EYE TRACKING BASED SALIENCY FOR AUTOMATIC CONTENT AWARE IMAGE PROCESSING

Steven Scher, Joshua Gaunt**, Bruce Bridgeman**, Sriram Swaminarayan***,James Davis**

*University of California Santa Cruz, Computer Science Department
**University of California Santa Cruz, Psychology Department
***Los Alamos National Laboratory, CCS-2

ABSTRACT

Photography provides tangible and visceral mementos of important experiences. Recent research in content-aware image processing to automatically improve photos relies heavily on automatically identifying salient areas in images. While automatic saliency estimation has achieved estimable success, it will always face inherent challenges. Tracking the photographer's eyes allows a direct, passive means to estimate scene saliency. We show that saliency estimation is sometimes an ill-posed posed problem for automatic algorithms, made well-posed by the availability of recorded eye tracks. We instrument several content-aware image processing algorithms with eye track based saliency estimation, producing photos that accentuate the parts of the image originally viewed.

Index Terms— Eye Tracking, Saliency, Computational Photography, Content Aware Resizing, Seam Carving

1. INTRODUCTION

Photos and videos are a powerful medium for capturing a moment's fleeting experience and later sharing it with others. The best photography does not merely faithfully document the scene in front of the camera. Rather, the photographer uses various artifices to influence the viewer's perception of the scene, directing the viewer to notice certain aspects of the image. This ability is often reserved only for professional photographers, and achieved at the time of image capture through framing, exposure, and focus, or aferward with image editing software.

A casual photographer, while wishing to preserve 'what they noticed,' typically settles for simply recording an accurate portrait of what is in front of them. Recent research in content-aware image processing has dramatically improved the ability of the amateur photographer to apply software that automatically or semi-automatically modifies their photo to accentuate some region of the photo.

Many such algorithms rely crucially on an estimated saliency map of the image: which regions are important, and which are not? Automatic saliency estimation faces two important challenges. First, determining important and unimportant regions of some photos requires high-level scene analysis beyond current capabilities. Second, objective saliency may be elusive when two photographers disagree as to the salient parts of the same scene. The two photographers may have different motives in taking their pictures, or differing knowledge of the semantic content scene.

While objective saliency may sometimes be ill-posed, personal saliency is not. We propose to record the photographer's eye movements to identify the parts of the scene they notice, and to later manipulate the image in order to draw viewers' eyes to those same regions. Photographs of the same object, taken from the same place, with the same camera, should differ depending on the photographer, and what caught *their* eye.

We show that automatic saliency algorithms can fail to account for semantic scene content, where eye tracking supplies useful saliency maps. We further apply content-aware image processing algorithms using saliency maps derived from eye tracking.

We believe that the ability to record photographer's eye movements is within reach of camera manufacturers, noting that Canon included an "Eye Controlled Focus" option in several film-based SLR cameras from 1992 to 2004: an eye-tracker built into the viewfinder directed the camera's autofocus. To our knowledge, however, no camera has recorded these eyetracks along with the photo. We hope this work inspires manufacturers to do so in the future.

The primary contribution of this paper is the demonstration that eyetrack data may be used to esimate image saliency for content-aware image processing algorithms that emphasize those parts of the scene that most struck the viewer's eye.

2. RELATED WORK

Unfortunately, eye tracking has recieved little attention with regard to saliency estimation in content aware image processing.

