Proceedings of the
2007 IEEE Intelligent Vehicles Symposium
Istanbul, Turkey, June 13-15, 2007

Boosting with Multiple Classifier Families

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 the importance of creating an even playing field between weak classifiers and classifier families in the RealBoost boosting algorithm. Classifier families are constructed based on Haar-like features in various color spaces, which are then trained simultaneously in RealBoost to create a strong classifier rule. It is shown that the usual method for minimising error at each RealBoost round may express a bias against some weak classifier families. A particular bias toward overfitting features is found. An initial method for achieving parity between families of weak classifiers is applied to improve classification.**

Classification results for various groups of classifier families are shown on pedestrian and sign detection tasks. Particular attention is given to the effect of recently proposed model improvements, including response binning and smoothed response binning. The final system yields significantly lower error rates on classification tasks, and demonstrates the value of color information within the context of the improved methods.

I. INTRODUCTION

Recently, many pattern recognition methods have been introduced and applied to tasks, such as, face, pedestrian or road sign detection. In particular, the works of Viola and Jones [1] has led to a number of descendant methods based on the popular Haar-like features. Many of these methods improve classification by employing various additional cues, such as, background motion [2] for pedestrian detection, color characteristics for sign [3] or face detection [4], optical flow and depth motion for human detection [5], and histograms of oriented gradients for human detection [6].

More recently, Rasolzadeh et al. [7] has proposed improvements to the way the responses from Haar-like features are modelled to form the weak hypotheses. This involves a method of *response binning* to form a histogram-like distribution of the positive and negative responses. A similar use of histogram-like models is found in Liu et al. [8]. These models increase the discriminative power of individual weak classifiers. Unfortunately, they are prone to overfitting. Rasolzadeh's [7] solution was to select a number of bins for the models which minimises overfitting and underfitting. Overett et al. [9] showed that such models may both overfit and underfit at different sections of the model near and far from the modal peaks. The proposed solution was to smooth the response binned models based on the weight of the training data represented in each bin. This allowed more detailed models with more bins and further increased the discriminative power of individual weak classifiers.

Both [7] and [9] used the real-valued variant of AdaBoost, which is generally called RealBoost. This replaced the binary thresholding method in AdaBoost to incorporate a real-valued confidence measure as proposed by Schapire et al. [10]. RealBoost considers the confidence of a weak classifier in a particular decision given a particular response.

It is the combination of the new Response Binning Method (RB-method) and the Smoothed Response Binning Method (SRB-method) with the real-valued confidence measures of [10], which lead to the issues discussed in this paper. The RealBoost algorithm assumes the validity of the confidence measure, however characterised, in expressing confidences which are at least valid comparable to each other. In other words, RealBoost will choose classifiers appropriately *iff* the classifiers can be meaningfully ordered in terms of their discriminative confidence over the training set. This is usually true when only one kind of classifier family is used. If two classifier families are used (i.e, RealBoost has two families of features to select from) and one family tends to overfit more than the competing family then RealBoost will *favour the overfitting classifier* family. To our knowledge, no prior investigation of these issues has been conducted for such feature families or model implementations in RealBoost.

This paper outlines an initial method for overcoming this bias toward overfitting features and discusses some of the major issues which need to be overcome to create a robust solution to this issue.

Additionally, the richness of color information is explored using the improved methods. An example, of color pedestrian images from the training database is shown in Figure 1.

II. FAMILIES OF WEAK CLASSIFIERS

Families of weak classifiers are created to make up a pool of classifiers with similar characteristics from which the RealBoost algorithm will choose. The families are limited to one color space and one method for constructing weak hypotheses (one modelling method). We will consider the subset of the families, supported by our experimental system. These are produced by creating tuples of color space choice with either the RB-method or SRB-method. Table I shows the 8 families which are considered in this paper. More detail on the families is covered below.

This selection of family boundaries is not necessarily definitive. It is based on our experimental implementation

1-4244-1068-1/07/$25.00 ©2007 IEEE.

Fig. 1. Unnormalised 32x80 color pedestrian images from the training database.

