Fast Asymmetric Learning for Cascade Face Detection

Jianxin Wu, S. Charles Brubaker, Matthew D. Mullin, and James M. Rehg
College of Computing and GVU Center, Georgia Institute of Technology
TSRB/85 5th ST NW, Atlanta, GA 30332, USA
{wujx,brubaker,mdmullin,rehg}@cc.gatech.edu

Abstract

A cascade face detector uses a sequence of node classifiers to distinguish faces from non-faces. This paper presents a new approach to design node classifiers in the cascade detector. Previous methods used machine learning algorithms that simultaneously select features and form ensemble classifiers. We argue that if these two parts are decoupled, we have the freedom to design a classifier that explicitly addresses the difficulties caused by the asymmetric learning goal. There are three contributions in this paper. The first is a categorization of asymmetries in the learning goal, and why they make face detection hard. The second is the Forward Feature Selection (FFS) algorithm and a fast caching strategy for AdaBoost. FFS and the fast AdaBoost can reduce the training time by approximately 100 and 50 times, in comparison to a naive implementation of the AdaBoost feature selection method. The last contribution is Linear Asymmetric Classifier (LAC), a classifier that explicitly handles the asymmetric learning goal as a well-defined constrained optimization problem. We demonstrated experimentally that LAC results in improved ensemble classifier performance.

1 Introduction

There has been much progress in frontal face detection in recent years. State of the art face detection systems can reliably detect frontal faces at video rate. Various face detection methods have been proposed [33, 28, 31, 45, 24, 11, 37].

Most face detectors use a pattern classification approach. A classifier that can discriminate face patches from background non-face patches is trained from a set of training examples. When a new test image is presented, patches of all possible sizes and positions are extracted and scaled to the same size as the training samples. The trained classifier then decides whether a patch is a face or not. This brute-force search strategy is used in most of the face detection methods.

The classifiers used in early work on face detection, e.g. neural networks [28] and SVM [24], were complex and computationally expensive. Instead of designing a complex classifier, Viola and Jones [37] employed a cascade of simpler classifiers, illustrated in figure 1. An input patch was classified as a face only if it passed tests in all the nodes. Most non-face patches were quickly rejected by the early nodes. Cascade detectors have demonstrated impressive detection speed and high detection rates. In this paper we use the cascade structure, in order to ensure high testing speed.

There were three contributions in the Viola-Jones face detection system: the integral image representation, the cascade framework, and the use of AdaBoost to train cascade nodes. The cascade framework allows background patches to be filtered away quickly. The integral image representation can calculate the image features extremely fast, which are called 'rectangle features' and are then used in the node classifiers. The AdaBoost algorithm [29] is used to select rectangle features and combine them into an ensemble classifier in a cascade node. The integral image and the cascade framework make the detector run

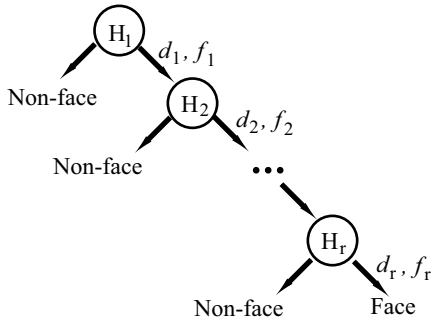


Figure 1: Illustration of the cascade structure with r nodes, where H_i is the ith node classifier, and d_i and f_i are the detection rate and false positive rate of the ith node, respectively.

fast, and AdaBoost is the key to a cascade's high detection rate.

AdaBoost performs two tasks simultaneously when it trains a node classifier: selecting several rectangle features, and forming an ensemble classifier using linear combination of these features. However, these two processes are not necessarily tied together. In this paper, we show that by decoupling the problems of feature selection and ensemble classifier design, we can address the fundamental difficulties in the learning problem explicitly. The result is improved classification performance and faster training times. The contributions of this paper are summarized as three points.

First, we categorize different forms of asymmetries in the face detection problem and explain how they make face detection hard. For example, while the positive class contains only faces and requires only thousands of training images, the negative class contains image patches from all over the world and requires billions of training samples.

Second, we propose the Forward Feature Selection (FFS) algorithm as an alternative way to select features in a cascade node. FFS has similar detection accuracy as AdaBoost. It also reduces the cascade's training time by two orders of magnitude compared to a naive implementation of AdaBoost. We also present a faster implementation for the AdaBoost algorithm. FFS is approximately 2 to 3 times faster than this new implementation of AdaBoost. In addition, FFS only requires about 3% memory usage as that of the faster AdaBoost implementation, which makes it a natural choice in applications with a huge training set and a big pool of features.

Third, we then propose the Linear Asymmetric Classifier (LAC) to form ensemble classifiers. Decoupled from the feature selection process, LAC has the freedom to explicitly address the asymmetries in the face detection problem. Thus, LAC can improve the node classifiers' performances. For example, applying LAC to features selected by AdaBoost, we obtain a better classifier than that provided by AdaBoost itself. LAC runs very fast and is easy to implement. The derivation of LAC also explains the empirical phenomenon that setting the false positive rate of each node to 50% gives the best performance.

The rest of this paper is organized as follows Section 2 explains how the cascade framework helps solving difficulties in the face detection problem and section 3 provides a survey of related methods. FFS and a faster implementation for AdaBoost are introduced in section 4, and LAC is described in section 5. Experimental results comparing FFS/LAC to other methods are also presented. Section 6 concludes this paper with discussions of future work.

Preliminary versions of portions of this work has been published in [42] and [41]. The FFS algorithm presented in this paper is an improved version of the FFS algorithm in [42]. Some new results are also presented in this paper, including analysis of asymmetries in the face detection problem (section 2 and 3), an fast implementation of the AdaBoost method (section 4.3), validity of LAC's assumptions (section 5.2), and additional experimental results (figure 7, 10(c), and 11).

2 Analysis of the cascade face detector

2.1 Asymmetries in Face Detection

We observe three asymmetries in the face detection problem: uneven data distribution, goal asymmetry, and the unequal complexity within the positive and

negative classes. In this section, we will discuss why these asymmetries make the classifier design problem difficult. We want to point out that these asymmetries also apply to the detection of all other objects (e.g. car detection).

The first asymmetry comes with the uneven data distribution. Among the millions of image patches generated from an input image, only very few contain faces. The occurrence of a face in an image is a rare event. In this sense, face detection (and all other detection problems in vision) are rare event detection problems. Methods designed to minimize error rate will classify all instances as negative on such extremely uneven data sets. Thus all faces will be missed. Another difficulty associated with this unevenness is that the negative class usually has huge amount of data. Approximately 350 million non-face patches are used in [37] to train a cascade. The flood of non-face patches makes the learning algorithm train very slowly.

The second asymmetry comes from the difference in positive and negative class learning goals. A high (e.g. 95%) detection rate is required, because we do not want to lose any faces. However, because of the huge amount of non-face data, an extremely low false positive rate (e.g. 10^{-7}) is necessary for reliable detection. It is difficult for a single classifier to achieve such a learning goal.

The last asymmetry comes from the different composition of the two classes. The positive class consists of only faces. However, the negative class consists of image patches from a nearly infinite number of different object categories: animals, trees, man-made objects, buildings, and more. It is not hard to distinguish faces from cars. However, it is much harder to distinguish faces from all other objects.

2.2 Cascade Approach to Asymmetric Problems

The cascade structure alleviates the difficulties associated with the three asymmetries described above.

Cascade classifier deals with the uneven data set with sampling. The training set for each node classifier is balanced by sampling roughly the same amount of non-face patches as the number of faces. After a new node is trained, all non-face patches which are correctly classified by this node are removed from the pool of non-faces. So the number of non-face patches in the pool decreases at an exponential speed. This data bootstrapping strategy deals effectively with the huge amount of non-face data. It is also a way to find non-face samples that are difficult to separate from faces [33]. It is worth noting that sampling is a widely used strategy to deal with uneven data sets in the machine learning and data mining domains [40].

Goal asymmetry is also addressed by the cascade classifier. Consider a cascade consists of a set of nodes $H_1, H_2, \ldots, H_r$. Let M be the event that the testing instance is a true face, and A_i be the event that H_i classifies it as a face. Then, the detection rate D and false positive rate F of the cascade are computed as

$$\begin{array}{rcccl} D & = & \Pr\left[A_r, \ldots, A_1 | M\right] & = & \prod_{i=1}^{r} d_i \\ F & = & \Pr\left[A_r, \ldots, A_1 | \bar{M}\right] & = & \prod_{i=1}^{r} f_i \end{array} \quad (1)$$

by the chain rule, where $d_i = \Pr\left[A_i | A_{i-1}, \ldots, A_1, M\right]$ and $f_i = \Pr\left[A_i | A_{i-1}, \ldots, A_1, \bar{M}\right]$ are the detection rate and false positive rate of the ith node. The above equations do not assume that the nodes make independent errors. The false positive rate F drops to 0 exponentially with the depth of the cascade.

As a consequence of Eq. (1), it is natural to define the learning goal of a cascade as: for every node, design a classifier with very high (e.g. 99.9%) detection rate and only moderate (e.g. 50%) false positive rate.

This node learning goal is in principle easier to achieve in comparison to the daunting 10^{-7} false positive rate goal for a single classifier. However, it is a cost-sensitive learning problem. A false negative clearly costs more than a false positive since we allow about 50% errors in the negative class, but nearly no error can be allowed in the face class. Many machine learning algorithms (including AdaBoost) are designed to minimize error rates and usually do not work well on cost-sensitive problems [40]. Viola and Jones proposed the AsymBoost method [36] to handle this asymmetry. We will discuss the drawback of AsymBoost and other asymmetric learning methods in the related works (section 3). We then propose the Linear Asymmetric Classifier to deal with this cost asymmetry in section 5.

The asymmetry in class composition is taken care of by increasing the complexity of the node classifiers. When more nodes are used, the data bootstrapping process will include instances from more object categories, thus making the node negative training set contain non-face patches which are hard to separate from faces. The node classifiers become more complex, consequently. In [37], only 2 features were used in the first node, while 200 features were used for the last node. Since most non-face patches are rejected by early nodes, a small number of features are evaluated for these patches. This fact enables the cascade to run at video rate.

3 Related works

Our goal is to solve the node learning goal quickly and robustly. However, before presenting our solution, we will first discuss some works that are related to the cascade structured face or general object detectors. A comprehensive review on other face detection systems can be found in [44].