Santella et al [1] created a user interface allowing a computer user to semi-automatically crop an image by record-

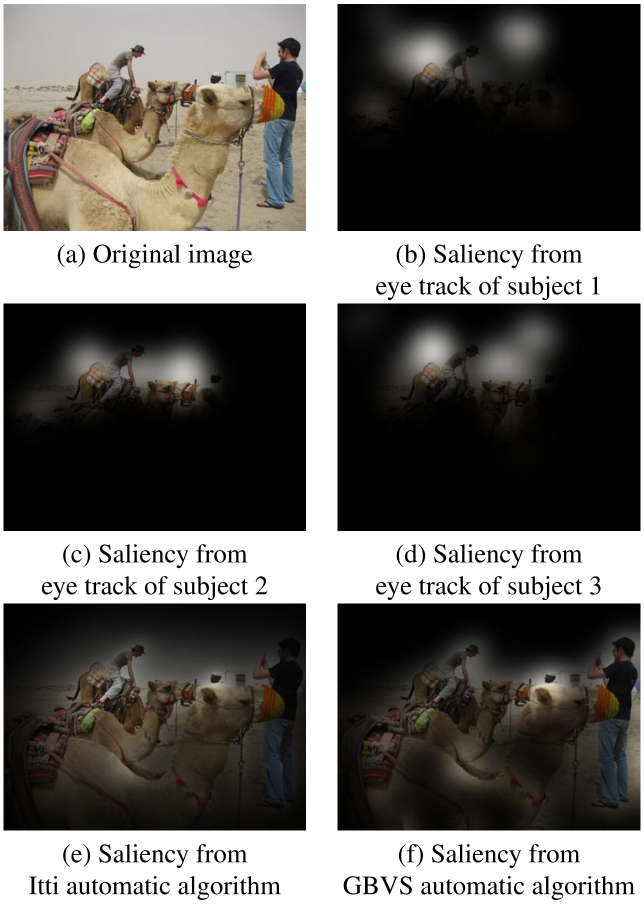
corresponding author: Steven Scher, sscher@ucsc.edu

ing their eye tracks while using image editing software. Our intended application targets photographers at image capture time, and considers several content aware image procesing techniques rather than cropping.

In another line of research, Santella et. al. [2], [3],[4] strive toward an artistic goal, seeking to automate the creation of stylized cartoons. Conversely, we seek to preserve the appearance of an authentic image while redirecting a new viewer's eye to match. Like us though, they use eyetracks of individuals to identify regions of interest in images, and use this information to modify the image.

Several content aware image processing techniques may be used to direct a viewer's attention in an image. For example, the brightness, contrast, and color saturation may be selectively diminished or enhanced, or the image may be cropped the image to limit the viewer's attention to the areas desired. More flexible tools of recent interest are content-aware resizing algorithms, such as Seam Carving [5] or related methods [6] [7] [8] [9] [10] that selectively enlarge or shrink different regions of the image.

Content-aware resizing has received extensive attention since the Seam Carving paper of 2007. Most work focuses on one of two distinct challenges. First, a saliency map must be constructed to determine which parts of the image should be emphasized, and which de-emphasized or removed. Second, and separately, the image is nonuniformly resampled to remove those image regions deemed least important, leaving the important regions behind. This paper responds to the first challenge. While typical automatic methods find strong edges or high-frequency content [11],[12], [5], passively-collected eye tracks allow a new answer. What does the photographer want the viewer to see? What the photographer saw.



(a) Original image

(b) Saliency from eye track of subject 1

(c) Saliency from eye track of subject 2

(d) Saliency from eye track of subject 3

(e) Saliency from Itti automatic algorithm

(f) Saliency from GBVS automatic algorithm

Fig. 1. Saliency of an image is estimated from recorded eye tracks, and from two automatic saliency estimation algorithms. Note that the automatic algorithms find most of the image salient, while all three subjects' eyes concentrate on the camel's rider.

