Improved Response Modelling on Weak Classifiers for Boosting

Gary Overett
RSISE
Australian National University
ACT 0200, Australia
Email: gary.overett@rsise.anu.edu.au

Lars Petersson
National ICT Australia
Locked Bag 8001
Canberra, Australia
Email: lars.petersson@nicta.com.au

***Abstract*— This paper demonstrates a method of increasing the quality of weak classifiers in the boosting context by using improved response modelling. The new method improves upon the results of a recent response binning approach proposed by Rasolzadeh et al. [1]. For experimental purposes the improved method is applied to the familiar Haar features as used by Viola and Jones in their face/pedestrian detection systems. However, the methods benefits are general and therefore not restricted to this particular feature type. Unlike many previous methods, this method is suitable for modelling multi-modal responses and is highly resistant to overfitting. It does this by adaptively choosing suitable support regions around the values taken by the standard response binning method. More accurate models are produced, with particular improvement around the final decision boundary. It is shown that the new method can be trained with one tenth of the training data required to achieve similar results on previous methods. This substantially lowers the overall training time of the system. The method's ability to consistently produce better hypotheses over a variety of pedestrian detection tasks is shown.**

I. INTRODUCTION

This paper concerns itself with the problem of pedestrian detection using an improved version of the well known method of Viola and Jones [2]. The detector presented here is for use aboard a moving smart car [3].

There are many new and successful visual pattern recognition methods in the field of visual detection, e.g. [4], [2], [5] and [6]. These offer a range of ubiquitous applications in fields like surveillance [7], and pedestrian detection for smart cars [3]. Previous work by Viola et. al outlined a successful pedestrian detector [4] which relied on a static background. Such cues are not suitable for use aboard a moving vehicle. Our AdaBoost variant is built upon the advancements in Rasolzadeh et al. [1], including the response binning approach. Specifically, our implementation includes the use of real-valued weak classifiers [8] as seen in Viola and Jones [9] and referred to there as the *RealBoost* algorithm. See Algorithm 1.

The aim of this work is to extract as much information from the same weak classifiers without adding a significant processing load to the final real-time system. Many of the previous methods for forming hypothesis do not extract as much information from the base features as is available. By constructing improved models, and therefore improved hypotheses, we are able to greatly improve the strength of the final classifier.

```
$trainClassifiers(X, Y, F)$ :
$X = \{x_1, x_2, ..., x_N\}$, the set of example windows
$Y = \{y_1, y_2, ..., y_N\}, y_i \in -1, 1$, are the corresponding labels
$F = \{f_1, f_2, ..., f_M\}$, the set of filters
$D_1(i) = 1/N$
For $t = 1, ..., T$ (or until the desired rate is met)
  1) Train classifiers $h_j$ using distribution $D_t$. The classifier takes on two possible values: $h_+ = \frac{1}{2}ln\left(\frac{W_{++}}{W_{+-}}\right)$ and $h_- = \frac{1}{2}ln\left(\frac{W_{-+}}{W_{--}}\right)$ for positive and negative examples respectively. $W_{pq}$ is the weight of the examples given the label $p$ which have true label $q$.
  2) Select the classifier $h_t$ which minimises
       $$Z_t = \sum_{i=1}^{N} D_t(i)exp(-y_i h_t(x_i))$$
  3) Update distribution $D_{t+1}(i) = \frac{D_t(i)exp(-y_i h_t(x_i))}{Z_t}$
The final strong classifier (cascade stage) is
       $$H(x) = sign\left(\sum_{t=1}^{T} h_t(x)\right)$$
```

Alg. 1: The RealBoost version of classifier training and boosting.

II. REAL-VALUED ADABOOST WITH RESPONSE BINNING

Real-valued classification methods can be more suitable than binary AdaBoost because they gain significant strength by providing a useful confidence rating for each weak hypothesis [10]. A confidence measure is attained by taking the magnitude of the response to the weak hypothesis. This is used to rank potential classifiers and assemble a strong classifier. When classification speed is required one usually builds a cascade of strong classifiers as shown in Figure 1. Classifiers later in the cascade often have more complex features and are slower to evaluate.

A. Optimal thresholds and General classifiers

In a general classification task a *Maximum A Posteriori* (MAP) rule, which seeks to minimise the error rate over the training data, is often applied.

1-4244-0602-1/07/$20.00 ©2007 IEEE.

