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



# Multi-Sensor Classification using a boosted Cascade Detector

Leonhard Walchshäusl and Rudi Lindl
BMW Group Research and Technology
Vehicle Sensors and Perception Systems
Hanauerstr. 46, Munich, Germany
{leonhard.walchshaeusl,rudi.lindl}@bmw.de

*Abstract*— **This paper provides a description of a new low-level feature-fusion approach for real-time object recognition utilising an arbitrary number of imaging sensors and based on a boosted cascade of simple features [1].**

**The approach is demonstrated by means of a vehicle detection system. The application utilises laser scanner responses for hypotheses generation and low-level features from both colour and far infrared images for hypotheses classification. A first evaluation shows promising results.**

## I. INTRODUCTION

Multi-sensor systems offer considerable advantage over single-sensor systems. They benefit from both an increased spatial and temporal coverage due to complementary sensor ranges and data acquisition frequencies. They show a higher reliability performance on account of built-in redundancy. In addition, they offer an improved operational robustness, because distinct physical sensing principles compensate for particular perception shortcomings. Altogether, multiple sensors are able to gather more information about the reality especially in varying environmental conditions.

The literature offers several approaches to consult more than one sensor for object recognition purposes (e.g. [2], [3]). A decision-based approach, performing the classification for each sensor separately reaches the final decision by voting or weighted averaging [4]. That output could suffer from an absence of significance, if only a few sensors are combined or the information gain of the involved sensors is unbalanced. Moreover, sub-threshold information derived from one sensor might be discarded. Consequently, the information is lost and cannot contribute to the overall result. Contrary to that, early-fusion approaches operate on low-level features derived from all available sensors. A common course of action is a static feature vector composed of features from all sensors that are judged by learners like neuronal networks or support vector machines.

Not long ago, Viola et al. [1] have introduced a well-established approach for rapid object-detection on a single video sensor. They utilise a boosted cascade of simple Haar-like image-features. Weak classifiers which are extracted from these features are selected and combined with AdaBoost [5] to form strong stage classifiers which are organised as degenerated decision tree.

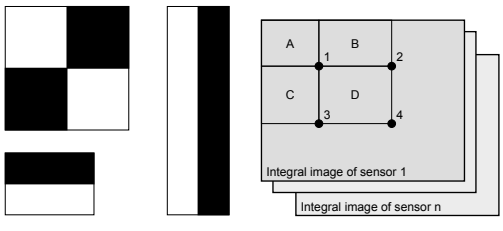
Recently, several improvements regarding computing time, detection rate and multi-class enhancements have been proposed. Huang et al. [6], [7] have employed nested degenerated decision trees of strong classifiers to hand on effective information from the current to the next tree layer. They have addressed multi-class demands for multi-view face detection by feature flipping and eight extra classifier cascades. Lienhart et al. [8] have suggested a clustering-and-splitting approach to cope with diverse clusters of object patterns. They have validated their approach for human mouth tracking and improved the computation time in comparison to a single detector cascade.

A further increase in detection performance could be reached by utilizing more than one imaging sensor. The contribution of this paper is an extension to the object recognition approach of Viola et al. in such a way that an arbitrary amount of imaging sensors can be employed in the detection process to improve performance.

## II. BOOSTED CLASSIFIERS FOR MULTI-SENSOR OBJECT DETECTION

### A. Features

The object detection approach of this paper operates on simple image features which were introduced as Haar-like features by Viola et al. [1] (see Figure 1(a)). These features are typically bound to a single sensor. In order to satisfy multi-sensor demands, an independent feature set $(f_{1,s},\ldots,f_{n,s})$ for every sensor is defined. For each feature $f_{i,s}$ taken from sensor $s$ a weak classifier $h_{i,s}$ is composed of a threshold $\theta_{i,s}$, a feature function $f_{i,s}$ (which is defined as a subtraction of adjacent rectangular image regions) and a parity function $p_{i,s}$.




(a) (b)

Fig. 1. (a) Three exemplary Haar-like features. The sum of pixel intensities which lie in the black region are subtracted from the pixel sum in the white region. (b) Imaging sensor dependent stack of integral image representation. This allows for an efficient feature calculation. The pixel sum inside area $D$ e.g. results in $4+1-(2+3)$, since the value of the integral image at position 1 is the sum of the intensities in rectangle $A$, at location 2 rectangle $A+B$, at location 3 rectangle $A+C$ and at location 4 rectangle $A+B+C+D$.

$$h_{i,s}(x_s) = \begin{cases} 1 & \text{if } p_{i,s} f_{i,s}(x_s) \leq p_{i,s}\theta_{i,s} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The discrete weak classifier acts on a sub-window $x_s$ of a single sensor image. In order to achieve an efficient feature calculation, Viola et al. introduced an integral image representation. Thereby, an arbitrary rectangular sum can be calculated with only four additions. We propose a stack of integral images to facilitate a rapid feature calculation on multiple images simultaneously (see Figure 1(b)).