TABLE I
8 WEAK CLASSIFIER FAMILIES

	S	**LAB**	**RGB**	**rgb**
RB-method	S_{RB}	LAB_{RB}	RGB_{RB}	rgb_{RB}
SRB-method	S_{SRB}	LAB_{SRB}	RGB_{SRB}	rgb_{SRB}

and consideration of the likely selection parity of the families in RealBoost. For example, the RGB color space could be further divided into Red, Green and Blue families.

A. Feature Types within Families

For simplicity, we limit all families to the five well known Haar-like feature types, see Figure 2. These describe horizontal and vertical edges and lines, as well as diagonal lines [1].

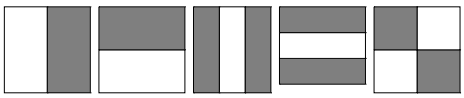

Fig. 2. Haar-like Basic Feature types

B. Color Spaces

While many color spaces may be considered for recognition tasks, not all are suitable for fast real time evaluation using precomputed methods such as the integral image and Haar-like features. It is difficult, for example, to evaluate the Hue values from the HSV color space using an integral image. This is because the hue channel is circular and evaluating a sum of hue values for an image region tells us little about the 'average' hue in that region.

Thus, we constrain ourselves to the following color spaces:

1) **S** - greyscale intensity.
2) **RGB** - consisting of the usual Red, Green and Blue color channels.
3) **rgb** - Intensity normalised values where r=R/S, g=G/S, and b=B/S.
4) **LAB** - Uses the A and B color channels of the CIE L*A*B* [11] color space and drops luminosity, which is considered to be covered by the S (intensity) color space.

C. Improved Models

Features such as those shown in Figure 2 produce a raw feature response x by subtracting the sum of intensities in the shaded region from the sum of intensities in the light region. This is then compared to a trained model of the positive and negative training sets which yield a final real-valued response indicating the likelihood of a given outcome (a hypothesis $h_t(x)$). Figure 3 shows four possible methods of evaluating a feature response.

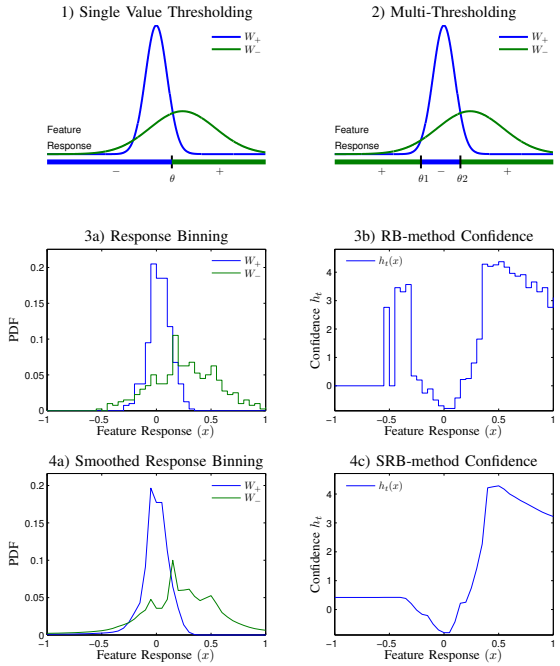

Fig. 3. Various Hypothesis Modelling Methods

1) **Single Value Thresholding** - This method uses a single threshold to discriminate between the positive and negative training sets. It relies on a difference of means in the training sets.
2) **Multi-Thresholding** - This method uses two thresholds to discriminate between outcomes. A discriminative result can still be found when modal peaks of the training sets overlap (which Rasolzadeh [7] found was often the case).
3) **Response Binning** - Here the raw feature responses x, from training data, are placed into a histogram-like model. This method tends to suffer from overfitting. See 3a and 3b in Figure 3. More information can be found in [7].
4) **Smoothed Response Binning** - This method takes the model from the previous method and adaptively smooths it based on the amount of training weight represented in each region. This has the effect of ensuring that no bin forms a hypothesis based on too little training data. This method is outlined in [9].

