Labeling images by interpretation from Natural Viewing

Karen Guo
Department of Computer Science
University of Minnesota
guoxx431@umn.edu

Danielle N Pratt
Department of Psychology
University of Minnesota
pratt308@umn.edu

Angus MacDonald III
Department of Psychology
University of Minnesota
angus@umn.edu

Paul R Schrater
Department of Computer Science
University of Minnesota
schrater@umn.edu

ABSTRACT

In this paper, we would like to discuss the connection between visual processing and the understanding of an image. While the information of image viewing can be obtained from subjects' eye fixation, the understanding of an image can be obtained from the subjects' description of the given image. Furthermore, we proposed a new image labeling method based on the connection between eye fixation and image description by humans. By generating this new kind of labeling method, we can construct an image dataset with labels that are closer to how humans understand the incoming image. In addition, we would like to discuss the proof that the proposed labels better describe the image compared to other types of labeling systems.

Research about the relationship between images and human descriptions can be applied to several different applications. For instance, by analyzing the pairwise similarity of user descriptions, we could have a measurement of the complexity of image content. Another possible application is to use this dataset as a criterion to find the difference in visual processing of individuals with or without certain psychological characteristic.

Author Keywords

Image Representation; Scene Analysis; Computer Vision; Vision and Scene Understanding; Visual Attention; Eye Fixation; Image Annotation

INTRODUCTION

Understanding an image is straightforward for a human: humans view an image and describe both the content and what is happening in the image. Teaching the computer to learn to understand an image the way that human does is an interesting question since there are lots of potential applications in artificial intelligence fields. One of the most well-known image understanding methods is to recognize objects existing in the image. For example, ImageNet [1] is an image dataset that contains thousands of object classes and is used to train the computer to detect and recognize these objects. And Alexnet [6] is one of the famous deep neural networks for retrieving the information from the dataset and performing well on object detection and recognition tasks. This method uses one aspect of human visual processing: object recognition.

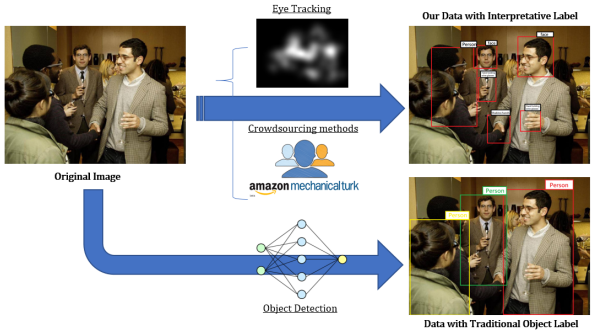

Figure 1. Overview of the data collection concept. Comparing to the previous image labeling methods, our method considers both objects and the interaction or relation between objects. Moreover, our annotation results are more helpful in understanding the whole image since we consider human eye movement while located these regions.

However, human understanding of an image is not limited to the object recognition in the image. Humans consider not only the objects in the image but also the details or distortion of them. Moreover, humans may also focus on the interactions or the relations between objects or smaller entities. Take figure 1 as an example, if we apply an object recognition method such as [10, 9] to this image, these methods result in several *person* objects and their positions in the image. But ideally, we would want to focus on descriptions beyond the objects, such as *"the crowd in the convention room"* or *"the shaking hands of two people at the front."*

In this paper, we introduce a procedure of collecting image information from a more natural perspective of human visual processing. In comparison with an object-oriented dataset, we asked subjects to describe the whole image before labeling partial regions in the image. This way, we could simulate the order of human visual processing while a new scene

ExSS '18, March 11, Tokyo, Japan.