3. SALIENCY

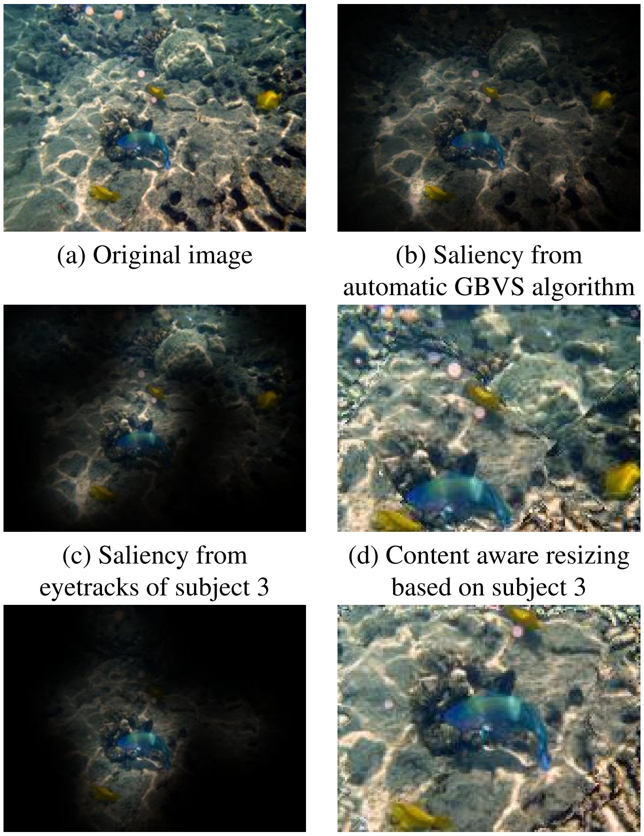
Tracking a photographer's eye movements allows the consruction of a saliency map indicating the parts of the scene most noticed. Looking ahead, we expect that future cameras will soon be equipped with eye trackers built directly into their viewfinders. Our present experiment, however, was conducted with off-the-shelf equipment in a laboratory setting. Rather than a camera's viewfinder, subjects peered through a half-mirror to see a computer monitor while a Bouis infrared eye tracker recorded their eye movements. Before viewing a photo, the subject viewed a sequence of 25 calibration images consisting of points on a 5x5 grid. This calibration typically provided an accuracy of 20-50 pixels on an 800x600 screen, with an accompanying accuracy estimate for each session.

We sample eye gaze directions at 1kHz and estimate the average time spent looking at each pixel by convolving with a gaussian filter that spreads the contribution of each measurement over an area matched to the accuracy of the measurement. Santella et al [1] used a more sophisticated methodology to better segment complex objects from their backgrounds, but we have found our simple technique sufficient for the tasks at hand.

We compare the observed saliency maps to two automatic methods. The 'Itti' algorithm [12] begins by applying a filter bank to the image. These filter responses are then normalized and averaged.

The Graph Based Visual Saliency (GBVS) algorithm [11] constructs a fully-connected graph with a node for each pixel, with directed edges weighted according to the dissimilarity between the pixels' responses to filters and their distance. The stationary distribution is obtained through the power method to find 'interesting' pixels. A new graph is then constructed, also with a node for each pixel, with connections only between neighboring nodes, and weighted by the similarity of their interestingness (as found by the first graph). The power method is again used to find the stationary distribution, concentrating the mass into localized regions. The authors [11] have kindly provided implementations of the GBVS and Itti algorithms.

Figure 1 compares the GBVS and Itti algorithms to saliency maps derived from recorded eye tracks. Note that in this case all three subjects' recorded eye tracks focus on the person riding the camel, while both saliency algorithms distributed their attention over a large region of the photo. The visual cues that make the camel's rider so interesting to human viewers are high level semantic cues difficult for any automatic saliency algorithm to identify.



(a) Original image

(b) Saliency from automatic GBVS algorithm

(c) Saliency from eyetracks of subject 3

(d) Content aware resizing based on subject 3

(e) Saliency from eyetracks of subject 4

(f) Content aware resizing based on subject 4

Fig. 2. Saliency maps derived from eyetracks of two subjects distinctly differ, and the result of content aware resizing thus differs as well. In this case, the automatic saliency algorithm finds most of the image to be salient.