3.1 'Project and Reject' detectors

One key aspect of the cascade detector is the ability to quickly reject some candidate image patches. This intuition is utilized in many face detection systems implicitly or explicitly. For example, the neural network-based detector of Rowley et. al. [28] incorporated a manually-designed two node cascade structure for improving the detection speed.

There are some other detectors which explicitly exploited the idea of rejecting non-face patches in a cascade-like structure. Most of these methods follow a 'Project and Reject' approach. In the maximal rejection classifier approach [8] and the Antiface approach [18], in each stage the input patch was projected to a given direction and was rejected if the projected value was beyond certain thresholds. In [27], a set of reduced set vectors was calculated from a support vector machine. These vectors were applied as projection directions sequentially. An alternative cascade framework for SVM classifiers was proposed by Heisele et. al. [14]. Baker and Nayar proposed a theory of pattern rejection for object recognition [3] using the project and reject procedure.

There are two major differences between these methods and the Viola-Jones detector. First, cascade detectors used Haar-like rectangle features [25]. Rectangle features can be extracted more quickly, in comparison to the projection operation. Second, earlier nodes in a cascade have smaller complexity than deeper nodes. Since most non-face patches are rejected by these early nodes, the cascade is able to run faster than those detectors whose nodes all have the same complexity.

3.2 Node training in a cascade framework

From now on we focus on the Viola-Jones cascade framework in order to ensure video rate testing speed. The central learning problem is to construct a single node which satisfies the node learning goal. Successive iterations of this procedure will result in a cascade.

As discussed above, it is the cost asymmetry (or, cost-sensitive) nature of this learning goal that makes it difficult. Many cost-sensitive learning methods have been proposed. We are specifically interested in those variants of AdaBoost since the Viola-Jones work has demonstrated that AdaBoost is an effective method to learn a node. The naive way, by modifying the initial weight distribution, was used by Schapire et al. in text filtering [30]. However, it is pointed out by Viola and Jones in [36] that, even when positive examples are assigned much higher initial weights than negative examples, the difference is absorbed quickly by AdaBoost. They proposed AsymBoost [36] as a remedy, which continuously gave positive examples higher weights at every training round. They applied AsymBoost to face detection and showed that it had fewer false positives than standard AdaBoost. The key idea is to put more weights on positive examples than negative ones. Many other strategies to apply this idea have been proposed in the machine learning and data mining literature, e.g. AdaUBoost [17], AdaCost [9], and the CSB family [34]: CSB0, CSB1 and CSB2. One characteristic of these methods is that they all

conflate the problem of selecting features with the problem of designing an ensemble classifier.

There are other methods that are related to the node learning goal, e.g. BMPM [15] and MRC [8]. BMPM is an asymmetric learning algorithm, which maximizes an lower bound of the detection rate while keeping the false positive rate smaller than a constant. Additional discussions on these methods will be presented in section 5.4.

Finally, we mention in pass that there are tools to analyze the generalization ability of a cascade. The cascade structure is a special case of decision list, a data structure introduced by Rivest [26]. Anthony gave generalization bounds for threshold decision lists [1], which can be applied to cascade detectors.

3.3 Other methods related to the cascade detector

We also briefly mention some other improvements to the cascade detector, including new features, node and cascade classifiers designing methods, and applications.

New features have been proposed, e.g. the rotated Haar-like features [21], features based on the modified Census Transform [12], and the diagonal features [16]. Levi and Weiss studied various features to reduce the number of training images [19]. Torralba et al. proposed a way to efficiently share features for multi-class object detection, although they did not use a cascade [35].

The learning algorithm used to train node classifiers was another topic of interest. Lienhart et al. [21] experimentally evaluated different boosting algorithms and different weak classifiers. They argued that Gentle AdaBoost and CART decision trees had the best performance. Xiao et al. proposed the Boosting Chain algorithm [43] to integrate historical knowledge into the ensemble classifier. Li et al. incorporated floating search into the AdaBoost algorithm (FloatBoost) for detecting multi-view faces [20]. Liu and Shum proposed KLBoosting to train a node classifier, in which the weak classifiers were based on histogram divergence of linear features [22]. It is worth noting that in KLBoosting classifier designing was decoupled from feature selection. In KLBoosting, the linear classifier was learned using gradient descent after features were selected.

The aforementioned research focused on improving a single node classifier. In [32], Sun et al. considered the problem of connecting the learning objectives of a node to the overall cascade performance. They proposed a cascade indifference curve framework to select the point of operation in a node classifier's ROC curve.

Besides improving the cascade face detector, researchers have used the framework in other fields. A cascade approach has been used (in some cases with substantial modifications) to detect multi-view faces [16, 20], text [6], wiry objects [5], and pedestrians [38].

4 Fast Feature Selection

Algorithm 1 The cascade framework

1: {Given a set of positive examples $\mathcal{P}$, an set of initial negative examples $\mathcal{N}$, and a database of bootstrapping negative examples $\mathcal{D}$. }
2: {Given a learning goal $\mathcal{G}$ for the cascade}
3: {The output is a cascade $H = (H_1, H_2, \ldots, H_r)$}
4: $i \Leftarrow 0$,$H \Leftarrow \emptyset$,
5: **repeat**
6: $\quad i \Leftarrow i + 1$
7: $\quad$ *NodeLearning* { Learn H_i using $\mathcal{P}$ and $\mathcal{N}$, add H_i to H }
8: $\quad \mathcal{N} \Leftarrow \emptyset$
9: $\quad$ Run the current cascade H on $\mathcal{D}$, add any false detection to $\mathcal{N}$ until $\mathcal{N}$ reaches the same size as the initial set.
10: **until** The learning goal $\mathcal{G}$ is satisfied

The cascade framework is shown in algorithm 1. This is an abstract description of the cascade learning algorithm. As showed in algorithm 1, there are two major blocks in training a cascade: line 7 is a node learning algorithm and line 9 is the data bootstrapping process.

In the Viola-Jones detector, the node learning algorithm is AdaBoost, which do feature selection and classifier designing simultaneously. We decouple '*NodeLearning*' into two separate parts and propose Forward Feature Selection to perform the feature selection task. We also propose a fast implementation of AdaBoost to select features.

4.1 The Forward Feature Selection algorithm

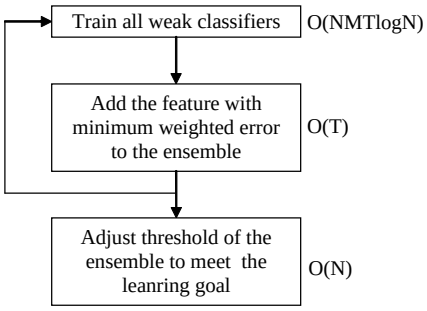
In [37], AdaBoost was used to train the node classifier H_i in line 7 of the algorithm 1. AdaBoost is an iterative method for obtaining an ensemble of weak classifiers by evolving a distribution of weights, D_t, over the training set. In the Viola-Jones approach, each iteration t of the boosting process added the rectangle feature h_t with the lowest weighted error to the ensemble classifier. After T rounds of boosting, the decision of the AdaBoost ensemble classifier is defined as $H(\mathbf{x}) = \text{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{x}) - \theta\right)$, where the α_t's are the ensemble weights obtained by the AdaBoost algorithm and θ is threshold of the ensemble. The flowchart of this algorithm is shown in figure 2(a).

AdaBoost picked the rectangle feature with the smallest *weighted* error with respect to the weight distribution D_t. D_t was updated every round, which in turn required that the weak classifiers were re-trained at every round, as indicated by figure 2(a). In the face detection application, the number of training examples and rectangle features are both in the order of thousands. Thus, the re-training of rectangle features is the most time consuming component in the algorithm. In [37], it took a few weeks to train a complete cascade face detector.

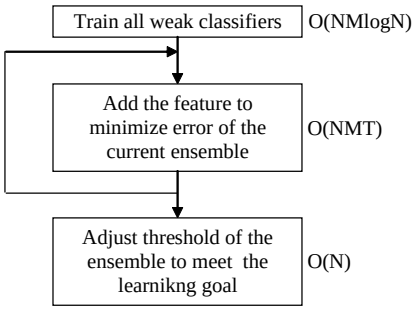
We propose a new feature selection method based on Forward Feature Selection (FFS), a greedy feature selection method [39]. After the features are selected, an ensemble classifier can be formed by simple voting of the selected features. Pseudo-code for the FFS algorithm for selecting features and building an ensemble classifier for a single cascade node is given in algorithm 2. The corresponding flowchart is illustrated in figure 2(b).

Algorithm 2 The FFS algorithm as a new feature selector and node learning algorithm

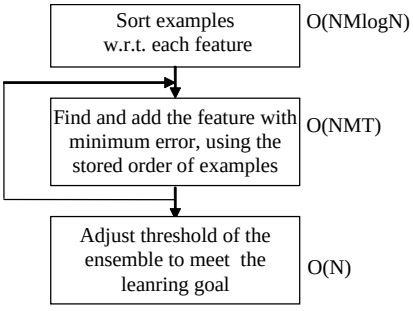
1: {Given a set of examples $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$, where N is the size of the training set.}
2: {Given a set of rectangle features $\{h_i\}_{i=1}^{M}$, where M is the number of rectangle features.}
3: {The output is an ensemble classifier, whose false positive rate is 0.5. S is the set of selected features.}
4: **for** $i = 1$ to M **do**
5: Pick appropriate threshold for rectangle feature h_i, such that h_i has smallest error on the training set
6: **end for**
7: Make a table V_{ij} such that $V_{ij} = h_i(x_j)$, $1 \le i \le M, 1 \le j \le N$
8: $S \Leftarrow \emptyset, v \Leftarrow \mathbf{0}_{1\times N}$, where $\mathbf{0}_{1\times N}$ is a row vector of zeros.
9: **for** $t = 1$ to T **do**
10: **for** $i = 1$ to M **do**
11: $S' \Leftarrow S \cup h_i, v' \Leftarrow v + V_{i:}$, where $V_{i:}$ is the ith row of V.
12: {The classifier associated with S' is $H'(\mathbf{x}) = \text{sgn}\left(\sum_{h\in S'} h\left(\mathbf{x}\right) - \theta\right)$, and we have $H'(\mathbf{x}_i) = \text{sgn}(v'_i - \theta)$.}
13: Find the θ that makes H' has the smallest error rate on the training set
14: $\epsilon_i \Leftarrow$ the error rate of H' with the chosen θ value
15: **end for**
16: $k \Leftarrow \arg\min_{1\le i\le M} \epsilon_i$
17: $S \Leftarrow S \cup h_k$, $v \Leftarrow v + V_{k:}$
18: **end for**
19: {The output is $H(\mathbf{x}) = \text{sgn}\left(\sum_{h\in S} h\left(\mathbf{x}\right) - \theta\right)$}
20: Adjust the value of θ such that H has a 50% false positive rate on the training set.



