Histograms of Oriented Gradients for Human Detection

Summary written by Nicholas Hinke
February 01, 2022

Summary: This paper studies the classical problem of shape-based object recognition, specifically by addressing the detection of human pedestrians in images. In order to solve this problem, the proposed detection scheme must correctly determine if an image contains an upright human figure while navigating cluttered backgrounds, variations of illumination, partial occlusions, *etc*. Inspired by the then-recent work on *Scale-Invariant Feature Transform* (SIFT) descriptors [5] and other previous works utilizing orientation histograms [3, 4, 6], the authors seek to demonstrate both the performance and simplicity of their novel keypoint-free approach when compared with other methods. Using a dense overlapping grid of what they denote as *Histogram of Oriented Gradient* (HOG) descriptors, the authors claim to achieve a much greater success rate on the same image set in comparison to other prior works [2].

Approach: Whereas many previous works use wavelets or PCA-based descriptors when representing an image, the authors chose to instead use a "locally well-normalized" dense overlapping grid of HOG descriptors. This choice was made primarily with the goal of recognizing highly-characteristic local shapes with a relative degree of pose-invariance, but also to demonstrate the effectiveness of a comparatively simple dense grid representation which does not require the detection of any particular keypoints [2].

In order to construct this HOG representation of an image and subsequently classify it, the authors implemented a six-stage pipeline in which they varied the corresponding parameters for each stage to maximize the detector's performance. The first stage involved gamma and color normalization, in which they chose to use the RGB color space (when available) with $\gamma = 0$. The second stage required gradient computation, where it was determined that optimal performance was achieved when using 1D centered derivatives on each color channel with no Gaussian smoothing ($\sigma = 0$). The third stage involved binning of the unsigned gradient orientations of 8×8 pixel *cells* using weighted votes of the image gradients centered at each individual pixel–note that 9 bins of 20° increments was shown to perform best. Crucially, the fourth stage required constructing overlapping 2×2 cell *blocks* (*i.e.* 16×16 pixels) such that every cell was a member of four distinct blocks. A Gaussian spatial window of $\sigma = 8$ pixels was then applied to each pixel within every block, before subsequently contrast-normalizing each block using the Lowe-style clipped L2 norm as in [5]. The fifth stage involved concatenating all of the cell responses–each normalized with respect to its four corresponding blocks–over the entire image into one unified HOG representation. Finally, the sixth stage utilized a linear support vector machine (SVM) classifier with a single detection window of the same size as the image to determine whether or not a human figure was present [2].

After constructing the detector, the authors tested it on the existing MIT pedestrian database [7] with near perfect results. The authors then developed a more challenging data set known as 'INRIA' [1] in order to compare the performance of their detector with other current detection methods. The resulting detector vastly outperformed any of the other methods, achieving a miss rate of approximately 0.1 at 10^{-4} false positives per window (FPPW) compared to the next-best Generalized Haar Wavelet method which reached a miss rate of only about 0.25 at 10^{-4} FPPW [2].

Strengths: Most importantly, the detector as implemented in this paper demonstrates superior performance when compared to any other existing methods of the time. Additionally, the authors were quite successful in demonstrating the "power and simplicity" of their dense overlapping grid approach especially in contrast to other keypoint-based methods [2]. Regarding their choice of hyperparameters, the authors were wise to independently vary the parameters of each stage in order to maximize the detector's performance. It was also shrewd of the authors to develop a more challenging data set in order to yield a more interesting comparison of results between detection schemes.

Weaknesses: The most glaring drawback of the detector is the choice of classifier. As mentioned by the authors, a non-trivial performance increase could be attained by using a kernel SVM, at the severe cost of efficiency. Additionally, it seems as though some of the choices of hyperparameter (*e.g.* cell size) may be highly dependent on the image resolution and scale, as the authors state that human limbs within their images are typically about the length of one cell (6-8 pixels) [2].

Reflections: As discussed above, the most obvious potential future work would be to improve the efficiency of the detector such that the performance gains of using a kernel SVM classifier could be realized. Moreover, as mentioned by the authors, a more parts-based approach to human detection may improve the detector's robustness to occlusions as well as variations in scale and pose [2]. Finally, the authors could further prove the effectiveness of their detector by training and testing it on a variety of other object classes.

References

[1] Inria person data set. `http://pascal.inrialpes.fr/data/human/`, 2005.

[2] 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.

[3] W. T. Freeman and M. Roth. Orientation histograms for hand gesture recognition. In *Proceedings of the 1995 IEEE Computer Society International Conference on Automatic Face and Gesture Recognition*, pages 296–301, Zurich, Switzerland, June 1995. IEEE.

[4] W. T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma. Computer vision for computer games. In *Proceedings of the 1996 IEEE Computer Society International Conference on Automatic Face and Gesture Recognition*, pages 100–105, Killington, VT, USA, October 1996. IEEE.

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[6] R. K. McConnell. Method of and apparatus for pattern recognition, U.S. Patent No. 4,567,610, January 1986.

[7] C. Papageorgiou and T. Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33, 2000.