Door Handle Detection for the Stanford AI Robot (STAIR)

Benjamin J Sapp **Jingshen Jimmy Zhang**
Department of Computer Science
Stanford University
Stanford, CA 94305
{*bensapp,jimmy.zhang*}*@stanford.edu*

Abstract

We adapted a visual object detection framework, originally applied to face recognition, to perform fast and robust door handle detection. We detect objects in real time using a cascade of weak classifiers. The framework rejects most image sub-windows early on in the cascade and a very small minority ever reaches the later nodes. Our system achieves a high detection rate (approximately 95%) while maintaining a very low number of false positives, even though door handles are intrinsically less feature-rich than faces.

1 Introduction

The Stanford AI Robot is designed to perform a variety of tasks, including navigating through doorways. This requires real time door handle detection for different environments and many types of handles.

Finding door handles is a rare event detection problem. In practice, millions of windows must be examined of which only a few are door handles. Performance of rare event detection systems can be evaluated by the event detection (true positive) rate d, and false positive rate, f.

We adapted an object detection frame work implemented by Wu et al. [5]. The original cascade architecture was developed by Viola and Jones [2], in which a cascade of weak classifiers is constructed. Each classifier itself is chosen to achieve a very high detection rate while allowing a moderate false positives rate. Upon evaluating an image, sub-windows must avoid rejection in each classifier of the cascade sequentially. Thus, the overall false positive rate would be the product of all the individual rates. Similarly, the overall detection rate would be the product of all the individual detection rates. Consider a cascade with 10 classifiers, where each individual classifier has detection rate .99 and false positive rate 0.5. The overall detection rate is $(0.99)^{10} \approx 0.9$ and corresponding false positive rate is $(0.5)^{10} \approx 9.7\text{E-}4$.

Viola and Jones used this approach for face detection. Wu et al improved upon the algorithm by making the training phase two orders of magnitude more efficient by modifying the feature selection algorithm for each classifier in the cascade.

This paper discusses our application of Wu et al's implementation to door handle detection. We show that though door handles themselves are not as feature-rich as faces, the same framework performs as well on handles as faces. This work gives insight into the tradeoff between detection and false positive rates in our problem, shows the effectiveness of simple rectangular features, and compares our results to previous work in face detection.

2 The cascade classifier architecture

A short discussion of the cascade architecture follows. For a more in depth description, refer to [2].

Each node in the cascade is an ensemble classifier. The classifiers are composed of 2-200 simple continuous features combined with a threshold. AdaBoost is used to select classifiers (or features) during every iteration to lower the ensemble's error rate. Direct feature selection is an alternative to Adaboost, which adds a classifier to the ensemble at each iteration that either improves the detection rate, or lowers the false positive rate. The final classification from direct feature selection is a majority vote of all classifiers.

Rectangular features are used to compute the difference between the sum of intensities in adjacent rectangles of equal size in the image. The rectangles can vary in size and number, yielding an over-omplete set of features for each image. For our setting, we chose the same 5 types of rectangle features as Viola-Jones. These features can be computed extremely quickly using integral images.

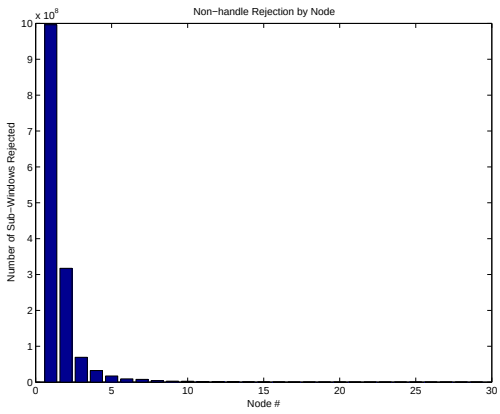

Figure 1

Classification using this cascade is fast since test windows are applied sequentially, and at least 50% of test windows are rejected at each node. Figure 1 shows the number of rejections by node in empirical tests of nearly 1 billion rejections.