(a) Naive AdaBoost feature selection implementation



(b) Forward Feature Selection



(c) Faster AdaBoost feature selection implementation

Figure 2: Diagram comparing the naive AdaBoost implementation, the FFS algorithm, and the faster implementation of AdaBoost for selecting features and forming a single node classifiers. In these diagrams, N is the number of training examples, M is the total number of features, and T is the number of features that is needed to be selected in a single node classifier.

Before getting into details of FFS, the flowcharts reveal the intuition behind it. In 2(b), 'Train all weak classifiers', the most time-consuming component, is moved out of the loop. That is, the weak classifiers will be trained for only once. In AdaBoost, they need to be trained T times, where T is the number of features in the AdaBoost ensemble. Since most of the training time is used to train weak classifiers, FFS requires only $1/T$ training time as that of AdaBoost.

In summary, the key intuition in FFS is *caching*: the results of trained weak classifiers are stored and re-used.

Both AdaBoost and FFS are greedy feature selection methods: they pick up the 'best' feature they find in every round, based on different criteria. There are three major differences between FFS and AdaBoost.

First, there is no distribution maintained over the training set in FFS. Each training example is treated equally. Thus, the rectangle features are trained only once. Their corresponding weak classifiers' classification results are stored into a table V. In the following feature selection part, table lookup operations provides all the information needed about the rectangle features. Thus, FFS greatly expedites the training process, with the cost of the storage of a table V.

Second, after the features are selected, the output ensemble classifier $H(\mathbf{x}) = \text{sgn}\left(\sum_{h\in S} h(\mathbf{x}) - \theta\right)$ is a voting of these features in S. The total vote $\sum_{h\in S} h(\mathbf{x})$ is an integer between 0 and T, the number of selected features. The total vote in AdaBoost is a real number instead.

Third, the criterion used in FFS to select features is that the selected feature should make the ensemble classifier has smallest error. In AdaBoost, the criterion is to choose a single feature with smallest weighted error.

The FFS algorithm operates as follows. The first step is to train all weak classifiers and store their classification results into a table V. The set of selected features, S, is initialized to an empty set. At every round, we examine every possible feature and select the feature that mostly reduce the ensemble classifier's error.

Given the candidate feature h_i and S, it is easy to compute the error rates of $S' = S \cup h_i$. The sum of votes v' for S' is obtained by a vector addition (line 11 in algorithm 2). As discussed above, the components of v' are integers between 0 and t, the round index number. Thus, the threshold θ can only take values in the set $\{0, 1, \ldots, t\}$. A histogram for v' can be built (how many positive/negative examples have 0 votes, 1 votes, etc.) With this histogram, it is trivial to find the optimal value for θ, which makes the ensemble classifier associated with S' has smallest error rate (line 13).

As in Algorithm 2, N, M, T denotes the number of training examples, the number of rectangle features, and the number of selected features, respectively. It takes $O(N)$ to compute v'. It takes $O(N)$ steps to build a histogram, and $O(t)$ time to find θ (usually $t \ll N$). The code between line 11 and line 14 has complexity $O(N)$. The outer loop between line 9 and line 18 is $O(NMT)$. We will later show that to train a single feature (line 5) requires $O(N \log N)$ time steps (refer to section 4.3). So the overall time complexity of the FFS algorithm is $O(NMT+NM \log N)$. These computational complexity results are also shown in figure 2(a).

4.2 Complexity of the AdaBoost algorithm

Similarly, we can analyze the complexity of the AdaBoost algorithm.

The first step is to analyze the complexity of the weak classifier training algorithm. When a rectangle feature is given, the optimal threshold τ can only take value from a finite set: the feature values at the training samples. After the feature values are sorted, the error rate of a weak classifier with different thresholds can be updated sequentially at all possible threshold values. This algorithm is described in Algorithm 3.

The sorting can be done in $O(N \log N)$. Thus, the complexity of training a single weak classifier is $O(N \log N)$. The AdaBoost algorithm runs T rounds. In every rounds, almost all time is used to train the M weak classifiers. So the time complexity of the AdaBoost algorithm for training a single node is $O(NMT \log N)$. As discussed in section 4.1, FFS has complexity $O(NMT + NM \log N)$. In face detection we have $N \gg T \gg \log N$, which means that FFS

Algorithm 3 Training a weak classifier

1: {Given a training set $\{\mathbf{x}_i, y_i\}_{i=1}^N$ with weights $\{w_i\}_{i=1}^N$, a rectangle feature h, and its corresponding mask $\mathbf{m}$}
2: Compute the feature values, $v_1, \ldots, v_N$, where $v_i = \mathbf{x}_i^T \mathbf{m}$
3: Sort the feature values into a new vector $v_{i_1}, \ldots, v_{i_N}$ such that $(i_1, \ldots, i_N)$ is a permutation of $(1, \ldots, N)$, and $v_{i_1} \leq \ldots \leq v_{i_N}$
4: $\epsilon \Leftarrow \sum_{y_i=-1} w_i$
5: **for** $k = 1$ to N **do**
6: **if** $y_{i_k} = -1$ **then**
7: $\epsilon \Leftarrow \epsilon - w_{i_k}, \epsilon_i \Leftarrow \epsilon$
8: **else**
9: $\epsilon \Leftarrow \epsilon + w_{i_k}, \epsilon_i \Leftarrow \epsilon$
10: **end if**
11: **end for**
12: $k = \arg\min_{1 \leq i \leq N} \epsilon_i$, $\tau = \mathbf{x}_{i_k}^T \mathbf{m}$
13: The output is a weak classifier $h(\mathbf{x}) = \text{sgn}(\mathbf{x}^T \mathbf{m} - \tau)$

requires approximately $1/T$ of the training time of AdaBoost.

4.3 Faster AdaBoost implementation

Caching is the key to the speed improvement of FFS. Although we have to store a big table V in memory, we never need to train the classifiers again, because all pertinent information is stored in V. We trade space for time.

Careful examination of the Algorithm 3 revealed that AdaBoost can also be expedited using the same caching strategy. In different rounds of the AdaBoost algorithm, the weight distribution $\{w_i\}_{i=1}^T$ changes, but the training set $\{\mathbf{x}_i, y_i\}_{i=1}^N$ remains constant. Furthermore, the feature values v_i and the permutation vector $i_1, \ldots, i_N$ do not change in different rounds of AdaBoost. This part (line 2 and line 3 in Algorithm 3) only needs to be done once and the permutation vectors can be stored in a $M \times N$ table V for future use. Using such a permutation table V, the weak classifier training algorithm (now only line 4-13 of Algorithm 3) is $O(N)$.

With the permutation table, the complexity of the AdaBoost algorithm is greatly reduced. In order to obtain the permutation table V, we need $O(NM \log N)$ time steps. The following T rounds of feature selection each takes $O(NM)$. So the overall complexity of the faster implementation of AdaBoost is $O(NMT + NM \log N)$, which is the same as the FFS algorithm. The flowchart of this implementation is shown in figure 2(c). Avidan and Butman suggested using a similar caching idea independently in [2].

We want to point out that FFS has much lower memory requirement than the faster AdaBoost implementation. In FFS, every entry in the table V is a binary value and requires only 1 bit. However, every entry in the table V in AdaBoost is an integer which uses 32 bits (in a 32-bit CPU). In applications with large number of training examples and a large feature set (and consequently a large table V), FFS is the preferred feature selection method.

4.4 Experiments comparing FFS and AdaBoost

Although the computational complexity of the FFS and AdaBoost algorithms can be analyzed, their detection performances need to be compared by experiments. Both algorithms are used to build cascade face detectors, and their performances are compared on the MIT+CMU test set [28]. We did controlled experiments. Two cascade face detectors were trained using the FFS and AdaBoost algorithm, respectively. All other aspects of the experimental setup were the same: the two cascades were trained with the same training set, validation set, abstract cascade algorithm, and learning goal. In all cascades we trained, cascade nodes at the same depth all have the same number of features. All cascades were evaluated using the same test set and post-processing step.

Our training set contained 5000 example face images and 5000 initial non-face examples, all of size 24x24. We had a set of 4832 face images for validation purposes. We used approximately 2284 million non-face patches to bootstrap the non-face examples between nodes (line 9 of Algorithm 1). We used 16233 features sampled uniformly from the en-

tire set of rectangle features. For testing purposes we used the MIT+CMU frontal face test set in all experiments. Although many researchers use automatic procedures to evaluate their algorithm, we decided to manually count the missed faces and false positives.[1] When scanning a test image at different scales, the image is re-scaled repeatedly by a factor of 1.25. Post-processing is similar to [37].

In the AdaBoost cascade, every node classifier is an AdaBoost ensemble. In each node classifier, the weights of selected features are set by the AdaBoost training algorithm. In the FFS cascade, every node is an ensemble classifier too. In each node classifier, we use FFS to select features, and the weights for all of the selected features are set to 1. In both cascades, we adjust thresholds of the node classifiers such that they have 50% false positive rates.

Our cascade training algorithm terminates when the bootstrapping non-face image database are depleted. Since our learning goal requires that every cascade node has a false positive rate of 50%, all cascades should have approximately the same number of nodes. However, we find that the AdaBoost cascade has 21 nodes, but the FFS cascade has only 17 nodes. This discrepancy comes from the fact that in FFS, every weak classifier has integer votes (1 or -1). With only integer votes, FFS can not achieve an exact 50% false positive rate and we choose lower false positive rates for each node in our experiments. The average false positive rate of all nodes in the FFS cascade is 43.48%. So the FFS cascade has fewer nodes[2].

ROC curves of the AdaBoost cascade and the FFS cascade are shown in figure 3. We construct the ROC curves by repeatedly removing nodes from the cascades to generate points with increasing detection and false positive rates. The curve between two consecutive points is approximated by a line segment. The ROC curves show that the FFS cascade has very close performance to the AdaBoost cascade. In regions with more false positives (>100), the AdaBoost classifier's performance is slight better than that of the FFS classifier. In regions with less false positives, the FFS cascade has slightly better performance.

Experiments also showed other properties of the FFS algorithm. Since the same cascade framework was used in the FFS cascade and each node had the same number of features as the AdaBoost cascade, the detection speed (test speed) should be about the same as the AdaBoost cascade. The experiments showed that both cascade detectors did have very close testing speed.

