Neurocomputing 168 (2015) 917–929

Contents lists available at ScienceDirect

Neurocomputing

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

On semantic-instructed attention: From video eye-tracking dataset to memory-guided probabilistic saliency model

Yan Hua [a], Meng Yang [b], Zhicheng Zhao [a,*], Renlai Zhou [b], Anni Cai [a]

[a] Beijing University of Posts and Telecommunications, Beijing, China
[b] Beijing Normal University, Beijing, China

ARTICLE INFO

Article history:
Received 17 September 2014
Received in revised form
25 April 2015
Accepted 9 May 2015
Communicated by Yongdong Zhang
Available online 19 May 2015

Keywords:
Semantic-instructed viewing
Eye-tracking
Top-down attention
Memory
Saliency model

ABSTRACT

Visual attention influenced by example images and predefined targets are widely studied in both cognitive and computer vision fields. Nevertheless, semantics, known to be related to high-level human perception, have a great influence on top-down attention process. Understanding the impact of semantics on visual attention is beneficial for providing psychological and computational guidance on many real-world applications, e.g., semantic video retrieval. In this paper, we intend to study the mechanisms of attention control and computational modeling of saliency detection for dynamic scenes under semantic-instructed viewing conditions. We start our study by establishing a dataset REMoT, the first video eye-tracking dataset with semantic instructions to our best knowledge. We collect the fixation locations of subjects when they are given four kinds of instructions with different levels of noise. The fixation behavior analysis on REMoT shows that the process of semantic-instructed attention can be explained with long-term memory and short-term memory of human visual system. Inspired by this finding, we propose a memory-guided probabilistic model to exploit the semantic-instructed top-down attention. The experience of attention distribution to similar scenes in long-term memory is simulated by linear mapping of global scene features. An HMM-like conditional probabilistic chain is constructed to model the dynamic fixation patterns among neighboring frames in short-term memory. Then, a generative saliency model is constructed which probabilistically combines the top-down and a bottom-up modules for semantic-instructed saliency detection. We compare our model to state-of-the-art models on REMoT and a widely used dataset RSD. Experimental results show that our model achieves significant improvements not only in predicting visual attention under correct instructions, but also in detecting saliency for free viewing.

1. Introduction

Although receiving a large amount of 10^8–10^9 bits every second from eyes [1], human visual system (HVS) is capable of rapidly catching the prime information by selectively paying attention to regions with salient patterns. This visual attention mechanism plays a key role in visual processing tasks such as object recognition, semantic extraction and scene analysis. Nevertheless, the attention mechanism has not been fully understood due to the agnosticism on the collaboration of different parts in human brain. How to simulate such a biological cognitive process with computational models is even more challenging for practitioners in computer vision.

Generally, visual attention is induced by two interrelated and indivisible information processing stages, bottom-up and top-down. Bottom-up is an involuntary attention that is only attracted by physical features of stimulus. Most studies on bottom-up attention are focused on identifying the physical features that may attract attention [2–6]. The achievements in bottom-up studies have been widely utilized to extract saliency visual regions for computer vision tasks, such as image segmentation [7], object recognition [8] and scene classification [9]. Top-down is a voluntary attention processing, which requires active awareness and directs intentional searching in our daily life [10]. There are many cognitive functions driving top-down attention, such as knowledge, memory, expectation, reward and current goal [11]. The mechanisms of top-down attention still need to be further investigated.

Existing studies on top-down control of visual attention mainly focus on guiding with visual aspect of objects and scenes [10,12–15]. For example, previous inspected visual item shortens the searching time [10]. How one's attention affected by simple semantic instructions, e.g., a target object name or its simple description, has also been investigated [16–18]. Cognitive experiments show that targets and their semantically similar regions are

* Corresponding author.
E-mail addresses: huayan@bupt.edu.cn (Y. Hua), pandamengbnu@gmail.com (M. Yang), zhaozc@bupt.edu.cn (Z. Zhao), rlzhou@bnu.edu.cn (R. Zhou), annicai@bupt.edu.cn (A. Cai).

http://dx.doi.org/10.1016/j.neucom.2015.05.033
0925-2312/

recalled more often, thus they can be recognized more accurately than unrelated distractors. However, in practical applications users would prefer to provide complicated textual query inputs, e.g., a sentence or multiple keywords, for a visual search task. Unfortunately, previous work on visual examples and simple semantic instructions cannot be directly transferred to explain visual attention mechanisms under such situations. To facilitate the needs of practical applications such as semantic video retrieval, it is necessary to study the deployment as well as computational modeling of visual selective attention given complex instructions.

