Using humans to make mid-level features

Genevieve Patterson[1] Tsung-Yi Lin[2] James Hays[1]
Brown University[1] University of California, San Diego[2]

Abstract

Abstract:

Section 1, Intro: (1) Discriminative patches are hot. [6, 2, 4] (2) They have demonstrated state of the art performance for various recognition tasks. [examples, citations]. (3) What is the big idea behind these representations? How do they relate to bag of words models? (4) While there are several proposed methods to discover discriminative patches [examples, citations], we examine an interesting alternative – having non-expert humans-in-the-loop to tell us which visual elements they think are discriminative.

(5) Relation to other human-in-the-loop methods [3, 1, 5],. Humans (possibly non expert, crowdsourced humans) are commonly used in vision algorithms at two stages (a) annotation time, either exhaustively annotating a dataset or providing the most informative annotations in an ”active learning” framework or (b) test time, coupled with a computational method (e.g. visipedia) to improve human accuracy and / or reduce human effort. In contrast, we put humans in the loop at neither annotation time or test time, but rather at ”representation discovery” time. The humans are directly telling the computer which visual elements should be discriminative. This has some similarity to part-based annotation of visual phenomena, except we don’t require any explicit semantic meaning for the parts, and in our initial experiments the humans never even see entire images. Putting humans ”in the loop” at this stage in a recognition algorithm, is, to the best of our knowledge, never before studied.

Section 2, Approach: More specifics about our starting point (CMU method), and how we put humans in the loop, and how we learn from human annotations, and how the human-based learning compares to the cross-validation-based learning (which is probably the most complicated, most accurate way to learn these models)

How do we actually turn the human filtering into discriminative patch classifiers? One classifier per HIT? Or have a stronger requirement on consensus?

What if the human filtered patches are consistent concepts, but they’re not close enough together in the feature space? Can we do an optimization to translate (or rotate, or scale) each example in order to maximize their similarity in feature space before training the classifier? Can we train a non-linear classifier?

Results: With all computer discovered patches, X

Acknowledgements. We thank Hari Narayanan (Brown Univ.) for his insights and contributions in the data annotation process. Genevieve Patterson is supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. This work is also funded by NSF CAREER Award 1149853 to James Hays.

References

[1] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *Computer Vision–ECCV 2010*, pages 438–451. Springer, 2010. 1

[2] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM Transactions on Graphics (TOG)*, 31(4):101, 2012. 1

[3] Y. Gingold, A. Shamir, and D. Cohen-Or. Micro perceptual human computation for visual tasks. *ACM Transactions on Graphics (TOG)*, 31(5):119, 2012. 1

[4] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 1

[5] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2973–2980. IEEE, 2012. 1

[6] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *Computer Vision–ECCV 2012*, pages 73–86. Springer, 2012. 1