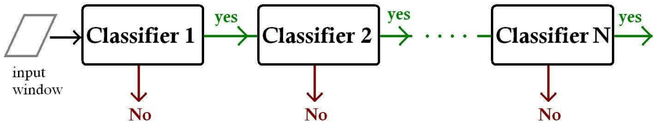

Fig. 1. The cascade architecture. A series of image subwindows are passed into the cascade for evaluation. Each strong classifier will either reject the image as a negative sample or pass it onto the next classifier in the cascade. The stages are trained using *RealBoost* and the outer threshold is adjusted to the desired level for that stage.

Consider a multiclass $\omega_1, \omega_2, ..., \omega_C$ classifier dividing the feature space Ω into C separate regions $R_1, R_2, ..., R_C$, where all points within R_i are labelled ω_i and that $\Omega \supseteq \bigcup_{i=1}^{C} R_i$.

The probability of error for the classifier is then given by:

$$P(error) = \sum_{i=1}^{C} \sum_{\substack{j=1 \\ j \neq i}}^{C} P(x \in R_i, \omega_j) =$$

$$= \sum_{i=1}^{C} \sum_{\substack{j=1 \\ j \neq i}}^{C} \int_{R_i} P(\omega_j|x) p(x) dx$$

Since the total probability is $\sum_{i=1}^{C} \sum_{j=1}^{C} P(x \in R_i, \omega_j) = 1$ we have that,

$$P(error) = 1 - \sum_{i=1}^{C} \int_{R_i} P(\omega_i|x) p(x) dx$$

For the two class problem, the error is minimised by choosing $x \in R_1$ when $P(\omega_1|x) \geq P(\omega_2|x)$ and vice versa. However, it must be accepted that the gained values for $P(\omega_1|x)$ and $P(\omega_2|x)$ are dependant upon the degree to which the model represents the true relative probability of the positive and negative classes given infinite knowledge. A key issue is that many classifiers base their discriminative decisions on oversimplified models of the training sets.

B. Single Threshold Discrimination

Classical implementations for boosting Viola-Jones features, prior to [1] tended to choose a single threshold as the decision boundary for the feature response. See Figure 2.

The boosting algorithm chooses features according to this single threshold rule, which can fool an observer into thinking that the method is selecting all of the most discriminant features. Rasolzadeh et al. [1] notes statistical analysis shows that more than 90% of features are non-discriminative under single threshold discrimination. This is usually due to multimodal distributions or overlapping modal peaks. See Figure 3. While single thresholding sees these features as non-discriminative they are often more discriminative given a better model.

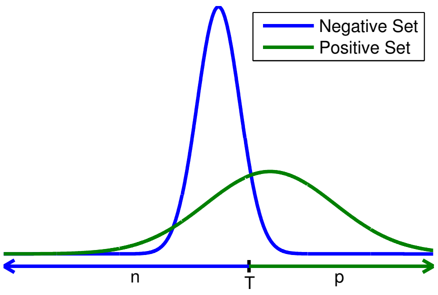

Fig. 2. Classical Single Threshold Hypotheses. This method performs well when the positive and negative probability distributions are unimodal and where modal peaks are separated by significant distance.This allows us to divide the positive and negatives sets at some threshold T.

C. Parameterising the Gaussian case

Rasolzadeh et al. [1] propose two improved methods. The first and simplest is to parameterise the Gaussian case and the second is the response binning method, see Section II-D. By assuming the distributions to be Gaussian, two improved thresholds can be found. See Figure 3.

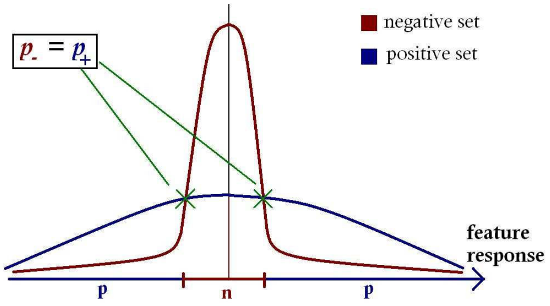

Fig. 3. By parameterising the two distributions with a Gaussian model, two intersection points can be found which will represent the two thresholds. This image is taken courtesy of [1]

The assumption that the distributions are Gaussian, while not entirely valid, is a fairly good representation of the distributions and does lead to far more accurate results. However, for a truly accurate model of the positive and negative distributions which allows us to construct the best possible real-valued response, a non-parametric method such as response binning is adequate.

