Visual Task Inference Using Hidden Markov Models

Abstract

It has been known for a long time that visual task, such as reading, counting and searching, greatly influences eye movement patterns. Perhaps the best known demonstration of this is the celebrated study of Yarbus showing that different eye movement trajectories emerge depending on the visual task that the viewers are given. The objective of this paper is to develop an *inverse* Yarbus process whereby we can infer the visual task by observing the measurements of a viewer's eye movements while executing the visual task. The method we are proposing is to use Hidden Markov Models (HMMs) to create a probabilistic framework to infer the viewer's task from eye movements.

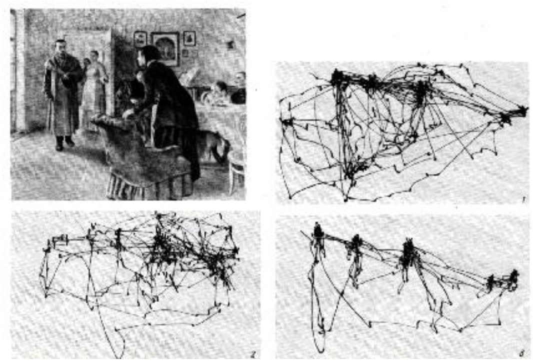


Figure 1: Eye trajectories measured by Yarbus by viewers carrying out different tasks. Upper right - no specific task, lower left - estimate the wealth of the family, lower right - give the ages of the people in the painting.

1 Introduction

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. As a support of this view, a study in change detection [Rensink *et al.*, 1997] suggests that the internal scene representations do not contain complete knowledge of the scene and the changes introduced in video images at locations that were not being attended-to were very difficult for people to notice.

The effect of visual task on attention has been long studied in the literature related to attention cognitive process of humans. In an early study, Yarbus [1967] showed that visual task has a great influence on the viewer's eye trajectory. In Figure 1 we can see how visual task can modulate the conspicuity of different regions and as a result change the pattern of eye movements. Perhaps Yarbus effect has been best studied for the task of reading. Clark and O'Regan [1998] have shown that when reading a text, the *centre of gaze* (COG) lands on the locations that minimizes the ambiguity of the word arising from the incomplete recognition of the letters. In another study Castelhano et al [2009] have shown that two tasks of visual search and memorization show significantly different eye movement patterns. Bulling et al [2009] have also shown that eye movement analysis is a rich modality for activity recognition.

Although the effect of visual task on eye movement pattern has been investigated for various tasks, there is not much done in the area of visual task inference from eye movement. In other words, in a forward Yarbus process the visual task is given as an input and the output is task-dependent scanpaths of eye movements. However, in this work we develop a HMM-based method to realize an inverse Yarbus process where eye patterns are the inputs to the system and the underlying visual task is given as the output. This task information, then, can serve as an input to the forward Yarbus process to predict the next attention grabbing location in the scene.

2 An Inverse Yarbus Process

As we know, *generative learning* is a class of supervised learning that classifies data in a probabilistic manner. By applying probabilistic inference we can develop a mathematical framework for merging all sources of information and presenting the result in the form of a probability density function. In our case of developing an inverse projection from eye movement space to visual task space, we need a probabilistic framework that can account for prior knowledge about the tasks, too. Moreover, we need the inference to give us a probability distribution over different possible tasks rather than a single task as the output. This way we can design a high level process that decides about the task and provides us with the degree of confidence in choosing various tasks as the output.

Suppose we have data of the form $< \mathbf{Q}, y >$, where $y \in Y$ is a task label (e.g., reading, counting, searching, etc.) and $\mathbf{Q}$ is the vector containing the observation sequence of fixation locations $(q_1, q_2, ..., q_T)$ sampled from a stochastic process $\{q_t\}$ in discrete time $t = \{1, 2, ..., T\}$ over random image locations; hence discrete-state-space $S = [s_1, s_2, ..., s_N]$, so that

$$q_i = s_j | i \in [1, T], j \in [1, N]. \quad (1)$$

In general, generative learning algorithms model two entities:

- $P(y)$: The *prior probability* of each task $y \in Y$.
- $P(\mathbf{Q}|y)$: The task conditional distributions which is also referred to as the *likelihood* function.