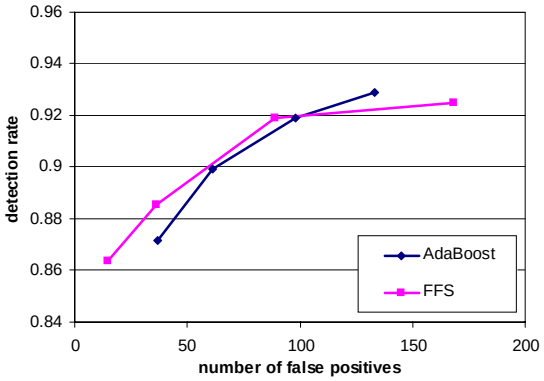


Figure 3: ROC curves comparing cascade detectors using the AdaBoost and FFS algorithm on the MIT+CMU test set.

Furthermore, although both FFS and faster implementation of the AdaBoost algorithm has the same complexity $O(NMT + NM \log N)$, the constant factor in FFS is smaller. To train a node having the same number of features, experiments showed that the faster implementation of AdaBoost usually took 2.5–3.5 times of the training time of FFS, and the original AdaBoost implementation needed 50–150 times of the training time of FFS.

The fact that FFS has similar performance as AdaBoost suggests that AdaBoost is not the only choice for the feature selection method in training a cascade node. Other feature selection methods, such as FFS, can be used in place of AdaBoost and still get good or even better performance. For a complete comparison of feature selection methods in the cascade framework (including FFS, CMIM [10] and various boosting methods), we refer the readers to [4].

5 Linear Asymmetric Classifier

We have presented FFS as an alternative to AdaBoost for feature selection in cascade classifiers. This raises the question of whether alternative methods for forming an ensemble classifier from selected features could lead to performance improvements. In particular, neither FFS or AdaBoost explicitly addresses the difficulties caused by the asymmetries discussed in section 2.1. In this section, we propose the Linear Asymmetric Classifier (LAC) which is designed to handle the asymmetric node learning goal in the cascade framework: for every node, design a classifier with very high (e.g. 99.9%) detection rate and only moderate (e.g. 50%) false positive rate.

5.1 The Linear Asymmetric Classifier

We do not handle the feature selection problem in LAC, that is, we assume that appropriate features have already been selected by other algorithms. For example, AdaBoost, FFS, or information theory based method [10] can be used to select features. We study the problem of how to find a linear classifier that's optimal for the node learning goal with these features.

The problem formulation can be formalized as follows. Let $\mathbf{x} \sim (\bar{\mathbf{x}}, \Sigma_{\mathbf{x}})$ denote that $\mathbf{x}$ is drawn from a distribution with mean $\bar{\mathbf{x}}$ and covariance matrix $\Sigma_{\mathbf{x}}$. Note that we do not assume any specific form of the distribution. The only assumption is that its mean and covariance can be estimated from samples. We are dealing with binary classification problems with two classes $\mathbf{x} \sim (\bar{\mathbf{x}}, \Sigma_{\mathbf{x}}), \mathbf{y} \sim (\bar{\mathbf{y}}, \Sigma_{\mathbf{y}})$, which are fixed but unknown. Here $\mathbf{x}$ denotes a vector of feature values of a positive example and $\mathbf{y}$ denotes a vector of feature values of a negative example. Note that the notations in this section are different from previous sections. Both $\mathbf{x}$ and $\mathbf{y}$ are used to denote feature vectors, in order to emphasis the fact the learning goals are asymmetric and to make presentations more clear. Class labels of training examples are obvious from the notation ($\mathbf{x}$ for positive and $\mathbf{y}$ for negative) and are ignored. We use $\mathbf{z}$ to denote an example with unknown class label. The linear classifier to be learned can be written as $H = (\mathbf{a}, b)$:

$$H(\mathbf{z}) = \begin{cases} +1 & \text{if } \mathbf{a}^T\mathbf{z} \geq b \\ -1 & \text{if } \mathbf{a}^T\mathbf{z} < b \end{cases}.$$

The asymmetric node learning goal is expressed as:

$$\begin{array}{ll} \max\limits_{\mathbf{a}\neq 0, b} & \Pr\limits_{\mathbf{x}\sim(\bar{\mathbf{x}},\Sigma_{\mathbf{x}})}\{\mathbf{a}^T\mathbf{x} \geq b\} \\ \text{s.t.} & \Pr\limits_{\mathbf{y}\sim(\bar{\mathbf{y}},\Sigma_{\mathbf{y}})}\{\mathbf{a}^T\mathbf{y} \leq b\} = \beta. \end{array} \quad (2)$$

In general this problem has no closed-form solution. In this section, we will develop an approximate solution for it. Empirical results showed that it is effective to set $\beta = 0.5$ for all cascade nodes. Thus, we will give a closed-form (approximate) solution when $\beta = 0.5$.

Note that an AdaBoost classifier is a linear combination of weak classifiers:

$$H(\mathbf{x}) = \text{sgn}(\Sigma_{t=1}^{T} a_t h_t(\mathbf{x}) - b) = \text{sgn}(\mathbf{a}^T\mathbf{h}(\mathbf{x}) - b) \quad (3)$$

in which $\mathbf{h}(\mathbf{x})$ is the vector of weak classifiers' outputs. Thus $H(\mathbf{x})$ is a linear classifier in the feature space defined by $\mathbf{h}(\mathbf{x})$. However, there is no guarantee that the $(\mathbf{a}, b)$ selected by AdaBoost will satisfy Eq. (2) for a given choice of β. The same argument applies to FFS. We seek a linear discriminant $(\mathbf{a}, b)$ which maximizes the node learning goal in Eq. (2).

The key idea to solve this learning problem is to use the cumulative distribution functions of $\mathbf{a}^T\mathbf{x}$ and $\mathbf{a}^T\mathbf{y}$ to replace the Pr{} function.

Let $\mathbf{x_a}$ denote the standardized version of $\mathbf{a}^T\mathbf{x}$ ($\mathbf{x}$ projected onto the direction of $\mathbf{a}$), i.e.

$$\mathbf{x_a} = \frac{\mathbf{a}^T(\mathbf{x} - \bar{\mathbf{x}})}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}, \quad (4)$$

obviously we have $\mathbf{x_a} \sim (0, 1)$. Let $\Phi_{\mathbf{x,a}}$ denotes the cumulative distribution function (c.d.f.) of $\mathbf{x_a}$, i.e.

$$\Phi_{\mathbf{x,a}}(b) = \Pr\{\mathbf{x_a} \leq b\}. \quad (5)$$

$\mathbf{y_a}$ and $\Phi_{\mathbf{y,a}}$ are defined similarly as

$$\mathbf{y_a} = \frac{\mathbf{a}^T(\mathbf{y} - \bar{\mathbf{y}})}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}}, \quad (6)$$

$$\Phi_{\mathbf{y,a}}(b) = \Pr\{\mathbf{y_a} \leq b\}. \quad (7)$$

Thus, the constraint in Eq. (2) can be re-written as

$$\begin{aligned}\beta &= \Pr\left\{\mathbf{a}^T\mathbf{y} \le b\right\} \\ &= \Pr\left\{\frac{\mathbf{a}^T(\mathbf{y}-\bar{\mathbf{y}})}{\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}} \le \frac{b-\mathbf{a}^T\bar{\mathbf{y}}}{\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}}\right\} \\ &= \Phi_{\mathbf{y},\mathbf{a}}\left(\frac{b-\mathbf{a}^T\bar{\mathbf{y}}}{\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}}\right),\end{aligned}$$

which in turn gives an expression for the optimal value of b:

$$b = \mathbf{a}^T\bar{\mathbf{y}} + \Phi^{-1}_{\mathbf{y},\mathbf{a}}(\beta)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}} \tag{8}$$

where $\Phi^{-1}_{\mathbf{y},\mathbf{a}}$ is the inverse function of $\Phi_{\mathbf{y},\mathbf{a}}$. Note that $\Phi^{-1}_{\mathbf{y},\mathbf{a}}$ depends on both $\mathbf{y}$ and $\mathbf{a}$. Similarly, the objective function in Eq. (2) can be re-written as

$$1 - \Phi_{\mathbf{x},\mathbf{a}}\left(\frac{b-\mathbf{a}^T\bar{\mathbf{x}}}{\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}}}\right)$$

Using Eq. (8) to eliminate b and we obtain

$$1 - \Phi_{\mathbf{x},\mathbf{a}}\left(\frac{\mathbf{a}^T(\bar{\mathbf{y}}-\bar{\mathbf{x}}) + \Phi^{-1}_{\mathbf{y},\mathbf{a}}(\beta)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}}}\right).$$

Thus the constrained optimization problem (2) is equivalent to

$$\min_{\mathbf{a}\neq 0} \Phi_{\mathbf{x},\mathbf{a}}\left(\frac{\mathbf{a}^T(\bar{\mathbf{y}}-\bar{\mathbf{x}}) + \Phi^{-1}_{\mathbf{y},\mathbf{a}}(\beta)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}}}\right). \tag{9}$$

In Eq. (9), $\Phi_{\mathbf{x},\mathbf{a}}$ and $\Phi^{-1}_{\mathbf{y},\mathbf{a}}$ depend on the distributions of $\mathbf{x}$ and $\mathbf{y}$, in addition to the projection direction $\mathbf{a}$. Because we have no knowledge of these distributions, we cannot solve Eq. (9) analytically. We need to make some approximations to simplify it.

