Gaze-guided Image Classification for Reflecting Perceptual Class Ambiguity

Tatsuya Ishibashi **Yusuke Sugano** **Yasuyuki Matsushita**
Graduate School of Information Science and Technology, Osaka University
{ishibashi.tatsuya, sugano, yasumat}@ist.osaka-u.ac.jp

ABSTRACT

Despite advances in machine learning and deep neural networks, there is still a huge gap between machine and human image understanding. One of the causes is the annotation process used to label training images. In most image categorization tasks, there is a fundamental ambiguity between some image categories and the underlying class probability differs from very obvious cases to ambiguous ones. However, current machine learning systems and applications usually work with discrete annotation processes and the training labels do not reflect this ambiguity. To address this issue, we propose an new image annotation framework where labeling incorporates human gaze behavior. In this framework, gaze behavior is used to predict image labeling difficulty. The image classifier is then trained with sample weights defined by the predicted difficulty. We demonstrate our approach's effectiveness on four-class image classification tasks.

CCS Concepts

•Human-centered computing → Human computer interaction (HCI); •Computing methodologies → Computer vision;

Author Keywords

Eye tracking; Machine learning; Computer vision

INTRODUCTION

Machine learning-based computer vision methods have been growing rapidly and the state-of-the-art algorithms even outperform humans at some image recognition tasks [9, 10]. However, their performance is still lower than humans' when the training data is limited or the task is complex [1]. Further, the errors that machines make are often different from the ones humans make [12].

One approach to overcome this difficulty is to incorporate humans in the loop via human-computer interaction [4, 6, 7]. Some prior examples include using human brain activities to infer perceptual class ambiguities in image recognition and assigning difficulty-based sample weights to the training images [8, 15]. However, while gaze is also known to reflect internal states of humans, is much cheaper to measure than brain activity, and has been used as a cue to infer user properties related to visual perception [3, 14, 16, 17], there has not been much research on using gaze data for guiding machine learning processes.

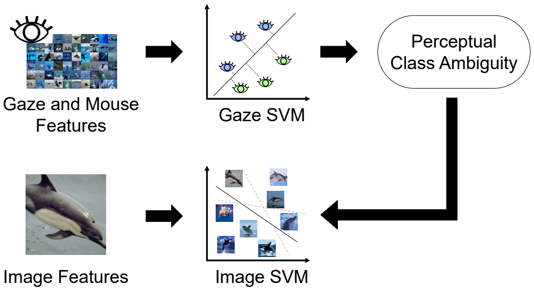

Figure 1. Overview of the proposed method. The first gaze SVM is trained using gaze and mouse features during image annotation and the second image SVM is trained so that it behaves similarly to the gaze SVM and reflects the perceptual class ambiguity for humans.

This work proposes an approach for gaze guided an image classification that better reflects the class ambiguities in human perception. An overview is given in Fig. 1. First, we collect gaze and mouse interaction data when participants work on a visual search and annotation task. We train a support vector machine (SVM) [2] using features extracted from these gaze and mouse data and use its decision function to infer perceptual class ambiguities when assigning the target image classes. The ambiguity scores are used to assign sample weights for training a second SVM with image features. This results in an image classifier that behaves similarly to the gaze-based classifier.

GAZE-GUIDED IMAGE RECOGNITION

The basic idea of our method is that the behavior of the image annotator reflects the difficulty of assigning class labels. Gaze behavior on annotated images is more distinctive if the image clearly belongs to the target or non-target classes, while it becomes more indistinctive on ambiguous cases. Therefore, the decision function of an SVM classifier trained on gaze and mouse features can be used to estimate the underlying perceptual class ambiguity of the training images.

Our method uses gaze data recorded during a visual search and annotation task on an image dataset with pre-defined image

UIST '18 Adjunct October 14–17, 2018, Berlin, Germany

ACM ISBN 978-1-4503-5949-8/18/10.
DOI: https://doi.org/10.1145/3266037.3266090

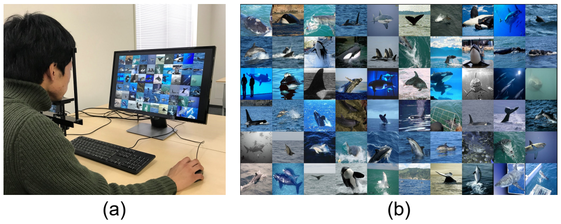

Figure 2. (a) Data collection setup. (b) Example of images displayed in the annotation task.