We can, then, estimate the probability of task y given a new sequence $\mathbf{Q}$ by a simple application of Bayes's rule:

$$P(y|\mathbf{Q}) = \frac{P(\mathbf{Q}|y)P(y)}{P(\mathbf{Q})} = \frac{P(\mathbf{Q}|y)P(y)}{\sum_{y' \in Y} P(\mathbf{Q}|y')P(y')}. \quad (2)$$

Thus, in order to make an inference, we need to obtain the likelihood term and modulate it by our prior knowledge about the tasks represented by the a-priori term. The likelihood term can be considered as an objective evaluation of forward Yarbus process in the sense that it gives us the probability of projections from visual task space to eye movement space. The likelihood term can be estimated as follows:

$$\begin{aligned} &P(\mathbf{Q}|y) = P(q_1, q_2, \ldots, q_T|y) \\ &= P(q_1|y)P(q_2|q_1, y) \ldots P(q_T|q_1 \ldots q_{T-1}, y). \end{aligned} \quad (3)$$

The standard approach for determining this term is to use a *saliency map* as an indicator of how attractive a given part of the field-of-view is to attention. 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 map. In the rest of this section we will show how these models posit different assumptions to estimate the likelihood term.

2.1 Bottom-Up Models

In bottom-up models, the allocation of attention is merely based on the characteristics of the visual stimuli and does not require any top-down guidance (task information) to shift attention. Moreover, in this model we assume that observations q_i are conditionally independent which reduces the likelihood term to:

$$\begin{aligned} P(\mathbf{Q}|y) &= P(q_1, q_2, \ldots, q_T|y) \\ &= P(q_1)P(q_2) \ldots P(q_T). \end{aligned} \quad (4)$$

This assumption is called the *naïve Bayes assumption* and only needs the probability of directing the *focus of attention* (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 [2001a]. 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 overlaying outputs from different filters tuned to simple visual attributes (color, intensity, orientation, etc.) and can generate a map for an input image without requiring any training.

Although image 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 by humans [Tatler, 2007]. In particular, when a visual task is involved, image salience models can fail almost completely [Einhäuser *et al.*, 2008]. In our view the bulk of this shortfall is due to a lack of task-dependence in the models.

2.2 Top-Down Models

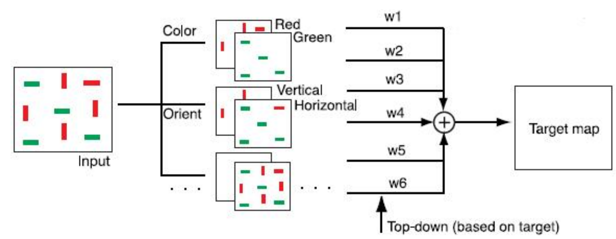


Figure 2: Influence of top-down, task dependent priors on bottom-up attention models. The influence can be modeled as a weight vector modulating the linear combination of the feature maps (after Rutishauser and Koch).

As highlighted in the previous section, attention is not just a passive, saliency-based enhancement of the visual stimuli; rather, it actively selects certain parts of a scene based on the ongoing task and saliency of the targets. The second major group of human attention models is top-down, task dependent method that modulates the saliency maps according to the viewer's visual task. Perhaps the best illustration of the interaction between top-down and bottom-up models is done by Itti and Koch [2001b], and Rutishauser and Koch [2007]. In these models (see Figure 2) different tasks enforce different weight vectors on the combination phase.

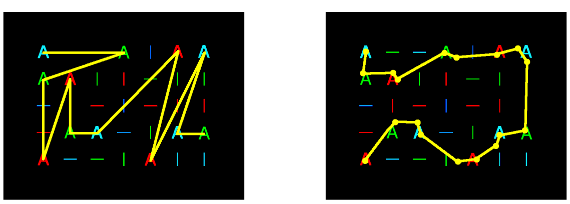
Overall, the model improves the bottom-up model by incorporating the task dependency and can be used to generate the likelihood term in equation 3 by the following equation:

$$\begin{aligned} P(\mathbf{Q}|y) &= P(q_1, q_2, \ldots, q_T|y) \\ &= P(q_1|y)P(q_2|y)\ldots P(q_T|y), \end{aligned} \tag{5}$$

As we can see in top-down models we still assume that the *naïve Bayes assumption* still holds and thus the observations are conditionally independent.

