Confidence Measurements for Adaptive Bayes Decision Classifier Cascades and their Application to US Speed Limit Detection

Armin Staudenmaier[1,2,3], Ulrich Klauck[1], Ulrich Kreßel[2], Frank Lindner[2] and Christian Wöhler[3]

[1]Hochschule Aalen, Beethovenstraße 1, 73430 Aalen, Germany
[2]Daimler AG, Group Research, P. O. Box 2360, 89013 Ulm, Germany
[3]Technische Universität Dortmund, Otto-Hahn-Str. 4, 44221 Dortmund, Germany

Abstract. This article presents an adaptive Bayes model for the decision logic of cascade classifier structures. The proposed method is fast and robust with respect to multimodal and overlapping distributions and can be applied to arbitrary stage classifiers with continuous outputs. The method consists of an adaptive computation of thresholds and probability density functions which outperform the threshold based decision. It furthermore guarantees high detection rates independent of the number of stage classifiers. Based on this Bayes model different confidence measures are proposed and evaluated statistically and used for merging detection windows. The algorithm is applied to the detection of US speed limit signs under typical driving conditions. Results show that on a single CPU with 3.3 GHz the proposed method yields single image detection rates of 97 % with 0.2 false positives per image running at 13 Hz, and for a different setup a detection rate of 93 % with 0.2 false positives per image performing with 43 Hz for scanning the whole image (752x480 pixels).

1 Introduction

Robust and fast object detection is applicable to a very large set of computer vision based applications. This paper presents a machine learning method for the detection of US speed limit signs. This is a challenging task since strong consistent features like circles or thick edges are missing, in contrast to circular signs. Even aspect ratios and digit types vary.

Most systems for US traffic sign recognition build on the work of Viola and Jones [12, 11] who use strong classifiers composed of weak classifiers boosted [14] in a classifier cascade with an implicit iterative method for setting thresholds. The performance of each stage classifier has influence on all subsequent stage classifiers and is therefore essential for a good overall performance. Most authors therefore try to reduce the number of stage classifiers. Overett et al. [10] argue that each stage likely reduces the detection rate and thus use only three to five stages. Rasolzadeh et al. [3] present a method for the computation of thresholds

and a histogram based method for a single stage classifier and the RealBoost Algorithm.

This paper focuses on the cascaded structure and presents a novel modular approach for an adaptive Bayesian Classifier Cascade. It is shown how the optimal decision values and thresholds of these stage classifiers can be computed with low effort independently of the number and type of stage classifiers while maintaining high detection rates with low false positive rates. Zaragoza et al. [4] present an overview of existing confidence measures with probability based confidences yielding remarkable performance. The proposal is therefore for three different methods of confidence measures based on the Bayesian approach for binary classification problems which can be extracted with little effort by fusion of the confidence values of each stage classifier. The best performing confidence measure is used for merging different detections into one by means of linear combination.

This paper is organized as follows. Section 2 explains the algorithm for the Bayes decision for cascaded classifier structures and propose a confidence measure depending on the stage classifiers. Different settings of the cascaded Bayes decision algorithm are evaluated with respect to their detection and run-time performances in Section 3.1. Additional confidence values are introduced and compared in Section 3.2.

2 Bayes Decision for Cascaded Classifier Structures

In principle, a cascade classifier is a series of ordered stage classifiers. According to [11] the decision function of a cascade classifier h_C containing $h_1, \ldots, h_T$ stage classifiers is as follows:

$$h_C^T = (h_1, \ldots, h_T) = \begin{cases} \omega_0 & \text{if } h_j = \omega_0, \text{ for } 1 \leq j \leq T \\ \omega_1 & \text{else} \end{cases} \tag{1}$$

