Computer Vision and Image Understanding 131 (2015) 28–41

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Discovering salient regions on 3D photo-textured maps: Crowdsourcing interaction data from multitouch smartphones and tablets

Matthew Johnson-Roberson [a,*], Mitch Bryson [c], Bertrand Douillard [b], Oscar Pizarro [c], Stefan B. Williams [c]

[a] Naval Architecture & Marine Engineering, University of Michigan, Ann Arbor, MI, USA
[b] NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
[c] The Australian Centre for Field Robotics, The University of Sydney, NSW, Australia

ARTICLE INFO

Article history:
Received 3 December 2013
Accepted 15 July 2014

Keywords:
Crowdsourcing
Visual saliency
3D maps
Multitouch interaction
HMM
Gaze-tracking
Smartphones
Mobile devices

ABSTRACT

This paper presents a system for crowdsourcing saliency interest points for 3D photo-textured maps rendered on smartphones and tablets. An app was created that is capable of interactively rendering 3D reconstructions gathered with an Autonomous Underwater Vehicle. Through hundreds of thousands of logged user interactions with the models we attempt to data-mine salient interest points. To this end we propose two models for calculating saliency from human interaction with the data. The first uses the view frustum of the camera to track the amount of time points are on screen. The second uses the velocity of the camera as an indicator of saliency and uses a Hidden Markov model to learn the classification of salient and non-salient points. To provide a comparison to existing techniques several traditional visual saliency approaches are applied to orthographic views of the models' photo-texturing. The results of all approaches are validated with human attention ground truth gathered using a remote gaze-tracking system that recorded the locations of the person's attention while exploring the models.

1. Introduction

We have developed a smartphone/tablet app for the viewing and manipulation of 3D models gathered with an Autonomous Underwater Vehicle (AUV). This app is freely available and has been downloaded and used by a large number of users. The question this paper is attempting to answer is "Can we employ crowdsourcing to perform salient interest point detection from users not specifically tasked to find these points?" A diagram depicting the high-level system presented in this work is shown in Fig. 1.

Saliency, particularly visual saliency is a popular construct from the field of biological vision and broadly describes an organisms ability to focus attention on a subset of its sensory input for further processing. In this work data subsetting is the most relevant part of the visual saliency process. While scientists and non-experts will have differing opinions on the high level top-down definitions of saliency, rapid bottom-up visual saliency is much less task and operator dependent [49]. This paper is focused on such processing in the context of a long-term environment-monitoring program using AUVs. At the Australian Centre for Field Robotics there is an ongoing program to perform benthic monitoring with an AUV [73]. This program deploys an AUV in unstructured natural environments where it gathers data for human review. One of the major bottlenecks in this process is the vast amount of data gathered by the AUV. The AUV is capable of gathering orders of magnitude more data than previous techniques. Traditionally divers used hand-held cameras to gather visual data in underwater environments and issues of decompression, airtime, and safety severely limited the quantity of data that scientists could gather. With the AUV in its current configuration, monitoring images can be gathered at up to 4 Hz. A typical field campaign lasting two weeks can result in hundreds of thousands of images requiring review.

The challenge of how to deal with this massive image archive is being explored on several fronts. A large effort has gone into unsupervised clustering [64], human hand labeling [46], and supervised classification [4]. This work presents an alternative for gathering large amounts of human review data quickly and inexpensively. The assertion we present in this paper is that human visual saliency can be modeled by proxy through the exploratory motions of a large number of users in a 3-D environment.

Capturing human curiosity and exploration for robotic platforms is non-trivial. The well established approach is to use visual saliency measures but it is not necessarily clear that they can predict what people find interesting in a 3D scene and how they will choose to interact with it. This paper presents two alternative

* Corresponding author.
E-mail addresses: mattjr@umich.edu (M. Johnson-Roberson), m.bryson@acfr.usyd.edu.au (M. Bryson), Bertrand.Douillard@jpl.nasa.gov (B. Douillard), o.pizarro@acfr.usyd.edu.au (O. Pizarro), stefanw@acfr.usyd.edu.au (S.B. Williams).