III. Real-Valued AdaBoost

In the Real-Valued AdaBoost system 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 classification rule. This algorithm is outlined in Algorithm 1.

$trainClassifiers(X, Y, F)$:
$X = \{x_1, x_2, ..., x_N\}$, 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

$$Z_t = \sum_{i=1}^{N} D_t(i) exp(-y_i h_t(x_i)) \quad (1)$$

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: RealBoost

Of particular interest to us is the feature selection in Step 2 (see Algorithm 1). Here the algorithm selects the classifier h_t which minimises Equation 1.

This selection criteria is shown in [10] to be an optimal choice given the assumption that the confidence measure $h_t(x_i)$ is valid and the weights are distributed as per Step 3. However, confidence measures are inherently flawed because they are only based on training data and the success of the modelling method used (see Section II-C). Some modelling methods tend to overfit or underfit more than others. The degree to which this happens is unique to the color space used. It as also sensitive to the redistribution of weight during each round and the difficulty of the 'problem at hand' in a given round of the RealBoost algorithm.

A. RealBoost Bias Towards Overfitting Features

Consider what happens when RealBoost chooses between two families of classifiers where one family (A) is overfitting on the training data while the other (B) has a more representative model of the real-world distribution. For overfitting family A, the values of $exp(-y_i h_t(x_i))$ will be very small thus lowering Z_t^A for all possible t's from set A. Alternatively, the more representative models in set B will exhibit higher values for Z_t^B.

This effect can be most easily observed when we run RealBoost with two Families made up of the same feature types and in the same color space but with the RB-method and the SRB-method respectively. The RB method is known to overfit while the SRB-method produces a more representative model. Figure 4 shows the disparity between the potential Z_t values in features from the S_{RB} and S_{SRB} families respectively. From this figure it is clear that the overfitting of the S_{RB} family clearly biases the RealBoost system toward choosing the overfitting features.

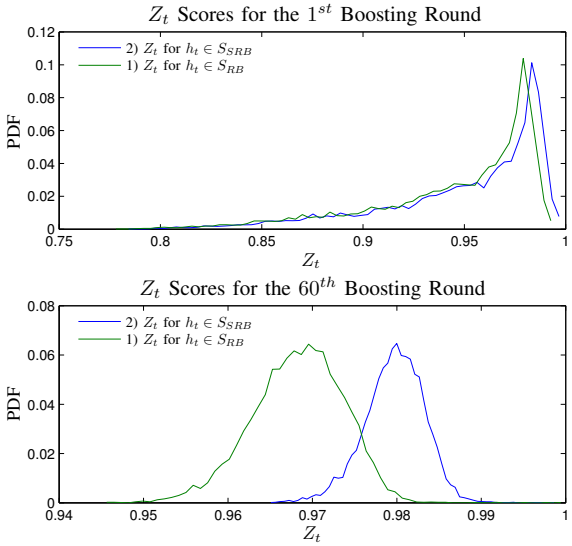

Fig. 4. Disparity between Z_t scores from each of the 2 families. Plots are made in the 1^{st} and 60^{th} training round to show how this disparity grows during training. Greater detail of this changing disparity is found in Figures 6 and 7. Experimental parameters used are shown in Table II.

IV. Creating Parity Among Feature Families

Figure 5 shows a high-level schematic of the problem RealBoost has in choosing features from multiple families.

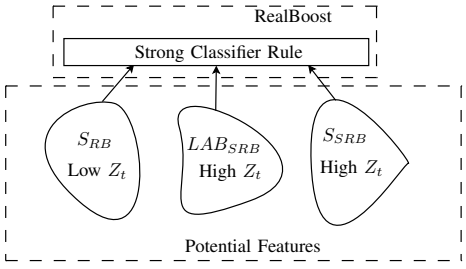

Fig. 5. RealBoost Feature Selection. Optimistically low Z_t scores in S_{RB} will create selection bias for this feature due to its greater overfitting.