3 Training

Our dataset consists of 506 images of exterior and interior door handles from approximately 20 buildings at Stanford. We chose a diverse set of door handles and backgrounds consisting of the door itself and nearby objects. Interesting extraneous objects in this dataset include door frames, separate locks, keyholes, thermostats, trash cans, and bookshelves. Several images even feature glass doors or metal mesh, adding to the background noise. Handle and door color also vary. Pictures were taken from different vantage points to allow detection at slightly different approaches.

We train using 24x24 pixel grayscale images. Formatting raw data into positive training examples took trial and error to achieve good results. We found that the best technique was to crop the door handle as tightly as possible and maintain the same aspect ratio. Attempts to include more visual cues from the handle surroundings into the data failed. For example, we thought it would be intuitive to add the door's edge in the image, which is always close to the door handle and most often shown as a strong dark vertical line in an image. However, because of inconsistent cropping, alignment and the distances between the handle and door

edge, it was difficult to capture this visual cue consistently in our feature set. The result was inconsistent and arbitrary classification.

For all handles, we attempted to align the axis of rotation of the handle with the center of the image patch. This allowed features that were consistent across all door handles. For the majority of door handles, this axis of rotation is the center of a circular disk component of the handle, and thus a very helpful indicator of a door handle. To also maintain consistency across different doors, intensity variance was normalized to minimize lighting discrepancies.

The cascade architecture is only effective at achieving a low false positive rate if each node in the cascade rejects unique sets of features, i.e., makes errors independent of other nodes. Thus, each node is trained on negative examples that were not rejected by any of the previous nodes. A large bootstrap database of negative examples is searched after every node is added to the cascade, and a subset of the database becomes the set of negative examples for the next node. The bootstrap database is a collection of arbitrary images found by crawling the Internet, making sure no door handles were present.

4 Classification

An image to be classified is iteratively scanned at multiple scales and locations by a moving detector window. For 320x240 images, this classification process runs at approximately three frames per second.

It is often the case that sub-windows adjacent in scaling and location will all be classified as door handles As a result, classification often produces many overlapping detections. To obtain a more coherent classification, post-processing is done. In Viola and Jones' work, each set of overlapping detections, represented as bounding boxes in the image, is averaged together to make one final detection. We attained best results by taking the smallest bounding box that enclosed any set of overlapping detections.

Post-processing has some interesting effects on result analysis, as noted in Viola and Jones' work. In our case, while it does a good job in giving an accurate bounding box around door handles, it also reduces the number of false positives. In extreme cases, this has the effect of saturating the number of false positives. As the number of false positives in an image increases, more of them are likely to overlap, and become clustered into one giant false positive.

5 Results

Our full data set for testing contained 1012 images of door handles, and 1720 initial negative examples, combined with a bootstrap set of 1500-1700 additional images. We performed cross validation on this data, training 80% of the examples and reserving 20% for for testing. This was done three times with random dataset partitions and results were subsequently averaged.

The resulting cascade consists of 28 nodes, with a total of 4096 features derived from a pool of 16233 total features. Testing was performed manually for counts of false positives and handle detection rates.

The number of false positives, rather than the false positive rate, was plotted for two reasons. First, it allows us to compare our results to those of face recognition papers. Second, because of post-processing, it is difficult to relate the final number of false positives with the total number of sub-windows examined. In classifying a

typical test set, approximately 300 million sub-windows are examined, and the final number of false positives is six orders of magnitude smaller.

A receiver operating characteristic (ROC) curve is shown in Figure 2. Points were generated by removing nodes from the end of the 28 node cascade to show the relationship between edge detection rate and false positives. The justification for this is that removing a node from the cascade is the same as setting its false positive rate and its detection rate to one.