Although top-down models have somewhat addressed the problem of task independency of bottom-up models, they still suffer from some major shortcomings that disqualify them as practical solutions for obtaining the likelihood term.

One of the shortcomings of top-down models is a lingering problem from bottom-up models and arises from independence assumption in equation 4. If we assume that the consecutive fixations are independent from each other, the probability of fixating on a certain location in an 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 human visual psychophysics [Posner and Cohen, 1984]. For instance, in an early study by Engel [1971] it is indicated that in visual search task, the probability of detecting a target depends on its proximity to the location currently being fixated (*proximity preference*). Although Dorr et al. [2009] suggest that in a free viewing of a scene, low-level features at fixation contribute little to the choice of the next saccade, Koch and Ullman [1985] and Geiger and Lettvin [1986] 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 (generated by top-down models) and natural eye trajectories can best be demonstrated by comparing two trajectories from a recording of eye movements during a visual search task and an artificial trajectory produced by the top-down model (see Figure 3). As we can see the location of fixation points in Figure 3a (artificial) are sparse, while those of Figure 3b (natural) are more correlated to their predecessors'.



(a) (b)

Figure 3: Eye trajectories in visual search task. In these images the task is to count the number of "A"s. a) The trajectory produced by the top-down model. b) The trajectory obtained by recording eye movements while performing the task.

3 Sequential Analysis: Hidden Markov Models

An alternative approach to obtain the density function of the likelihood term in equation 3 is to allow for dependency between the attributes $q_1 \ldots q_T$. In an early study Hacisalihzade et al. [1992] used Markov processes to model 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[1] rather than crisscrossing it at random. In another study, Elhelw et al. [2008] also have successfully used a first-order, discrete-time, discrete-state-space Markov chain to model eye-movement dynamics.

Based on the observed dependencies between the elements of eye movement trajectories (i.e., $q_1, q_2, \ldots, q_T$) and successful application of Markov processes to model them, we propose to use first-order, discrete-time, discrete-state-space Markov chains to model the attention cognitive process of the human brain. Given this assumption the likelihood term will be:

$$\begin{aligned} P(Q|y) &= P(q_1, q_2, \ldots, q_T) \\ &= P(q_1|y)P(q_2|q_1, y)\ldots P(q_T|q_{T-1}, y). \end{aligned} \tag{6}$$

By choosing Markov chains as the underlying process model, we can develop a model that can estimate the likelihood term of $(P(\mathbf{Q}|y))$; while allowing for the dependencies between the consecutive fixations; and make inferences about the task (given the eye trajectories) by plugging the likelihood term into equation 2. However, 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[2].

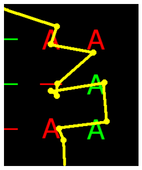
Eye position does not tell the whole story when it comes to tracking attention. While it is well known that there is a strong link between eye movements and attention [Rizzolatti *et al.*, 1994], the attentional focus is nevertheless frequently well away from the current eye position [Fischer and Weber, 1993]. Eye tracking methods may be appropriate when the subject is carrying out a task that requires foveation. However, these methods are of little use (and even counterproductive) when the subject is engaged in tasks requiring peripheral vigilance. Moreover, due to the noisy nature of the eye-tracking equipments, actual eye position itself is usually different from what the eye-tracker shows, which will introduce systematic error into the system.

In Figure 4 different eye trajectories of a subject executing the task of counting the number of "A"s in an image are demonstrated. As we can see, these two images illustrate different levels of linkage between COG and FOA. In Figure 4a, fixation points mainly land on the targets of interest (*overt attention*), whereas in the fast counting version of the same task with the same stimulus (Figure 4b), the COG does not necessarily follow the FOA and sometimes our *awareness* about a target does not imply foveation on that target (*covert attention*).