A number of schemes may be used to create an even playing field between classifier families. The most naive approach would be to estimate the average overfitting and its effect on Z_t scores between two families and apply some scaling factor to all Z_t scores from one family. For example, we might take:

$$Z_t^*(Z_t) = \begin{cases} Z_t & \text{if } h_t \in S_{SRB} \\ \delta Z_t & \text{if } h_t \in S_{RB} \end{cases}$$

Unfortunately, the problem at each boosting round becomes more difficult as the previous classifiers strengthen the strong classifier rule $H_t(x)$. Early in the boosting rounds RealBoost is able to find very discriminant features. When distributions of the positive and negative data sets are moderately disjoint the effect of overfitting is minor. In later rounds, when RealBoost is learning more subtle features we find the effect of the overfitting is much greater. Hence, this naive approach does not work very well.

A. *Parity Between Two Families with the Same Base Feature*

Our initial solution to the changing unfairness/disparity between the Z_t scores for the S_{RB} and S_{SRB} families is to create a simple model function $\delta(R)$ of the disparities. This adjusts the Z_t score to a less optimistic value based on the boosting round number R. We find the new Z_t^* score as follows:

$$Z_t^*(Z_t, R) = \begin{cases} Z_t & \text{if } h_t \in S_{SRB} \\ \delta(R) Z_t & \text{if } h_t \in S_{RB} \end{cases}$$

To create a model $\delta(R)$ of the disparity as the rounds progressed we ran RealBoost through 150 rounds ($T = 150$) with either the S_{RB} or S_{SRB} families. Figure 6 shows the resulting graph of the minimum Z_t scores in each round.

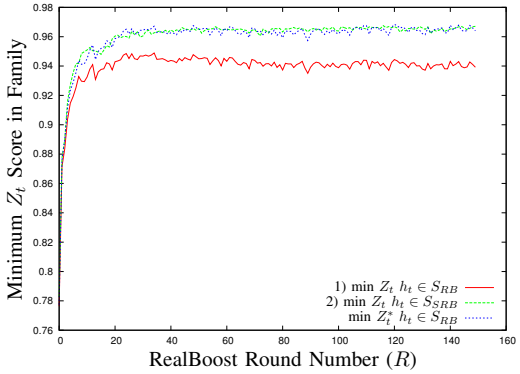

Fig. 6. Minimum Z_t scores from each family as training rounds progress. We see that the minimum Z_t scores are consistently lower from the S_{RB} family than the S_{SRB} family. A third line showing the adjusted $Z_t^* = \delta(R)Z_t$ scores for the overfitting feature is shown. The formation of the function $\delta(R)$ is shown in Figure 7. Experimental parameters are shown in Table II.

Figure 7 shows the percentage difference in Z_t scores from either family. This is used to create a simple function $\delta(R) = \alpha log(R/\beta) + c)$ to adjust the values in each round. Where α, β and c are tuned to generate the best fit possible. Several different δ functions were tested for suitability and it appears that the success of the method is not sensitive to minor changes in the function as long is it smoothly captures the disparity between Z_t scores with reasonable accuracy.

The success of this adjustment in lowering the overall error rate is shown in Figure 8.

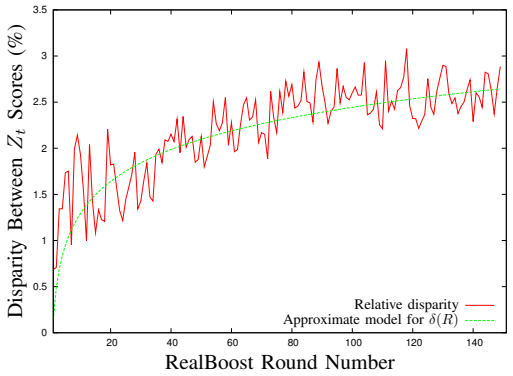

Fig. 7. Relative disparity between Z_t scores from the S_{RB} and S_{SRB} families. This is shown next to a basic approximation $\delta(R)$ of the disparity between the sets.