4. CONTENT AWARE IMAGE PROCESSING

Content aware image resizing distorts the sizes of different parts of an image, enlarging or shrinking some more than others in order to emphasize salient regions. Differing saliency maps will emphasize different areas in the resulting image. In the popular *seam carving* algorithm, a subset of pixels in the original image is chosen to appear in the resulting image. To achieve this, the original image is iteratively shrunk by one row or one column. Rather than an intact column, a seam is removed - a set of pixels that are all diagonally or vertically adjacent, with one pixel from each row. The seam is chosen to preserve the parts of the image weighted highly by the saliency map and remove the parts given low weight.

Attention can also be drawn to one part of an image by selectively defocusing other parts. This effect is commonly used by photographers when capturing photos, by using a shallow depth of field to keep their subject in focus while other objects are out of focus. A similar effect can be achieved after image capture by blurring some parts of the image with a gaussian filter. We applied a different level of gaussian blur at each pixel, with the kernel's width smaller for more salient pixels.

We now compare saliency maps from viewers with distinct ideas of what in a scene is salient. In the previous section, the three human subjects showed remarkable agreement in Figure 1 that the camel rider was the most interesting part of the photo. In contrast, the subject in Figure 3(c) attended to each of the fish and a rock, while the subject in Figure 3(e) concentrated only on the large blue fish. What is ”interesting” varies from person to person. This difference in judged saliency leads to two very different seam carved results. Figure 3(d) includes all four fish and regions from the top of the photo, while 3(f) centers tightly around the blue fish. The GBVS algorithm's saliency in Figure 3(b), meanwhile, encompasses a large part of the image.

Consider the scene of four ultimate frisbee players in Figure 3. While many viewers will find the players more salient than the background, viewers will disagree as to whether some players are more important to the photo than others. To demonstrate the ability of selective defocus to capture the photographer's experience, a subject was asked to look at each of four players in the photo, in turn. Their eye tracks were recorded, giving four separate saliency masks, and four selectively defocused images. Each leaves a different player in focus while the rest of the image is slightly blurred.

5. CONCLUSION

Content-aware image processing provides exciting and useful tools to photographers, and depends crucially on estimating image saliency. We have demostrated that passively tracking the eyes of photographers would provide personalized saliency maps for use in such algorithms.

6. ACKNOWLEDGEMENTS

We would like to thank LANL ISSDM and NSF #CCF-0746690 for funding this work.

7. REFERENCES

[1] Anthony Santella, Maneesh Agrawala, Doug Decarlo, David Salesin, and Michael Cohen, “Gaze-based interaction for semi-automatic photo cropping,” in *CHI '06:*

Proceedings of the SIGCHI confernce on Human Factors in computing systems, 2006, pp. 771–780.

[2] Anthony Santella and Doug DeCarlo, “Abstracted painterly renderings using eye-tracking data,” *Non-Photorealistic Animation and Rendering 2002)*, pp. 75–82, 2002.

[3] Anthony Santella and Doug DeCarlo, “Stylization and abstraction of photographs,” *ACM Transactions on Graphics, (Proceedings SIGGRAPH 2002)*, pp. 769–776, 2002.

[4] Anthony Santella and Doug DeCarlo, “Visual interest and npr an evaluation and manifesto,” *Non-Photorealistic Animation and Rendering 2004*, pp. 71–78, 2004.

[5] Shai Avidan and Ariel Shamir, “Seam carving for content-aware image resizing,” *ACM Transactions on Graphics, (Proceedings SIGGRAPH 2007)*, vol. 26, no. 3, 2007.

[6] Michael Rubinstein, Ariel Shamir, and Shai Avidan, “Improved seam carving for video retargeting,” *ACM Transactions on Graphics, (Proceedings SIGGRAPH 2008)*, vol. 27, no. 3, 2008.

[7] Ariel Shamir and Shai Avidan, “Seam carving for media retargeting,” *Commun. ACM*, vol. 52, no. 1, pp. 77–85, 2009.