First, let us give a bound for Φ and Φ^{-1}. Let $Z \sim (0,1)$, applying the one-tailed version of the Chebyshev inequality, for $z > 0$ we get

$$\begin{aligned}\Phi(z) &= \Pr\{Z \le z\} \\ &= 1 - \Pr\{Z \ge z\} \\ &\ge 1 - \tfrac{1}{1+z^2} \\ &= \tfrac{z^2}{1+z^2}.\end{aligned} \tag{10}$$

Since Φ^{-1} is increasing, we have

$$\Phi^{-1}(\Phi(z)) = z \ge \Phi^{-1}\left(\frac{z^2}{1+z^2}\right)$$

A little bit of algebraic manipulation gives us the following bound:

$$\Phi^{-1}(\beta) \le \kappa(\beta), \text{ where } \kappa(\beta) = \sqrt{\frac{\beta}{1-\beta}} \tag{11}$$

From the definition, it is obvious that $\mathbf{x}_\mathbf{a} \sim (0,1)$. Thus, instead of minimizing $\Phi_{\mathbf{x},\mathbf{a}}(z)$ in Eq. (9), we can instead minimize its upper bound $\kappa(z)$. Furthermore, since $\kappa(z)$ is an increasing function, it is equivalent to minimizing z. Thus, we can *approximately* solve Eq. (9) by solving

$$\min_{\mathbf{a}\neq 0} \frac{\mathbf{a}^T(\bar{\mathbf{y}}-\bar{\mathbf{x}}) + \Phi^{-1}_{\mathbf{y},\mathbf{a}}(\beta)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}}}, \tag{12}$$

or, equivalently,

$$\max_{\mathbf{a}\neq 0} \frac{\mathbf{a}^T(\bar{\mathbf{x}}-\bar{\mathbf{y}}) - \Phi^{-1}_{\mathbf{y},\mathbf{a}}(\beta)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}}}. \tag{13}$$

This transformation is approximate because in Eq. (9), the function $\Phi_{\mathbf{x},\mathbf{a}}$ depends on $\mathbf{a}$, while $\mathbf{a}$ also appears in the argument of $\Phi_{\mathbf{x},\mathbf{a}}$. However, if we assume that $\mathbf{a}^T\mathbf{x}$ is Gaussian for any $\mathbf{a}$, then $\mathbf{x}_\mathbf{a}$ is the standard normal distribution. Under this assumption, $\Phi_{\mathbf{x},\mathbf{a}}$ does not depend on $\mathbf{a}$ any more, and Eq. (13) is exactly equivalent to Eq. (2).

Second, we assume that the median value of the distribution $\mathbf{y}_\mathbf{a}$ is close to its mean. This assumption is true for all symmetric distributions and is reasonable for many others. Under this assumption, we have $\Phi^{-1}_{\mathbf{y},\mathbf{a}}(0.5) \approx 0$. Thus for $\beta = 0.5$ (which is used in the cascade framework), Eq. (13) can be further approximated by

$$\max_{\mathbf{a}\neq 0} \frac{\mathbf{a}^T(\bar{\mathbf{x}}-\bar{\mathbf{y}})}{\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}}}. \tag{14}$$

If in addition we can assume that $\mathbf{y}_\mathbf{a}$ is a symmetric distribution and $\beta = 0.5$, we have $\Phi^{-1}_{\mathbf{y},\mathbf{a}}(0.5) = 0$ in addition to Eq. (13). The implication is that under these assumptions, Eq. (14) is exactly equivalent to the node learning goal in Eq. (2). We call the linear discriminant function determined by Eq. (14) the Linear Asymmetric Classifier (LAC) and use it in the cascade learning framework.

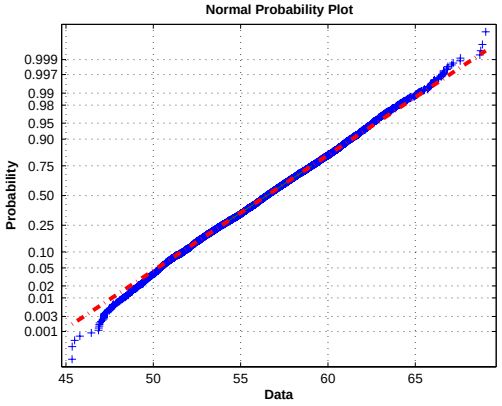


Figure 4: Normality test for $\mathbf{a}^T\mathbf{x}$, in which $\mathbf{x}$ is features extracted from face data, and $\mathbf{a}$ is drawn from the uniform distribution $[0\ 1]^T$.

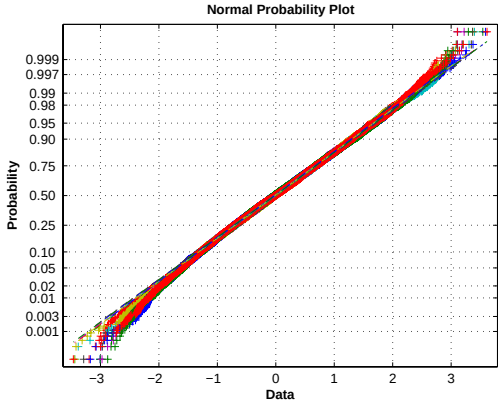


Figure 5: Normality test for $\mathbf{a}^T\mathbf{x}$. This figure shows overlapped results for 10 different $\mathbf{a}$'s, all drawn from the uniform distribution $[0\ 1]^T$. The data in this figure are centered (i.e. means are subtracted).

The form of Eq. (14) is similar to the Fisher Discriminant Analysis (FDA) [13], which can be written as:

$$\max_{\mathbf{a}\neq 0} \frac{\mathbf{a}^T(\bar{\mathbf{x}}-\bar{\mathbf{y}})}{\sqrt{\mathbf{a}^T(\Sigma_{\mathbf{x}}+\Sigma_{\mathbf{y}})\mathbf{a}}}. \tag{15}$$

The only difference between FDA and LAC is that the pooled covariance matrix $\Sigma_{\mathbf{x}}+\Sigma_{\mathbf{y}}$ is replaced by $\Sigma_{\mathbf{x}}$. This analogy immediately gives us the solution to Eq. (14) as:

$$\mathbf{a}^* = \Sigma_{\mathbf{x}}^{-1}(\bar{\mathbf{x}}-\bar{\mathbf{y}}), b^* = \mathbf{a}^{*T}\bar{\mathbf{y}}, \tag{16}$$

under the assumption that $\Sigma_{\mathbf{x}}$ is positive definite. In applications where $\Sigma_{\mathbf{x}}$ happens to be positive semi-definite, $\Sigma_{\mathbf{x}}+\lambda I$ can be used to replace $\Sigma_{\mathbf{x}}$, where λ is a small positive number.

5.2 Empirical Support for Gaussianity

We have shown that if for any $\mathbf{a}$, $\mathbf{a}^T\mathbf{x}$ is Gaussian and $\mathbf{a}^T\mathbf{y}$ is symmetric, then LAC is guaranteed to be the optimal linear classifier for the node learning goal. In this section we verify that these assumptions are valid in the cascade face detector.

Probability theory shows that $\mathbf{x}$ is Gaussian if and only if $\mathbf{a}^T\mathbf{x}$ is Gaussian for all $\mathbf{a}$'s [7]. Obviously $\mathbf{x}$ is not Gaussian since all of its components are binary random variables. However, experiments show that $\mathbf{a}^T\mathbf{x}$ is approximately Gaussian for most reasonable instantiations of $\mathbf{a}$. Figure 4 shows the normal probability plot of $\mathbf{a}^T\mathbf{x}$ for $\mathbf{a}$ randomly drawn from the uniform distribution. $\mathbf{a}^T\mathbf{x}$ fits closely to a normal distribution, only with small deviations at the tails. Figure 5 shows that $\mathbf{a}^T\mathbf{x}$ is approximately Gaussian for all different $\mathbf{a}$'s in our experiments.

Experiments also show that $\mathbf{a}^T\mathbf{y}$ fits nearly exactly to a Gaussian (refer to Figure 6). We tested the normality of $\mathbf{a}^T\mathbf{y}$ for many non-face training data sets and different instantiations of $\mathbf{a}$, and $\mathbf{a}^T\mathbf{y}$ always fits a Gaussian distribution. Since centered Gaussian distributions are symmetric, we can safely assume that $\mathbf{a}^T\mathbf{y}$ is symmetric for all $\mathbf{a}$'s.

The normal probability plot is a way to visually examine whether a distribution is normal or not. The kurtosis of a one dimensional distribution provides a numerical evaluation of the normality, since the kurtosis of a normal distribution is 0 and nearly all non-Gaussian distributions have non-zero kurtosis. Table 1 summarizes the kurtosis value of $\mathbf{a}^T\mathbf{x}$ and $\mathbf{a}^T\mathbf{y}$ for 1000 different $\mathbf{a}$'s, all of which drawn randomly from the uniform distribution $[0\ 1]^T$. The result in table 1 confirms that for most reasonable $\mathbf{a}$, $\mathbf{a}^T\mathbf{x}$ is approxi-

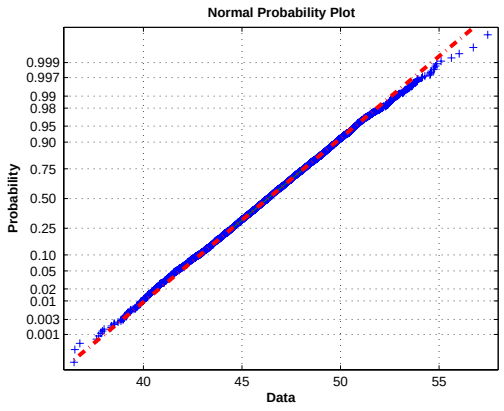


Figure 6: Normality test for $\mathbf{a}^T\mathbf{y}$, in which $\mathbf{y}$ is features extracted from non-face data, and $\mathbf{a}$ is drawn from the uniform distribution $[0\ 1]^T$.

mately Gaussian and $\mathbf{a}^T\mathbf{y}$ fits very close to a normal distribution.

	Kurt($\mathbf{a}^T\mathbf{x}$)	Kurt($\mathbf{a}^T\mathbf{y}$)
mean	-0.23	-0.02
standard deviation	0.05	0.07
min	-0.38	-0.06
max	-0.22	0.23

Table 1: Summary of the kurtosis of $\mathbf{a}^T\mathbf{x}$ and $\mathbf{a}^T\mathbf{y}$ for 1000 different $\mathbf{a}$'s randomly drawn from the uniform distribution $[0\ 1]^T$.

5.3 LAC in the cascade framework

Since most visual classification tasks are not linearly separable, we need to inject some non-linearity into the linear asymmetric classifier. When the rectangle features are used, the vector of their corresponding weak classifiers' output $\mathbf{h}(\mathbf{x})$ is used as features (refer to Eq. 3). The rectangle features can be selected by any feature selection method, e.g. AdaBoost, AsymBoost, or FFS. The abstract cascade learning algorithm remains unchanged (Algorithm 1), while the node learning algorithm is replaced by a feature selector plus LAC. The new node learning algorithm is shown in Algorithm 4.