D. Response Binning

In Rasolzadeh et al. [1] the discriminating strength of a given hypothesis is improved using a response binning method, hereafter referred to as the RB method. The RB method models the positive and negative training sets more accurately. It includes a suitable model which covers the possibility that the positive and negative response distributions are multi-modal. However, the RB method is prone to overfitting if too many bins are used and underfitting if too few bins are used. Indeed, it is even possible that on some particularly 'unfriendly' response distributions that the bins are underfitting at the modal peaks AND overfitting in sparse regions of the distribution. Many methods from the statistical toolbox can be used to overcome such problems, however, the RB method is designed to cost little more than the extremely fast method given in [4]. Thus, any solution to the problem of overfitting and underfitting should avoid slowing down the computation further.

In the RB method a model is constructed by summing the weight of the positive and negative distributions in C separate bins $R_1, R_2, ..., R_C$. This is similar to constructing a histogram of the distribution with C bins, however, instead of summing the instances of training examples with a particular response, we calculate the sum of the weight falling in a given bin. The positive and negative weight distributions $H_+(f_i)$ and $H_-(f_i)$ are computed for C bins and their relative inequality is stored. This inequality of the histogram bins makes up the final MAP rule for the feature.

$$MAP(f_i) = \begin{cases} 1 & \text{if } H_+(f_i) > H_-(f_i) \\ -1 & \text{otherwise} \end{cases} \tag{1}$$

A fast evaluation can then be made for $MAP_i(f_i)$ for any x. This new classifier seamlessly replaces the old one as the new hypothesis. This method significantly improves the quality of the final hypothesis and the robustness of the final linear cascade.

Furthermore, this binary rule can then be expressed as a confidence measure to be used with *RealBoost* , as noted by Schapire&Singer [8]. As in Rasolzadeh et al. [1], W_c^j is defined as the sum of the weights of examples in class c that end up in bin j when evaluated/mapped by features f_m. This gives us:

$$h_m(j) = \frac{1}{2}\log\left(\frac{W_+^j(m)}{W_-^j(m)}\right) \tag{2}$$

III. Adaptive Smoothing for Response Bins

Rasolzadeh includes some notes on choosing an appropriate number of bins for the hypothesis. It is noted that choosing a large number of bins (≡ high resolution) leads to a *noisy* distribution while fewer bins (≡ low resolution) may lead to a less noisy distribution with a loss of accuracy. This, however, does not give the complete picture.

For very large numbers of bins, e.g. the case where the number of bins exceeds the training data, one finds that most bins contain few training examples. In fact, for many reasonable choices for C some sparse regions actually contain no training examples at all. This means that the final confidence measure, Equation 2, makes a decision based on too few training examples for a statistically robust inference. On the other hand, fewer response bins can lead to significant underfitting. For example, near a sharp modal peak the smoothing effect of using too few response bins can mean that the peak is poorly modelled. See Figures 4 and 5.

The problem can be compounded as several weak classifiers are placed together in a single strong classifier. Two compounding effects have been identified which produce poor weak classifiers when too few or too many response bins are used.

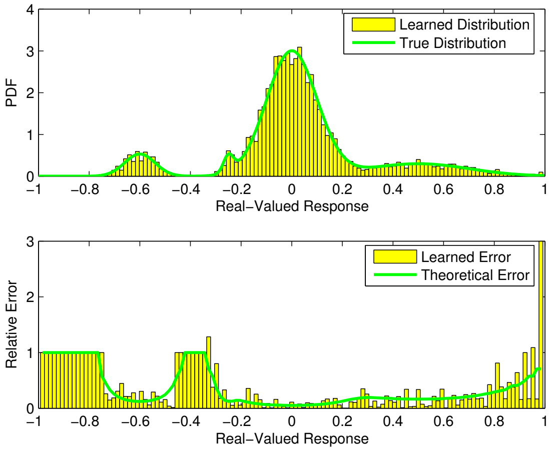

Fig. 4. When using 128 Bins the relative error over a population of 4000 samples is large over sparse regions of the distribution. This leads to a poorly trained hypothesis for these sparsely trained regions.