http://dx.doi.org/10.1016/j.cviu.2014.07.006
1077-3142/

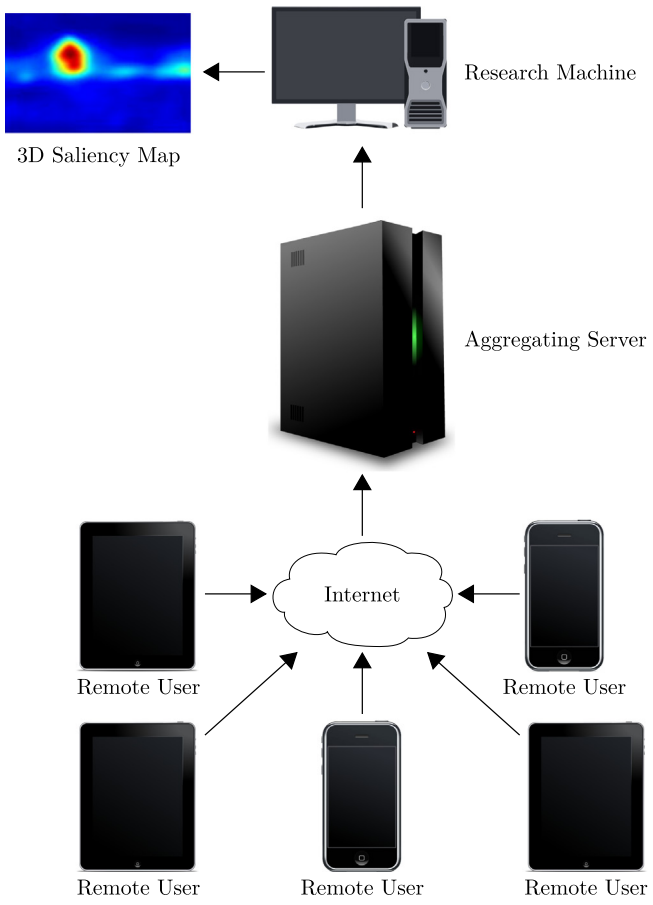

Fig. 1. Diagram of network architecture for crowdsourcing saliency.

measures of human interest both based on the motion of the viewpoint used by the operator and compares them to traditional saliency measures. Through the crowdsourcing of many remote smart phone/tablet users we gather data to perform the identification of visual saliency on 3-D photo mosaic maps. Human experiments with ground truth from eye tracking are used to validate our results.

1.1. Crowdsourcing as a model for problem solving

Crowdsourcing has emerged as a successful model for solving tasks by leveraging the human intelligence of large groups of remote users in a distributed fashion. The term crowdsourcing was coined in 2006 and first appeared in scientific literature in 2008 as "an online, distributed problem-solving and production model" [6]. The crowdsourcing model has since been adapted to outsource difficult steps in many computational tasks [32]. Recently the computer vision community has begun using crowdsourcing to solve challenging vision problems.

In parallel, researchers have started harnessing the power of data-mining over massive user bases to answer many new questions. Search engines and social networking use the interaction from millions of users to refine and improve advertising and site usability [17]. Researchers have used this data to learn about the demographics of users, social trends, and behavioral patterns [41,65]. With the rise of smart phones and 'app stores' mobile platforms have quickly become a practical means of gathering massive amounts of user data. App analytics is attempting to turn the millions of smart phones in use into a distributed network of data sources.

Traditional crowdsourcing of vision tasks relies upon motivating users through community good will [59], financial incentives (Mechanical Turk) or competitive/entertainment incentives [1,2] by turning a task into a game. The intended motivation for users of our app was education and entertainment. The app was advertised in the education section of the Apple iTunes app store and in its description and screenshots offered the promise of exploration of images from the deep sea. We attempted to capitalize on public interest in science, especially exploratory science, to motivate downloads. A novel aspect of our approach is that the motivation of users was somewhat more decoupled from the task than in a traditional crowdsourcing model. To work with such user data we propose the use of a novel paradigm from big data analytics where the answers to questions can be inferred from the data of many users. The power of our data-mining approach to crowdsourcing is that data is collected from a much larger pool of users. A full discussion of the motivations and demographics of users on various crowdsourcing platforms is beyond the scope of this paper, however Kaufmann et al. [30] present a review of the studies on Mechanical Turk. While these studies reflect a diverse user pool they also show that the Mechanical Turk user base is a fraction the size of the potential smart phone app user pool [31,56]. Using the smart phone platform gives us access to a much more general audience. To further the general appeal of the app we do not ask users to explicitly identify things they find interesting. Rather, we attempt to infer interest from patterns of interaction and in doing so free the user from an artificially constrained task. Without asking users to answer a specific question, their motivations for participating can be much more varied. This potentially gives access to a much larger 'crowd'.

1.2. Paper layout

We will be presenting two novel metrics to calculate saliency from human user interaction data. One employs the use of the camera's frustum to histogram observed points, while the other leverages a Hidden Markov model (HMM) to classify interaction data spatially into a saliency map. These techniques are compared to several state-of-the-art visual saliency techniques and validated using human gaze tracking data. The paper is laid out as follows. Section 2 discusses prior work. Section 3 presents the developed app as a platform for crowdsourcing. In Section 4 the two interaction-based formulations for saliency are laid out. The human trials for validation are discussed in Section 5. Results are presented in Section 6 and finally Section 7 concludes and presents future work.

2. Prior work

2.1. Crowdsourced vision

Tools such as LabelMe, ImageNet, BUBL and other systems which leverage Amazon's Mechanical Turk have provided solutions to the problem of image-labeling using human computation [18,14,33]. Mechanical Turk has become a particularly popular platform for crowdsourcing for vision. It offers flexibility and there has been research into assessing, processing, and rectifying image labelings from large groups of human sources [63,71,55]. All the aforementioned systems deal with image annotation with various types of semantic information ranging from object identification, object classification, and object segmentation.

In the field of gaze tracking Rudoy et al. propose a relevant model of crowdsourcing gaze tracking. They project a pattern over video or image data. Then a 'crowd' of remote users enter the subsection of the pattern viewed providing a proxy for direct gaze tracking [58].