Algorithm 4 The LAC algorithm as a new node learning algorithm

1: {Given a training set composed of positive examples $\{\mathbf{x}_i\}_{i=1}^{n_\mathbf{x}}$ and negative examples $\{\mathbf{y}_i\}_{i=1}^{n_\mathbf{y}}$, and a set of rectangle features.}
2: {Given a feature selection method $\mathcal{F}$}
3: {The output is a classifier with false positive rate 0.5}
4: Use $\mathcal{F}$ to select T weak classifiers $\mathbf{h} = (h_1, h_2, \ldots, h_T)$ where $h_i(\mathbf{z}) = \text{sgn}(\mathbf{z}^T\mathbf{m}_i - \tau_i)$
5: Build a feature vector $\mathbf{h}(\mathbf{z}) = (h_1(\mathbf{z}), h_2(\mathbf{z}), \ldots, h_T(\mathbf{z}))$ for each training example.
6: Estimate the mean and covariance: $\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n_\mathbf{x}} \mathbf{h}(\mathbf{x}_i)}{n_\mathbf{x}}$, $\bar{\mathbf{y}} = \frac{\sum_{i=1}^{n_\mathbf{y}} \mathbf{h}(\mathbf{y}_i)}{n_\mathbf{y}}$, $\Sigma_\mathbf{x} = \frac{\sum_{i=1}^{n_\mathbf{x}} (\mathbf{h}(\mathbf{x}_i)-\bar{\mathbf{x}})(\mathbf{h}(\mathbf{x}_i)-\bar{\mathbf{x}})^T}{n_\mathbf{x}}$, $\Sigma_\mathbf{y} = \frac{\sum_{i=1}^{n_\mathbf{y}} (\mathbf{h}(\mathbf{y}_i)-\bar{\mathbf{y}})(\mathbf{h}(\mathbf{y}_i)-\bar{\mathbf{y}})^T}{n_\mathbf{y}}$
7: Applying Eq. (16) to get

$$\mathbf{a} = \Sigma_\mathbf{x}^{-1}(\bar{\mathbf{x}} - \bar{\mathbf{y}}), b = \mathbf{a}^T\bar{\mathbf{y}}$$

8: The output is a classifier

$$H(\mathbf{z}) = \text{sgn}\left(\sum_{t=1}^{T} \mathbf{a}_t\mathbf{h}_t(\mathbf{z}) - b\right) = \text{sgn}\left(\mathbf{a}^T\mathbf{h}(\mathbf{z}) - b\right)$$

When we want to apply FDA instead of LAC, we replace Eq. (16) with the following FDA solution:

$$\mathbf{a} = (\Sigma_\mathbf{x} + \Sigma_\mathbf{y})^{-1}(\bar{\mathbf{x}} - \bar{\mathbf{y}}). \tag{17}$$

It is tempting to use the integral feature values $\mathbf{z}^T\mathbf{m}_k - \tau_k$ directly as features in the Linear Asymmetric Classifier or FDA. However, since $\mathbf{z}^T\mathbf{m}_t - \tau_t$ is a linear function of the input $\mathbf{z}$, $H(\mathbf{z}) = \text{sgn}\left(\sum_{t=1}^{T} \mathbf{a}_t(\mathbf{z}^T\mathbf{m}_t - \tau_t) - b\right)$ is still a linear discriminant function and will not work for problems that are not linearly separable. The sgn function introduces the necessary non-linearity into the features.

Both AdaBoost and LAC have the same form $H(\mathbf{z}) = \text{sgn}\left(\mathbf{a}^T\mathbf{h}(\mathbf{z}) - b\right)$ and they share the same feature vector $\mathbf{h}(\mathbf{z})$. The only difference between

these two classifiers are parameters of the linear discriminant $(\mathbf{a}, b)$. In AdaBoost, a_i is chosen in step i of the AdaBoost procedure to minimize a margin-based cost function [23]. This is a greedy procedure and a_i is never changed after its value is determined. Furthermore, AdaBoost does not take into account the fact that the two classes are asymmetric. The linear asymmetric classifier, on the contrary, is a global procedure to seek the optimal vector $\mathbf{a}$ which optimizes the asymmetric loss in Eq. (2).

Viola and Jones proposed AsymBoost [36] to accommodate the asymmetry. In AsymBoost, sample weights were updated using

$$D_{t+1}(i) = \frac{D_t(i)\exp\left(-y_i h_t(x_i)\right)\exp\left(y_i \log\sqrt{k}\right)}{Z_t}$$

instead of the standard AdaBoost updating rule

$$D_{t+1}(i) = \frac{D_t(i)\exp\left(-y_i h_t(x_i)\right)}{Z_t}.$$

The extra term $\exp\left(y_i \log\sqrt{k}\right)$ causes the algorithm to gradually pay more attention to positive samples in each round of boosting, in which k is a parameter representing the level of asymmetry. However, the resulting linear discriminant $(\mathbf{a}, b)$ is determined in the same way as ordinary AdaBoost.

5.4 Comparison to previous work

There are other methods that are similar to the form of Eq. (14). Researchers have also presented methods that are related to the node learning goal. However, the node learning goal was not explicitly defined and solved in these methods. In this section we will examine the relationship between the proposed LAC and other related classifiers.

By applying Eq. (11) into Eq. (9), we get another approximation to Eq. (9):

$$\max_{\mathbf{a}\neq 0} \frac{\mathbf{a}^T\left(\bar{\mathbf{x}} - \bar{\mathbf{y}}\right) - \kappa\left(\beta\right)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}, \qquad (18)$$

which is the Biased Minimax Probability Machine [15]. Eq. (18) is a worst case lower bound of our objective function Eq. (9). The solution of a BMPM used fractional programming which is computationally expensive. The BMPM solver in [15] often takes thousands of iterations to converge, while the solution in Eq. (16) requires only a single matrix inversion. BMPM solves a more general class of problems than LAC (β is not confined to 0.5), while LAC incorporates more domain knowledge. This difference is clear from the derivations of these method. BMPM is based on the Chebyshev inequality, while LAC is based prior knowledge (assumptions) about the data.

Another related objective function comes from the Maximum Rejection Classifier (MRC) [8], which can be written:

$$\max_{\mathbf{a}\neq 0} \frac{\left(\mathbf{a}^T\bar{\mathbf{y}} - \mathbf{a}^T\bar{\mathbf{x}}\right)^2 + \mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}. \qquad (19)$$

The solution of Eq. (19) requires solving a generalized eigenvalue problem. The intuition behind Eq. (19) is to make the overlap between the projections $\mathbf{x_a}$ and $\mathbf{y_a}$ small. The derivation of Eq. (19) in [8] treats the two classes equally. Asymmetry in the MRC framework results from the fact that the two classes have different prior probabilities with $P(\mathbf{x}) \ll P(\mathbf{y})$. However, the effect of the prior on $\mathbf{y}$ is reduced quickly as the stage-wise rejection process continues. After a few rejections, $P(\mathbf{x})$ is not negligible any more in comparison to $P(\mathbf{y})$. Under such conditions, Eq. (19) is not an appropriate objective function.

A final comparison can be made between LAC and FDA. FDA and LAC both have their own merits and drawbacks. We have shown that when $\mathbf{x_a}$ is normal, $\mathbf{y_a}$ is symmetric, and $\beta = 0.5$, LAC is indeed the optimal solution to the node learning goal. However, when these assumptions are broken, LAC may be suboptimal. The intuition in FDA is to maximize the (normalized) separation between the two class means. It does not minimize the error rate or the node learning goal. The advantage of FDA is that it does not have constraints – performance will be reasonably good if the class means are far apart. If we assume that $\mathbf{x}$ and $\mathbf{y}$ have equal covariance matrices, LAC is equivalent to FDA.

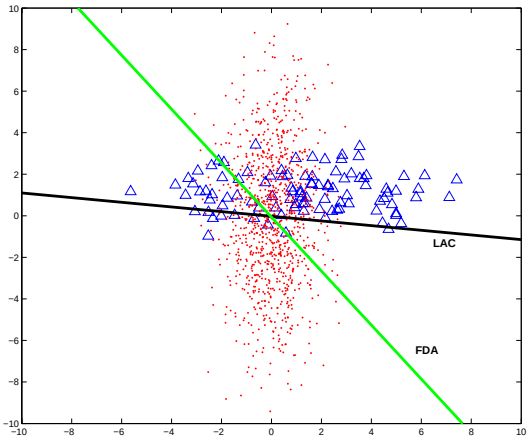


Figure 7: Comparing LAC and FDA on synthetic data set when both $\mathbf{x}$ and $\mathbf{y}$ are Gaussians.

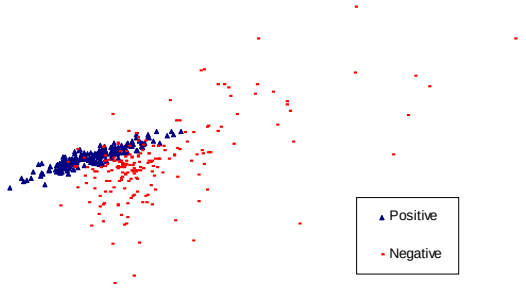


Figure 8: Example of a synthetic data set where $\mathbf{y}$ is not symmetric.

5.5 Experiments on LAC

We tested the performance of the linear asymmetric classifier on both a synthetic data set and the face detection task. In the synthetic data set, LAC is compared against BMPM, MRC and FDA. For detection of faces, the cascade framework is used. Three feature selectors are used: AdaBoost, AsymBoost, and FFS. We compare three different ways to determine the linear discriminant $(\mathbf{a}, b)$ after the features are selected. The first method is to use the weights $\mathbf{a}$ and threshold b found by AdaBoost or AsymBoost; the second method uses the proposed linear asymmetric classifier; the third method uses Fisher Discriminant Analysis. We use "X+Y" to denote the methods used in experiment, e.g. AdaBoost+LAC means that the features are selected by AdaBoost and the linear discriminant function is trained by LAC.

5.5.1 Results on Synthetic Data

Figure 7 gives some intuition of the difference between LAC and FDA. The positive data is drawn from a normal distribution with mean $(1.23, 1.23)$ and covariance $[10\ 0; 0\ 1]$. The negative data is drawn from a normal distribution with mean $(0, 0)$ and covariance $[1\ 0; 0\ 10]$. LAC is guaranteed to be optimal in this data set. It rejects 50% of the negative data, while keeping almost all positive data. On the contrary, FDA returns a boundary that also rejects 50% of the negative data, but more than 13% of the positive data are rejected.

We also tested LAC on data sets where its assumptions were broken. The synthetic data set was generated using the following steps. Three distributions were created as: $\mathbf{d}_i = A_i\mathbf{c}_i + \mathbf{m}_i, i = 1, 2, 3$, where $\mathbf{c}_i \sim N(0, I), \mathbf{m}_i \sim N(0, 0.1I)$, and the elements of A were drawn randomly from a uniform distribution in $[0, 1]$. The positive examples $\mathbf{x}$ were drawn from $\mathbf{X} = \mathbf{d}_1$, and negative examples $\mathbf{y}$ were drawn from $\mathbf{Y} = \mathbf{d}_2^2 - \mathbf{d}_3^2$. This choice produced a $\mathbf{Y}$ which is not symmetric, and has a reasonable overlap with $\mathbf{X}$. One example of such a data set is shown in figure 8. The training and test sets both contain 1000 samples, including 500 positive and 500 negative samples. Four linear discriminant methods (LAC, FDA, MRC, and BMPM) were compared[3]. In each method, we determined the projection direction $\mathbf{a}$ using the corresponding method. The threshold b was determined such that on the training set the false positive rate was 50%. For every method, the experiments were repeated 100 times. The averaged test set accuracy on both classes are reported in table 2.