Single Haar-like features are not meaningful enough to achieve low error rates. Therefore, an adaptive boosting technique AdaBoost [5] is used to form a strong classifier.

### B. Training

The training process trains a weak classifier $h_{j,s}$ for each Haar-like feature $f_{j,s}$ taken from sensor $s$ based on negative and positive sample image tuples $\vec{x}_i$. Thereafter, each boosting cycle $t$ determines the weak classifier $h_t$ which results in the smallest training error $\varepsilon_{j,s}$ considering the current weight distribution $w_{t,i}$. The weights $w_{t,i}$ of all misclassified examples are increased (3) and normalized (2). Consequently, the algorithm concentrates more and more on the difficult cases. Furthermore, each cycle selects a feature $f_{j,s}$, which offers the best separation of the training set. The strong classifier $h(\vec{x})$ is composed of all weak classifiers $h_t$ (Compare algorithm 1).

Samples which are misclassified by $h(\vec{x})$ build the input for the training of subsequent strong classifiers. The outcome of this is a degenerated decision tree, also referred to as cascade, composed of strong classifier stages. A detailed description of the cascade training can be found in [9].

### C. Classification

The resulting multi-sensor cascade operates on a region stack $R$ which is composed of rectangular regions of interest $R_s(x_s, y_s, w_s, h_s)$ deduced from every imaging sensor $s$, in which $(w_s, h_s)$ denotes the dimension of the regions and $(x_s, y_s)$ the location in the image plane, respectively. It has to be ensured, that every region $R_s$ represents the projection of the same real-world object. Accordingly, only objects entirely observable by all employed imaging sensors $s$ can be classified. Thus, only the maximum sensing overlap can generate a complete region stack and serve as a possible input area for the multi-sensor cascade.

In order to allow a high-performance evaluation of all weak classifiers, the complete region stack $R$ is converted into a stack of integral images $R_i$ composed of $R_s^i$ (see Figure 1(b)), where

$$R_s^i(x_s, y_s) = \sum_{x' \leq x_s, y' \leq y_s} R_s(x', y'). \quad (5)$$

All weak classifiers of one stage are evaluated on this stack of integral images according to the strong classifier (4).

The evaluation of the complete cascade is performed from the first stage classifier to the last stage classifier. Each positive result of one stage triggers the succeeding stage. A negative outcome at any stage immediately rejects the region stack and thus the underlying object. The object to be classified is only accepted if all stages of the cascade are passed successfully.

**Algorithm 1**: The AdaBoost algorithm for multi-sensor classifier training. Features are taken from all available sensors.

**Input**: Set of training images $(\vec{x}_i, y_i), \ldots, (\vec{x}_n, y_n)$, where $\vec{x}_i$ denotes a tuple of corresponding sensor views $x_{i,s}$ for training sample $i$ and $y_i \in \{0,1\}$ indicates negative and positive samples respectively.

**Output**: A strong classifier $h(\vec{x})$

**for** *t=1* **to** $T$ **do**

  Normalize the weights

$$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}} \quad (2)$$

  **foreach** *Feature j* **do**

    **foreach** *Sensor s* **do**

      Train a classifier $h_{j,s}$ which is restricted to use a single feature.

      $\varepsilon_{j,s} = \sum_i w_i |h_{j,s}(x_{i,s}) - y_i|$.

    **end**

  **end**

  Choose the classifier, $h_t = h_{j,s}$, with the lowest error $\varepsilon_{j,s}$. Update the weights:

$$w_{t,i} = w_{t,i} \beta_t^{1-e_i} \quad (3)$$

  where $e_i = 0$ if example $x_{i,u}$ is classified correctly, $e_i = 1$ otherwise, and $\beta_t = \frac{\varepsilon_{t,u}}{1-\varepsilon_{t,u}}$

**end**

The strong classifier is:

$$h(\vec{x}) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_{t,u}(x_u) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $\alpha_t = \log \frac{1}{\beta_t}$

## III. MULTI-SENSOR VEHICLE RECOGNITION

One important aspect for advanced driver assistance systems is a robust, accurate and reliable perception of the vehicle's environment. In order to meet these requirements a multi-sensor perception system composed of a far infrared camera (FIR), a grey-scale camera, a laser scanner and a radar has been set up to detect and track vehicles. Our detection system (see Figure 2) can be split in two parts which are explained shortly in the following. Refer to [10] for a more detailed description of the overall system.