where ω_1 is the class for objects and ω_0 is the background class. An object is detected only if it is accepted by each stage classifier. A stage classifier typically extracts specific features from the region to be detected and then applies a classification function to obtain a decision. Only samples which are not rejected are passed to the next stage, such that each stage classifier changes the prior probabilities of successive stages. Due to the structure of the discrete decision of the cascade the detection rates of each classifier are multiplicative and therefore adding more classifiers without adaptation of decision functions results in a degradation of the detection performance. By using a Bayesian approach, the prior probabilities of object and background classes can be tracked and the stage classifier detection rates are maintained for the cascade classifier. The prior probability for classes ω_i induced by a classifier h with output $o(x)$ and and its probability density function p is:

$$P_h(\omega_i) = \frac{1}{|\Omega^i|} \int_{x \in \Omega^i \wedge h(x) = \omega_1} p(o(x)|\omega_i) \, \mathrm{d}o \tag{2}$$

with $i \in \{0, 1\}$ and with the set Ω^1 for objects and Ω^0 for background. For each sample in Ω^i the output is accumulated in a frequency table over the output and then normalized and stored which allows estimating $p(o(x)|\omega_i)$ and calculating P_h with a sum. The rule for updating the prior probabilities $P_j(\omega_i)$ after the classification with stage classifier h_j is:

$$P_{j+1}(\omega_i) = P_{h_j}(\omega_i) P_j(\omega_i) \tag{3}$$

where $\Omega_j^1 = \Omega^1$ and $\Omega_j^0 = \{x \in \Omega^0 | h_C(x) = \omega_1\}$ are classified as objects in stage j, therefore the complexity factors for samples are increasing at each stage. The priors used in Eq. (3) correspond to the detection and false positive rates [5] which is shown in Fig. 1(a). A more accurate but also computationally more expensive way is to use Eq. (2) with the actual cascade h_C^j. Initial prior probabilities are set to one since optical transfer is identical for each sample. To prevent vanishing prior probabilities, $P(\omega_0) = \max(P(\omega_0), P_{min})$ with $P_{min} \simeq 10^{-9}$. The final Bayes decision function is then given by:

$$h_j(x) = \begin{cases} \omega_1 & \text{if } p_j(o(x)|\omega_1) P_j(\omega_1) - p_j(o(x)|\omega_0) P_j(\omega_0) \geq 0 \\ \omega_0 & \text{else} \end{cases} \tag{4}$$

where $o(x)$ is the output value and $p_j(o(x)|\omega_i)$ is obtained from the frequency table (see Eq. (2) above). Fig. 1(b) shows the feature extraction part followed

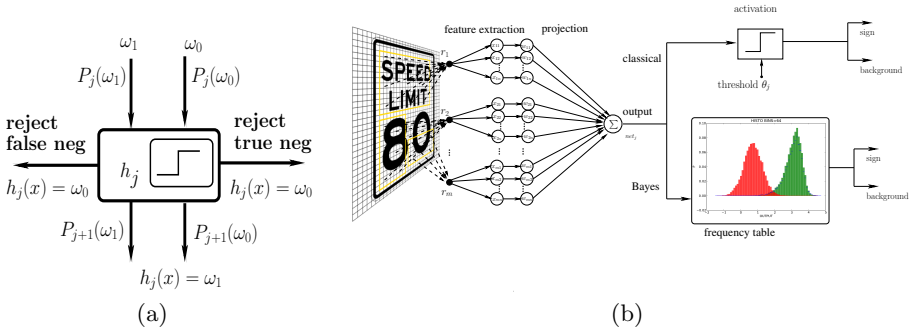


Fig. 1. 1(a): Change of prior probability induced by a stage classifier. 1(b): Extraction of features and decision for a stage classifier. The upper row shows the decision with a classical threshold perceptron while the lower row shows a frequency table used for the decision and tracking of probabilities.

by the decision part which is the classical threshold decision function in the upper row and the Bayes decision function modeled with a frequency table in the lower row as described above. The feature extraction scheme and feature types used are a combination of intensity based and structure tensor features in combination with linear perceptrons [13] as described in [1]. The weights of

the perceptron are computed based on Fisher LDA [6, 8] in combination with a sequential forward selection of features [2] using the training set $\{\Omega_j^0, \Omega_j^1\}$ since no external parameters are involved and the training procedure requires about 50 ms on average. Besides, it can be shown that it is directly related to a minimum squared error procedure [7]. This training procedure is repeated for each new stage j and stage classifiers are added until the prior probability of a negative $P_{h_j}(\omega_0)$ approaches a value of 0.95. If this threshold is lowered, the training will stop earlier resulting in a higher overall false positive rate.