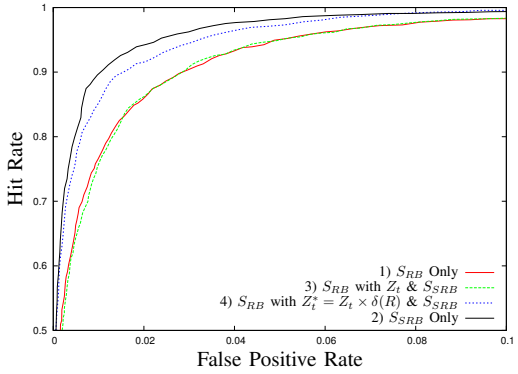

Fig. 8. Pedestrian Detection ROC curves for 2 family feature pool with Z_t^* score replacing the overly optimistic score against an unchanged 2 family feature pool. The graph also shows the ROC curve when the S_{RB} or S_{SRB} family is used exclusively by RealBoost. In experiment 3 the bias is so severe in favour of choosing the poorer feature that the existence of the improved feature makes no significant difference at all. Skeptics will note that the two family S_{RB} with $Z_t^* = Z_t \times \delta(R)$ & S_{SRB} (curve 4) is outperformed by the single family S_{SRB} experiment (curve 2). This is because the S_RB family doesn't not contain any new information. A demonstration of the δ compensation methods ability to improve results between two complementary color features is shown in Figure 9. Experimental parameters are shown in Table II.

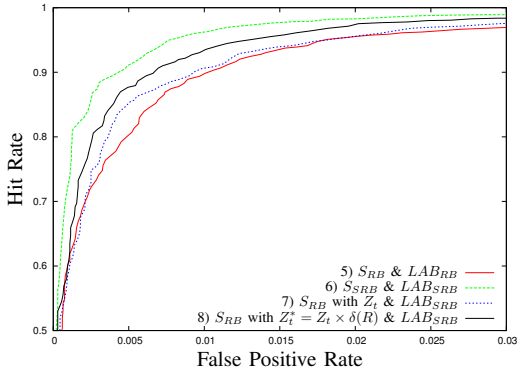

Fig. 9. Pedestrian detection ROC curves for 2 family feature pool with S and LAB family variants. Using $\delta(R)$ to even the playing field between two classifier families clearly lowers the negative effect of the bias toward the S_{RB} family (compare experiments 7 & 8). The best result is achieved by improving the models of both families, i.e, by using the SRB-method, in experiment 6, see Section IV-B. Experimental parameters are found in Table II.

TABLE II
EXPERIMENTAL PARAMETERS

Exp#	Features
1	10KS_{RB}
2	10KS_{SRB}
3	10KS_{RB} + 10KS_{SRB}
4	(10KS_{RB} with Z_t^*) + 10KS_{SRB}
5	10KS_{RB} + 20KLAB_{RB}
6	10KS_{SRB} + 20KLAB_{SRB}
7	(10KS_{RB} with Z_t) + 20KLAB_{SRB}
8	(10KS_{RB} with Z_t^*) + 20KLAB_{SRB}
9	10KS_{SRB} + 30KRGB_{SRB}
10	10KS_{SRB} + 30Krgb_{SRB}
11	10KS_{SRB}
12	10KS_{SRB} + 20KLAB_{SRB}
13	10KS_{SRB} + 30KRGB_{SRB}
14	10KS_{SRB} + 30Krgb_{SRB}

Exp#'s	Training	Rounds	ROC-Valid	Subject
1-2	9K+,9K−	150	2K+,20K−	Pedestrians
3-10	9K+,9K−	200	2K+,20K−	Pedestrians
11-14	9K+,9K−	10	2K+,20K−	Signs

B. Parity Between Families with Non-Parametric Models

Both the RB-method and the SRB-method provide strong parametric models of the training data for Haar-like features in the trialled color spaces. So what happens when all feature families available to RealBoost are exclusively using either method? Is it valid to assume that the degree of overfitting in all models is now similar, and therefore, that we do not need to adjust the Z_t score?

In the case of the **RB-method**, the number of response bins used is known to affect the degree of overfitting [7], [9]. Generally, a lower number of bins is used than in the SRB-method. Furthermore, some features in the family will have a narrower distribution than others and will use fewer bins. The greatest overfitting occurs for features within a family which spreads more evenly over the model. These overfitting features will be selected by the RealBoost algorithm beyond their true usefulness in building a strong classifier. This means that the Z_t scores are not necessarily comparable within a family.