Table 1. Gaze and mouse features. Median and variance were computed across all participants.

Gaze	Fixation count	Median / Variance
	Total fixation duration	Median / Variance
	Timestamp of the first fixation	Median / Variance
Mouse	Mouseover count	Median / Variance
	Total mouseover duration	Median / Variance
	Timestamp of the first mouseover	Median / Variance
	Timestamp of the first click	Median / Variance
	Proportion of participants who clicked	

categories. Figure 2 shows the setup for the data collection process. In our experiments, we prepared image datasets that consist of four different categories and sequentially showed subsets of 60 images as shown in Fig. 2 (b). Participants were instructed about the four classes beforehand and asked to search for and click 15 images corresponding to one target class out of the four classes within a time limit of 45 seconds. We recorded locations of fixation, mouse cursor and associated timestamps[1]. For each image, we extracted 15 types of gaze and mouse features listed in Table 1. We obtained the median and the variance from the data of all participants.

The gaze and mouse features were then used to estimate the perceptual ambiguity of target labels. We first train an SVM to classify the four image categories with only gaze and mouse features. The distance from the decision boundary to each sample approximately represents the ease of category prediction. We converted the distance to the perceptual class ambiguity score c through a sigmoid function. The score is designed to be small when the sample is misclassified or close to the decision boundary and large when the sample is classified correctly and far from the decision boundary.

The image classifier SVM was trained using a weighted loss function instead of the standard hinge loss function. In the weighted loss function, the loss of the i-th image is $1+c_i$ times the hinge loss, where c_i is the perceptual class ambiguity score of the i-th image. The weighted function assigns a larger loss in proportion to the perceptual class ambiguity estimated from the gaze and mouse behavior, and gives more misclassification penalty to images that are easy for humans to assign target labels.

EXPERIMENTAL RESULTS

We compared the performance of our approach with a standard hinge-loss SVM. We picked four visually similar object classes (dolphin, whale, killer whale, shark) from the ImageNet dataset [5] and four similar scene classes (corn field, golf course, pasture, rice paddy) from the Places205 dataset [18].

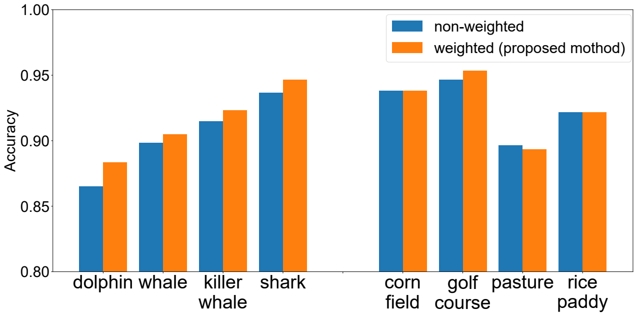

Figure 3. The classification accuracy of SVMs for the object (left) and scene (right) datasets.

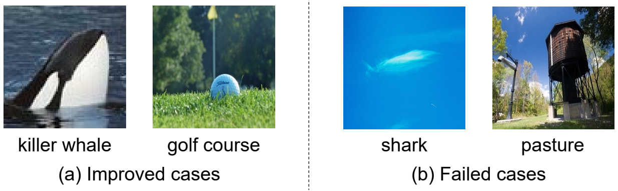

Figure 4. Examples whose classification result has changed through our approach. Texts indicate their ground-truth labels.

Each of the four object/scene classes contains 600 training images and 150 test images. A total of 9 male and 1 female university students (22-24 years old) participated in the annotation task on training images. Image features were extracted from the middle convolutional layer of the AlexNet [11] pre-trained on ILSVRC2012 dataset [13]. Hyperparameter C of the SVM was optimized via 10-fold cross validation on the training data and γ of the RBF kernel was set to $1/600$.

Figure 3 shows the classification accuracies on each dataset. Our proposed method yields performance improvements especially on the object dataset. In the case of dolphin, our approach results in a significant performance improvement ($p < 0.01$, Wilcoxon signed-rank test). Furthermore, Fig. 4 shows some example images whose estimated labels changed with our proposed method. While our proposed method could make correct predictions on obvious cases, it also made false predictions on ambiguous cases reflecting the perceptual class ambiguity for humans.