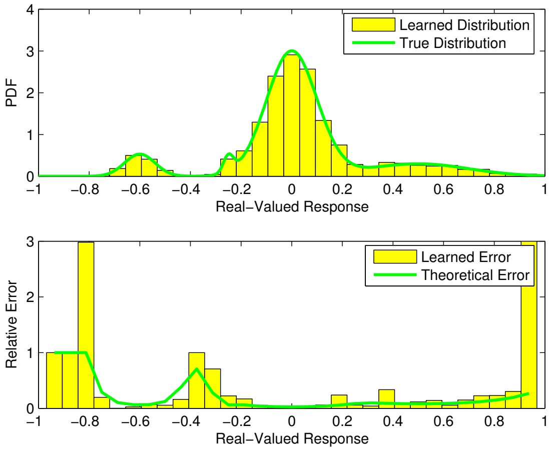

Fig. 5. When using 32 Bins the relative error over a population of 4000 samples is less significant in sparse regions of the distribution. This improves accuracy in the sparse regions with a loss of resolution at the modal peaks.

Firstly, if few bins are used, resolution is lost around the modal peaks and in the case of Haar features there is often a sharp modal peak in the negative set around zero. When the final confidence measure is calculated, using Equation 2, the maximum confidence that an image is a member of the negative set is, as expected, located around the negative modal peak at zero. See Figure 6. When boosting is performed in a linear cascade of strong classifiers one often sets the thresholds so as to have low false negatives and this shifts the decision boundary to the regions of greatest confidence in a negative result. Thus, the decision boundary is actually moved to the poorly modelled region of the hypothesis.

Secondly, when too many bins are used, one poorly models the sparse regions of the distribution. Initially this might

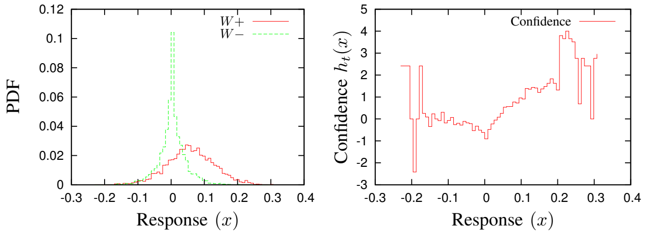

Fig. 6. A discriminative feature response with 64 bins: The left graph shows the positive and negative feature set distributions. The right graph shows the real-valued confidence measure for the same feature. Note that the positive and negative confidence is greatest furthest from the modal peaks. This is due to the lack of training for these response bins.

be dismissed because the low number of training samples implies that few real world images will exhibit that particular response. That is, one poorly models the "rare event" regions of the distribution. However, as more and more weak classifiers are added to a stage the compounded effect of these poorly modelled regions grows in significance. Since the final strong rule is given by $H(x) = sign\left(\sum_{t=1}^{T} h_t(x)\right)$, it is clear that if any of the weak classifiers $h_t(x)$ contain an extremely inaccurate confidence measure, for some x, then the final classifier $H(x)$ may also be inaccurate. Figure 7 shows the effect of using many bins on the positive and negative distributions as well as the final hypothesis $h_t(x)$.

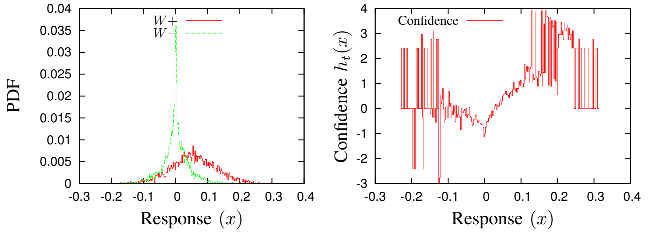

Fig. 7. A discriminative feature response with 256 response bins: The left graph shows the positive and negative feature set distributions. The right graph shows the real valued confidence measure for the same feature. Note how noisy the confidence measure becomes farther from the modal peaks. This is particularly significant for this larger number of bins.

Given the problems detailed above, our aim is to solve the problem of overfitting and adequately model the feature response. It is also very important that any solution evaluates quickly. This excludes a number of the classical response modelling solutions provided by the statistical machine learning toolbox. For this reason, the solution suggested here uses the same method of storing the response and similar means of evaluating it as in Rasolzadeh et al. [1].

