Summary of "Rapid Object Detection using a Boosted Cascade of Simple Features"

Matt Dickenson `mcd31`

STA571/CS590.01 Due: 31 March, 2014

Viola and Jones (2001) use boosting in the context of objected detection in images. The first step in their process is the computation of image features that can be used for classification. They focus on rectangular features, including a two-rectangle feature (the difference between the sum of pixels within two rectangular regions), a three-rectangle feature (subtracting the sum of two outside rectangles from the sum in the center), and a four-rectangle feature. To facilitate rapid computation of rectangular features, the authors use a representation they refer to as the "integral image":

$$\begin{aligned} ii(x,y) &= \sum_{x' \leq x, y' \leq y} i(x', y') \\ &= ii(x-1, y) + s(x, y) \\ s(x,y) &= s(x, y-1) + i(x, y) \end{aligned}$$

where $i(x, y)$ is the original image and $s(x, y)$ is the cumulative row sum, for recurrence relations with base cases $s(x, -1) = 0$ and $ii(-1, y) = 0$. The computation time for the integral image is linear in the number of pixels. It also allows the computation of any rectangular sum in four (constant-time) array references, simplifying the computation of the two-, three-, and four-rectangle features.

The resolution (24×24 pixels) would allow over 180,000 rectangular features can be computed, so an intermediate feature selection step is important. For this the authors use a "cascade" of AdaBoost classifiers, each of which works for a single feature as follows. For example images $x_i, i \in 1, \ldots, n$ and a binary classification target $y_i \in \{0, 1\}$, initialize weights $w_{1,i} = \frac{1}{2 \times \#negatives}$ for $y = 0$ and $w_{1,i} = \frac{1}{2 \times \#positives}$ for $y = 0$. Iterating for $t = 1 \ldots T$, normalize these weights so that each w_t is a probability distribution. For each feature j train h_j using only a single feature, choose the h_t with the lowest error rate $\epsilon_t = \sum_i w_i |h_j(x_i) - y_i|$, and update the weights $w_{t+1,i} = w_{t,i} \beta_t^{1-e_i}$, where e_i is 0 if x_i is classified correctly and 1 otherwise, and $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$. The final classifier is $h(x) = 1$ if $\sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum t = 1^T \alpha_t$ and 0 otherwise.

The use case in view in this paper is face detection in images. Each image is larger than the detector resolution (384×288 pixels in the example), so there are many sub-images to be analyzed and classified. To reduce computation time, we wish to quickly reject negative examples (not a face) while recognizing as many positive examples (faces) as possible. The authors' use of a series ("cascade") of boosted classifiers, with decreasing thresholds (lower thresholds yield higher detection rates and higher false positives). The cascade quickly rejects negative examples early in the process, and applies more computational effort to cases that are more likely to be positive examples. At each stage of the cascade a target is selected for the minimum reduction in false positives and the maximum decrease in detection. The amortized cost of this detector at every scale and location is much faster than other methods, and maintains accuracy using simple features. Some improvements could be made using an odd number of cascades of classifiers (each trained separately) and taking the majority vote as the classification rule.