CONCLUSION

This work explored a gaze-guided image classification approach that incorporates perceptual class ambiguities. Although overall improvements on classification accuracy were relatively marginal, experimental results showed promise that our approach can influence classification algorithm to reflect the underlying ambiguity of image categories. It is expected that the proposed approach will have a larger impact on more challenging classification tasks, possibly with highly subjective labels.

ACKNOWLEDGMENTS

This work was supported by JST CREST Grant Number JPMJCR1781, Japan.

[1] We used a Tobii Pro X3-120 eye tracker in the experiments.

REFERENCES

1. A. Borji and L. Itti. 2014. Human vs. computer in scene and object recognition. In *Proc. CVPR*. 113–120. DOI: http://dx.doi.org/10.1109/CVPR.2014.22
2. B.E. Boser, I.M. Guyon, and V.N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proc. COLT*. 144–152. DOI: http://dx.doi.org/10.1145/130385.130401
3. A. Bulling, C. Weichel, and H. Gellersen. 2013. EyeContext: Recognition of high-level contextual cues from human visual behaviour. In *Proc. CHI*. 305–308. DOI:http://dx.doi.org/10.1145/2470654.2470697
4. Y. Cui, F. Zhou, Y. Lin, and S. Belongie. 2016. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *Proc. CVPR*. 1153–1162. DOI: http://dx.doi.org/10.1109/CVPR.2016.130
5. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*. DOI: http://dx.doi.org/10.1109/CVPR.2009.5206848
6. J.A. Fails and D.R. Olsen Jr. 2003. Interactive machine learning. In *Proc. IUI*. 39–45. DOI: http://dx.doi.org/10.1145/604045.604056
7. J. Fogarty, D. Tan, A. Kapoor, and S. Winder. 2008. CueFlik: Interactive concept learning in image search. In *Proc. CHI*. 29–38. DOI: http://dx.doi.org/10.1145/1357054.1357061
8. R.C. Fong, W.J. Scheirer, and D.D. Cox. 2018. Using human brain activity to guide machine learning. *Scientific reports* 8, 1 (2018), 5397. DOI: http://dx.doi.org/10.1038/s41598-018-23618-6
9. K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proc. CVPR*. 770–778. DOI:http://dx.doi.org/10.1109/CVPR.2016.90
10. A. Karpathy, G. Toderici, S. Shetty ans T. Leung, R. Sukthankar, and L. Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proc. CVPR*. 1725–1732. DOI: http://dx.doi.org/10.1109/CVPR.2014.223
11. A. Krizhevsky, I. Sutskever, and G.E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*. 1097–1105.
12. R.T. Pramod and S.P. Arun. 2016. Do computational models differ systematically from human object perception?. In *Proc. CVPR*. 1601–1609. DOI: http://dx.doi.org/10.1109/CVPR.2016.177
13. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *IJCV* 115, 3 (2015), 211–252. DOI: http://dx.doi.org/10.1007/s11263-015-0816-y
14. H. Sattar, S. Muller, M. Fritz, and A. Bulling. 2015. Prediction of search targets from fixations in open-world settings. In *Proc. CVPR*. 981–990. DOI: http://dx.doi.org/10.1109/CVPR.2015.7298700
15. W.J. Scheirer, S.E. Anthony, K. Nakayama, and D.D. Cox. 2014. Perceptual annotation: Measuring human vision to improve computer vision. *IEEE TPAMI* 36, 8 (2014), 1679–1686. DOI: http://dx.doi.org/10.1109/TPAMI.2013.2297711
16. S. Shimojo, C. Simion, E. Shimojo, and C. Scheier. 2003. Gaze bias both reflects and influences preference. *Nature Neuroscience* 6, 12 (2003), 1317–1322. DOI: http://dx.doi.org/10.1038/nn1150
17. Y. Sugano, Y. Ozaki, H. Kasai, K. Ogaki, and Y. Sato. 2014. Image preference estimation with a data-driven approach: A comparative study between gaze and image features. *JEMR* 7, 3 (2014). DOI: http://dx.doi.org/10.16910/jemr.7.3.5
18. B. Zhou, A. Lapedrizaa, J. Xiao, A. Torralba, and A. Oliva. 2014. Learning deep features for scene recognition using places database. In *Proc. NIPS*. 487–495.