Smoother response distributions are acquired as follows: Let $W_c^j(m)$ denote the sum of the weights of samples in class c that fall in bin j for some feature f_m.

$$W_c^j(m) = \sum_{\substack{f_m(x_i) \in bin_j \\ y_i = c}} \omega_i$$

The weights ω_i are used instead of summing the training samples in keeping with the standard adaptive boosting method in the AdaBoost algorithm. This method prioritises

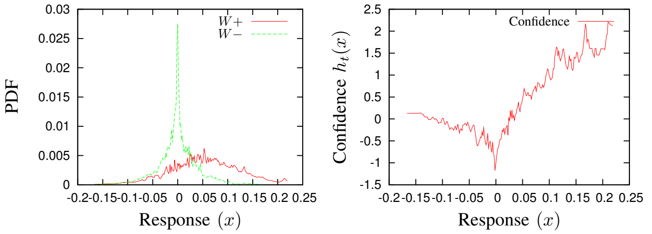

Fig. 8. A smoothed feature response with 256 response bins: The left graph shows the positive and negative feature set distributions. The right graph shows the real valued confidence measure for the same feature. Compare the confidence noise levels to those found in Figure 7.

the importance of some harder training examples by assigning it a weight as the algorithm adds classifiers into a stage. More information on the specific use of the weights can be found in [8] and [1]. If any weight W_c^j is found to contain insufficient weight for the bin to hold a meaningful representation of the true distribution, then the value of W_c^j is altered to contain an average of the weights ω_i in the neighbourhood N of minimum radius r centred at bin j such that the new W_c^j is greater than some threshold T.

$$W_c^{*j}(m) = \sum_{\substack{f_m(x_i) \in N \\ y_i = c}} \omega_i$$

As formulated above, this method would be unreasonably slow. This is because it requires examination of all the training responses $f_m(x_i)$ for each bin requiring smoothing and for a number of possible radii, until some suitable r can be found. Instead, this method tests only the set of radii r, such that the radius exactly includes bin j, and some integer n of its neighbouring bins on each side. Once a radius containing enough weight is found, the weight of the outermost two bins of the neighbourhood is scaled by a scaling factor α until $W_c^{*j} = T$.

$$W_c^{*j} = \sum_{k=1-r}^{r-1} W_c^{j+k} + \alpha(w_c^{j-r} + w_c^{j+r})$$

This method quickly finds a support region around the bin in question which allows a meaningful hypothesis $h_t(x)$ to be constructed for all bin ranges R_i. An appropriate threshold T can be found empirically. For our pedestrian training set it has been found that a threshold T requiring about 1% of the total weight is appropriate for a training set of 10000 positive and 10000 negative samples. This is roughly equivalent to requiring at least 200 samples in any support region.

This process requires a small amount of extra work to be done during the training of a classifier. It stores the final hypothesis in the same format as the RB method and can therefore be evaluated in similar fashion at run-time. However, in the interest of increasing our accuracy when referencing the hypothesis, the new method also applies linear interpolation to the final hypothesis. Doing this requires memory references of two adjacent memory addresses instead of just one. However, the additional cost of referring to the second value in memory is usually minor, making the new method almost as fast as the RB method.

IV. Experimental Evaluation

While the method presented here will provide benefits to almost any classification task using Haar-like features and a variety of other features, we evaluate our method on a pedestrian training set. Pedestrian detection is generally considered harder than front-on face detection or road sign detection. This is because of the variety of poses exhibited by pedestrians and the tendency for clothes and movement to morph appearance. Such a dynamic training task is ideal for testing the new method since it is aimed at learning hypotheses which are as general as possible given any emerging patterns. For example, on the task of front-on face detection, positive and negative distributions tend to separate clearly which allows methods such as single thresholding to perform well. However, for pedestrians the separation of positive and negative sets is often far more subtle.

These results are compared to those found using the RB method in order to show the effect of smoothing when compared to the unsmoothed method. They are not compared to the naive single threshold method as it has already been shown to perform worse [1]. In order to gain the most thorough insight possible these results are compared to Rasolzadeh's RB method over a number of different training tasks. Note, that the underlying system being used is the same system as that used by Rasolzadeh et al, with a few minor improvements. A comparison to the result shown in [1] is not shown as some improvements to the training data and *RealBoost* have since been made. Instead, the best possible RB method results are compared to the Smoothed Response Binning results using the same experimental system.