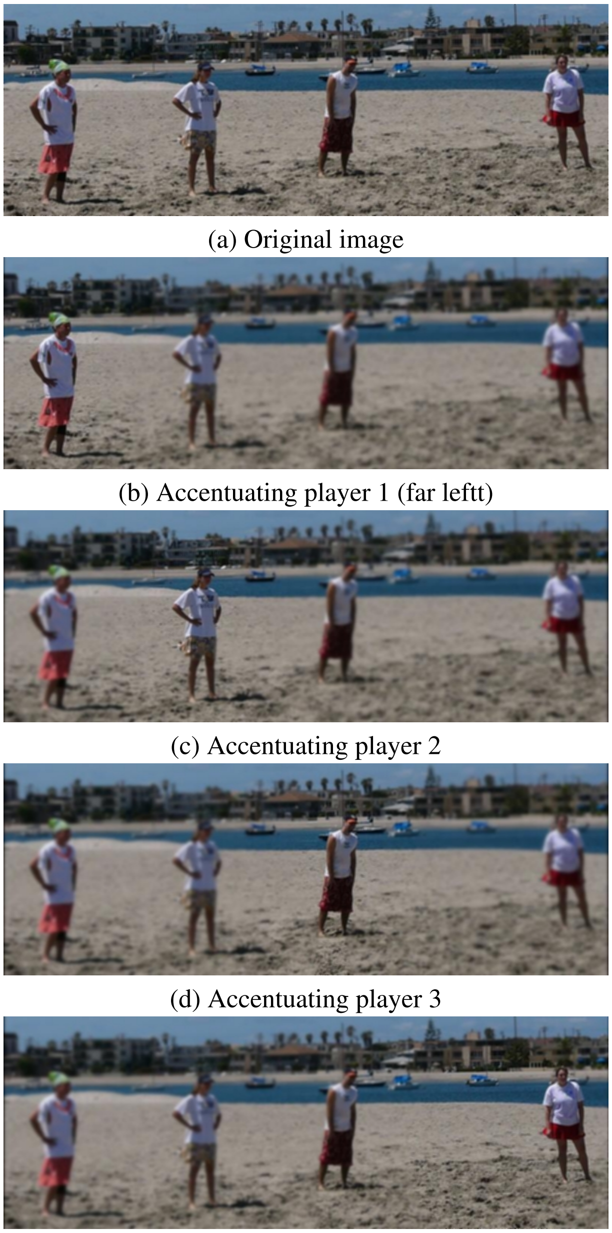
[8] Michael Rubinstein, Ariel Shamir, and Shai Avidan, “Multi-operator media retargeting,” *ACM Transactions on Graphics, (Proceedings SIGGRAPH 2009)*, vol. 28, no. 3, 2009.

[9] Lior Wolf, Moshe Guttmann, and Daniel Cohen-Or, “Non-homogeneous content-driven video-retargeting,” in *Proceedings of the Eleventh IEEE International Conference on Computer Vision (ICCV-07)*, 2007.

[10] Vidya Setlur, Ramesh Raskar, Saeko Takagi, Michael Gleicher, and Bruce Gooch, “Automatic image retargeting,” in *In In the Mobile and Ubiquitous Multimedia (MUM), ACM*. 2005, Press.

[11] J. Harel, C. Koch, and P. Perona, “Graph based visual saliency,” *Proceedings of Neural Information Processing Systems (NIPS)*, 2006.

[12] L. Itti, C. Koch, and E. Niebur, “A model of saliency based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine*, 1998.



(a) Original image

(b) Accentuating player 1 (far leftt)

(c) Accentuating player 2

(d) Accentuating player 3

(e) Accentuating player 4 (far right)

Fig. 3. Viewers may disagree with regard to the salient parts of an image. This image contains four players, any or all of whom may be salient, depending on the viewer. To simulate this, a subject was asked to look at each of the four people in the photo, in turn. Eye movements during each of those glances were recorded separately, and were used to render four different images, each drawing attention to one person by selectively defocusing the non-salient regions.