Although the negative class is not symmetric and LAC is not optimal, it works the best while FDA follows closely. Two-tailed paired t-test shows that

there is no significant difference between LAC and FDA. Both the difference between LAC and MRC, and the difference between LAC and BMPM are significant, at the 0.01 level.

Classifier	Positive Accuracy	Negative Accuracy
LAC	96.11	50.04
FDA	95.12	49.96
MRC	90.19	49.96
BMPM	87.99	50.26

Table 2: Results on synthetic data set.

In cascaded classifiers, the imbalance between classes is absorbed by the cascade structure. In each node of a cascade, balanced training sets are usually used. This is why we used a balanced training set in the above synthetic data set. We also tested the performance of these classifiers on imbalanced training set. Two extra sets of experiments were performed. The training sets still had 500 positive examples, but the negative class had 1000 and 1500 examples, respectively. The examples were drawn from the same distributions as described above. All four classifiers' performances remained approximately the same, despite of the increasing of negative training examples. Thus detailed error rates are not presented. Under both imbalance level, LAC performed about the same as FDA, and both LAC and FDA were better than MRC and BMPM.

5.5.2 Results on Face Detection

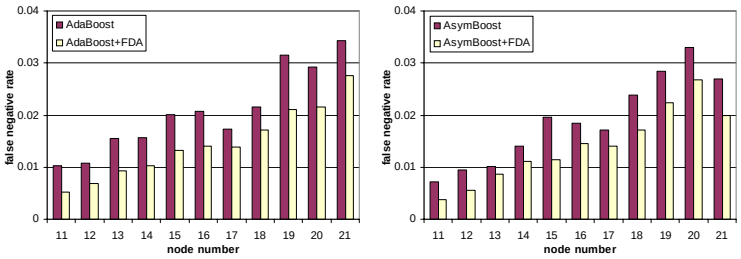
For face detection, we trained 9 different cascades, using the three feature selectors (AdaBoost, AsymBoost, and FFS) and three linear discriminant functions (using weights provided by the feature selector, LAC, or FDA). Each cascade has 21 nodes, except that the AsymBoost+LAC cascade has 22 nodes and the FFS cascade has 17 nodes. We require that every node have 50% false positives and the cascade training process is finished when there are not enough non-face patches to bootstrap.

In order to make the face detector run at video speed, the first node used only 7 features. We used more features while the node index increases (the last node used 200 features).

We consider two types of performance measures: node and cascade. The node performance measure is the classifiers' ability to achieve the node learning goal. Given a trained cascade, each node has an associated training set, which is generated by the bootstrapping process (refer to algorithm 1). We collected all such training sets from the 9 trained cascades. Given one such training set, different algorithms are required to achieve the criteria in Eq. (2). Their performance is evaluated using the validation set. The node performance measure is useful because it directly compares the ability of each method to achieve the node learning goal. The cascade performance measure compares the performance of the entire cascade. The performance of a cascade depends on more than the classifier that is used to train the nodes. The background data bootstrapping step and post processing step in face detection also have significant effects on the cascades' performance. The cascade performance measure is evaluated using the MIT+CMU benchmark test set.

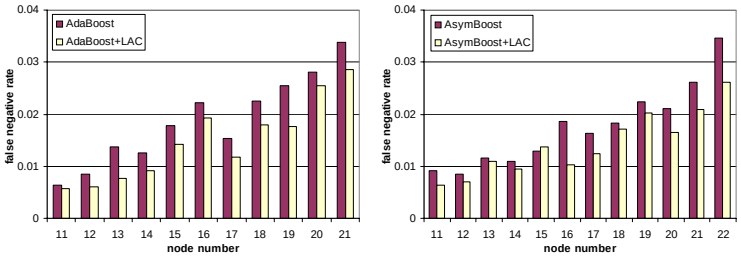
The node comparison results are shown in figure 9. We did not perform the node comparison for the FFS algorithm, since FFS enforced all weights to be 1 and was not in the same hypothesis space as AdaBoost, FDA, or LAC. We collected the training set from the remaining 6 cascades, using the other two feature selectors (AdaBoost and AsymBoost) and three linear discriminant functions. The AdaBoost cascade is the same one as used in section 4.4).[4]

We are able to observe the effects of using FDA or LAC to train a linear discriminant function instead of using the values provided by the AdaBoost (or AsymBoost) algorithm. From the results in figure 9, it is obvious that both FDA and LAC can greatly reduce the false negative rates (i.e. increase the detection rates). In figure 9(a), averaged over the 11 nodes shown, AdaBoost+FDA reduces the false negative rates by 31.5% compared to AdaBoost, while in 9(c) AdaBoost+LAC reduces it by 22.5%. When AsymBoost is used as the feature selector, the reductions are 27.3% and 17.3%, respectively. In figure 9(a) to 9(d), training sets came from FDA or LAC cascades. We also compare node performance when the training sets came from the AdaBoost or Asym-



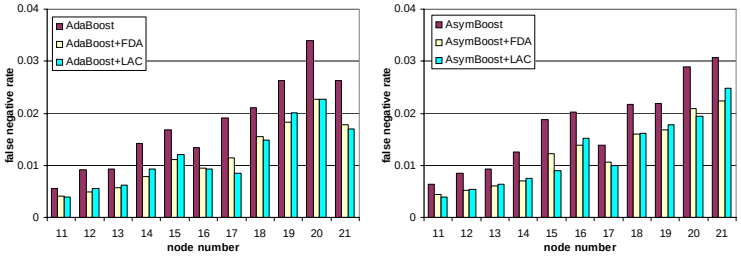
(a) AdaBoost vs FDA

(b) AsymBoost vs FDA



(c) AdaBoost vs LAC

(d) AsymBoost vs LAC



(e) AdaBoost vs FDA & LAC

(f) AsymBoost vs FDA & LAC

Figure 9: Experiments comparing different linear discriminant functions. The y axis shows the false negative rate when $\beta = 0.5$. In 9(a), training sets are collected from the AdaBoost+FDA cascades' node 11 to 21 (x axis shows the node number). AdaBoost and AdaBoost+FDA are compared using these training sets. Similarly, 9(b)-9(f) used training sets from the AsymBoost+FDA, AdaBoost+LAC, AsymBoost+LAC, AdaBoost, and AsymBoost cascades respectively. We do not show results when the data set index is less than 11, for space constraint.

Boost cascade. Results are shown in figure 9(e) and 9(f). Both FDA and LAC work better than the original AdaBoost and AsymBoost.

Cascade comparison results are shown in figure 10. The x axis is the total number of false positives in the MIT+CMU test set. The y axis is the correct detection rate. Figure 10(a)-10(c) show the results when AdaBoost, AsymBoost, or FFS is used as the feature selector, respectively. These ROC curves show that both FDA and LAC have significant advantages over the linear discriminant provided by AdaBoost, AsymBoost, and FFS. It coincides well with the node performances in figure 9.

Another way to interpret figure 10 is to compare the number of false positives at a same detection rate. LAC can greatly reduce false positives. For example, in figure 10(a), averaged over the range of possible detection rates, AdaBoost+LAC reduces the number of false positives by 36.1% compared with AdaBoost.

5.5.3 Discussions

There are a few interesting observations from the experimental results that are worth discussion.

First, the improvement of LAC over AsymBoost is smaller than that of LAC over AdaBoost. Our conjecture is that since AsymBoost already takes into account the asymmetry when it selects features, LAC has smaller space to improve.

Second, the error reduction effects of FDA or LAC in figure 9 is more significant than those in figure 10. We conjecture that the background data bootstrapping and post processing remove part of the error reducing effects.

The effect of post-processing can be studied by evaluating the cascade detector without performing the post-processing step. The results without post-processing when AdaBoost was used is shown in figure 11. Comparing figure 10(a) and figure 11, we find that the post-processing step does not change the relative performances of different algorithms, since these two figures are very similar to each other. However, post-processing can greatly reduce the absolute number of false positives.

Third, we find that FDA works better than LAC in a few cases. LAC is derived under the assumptions

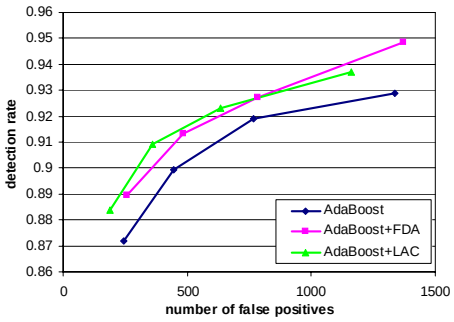


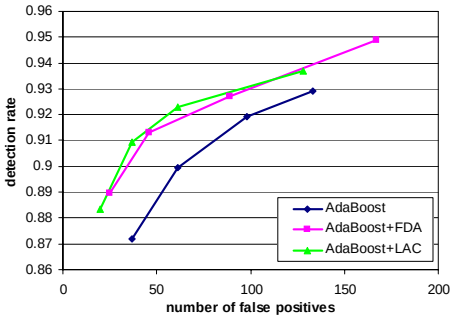
Figure 11: Experiments comparing cascade performances using the MIT+CMU benchmark test set. The x axis is the total number of false positives without performing the post-processing step.

that $\mathbf{x_a}$ is Gaussian and $\mathbf{y_a}$ is symmetric. However, from figure 4 and figure 6 we find that although $\mathbf{y_a}$ is always symmetric, $\mathbf{x_a}$ slightly deviates from the Gaussian distribution at the tails. This might be the reason why LAC does not perform the best in some cases.