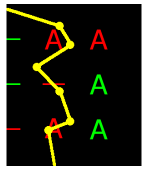
In real life, human attention usually deviates from the locus of fixation to give us knowledge about the parafoveal

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

[2] So far we have used the terms centre of gaze (COG) and focus of attention (FOA) interchangeably, but from now on, after explaining the difference between these two phenomena, we will distinguish between these two terms.



(a)



(b)

Figure 4: Eye trajectories recorded while executing a task given 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 "A"s is shown. Notice that the results from counting the characters were correct for both cases. Thus, the target that seems to be skipped over (the middle right "A") has been attended at some point.

and peripheral environment. This knowledge can help the brain to decide about the location of the next fixation that is most informative for building the internal representation of the scene. Therefore, this discrepancy between the focus of attention and fixation location helps us efficiently investigate a scene but at the same time makes the FOA covert and consequently hard to track. This off-target fixation can also be attributed to an accidental, attention-independent movement of eye, equipment bias, overshooting of the target or individual variability and will cause the occurrence of similar eye trajectories from multiple forward mappings from the task space (Yarbus) and therefore makes the inverse mapping from eye movement space to visual-task space (inverse Yarbus) an ill-posed problem.

Although, different tasks can generate similar eye movements, they always have their own targets of interest in an image and the covert attention will reveal these targets to us. Thus, if we could find a way to track the covert attention given the eye movements, we could regularize the ill-posed inverse problem and find a solution that can infer the task from the eye movements. The solution we are proposing to track the covert attention is to use *hidden Markov models* (HMMs). HMM is a statistical model based on Markov process in which the states are unobservable. In other words, HMMs model situations in which we receive a sequence of observations (that depend on a dynamic system), but we do not observe the state of the system itself. HMMs have been successfully applied in speech recognition [Rabiner, 1990], anomaly detection in video surveillance [Nair and Clark, 2002] and handwriting recognition [Hu *et al.*, 1996]. In another study Salvucci and Anderson [2001] have developed a method for automated analysis of eye movement trajectories in the task of equation solving by using HMMs.

A typical discrete-time, continuous HMM λ can be defined by a set of parameters $\lambda = \{A, B, \Pi\}$ where A is the *state transition probabilities* and governs the transitions between the states; B is the *observation pdf* and defines the parameters of the observation probability density function of each state; and Π is the *initial state distribution* and defines the probability of starting an observation from each of the states.

In the literature related to HMMs we can always find three major problems: evaluation, decoding and training. Assume we have HMM λ and a sequence of observations $\mathbf{O}$. Evaluation or scoring is the computation of the probability of observing the 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, Viterbi and Baum-Welch algorithm, respectively (details about the algorithms can be found in [Rabiner, 1990]).

In our proposed model, covert attention is represented by the hidden states of a task-dependent HMM λ_y (where λ_y is effectively equivalent to y in equation 2 in that they both designate different tasks and λ is merely an indication of using HMMs for modelling the tasks). Fixation locations, thus, will correspond to the observations of an HMM and can be used in training task-dependent models and evaluating the probability $P(\mathbf{O}|\lambda_y)$. By this interpretation of variables we can use a sequence of eye positions ($\mathbf{O}$) to represent the hidden sequence of covert attention locations ($\mathbf{Q}$) in the Bayesian inference and modify equation 2 to:

$$P(\lambda_y|\mathbf{O}) = \frac{P(\mathbf{O}|\lambda_y)P(\lambda_y)}{P(\mathbf{O})}. \quad (7)$$

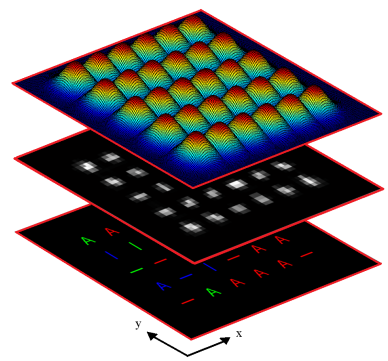


Figure 5: Observation pdfs give us the probability of seeing an observation given a hidden state. In this figure we have put the fixation location pdfs of all the targets together and superimposed them on the original image and its corresponding bottom-up saliency map.