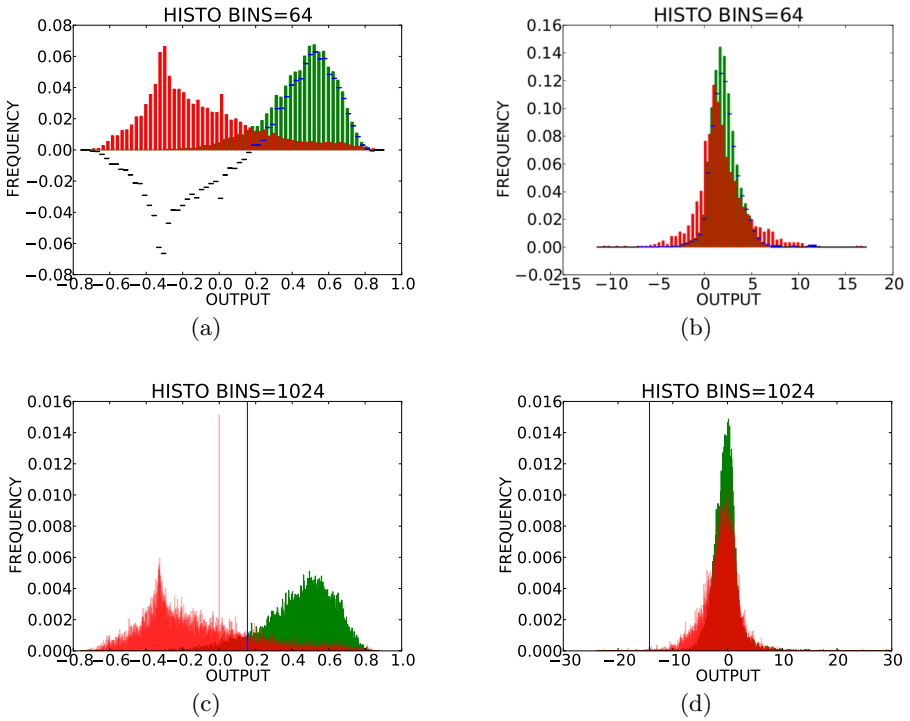


Fig. 2. First row: Frequency tables over the output values of stage classifiers 1 and 79 (see 2(a), 2(b)) out of $T = 211$ with 64 bins. The blue bars are the calculated decision values for a positive decision and black bars for a negative decision for each bin. Second row: Frequency tables over the output of classical threshold perceptron classifier 1 and 65 (see 2(c),2(d)) out of $T = 81$. The threshold is marked as a blue line. To enable more precise positioning, 1024 bins are used. Green bars correspond to output belonging to objects, red bars to output associated with background patches.

Fig. 2 shows frequency tables of stage classifiers. Green bars model the pdf of positives and red bars model the pdf of negatives. In the first row blue bars correspond to detected objects and black bars which are less than zero are rejec-

tions (see Fig. 2(a) and 2(b)). Due to the modified prior probabilities the black bars are close to zero in Fig. 2(b). The lower row shows the threshold decision method. The blue line marks the computed threshold and all values to the right of it are classified as objects. The two peaks in Fig. 2(a) and 2(c) near the output 0.0 are due to homogeneous patches (e.g. sky) which are mapped to zero. To determine the threshold, the frequencies are multiplied by their prior probabilities and the threshold is placed at the point of minimum error. Comparing Fig. 2(b) and 2(d), which correspond to stage classifiers 79 and 65, one can see the advantage of the histogram-based decision which contains three positive decision regions (two small additional regions near output 9.5 and 12.0) and four negative, yielding a lower FPR than the threshold classifier which has a FPR of approximately 99 %.