### A. Hypotheses Generation

Based on laser scanner and radar measurements several cubic object hypotheses are initialized and tracked as stated in [10]. These two ranging sensors allow for an estimation of the objects' position and dimensions $R_{veh}(x_{veh}, y_{veh}, w_{veh}, h_{veh})$. An exact calibration of all participating sensors is mandatory. The estimation of $R_{veh}$ enables

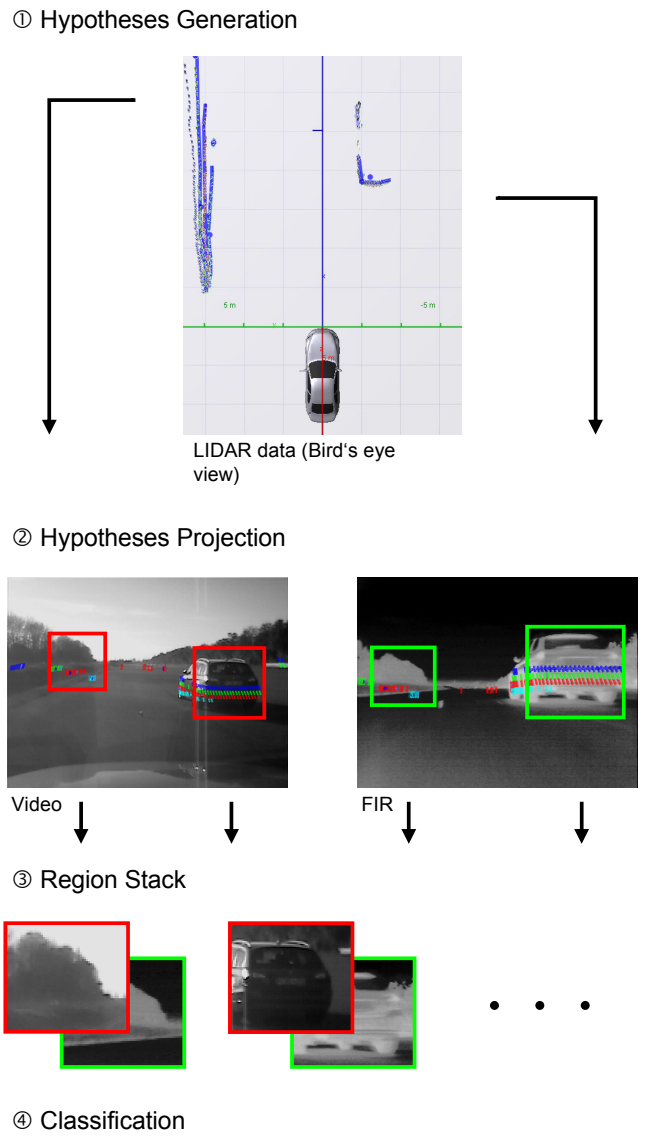



Fig. 2. Flowchart of the vehicle recognition application. (1) Bird's eye view showing LIDAR responses of a car running ahead, (2) projection of the hypotheses into the video and FIR image, (3) image stacks of regions of interests, (4) classification.

an unambiguous projection of the hypotheses onto the image planes of the FIR and grey-scale cameras resulting in regions of interest $R_{FIR}$ and $R_{grey}$. In order to compensate the uncertainties of the hypothesis's position, width and height estimation an additional confidence area is added to $R_{veh}$. This in turn leads to slightly altered regions $\tilde{R}_{FIR}$ and $\tilde{R}_{grey}$ respectively. Accordingly, the bounds of a sensor dependent scale factor range $\vec{\tau}_s$ are adapted. Finally, the region stack $R$ is composed of $\tilde{R}_{FIR}$ and $\tilde{R}_{grey}$.

A multi-sensor cascade $\Upsilon$ (see section II-C) working on a $30 \times 24$ pixel sub-window rejects an implausible region stack. In the majority of cases the dimension of $R$ significantly exceeds the window size of $\Upsilon$. Therefore, the cascade has to be scaled. For every scale factor $\zeta \in \vec{\tau}_s$ a scaled cascade $\tilde{\Upsilon} = \zeta\Upsilon$ is applied to all remaining sub-windows of $R$. The hypothesis is verified as a vehicle if at least one classification step on a sub-window successfully returns. Once a hypothesis is correctly classified a more specific vehicle model is instantiated. Using this extended object model further tracking of the vehicle is performed by common Kalman filtering. Currently, the classification is limited to rear ends of vehicles.

