McGill University in McConnell Engineering building.

The purpose of these experiments is to study how visual task affects the allocation of visual attention. In particular, this study examines the interactions amongst attention, eye and head movements, and visual task. Participants in the study will be asked to sit in a chair and their eye movements will be recorded while performing simple visual tasks. Eye and head movements will be measured using a remote camera system while looking at a LCD screen. No health hazard is associated with the study.

This study uses a commercial device that will track one's eye and head movements. Subjects should be aware that the eye tracker shines a low-power infrared (IR) beam on both eyes used for proper illumination of the eye for gaze tracking purposes. It is very safe. In order to avoid muscle and mental fatigue (boredom) from the procedure, subjects are strongly encouraged to rest as much as needed. The beginning of a trial is controlled by the subject. There will also be mandatory rest periods.

Subject identity, age and gender will remain completely confidential in any report of the results of this study and the data will be stored on a password protected computer which is only accessible to the principal experimenter. We anticipate disseminating the information of this research through publication in scientific conference proceedings, journals, as well as a thesis. We may also have several talks in scientific seminars. Subjects will receive monetary compensation of $10/hour for participating within the experimental protocol.

Should I decide at anytime, prior or during experimentation, to withdraw from the study, I may do so without any prejudice to my present or future activities at McGill University.

I am not aware of any personal or hereditary family ailments, in particular epilepsy, and to the best of my knowledge have no deficits in my balance or orienting systems; I consider myself in good health at the present time. I am not at present under the influence of any therapeutic or other drugs.

Age: _____ Sex: M / F

Date: _____

Address: _____

Phone: _____ Signature: _____