Our full 28 node cascade, when tested on a set of 203 full size images of handles with background, had a detection rate of 95.3% with an average of 135.67 false positives. In the ROC curve, our detection rate was higher and our false positive number lower than any cascade found in [5] and [2] used for face recognition. Furthermore, the face test set [1] consisted of only 130 images, giving them an even higher number of false positives per image.

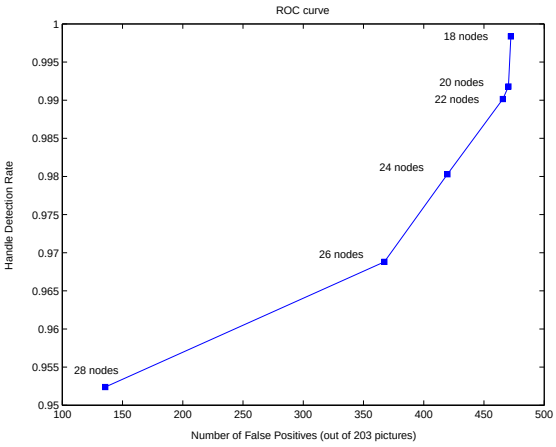

Figure 2

Many attempts were made to further lower the number of false positives by adding in more relevant examples into the bootstrap database – replacing images from the web with objects known to give our current classifier difficulties, in particular shiny and circular objects. These attempts failed to achieve any significant benefits. This may indicate that the set of features used is not rich enough to make any finer distinctions. This is also confirmed upon running tests on a much larger database of bootstrap images from the Internet: using some 7000 images, the cascade was approximately the same number of nodes, and performed comparably to our original cascade.

6 Conclusions

We have demonstrated that we can apply rare event detection to door handle recognition with results comparable to previous applications to face recognition. In real time tests, door handle detection performs well, but with a mediocre false positive rate.

There are many possibilities for future work focusing on how individual nodes are constructed. Wu et al. in [4] explore better ways to combine the weak classifiers in a single node, rather than using AdaBoost's weight setting algorithm. They develop a constrained optimization problem to directly minimize the false positive error of a node with respect to a set of weights, and show empirically that it outperforms AdaBoost's weights.

On a larger scale, using a less primitive set of features may further improve accuracy. One of the major benefits of rectangle features is their ease of computation, which allows training a cascade to take only a matter of hours. A new class of features should try to maintain this efficiency. We have discussd the possibility of applying Sobel operators for finer intensity gradient information, triangle features (a variant of rectangle features but lines are diagonal instead of axis-aligned) and Scale Invariant Feature Transform (SIFT).

Substituting Random Forests for ensemble classifiers to construct a node is another possibility. Random Forests is similar in spirit to the current architecture, in that it takes an ensemble of decision trees as weak classifiers, and takes a majority vote to make a decision. Random Forests has had favorable performance over other algorithms in problems with imbalanced data.

We have established that the cascade architecture not only excels at face detection, but can also be applied to a type of object that is not as rich in visual features. Ultimately, we hope to use this rare detection framework to train cascades for a host of other objects, which would provide robots like STAIR an efficient method for real time visual object detection.

7 Acknowledgements

We are very grateful for Jianxin Wu for providing his rare event detection code, help with debugging, and insight into our specific problem. We would also like to thank Gary Bradski and Prof. Andrew Ng for invaluable discussion and guidance.

References

[1] Rowley, H.A., Baluja, S. & Kanade, T. (1998) Neural network-based face detection. *IEEE Trans. onPattern Analysis and Machine Intelligence*, 20(1):23–38.

[2] Viola, P. & Jones, M. (2001) Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, pages 511–518.

[3] Viola, P., & Jones, M. (2004). Robust real-time face detection. IJCV, 57, 137–154.

[4] Wu, J., Regh, J.M., & Mullin, M.D. (2005) Linear Asymmetric Classifier for Cascade Detectors, ICML.

[5] Wu, J., Regh, J.M., & Mullin, M.D. (2004) Learning a rare event detection cascade by direct feature selection, NIPS.