In order to obtain the HMMs for each visual task, we need to train the parameters by using an eye movement database of the corresponding task. To do so, first we need to define the structure of our desired HMMs by creating a generic HMM. For the generic HMM we have assigned an ergodic or fully connected structure in which we can go to any state of the model in a single step, no matter what the current state of the model is. Since the salient areas are more likely to grab the subjects' attention and consequently redirect subjects' eye movements towards themselves, we have assigned one state to each conspicuous target in the bottom-up saliency map. Moreover, we postulate that in each state the observations are random outcomes of a 2-D Gaussian probability density function with features x and y in a Cartesian coordinates. In Figure 5 we have put the observation pdfs of all the states together and have superimposed them on the original image and its corresponding bottom-up saliency map. According to [Rabiner, 1990] 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 the nonzero value constraints). Also in order to reduce the training burden, we have used a technique called *parameter tieing* [Rabiner, 1990] to assign to each state a diagonal covariance matrix with equal variances in both x and y directions.

Having defined the general structure of the HMMs, we can obtain task-dependent HMMs by training the generic HMM with task-specific eye trajectories by using the Baum-Welch algorithm. The parameters to be trained are the initial state distribution (Π), state transition probabilities (A), mean and covariance of the observation pdfs ($\boldsymbol{\mu}$ and C).

After training the task-dependent HMM λ_y for each task, we can calculate the likelihood term $(P(\mathbf{O}|\lambda_y))$ by applying the parameters of λ_y to the forward algorithm. This way, we will be able to make inferences about the tasks given an eye trajectory by plugging the likelihood term in equation 7.

4 Experiments

In this section we will evaluate the performance of our proposed method in inferring the viewer's task. The inference is made by applying the Bayes rule (equation 7 for HMMs and equation 2 for the other methods) to the observation sequences of our database of task-dependent eye movements. Equal probabilities are used as the task priors which results in a maximum likelihood (ML) estimator. The inferences will tell us how well the cognitive models developed by different techniques can classify a task-dependent eye trajectory as its correct task category.

In order to perform the evaluation, we compare three different methods: HMM, sequential and top-down models. To build a database of task-dependent eye trajectories, we ran some 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 and displayed on a 1280$\times$800 pixel screen at a distance of 18 inches (1 degree of visual angle corresponds to 30 pixels, approximately). Each stimulus was composed of 35 objects each randomly selected from a set of nine objects (horizontal bar, vertical bar and character "A" in red, green and blue colors) that were placed at the nodes of an imaginary 5$\times$7 grid superimposed on a black background (see the lower layer of Figure 5). At the beginning of each trial, a message defined the visual task to be performed in the image.

The visual tasks were counting red bars, green bars, blue bars, horizontal bars, vertical bars or characters; hence six tasks in total. Each subject undergoes six segments of experiments, each of which comprises performing the six tasks on five stimuli resulting in 180 trials for each subject (1080 trials in total). In order to remove the memory effects, we have designed each segment so that the subject sees each stimulus just once in every five consecutive trials and the visual task changes after each trail. This segment is repeated for five more times to build the eye movement database of that subject.

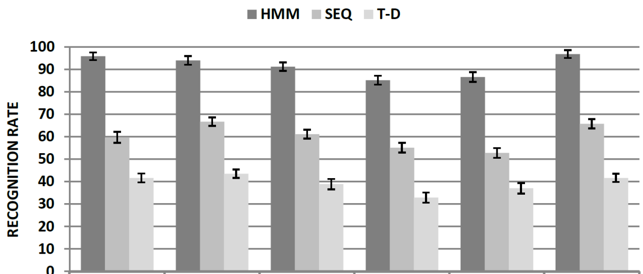
In training the HMMs good initial estimates for the mean and covariance matrices of the state observation matrices are helpful in skipping the local minima in training the task-dependent models. Since the salient areas are more likely to grab the subjects' attention and consequently redirect subjects' eye movements towards themselves, in the generic HMM, we set the initial value of the means equal to the centroids of the conspicuous targets on the bottom-up saliency map (Figure 5) averaged over all five stimuli. Then we use nearest neighbour 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. For the transition probabilities, we created the generic state transition distribution matrix A by giving a uniform distribution to all transitions from each state. Since we always started the experiment from the center point in the images, the initial state probability was assigned deterministically for all four methods.