A. Choosing a Suitable Number of Response Bins

Given the tradeoff between high and low resolution response distributions, it is worth comparing the effect of various resolutions on a large training task. This first set of experiments was performed for 32, 48, 64, 128, 256 and 512 bins for each method. The training set was made up of 8000 positive and 8000 negative samples. Validation is performed using 2000 positive and 2000 negative unused samples. Classification is performed by a classifier cascade with 6 stages. The stages were comprised of 6, 10, 12, 14, 18 and 40 features respectively. It is impractical to show all the results for each of the 12 experiments. Shown are a selection of the most relevant results, see Figure 9.

For this experiment the RB method peaks in accuracy at 64 bins, while the new method continues to improve with increasing resolution. However, the benefit of using higher resolutions becomes insignificant after 256 bins. In many cases the error rate is halved by using the smoothed method. This is a very significant reduction in error because the error rates are already quite low.

It was also found that the smoothed response binning method does not overtrain as quickly as the original method. This means that many more features can be added to a stage before the *RealBoost* algorithm suffers starvation. To test this, a large stage of 120 features was constructed with the same training set and resolutions as above. See Figure 10.

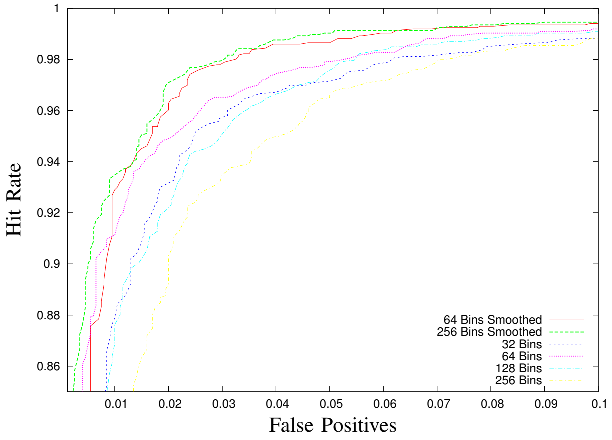

Fig. 9. The ROC-curves for a 6 stage classifier. The smoothed response binning and simple response binning methods are examined for different response bin resolutions. The superiority of using smoothed hypotheses is clear. While the classifier using 64 bins is fairly successful the detrimental effect of increasing or decreasing the resolution is clearly shown.

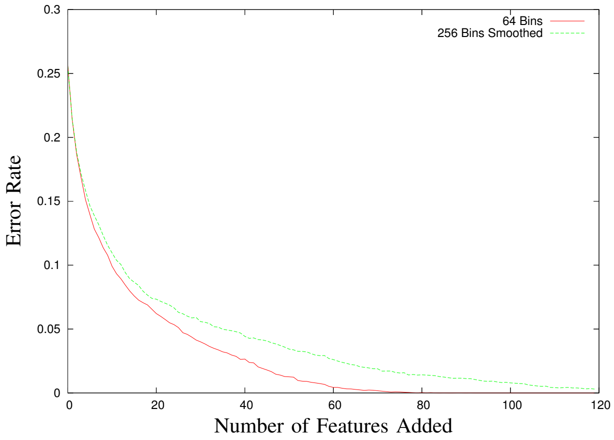

Fig. 10. Estimated Errors for new features as they are added to a large 120 feature stage. The error estimates for smoothed hypotheses are higher because the hypotheses are not overfitting. For the unsmoothed response binning method the hypotheses overfit, which leads to overly optimistic error estimates. Where the error reaches zero the *RealBoost* algorithm is basically starved and is gaining very little as additional features are added.

Here we see that the new method continues to improve even after the addition of 120 features while the RB method has already become almost overtrained.

B. Training the Smoothed Response Method

The original RB method requires that sufficient data is available to effectively train each bin. Even on 16000 samples (8000 positive and 8000 negative) the method still overfits. This is not the case for the new Smoothed Response Binning method, which leads to a dramatic decrease in the amount of training data required to create an accurate hypothesis. To show this, we trained single stage classifiers with 120 features on training sets of 8000, 4000, 1600 and 800 positive and negative samples. See Figure 11.

