Vision Research 103 (2014) 127–142

Contents lists available at ScienceDirect

Vision Research

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

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

Amin Haji-Abolhassani *, James J. Clark

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

ARTICLE INFO

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

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

ABSTRACT

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

1. Introduction

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

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

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

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

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

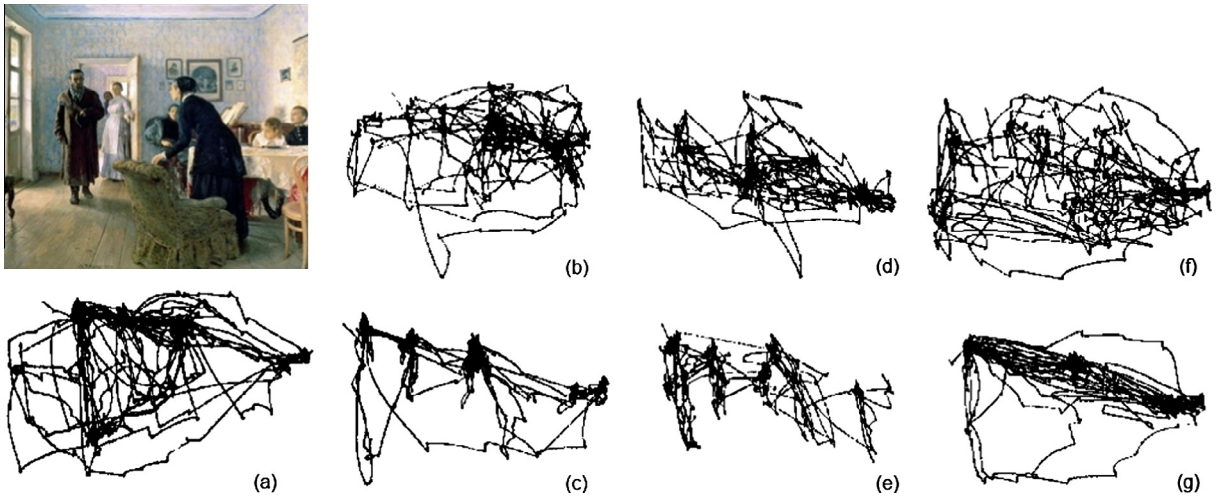

Fig. 1. Eye trajectories measured by Yarbus by viewers carrying out different tasks. (a) No specific task. (b) Estimate the wealth of the family. (c) Give the ages of the people in the painting. (d) Summarize what the family had been doing before the arrival of the "unexpected visitor". (e) Remember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

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

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

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

1.1. Attention modeling

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

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

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

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