The Bayes approach described above can also be used for the effective computation of confidences. For high frequencies the classifier is confident since it has gathered a large number of evidences. If the frequencies of both classes are high, there is a conflict and the confidence should be low. By using the prior probabilities for each stage the confidences for negatives are decreasing since it is increasingly likely to encounter positives in higher stages, resulting in a confidence value:

$$\gamma_1(h_c^T, x) = \frac{1}{T} \sum_{j=1\,,\,h_j(x)=\omega_1}^{T} |p_j(o(x)|\omega_1)P_j(\omega_1) - p_j(o(x)|\omega_0)P_j(\omega_0)| \quad (5)$$

3 Experimental Evaluation

The evaluation of different classifiers concerning the overall detection performances is performed in Section 3.1, while in Section 3.2 additional confidence measures which are compared to the confidence in Eq. (5) are suggested and evaluated using two statistical methods.

3.1 Evaluation of Detection Performances

The training uses a set composed of 10000 grayscale images acquired from a test vehicle and displaying US speed limit signs. The images show strongly varying lighting conditions, the presence of cast shadows, as well as structures resulting from the Bayer pattern of the utilized camera. Some typical examples are shown in Fig. 3. The two-digit block with a small border and base widths ranging from 24 to 60 pixels are cut out and used for training, since the positioning of other attributes like the words “speed” or “limit” are variable and additional plate information is integrated into some signs. As training examples for the negative background class ω_0, the images containing signs are scanned and all windows ranging from 24 to 60 pixels which do not overlap any sign are collected, yielding set sizes of 10815 positives and 5×10^8 negatives. For the stage classifiers

the complexity of negative training samples increases since there are two digit combinations, for example on road signs or billboards which can hardly be distinguished from the digits of a speed limit sign. Fig. 3 shows some typical examples

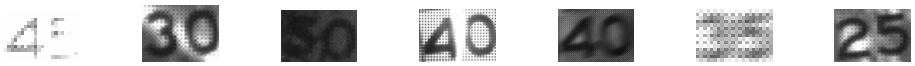


Fig. 3. Speed limit signs acquired under difficult illumination conditions. The two-digit block is used for the detector. The squared structure is induced by the Bayer pattern of the camera.

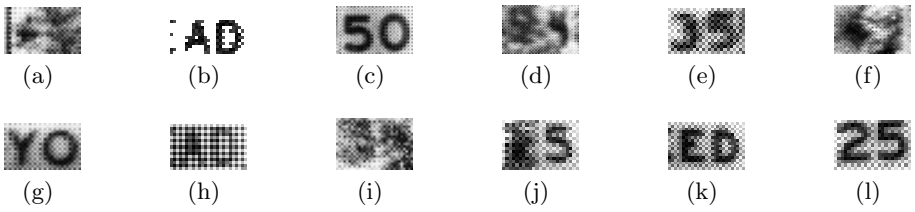


Fig. 4. Normalized false positives for stage classifier 113 out of 118. The patches 4(c), 4(e) and 4(l) originate from images with two signs (left and right) where one has been missed in the labeling process. The patch 4(k) represents the letters "E D" from "S P E E D".

displaying difficult illumination conditions whereas Fig. 4 shows normalized false positives of stage classifier 113 out of 118 stages in total. Very similar patches are classified as signs in high stages, even some patches from real signs which were missed in the labeling process are detected by the classifier. Analyzing the original images corresponding to the false positives, reveals that patches 4(c), 4(e) and 4(l) originate from images with two signs located at each side of the street where one has been missed in the labeling process.