In the case of the **SRB-method**, we are able to use a higher number of response bins as this does not lead to overfitting. Exceptionally unfriendly distributions can be imagined which would cause this method to overfit, underfit or produce significant artifacts. Such distributions have not, in our experience, been observed in Haar-like feature based families in any color space. Rather, all distributions observed appear to be well suited to this method. For this reason, we conclude that the SRB-method feature families are indeed comparable in terms of Z_t scores. Figure 9 shows the effect of using the SRB-method on two families in different color spaces.

C. Parity Between Families with Parametric Models

We have investigated the suitability of Real-Valued classifiers based on Gaussian models for the positive and negative distributions. Significant success was had [7] in using Gaussian models with the binary Multi-Thresholding method and AdaBoost (see Figure 3 and Section II-C). An attempt was made by us to formulate a version of the same technique using RealBoost. Causing initial surprise, this was found to be very poor. Closer examination revealed that the Gaussian models produced precisely predicted the location of the two optimal thresholds θ_1 and θ_2. However, the Gaussian models produced very poor real-valued confidence measures at points on the distributions far from θ_1 and θ_2. This inserted major inaccuracies into the RealBoost strong classifier rule.

Clearly, some feature responses, including those based on Haar-like features, will exhibit distributions which are more Gaussian than others. Similar behaviours are likely to occur with other parametric models. The degree to which any parametric model overfits or underfits may be *individual* to a single feature or *associated* with a particular family. For example, some features may produce fairly Gaussian distributions in one color space but not another. If Gaussian models are used in both families then we may want to compensate for the greater failing of the model in the less Gaussian family. In the typical case of a Haar-like feature in some color space we believe this to be an inadequate solution. The features, within a family, vary so much in their distributions that attempting a 'one size fits all' adjustment $\delta(R)$ for a whole family is hardly meaningful.

D. Modelling the Overfitting

The adjustment of Z_t by $\delta(R)$ was formed by comparing two feature families with similar base features over several experimental boosting rounds. This is able to guide RealBoost toward better choices. The downside is that $\delta(R)$ has a limited ability to predict the effect of overfitting when RealBoost is building the strong classifier rule with other feature families of unknown effect on the strong classifier rule.

A possible solution to this is to validate all hypotheses against validation data to produce a per feature offset. However, a thorough validation stage prior to RealBoost's selection stage would be cumbersome even for offline training.

A more feasible solution may be to estimate the uncertainty of a model by measuring the rate at which it converges to a stable hypotheses during training.

V. COLOR SPACE SUITABILITY

While much of this paper concerns itself with issues of parity between Z_t scores in RealBoost, it also contains the results of our experimental work in combining a number of our improvements to our RealBoost experimental system. Particularly, it is the first examination of the richness of information to be gained from color when using our improved modelling methods. Figure 10 shows ROC curves for a variety of combinations of the feature families. Details of the experimental parameters used in these experiments can be found in Table II. Please note, that these experiments were designed to show the benefits of using a mixed pool

of features, hence the use of a single stage only. In a real system, a cascaded approach, as in [1], should be used.

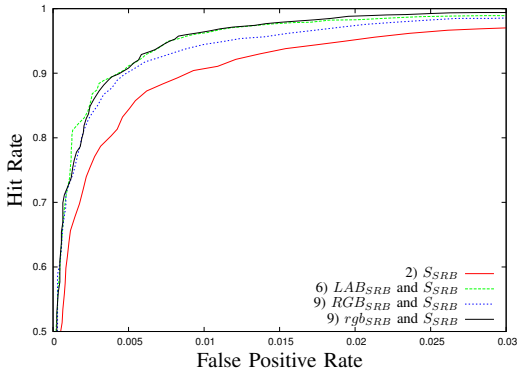

Fig. 10. Pedestrian detection ROC curves comparing the contribution of color to overall recognition. Experimental parameters are shown in Table II. For a false positive rate of 1% color improves detection from 90.7% to 96.4%.