As is shown in Figure 11, the newer method can produce better hypotheses on one tenth of the training data. This leads to a significant decrease in the final training time of

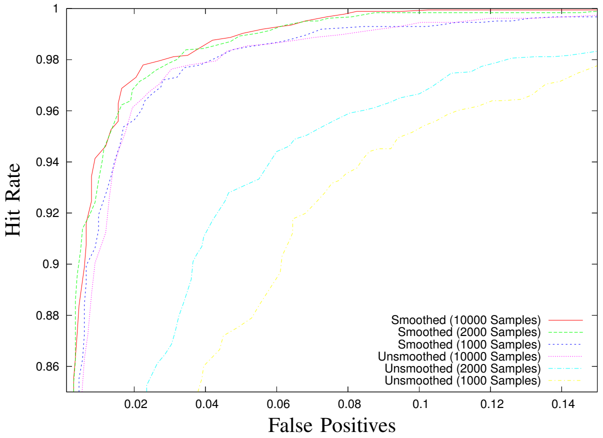

Fig. 11. ROC-curves for a single stage classifier, trained on various sizes of training sets using both the RB method and the smoothed response binning method. The classifier using the unsmoothed RB method quickly looses accuracy as training samples decrease, while the Smoothed RB method continues to perform well.

the system.

V. CONCLUSIONS AND FUTURE WORKS

An improved method for modelling the training responses of weak classifiers for boosting was shown using a smoothed response binning method. This method has been shown to outperform the previous simple response binning method and in turn the classical single threshold approach. It directly answers the problems found in simple response binning and does not overfit as easily. The new hypotheses can be evaluated with only a minimal additional computational cost when compared to the previous method.

The ability of the new method to produce accurate models on less training data is also shown. The new Response Binning Method can be trained on less than one tenth of the data while still achieving the accuracy of the previous method. Thus the new system can be trained in substantially less time.

The new smoothed response binning method seamlessly replaces other real-valued confidence measures that may be used in the *RealBoost* algorithm and the new method could in principle be used to improve any type of weak classifier.

A. Future Works

Improving the models behind a hypothesis can be done in several ways. Other methods may lead to an even more accurate model or a method which is faster to evaluate. In particular, a sum of Gaussians or otherwise parameterised model may succeed in modelling the distribution more accurately. The affect of increased model accuracy has already been explored through the use of higher numbers of bins and thus results are likely to be similar. However, evaluation speed may be improved using different models on certain architectures, e.g. FPGA implementation.

Variations of this method are likely to improve the quality of hypotheses for multidimensional features. This includes multidimensional features which are constructed from several other simple features.

A number of other features for recognition, such as Histograms of Oriented Gradients [11], suffer from the same over training in sparse regions of their distributions. The benefits of a similar method in such cases merits further study.

REFERENCES

[1] B. Rasolzadeh, L. Petersson, and N. Pettersson, “Response binning: Improved weak classifiers for boosting,” *IEEE Intelligent Vehicle Symposium*, 2006.

[2] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” *Computer Vision and Pattern Recognition*, vol. 01, p. 511, 2001.

[3] L. Petersson, L. Fletcher, A. Zelinsky, N. Barnes, and F. Arnell, “Towards Safer Roads by Integration of Road Scene Monitoring and Vehicle Control,” *The International Journal of Robotics Research*, vol. 25, no. 1, pp. 53–72, 2006.

[4] P. Viola, M. J. Jones, and D. Snow, “Detecting pedestrians using patterns of motion and appearance,” *iccv*, vol. 02, p. 734, 2003.

[5] P. Papageorgiou, M. Oren, and T. Poggio, “A general framework for object detection,” *In Proc. ICCV, pp. 555-562*, 1997.

[6] H. A. Rowley, S. Baluja, and T. Kanade, “Neural network-based face detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, 1998.

[7] E. Andrade, S. Blunsden, and R. Fisher, “Characterisation of optical flow anomalies in pedestrian traffic,” *Imaging for Crime Detection and Prevention*, pp. 73–78, 2005.

[8] R. E. Schapire and Y. Singer, “Improved boosting using confidence-rated predictions,” *Machine Learning*, vol. 37, no. 3, pp. 297–336, 1999.

[9] P. Viola and M. Jones, “Fast and robust classification using asymmetric adaboost and a detector cascade,” *Neural Information Processing Systems*, 2001.

[10] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: a statistical view of boosting,” *Technical report, Department of Statistics, Sequoia Hall, Stanford Univerity*, 1998.

[11] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *Computer Vision and Pattern Recognition*, vol. 01, pp. 886–893, 2005.