Different classifiers with Bayes decision functions differing in their number of bins for the frequency tables have been trained. For evaluation, Receiver Operating Characteristics (ROC) curves are constructed by successively removing stage classifiers. We also determine the average computational burden of the detection in milliseconds CPU time for scanning the image for signs over the false positives. Fig. 5 shows the ROC curve with detection rates on the y-axis and false positives per image (FPPI) on the x-axis on the left and the cost diagram with average time per image in ms on the y-axis and FPPI on the x-axis. The number of bins used for the frequency tables has different effects on the performance. The prefix "FRQ" means that Bayes decision classifiers with frequency

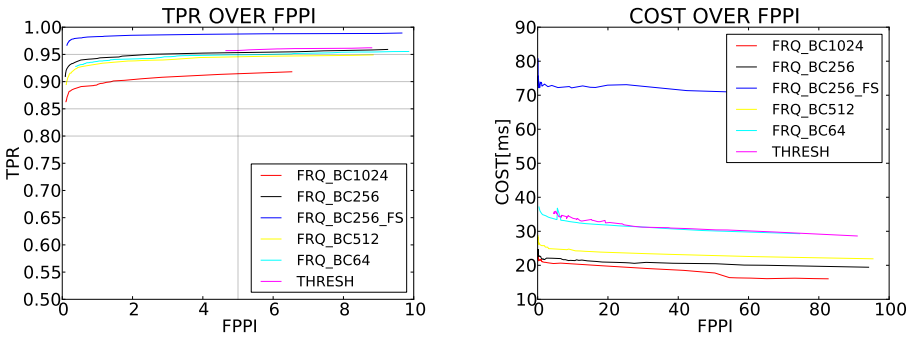


Fig. 5. ROC curves with detection rates (TPR) on the y-axis and false positives per image (FPPI) on the x-axis. The cascades starting with FRQ contain frequency tables with different numbers of bins. The cascade with ID FRQ_BC256_FS e.g. contains 256 bins and uses a finer sampling method. The threshold perceptron cascade with ID THRESH terminates at 5 false positives per image.

tables are used and the number following "BC" corresponds to the number of bins used. The classifier with ID FRQ_BC256 seems to mark a "sweet spot" since the system performs at 43 Hz and the single image detection rates yield 93 % and 0.2 FPPI scanning the whole image (752x480 pixels) on a single processor with 3.3 GHz. Using a finer sampling, detection rates of of 97 % and 0.2 FPPI are achieved running at 13 Hz (see FRQ_BC256_FS). The classical threshold classifier has high detection rates but training stopped at 5 FPPI since the error of the stage classifier was too high. The total number of classifiers used are: FRQ_BC1024=87, FRQ_BC512=159, FRQ_BC256=118, FRQ_BC256_FS=168, FRQ_BC64=211, THRESH=81. This shows that very high detection rates are obtained even for large numbers of stages. The bend appearing for nearly all ROC-curves at low FPPIs can be explained by the fact that the stage classifier encounters positive sign samples in the negative background set as described above. The training times of each classifier vary, e.g. FRQ_BC256=25 hours and FRQ_BC256_FS=29 hours.

3.2 Evaluation of Confidences

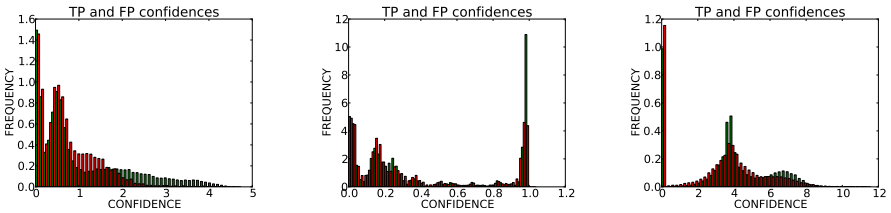
The evaluation of confidence measurements is done in two ways. The first method uses the fact that the detector is statistically more confident on true positives than on false positive detections. The idea is to accumulate the confidence values for FPs and TPs over the whole range. Then the distribution of the confidence values should be high around low values for false positive samples since the detection is not correct and it should be high for large values for true positive samples since the classifier should then be confident about its decision. To obtain the statistics, around 6000 FPs and about 6000 TPs are gathered from the de-

tector, the confidences are accumulated and the resulting histogram is equalized. For comparison we introduce further confidence values:

$$\gamma_2(h_c^T, x) = \frac{1}{T} \sum_{j=1, h_j=\omega_1}^{T} \frac{|p_j(o(x)|\omega_1)P_j(\omega_1) - p_j(o(x)|\omega_0)P_j(\omega_0)|}{p_j(o(x)|\omega_1)P_j(\omega_1) + p_j(o(x)|\omega_0)P_j(\omega_0)} \tag{6}$$