## IV. EXPERIMENTAL RESULTS

This sections describes some results concerning the low-level feature fusion approach. To illustrate the potential of the proposed feature fusion a multi-sensor cascade is compared to its single sensor cascades (FIR and grey-scale). In order to achieve comparable results between all involved classifiers they share the same training- and test data. As described in section II-C only overlapping sensing areas are able to serve as input for the multi-sensor cascade classifier. Thus, the training data is limited to data where this assumption holds.

The training data contains 650 manually extracted image pairs (FIR and grey-scale) which were taken from 150 different vehicle rears (see Figure 3) in front of a very similar background. All traffic scenes were recorded in the urban area with a minimal different camera perspective at a temperature of about 10 degree above zero. Furthermore, 1200 negative samples without vehicles have been collected.

Fig. 3. Small subset of the vehicle training set. Grey-scale training samples are in odd rows. Even rows contain the corresponding FIR samples.

All three cascades (FIR, grey-scale, multi-sensor) were trained until the false positive rate on the training set reached $10^{-5}$. The training of the FIR classifier finished after 10 stages and 38 involved features. The grey-scale cascade needed 12 stages and 45 features to achieve the desired error rate. Finally, the multi-sensor cascade consists of only 9 stages and 25 different features (13 from FIR and 12 from grey-scale) since it selected the most distinctive features (see Figure 4) from both sensors to reach the desired error rate.

With regard to the detection rates, the single sensor FIR cascade achieves better results than the grey-scale cascade as long as only a few features are utilised. That is due to

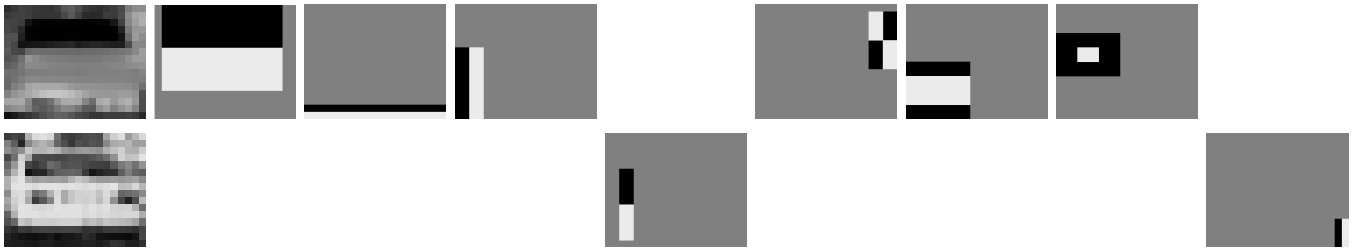


Fig. 4. The first eight Haar-like features (6 from FIR and 2 from the grey-scale sensor) in the order of their selection (from left to right) during the training process for the multi-sensor cascade (FIR features depicted in the top row and grey-scale features in the bottom row). The first column contains typical FIR and grey-scale sample images taken from the training set. Obviously, feature one, e.g., corresponds to the "cold" rear window and the "warm" vehicle's rear end.

the fact, that on the one hand textureless FIR images only cover coarse details of vehicle rear ends (see first column in Figure 4), which can be easily distinguished from most of the background objects. On the other hand the feature richness in consequence of the detailed textured grey-scale representation becomes more and more important in order to reject challenging background objects. Accordingly, the grey-scale cascade measures up to the FIR cascade and even gets ahead as more features are added to the single-sensor cascades (see Figure 5).

The entangled training process of the multi-sensor cascade selects the most distinctive feature, independent of the sensor. Thus, a seamless transition from coarse FIR to fine grey-scale features is achieved automatically. Altogether, the multi-sensor cascade slightly outperforms both single-sensor classifiers concerning detection rate and computational performance as it achieves better classification rates with fewer features (compare Figure 5 and 6).

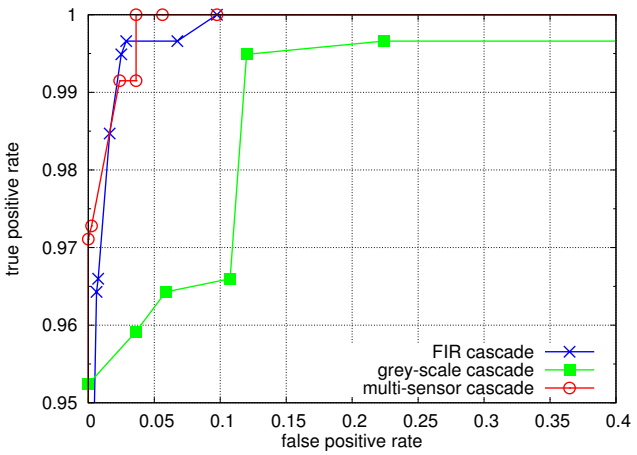



Fig. 5. ROC curve of the single-sensor cascades as well as of the multi-sensor cascade. The ROC curve has been generated by varying the stages of the cascades from one to the total amount of trained stages.