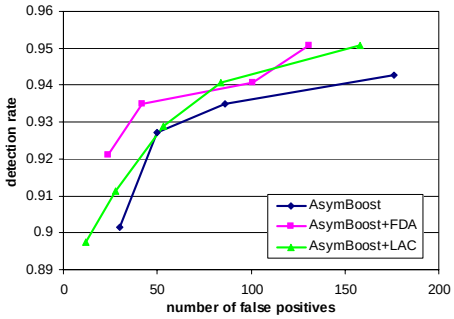
One final point is that the derivation of LAC gives an explanation for setting $\beta = 0.5$. Experiments show that $\beta = 0.5$ gives the best empirical results. The cascade framework itself does not explain this phenomenon. Our explanation is that setting $\beta = 0.5$ helps us incorporate the domain knowledge. The negative data sets $\mathbf{y}$ are successive samples of the non-face images, which basically includes every possible image in the world, except faces. The characteristics of these negative data sets are very complex to model. However, it is reasonable to assume that $\mathbf{y}$ follows a symmetric distribution. By setting $\beta = 0.5$ and consequently using $\Phi^{-1}_{\mathbf{y},\mathbf{a}}(0.5) = 0$, we have incorporated all the domain knowledge we have about $\mathbf{y}$, without any further assumptions.

6 Conclusions

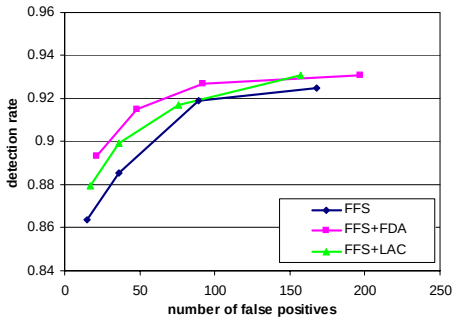
We have presented a new approach to design a node classifier in a cascade detector. Previous methods used machine learning algorithms that simultaneously select features and form ensemble classifiers.



(a) AdaBoost vs FDA & LAC



(b) AsymBoost vs FDA & LAC



(c) FFS vs FDA & LAC

Figure 10: Experiments comparing cascade performances using the MIT+CMU benchmark test set. The x axis is the total number of false positives. The y axis is the detection rate.

We analyzed the asymmetries inherent in the face detection problem and why these asymmetries make the problem difficult. We then argued that if we decouple the feature selection step from the classifier design step, we have the freedom to use different feature selection methods. More importantly, we have the freedom to design an ensemble classifier that explicitly address the asymmetries in its learning goal. We proposed FFS as the new feature selection method, and LAC as the new classifier.

The contributions of this paper can be summarized into three points. The first contribution is an analysis of the cascade detector. Three types of asymmetries are categorized: uneven data distribution, goal asymmetry, and the unequal complexity of the positive and negative class. We argued that these asymmetries are the characteristics of the face detection problem that make it hard to solve. A literature survey of computer vision and machine learning researches to deal with these asymmetries are provided too.

The second contribution is Forward Feature Selection, the feature selection part of our new decoupled node learning algorithm. We also propose a faster implementation method for the AdaBoost algorithm. On one hand, FFS provides an alternative feature selection method. The classifier formed by voting the FFS features has similar accuracy as the AdaBoost method. On the other hand, FFS is computationally attractive. FFS is two orders of magnitude faster than the naive implementation of AdaBoost. FFS is also 2.5 to 3.5 times faster than the faster implementation of AdaBoost, but only requires about 3% memory usage as that of AdaBoost.

The third contribution is Linear Asymmetric Classifier, the classifier design part of the decoupled node learning algorithm. The asymmetries is taken care of by LAC as a well-defined constrained optimization problem. By incorporating domain knowledge (or assumptions about the data), LAC solves this complex optimization problem approximately in closed form and a computationally efficient manner. The derivation of LAC also gives some hints to interesting characteristics of the face detection data sets and the phenomenon that setting $\beta = 0.5$ gives best empirical result. Experiments on both synthetic and MIT+CMU benchmark show that LAC can greatly reduce the errors. In addition, we also applied Fisher's Discriminant Analysis to features extracted by AdaBoost and got improved results.

Despite its effectiveness, there are limitations in our node learning algorithm. We describe these limitations and propose some future work that are possible ways to address these limitations.

- One of the conditions for LAC to be optimal is $\beta = 0.5$. However, in applications other than face detection, $\beta \neq 0.5$ may be required. Other asymmetric learning methods, such as BMPM [15], might be used in these cases.

- LAC is a linear classifier. It is possible to extend LAC to a non-linear classifier (for example, use the kernel method). Will this non-linear extension improve the classifier's performance?

- We observed that FDA outperformed LAC in some cases and we conjecture that this is because $\mathbf{a}^T\mathbf{x}$ is only approximately Gaussian. Is it possible to design a new feature selection method which guarantees $\mathbf{a}^T\mathbf{x}$ to be normally distributed?

- It is also desirable to have a feature selection method that take into account the asymmetric learning goal in a principled way.

Notes

[1]We found that the criterion for automatically finding detection errors in [21] was too loose. This criterion yielded higher detection rates and lower false positive rates than manual counting.

[2]Since every node in the FFS cascade has a false positive rate lower than 50%, it will consume more bootstrapping non-face data than a cascade node trained by AdaBoost (with 50% false positive rate). Thus the FFS cascade consumes the bootstrapping non-face data more quickly and consequently has fewer number of nodes.

[3]We used the Matlab toolbox for BMPM from http://www.cse.cuhk.edu.hk/~miplab/mempm_toolbox/.

[4]The source code for training a cascade using methods described in this paper is available online at http://www.cc.gatech.edu/~wujx. We also provide trained cascades, demo executables, and a video showing testing results.

References

[1] M. Anthony. Generalization error bounds for threshold decision lists. *Journal of Machine Learning Research*, 5:189–217, Feburary 2004.

[2] S. Avidan and M. Butman. The power of feature clustering: An application to object detection. In *NIPS 17*, 2004.

[3] S. Baker and S. Nayar. Pattern rejection. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 544–549, 1996.

[4] S.C. Brubaker, J. Wu, J. Sun, M.D. Mullin, and J.M. Rehg. On the design of cascades of boosted ensembles for face detection. Technical report, GIT-GVU-05-28, GVU center, Georgia Institute of Technology, 2005.

[5] O. Carmichael and M. Hebert. Shape-based recognition of wiry objects. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages II:401–408, 2003.

[6] X. Chen and A.L. Yuille. Detecting and reading text in natural scenes. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages II:366–373, 2004.

[7] E. J. Dudewicz and S. N. Mishra. *Modern Mathematical Statistics*. Wiley, 1988.

[8] M. Elad, Y. Hel-Or, and R. Keshet. Pattern detection using a maximal rejection classifier. *Pattern Recognition Letters*, 23(12):1459–1471, 2002.

[9] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan. Adacost: Misclassification cost-sensitive boosting. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 97–105, 1999.

[10] F. Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5:1531–1555, 2004.

[11] F. Fleuret and D. Geman. Coarse-to-fine face detection. *International Journal of Computer Vision*, 41(1-2):85–107, 2001.

[12] B. Froba and A. Ernst. Face detection with the modified census transform. In *The Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 91–96, May 2004.

[13] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, 1990.

[14] B. Heisele, T. Serre, S. Mukherjee, and T. Poggio. Feature reduction and hierarchy of classifiers for fast object detection in video images. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages II:18–24, 2001.

[15] K. Huang, H. Yang, I. King, and M. R. Lyu. Learning classifiers from imbalanced data based on biased minimax probability machine. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages II:558–563, 2004.

[16] M. J. Jones and P. Viola. Fast multi-view face detection. Technical report, TR2003-96, MERL, 2003.

[17] G. J. Karakoulas and J. Shawe-Taylor. Optimizing classifiers for imbalanced training sets. In *NIPS 11*, pages 253–259, 1999.

[18] D. Keren, M. Osadchy, and C. Gotsman. Antifaces: A novel, fast method for image detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(7):747–761, 2001.

[19] K. Levi and Y. Weiss. Learning object detection from a small number of examples: the importance of good features. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages II:53–60, 2004.

[20] S.Z. Li, Z.Q. Zhang, H.-Y. Shum, and H.J. Zhang. FloatBoost learning for classification. In S. Thrun S. Becker and K. Obermayer, editors, *NIPS 15*. MIT Press, December 2002.

[21] R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. Technical report, MRL, Intel Labs, 2002.

[22] C. Liu and H.-Y. Shum. Kullback-leibler boosting. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages I:587–594, 2003.

[23] L. Mason, J. Baxter, P. L. Bartlett, and M. Frean. Functional gradient techniques for combining hypotheses. In *Advances in Large Margin Classifiers*, pages 221–246. MIT Press, 2000.

[24] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 130–136, 1997.

[25] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proc. Intl. Conf. Computer Vision*, pages 555–562, 1998.

[26] R. L. Rivest. Learning decision lists. *Machine Learning*, 2(3):229–246, 1987.

[27] S. Romdhani, P. Torr, B. Schoelkopf, and A. Blake. Computationally efficient face detection. In *Proc. Intl. Conf. Computer Vision*, pages 695–700, 2001.

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

[29] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[30] R. E. Schapire, Y. Singer, and A. Singhal. Boosting and rocchio applied to text filtering. In *SIGIR*, pages 215–223, 1998.

[31] H. Schneiderman and T. Kanade. A statistical model for 3d object detection applied to faces and cars. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 746–751, 2000.

[32] J. Sun, J. M. Rehg, and A. F. Bobick. Automatic cascade training with perturbation bias. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages II:276–283, 2004.

[33] K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

[34] K. M. Ting. A comparative study of cost-sensitive boosting algorithms. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 983–990, 2000.

[35] A. Torralba, K.P. Murphy, and W.T. Freeman. Sharing features: Efficient boosting procedures for multiclass object detection. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages II:762–769, 2004.

[36] P. Viola and M. Jones. Fast and robust classification using asymmetric AdaBoost and a detector cascade. In *NIPS 14*, 2002.

[37] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[38] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. Intl. Conf. Computer Vision*, pages 734–741, 2003.

[39] A. R. Webb. *Statistical Pattern Recognition*. Oxford University Press, New York, 1999.

[40] G. M. Weiss. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, 6(1):7–19, 2004.

[41] J. Wu, M.D. Mullin, and J.M. Rehg. Linear asymmetric classifier for cascade detectors.

In *Proc. 22nd International Conference on Machine Learning*, 2005.

[42] J. Wu, J.M. Rehg, and M.D. Mullin. Learning a rare event detection cascade by direct feature selection. In *NIPS 16*, 2003.

[43] R. Xiao, L. Zhu, and H.-J. Zhang. Boosting chain learning for object detection. In *Proc. Intl. Conf. Computer Vision*, pages 709–715, 2003.

[44] M.-H. Yang, D. J. Kriegman, and N. Ahujua. Detecting faces in images: a survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[45] M.-H. Yang, D. Roth, and N Ahuja. A snow-based face detector. In *NIPS 12*, pages 862–868, 1999.