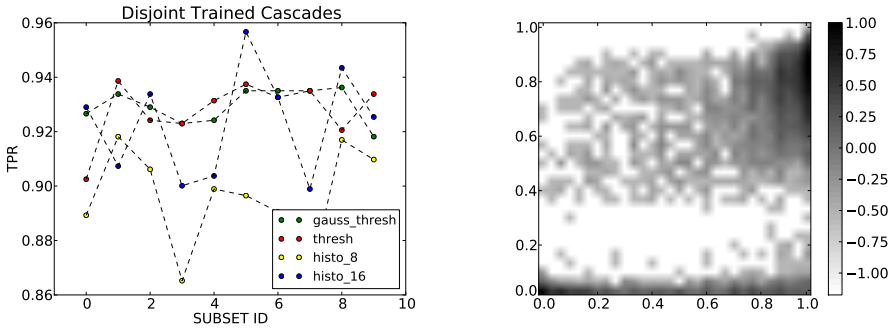
$$\gamma_3(h_c^T, x) = \frac{1}{T} \sum_{j=1, h_j=\omega_1}^{T} |o(x) - \theta| \tag{7}$$

and their corresponding statistics are shown in Fig. 6. There it can be seen



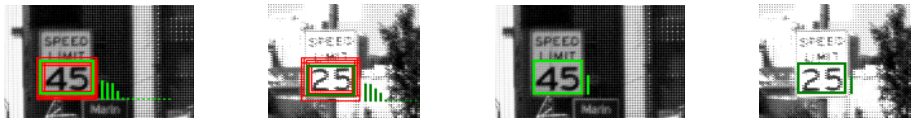
(a) γ_1 (Eq. 5) (b) γ_2 (Eq. 6) (c) γ_3 (Eq. 7)

Fig. 6. Accumulated confidence values for true positives (green bars) and false positives (red bars)



(a) (b)

Fig. 7. 7(a): Detection rates of 40 classifiers trained on 10 disjoint training sets consisting of 1300 images each. Each point corresponds to a cascade classifier. The x-axis denotes the subsets. 7(b): Statistical confidence values on the x-axis vs. the proposed confidence measure according to (Eq. 5) on the y-axis using a logarithmic scale.



(a) (b) (c) (d)

Fig. 8. Visualization of confidence values. 8(a) and 8(b): Detections are visualized as rectangles and their confidences are shown as vertical bars, sorted by decreasing order. The most confident rectangle is colored green, all others are colored red. 8(c) and 8(d): Merged windows obtained by linear combinations of confidence values.

clearly that the decision-based confidence according to Eq. (5) outperforms those computed by a distance to the threshold (Eq. (7)) and the normalized version (Eq. (6)) since the overlap between the distributions is smaller. The second evaluation method is based on a statistically exact measurement of confidence values [9] which states that for K classifiers trained on disjoint data sets an exact confidence measure can be obtained by averaging the classifiers decisions. For training of the cascades 10 disjoint sets consisting of 1300 images each were used. To generate more diverse classifiers, four classifiers with different parameters were trained per set. Then for a set of 6000 test signs the confidence values obtained by averaging the 40 classifier decisions for each classification were compared to the confidence γ_1 from Eq. (5) generated by the single cascade with ID FRQ_BC256_FS from Fig. 5 with a logarithmic plot shown in Fig. 7(b). The result shows that one cluster appears for positive and negative decisions, respectively. Fig. 7(a) shows the detection performance of each single cascade trained with the corresponding subset. The last evaluation is a visual inspection of the confidence values of detected signs. Each detection of a sign is visualized as a rectangle and its confidence value as a vertical bar. The bars are sorted by decreasing confidence and the rectangle with highest confidence is colored in green, all others in red, as shown in Fig. 8(a) and 8(b). A movie consisting of 6000 images was generated and visually inspected. The "most confident" detection was virtually always the central detection which is placed on the perfect digit block. Another movie was generated with a merging algorithm that combines multiple detection windows into one by means of confidence weighted linear combination of the rectangles coordinates which is shown in Fig. 8(c) and 8(d).