## V. CONCLUSIONS AND FUTURE WORK

### A. Conclusions

This paper presented a novel multi-sensor extension to the object recognition approach of Viola et al. [1]. The Haar-like feature space and the AdaBoost [5] training algorithm are extended to act on multiple imaging sensors. The classification process, which is demonstrated by means of vehicle hypothesis verification, operates on a stack of image regions which are extracted from every sensor. Only the maximum sensing overlap can generate a complete region stack and serve as possible input for a multi-sensor cascade. A first experimental evaluation has shown that the introduced approach leads to slightly higher classification rates compared to the results achieved by a single imaging sensor.

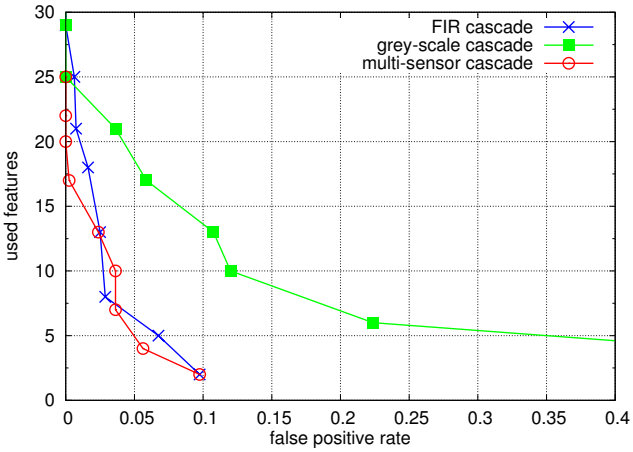



Fig. 6. The number of used features of the trained cascades in comparison to the false positive rate.

### B. Future Work

Currently, the sensor fusion approach is limited to imaging sensors. Further research is needed to evaluate the suitability of the low-level feature fusion with respect to the integration of different sensor principles. Laser scanner intensities for example could be added to the feature space in order to further enhance the detection quality. Avoiding sensor failures is crucial for the detection performance as the cascade classifier relies heavily on every single sensor input. Therefore, these drop outs should be considered and treated separately. Finally, further evaluation has to be accomplished to acquire recognition rates on a more versatile training set and to compare the performance of the introduced multi-sensor feature selection to decision based late-fusion approaches.

## VI. ACKNOWLEDGMENTS

The low-level feature-fusion approach for real-time object recognition presented in this publication is part of the results achieved in the *COMPOSE*-project which is an application-driven subproject of the *PReVENT* Integrated Project, an automotive initiative co-funded by the European Commission's Sixth Framework Programme for active road safety.

## REFERENCES

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proc. CVPR*, vol. 1, pp. 511–518, 2001.

[2] R. Chellappa, G. Qian, and Q. Zheng, "Vehicle Tracking using Acoustic and Video Sensors," *Proc. IEEE Intl Conf. Acoustics, Speech, and Signal Processing*, pp. 793–796, 2004.

[3] C. Stiller, J. Hipp, C. Rossig, and A. Ewald, "Multisensor obstacle detection and tracking," *Image and Vision Computing*, vol. 18, no. 5, pp. 389–396, 2000.

[4] Y. Chen, C. Shahabi, and G. Burns, "Two-Phase Decision Fusion Based On User Preference," *Brain Res*, vol. 60, pp. 2–0, 2004.

[5] Y. Freund and R. Schapire, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, 1999.

[6] C. Huang, H. Al, B. Wu, and S. Lao, "Boosting nested cascade detector for multi-view face detection," *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 2, 2004.

[7] B. Wu, H. Ai, C. Huang, and S. Lao, "Fast rotation invariant multi-view face detection based on real Adaboost," *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pp. 79–84, 2004.

[8] R. Lienhart, L. Liang, and A. Kuranov, "A detector tree of boosted classifiers for real-time object detection and tracking," *IEEE Int. Conf. on Multimedia and Systems (ICME2003)*, 2003.

[9] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 1, no. 2, 2002.

[10] R. Lindl and L. Walchshäusl, "Three-Level Early Fusion for Road User Detection," *PReVENT Fusion Forum e-Journal*, no. 1, pp. 19–24, 2006.