In this paper, we make the first attempt to study mechanisms of attention control and modeling of saliency detection for dynamic scenes under complex semantic-instructed viewing conditions. The *semantic-instructed viewing* means that subjects are given complex semantic instructions before they watch a video clip. For example, users may be asked to judge whether a video is about "a boy wearing red coat playing basketball on the playground". The complex instructions in our study are a set of sentences describing events that may be composed of subjects, actions, objects, scenes, attributes and other descriptions. Compared to simple semantic instructions, complex instructions deliver richer semantics which can more precisely describe the user's intention. To further investigate how attention mechanisms are affected by different levels of noise in complex semantic instructions, participants in our studies will be given either no instruction (free viewing), or an instruction that may be correct, partially correct or totally incorrect before viewing a video. We provide diversified types of semantic instructions that well simulate the conditions of real world semantic video search.

In HVS, selective visual attention is exploited by actively controlling gaze to direct eye fixation towards informative regions in real time. Fixation maps of eye-movement detected and recorded by eye tracker in cognitive experiments are essential for analyzing visual attention patterns. Therefore, we start our study by establishing a new eye-tracking dataset REMoT, in which subjects' fixations are recorded under given complex semantic instructions. We then analyze differences of fixation positions between different kinds of semantic instructions and compare fixations with target locations based on this new eye-tracking dataset. By the experiments and analyses, we find the positive signals of visual attention influenced by complex semantic instructions. More specifically, the fixation patterns obtained in our experiments can be partially explained with declarative memory, procedural memory and short-term memory of HVS.

The declarative memory is a type of long-term memory [19]. What stored in declarative memory is the knowledge and facts about concerned objects and their representations. It has been proved that declarative memory is involved in single target search task in cognitive studies [20]. Our experimental results also show that subjects identify the target regions more accurately and more quickly with correct instructions, indicating the guidance of declarative long-term memory on attention towards targets. The procedural memory is another kind of long-term memory for performing particular type of actions [19]. Behavioral analysis [21] indicates that the effects of memory-guided spatial orienting concur with the intuition that in everyday life our predictions about where relevant or interesting events will be unfolded are largely shaped by our memory experience. Our analyses show that the average fixation patterns under different instructed-viewing conditions trend to be similar, implying the impact of procedural long-term memory on attention in visual search actions. The short-term memory has been demonstrated [22] by the fact that subjects were able to report whether a change had occurred when they viewed two image patterns separated by a short temporal interval. Experiments also demonstrate that in the cognition process [10], visual search is more efficient when the memory cue matches the visual item containing the target. Our analyses on REMoT show that fixations shift from target regions to other regions along with the time after target objects appearing, depicting the influence of short-term memory on attention dynamics.

Inspired by the above findings on REMoT, we propose a memory-guided probabilistic saliency model in this paper. There has been a huge body of research work on computational modeling of visual attention from bottom-up [2,23–28] and top-down [29–34] perspectives. Bottom-up models define saliency mainly based on bio-inspired measurements on low-level image features. Top-down models generally detect salient regions with different learning methods. There have been some top-down methods which aim to saliency detection, but specifically for particular tasks such as video game playing [30,34]. In addition, there have been a number of dynamic models [31,32,23] which integrate the motion feature along with other static low-level features to estimate the saliency on a video frame. Unfortunately, they all lack the mechanisms for memory-guided saliency which has been proved important by analysis on REMoT for semantic-instructed attention. Previous psychology study [35] also supports that a robust memory of visual content guides the spatial attention.

The proposed model is a three-layer generative graphical model which combines a top-down module and a bottom-up module. The former module is procedural memory-driven, while the latter is visual feature-activated for static scenes (i.e., each video key-frame). We employ gist of a scene [36] to define the global scene context, and the relationship between gist and semantic-instructed saliency is learned to simulate the content stored in procedural long-term memory. To model the short-term memory of dynamic attention, we construct an HMM-like conditional probabilistic chain on the saliency maps of a video key-frame sequence. It learns the diversified saliency patterns from true fixation data collected from the semantic-instructed psychological experiments. By incorporating both top-down and bottom-up cues, our model outperforms existing saliency models on both RSD [37], a widely used dataset with manually labeled salient objects, and our REMoT dataset with different types of instructions.

The contributions of our work are summarized as follows:

- We undertake a significant effort on establishing a new eye-tracking dataset REMoT and making it available to the research community. In REMoT, the fixations are recorded by eye tracker under free viewing and three kinds of semantic instructions (correct, incorrect and partially correct). To our best knowledge, this is the first human eye-tracking dataset with semantic instructions on real-world videos.
- Detailed analyses on fixation locations of the recorded data are performed under different kinds of viewing instructions. The analyses show the evidences of attention affected by long-term and short-term memories of HVS.
- Inspired by our analyses, we introduce long-term memory and short-term memory as the top-down cues into attention modeling for semantic-instructed viewing. A three-layer generative graphical model is proposed to combine bottom-up and top-down cues. Our proposed model achieves a significant improvement in predicting saliency compared to state-of-the-art models on both REMoT and RSD.

Roadmap: Section 2 gives a brief review on related work. Section 3 describes the collection of eye-tracking dataset REMoT. Section 4 introduces the analyses of fixation patterns. Section 5 presents the memory-guided probabilistic saliency model. Section 6 demonstrates the experimental results of saliency models. Section 7 gives the conclusion and future work.