4 Conclusion

This article presents a modular approach for an adaptive Bayesian classifier cascade which was used in the real world detection of US speed limits. A method for computing Bayes decision functions, which are known to be optimal, is proposed which is applicable to cascaded structures by updating the prior probabilities and modeling the probability density functions in an adaptive way independent of the number and type of stage classifiers. In addition, different confidence values have been suggested and compared. The measure most closely related to

the Bayes decision, yields the best performance. A merging algorithm combines multiple detections into a single precise detection based on their confidence values. The overall single image detection rates aggregate to 97 % with 0.2 false positives per image running at 13 Hz and 93 % with 0.2 false positives per image running at 43 Hz for scanning the whole image.

References

1. A. Staudenmaier, U. Klauck, F. Lindner, U. Kreßel, C. Wöhler: *Resource Optimized Cascaded Perceptron Classifiers using Structure Tensor Features for US Speed Limit Detection.* 8th International Workshop on Intelligent Transportation (WIT 2011) (2011)
2. A. W. Whitney: *A direct method of nonparametric measurement selection.* IEEE Trans. Computers 20, 1100–1103 (1971)
3. B. Rasolzadeh, L. Petersson, N. Pettersson: *Response Binning: Improved Weak Classifiers for Boosting.* Intelligent Vehicles Symposium (IEEE IV) pp. 344–349 (2006)
4. H. Zaragoza, F. d'Alché-Buc: *Confidence Measures for Neural Network Classifiers.* Proceedings of the Seventh Int. Conf. Information Processing and Management of Uncertainty in Knowlegde Based Systems (1998)
5. N. Toth, B. Pataki: *On Classification Confidence and Ranking Using Decision Trees.* Intelligent Engineering Systems (INES '07) pp. 133–138 (June-July 2007)
6. R. A. Fisher: *The Statistical Utilization of Multiple Measurements.* Annals of Eugenics 8, 376–386 (1938)
7. R. O. Duda, P. E. Hart, D. G. Stork : *Pattern Classification (2nd Edition).* Wiley-Interscience (2001)
8. S.-J. K. Alessandro, A. Magnani, S. P. Boyd: *Robust Fisher Discriminant Analysis.* Advances in Neural Information Processing Systems pp. 659–666 (2006)
9. T. Cui, A. Grumpe, M. Hillebrand, U. Kreßel, F. Kummert, C. Wöhler: *Analytically tractable sample-specific confidence measures for semi-supervised learning.* Proc. Workshop Computational Intelligence pp. 171–186 (2011)
10. G. Overett, L.P.: *Large Scale Sign Detection using HOG Feature Variants.* In: Intelligent Vehicles Symposium (IEEE IV). pp. 1–6. Baden-Baden, Germany (June 2011)
11. P. Viola, M.J.: *Rapid object detection using a boosted cascade of simple features.* Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2001 1(C), I–511–I–518 (2001)
12. P. Viola, M.J.: *Robust Real Time Object Detection* . IEEE ICCV Workshop Statistical and Computational Theories of Vision July 57(2), 137–154 (2001)
13. Rosenblatt, F.: *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.* Psychological Review 65(6), 386–408 (1958)
14. Y. Freund, R.E.S.: *A decision-theoretic generalization of on-line learning and an application to boosting.* In: Proceedings of the Second European Conference on Computational Learning Theory. pp. 23–37. EuroCOLT '95, London, UK, UK (1995)