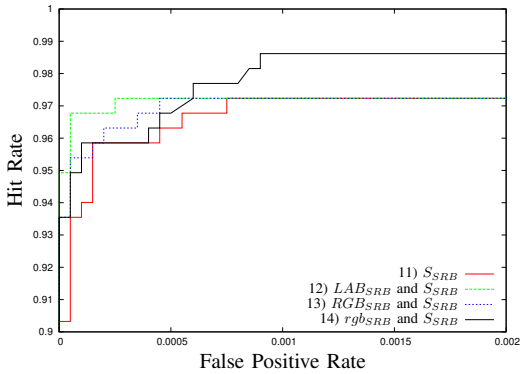

Fig. 11. Sign detection ROC curves comparing suitability of different color spaces. Experimental parameters are shown in Table II. For a false positive rate of 0.01% the LAB color information improves detection from 94% to 96.7%.

Figures 10 and 11 show clearly that color information is a useful cue for improving the quality of an overall classifier for both pedestrian and sign detection tasks.

The addition CIE LAB color space to monochrome cues produces the best results on the pedestrian detection task. It is possible to make all the SRB-method based feature families from Table I available for selection by RealBoost, however, the gains are likely to be small. There is also a high cost to preparing the various precomputed integral images of the different color spaces. Thus we suggest the use of the CIE LAB color space as an additional color cue for pedestrian detection.

For very low false positive rates the CIE LAB color space is once again the best addition to monochrome information. For higher rates the rgb_{SRB} family contributes more information. However, for live applications one usually requires the false positive rate to be very low due to the large number of possible input windows in a single frame of video.

Since the CIE LAB color space requires only 2 more precomputed images for the A and B color channels it is also preferable to the 3 channel RGB and rgb color spaces.

VI. Conclusions

The parity of features available to RealBoost must be considered carefully as RealBoost will exhibit bias toward optimistic and overfitting features. Models can be built to predict the bias and compensate prior to RealBoost features selection. Implementations of such compensation mechanisms yield very positive results. These compensate for the major differences in optimism between different families of features. However, results point towards additional work on an even more robust compensation mechanism, particularly one which is able to compensate for differing levels of overfitting between features of the same family.

Several color spaces are able to provide greater robustness on pedestrian and sign detection tasks. The A and B channels of the CIE LAB color space are particularly good as an additional cue to monochrome Haar-like features on both pedestrian and sign detection. The final strong classifiers, using both monochrome and color cues, achieve much improved accuracy and provide notable robustness on difficult recognition tasks.

References

[1] 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.

[2] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *IJCV*, vol. 63, no. 2, pp. 153–161, 2005.

[3] C. Bahlmann, Y. Zhu, V. Ramesh, M. Pellkofer, and T. Koehler, "A system for traffic sign detection, tracking, and recognition using color, shape, and motion information," in *IEEE Intelligent Vehicles Symposium (IV 2005)*, 2005.

[4] S.-H. Huang and S.-H. Lai, "Detecting faces from color video by using paired wavelet features," *cvprw*, vol. 05, p. 64, 2004.

[5] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance." in *ECCV (2)*, ser. Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, Eds., vol. 3952. Springer, 2006, pp. 428–441.

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

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

[8] C. Liu and H.-Y. Shum, "Kullback-leibler boosting," *Computer Vision and Pattern Recognition*, vol. 01, pp. 587–594, 2003.

[9] G. Overett and L. Petersson, "Response binning: Improved response modelling on weak classifiers for boosting," *IEEE International Conference on Robotics and Automation*, 2007.

[10] R. E. Schapire and Y. Singer, "Improved boosting using confidence-rated predictions," *Machine Learning*, vol. 37, no. 3, pp. 297–336, 1999. [Online]. Available: citeseer.ist.psu.edu/article/singer99improved.html

[11] *Photoshop Lab Color: The Canyon Conundrum and Other Adventures in the Most Powerful Colorspace*. Addison-Wesley, 2005.