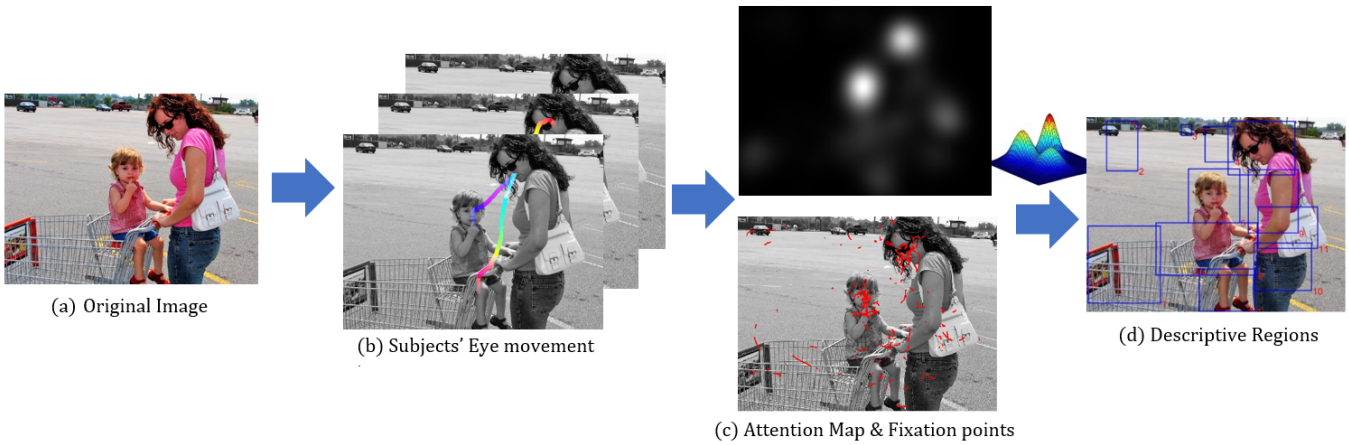

Figure 2. The procedure of generating descriptive regions from human visual attention for image annotation.

incoming. Moreover, we involve eye fixation data as visual attention prior to labeling process. Both the information from feature extraction and the eye fixation traces of the image are included to calculate the important regions of the image. Therefore, these regions are more critical for understanding the image. After regenerating the region-need-to-annotate, we construct an annotation interface for crowdworkers considering the balance between efficiency and fatigue, and simulation of human vision. Our annotation result is closer to how we view an image than previous datasets. These descriptions we collected provide not only the name of entities but also the relations between entities with an overall and natural understanding of the image.

In the following sections, we will discuss the details of our annotation methods and the potential applications based on our dataset.

DATA ANNOTATION

In this section, we introduce how we combine subjects' eye fixation and their descriptions upon an image to generate our new labeling on the image. Our stimuli images for annotation are originated from MS-COCO dataset [7] as a reasonable subset containing different scenes and situations.

Visual Attention Clusters

In order to involve human eye movement to our annotation method, we first recorded 100 subjects' mouse traces on given images with SALICON [4] to simulate their eye movements. SALICON is a tool to approximate visual attention via mouse traces. They first applied Gaussian blur filter on every image and uploaded them to Amazon Mechanical Turk to collect large-scale mouse-tracking data. The collected mouse traces on the blurred images can be transformed into simulated eye movement maps on the images (Figure 2. (b)). In this way, the visual attention map of an image can be approximated from a large amount of subjects' mouse traces instead of using an eye tracking machine (Figure 2.(c)). Furthermore, in order to emphasize on the eye fixation on an image, we collected "fixation points" from the mouse traces. These fixation points are defined and filtered by the length of time that the mouse stays at the certain position. From their analysis in [4], these maps are closer to the real visual attention of human than the attention maps generated by image-oriented saliency detection methods such as [3] and [11].

After obtaining the fixation points in an image, we assume that these points belong to some regions where humans would focus to within their viewing processes. We approximated these regions with a mixture Gaussian model and clustered these fixation points into regions. This way, we generated a set of regions, or *descriptive regions*, that include information related to human visual attention while viewing and understanding an image.

Crowdworkers' Annotation and Postprocessing

After discovering the descriptive regions from aggregating fixation points, we next designed an annotation interface for these regions. Our instructions in the annotation interface lead users to describe the whole image first. Then the interface provides users these descriptive regions in the given image for labeling. With this questions order, we ensure that the descriptions are similar to human natural viewing process. Here we uploaded our annotation interface to Amazon Mechanical Turk (Mturk), which is a crowdworker platform, and collect users' descriptions on it.