Since Top-down and sequential methods posit that attention is always overt, we used nearest neighbour to find the closest target to each fixation location to represent it's attentional allocation. In order to develop the top-down method, we estimated the likelihood term of equation 5 for each observation sequence Q. To do so, we trained the terms $P(q_t = s_i|y)$ by calculating the frequency of seeing state i (s_i) in the training set of eye trajectories pertaining to task y. For the sequential model, we used the same training set to estimate the likelihood term of equation 6 by calculating the frequency of seeing different combinations (s_i, s_j) for each $i, j \in [1, N]$. In both top-down and sequential methods, we avoided zero probabilities for the states or state combinations not seen in the training data by increasing all the count numbers by one (smoothing).

Figure 6 demonstrates the accuracy of hidden Markovian, sequential, and top-down methods in inferring viewer's task. Each bar summarizes the accuracy of a model by representing the mean accuracy along with its standard error of the mean (SEM) in inferring the corresponding visual task by using the trained model. For each bar we have run a six-fold cross-validation on a dataset of 180 task specific eye trajectories to train/test the model and have compared the performance of different methods by drawing their corresponding bars around each visual task. As we can see, HMMs significantly outperform other methods in all six cases. This major improvement is due to the sequential nature of the HMMs and relaxing the constraint of overtness of attention in the model. The effect of the former can be highlighted by comparing sequential and top-down models and the latter accounts for the relatively huge gap between the accuracy of sequential models and that of the HMMs.

5 Conclusion

In this paper we presented a probabilistic framework for the problem of visual task inference and tracking covert attention. We have used Hidden Markov Models (HMMs) with Gaussian observation distributions to account for the disparity between the overt centre of gaze (COG) and covert focus of attention (FOA). The hidden states represent the covert FOA



		R	G	B	H	V	C
MEAN	HMM	95.83	93.98	91.20	85.18	86.57	96.76
	SEQ	59.73	66.68	61.12	55.10	52.78	65.75
	T-D	41.65	43.51	38.88	32.87	37.03	41.65
SEM	HMM	1.68	1.90	1.93	1.97	2.18	1.73
	SEQ	2.43	1.88	1.99	2.19	2.16	2.09
	T-D	1.94	1.91	2.30	2.25	2.31	1.82

TASKS

Figure 6: Comparison of the accuracy of visual task inference using HMM, Sequential (SEQ) and Top-Down (T-D) 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.

and the observations stand for overt COG. We have trained task-specific HMMs and used the Bayes rule to infer the visual task given the eye movement data. The data analysis conducted on our task dependent eye movement database shows that the idea of using HMMs surpasses the classic attention tracking methods in inferring the visual task. While the results presented in this report seem to be very promising, further investigations should be done to extend the idea to natural scenes and more realistic situations.

References

[Bulling *et al.*, 2009] 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.

[Castelhano *et al.*, 2009] 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.

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

[Dorr *et al.*, 2009] 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.

[Einhäuser *et al.*, 2008] 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.

[Elhelw *et al.*, 2008] 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.

[Engel, 1971] 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.

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

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

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

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

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

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

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

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

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

[Posner and Cohen, 1984] M.I. Posner and Y. Cohen. Components of visual orienting. *Attention and performance X: Control of language processes*, 32:531–556, 1984.

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

[Rensink *et al.*, 1997] R.A. Rensink, J.K. O'Regan, and J.J. Clark. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, pages 368–373, 1997.

[Rizzolatti *et al.*, 1994] G. Rizzolatti, L. Riggio, and B.M. Sheliga. Space and selective attention. *Attention and performance XV: Conscious and nonconscious information processing*, pages 231–265, 1994.

[Rutishauser and Koch, 2007] 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.

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

[Tatler, 2007] B.W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4, 2007.

[Yarbus, 1967] A.L. Yarbus. Eye movements during perception of complex objects. *Eye movements and vision*, 7:171–196, 1967.