HOGgles: Visualizing Object Detection Features

Summary written by Nicholas Hinke
February 15, 2022

Summary: This paper presents a novel method of inverting and visualizing features used in object detection algorithms in order to better understand their workings and shortcomings. In other words, the authors seek to answer two questions using their new approach: 1) why do object detectors sometimes so blatantly contradict what humans can obviously see in the image, and 2) how much more room for improvement is there within object detectors that employ these popular visual features? Inspired by this quest, as well as by the then-recent work on feature inversion and visualization [2, 7, 11] and object detector performance and limitations in general [3, 5, 8, 12], the authors attempt to provide some meaningful insights to the above questions. Specifically, this paper studies the popular feature known as the *Histogram of Oriented Gradients* (HOG) descriptor [1], although the authors claim their method to be general enough to apply to essentially any visual feature [9].

Approach: As noted above, the goal of this paper was two-fold: first, to produce useful visualizations of a particular feature; and second, to use those visualizations to better understand the performance of object detectors as a whole. To achieve the first goal, the authors implemented four algorithms–the last of which is considered to be the main result of the work–and even made them publicly available as a feature visualization toolbox [9].

The first three algorithms implemented by the authors all closely build on other related works, and thus serve as "baselines" to which the fourth algorithm was compared. The first baseline algorithm was an exemplar linear discriminant analysis (eLDA) detector which was trained on a large database. This method proved to be quite effective at capturing the low-frequency image data despite its simplicity, although suffered greatly in terms of computational–and time–complexity [9]. The second baseline algorithm was ridge regression-based inversion where images and HOG descriptors were treated as random variables sampled from a Gaussian distribution. This method was very fast due to its cheap computational complexity, but failed to capture much of any high frequency image data [9]. The third baseline algorithm involved directly solving a constrained optimization problem using the coordinate descent method. This method was much better than the previous two at capturing high frequency image data, but was also found to be overly sensitive to noise in the original image [9].

Finally, the fourth algorithm utilized paired dictionary learning to invert and visualize the HOG descriptors. This was done by deconstructing the image and its corresponding HOG descriptor into two bases with a shared set of coefficients. Since the bases can be found by solving a typical paired dictionary learning problem, feature inversion can be accomplished by first finding the shared coefficients using the HOG descriptor and its basis, and then projecting those coefficients onto the original image basis [9]. Moreover, this can be done quite efficiently due to the results of other works in the field of dictionary learning [4, 6, 10].

In order to evaluate the efficacy of their feature visualizations, the authors evaluated their algorithm using both automated and human-experimental means. First, they tested how accurately their visualizations reconstruct the original image pixels, and found satisfactory performance of their algorithm which trailed eLDA by 3.4% on average. More importantly, however, the authors tested their algorithm using humans by asking subjects to classify the resulting feature visualizations. Remarkably, their algorithm was found to be superior to all others, including a 19.2% boost in performance over the popular HOG glyphs [9].

Strengths: Most importantly, the authors' algorithm for constructing feature visualizations is far more easily understood by humans, thus allowing researchers to better understand the performance and limitations of various object detectors. Additionally, the authors claim that their visualization methods are feature-independent, which would make their algorithm quite general and versatile. Lastly, it is certainly worth mentioning that this paper demonstrated the first successful feature inversion of HOG descriptors [9].

Weaknesses: Arguably the biggest claim of the paper, the authors assert that their method is generally-applicable to a wide variety of visual features; however, they fail to provide any empirical evidence to support it. Additionally, in reference to its application to HOG features in particular, the authors state that their algorithm is notably sensitive to HOG template size, thus suggesting another potential limitation to the method's generality [9].

Reflections: As mentioned above, it would be quite valuable to study the efficacy of the proposed method on other popular visual features such as LBP or SIFT descriptors. Moreover, as very briefly discussed in the paper, it may prove beneficial to further analyze the ability of this algorithm to reconstruct information from each color channel of a non-monochromatic image [9]. Finally, it would be interesting to test the runtime efficiency of this algorithm on today's modern computer hardware.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, June 2005. IEEE.

[2] E. d'Angelo, A. Alahi, and P. Vandergheynst. Beyond bits: Reconstructing images from local binary descriptors. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 935–938, 2012.

[3] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *ECCV*, 2012.

[4] H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.

[5] L. Liu and L. Wang. What has my classifier learned? visualizing the classification rules of bag-of-feature model by support region detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2012.

[6] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. New York, NY, USA, 2009. Association for Computing Machinery.

[7] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2004.

[8] D. Parikh and C. L. Zitnick. The role of features, algorithms and data in visual recognition. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2328–2335, 2010.

[9] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. In *2013 IEEE International Conference on Computer Vision*, pages 1–8, 2013.

[10] S. Wang, L. Zhang, Y. Liang, and Q. Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2216–2223, 2012.

[11] P. Weinzaepfel, H. Jégou, and P. Pérez. Reconstructing an image from its local descriptors. In *CVPR 2011*, pages 337–344, 2011.

[12] X. Zhu, C. Vondrick, D. Ramanan, and C. C. Fowlkes. Do we need more training data or better models for object detection? In *BMVC*, 2012.