In order to refine the descriptions collected from Mturk, we use natural language processing (NLP) tools to postprocess descriptions of each image. Currently one of the NLP tools we use is *Wordnet* [2]. *Wordnet* is an English dictionary dataset that has a tree-like structure for every word. By applying a dictionary to the collected descriptions, we clear the incomprehensible descriptions and merge nouns that have similar meaning, which is defined by both the nouns and their hypernyms that are related by wordnet. Figure 3. shows an example of 10 subjects' descriptions from Mturk and one refined description of the red box in the image. With more descriptions collected and more NLP tools involved in the future, we could generate more detailed and informative annotations for these descriptive regions.

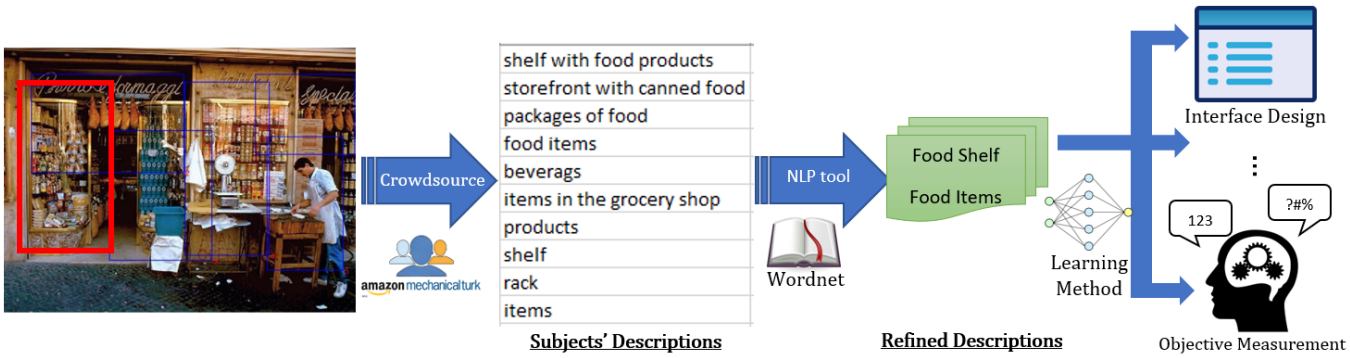

Figure 3. The procedure of generating descriptive regions from human visual attention for image annotation.

DISCUSSION AND NUMERICAL ANALYSIS

Our annotation method represents a different way for computers to learn and understand an image. Currently, we have collected 10 subjects' annotation results on 113 images as a pilot dataset. As more images are annotated using this novel method, computers can learn to generate a more human-like description for a new image by applying neural network methods with structures such as [8] and [5]. Furthermore, this dataset can be used for multiple applications such as object detection, foreground-background separating, scene recognition and image caption generation. Moreover, deep learning methods can exploit these new annotations with both directed and generative models. Given the interest-defined regions R, labels L and a set of images I, a computer can learn the mapping between them and make complex predictions between R, L and I. For example, we can learn to predict what is interesting in an image, or generate novel image from labels and/or regions. We could also find related images using a single input region after we learn the relation between R and I. This connection can also aid in recognizing different objects from a more contextual view.

In addition to direct applications in computer vision fields, we also want to identify human interest in an image. People may naturally gravitate toward only a few regions of interest in some objectively complex images. We want to quantify this *Interest Complexity*, as we believe it is a better measure of how much information people actually extract from an image, independent of the image's objective complexity. In order to generate this *Interest Complexity* (Ω), we consider the users' descriptions (or labels L) collected for regions R from the Mturk platform. Different subjects' descriptions of each image can vary widely, providing considerable information for the annotation procedure. Comparing the content of subjects' descriptions of the same image pairwise is a way to retrieve information. We could not only know whether a subject is answering correctly, but also know how many subjects give simultaneous descriptions. In our analysis, We use spectral clustering to find the description groups that the belonging subjects have similar descriptions. If the image has more sparse groups, the image has descriptive regions or content that is hard to describe with the same words. Following this criterion, we generate Ω for each image based on the collected descriptions from the crowdworkers. Figure 4. shows Ω generated from our current pilot dataset with 10 subjects for each image. In Figure 4. (b), the image contains only one woman with a golf club and the background is clean, which results in a higher magnitude of the simplicity weight. On the other hand, Figure 4. (c) contains an enormous amount of information as it has a cluttered background. The sample description of one descriptive region can be found in Figure 3. Since there are many different descriptions among subjects for one region, the resulting Ω has a lower magnitude. Figure 4 (a) contains an overall example Ω of 20 images. Generally speaking, there maybe two set of ambiguous descriptions of an image, but most user descriptions fall into these two groups, which suggests an objectively complex image and results in a concentrating interest (higher Ω). There may also be 10 sets of different descriptions with each set containing only one user, which would be a case of lower Ω. With more and more user descriptions collected, we could stably generate this interest complexity measurement Ω.

After measuring the interest complexity Ω of each image, a variety of applications in different fields can be developed or improved. For example, we could explore choices for the design of a new user interface based on the testing groups, such as using a simpler image set when the test has a time limitation for users to understand the image. This can also be used as an objective measurement of detecting a certain psychological characteristic. For instance, we could collect eye movement data from individuals with and without a psychological characteristic. By comparing and analyzing these data with our annotated dataset, we could find a more numerical way to distinguish whether a new individual has this characteristic. It could also aid in some clinical tests that currently required clients to take a subjective test, the results of which need to be scored by experienced professionals. Through the above-mentioned process, we could run an objective measurement and facilitate professionals testing. This could also possibly simplify the testing pipeline if image viewing is accessible and more comfortable for clients.

(a) *Interest Complexity* Ω generated from subjects' descriptions.

(b) $\Omega = -0.46$ (Simple)

(c) $\Omega = -0.12$ (Complex)

Figure 4. simplicity weight for an image: the image with a lower weight magnitude means that the image contains more complicated content and has less similarity among the subjects' descriptions.

CONCLUSION

In this paper, we present an image annotation method that includes information from human visual processing. This annotation method shows a novel concept for a computer to understand an image using different aspect from current image processing methods. By using human interest to guide annotations, we get annotations that focus on the naturally interesting aspects of an image and the relations between them. Our annotations are designed to be closer to the way people generate explanations for the content of images. We elicit fixations and annotations based on the user explicitly looking at an image to explain what is happening. We also discuss different applications and effects in a variety of fields to show that this distinctive annotation concept is useful and required for future development of fields such as artificial intelligent, user interface, and psychology. With more and more data being collected within our method and analyzed, computers can learn and achieve a more human-like perspective of the surrounding entities and how they interact or relate to each other.

ACKNOWLEDGMENTS

First of all, we would like to give our appreciation to Professor Angus MacDonald III and his student Danielle Pratt of Clinical Psychology Department for the discussion and eye tracking data collection. We would also make our big thanks to Sewon Oh and all the colleague in CoMoCo Lab of University of Minnesota for the contribution on pre-testing the annotation interface. The last but not the least, we would like to thank to Minnesota Supercomputing Institute (MSI) and College of Liberal Art (CLA) server in University of Minnesota for providing server computation and storage.

REFERENCES

1. Deng, J. D. J., Dong, W. D. W., Socher, R., Li, L.-J. L. L.-J., Li, K. L. K., and Fei-Fei, L. F.-F. L. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), 2–9.
2. Fellbaum, C. *WordNet : an electronic lexical database*. MIT Press, 1998.
3. Goferman, S., Zelnik-Manor, L., and Tal, A. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*, 10 (2012), 1915–1926.
4. Jiang, M., Huang, S., Duan, J., and Zhao, Q. SALICON: Saliency in Context. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June (2015), 1072–1080.
5. Kendall, A., Badrinarayanan, V., and Cipolla, R. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding.
6. Krizhevsky, A., Sutskever, I., and Geoffrey E., H. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25 (NIPS2012)* (2012), 1–9.
7. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8693 LNCS (2014), 740–755.
8. Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:* (2015).
9. Ren, S., He, K., Girshick, R., and Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information* (2015).
10. Simonyan, K., and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ImageNet Challenge* (2014), 1–10.
11. Yang, J., and Yang, M.-H. Top-Down Visual Saliency via Joint CRF and Dictionary Learning. In *CVPR* (2012).